FROM GENES TO SPECIES: NOVEL INSIGHTS FROM METAGENOMICS

EDITED BY: Eamonn P. Culligan and Roy D. Sleator
PUBLISHED IN: Frontiers in Microbiology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-975-4
DOI 10.3389/978-2-88919-975-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

FROM GENES TO SPECIES: NOVEL INSIGHTS FROM METAGENOMICS

Topic Editors:
Eamonn P. Culligan, Cork Institute of Technology, Ireland
Roy D. Sleator, Cork Institute of Technology, Ireland

The majority of microbes in many environments are considered "as yet uncultured" and were traditionally considered inaccessible for study through the microbiological gold standard of pure culture. The emergence of metagenomic approaches has allowed researchers to access and study these microbes in a culture-independent manner through DNA sequencing and functional expression of metagenomic DNA in a heterologous host. Metagenomics has revealed an extraordinary degree of diversity and novelty, not only among microbial communities themselves, but also within the genomes of these microbes. This Research Topic aims to showcase the utility of metagenomics to gain insights on the microbial and genomic diversity in different environments by revealing the breadth of novelty that was in the past, largely untapped.

Citation: Culligan, E. P., Sleator, R. D., eds. (2016). From Genes to Species: Novel Insights from Metagenomics. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-975-4

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 03 August 2016
doi: 10.3389/fmicb.2016.01181

Editorial: From Genes to Species: Novel Insights from Metagenomics

*Eamonn P. Culligan * and Roy D. Sleator **

Department of Biological Sciences, Cork Institute of Technology, Cork, Ireland

Keywords: metagenomics, functional metagenomics, metatranscriptomics, next generation sequencing, microbiome

The Editorial on the Research Topic

From Genes to Species: Novel Insights from Metagenomics

The majority of microbes in many environments are considered "as yet uncultured" and were traditionally considered inaccessible for study through the microbiological gold standard of pure culture. The emergence of metagenomic approaches has allowed researchers to access and study these microbes in a culture-independent manner through DNA sequencing and functional expression of metagenomic DNA in a heterologous host. Metagenomics has revealed an extraordinary degree of diversity and novelty, not only among microbial communities themselves, but also within the genomes of these microbes. Metagenomic analysis can involve sequence-based or functional approaches (or a combination of both). The continuous improvements to DNA sequencing technologies coupled with dramatic reductions in cost have allowed the field of metagenomics to grow at a rapid rate. Many novel insights on microbial community composition, structure, and functional capacity have been gained from sequence-based metagenomics. Functional metagenomics has been utilized, with much success, to identify many novel genes, proteins, and secondary metabolites such as antibiotics with industrial, biotechnological, pharmaceutical, and medical relevance. Future improvements and developments in sequencing technologies, expression vectors, alternative host systems, and novel screening assays will help advance the field further by revealing novel taxonomic and genetic diversity. This Research Topic aims to showcase the utility of metagenomics to gain insights on the microbial and genomic diversity in different environments by revealing the breadth of novelty that was in the past, largely untapped. This Research Topic comprises 19 submissions from experts in the field and covers a broad range of themes and article types (Review, Methods, Perspective, Opinion, and Original Research articles). We have broadly grouped the articles under four themes; functional metagenomics, targeted metagenomics, sequence-based metagenomics, and host-associated.

We begin with a number of articles focusing on functional metagenomics. A review by Coughlan et al. gives an overview of metagenomics and focuses on the utility of functional metagenomics for the discovery of proteins and antimicrobial compounds with relevance to the food and pharmaceutical industries. Continuing this theme Uchiyama et al. report the discovery of a glucose-tolerant β-glucosidase from screening ∼10,000 clones from a metagenomic library created from Kusaya gravy (a traditional Japanese fermentate made from fish). β-glucosidases are often sensitive to glucose inhibition, therefore glucose-tolerant variants are desirable to improve enzymatic efficiency. Mirete et al. also used a functional metagenomic approach to identify novel salt tolerance genes from brine and rhizosphere-associated communities in a hypersaline saltern. A number of the genes had not previously been known to play a role in salt tolerance. This approach demonstrates one of the main advantages of functional metagenomics; assigning function to unknown genes or new functions to annotated genes.

OPEN ACCESS

Edited by:
Ludmila Chistoserdova, University of Washington, USA

Reviewed by:
Susannah Green Tringe, U.S. Department of Energy Joint Genome Institute, USA

****Correspondence:***
Eamonn P. Culligan
eamonn.culligan@cit.ie
Roy D. Sleator
roy.sleator@cit.ie

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *10 June 2016*
Accepted: *18 July 2016*
Published: *03 August 2016*

Citation:
Culligan EP and Sleator RD (2016) Editorial: From Genes to Species: Novel Insights from Metagenomics. Front. Microbiol. 7:1181. doi: 10.3389/fmicb.2016.01181

As with any technology, there are advantages and disadvantages. In their Perspective article, Lam et al. present the main challenges and potential solutions associated with functional metagenomics. Biases may be introduced at different stages of the process, from DNA extraction, library construction, cloning, and choice of expression vector and heterologous host. The authors discuss advances to improve each step and provide helpful comments based on their own considerable experience. They also present data, which suggests cloning bias is occurring at the level of individual operational taxonomic units (OTUs). Finally, it is suggested that moving beyond *Escherichia coli* as a cloning host will increase the diversity of hits from functional screens. An additional issue associated with metagenomics it that there is still a dearth of functional information for a large proportion of protein families; a problem which is increasing due to the enormous amounts of sequencing data that continues to be generated and deposited in databases. Ufarté et al. review sequence-based and activity screening approaches in metagenomics to assign functions to novel genes. The authors also discuss recent developments in microfluidic approaches for ultra-high-throughput screening, where up to 1 million clones can be assessed in a single day.

On a similar theme, Suenaga discusses the role of "targeted" metagenomics in compiling specific groups of enzymes to study their adaptive evolution, and echo the importance of the microfluidics approach mentioned above, as well as technologies such as cell compartmentalisation, flow cytometry, and fluorescent cell sorting in the future for high-throughput screening. Trindade et al. reviews how targeted metagenomics may be used to identify natural products from marine organisms and microbes, which have the potential to treat human disease. The authors explain why functional screening approaches have been largely unsuccessful in this regard. However, using targeted metagenomic approaches, guided by well-known structural and functional characteristics of natural products, a number of clinically relevant compounds have been successfully isolated; including several potent anti-cancer and anti-fungal compounds such as, bryostatins, patellazoles, polytheonamides, ecteinascidin 743, pederin, psymberin, and calyculin A. Dziewit et al. describe a targeted approach to detect methanogenic archaea. Methanogenic archaea are important community members of many diverse environments including peatlands, freshwater sediments, and the intestinal tract of animals and humans. Many members have proved difficult to culture and previous studies have relied on metagenomic, 16S rDNA, and *mcrA* gene sequencing. The authors present a methods paper detailing the development of a number of sets of degenerate primers for methanogenic archaea based on the *mcrB, mcrG, mtbA,* and *mtbB* genes, which are involved the process of methanogenesis. These novel molecular markers will provide additional information on the biology, diversity, and phylogenetic relationships of these organisms.

Sequence-based metagenomics can provide unprecedented information on composition, diversity, and functional capacity of microbial communities. One of the main challenges associated with sequence-based metagenomics is *de novo* assembly of reads following sequencing. Howe et al. outline some of the main issues with such assemblies. The authors also include a unique iPython notebook tutorial that allows readers to follow the steps of this process and execute assembly of a mock metagenome.

Wemheuer et al. assessed the effect of phytoplankton *Phaeocystis globosa* algal bloom on microbial communities in the North Sea, using metagenomic, and metatranscriptomic approaches. Changes in community composition were identified inside the bloom in comparison to outside the bloom, most likely due to changing nutrient availabilities during algal bloom growth. Indeed, metatranscriptomic data revealed changes in gene expression in response to the bloom. Genes for incorporation of leucine and isoleucine were significantly upregulated and many genes encoding transposases were overexpressed inside the bloom. It is suggested that genome rearrangement via expression of transposases enables increased stress resistance and enhanced adaptation to changing environmental conditions.

Using a similar metagenomic and metatranscriptomic approach, He et al. investigated microbial sulfur cycling and carbon and nitrogen metabolism in a hydrothermal chimney. The genes identified were used to unravel potential pathways for sulfur and carbon metabolism, which play an important role for survival in this environment. Furthermore, γ-proteobacteria, and ε-proteobacteria are proposed as community members capable of denitrification, using electrons generated from oxidation of reduced sulfur. Bargiela et al. report a bioinformatic analysis of a previously published metagenomic dataset to identify genes enriched in a crude-oil-contaminated marine environment. Specifically, genes enriched following ammonium and uric acid (bio-stimulants) treatment were identified. Differences in taxonomic composition, presence of genes and metabolic pathway constituents and biodegradation were noted following bio-stimulant treatment. Both bio-stimulants appeared to increase the capacity for microbial degradation of crude oil.

Rosario et al. present research on the area of viral metagenomics. Twenty-seven novel CRESS-DNA (circular Rep-encoding ssDNA) viruses were identified and sequenced from marine invertebrates, some of which may represent a novel family. Intrinsically disordered regions (IDRs) within proteins were also investigated. IDRs lack rigid structure and allow the protein to exist in different states, which may allow multi-functionality in such proteins. Different IDRs are commonly found in proteins encoded by CRESS-DNA viruses and may be useful to characterize divergent structural proteins, though at present the importance of the different IDRs remains to be confirmed.

Bao et al. used strand-specific metatranscriptomics in a novel way to identify anti-sense transcription among members of the human gut microbiota. Anti-sense RNAs are encoded on the opposite strand of DNA from the mRNA transcript and may have important regulatory functions in gene expression. Most of the species tested displayed anti-sense transcription (ranged from 0 to 38.5% for protein coding genes between different species). Interestingly, the functional category of genes most over-represented with anti-sense transcription included prophage-associated and transposon genes.

Metagenomic approaches have provided a wealth of information about the microbes on and in the human body

(microbiota) and their potential role in human health and disease. Belizario and Napolitano review current information on a number of human microbiomes (gut, oral, skin, placental), and discuss how targeting and mining the microbiota is opening a new area of microbiome-based therapeutics. For example, the use of probiotics and prebiotics, phage therapy and CRISPR technology are exciting areas of research, while faecal microbiota transplantation (FMT) has shown promising results for the treatment of *Clostridium difficile* infection (CDI). Kinumaki et al. use metagenomic sequencing to profile the gut microbiota of patients with Kawasaki disease (KD), an acute childhood illness characterized by vascular inflammation, which is a leading cause of acquired heart disease. The precise cause of KD is unknown, but a possible microbial influence has been suggested to play a role in its pathogenesis. Metagenomic sequencing revealed differences in gut microbiota composition between KD patients during acute and non-acute phases of the disease. In particular, a number of species from the genus *Streptococcus* were significantly increased during the acute phase of KD. The authors suggest that species of *Streptococcus* may play a role in KD pathogenesis, but more research is required to conclusively demonstrate a causal link.

Brito and Alm, review strain-level tracking of microbes using metagenomics. The authors state that transmission has primarily focused on pathogenic organisms, but very little is known about transmission of commensal species. With significant emerging evidence for the roles that commensal microbes play in human health and disease, the ability to track, and differentiate microbes at the strain level is important. Metagenomic sequencing provides advantages over 16S rDNA sequencing in this regard for example, and long-read sequencing (e.g., Oxford Nanopore's MinION) and proximity ligation (enables detection of protein-protein and protein-DNA interactions, as well as post-translational modifications) may help improve this in the future. The ability to track strain-level transmission will be key to monitor live microbial therapeutics and the biological containment of engineered microorganisms, while longitudinal studies could reveal how transmission affects daily or intermittent changes to the microbiota.

Voss et al. propose the "pawnobiome" as a "subset of the microbiome that is purposefully managed for manipulation of the host phenotype, which includes individual microbes named pawnobes." Different from the hologenome theory of evolution, where the unit of selection is the holobiont (i.e., both the host and its associated microbiota); the pawnobiome can evolve independently and faster than the host and is not wholly reliant on host survival. It is also proposed that the pawnobiome can affect host phenotype and can be independently/artificially selected; thus having implications for health and disease, biotechnology, and evolutionary biology.

Finally, Simm et al. present an analysis of the core- and pan-genome of cyanobacteria. Using 58 sequenced cyanobacterial genomes, the authors identify 559 genes that define the core-genome. Furthermore, 3 genes specific to thermophilic cyanobacteria and 57 genes specific to heterocyst-forming cyanobacteria were also defined. Additionally, outer membrane β-barrel proteins were investigated. It was found that most of these proteins are not globally conserved and exhibit strain specificity, indicating cyanobacteria have evolved individual strategies for environmental adaptation and interaction.

Overall, this Research Topic showcases a broad range of articles which illustrate the utility of both sequence-based and functional metagenomic approaches to investigate what were once inaccessible and undiscovered areas of microbial genomics, physiology, evolution, and ecology. Future advances in metagenomic research and technology will undoubtedly reveal further novelty and diversity from genes to species and beyond.

AUTHOR CONTRIBUTIONS

EC and RS co-edited the Research Topic. Both authors wrote, edited, and approved the final version of the Editorial.

ACKNOWLEDGMENTS

We thank the Frontiers Editorial Office for their assistance in completing this Research Topic, the reviewers for their time and expertise and the authors for their submissions. EC is funded by an Irish Research Council Government of Ireland Postdoctoral Fellowship (GOIPD/2015/53). RS is Coordinator of the EU FP7 project ClouDx-i.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 30 June 2015
doi: 10.3389/fmicb.2015.00672

Biotechnological applications of functional metagenomics in the food and pharmaceutical industries

Laura M. Coughlan[1], *Paul D. Cotter*[1,2], *Colin Hill*[2,3] *and Avelino Alvarez-Ordóñez*[1]*

[1] Teagasc Food Research Centre, Cork, Ireland, [2] Alimentary Pharmabiotic Centre, Cork, Ireland, [3] School of Microbiology, University College Cork, Cork, Ireland

Microorganisms are found throughout nature, thriving in a vast range of environmental conditions. The majority of them are unculturable or difficult to culture by traditional methods. Metagenomics enables the study of all microorganisms, regardless of whether they can be cultured or not, through the analysis of genomic data obtained directly from an environmental sample, providing knowledge of the species present, and allowing the extraction of information regarding the functionality of microbial communities in their natural habitat. Function-based screenings, following the cloning and expression of metagenomic DNA in a heterologous host, can be applied to the discovery of novel proteins of industrial interest encoded by the genes of previously inaccessible microorganisms. Functional metagenomics has considerable potential in the food and pharmaceutical industries, where it can, for instance, aid (i) the identification of enzymes with desirable technological properties, capable of catalyzing novel reactions or replacing existing chemically synthesized catalysts which may be difficult or expensive to produce, and able to work under a wide range of environmental conditions encountered in food and pharmaceutical processing cycles including extreme conditions of temperature, pH, osmolarity, etc; (ii) the discovery of novel bioactives including antimicrobials active against microorganisms of concern both in food and medical settings; (iii) the investigation of industrial and societal issues such as antibiotic resistance development. This review article summarizes the state-of-the-art functional metagenomic methods available and discusses the potential of functional metagenomic approaches to mine as yet unexplored environments to discover novel genes with biotechnological application in the food and pharmaceutical industries.

Keywords: functional metagenomics, industrial applications, food, pharmacological, catalysts, bioactives, antimicrobials

OPEN ACCESS

Edited by:
Eric Altermann, AgResearch Ltd., New Zealand

Reviewed by:
William John Kelly, AgResearch Ltd., New Zealand
Diego Mora, University of Milan, Italy

***Correspondence:**
Avelino Alvarez-Ordóñez, Teagasc Food Research Centre, Moorepark, Fermoy, Cork, Ireland
avelino.alvarez-ordonez@teagasc.ie

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *26 April 2015*
Accepted: *19 June 2015*
Published: *30 June 2015*

Citation:
Coughlan LM, Cotter PD, Hill C and Alvarez-Ordóñez A (2015) Biotechnological applications of functional metagenomics in the food and pharmaceutical industries. Front. Microbiol. 6:672. doi: 10.3389/fmicb.2015.00672

Introduction

Recent advances in molecular microbiology have revealed that the microbial world extends far beyond what can be revealed by traditional microbiological techniques. Environments once believed to be devoid of life have now been shown to support the growth of microbes. As a consequence, it is now accepted that microorganisms thrive throughout nature, and that at least some microorganisms can be found in almost all known environments. This is due to the fact that microbial life has adjusted to survive under a wide range of harsh or unaccommodating conditions,

resulting in a variety of diverse microorganisms adapted to specific niches. This review article explores the molecular methods that can provide access to these specially adapted microbes and, more specifically, their potentially useful genes/molecules and outlines how these approaches can be harnessed by the food and pharmaceutical industries.

Traditional microbiology generally involves obtaining a pure culture as a major step in any study. However, it is estimated that standard laboratory culturing techniques provide information on 1% or less of the bacterial diversity in a given environmental sample (Torsvik et al., 1990). This is most noticeable in what is known as the plate count anomaly, i.e., the discrepancy between the numbers of microorganisms detected by microscopy and the numbers obtained from pure colony counts of cultivated samples (Staley and Konopka, 1985). Although significant advances have been recently made in culturing as-yet-uncultured microbes, e.g., Ling et al. (2015), culture-independent techniques present a more promising effort to access the genetic information contained within the vast number of species in the environment.

Metagenomics presents a molecular tool to study microorganisms *via* the analysis of their DNA acquired directly from an environmental sample, without the requirement to obtain a pure culture. With this technology, the DNA of microorganisms in a population is analyzed as a whole. Sequencing and analysis of total metagenomic DNA can provide information about several aspects of the sample, allowing one to better characterize the microbial life in a given environment. It can not only reveal the identity of species present but also can provide insight into the metabolic activities and functional roles of the microbes present in a given population (Langille et al., 2013). Expression of the genetic information from an environmental sample in a routinely culturable surrogate host can also overcome in part the barriers faced when dealing with as yet uncultured bacteria. The coupling of this approach with function-based screening of the subsequent colonies to uncover a desired activity that has been conferred onto the host by the inserted environmental DNA in a functional metagenomics approach is a powerful technique for the discovery of novel functional genes from uncultured microorganisms.

In this review article, functional metagenomics is discussed as an emerging molecular technique with potential applications in industrial settings. An overview of the current methodological strategies employed for functional metagenomic analysis of microbial populations, with emphasis on the use of phenotypic-based metagenomic screens for the discovery of novel small molecules, enzymes, and bioactives is provided. The applications of such compounds to the food and pharmaceutical industries are discussed, while highlighting recent successes in this area.

Functional Metagenomics: Methodological Approaches

Sequencing-based Strategies

Metagenomic analyses begin with the isolation of microbial DNA from an environmental sample. The acquired metagenomic DNA specimen should be as pure and of as high quality as possible, and should accurately represent all species present both qualitatively and quantitatively. Direct sequencing of extracted metagenomic DNA, followed by appropriate bioinformatics analyses, can facilitate the elucidation of the functional traits of microorganisms colonizing particular environments (**Figure 1**).

The initial break from culture-dependent to culture-independent approaches for the microbiological analysis of an environmental sample involved the sequencing of genes encoding microbial ribosomal RNAs (rRNAs). Highly conserved primer binding sites within the bacterial 16S rRNA gene facilitate the amplification and sequencing of hypervariable regions that can provide species-specific signature sequences useful for bacterial identification in an environmental sample (Lane et al., 1985). This technology enables microbiologists to determine phylogenetic relationships between unculturable bacteria and assess and quantify the microbial consistency of a sample. In addition, through 16S rRNA gene sequencing of a metagenomic sample, a functional profile of the bacteria present in a given environment can also be obtained. Information regarding the functional roles of already studied bacterial species is available in database archives, including both cultured bacteria whose functional proteins have been extensively characterized as well as functions assigned to bacterial proteins produced by uncultured bacteria through previous metagenomic studies. Once a member of a previously described bacterial family has been identified in an environmental sample, or an appropriate closest known relative has been appointed, phylogenetic analysis may assign predicted functions to an identified bacterial species by referring to the functional information available regarding that particular taxonomic group. This process can be applied to potentially most, if not all, of the different bacterial species encountered in a sample and therefore community roles can be predicted for the microbes dwelling in the sampled niche without the need for shotgun sequencing (described below). Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt) is a computational approach developed by Langille et al. (2013) which can be used to predict the functional properties of microorganisms in a metagenomic sample from characterized relatives in available databases using 16S rRNA sequencing data. By quantifying the individual species abundance in a sample and, in doing so, quantifying the function(s) assigned to that family, PICRUSt can predict the overall functional composition of the community. Keller et al. (2014) also explored this concept through a combinatorial approach of 16S rDNA metabarcoding and single genomics for assessing the compositional and functional diversity of a microbial community. Although these authors were successful in validating their method, this innovative technique requires optimization prior to its introduction into larger and more challenging projects. Microbial eukaryotic communities may also be studied through similar strategies. Eukaryotic-specific primers homologous to the bacterial 16S rRNA can be used to target eukaryotic microbes present in an environmental sample. Bates et al. (2012) used bar-coded pyrosequencing of 18S rRNA to investigate the eukaryotic components of three different lichens, identifying members of the Alveolata, Metazoa, and Rhizaria taxonomic clades. Non-coding DNA located between the small

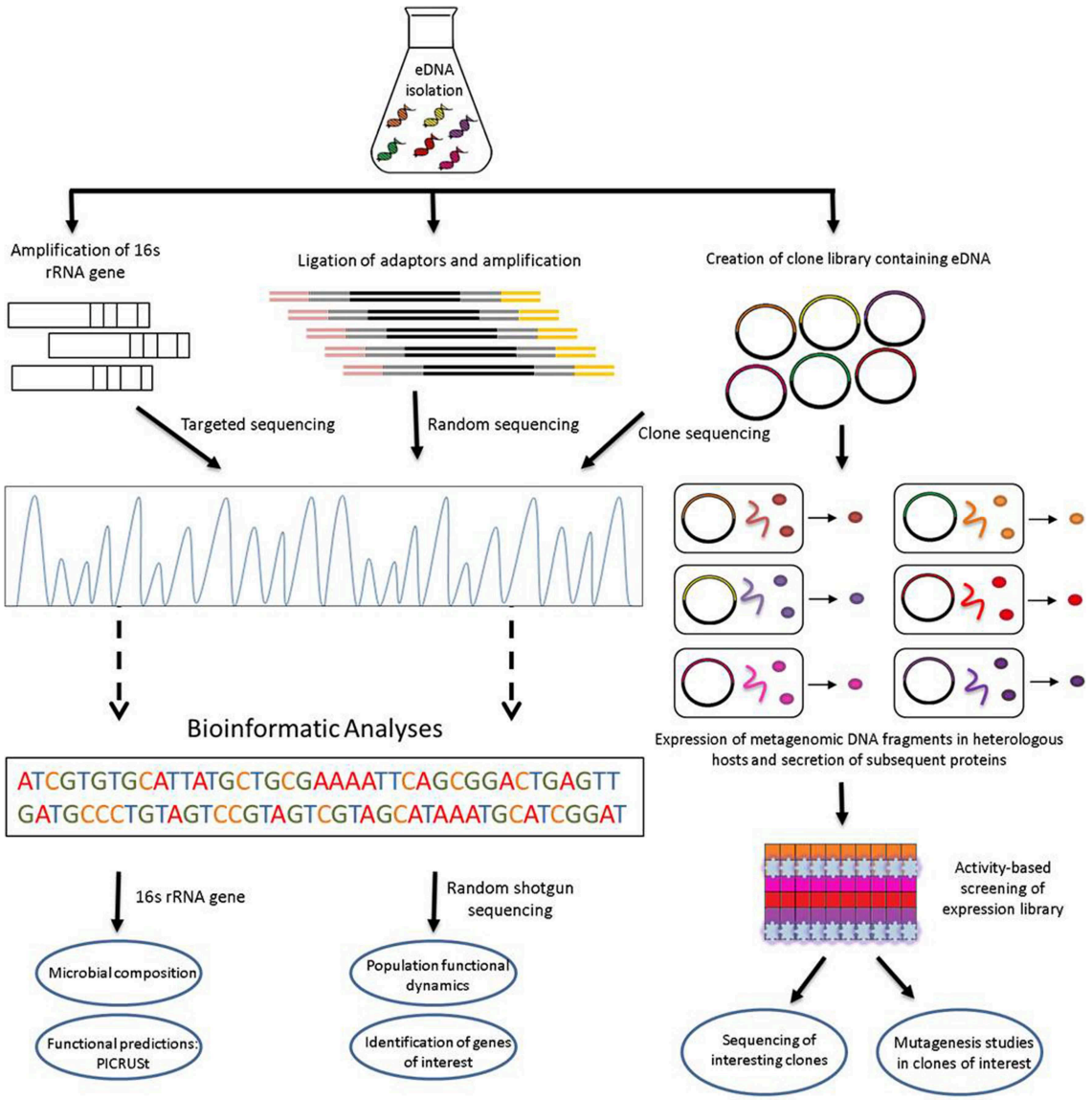

FIGURE 1 | Schematic view of functional metagenomic strategies for the identification of novel biocatalysts and bioactives from environmental DNA.

and large subunit eukaryotic rRNA genes, known as the Internal Transcribed Spacer (ITS) regions, are also targeted as a universal DNA marker in Fungi (Schoch et al., 2012). The environmental virome has also been explored through metagenomics by the coupling of sequence-independent amplification of viral nucleic acids with next generation sequencing technologies (Smits and Osterhaus, 2013), particularly in the areas of epidemiology and diagnostics. In addition, genes similar to those of metabolic cells, known as auxiliary metabolic genes (AMGs), have been discovered in viruses (reviewed by Rosario and Breitbart, 2011) and may have potential in the search for industrially relevant enzymes and bioactives.

Environmental DNA random shotgun sequencing, where total metagenomic DNA is sequenced, assembled and annotated, has been shown to be a more useful tool which may be used to analyse at a molecular/species level the metagenome of an environmental sample. In this instance, the functional potential of a microbial population is revealed by directly sequencing the environmental DNA rather than predicting its functional potential based on 16S rRNA data. Some examples of large scale metagenomic studies involving shotgun sequencing are those carried out by Venter et al. (2004), who characterized the microbial population of the Sargasso Sea identifying 1.2 million previously undescribed genes including the first assignment of rhodopsin-like photoreceptors to bacterial species, Warnecke et al. (2007), who analyzed the hindgut paunch microbiota of a *Nasutitermes* species of wood-feeding termite revealing unprecedented diversity of the microbial community

and identifying novel genes involved in cellulose and xylan hydrolysis, Oh et al. (2014), who analyzed the microbial content and subsequent functional capacity of the healthy human skin microbiome through shotgun metagenomic sequencing, and Hess et al. (2011), who deep sequenced 268 gigabases of metagenomic DNA obtained from the microbiota of cow rumen unveiling carbohydrate active genes encoding enzymes capable of degrading biomass, a desirable ability in the development of biofuels as a renewable energy source. Random sequencing of shotgun metagenomic DNA may reveal genes of interest, the probable phylogeny of which can be inferred through searches for homology in non-redundant databases, usually via Basic Local Alignment Search Tool (BLAST) analysis. Thus, random sequencing has the potential to identify the presence of already known genes, with reported beneficial functions, or their homologs in an uncultured microorganism, which can provide additional advantages and improve the functionality of in-use proteins/enzymes/catalysts, e.g., the new variant/homolog may encode a protein that is capable of carrying out a specific catalytic or metabolic function and may also be tolerant to an extreme environment habitually encountered in industrial processes. This approach is also useful for the study of the population dynamics of a community, including genomic evolution (Chandler et al., 2014; Kay et al., 2014) and the distribution and redundancy of functions throughout the community (Mendes et al., 2015).

Nevertheless, the sequence-based approaches to analysing environmental samples are limited to the study and identification of genes and DNA sequences homologous to those that are already known. Consequently, the possibility of using sequence-based methods for the discovery of proteins encoded by novel sequences is restricted. Phenotypic-based screening of constructed metagenomic expression libraries, described in the next section of the manuscript, is better suited to the unearthing of previously undescribed proteins and small molecules.

Phenotypic-based Strategies

Functional metagenomic analyses can be carried out on metagenomic libraries *via* the isolation and purification of DNA from an environmental sample, cloning of the DNA into a suitable vector, heterologous expression of the insert vector containing environmental DNA fragments in a suitable surrogate host (usually *Escherichia coli*), and analysis of subsequent transformants by either sequencing- or phenotypic-based approaches, or both (**Figure 1**). Screening of metagenomic libraries through phenotypic-based approaches is carried out to detect the expression of a particular phenotype conferred on the host by inserted DNA. Screening is usually performed on multiple clones simultaneously on a fixed matrix in which the entire group is assayed with an appropriate indicator to reveal the presence of a phenotypically relevant clone. Such assays require the functional protein to be secreted from the host cell to allow for extracellular detection. Metagenomic clones may be grown on specific indicator media, to allow visual identification of an active clone, e.g., hemolytic activity on blood agar (Rondon et al., 2000), lipolytic activity (Henne et al., 2000), etc. In other occasions, the presence of zones of inhibition in soft agar overlay assays using indicator microorganisms can reveal inhibitory or antimicrobial agents produced by an active clone (Tannieres et al., 2013; Iqbal et al., 2014). Libraries may also be screened based on selection approaches. In these circumstances only the clones onto which the activity of interest has been conferred by the metagenomic DNA insert will grow or survive. Selections include for instance the ability to metabolize a given substrate as a clone's sole carbon source (Entcheva et al., 2001), the ability to resist a potent antimicrobial agent (Donato et al., 2010) or the ability to grow in the presence of a lethal concentration of a heavy metal (Staley et al., 2015).

An alternative option for the identification of novel genes, the Substrate-Induced Gene EXpression screening (SIGEX), was developed by Uchiyama et al. (2005). It relies on the principle that catabolic gene expression is generally induced by a specific substrate or metabolite of catabolic enzymes and is controlled by regulatory elements situated close to these genes. With SIGEX, the environmental DNA inserts are fused with a reporter gene encoding green fluorescent protein (*gfp*) on an operon-trap vector and induced by a target substrate. SIGEX is combined with fluorescent-activated cell sorting (FACS) for the high-throughput selection of GFP-expressing clones. Additionally, the protocol eliminates the incorporation of clones containing self-ligated plasmids and those that are constitutively expressing GFP. Despite some limitations with regard to the applications of this method (reviewed by Yun and Ryu, 2005), SIGEX is an efficient process for the identification of novel catabolic substrate-induced genes. Uchiyama and Miyazaki (2010) went on to expand the capabilities of the SIGEX protocol and developed a reporter assay system for the screening of metagenomic libraries for enzymatic function called Product-Induced Gene EXpression (PIGEX). The system uses a transcriptional activator, which is sensitive to the product of the desired reaction, placed upstream of a *gfp* gene insert. Should a clone possess the activity of interest, upon exposure to an appropriate substrate, the product of this reaction activates transcription of the chosen transcriptional regulator and in turn *gfp* causing the clone to fluoresce, allowing easy detection of positive clones. Pooja et al. (2015) identified through PIGEX a periplasmic α-amylase from a cow dung-derived metagenomic library by isolating an active clone that fluoresced in response to a maltose substrate.

Despite the potential usefulness of such systems, phenotypic-based functional metagenomic approaches face a number of complications, to which potential resolutions are currently being devised. To successfully identify a useful gene or protein candidate a series of sequential steps in the cloning and screening process must occur adequately and effectively. Transcription of the entire gene, translation of its mRNA, correct protein folding, and secretion of the active protein from the surrogate host must all be achieved before functional screening even begins. Suitable and efficient screening methods must also be applied to detect the presence of an interesting gene within the metagenomic library. As the probability of identifying a metagenomic clone, among possibly thousands of others, with a specific desired activity is low (Uchiyama and Miyazaki, 2009), high-throughput screening (HTS) protocols may improve the chances of obtaining an active clone, by allowing higher numbers of clones to be screened simultaneously. An obstacle occurring at any of these stages may

result in the overlooking of an interesting clone which might have been detected under the correct circumstances.

One aspect of the methodological approach that can be particularly challenging relates to expressing DNA fragments isolated from microorganisms native to diverse and exotic environments in a relatively domesticated host such as *E. coli* (Banik and Brady, 2010). Even if the foreign DNA is successfully transcribed and translated (perhaps due to the presence of DNA regulatory elements placed on the vector), the correct chaperones required for proper protein folding in the original species may be absent from *E. coli*. A strategy being explored to overcome host related limitations is the generation of an alternative surrogate expression host that may be more suited to efficiently expressing the environmental DNA at hand. Craig et al. (2009) discovered two novel compounds through functional screening of a soil derived metagenomic library expressed in *Ralstonia metallidurans*. The library was constructed using *E. coli* as a heterologous host and then the DNA transferred to *R. metallidurans* for activity based screening. Two clones showed activity in *R. metallidurans*, one displaying antimicrobial activity through the expression of a polyketide synthase gene and a second yellow colored clone expressing a carotenoid gene cluster. Clones active in *R. metallidurans* did not confer the same metabolic abilities onto the *E. coli* host. This shows the importance of using additional heterologous hosts to identify active clones which may not be expressed in the standard *E. coli* host. After their initial success, this research group carried out a study to compare six different Proteobacteria as hosts for the same soil derived metagenomic cosmid library (Craig et al., 2010). Each host expressing the library was functionally screened for antimicrobial activity, pigment production and altered colony morphology conferred onto the host by the DNA insert. Bacterial species from common soil-dwelling phyla were chosen as experimental hosts. Five candidate hosts, *Agrobacterium tumefaciens, Burkholderia graminis, Caulobacter vibrioides, Pseudomonas putida,* and *Ralstonia metallidurans*, were compared to the standard and most commonly used host, *E. coli*. Active clones were recovered from the library, having been expressed by different heterologous hosts with minimal overlap between hosts. This study shows the usefulness of Broad-Host Range vectors for overcoming host expression related barriers. Biver et al. (2013a) carried out a study to evaluate the use of an *E. coli-Bacillus subtilis* shuttle vector to functionally screen a forest soil-derived metagenomic library for antimicrobial activity. Activity based screening identified a novel antimicrobial agent, shown to be proteinaceous in nature though not yet fully characterized, that is active against *Bacillus cereus*. The DNA fragment responsible for such activity was active in the *B. subtilis* host alone and no activity was observed when the fragment was expressed in *E. coli*. Again, the importance of developing multiple host expression systems is highlighted by these findings. Further studies similar to those mentioned above must be carried out to better characterize and therefore more fully understand potential alternative hosts. Another obstacle faced in heterologous expression is the possibility of a DNA fragment being too short to contain a functional gene cluster or operon. The availability of a vector able to accommodate large DNA inserts is also fundamental (Streit and Schmitz, 2004). The use of large insert vectors capable of accommodating biosynthetic gene clusters or operons, and the development of shuttle vectors capable of propagating in more than one heterologous host, are examples of strategies being explored to overcome methodological limitations.

Applications of Interest of Functional Metagenomics in Food and Pharmaceutical Industries

Discovery of Novel Bio-catalysts

Certain microbial enzymes are of particular interest to the food and pharmaceutical industries for the catalysis of reactions which may be difficult or expensive to maintain. This interest stems from the fact that there is often difficulty in synthesizing chemical catalysts that truly mimic the complexity of biological enzymes. Many industrial processes are associated with a large environmental burden. Substituting traditional chemical processes used to produce certain compounds or molecules with enzymatic pathways naturally sourced is a more environmentally friendly approach to large-scale production. As microorganisms can catalyze a vast range of reactions, they are an obvious source of enzymes for industrial applications. Several authors have explored this avenue in the last decade (**Table 1**).

Novel enzymes from natural sources are extremely useful in food processing reactions. Many of these relate to reactions that occur in nature to process food for energy but are difficult to mimic on an industrial level, e.g., degradation of starch. In other instances, the search has focused on enzymes that can carry out reactions under extreme conditions, which often prevail in food processing, e.g., high temperatures and extremes of pH. Indeed, microbial enzymes are used for brewing, baking, synthesis of sugar and corn syrups, starch and food processing, texture and flavoring, processing of fruit juices, and production of dairy products and fermented foods, among others, either as recombinant enzymes or by using starter cultures with desirable activities. The following are some examples of industrial food processes which have benefited (and may continue to do so) from access to the diverse repository of enzymes possessed by microorganisms.

In the food industry, starch harvested from sources such as maize, wheat, and potatoes is processed to yield food products such as glucose and fructose syrups, starch hydrolysates, maltodextrins, and cyclodextrins (reviewed by van der Maarel et al., 2002). In recent times, the chemical hydrolysis of starch, which involves acid treatment, is being replaced with enzymatic digestion by starch-hydrolyzing enzymes obtained from natural sources. Starch-modifying enzymes are also added to dough in the baking industry to act as bread anti-staling agents. These starch-converting enzymes usually originate from the α-amylase family or family 13 glycoside hydrolase. Amylases from microbial sources are used in starch processing such as α-amylases from *Geobacillus stearothermophilus* and *Bacillus licheniformis*. However, despite the advantages of using enzymatic over chemical hydrolysis (high specificity of enzymes, milder reaction

TABLE 1 | Some novel enzymes of industrial interest discovered through functional metagenomics.

Enzyme	Closest known homolog	Method/Host	Environment	References
Four lipolytic enzymes	Moderate identity (<50%) to lipolytic proteins from *Streptomyces, Moraxella, Acinetobacter,* and *Sulfolobus* sp.	Activity based screening of *E. coli* plasmid library	Soil from a meadow, a sugar beet field and the Nieme River valley, Germany	Henne et al., 2000
Low pH, thermostable α-amylase	High sequence similarity to α-amylase of *Pyrococcus* sp. KOD1	Function-based screening of *E. coli* plasmid library followed by expression of gene of interest in *Pseudomonas fluorescens* for functional evaluation	Deep sea and acid soil	Richardson et al., 2002
12 esterases, 9 endo-β-1,4-glucanases, and 1 cyclodextrinase	Various putative source organisms	Functional screening of lambda phage library transformed into *E. coli*	Rumen of dairy cow	Ferrer et al., 2005b
Three ß-glucanases	Low sequence identities to known ß-glucanases. Other sequences present in one of the inserts showed identity to *Bacteroides* sp.	Function-based screening of *E. coli* BAC library	Large bowel of mouse	Walter et al., 2005
β-agarase	77% identity to corresponding protein in *Pseudoalteromonas atlantica*	Activity based screening of *E. coli* plasmid library	Soil	Voget et al., 2003
Two esterases	One esterase showed 83% identity to metagenome-derived EstA3 (AAZ48934) and 59% identity to a betalactamase (YP_003266771) of *Haliangium ochraceum* DSM 14365. The other esterase showed 37% identity to a hypothetical protein from *Neisseria elongata*	Activity based screening of two separate libraries: (plasmid and fosmid) transformed into *E. coli*	Soil Water	Ouyang et al., 2013
Two esterases	One esterase showed 51% identity to a class C ß-lactamase from *Burkholderia pseudomallei* and was also 61% similar and 45% identical to a functional esterase (AAF59826) from *Burkholderia gladioli*. Second esterase showed 59% identity to a ß-lactamase from *Sphingopyxis alaskensis*	Activity based screening of two *E. coli* cosmid libraries	Soil Drinking water	Elend et al., 2006
Esterase	Unidentified mesophilic soil microbe	Activity based screening of *E. coli* plasmid library	Environmental soil samples: mudflats, beaches, forests	Kim et al., 2006
Thermostable esterase	64% similarity to an enzyme from *Pyrobaculum calidifontis*	Activity based screening of *E. coli* fosmid library	Mud Sediment-rich water	Rhee et al., 2005
Two esterases	One esterase showed highest identity (64.9%) to a putative esterase (YP_220901) from *Brucella abortus biovar* 1. The other esterase showed highest identity (40.4%) to a putative esterase (ZP_01658665) from *Parvibaculum lavamentivorans*	Activity based screening of *E. coli* BAC library	Surface seawater, South China Sea	Chu et al., 2008
Six lipolytic clones	The six clones individually showed highest identity to the following proteins: (i) Esterase/lipase (ZP_00034241), *Burkholderia fungorum*, (ii) Thermophilic carboxylesterase (1EVQA), *Alicyclobacillus acidocaldarius* (iii) Thermophilic carboxylesterase (1EVQA), *A. acidocaldarius* (iv) Esterase/lipase (ZP_00034303), *B. fungorum* (v) Esterase/lipase (ZP_00034303), *B. fungorum* (vi) Esterase HDE (BAA82510), petroleum-degrading bacterium HD-1	Activity based screening of *E. coli* fosmid library	Forest topsoil	Lee et al., 2004

(Continued)

TABLE 1 | Continued

Enzyme	Closest known homolog	Method/Host	Environment	References
Cellulase (β-glucosidase activity)	Low sequence identity to *Plasmodium* and *Borrelia* species	Function-based screening of *E. coli* library	Soil	Jiang et al., 2009
Glycosyl hydrolase	>60% identity to β-1-4-endoglucanase from *Prevotella bryantii* B14 (AAC97596) and β-1-4-xylanase from *Prevotella ruminicola* 23 (AAC36862). 100% identity to a partial sequence (AAB20175) of the N terminus B14 enzyme from *P. bryantii.*	Functional screening of lambda phage library transformed into *E. coli*	Cow rumen fluid	Palackal et al., 2007
137 nitrilase genes (Relevant in fine chemical synthesis in drug manufacture)	Varying degrees of amino acid sequence similarity to proteins from several sequence clades within the nitrilase subfamily	A phagemid library expressed in *E. coli* screened by selection for the ability to grow on a nitrile substrate	Soil Water	Robertson et al., 2004
Halotolerant and moderately thermostable tannase	New member of tannase superfamily	Activity-based screening of *E. coli* plasmid library	Cotton field soil	Yao et al., 2011
Three carboxylic ester hydrolases	77% amino acid identity to lipolytic enzyme (AEM45126) from German forest soil-derived metagenomic library	Activity-based screening of *E. coli* plasmid library	Forest soil	Biver and Vandenbol, 2013
Alkaline serine protease	Most closely related to an alkaline protease isolated from *Bacillus* sp.	Activity-based screening of IPTG-inducible vector library expressed in *E. coli*	Forest soil	Biver et al., 2013a
Fibrinolytic metalloprotease (zinc-dependent)	Amino acid sequence showed 46% identity to metallopeptidase from *Dechloromonas aromatica* (AAZ45577)	Activity-based screening of *E. coli* fosmid library	Mud, Korean west coast	Lee et al., 2007
Two serine proteases	First novel protease: 52% amino acid identity to a thermophilic alkaline protease from *Geobacillus stearothermophillus* (AAK29176). Second novel protease: 51% sequence identity with a putative protease of *Bacillus sphaericus* (CAB46075)	Activity-based screening of *E. coli* plasmid and fosmid libraries	Surface sand from Gobi and Death Valley deserts	Neveu et al., 2011
Alkaline serine protease	98% sequence similarity with uncharacterized proteases of various *Shewanella* sp.	Activity-based screening of *E. coli* plasmid library	Goat skin surface	Pushpam et al., 2011
Cold-active lipase	91% identity to a known lipase from *Pseudomonas fluorescens* B68 (AY694785)	Activity based screening of *E. coli* cosmid library	Oil-contaminated soil, Northern Germany	Elend et al., 2007
Moderately thermostable (and thermally activated) lipase	*Acidobacteria* phylum	Activity based screening of *E. coli* fosmid library	Soil, Brazilian Atlantic Forest	Faoro et al., 2012
Five esterases	Two did not show significant sequence identity to known esterases, the remaining genes showed low to moderate identity to known esterases	Activity based screening of *E. coli* phagemid vector library	Brine: seawater interface, Uranian hypersaline basin	Ferrer et al., 2005a
Thermostable family VII esterase with high stability in organic solvents	45% identity to *Haliangium ochraceum* DSM 14365 (ACY17267)	Activity based screening of *E. coli* fosmid library	Compost	Kang et al., 2011

(Continued)

TABLE 1 | Continued

Enzyme	Closest known homolog	Method/Host	Environment	References
Alkaline-stable family IV lipase	83% identity with a cold-active esterase from a deep-sea metagenomic library (ADA70028). 59% identity with an esterase from *Vibrio splendidus* LGP32 (YP_002394831)	Activity based screening of *E. coli* plasmid library	Marine sediment, South China Sea	Peng et al., 2014
Protease-insensitive feruloyl esterase	56% identity to predicted esterase from *Eubacterium siraeum* V10Sc8a (CBL34630). 55% identity to predicted esterase from *E. siraeum* (CBK96609)	Function-based screening of *E. coli* fosmid library	China Holstein cow rumen	Cheng et al., 2012a
Xylanase	44% identity to glycoside hydrolase family protein from *Clostridium thermocellum* ATCC 27405 (YP001038252)	Function-based screening of *E. coli* fosmid library	China Holstein cow rumen	Cheng et al., 2012b
Two UDP glycotransferase (UGT) genes. One is a novel macroside glycotransferase (MGT)	The first one is weakly similar (71% similarity) to hypothetical UGT from *Fibrisoma limi*. The second one is highly similar to a hypothetical MGT from *Bacillus thuringiensis*	Thin layer chromatography (TLC)-based functional screening of *E. coli* fosmid library	Elephant feces, Hagenbeck Zoo, Germany. Tidal flat sediment, Elbe river, Germany.	Rabausch et al., 2013
Cold-adapted ß-galactosidase	Highest percentage identities to β-galactosidases from *Planococcus* sp. "SOS Orange" (39%), *Planococcus* sp. L4 (39%), and *Bacillus halodurans* C-125 (39%)	Function-based screening of *E. coli* plasmid library followed by expression of gene of interest in *Pichia pastoris* for analysis and characterization	Topsoil samples, Daqing oil field, Heilongjiang Province in China	Wang et al., 2010
Cold-active ß-galactosidase	53% identity to β-galactosidases from *Clostridium hathewayi*	Function-based screening of *E. coli* plasmid library.	Ikaite columns SW Greenland	Vester et al., 2014
ß-galactosidase	Not available	Function-based screening of *E. coli* plasmid library followed by expression of gene of interest in *Pichia pastoris* for functional evaluation	Not available	Wang et al., 2012
11 amidase genes (Three novel)	Three novel amidases: the first showed highest identity (54%) to putative isochorismatase hydrolase from *Streptomyces* sp. strain AA4; the second showed 45% primary amino acid sequence identity with a hypothetical protein (further information not available); the third showed 57% primary amino acid sequence identity with a protein that contains a transmembrane ABC transporter signature motif and possibly encodes a polypeptide with amidase activity	PIGEX-based screening of benzoate-responsive sensor plasmid library transformed into *E. coli*	Activated sludge from aeration tank of a coke plant; wastewater treatment plant, Japan	Uchiyama and Miyazaki, 2010
Periplasmic α-amylase	100% similarity with *mal*S gene in *E. coli* (X58994.1)	PIGEX-based screening of maltose-induced plasmid library transformed into *E. coli*	Cow dung, India	Pooja et al., 2015
37 genes with lipolytic activity	29–90% sequence identity to known and putative proteins from numerous different species, including uncultured bacteria	Activity based screening of *E. coli* plasmid and fosmid libraries	Forest soil, Germany	Nacke et al., 2011

conditions, natural means of processing more acceptable to consumers and to the public), there are limitations with the enzymes currently being used. Starch hydrolysis is carried out at high temperatures, at which α-amylases are usually not active at a pH below 5.9. For the reaction to proceed efficiently, the pH must be raised by the addition of NaOH. As these enzymes also exhibit a Ca^{2+} dependency, Ca^{2+} must be added to the reaction in addition to adjusting the pH. Thermostable,

Ca^{2+}-independent α-amylases with low pH activity would be ideal for the starch hydrolyzing process. Richardson et al. (2002) identified an α-amylase optimal for the corn wet milling process. They carried out activity based screenings under conditions of temperature and pH similar to those of the corn wet milling process on a large library of metagenomic clones constructed from diverse environmental samples. The clones were also phylogenetically screened for homology to known α-amylases. Three clones were selected which performed well under the given conditions. Phylogenetic analysis revealed that all three enzymes were members of the glycosyl hydrolase family 13. They were expressed in *Pseudomonas fluorescens* and their activity was compared to the enzymes currently used in industry (from *B. licheniformis*). One clone was found to have better characteristics for application to the corn wet milling process than the enzyme currently in use. However, further research is needed to improve the low yield of enzyme produced under industrial conditions.

Lipases and esterases are hydrolytic enzymes which play important roles in the food and pharmaceutical industries. Lipases hydrolyze fats into fatty acids and glycerol at the water lipid interface and reverse the reaction in the non-aqueous phase (Gupta et al., 2004). Lipases are exploited by the dairy industry for the hydrolysis of milk fat, releasing short-chain and long-chain fatty acids, creating such features as richness, creaminess or cheesiness depending on the degree of lipolysis, as reviewed by Hasan et al. (2006). For this reason, it is important to use the correct lipolytic enzyme to achieve the right flavor in the final product. Peng et al. (2014) screened a metagenomic library constructed from a Chinese marine sediment for clones displaying lipolytic activity in an *E. coli* host. They discovered a novel highly alkaline-stable lipase with high specificity for butter milkfat esters. Treatment of butter with the newly identified lipase produced rich and distinctive flavors through the production of palmitic and myristic acids while maintaining the cheesy flavor of the short-chain fatty acids. As palmitic and myristic acids are added to food for their distinctive flavor, the hydrolysis of palmitate and myristate in the production of lipolysed milkfat (LMF) to flavor dairy products is a safe and economically viable potential application of the novel lipase identified in this study. Other dairy applications of lipases include the acceleration of cheese ripening and the enhancement of cheese flavor through the synthesis of short chain fatty acids (SCFAs) and alcohols. Lipases are also used in vegetable oil modification and preservation of baked goods (Hasan et al., 2006). Although in the past lipases used in the food industry were predominantly obtained from animal sources, the microbial world potentially holds a wide range of diverse lipases that can be used in many different industrial applications (**Table 1**). Examples of pharmaceutical applications of lipases sourced from microbes include the synthesis of an intermediate for the production of an anti-tumor agent (Zhu and Panek, 2001) and the synthesis of intermediates of antimicrobial agents (Kato et al., 1997). Also, through the screening of a metagenomic library constructed from an oil-contaminated German soil sample, Elend et al. (2007) identified a lipolytic cold-activated clone which showed high selectivity for esters of primary alcohols and (*R*) enantiomers of non-steroidal anti-inflammatory drugs such as ibuprofen. This enzyme has potential in the pharmaceutical industry for the conversion of such anti-inflammatories into an optically pure form.

Esterases catalyze the hydrolysis of an ester into its alcohol and an acid in aqueous solution. They are distinguished from lipases in that they hydrolyze short-chain over long-chain acylglycerols. In the food industry, esterases are used in fat and oil modification and in the fruit juices and alcoholic beverages industries to produce certain flavors and fragrances, as reviewed by Panda and Gowrishankar (2005). Feruloyl esterases hydrolyze the ester bond between ferulic acid (FA) and polysaccharides present in plant cell wall material. They have a dual usefulness as they not only break down plant biomass (which is useful in industrial waste management) but, in doing so, they de-esterify dietary fibers releasing bioactives with potential beneficial health effects (reviewed by Faulds, 2010). In a study carried out by Cheng et al. (2012a), a metagenomic library constructed from the microbial content of a Chinese Holstein cow rumen was functionally screened for feruloyl esterase activity, identifying a protease-insensitive feruloyl esterase capable of releasing FA from wheat straw. This novel enzyme is of particular industrial interest as it showed high thermal and pH stability and was resistant to several proteases including pepsin. A novel xylanase was isolated from the same metagenomic library (Cheng et al., 2012b) and its ability to work synergistically with the newly discovered feruloyl esterase to release xylooligosaccharides (XOS) and FA from wheat straw was assessed. XOS display prebiotic and gut modulatory activities and have other bioactive properties giving them value as food additives, as reviewed by Moure et al. (2006). The novel xylanase was not only effective in working with the feruloyl esterase, but additionally was capable of improving release of FA from wheat straw at a high dose. Esterases also play a role in the synthesis of chiral drugs including medications to relieve pain and reduce inflammation (Bornscheuer, 2002; Shen et al., 2002; Panda and Gowrishankar, 2005).

ß-galactosidases are widely used in the dairy industry for the hydrolysis of lactose to glucose and galactose. Lactose content in milk is reduced to improve taste (lactose is known to absorb undesirable flavors and odors), to accelerate the ripening of cheeses made from treated milk and for the removal of lactose for the production of lactose-free products for intolerant consumers (reviewed by Panesar et al., 2010). The currently commercially available ß-galactosidase for use in the dairy industry, from *Kluyveromyces lactis*, has a temperature optimum of 50°C and loses much of its enzymatic activity at temperatures below 20°C. Carrying out industrial reactions at lower temperatures is beneficial as it saves energy (and in turn is more economical), it prevents heat destruction of thermosensitive substances such as food compounds, molecules responsible for flavors, taste and nutritional value, etc., and it reduces contamination risks. Cold-active enzymes work at low temperature and can be easily inactivated by rising the temperature to a moderate condition. From a metagenomic library constructed from the ikaite columns of SW Greenland, Vester et al. (2014) isolated a cold-activated ß-galactosidase which can potentially be applied by the dairy industry. The discovered enzyme has an optimal pH of 6 (the natural pH of milk being pH 6.7–6.8) and a temperature

optimum of 37°C, but retains lactose hydrolytic activity at 5°C. These properties make it a good candidate for the hydrolysis of lactose into glucose and galactose in milk for the removal of lactose for production of lactose-free products for lactose-intolerant people. In a similar study by Wang et al. (2010) a cold-adapted ß-galactosidase was identified from a metagenomic library expressed in *E. coli*. The insert from the active clone (encoding a full-length ß-galactosidase) was expressed in *Pichia pastoris* to assess its candidacy for use in milk treatment and optimal activity was observed at a temperature of 38°C. The enzyme was active at the natural pH of milk.

Flavonoids are plant secondary metabolites found in numerous dietary fruits and vegetables and whose consumption is beneficial to human health (Ververidis et al., 2007a,b). Flavonoids are difficult to source as they are produced by plants at very low levels. Due to their structural complexity enzymatic modification is preferred over a chemical approach for industrial production. Glycosylation of flavonoids influences their water solubility and bioavailability, making glycosyltransferases that are active on flavonoids of great interest to the food and pharmaceutical industries. Rabausch et al. (2013) developed a novel thin-layer chromatography (TLC) based screening method for the identification of flavonoid-modifying enzymes from a metagenomic library. Two novel flavonoid-modifying enzymes with high activity on flavones, flavonols, flavanones, isoflavones, and stilbenes were discovered in this manner.

Proteases hydrolyze peptide bonds and therefore catalyze the degradation of proteins. They have numerous uses in the food industry, including the tenderizing of meat (Ashie et al., 2002), the coagulation of milk and flavor development in the dairy industry (Huang et al., 2011) and the proteolysis of gluten to achieve gluten-free products in the baking industry (Hamada et al., 2013). Proteases may also be used to release beneficial bioactive peptides from polypeptide chains in certain foods (Hafeez et al., 2014; Mora et al., 2014). Currently, commercial proteases used in the food industry are generally sourced from plants and culturable microorganisms. Proteases from as yet uncultured microbial extremophiles would be of use in the carrying out of proteolysis under unconventional reaction conditions. There have been several novel proteases discovered through functional metagenomic methods. For instance, Biver et al. (2013a) identified an oxidant-stable alkaline serine protease from a forest-soil metagenomic library. An alkaline serine protease was also identified in a metagenomic library constructed from goat skin surface samples by Pushpam et al. (2011). These alkaline proteases are examples of microbial enzymes with potential industrial applications, mainly in the detergent industry.

Tannins are naturally occurring water soluble polyphenols which constitute a large percentage of plant material. Tannases catalyze the hydrolysis of tannins, releasing gallic acid, and glucose. Tannases are used in the food industry as a clarifying agent in the manufacture of beverages such as instant teas, fruit juices, beer, and certain wines (Cantarelli et al., 1989; Boadi and Neufeld, 2001). Tannases are also important to the pharmaceutical industry for catalyzing the release of gallic acid (Sariozlu and Kivanc, 2009) which is used in the production of some antibacterial drugs. Additionally, gallic acid is used in the synthesis of propyl gallate, an antioxidant food additive. Tannases isolated from bacteria have typically been restricted to culturable strains, overlooking the diverse potential of those as yet uncultured. Yao et al. (2011) expressed a metagenomic clone library constructed from cotton field in *E. coli* and screened the transformants for tannase activity, revealing one active clone. Sequence analysis revealed that the active clone encoded a full length tannase gene, which was not found to be closely related to any currently known tannases. Analysis of tannase activity of the enzyme under various industrially relevant conditions was performed and a moderate thermostability of the identified enzyme, which may be useful for food industrial applications, was shown. The enzyme was also found to have a wide range of substrate specificity, making it suitable for applications in both the food and pharmaceutical industries. In 2014, this novel tannase was investigated by Yao et al. (2014) for its suitability for the removal of tannins from a green tea infusion. The presence of tannins in beverages such as green tea is problematic as the ability of tannins to precipitate proteins leads to the formation of a protein haze that is undesirable in terms of product taste and appearance (Wu and Bird, 2010). The tannase enzyme was recombinantly expressed in *E. coli* and immobilized to several matrices, identifying Ca-alginate beads as the most appropriate support. The immobilized enzyme was effective in the removal of tannins from green tea infusion and was found to possess properties distinct from those of the free enzyme, such as high operational and storage stabilities and a higher temperature and pH optimum.

Discovery of Novel Bioactives

As with the food industry, the use of microbial enzymes is of particular interest for the biosynthesis of pharmaceutical products previously synthesized *via* chemical means. Thus, functional metagenomics can be applied to the discovery of genes capable of carrying out reactions of interest for the obtaining of bioactives or the synthesis of intermediate compounds in the pharmaceutical industry. One avenue of interest has been the identification and heterologous expression of a microbial biosynthetic pathway capable of producing biotin for industrial purposes (Entcheva et al., 2001; Streit and Entcheva, 2003). Biotin (Vitamin H) is a human and animal dietary requirement and is currently chemically synthesized through industrial processes for addition to food and feed products, with associated negative environmental impacts. The use of biotin-producing microorganisms in place of chemical synthesis offers a greener alternative to conscientious industries. Other microbial biosynthetic genes of interest to the pharmaceutical industry capable of synthesizing other bioactives important for human health and medicine have been also identified by functional metagenomic strategies (listed in **Table 2**).

Walter et al. (2005) applied a functional metagenomic method to screen for lichenin-degrading activity in a Bacterial Artificial Chromosome (BAC) library constructed from bacteria obtained from the large-bowel microbiota of mice, identifying three clones with ß-glucanase activity. Glucans cannot be broken down by humans or monogastric animals and so, their hydrolysis relies

TABLE 2 | Some novel bioactives and biosynthetic pathways of industrial interest discovered through functional metagenomics.

Bioactive /Pathway	Closest known homolog	Method/Host	Environment	References
Pederin	>80% identity to sequences from *P. aeruginosa.* The discovered *ped* gene cluster is believed to be from a symbiont of the *Paederus* beetle from the genus *Pseudomonas*	Targeted sequencing-based strategy	*Paederus* beetles	Piel, 2002
Biotin	Highest identity to proteins from *Erwinia herbicola*. Significant identity also shown to proteins from *E. coli* and *Pseudomonas putida*	Selelction-based screening of enriched cosmid library in *E. coli* biotin auxotrophic strain	Horse excrement	Entcheva et al., 2001
Known siderophore: vibrioferrin	98% identity to proteins from *Vibrio parahaemolyticus* and *Vibrio alginolyticus*	Function-based screening of *E. coli* plasmid library	Tidal-flat sediment, Ariake Sea	Fujita et al., 2011
Polyketide synthase (PKS) gene	55–59% identity to hypothetical PKS from *Mycobacterium avium* (NP_961164)	Targeted sequencing-based strategy	Marine sponge *Discodermia dissoluta*, Netherlands Antilles	Schirmer et al., 2005
Novel serine protease inhibitor (serpin) gene	Moderate identities to serpins from *Salinibacter ruber* M8 and *Spirosoma linguale* DSM 74. Similarities with possible partial serpins from *Dyadobacter fermentans* DSM 18053, *Arthrospira maxima* CS-328 and *Cyanothece* sp. PCC 7822	Sequence-based screening of *E. coli* plasmid library.	Uncultured marine organisms	Jiang et al., 2011
Borregomycin A and B encoded by *bor* pathway (antiproliferative and antibiotic properties)	ORFs showing 32–86% identity to species from the following genera: *Micromonospora*, *Streptomyces*, *Actinoplanes*, *Corallococcus*, *Cellulomonas*, *Actinomadura*, *Salinispora*, *Microlunatus*, *Modestobacter*, *Frankia*, *Saccharomonospora*, *Nocardia*, *Phaeosphaeria*	Homology guided screening	Soil, Anza-Borrego Desert (CA)	Chang and Brady, 2013
Hypothetical protein with NF-kB pathway stimulatory activity	42% of predicted genes coverage to *B. vulgatus* ATCC 8482	Activity-based screening using a reporter cell line of an *E. coli* fosmid library	Human gut microbiota of Crohn's Disease patients	Lakhdari et al., 2010
Novel prebiotic degradation pathways (11 contigs)	Sequence homology to species of *Bifidobacterium*, *Eubacterium*, *Streptococcus*, *Bacteroides*, *Faecalibacterium*	Hydrolytic activity-based selective screening of two *E. coli* fosmid libraries	Human ileum mucosa and fecal microbiota samples	Cecchini et al., 2013
Five novel putative salt tolerance genes	Identity to hypothetical proteins from genus *Collinsella*, *Eggerthella,* and *Akkermansia*	Function-based screening of *E. coli* plasmid library	Human gut microbiota	Culligan et al., 2012
Novel salt tolerance gene	Not homologous to any sequence at time of study, highest BLAST score to hypothetical protein from *Caulobacter crescentus*	Function-based screening of *E. coli* plasmid library	Faecal sample, healthy 26 year old Caucasian male	Culligan et al., 2013
15 acid resistance genes	37–90% identity to proteins and hypothetical proteins from the following genera: *Thermosinus*, *Streptomyces*, *Candidatus*, *Hyphomicrobium*, *Methylococcus*, *Acidithiobacillus*, *Thioalkalivibrio*, *Nitrosococcus*, *Halorhodospira*, *Haliangium*, *Clostridium*, *Roseomonas*, *Acidiphilium*, *Gemmata*, *Terriglobus*, *Burkholderia*	Function-based screening of six *E.* coli plasmid libraries. Followed by expression in *Pseudomonas putida* and *Bacillus subtilis*	Planktonic and rhizosphere microbial communities of the Tinto River. Five libraries from *Erica andevalensis*, one from headwaters of Tinto River	Guazzaroni et al., 2013

on bacterial fermentation. As the consumption of glucans is associated with health benefits in humans (Abumweis et al., 2010), glucan hydrolyzing enzymes isolated from bowel-dwelling microbiota may be of interest to pharmaceutical and functional food related industrials. The feed industry may also benefit from the availability of ß-glucanases that improve the digestion of

barley-based feed diets by poultry livestock (Von Wettstein et al., 2000).

The development of novel therapeutic strategies relies heavily on gaining a better understanding of human commensals and host-microbe relationships. Lakhdari et al. (2010) established and validated a reporter system capable of detecting immune modulatory activity of metagenomic clones. A metagenomic library constructed from human fecal microbiota of Crohn's Disease (CD) patients was screened for NFkB modulatory activity (whether stimulatory or inhibitory) using an intestinal epithelial cell line transfected with a reporter gene. A clone displaying stimulatory activity of the NF-kB pathway was identified. Although the molecule responsible for the activity is not yet known, two potential candidate loci were determined through transposon mutagenesis: an efflux ABC type transport system and a putative lipoprotein. Phylogenetic analysis showed *Bacteroides vulgatus* to be the closest known homolog to the source of the insert of interest, an interesting finding as *B. vulgatus* is a human gut microbe found to be higher in abundance in CD patients than in a control population. This study presents the development of an innovative platform for screening metagenomic libraries and is likely to inspire the creation of other cell-based screening platforms from which a better understanding of human-microbe symbiotic communications can be obtained, advancing the development of novel therapeutic strategies promoting a healthy relationship with the gut microbiota and in turn the entire human microbiome.

Maintaining gut microbiota homeostasis has been shown to contribute to the overall sustaining of human gut health. Probiotics are an oral infusion of high numbers of live beneficial gut microbes formulated into various yogurts and dairy beverage products that, when ingested in adequate amounts, confer a health benefit on the host (Joint, 2001). As an oral formulation, these products face difficulties in efficacy due to insufficient cell numbers reaching the intestine, owing to the necessity of passing through the majority of the GI tract to reach their site of action in the bowel. The harsh pH and osmolarity of the upper GI tract can destroy a large proportion of the ingested cells. Novel acid and salt resistance mechanisms discovered through functional metagenomic studies similar to those of Guazzaroni et al. (2013), who identified an acid resistant metagenomic clone from the Tinto River environment, and Culligan et al. (2013), who discovered a gene conferring salt tolerance onto an *E. coli* host from a library derived from the human gut microbiota, may be of use in conferring stress resistance to probiotic products. However, this objective faces additional social challenges with respect to consumer acceptance of the use of genetically modified (GM) microorganisms to enhance food products. Although it is generally appreciated by the public that GM cells, organisms and microorganisms are necessary for the production of certain critical biologically active drugs, the thought of everyday food products having been prepared using GM materials is met with a sense of unease, especially in many EU member states. Thus, strict regulations involving the consumption of GM foods and the use of GM organisms in food production and processing have not been made more lenient, as they have in other countries, such as the USA, in recent years. Public transparency and an understanding of the extensive safety and efficacy testing of GM related food products may eventually lead to a change in consumer attitude to bioengineered goods.

Another avenue to maintain human gut health is to promote the growth of beneficial bacteria already present in one's lower GI tract through the use of prebiotics. Prebiotics are non-digestible oligosaccharides (NGOs), usually present in plant material, that are resistant to human digestion in the upper GI tract and are hydrolyzed in the gut by beneficial microbiota to produce SCFAs and organic acids that provide nutritional value to the human host (Gibson and Roberfroid, 1995). Cecchini et al. (2013) used a functional metagenomics approach to investigate the prebiotic hydrolyzing potential of the human gut microbiome by searching for novel prebiotic degradation pathways in a human ileum mucosa and a fecal microbiota derived metagenomic library. They identified high numbers of unknown gut microorganisms capable of hydrolyzing established prebiotics, indicating that the prebiotics tested are not specifically metabolized by their target microorganisms alone. Further investigations must be carried out to determine the effect (if any) of non-specific hydrolysis of prebiotic preparations in the human gut. Galacto-oligosaccharides (GOS) with prebiotic properties can be synthesized through the transgalactosylation activity of ß-galactosidase enzymes on lactose. Wang et al. (2012) validated the ability of a novel ß-galactosidase isolated from a metagenome-derived library for its ability to produce GOS. Carrying out the reaction in an organic-aqueous biphasic media was shown to improve GOS yield. The ß-galactosidase gene discovered in this study is a promising candidate for industrial production of GOS to be used as an additive in various food and dairy products. All of these studies highlight the flexibility of functional metagenomics as a molecular tool not only for identifying new metabolic pathways for biosynthesis of useful/industrially relevant compounds but also for evaluating the efficiency of current therapeutic strategies.

Discovery of Novel Antimicrobials

A major driving force behind the biotechnological applications of functional metagenomics is the search for novel antimicrobials effective in medical settings. Microorganisms produce antibiotic molecules to alleviate competitors in their natural habitat. Natural sources have proved fruitful in the past for providing antibiotic molecules, from the discovery of penicillin produced by *Penicillium rubens* in 1928 to date. Although most bacterial infections in humans are curable with current antibiotic therapies, the emergent problem of antimicrobial resistance has led to the prevalence of persistent untreatable infections caused by certain pathogens which have developed a resistance to the used antimicrobial therapy. Antibiotic resistance has challenged medical practitioners and researchers and has led to outbreaks of serious untreatable bacterial infections in clinical settings and even community outbreaks have occurred (Alanis, 2005), making antimicrobial resistance a serious threat to human health (World Health Organization, 2014). The rate of antimicrobial drug discovery has declined in recent years, owing in part to a low drug approval rate by governing bodies (Cooper and Shlaes, 2011) and

lesser rewards for manufacturers (Fischbach and Walsh, 2009). The exhaustion of products from culturable microorganisms and preferred use of chemical libraries of pure synthetic compounds over natural product exploration (Li and Vederas, 2009) have also contributed. New advances in metagenomics, high throughput screenings (HTS) and metabolic engineering, e.g., Jayasuriya et al. (2007), provide a new lease of life for natural product drug discovery. Functional metagenomic screens can be applied to the identification of novel antimicrobial molecules by screening microbial populations for antimicrobial activity against indicator or clinically relevant microorganisms. So far, this approach has led to the discovery of several novel antimicrobial compounds (**Table 3**). Gillespie et al. (2002) described the discovery of two novel antimicrobials (turbomycin A and B) exhibiting broad-spectrum activity against both gram-positive and gram-negative bacteria. These antibiotics were identified through activity-based screening of a metagenomic library from soil samples expressed in an *E. coli* host. Several metagenomic *E. coli* clones expressing antimicrobial activity were discovered by Macneil et al. (2001) through function-based screening of a BAC library constructed from soil microbial DNA. Metagenomic inserts from active clones were found to be related to the compound indirubin, a cyclin-dependent kinases (CDK) inhibitor used in the treatment of human chronic myelocytic leukemia (Hoessel et al., 1999; Marko et al., 2001). An indirubin compound with antimicrobial activity was also identified through activity-based screening of a forest soil metagenomic library by Lim et al. (2005). More recently, Scanlon et al. (2014) developed a HTS method which enabled them to co-culture recombinant clones from a native staphylococcal-derived metagenomic library with the bacterial pathogen *Staphylococcus aureus* in hydrogen-in-oil emulsions, with antibiotic activity being rapidly detected using a fluorescent viability assay. Six clones expressing a lysostaphin gene from *Staphylococcus simulans* with activity against *S. aureus* were identified in this way. Iqbal et al. (2014) constructed a metagenomic library from Arizona soil hosted by *Ralstonia metallidurans*. Functional screening for antimicrobial activity against *Bacillus subtilis* identified six positive clones encoding proteases, a lipase, and enzymes with cell wall lytic activity. These studies highlight the success of applying functional metagenomics to the discovery of novel natural antimicrobials with potential value to the pharmaceutical industry.

Certain cell-to-cell communication or quorum sensing molecules and agents with quorum sensing inhibitory (QSI) activities have been also discovered through function-based screening of metagenomic libraries (**Table 3**). An interesting study by Nasuno et al. (2012) identified two novel sets of quorum sensing (QS) genes from the LuxI family *N*-acyl-L-homoserine lactone (AHL) synthases and their paired LuxR family transcriptional regulators. These authors constructed metagenomic libraries from an activated sludge from a coke plant and forest soil samples and functionally screened them for the presence of QS genes using a modified *E. coli* host. This biosensor strain contained a *gfp* plasmid which produced unstable GFP in response to low levels of five different AHLs, enabling the detection of QS-regulated activity. Other studies which have applied metagenomics for the exploration of QS regulation are reviewed by Kimura (2014). When it comes to treating individuals infected with, or curbing outbreaks of, antimicrobial-resistant pathogens, in some cases quorum sensing inhibitors as an anti-virulence strategy may be a useful course of action. The concept of using quorum sensing inhibitors would also be of benefit to the food industry in the control of undesirable microorganisms in food preparations or food processing environments. Schipper et al. (2009) screened a soil metagenomic library, expressed in *E. coli*, for QSI activity using an *A. tumefaciens* based bioassay. The positive clones were expressed in *Pseudomonas aeruginosa* and were found to be most likely responsible for the reduced motility and biofilm formation observed in the *P. aeruginosa* host cells expressing the proteins of interest. Of the three active clones isolated, one was found to be similar to a known lactonase, and the remaining two clones were determined to encode novel lactonases.

Certain antimicrobial strategies used in clinical settings could also be applied to the control of bacterial persistence in food development and manufacturing processes. In industrial settings contamination of food products occurs at various stages throughout the food processing cycle. The raw food itself is usually a source of initial contamination. Food can also become contaminated or re-contaminated during its processing, e.g., re-contamination of milk post-pasteurization, resulting in an unsafe or spoiled product. The removal of harmful or spoilage microorganisms from food products and the prevention of microorganisms entering or persisting in food processing is highly desirable. This needs to occur without damaging the structure, texture, taste, and overall quality of food. A potentially powerful application of functional metagenomics with respect to the food industry is screening natural sources for bioactive molecules that function as antimicrobials or inhibitory compounds for use in food safety maintenance strategies. Once the compounds have been identified and mass produced, the ultimate goal is for them to be formulated into safe sanitization products that will not influence the quality of the food product. As microorganisms are widely used and often beneficial to the food industry (e.g., cheese manufacture, brewing), the aim would be to eliminate only those microorganisms which pose a threat to food safety and quality. Screening is performed in a targeted manner to identify isolates producing compounds that inhibit or eliminate the presence of a given problematic microorganism present in the food product or processing equipment. Due to their specificity, bioactives isolated from microorganisms may be used in combination with existing sanitization products. Extremophiles are of particular interest as these could target undesirable microorganisms in extreme environments, which are often present in food processing.

Functional metagenomics can be used to combat antimicrobial resistance via two strategies; through the discovery of novel antibiotics and anti-infectives (as mentioned above) and through the identification of resistance genes in microbial populations. As resistance is transferable, horizontal gene transfer (HGT) being the most common method by which resistance is acquired by previously susceptible strains, resistant genes possessed by environmental bacteria may be acquired by human pathogens. Functional metagenomics can be used

TABLE 3 | Some novel antimicrobials, anti-infectives and antimicrobial resistance genes discovered through functional metagenomics.

Antimicrobial	Closest known relationship (percentage homology)	Method/Host	Environment	References
Long-chain *N*-acyltyrosine synthase genes	No identity to bacteria cultured at that time. Some similarity to predicted proteins from *Nitrosomonas europaea*, *Desulfovibrio vulgaris,* and *D. desulfuricans*	Activity-based screening of *E. coli* cosmid library	Seven soil samples, Ithaca, NY Boston, MA Costa Rica	Brady et al., 2004
N-acyl amino acid biosynthesis gene	Highest similarity to hypothetical protein (MJ1207) from *Methanococcus jannaschii*	Activity-based screening of *E. coli* cosmid library	Soil	Brady and Clardy, 2000
Two isocyanide biosynthetic genes encoding isocyanide-containing antibiotic	Not available. Some identity to known and predicted proteins	Activity -based screening of *E. coli* cosmid library	Soil, Boston, MA	Brady and Clardy, 2005
Violacein biosynthetic gene cluster	Moderate identity to *Chromobacterium violaceum*	Activity -based screening of *E. coli* cosmid library	Soil, Ithaca, NY	Brady et al., 2001
Two ORFs within a clone encoding a transcriptional regulatory protein and a putative indole oxygenase	The indole oxygenase-like protein showed high identity to naphthocyclinone hydroxylase (NcnH) from *Streptomyces arenae*	Activity -based screening of *E. coli* fosmid library	Forest topsoil, Jindong Valley, Korea	Lim et al., 2005
Turbomycin A, B	The ORFs encoding the turbomycins A and B show 53% identity to legiolysin from *Legionella pneumophila,* 54% identity to hemolysin from *Vibrio vulnificus*, 49% identity to 4-hydroxyphenylpyruvate dioxygenase from *Pseudomonas* and 45% identity to MelA in *Shewanella colwelliana*	Activity-based screening of *E. coli* plasmid library	Soil	Gillespie et al., 2002
Uncharacterized protein with antimicrobial activity	Low to moderate sequence identity (26–58%) to proteins and hypothetical proteins from *Solitalea canadensis* DSM 3403 (38 and 46%), *Flavobacterium* sp. CF136 (26%), *Indibacter alkaliphilus* LW1 (40%), *Helicobacter bizzozeronii* CIII-1 (31%) and *Acidovorax* sp. JS42 (58%)	Activity-based screening of *E. coli-Bacillus subtilis* shuttle vector library	Soil sample from a deciduous forest, Belgium	Biver et al., 2013b
Novel chitinase with chitobiosidase activity (identified by the sequence-based approach)	45% identity to chitinase from an uncultured bacterium (Uchiyama and Watanabe, 2006) and amino acid identity to known proteins from *Chondromyces apiculatus* (41%), *Corallococcus coralloides* (40%), and *Myxococcus xanthus* (39%)	Targeted sequence-based analysis and activity-based screening of *E. coli* fosmid library	Soil, Swedish University of Agricultural Sciences, Uppsala, Sweden	Hjort et al., 2014
Six clones with antimicrobial activity: two with cell wall-degrading activity, three proteases and a lipolytic enzyme	54–31% identity to known amidase, lytic transglycosylase and proteases from *Desulfovibrio sp. U5L, Clostridium sp. CAG:1013, Myxococcus xanthus*, *Leptospira santarosai* and *Ferroglobus placidus* and to a putative lipolytic enzyme from an uncultured bacterium	Activity-based screening of broad-host cosmid shuttle vector library expressed in *Ralstonia metallidurans*	Soil, Sonoran Desert, Arizona, USA	Iqbal et al., 2014
Six clones encoding a lysostaphin gene	All six clones expressed the lysostaphin gene from the *Staphylococcus simulans* library strain	High throughput activity-based screening of *E. coli* and *Saccharomyces cerevisiae* plasmid libraries	Library derived from three native staphylococcal strains: *S. simulans*, *S. arlettae*, and *S. equorum*	Scanlon et al., 2014
ANTI INFECTIVE				
Two novel lactonases	One had 53% similarity to amino acid sequence from *Pseudomonas fluorescens.* The other, 57% similarity to *Nitrobacter* sp. Strain Nb-311A	Activity-based screening of *E. coli* phagemid vector, plasmid and broad-host-range vector library	Soil, University of Göttingen, Germany	Schipper et al., 2009

(Continued)

TABLE 3 | Continued

Antimicrobial	Closest known relationship (percentage homology)	Method/Host	Environment	References
Clone expressing NAHL-lactonase activity	Most closely related to Zn-dependent hydrolase from *Bradyrhizobium* sp.	Functional-based screening of *E. coli* fosmid library	Pasture soil, France	Riaz et al., 2008
Two novel pairs of LuxR/LuxI genes	QS pair 1: LuxI homolog: 42% amino acid similarity to putative LuxI in *Geobacter uraniireducens* Rf4. 38% protein sequence similarity to CviI in *Chromobacterium violaceum* ATCC 31532. LuxR homolog: 33% amino acid similarity to LuxR from *Geobacter* sp. strain FRC-32 and 31% to CviR from *C. violaceum* QS pair 2: LuxI homolog: 57% similar to LuxIQS6-1 of a metagenomic clone and 40% amino acid similarity to PpuI from *Pseudomonas putida*. LuxR homolog: 37% similarity to LuxRQS10-1 in a metagenomic clone and 35% similarity to BraR in *Burkholderia kururiensis*	Activity-based screening of two fosmid libraries expressed in a biosensor *E. coli* host	Activated sludge from a coke plant, Japan. Forest soil samples, Tsukuba city, Japan	Nasuno et al., 2012
Novel bacterial NAHLase	Most likely belonging to species of unknown Proteobacterium	Activity-based screening using an *Agrobacterium tumefaciens* biosensor strain of four *E. coli* fosmid libraries	Rhizosphere of *Solanum tuberosum* that was treated with γ-caprolactone	Tannieres et al., 2013
Three novel pair of LuxR/LuxI genes	QS pair 1: 47% identity to *Nitrosospira multiformis* ATCC 25196 and 34% to *Nitrococcus mobilis* Nb-231 QS pair 2: 51% and 32% identity to *Nitrosospira multiformis* ATCC 25196 QS pair 3: both genes had 37% identity to proteins from *Sphingomonas* sp. SKA58	Activity-based screening using an *Agrobacterium tumefaciens* biosensor strain of four *E. coli* plasmid libraries	Activated sludge Soil	Hao et al., 2010
Novel NADP-dependent short-chain dehydrogenase/reductase	61% identical to chromosome segregation protein SMC in *Acidobacterium* sp. MP5ACTX8	Activity-based screening of *E. coli* phagemid vector, plasmid, and broad-host-range vector library	Soil, University of Göttingen, Germany	Bijtenhoorn et al., 2011
ANTIBIOTIC RESISTANCE DETERMINANT				
Novel florfenicol and chloramphenicol resistance gene	33% amino acid identity to drug resistance transporters from *Wolbachia* spp. (YP_002726856, YP_198189, and NP_966057)	Function-based screening of *E. coli* fosmid library	Soil samples from an island in the Tanana River near Fairbanks, Alaska	Lang et al., 2010
Two novel genes conferring resistance to kanamycin and ceftazidime	Both showed highest similarity to uncultured soil microorganisms	Activity-based screening of *E. coli* fosmid library	Soil from apple orchard, southern Wisconsin	Donato et al., 2010
Resistance genes to chloramphenicol, ampicillin and kanamycin. Multidrug resistant clone conferring ampicillin and kanamycin resistance	Multidrug resistant clone showed highest identity (95%) to a ß-lactamase from *Bacillus* sp. BT-192. For chloramphenicol resistance, highest homology was seen to a hypothetical protein from *Methylibium petroleiphilum.* A kanamycin resistant clone showed 55% identity to a *Microscilla* sp. protein. An ampicillin resistant clone showed 66% identity to a ß-lactamase from *Spirosoma linguale*	Functional screening of metagenomic BAC, plasmid, and phagemid vector libraries expressed in *E. coli*. Sequencing of small insert libraries.	Activated sludge	Parsley et al., 2010
Novel chloramphenicol hydrolase (resistance to chloramphenicol and florfenicol)	14 ORFs varying in similarity (30–77%) to corresponding proteins from known microorganisms. Highest similarity overall to proteins from the bacterial phylum *Proteobacteria*	Activity-based screening of *E. coli* plasmid library	Alluvial soil	Tao et al., 2012
Novel carboxylesterase	Highest identity (58%) to ß-lactamase (YP_004154831) from *Variovorax paradoxus* EPS	Activity-based screening of *E. coli* cosmid library	Soil from the Upo wetland, South Korea	Jeon et al., 2011

(Continued)

TABLE 3 | Continued

Antimicrobial	Closest known relationship (percentage homology)	Method/Host	Environment	References
31 previously undescribed antibiotic resistance genes to ampicillin, amoxicillin, tetracycline, and penicillin. This includes class A and C β-lactamases and six different tetracycline resistance genes	Significant similarity to proteins from multiple genera from the ARDB and GenBank databases	Activity-based screening of *E. coli* plasmid library	Fecal samples of Herring gulsl, Appledore Island, ME and Rochester, NH, USA	Martiny et al., 2011
39 clones conferring resistance to kanamycin, gentamicin, chloramphenicol, rifampin, trimethoprim, and tetracycline	Highest homology to the following phyla: *Proteobacteria*, *Actinobacteria,* and *Firmicutes*	Activity-based screening of *E. coli* plasmid library	Urban soil, Seattle, WA, USA	McGarvey et al., 2012
110 antibiotic resistance genes conferring resistance to ß-lactams, aminoglycosides, amphenicols, sulfonamides, and tetracyclines, including 55 ß-lactamases	18 resistance genes showed 100% identity to known human pathogens	Activity-based screening of metagenomic library expressed in *E. coli* coupled with PARFuMS	11 soil samples, USA	Forsberg et al., 2012
95 unique antimicrobial resistance eDNA inserts. 10 novel β-lactamase gene families	Average of 69.5% nucleotide identity to GenBank sequences. 15 β-lactamase resistance genes showed high identity (>90%) to known human pathogens	Activity-based screening of metagenomic library expressed in *E. coli*	Human saliva and fecal samples	Sommer et al., 2009
A novel kanamycin resistance gene fusion (to a hypothetical protein domain)	N-terminus was 42% identical to AAC(6') from *Enterococcus hirae.* C-terminus was 35% identical to a hypothetical protein (CBL37632) from *Clostridiales* sp. SSC/2	Activity-based screening of *E. coli* fosmid library	Four human fecal samples	Cheng et al., 2012c
45 clones resistant to tetracycline, minocycline, aminoglycosides, streptomycin, gentamicin, kanamycin, amikacin, chloramphenicol, and rifampicin	26–92% similarity to known proteins in the GenBank database	Activity-based screening of *E. coli* plasmid library	Four agricultural soil samples, China	Su et al., 2014
Five clones conferring Fluoroquinolone resistance, cephalosporin resistance, and trimethoprim resistance	High similarity to homologs in species of *Bacillus*, *Erwinia*, *Exiguobacterium*, *Pseudomonas*	Activity-based screening of two *E. coli* plasmid libraries from cultured spinach microbiota and from uncultured spinach wash	Retail spinach	Berman and Riley, 2013
Ampicillin resistance and kanamycin resistance	Homology to *Streptococcus thermophilus* and *Lactobacillus helveticus*	Activity-based screening of an *E. coli* fosmid library	Mozzarella di Bufala Campana (MBC) Cheese, produced in Central and Southern Itlaly	Devirgiliis et al., 2014

to identify novel resistance mechanisms used by bacteria in nature which may not have manifested in the clinical setting yet and so can allow one to predict possible routes *via* which resistance to current antibiotic therapies could emerge. The studies discussed below provide insight into the diversity of antimicrobial resistance mechanisms, proposing new avenues of research for tackling antibiotic resistance. They also show the value of functional metagenomics as a tool for the investigation of antimicrobial resistance, as reviewed by Mullany (2014). Donato et al. (2010) screened a metagenomic apple orchard soil library for DNA fragments that conferred antibiotic resistance to their *E. coli* host. Clones were screened for resistance to a

selection of 10 antibiotics. The group reported the discovery of two novel enzymes. In one case, a metagenomic clone encoding an aminoglycoside acetyltransferase domain fused to a second acetyltransferase domain displayed resistance to kanamycin. Interestingly, sequence analysis of this clone did not predict antimicrobial resistance. The second interesting clone encoded a bifunctional protein containing a natural fusion of a ß-lactamase and a sigma factor conferring onto the host resistance to ceftazidime. Additional potential chloramphenicol resistance was predicted by sequencing this particular clone, which may evoke further analysis. Tao et al. (2012) used a TLC-based method to screen an alluvial soil-derived metagenomic library for chloramphenicol resistance. They identified a resistant clone harboring a hydrolysate which conferred to the host resistance to chloramphenicol and florfenicol, a synthetic form of chloramphenicol that was employed as a safe antibiotic treatment for use in farming. The enzyme was capable of hydrolyzing both chloramphenicol and florfenicol, with greater efficiency at hydrolyzing florfenicol. Various metagenomic studies have been carried out to identify antimicrobial resistance genes in certain foods. Antibiotic therapies for the treatment of bacterial infections in farm animals select for resistant microbes in food production chains (Hawkey, 2008). Although most microorganisms in foodstuffs are usually not pathogenic, resistant bacteria that survive on products for human consumption may transfer their resistance to opportunistic human pathogens or to the human microbiota. Certain foods (e.g., foods eaten raw) and the human gut microbiota itself may then potentially become a reservoir for antibiotic resistance genes. Retail spinach is commonly eaten raw and thus, has been linked to outbreaks of bacterial infections (Lynch et al., 2009; Wendel et al., 2009). Berman and Riley (2013) functionally screened two spinach-derived metagenomic libraries for resistance to 16 different antimicrobial agents, identifying numerous novel genes conferring resistance to ampicillin, aztreonam, ciprofloxacin, trimethoprim, and trimethoprim-sulfamethoxazole from five different active clones. Their study suggests that microorganisms in close contact with fresh food products, such as plant commensals and saprophytes, may serve as a reservoir of antimicrobial resistance genes. In a study with a similar objective, Devirgiliis et al. (2014) isolated clones displaying ampicillin and kanamycin resistance from a metagenomic library constructed from Mozzarella di Bufala Campana Italian cheese. These studies ultimately show that food products can potentially harbor bacterial species possessing clinically relevant antimicrobial resistance which may be horizontally transferred to pathogens, either directly or by an indirect route through the human microbiota.

Unusual or unexpected antimicrobial resistance mechanisms can be found in nature. Some studies investigating the resistome of uncultured bacteria have explored areas and environments which have not been previously exposed to clinical antibiotics and where endogenous microorganisms have therefore not faced selective pressure to develop antibiotic resistance. A recent study by Fouhy et al. (2014) examined the resistome of the naïve infant gut. A metagenomic library constructed from fecal samples of 22 six-month old infants who had not previously been exposed to antibiotics was screened for resistance to aminoglycoside and β-lactam antibiotics, identifying gentamicin and ampicillin resistant clones. PCR analyses were also carried out to detect DNA sequences encoding aminoglycoside and β-lactam resistance genes not successfully cloned and expressed in the library. One hundred ampicillin resistant clones were identified in their functional screen, conferring resistance *via* several β-lactamase genes. Aminoglycoside resistant clones were also identified, whose resistance was conferred by acetylation, adenylation, and phosphorylation genes. This study uncovered resistance to clinically relevant antibiotics in a naïve environment. Other studies assessing the resistome of microbial samples from remote areas where little or no antibiotic therapy has been practiced have also identified unexpected resistance (Pallecchi et al., 2008; Bartoloni et al., 2009). More recently, Clemente et al. (2015) examined the bacterial microbiome (from fecal, oral, and skin samples) of 34 Yanomami individuals from an isolated Amerindian village in South America. Among huge microbial diversity observed through 16S rRNA gene sequencing of DNA from the obtained samples, activity-based and culture-independent screening of functional and shotgun metagenomic libraries also revealed resistance genes to clinically relevant antibiotics. These studies further emphasize the diversity of the as yet uncultured microbial world by establishing that genes conveying resistance to current antibiotic therapies can be found in environments void of selective pressure.

Conclusions and Future Prospects

Metagenomics grants access to the huge diversity of the microbial world and has led to significant progress among research communities and in industrial settings with respect to understanding and benefitting from unculturable microbes. Functional metagenomics is a powerful tool for the discovery of novel enzymes and bioactives sourced from as yet uncultured microorganisms. As a relatively new technology, functional metagenomics faces challenges that have yet to be overcome. However, the promise of a technique that has already proven to be fruitful even in its early years suggests that there can be significant rewards if appropriate solutions and further optimization takes place. The development of new screening and selection techniques along with faster and cheaper sequencing technologies will allow for the expansion of a very promising field in microbiology, genetics and the food and pharmaceutical industries.

This article discusses the potential of functional metagenomics to facilitate the development of novel industrial products sourced from as yet uncultured microorganisms. Nonetheless, following the identification of useful proteins and bioactives, challenges ensue in another area, that being the development of a consumer friendly and commercially viable product that can be manufactured in industrially relevant quantities, retains its activity when scaled up (for example when present in high amounts in a large industrial reaction vessel), can be purified and formulated appropriately into a finished product and maintains its stability during shipping and storage. The product also

needs to be reasonably easy to use and must be applicable to current industrial demands, i.e., the product must perform efficiently under the proposed/outlined conditions to carry out the job it was bought to do. A successful reaction achieved under laboratory conditions may be difficult to reproduce on an industrial scale. Pilot plant studies must be carried out initially to identify any variables or short comings in the reaction that were not evident at the laboratory stages of development. These studies are a stepping stone between discovery of the interesting active agent and its formulation into a final commercial product. Once deficiencies and other problems have been corrected in the pilot plant phase, further studies must be conducted to qualify the agent at an industrial level and guarantee the development of a robust product that is efficient and true to its intended purpose. The acceptability of the novel enzyme or bioactive and its source microorganism to the relevant regulatory authorities must also be considered.

Once all these limitations are overcome, through access to the seemingly infinite diversity of the microbial world, functional metagenomics presents an opportunity to develop novel innovative products that offer something new and useful to industrial processes or even change for the better or make more convenient the way a current process is carried out.

Acknowledgments

The financial support of Science Foundation Ireland (SFI) under Grant Number 13/SIRG/2157 is acknowledged.

References

Abumweis, S. S., Jew, S., and Ames, N. P. (2010). Beta-glucan from barley and its lipid-lowering capacity: a meta-analysis of randomized, controlled trials. *Eur. J. Clin. Nutr.* 64, 1472–1480. doi: 10.1038/ejcn.2010.178

Alanis, A. J. (2005). Resistance to antibiotics: are we in the post-antibiotic era? *Arch. Med. Res.* 36, 697–705. doi: 10.1016/j.arcmed.2005.06.009

Ashie, I. N. A., Sorensen, T. L., and Nielsen, P. M. (2002). Effects of papain and a microbial enzyme on meat proteins and beef tenderness. *J. Food Sci.* 67, 2138–2142. doi: 10.1111/j.1365-2621.2002.tb09516.x

Banik, J. J., and Brady, S. F. (2010). Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr. Opin. Microbiol.* 13, 603–609. doi: 10.1016/j.mib.2010.08.012

Bartoloni, A., Pallecchi, L., Rodriguez, H., Fernandez, C., Mantella, A., Bartalesi, F., et al. (2009). Antibiotic resistance in a very remote Amazonas community. *Int. J. Antimicrob. Agents* 33, 125–129. doi: 10.1016/j.ijantimicag.2008.07.029

Bates, S. T., Berg-Lyons, D., Lauber, C. L., Walters, W. A., Knight, R., and Fierer, N. (2012). A preliminary survey of lichen associated eukaryotes using pyrosequencing. *Lichenologist* 44, 137–146. doi: 10.1017/S0024282911000648

Berman, H. F., and Riley, L. W. (2013). Identification of novel antimicrobial resistance genes from microbiota on retail spinach. *BMC Microbiol.* 13:272. doi: 10.1186/1471-2180-13-272

Bijtenhoorn, P., Mayerhofer, H., Muller-Dieckmann, J., Utpatel, C., Schipper, C., Hornung, C., et al. (2011). A novel metagenomic short-chain dehydrogenase/reductase attenuates Pseudomonas aeruginosa biofilm formation and virulence on *Caenorhabditis elegans*. *PLoS ONE* 6:e26278. doi: 10.1371/journal.pone.0026278

Biver, S., Portetelle, D., and Vandenbol, M. (2013a). Characterization of a new oxidant-stable serine protease isolated by functional metagenomics. *Springerplus* 2:410. doi: 10.1186/2193-1801-2-410

Biver, S., Steels, S., Portetelle, D., and Vandenbol, M. (2013b). Bacillus subtilis as a tool for screening soil metagenomic libraries for antimicrobial activities. *J. Microbiol. Biotechnol.* 23, 850–855. doi: 10.4014/jmb.1212.12008

Biver, S., and Vandenbol, M. (2013). Characterization of three new carboxylic ester hydrolases isolated by functional screening of a forest soil metagenomic library. *J. Ind. Microbiol. Biotechnol.* 40, 191–200. doi: 10.1007/s10295-012-1217-7

Boadi, D. K., and Neufeld, R. J. (2001). Encapsulation of tannase for the hydrolysis of tea tannins. *Enzyme Microb. Technol.* 28, 590–595. doi: 10.1016/S0141-0229(01)00295-2

Bornscheuer, U. T. (2002). Microbial carboxyl esterases: classification, properties and application in biocatalysis. *FEMS Microbiol. Rev.* 26, 73–81. doi: 10.1111/j.1574-6976.2002.tb00599.x

Brady, S. F., Chao, C. J., and Clardy, J. (2004). Long-chain N-acyltyrosine synthases from environmental DNA. *Appl. Environ. Microbiol.* 70, 6865–6870. doi: 10.1128/AEM.70.11.6865-6870.2004

Brady, S. F., Chao, C. J., Handelsman, J., and Clardy, J. (2001). Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org. Lett.* 3, 1981–1984. doi: 10.1021/ol015949k

Brady, S. F., and Clardy, J. (2000). Long-chain N-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. *J. Am. Chem. Soc.* 122, 12903–12904. doi: 10.1021/ja002990u

Brady, S. F., and Clardy, J. (2005). Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angew. Chem. Int. Ed Engl.* 44, 7063–7065. doi: 10.1002/anie.200501941

Cantarelli, C., Brenna, O., Giovanelli, G., and Rossi, M. (1989). Beverage stabilization through enzymatic removal of phenolics. *Food Biotechnol.* 3, 203–213. doi: 10.1080/08905438909549709

Cecchini, D. A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Dore, J., et al. (2013). Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. *PLoS ONE* 8:e72766. doi: 10.1371/journal.pone.0072766

Chandler, J. A., Thongsripong, P., Green, A., Kittayapong, P., Wilcox, B. A., Schroth, G. P., et al. (2014). Metagenomic shotgun sequencing of a Bunyavirus in wild-caught *Aedes aegypti* from Thailand informs the evolutionary and genomic history of the Phleboviruses. *Virology* 464–465, 312–319. doi: 10.1016/j.virol.2014.06.036

Chang, F. Y., and Brady, S. F. (2013). Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *Proc. Natl. Acad. Sci. U.S.A.* 110, 2478–2483. doi: 10.1073/pnas.1218073110

Cheng, F. S., Sheng, J. P., Cai, T., Jin, J., Liu, W. Z., Lin, Y. M., et al. (2012a). A protease-insensitive feruloyl esterase from China holstein cow rumen metagenomic library: expression, characterization, and utilization in ferulic acid release from wheat straw. *J. Agric. Food Chem.* 60, 2546–2553. doi: 10.1021/jf204556u

Cheng, F. S., Sheng, J. P., Dong, R. B., Men, Y. J., Gan, L., and Shen, L. (2012b). Novel xylanase from a holstein cattle rumen metagenomic library and its application in xylooligosaccharide and ferulic acid production from wheat straw. *J. Agric. Food Chem.* 60, 12516–12524. doi: 10.1021/jf302337w

Cheng, G., Hu, Y., Yin, Y., Yang, X., Xiang, C., Wang, B., et al. (2012c). Functional screening of antibiotic resistance genes from human gut microbiota reveals a novel gene fusion. *FEMS Microbiol. Lett.* 336, 11–16. doi: 10.1111/j.1574-6968.2012.02647.x

Chu, X. M., He, H. Z., Guo, C. Q., and Sun, B. L. (2008). Identification of two novel esterases from a marine metagenomic library derived from South China Sea. *Appl. Microbiol. Biotechnol.* 80, 615–625. doi: 10.1007/s00253-008-1566-3

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., et al. (2015). The microbiome of uncontacted Amerindians. *Sci. Adv.* 1:e1500183. doi: 10.1126/sciadv.1500183

Cooper, M. A., and Shlaes, D. (2011). Fix the antibiotics pipeline. *Nature* 472, 32. doi: 10.1038/472032a

Craig, J. W., Chang, F. Y., and Brady, S. F. (2009). Natural products from environmental DNA hosted in *Ralstonia metallidurans*. *ACS Chem. Biol.* 4, 23–28. doi: 10.1021/cb8002754

Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C., and Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries

in diverse proteobacteria. *Appl. Environ. Microbiol.* 76, 1633–1641. doi: 10.1128/AEM.02169-09

Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. (2012). Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. *ISME J.* 6, 1916–1925. doi: 10.1038/ismej.2012.38

Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. (2013). Functional environmental screening of a metagenomic library identifies stlA; a unique salt tolerance locus from the human gut microbiome. *PLoS ONE* 8:e82985. doi: 10.1371/journal.pone.0082985

Devirgiliis, C., Zinno, P., Stirpe, M., Barile, S., and Perozzi, G. (2014). Functional screening of antibiotic resistance genes from a representative metagenomic library of food fermenting microbiota. *Biomed Res. Int.* 2014:290967. doi: 10.1155/2014/290967

Donato, J. J., Moe, L. A., Converse, B. J., Smart, K. D., Berklein, F. C., McManus, P. S., et al. (2010). Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins. *Appl. Environ. Microbiol.* 76, 4396–4401. doi: 10.1128/AEM.01763-09

Elend, C., Schmeisser, C., Hoebenreich, H., Steele, H. L., and Streit, W. R. (2007). Isolation and characterization of a metagenome-derived and cold-active lipase with high stereospecificity for (R)-ibuprofen esters. *J. Biotechnol.* 130, 370–377. doi: 10.1016/j.jbiotec.2007.05.015

Elend, C., Schmeisser, C., Leggewie, C., Babiak, P., Carballeira, J. D., Steele, H. L., et al. (2006). Isolation and biochemical characterization of two novel metagenome-derived esterases. *Appl. Environ. Microbiol.* 72, 3637–3645. doi: 10.1128/AEM.72.5.3637-3645.2006

Entcheva, P., Liebl, W., Johann, A., Hartsch, T., and Streit, W. R. (2001). Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. *Appl. Environ. Microbiol.* 67, 89–99. doi: 10.1128/AEM.67.1.89-99.2001

Faoro, H., Glogauer, A., Couto, G. H., de Souza, E. M., Rigo, L. U., Cruz, L. M., et al. (2012). Characterization of a new Acidobacteria-derived moderately thermostable lipase from a Brazilian Atlantic Forest soil metagenome. *FEMS Microbiol. Ecol.* 81, 386–394. doi: 10.1111/j.1574-6941.2012.01361.x

Faulds, C. B. (2010). What can feruloyl esterases do for us? *Phytochem. Rev.* 9, 121–132. doi: 10.1007/s11101-009-9156-2

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Martins Dos Santos, V. A., Yakimov, M. M., et al. (2005a). Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem. Biol.* 12, 895–904. doi: 10.1016/j.chembiol.2005.05.020

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Reyes-Duarte, D., Dos Santos, V., et al. (2005b). Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ. Microbiol.* 7, 1996–2010. doi: 10.1111/j.1462-2920.2005.00920.x

Fischbach, M. A., and Walsh, C. T. (2009). Antibiotics for emerging pathogens. *Science* 325, 1089–1093. doi: 10.1126/science.1176667

Forsberg, K. J., Reyes, A., Wang, B., Selleck, E. M., Sommer, M. O., and Dantas, G. (2012). The shared antibiotic resistome of soil bacteria and human pathogens. *Science* 337, 1107–1111. doi: 10.1126/science.1220761

Fouhy, F., Ogilvie, L. A., Jones, B. V., Ross, R. P., Ryan, A. C., Dempsey, E. M., et al. (2014). Identification of aminoglycoside and beta-lactam resistance genes from within an infant gut functional metagenomic library. *PLoS ONE* 9:e0108016. doi: 10.1371/journal.pone.0108016

Fujita, M. J., Kimura, N., Sakai, A., Ichikawa, Y., Hanyu, T., and Otsuka, M. (2011). Cloning and heterologous expression of the vibrioferrin biosynthetic gene cluster from a marine metagenomic library. *Biosci. Biotechnol. Biochem.* 75, 2283–2287. doi: 10.1271/bbb.110379

Gibson, G. R., and Roberfroid, M. B. (1995). Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *J. Nutr.* 125, 1401–1412.

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., et al. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* 68, 4301–4306. doi: 10.1128/AEM.68.9.4301-4306.2002

Guazzaroni, M. E., Morgante, V., Mirete, S., and Gonzalez-Pastor, J. E. (2013). Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. *Environ. Microbiol.* 15, 1088–1102. doi: 10.1111/1462-2920.12021

Gupta, R., Gupta, N., and Rathi, P. (2004). Bacterial lipases: an overview of production, purification and biochemical properties. *Appl. Microbiol. Biotechnol.* 64, 763–781. doi: 10.1007/s00253-004-1568-8

Hafeez, Z., Cakir-Kiefer, C., Roux, E., Perrin, C., Mido, L., and Dary-Mourot, A. (2014). Strategies of producing bioactive peptides from milk proteins to functionalize fermented milk products. *Food Res. Int.* 63, 71–80. doi: 10.1016/j.foodres.2014.06.002

Hamada, S., Suzuki, K., Aoki, N., and Suzuki, Y. (2013). Improvements in the qualities of gluten-free bread after using a protease obtained from *Aspergillus oryzae*. *J. Cereal Sci.* 57, 91–97. doi: 10.1016/j.jcs.2012.10.008

Hao, Y., Winans, S. C., Glick, B. R., and Charles, T. C. (2010). Identification and characterization of new LuxR/LuxI-type quorum sensing systems from metagenomic libraries. *Environ. Microbiol.* 12, 105–117. doi: 10.1111/j.1462-2920.2009.02049.x

Hasan, F., Shah, A. A., and Hameed, A. (2006). Industrial applications of microbial lipases. *Enzyme Microb. Technol.* 39, 235–251. doi: 10.1016/j.enzmictec.2005.10.016

Hawkey, P. M. (2008). The growing burden of antimicrobial resistance. *J. Antimicrob. Chemother.* 62, I1–I9. doi: 10.1093/jac/dkn241

Henne, A., Schmitz, R. A., Bomeke, M., Gottschalk, G., and Daniel, R. (2000). Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl. Environ. Microbiol.* 66, 3113–3116. doi: 10.1128/AEM.66.7.3113-3116.2000

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331, 463–467. doi: 10.1126/science.1200387

Hjort, K., Presti, I., Elvang, A., Marinelli, F., and Sjoling, S. (2014). Bacterial chitinase with phytopathogen control capacity from suppressive soil revealed by functional metagenomics. *Appl. Microbiol. Biotechnol.* 98, 2819–2828. doi: 10.1007/s00253-013-5287-x

Hoessel, R., Leclerc, S., Endicott, J. A., Nobel, M. E. M., Lawrie, A., Tunnah, P., et al. (1999). Indirubin, the active constituent of a Chinese antileukaemia medicine, inhibits cyclin-dependent kinases. *Nat. Cell Biol.* 1, 60–67.

Huang, X. W., Chen, L. J., Luo, Y. B., Guo, H. Y., and Ren, F. Z. (2011). Purification, characterization, and milk coagulating properties of ginger proteases. *J. Dairy Sci.* 94, 2259–2269. doi: 10.3168/jds.2010-4024

Iqbal, H. A., Craig, J. W., and Brady, S. F. (2014). Antibacterial enzymes from the functional screening of metagenomic libraries hosted in *Ralstonia metallidurans*. *FEMS Microbiol. Lett.* 354, 19–26. doi: 10.1111/1574-6968.12431

Jayasuriya, H., Herath, K. B., Zhang, C., Zink, D. L., Basilio, A., Genilloud, O., et al. (2007). Isolation and structure of platencin: a FabH and FabF dual inhibitor with potent broad-spectrum antibiotic activity. *Angew. Chem. Int. Ed Engl.* 46, 4684–4688. doi: 10.1002/anie.200701058

Jeon, J. H., Kim, S. J., Lee, H. S., Cha, S. S., Lee, J. H., Yoon, S. H., et al. (2011). Novel metagenome-derived carboxylesterase that hydrolyzes beta-lactam antibiotics. *Appl. Environ. Microbiol.* 77, 7830–7836. doi: 10.1128/AEM.05363-11

Jiang, C. J., Hao, Z. Y., Zeng, R., Shen, P. H., Li, J. F., and Wu, B. (2011). Characterization of a novel serine protease inhibitor gene from a marine metagenome. *Mar. Drugs* 9, 1487–1501. doi: 10.3390/md9091487

Jiang, C. J., Ma, G. F., Li, S. X., Hu, T. T., Che, Z. Q., Shen, P. H., et al. (2009). Characterization of a novel beta-glucosidase-like activity from a soil metagenome. *J. Microbiol.* 47, 542–548. doi: 10.1007/s12275-009-0024-y

Joint, F. (2001). *WHO Expert Consultation on Evaluation of Health and Nutritional Properties of Probiotics in Food Including Powder Milk with Live Lactic Acid Bacteria*. Córdoba: Food and Agriculture Organization of the United Nations and the World Health Organization.

Kang, C. H., Oh, K. H., Lee, M. H., Oh, T. K., Kim, B. H., and Yoon, J. (2011). A novel family VII esterase with industrial potential from compost metagenomic library. *Microb. Cell Fact.* 10:41. doi: 10.1186/1475-2859-10-41

Kato, K., Ono, M., and Akita, H. (1997). New total synthesis of (-)- and (+)-chuangxinmycins. *Tetrahedron-Asymmetry* 8, 2295–2298. doi: 10.1016/S0957-4166(97)00253-X

Kay, G. L., Sergeant, M. J., Giuffra, V., Bandiera, P., Milanese, M., Bramanti, B., et al. (2014). Recovery of a medieval Brucella melitensis genome using shotgun metagenomics. *MBio* 5, e01337-14. doi: 10.1128/mBio.01337-14

Keller, A., Horn, H., Forster, F., and Schultz, J. (2014). Computational integration of genomic traits into 16S rDNA microbiota sequencing studies. *Gene* 549, 186–191. doi: 10.1016/j.gene.2014.07.066

Kim, Y. J., Choi, G. S., Kim, S. B., Yoon, G. S., Kim, Y. S., and Ryu, Y. W. (2006). Screening and characterization of a novel esterase from a metagenomic library. *Protein Expr. Purif.* 45, 315–323. doi: 10.1016/j.pep.2005.06.008

Kimura, N. (2014). Metagenomic approaches to understanding phylogenetic diversity in quorum sensing. *Virulence* 5, 433–442. doi: 10.4161/viru.27850

Lakhdari, O., Cultrone, A., Tap, J., Gloux, K., Bernard, F., Ehrlich, S. D., et al. (2010). Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-kappa B modulation in the human gut. *PLoS ONE* 5:e13092. doi: 10.1371/journal.pone.0013092

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid-determination of 16s Ribosomal-Rna sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U.S.A.* 82, 6955–6959. doi: 10.1073/pnas.82.20.6955

Lang, K. S., Anderson, J. M., Schwarz, S., Williamson, L., Handelsman, J., and Singer, R. S. (2010). Novel florfenicol and chloramphenicol resistance gene discovered in Alaskan soil by using functional metagenomics. *Appl. Environ. Microbiol.* 76, 5321–5326. doi: 10.1128/AEM.00323-10

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Lee, D. G., Jeon, J. H., Jang, M. K., Kim, N. Y., Lee, J. H., Lee, J. H., et al. (2007). Screening and characterization of a novel fibrinolytic metalloprotease from a metagenomic library. *Biotechnol. Lett.* 29, 465–472. doi: 10.1007/s10529-006-9263-8

Lee, S. W., Won, K., Lim, H. K., Kim, J. C., Choi, G. J., and Cho, K. Y. (2004). Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl. Microbiol. Biotechnol.* 65, 720–726. doi: 10.1007/s00253-004-1722-3

Li, J. W., and Vederas, J. C. (2009). Drug discovery and natural products: end of an era or an endless frontier? *Science* 325, 161–165. doi: 10.1126/science.1168243

Lim, H. K., Chung, E. J., Kim, J. C., Choi, G. J., Jang, K. S., Chung, Y. R., et al. (2005). Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Appl. Environ. Microbiol.* 71, 7768–7777. doi: 10.1128/AEM.71.12.7768-7777.2005

Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., et al. (2015). A new antibiotic kills pathogens without detectable resistance. *Nature* 517, 455–459. doi: 10.1038/nature14098

Lynch, M. F., Tauxe, R. V., and Hedberg, C. W. (2009). The growing burden of foodborne outbreaks due to contaminated fresh produce: risks and opportunities. *Epidemiol. Infect.* 137, 307–315. doi: 10.1017/S0950268808001969

Macneil, I. A., Tiong, C. L., Minor, C., August, P. R., Grossman, T. H., Loiacono, K. A., et al. (2001). Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J. Mol. Microbiol. Biotechnol.* 3, 301–308.

Marko, D., Schatzle, S., Friedel, A., Genzlinger, A., Zankl, H., Meijer, L., et al. (2001). Inhibition of cyclin-dependent kinase 1 (CDK1) by indirubin derivatives in human tumour cells. *Br. J. Cancer* 84, 283–289. doi: 10.1054/bjoc.2000.1546

Martiny, A. C., Martiny, J. B. H., Weihe, C., Field, A., and Ellis, J. C. (2011). Functional metagenomics reveals previously unrecognized diversity of antibiotic resistance genes in gulls. *Front. Microbiol.* 2:238. doi: 10.3389/fmicb.2011.00238

McGarvey, K. M., Queitsch, K., and Fields, S. (2012). Wide variation in antibiotic resistance proteins identified by functional metagenomic screening of a soil DNA library. *Appl. Environ. Microbiol.* 78, 1708–1714. doi: 10.1128/AEM.06759-11

Mendes, L. W., Tsai, S. M., Navarrete, A. A., de Hollander, M., van Veen, J. A., and Kuramae, E. E. (2015). Soil-borne microbiome: linking diversity to function. *Microb. Ecol.* 70, 255–265. doi: 10.1007/s00248-014-0559-2

Mora, L., Reig, M., and Toldra, F. (2014). Bioactive peptides generated from meat industry by-products. *Food Res. Int.* 65, 344–349. doi: 10.1016/j.foodres.2014.09.014

Moure, A., Gullon, P., Dominguez, H., and Parajo, J. C. (2006). Advances in the manufacture, purification and applications of xylo-oligosaccharides as food additives and nutraceuticals. *Process Biochem.* 41, 1913–1923. doi: 10.1016/j.procbio.2006.05.011

Mullany, P. (2014). Functional metagenomics for the investigation of antibiotic resistance. *Virulence* 5, 443–447. doi: 10.4161/viru.28196

Nacke, H., Will, C., Herzog, S., Nowka, B., Engelhaupt, M., and Daniel, R. (2011). Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories. *FEMS Microbiol. Ecol.* 78, 188–201. doi: 10.1111/j.1574-6941.2011.01088.x

Nasuno, E., Kimura, N., Fujita, M. J., Nakatsu, C. H., Kamagata, Y., and Hanada, S. (2012). Phylogenetically novel LuxI/LuxR-type quorum sensing systems isolated using a metagenomic approach. *Appl. Environ. Microbiol.* 78, 8067–8074. doi: 10.1128/AEM.01442-12

Neveu, J., Regeard, C., and Dubow, M. S. (2011). Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts. *Appl. Microbiol. Biotechnol.* 91, 635–644. doi: 10.1007/s00253-011-3256-9

Oh, J., Byrd, A. L., Deming, C., Conlan, S., Kong, H. H., Segre, J. A., et al. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59–64. doi: 10.1038/nature13786

Ouyang, L. M., Liu, J. Y., Qiao, M., and Xu, J. H. (2013). Isolation and biochemical characterization of two novel metagenome-derived esterases. *Appl. Biochem. Biotechnol.* 169, 15–28. doi: 10.1007/s12010-012-9949-4

Palackal, N., Lyon, C. S., Zaidi, S., Luginbuhl, P., Dupree, P., Goubet, F., et al. (2007). A multifunctional hybrid glycosyl hydrolase discovered in an uncultured microbial consortium from ruminant gut. *Appl. Microbiol. Biotechnol.* 74, 113–124. doi: 10.1007/s00253-006-0645-6

Pallecchi, L., Bartoloni, A., Paradisi, F., and Rossolini, G. M. (2008). Antibiotic resistance in the absence of antimicrobial use: mechanisms and implications. *Expert Rev. Anti Infect. Ther.* 6, 725–732. doi: 10.1586/14787210.6.5.725

Panda, T., and Gowrishankar, B. S. (2005). Production and applications of esterases. *Appl. Microbiol. Biotechnol.* 67, 160–169. doi: 10.1007/s00253-004-1840-y

Panesar, P. S., Kumari, S., and Panesar, R. (2010). Potential applications of immobilized beta-galactosidase in food processing industries. *Enzyme Res.* 2010:473137. doi: 10.4061/2010/473137

Parsley, L. C., Consuegra, E. J., Kakirde, K. S., Land, A. M., Harper, W. F. Jr., and Liles, M. R. (2010). Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl. Environ. Microbiol.* 76, 3753–3757. doi: 10.1128/AEM.03080-09

Peng, Q., Wang, X., Shang, M., Huang, J., Guan, G., Li, Y., et al. (2014). Isolation of a novel alkaline-stable lipase from a metagenomic library and its specific application for milkfat flavor production. *Microb. Cell Fact.* 13:1. doi: 10.1186/1475-2859-13-1

Piel, J. (2002). A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14002–14007. doi: 10.1073/pnas.222481399

Pooja, S., Pushpanathan, M., Jayashree, S., Gunasekaran, P., and Rajendhran, J. (2015). Identification of periplasmic alpha-amlyase from cow dung metagenome by product induced gene expression profiling (Pigex). *Indian J. Microbiol.* 55, 57–65. doi: 10.1007/s12088-014-0487-3

Pushpam, P. L., Rajesh, T., and Gunasekaran, P. (2011). Identification and characterization of alkaline serine protease from goat skin surface metagenome. *AMB Express* 1:3. doi: 10.1186/2191-0855-1-3

Rabausch, U., Juergensen, J., Ilmberger, N., Bohnke, S., Fischer, S., Schubach, B., et al. (2013). Functional screening of metagenome and genome libraries for detection of novel flavonoid-modifying enzymes. *Appl. Environ. Microbiol.* 79, 4551–4563. doi: 10.1128/AEM.01077-13

Rhee, J. K., Ahn, D. G., Kim, Y. G., and Oh, J. W. (2005). New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. *Appl. Environ. Microbiol.* 71, 817–825. doi: 10.1128/AEM.71.2.817-825.2005

Riaz, K., Elmerich, C., Moreira, D., Raffoux, A., Dessaux, Y., and Faure, D. (2008). A metagenomic analysis of soil bacteria extends the diversity of quorum-quenching lactonases. *Environ. Microbiol.* 10, 560–570. doi: 10.1111/j.1462-2920.2007.01475.x

Richardson, T. H., Tan, X. Q., Frey, G., Callen, W., Cabell, M., Lam, D., et al. (2002). A novel, high performance enzyme for starch liquefaction - Discovery and optimization of a low pH, thermostable alpha-amylase. *J. Biol. Chem.* 277, 26501–26507. doi: 10.1074/jbc.M203183200

Robertson, D. E., Chaplin, J. A., Desantis, G., Podar, M., Madden, M., Chi, E., et al. (2004). Exploring nitrilase sequence space for enantioselective catalysis. *Appl. Environ. Microbiol.* 70, 2429–2436. doi: 10.1128/AEM.70.4.2429-2436.2004

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., et al. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66, 2541–2547. doi: 10.1128/AEM.66.6.2541-2547.2000

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Sariozlu, N. Y., and Kivanc, M. (2009). Isolation of gallic acid-producing microorganisms and their use in the production of gallic acid from gall nuts and sumac. *Afr. J. Biotechnol.* 8, 1110–1115.

Scanlon, T. C., Dostal, S. M., and Griswold, K. E. (2014). A high-throughput screen for antibiotic drug discovery. *Biotechnol. Bioeng.* 111, 232–243. doi: 10.1002/bit.25019

Schipper, C., Hornung, C., Bijtenhoorn, P., Quitschau, M., Grond, S., and Streit, W. R. (2009). Metagenome-derived clones encoding two novel lactonase family proteins involved in biofilm inhibition in *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.* 75, 224–233. doi: 10.1128/AEM.01389-08

Schirmer, A., Gadkari, R., Reeves, C. D., Ibrahim, F., Delong, E. F., and Hutchinson, C. R. (2005). Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge Discodermia dissoluta. *Appl. Environ. Microbiol.* 71, 4840–4849. doi: 10.1128/AEM.71.8.4840-4849.2005

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* 109, 6241–6246. doi: 10.1073/pnas.1117018109

Shen, D., Xu, J. H., Wu, H. Y., and Liu, Y. Y. (2002). Significantly improved esterase activity of Trichosporon brassicae cells for ketoprofen resolution by 2-propanol treatment. *J. Mol. Catal. B Enzym.* 18, 219–224. doi: 10.1016/S1381-1177(02)00099-1

Smits, S. L., and Osterhaus, A. D. (2013). Virus discovery: one step beyond. *Curr. Opin. Virol.* 3, e1–e6. doi: 10.1016/j.coviro.2013.03.007

Sommer, M. O. A., Dantas, G., and Church, G. M. (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325, 1128–1131. doi: 10.1126/science.1176950

Staley, C., Johnson, D., Gould, T. J., Wang, P., Phillips, J., Cotner, J. B., et al. (2015). Frequencies of heavy metal resistance are associated with land cover type in the Upper Mississippi River. *Sci. Total Environ.* 511, 461–468. doi: 10.1016/j.scitotenv.2014.12.069

Staley, J. T., and Konopka, A. (1985). Measurement of *insitu* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39, 321–346. doi: 10.1146/annurev.mi.39.100185.001541

Streit, W. R., and Entcheva, P. (2003). Biotin in microbes, the genes involved in its biosynthesis, its biochemical role and perspectives for biotechnological production. *Appl. Microbiol. Biotechnol.* 61, 21–31. doi: 10.1007/s00253-002-1186-2

Streit, W. R., and Schmitz, R. A. (2004). Metagenomics - the key to the uncultured microbes. *Curr. Opin. Microbiol.* 7, 492–498. doi: 10.1016/j.mib.2004.08.002

Su, J. Q., Wei, B., Xu, C. Y., Qiao, M., and Zhu, Y. G. (2014). Functional metagenomic characterization of antibiotic resistance genes in agricultural soils from China. *Environ. Int.* 65, 9–15. doi: 10.1016/j.envint.2013.12.010

Tannieres, M., Beury-Cirou, A., Vigouroux, A., Mondy, S., Pellissier, F., Dessaux, Y., et al. (2013). A metagenomic study highlights phylogenetic proximity of quorum-quenching and xenobiotic-degrading amidases of the AS-family. *PLoS ONE* 8:e65473. doi: 10.1371/journal.pone.0065473

Tao, W., Lee, M. H., Wu, J., Kim, N. H., Kim, J. C., Chung, E., et al. (2012). Inactivation of chloramphenicol and florfenicol by a novel chloramphenicol hydrolase. *Appl. Environ. Microbiol.* 78, 6295–6301. doi: 10.1128/AEM.01154-12

Torsvik, V., Goksoyr, J., and Daae, F. L. (1990). High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol.* 56, 782–787.

Uchiyama, T., Abe, T., Ikemura, T., and Watanabe, K. (2005). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat. Biotechnol.* 23, 88–93. doi: 10.1038/nbt1048

Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* 20, 616–622. doi: 10.1016/j.copbio.2009.09.010

Uchiyama, T., and Miyazaki, K. (2010). Product-induced gene expression, a product-responsive reporter assay used to screen metagenomic libraries for enzyme-encoding genes. *Appl. Environ. Microbiol.* 76, 7029–7035. doi: 10.1128/AEM.00464-10

Uchiyama, T., and Watanabe, K. (2006). Improved inverse PCR scheme for metagenome walking. *Biotechniques* 41, 183–188. doi: 10.2144/000112210

van der Maarel, M. J., van der Veen, B., Uitdehaag, J. C., Leemhuis, H., and Dijkhuizen, L. (2002). Properties and applications of starch-converting enzymes of the alpha-amylase family. *J. Biotechnol.* 94, 137–155. doi: 10.1016/S0168-1656(01)00407-2

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74. doi: 10.1126/science.1093857

Ververidis, F., Trantas, E., Douglas, C., Vollmer, G., Kretzschmar, G., and Panopoulos, N. (2007a). Biotechnology of flavonoids and other phenylpropanoid-derived natural products. Part I: Chemical diversity, impacts on plant biology and human health. *Biotechnol. J.* 2, 1214–1234. doi: 10.1002/biot.200700084

Ververidis, F., Trantas, E., Douglas, C., Vollmer, G., Kretzschmar, G., and Panopoulos, N. (2007b). Biotechnology of flavonoids and other phenylpropanoid-derived natural products. Part II: Reconstruction of multienzyme pathways in plants and microbes. *Biotechnol. J.* 2, 1235–1249. doi: 10.1002/biot.200700184

Vester, J. K., Glaring, M. A., and Stougaard, P. (2014). Discovery of novel enzymes with industrial potential from a cold and alkaline environment by a combination of functional metagenomics and culturing. *Microb. Cell Fact.* 13:72. doi: 10.1186/1475-2859-13-72

Voget, S., Leggewie, C., Uesbeck, A., Raasch, C., Jaeger, K. E., and Streit, W. R. (2003). Prospecting for novel biocatalysts in a soil metagenome. *Appl. Environ. Microbiol.* 69, 6235–6242. doi: 10.1128/AEM.69.10.6235-6242.2003

Von Wettstein, D., Mikhaylenko, G., Froseth, J. A., and Kannangara, C. G. (2000). Improved barley broiler feed with transgenic malt containing heat-stable (1,3-1,4)-beta-glucanase. *Proc. Natl. Acad. Sci. U.S.A.* 97, 13512–13517. doi: 10.1073/pnas.97.25.13512

Walter, J., Mangold, M., and Tannock, G. W. (2005). Construction, analysis, and beta-glucanase screening of a bacterial artificial chromosome library from the large-bowel microbiota of mice. *Appl. Environ. Microbiol.* 71, 2347–2354. doi: 10.1128/AEM.71.5.2347-2354.2005

Wang, K., Li, G., Yu, S. Q., Zhang, C. T., and Liu, Y. H. (2010). A novel metagenome-derived beta-galactosidase: gene cloning, overexpression, purification and characterization. *Appl. Microbiol. Biotechnol.* 88, 155–165. doi: 10.1007/s00253-010-2744-7

Wang, K., Lu, Y., Liang, W. Q., Wang, S. D., Jiang, Y., Huang, R., et al. (2012). Enzymatic synthesis of galacto-oligosaccharides in an organic-aqueous biphasic system by a novel beta-galactosidase from a metagenomic library. *J. Agric. Food Chem.* 60, 3940–3946. doi: 10.1021/jf300890d

Warnecke, F., Luginbuhl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565. doi: 10.1038/nature06269

Wendel, A. M., Johnson, D. H., Sharapov, U., Grant, J., Archer, J. R., Monson, T., et al. (2009). Multistate Outbreak of *Escherichia coli* O157:H7 Infection associated with consumption of packaged spinach, August-September 2006: the wisconsin investigation. *Clin. Infect. Dis.* 48, 1079–1086. doi: 10.1086/597399

World Health Organization. (2014). *Antimicrobial Resistance: Global Report On Surveillance*. Geneva: World Health Organization.

Wu, D., and Bird, M. R. (2010). The interaction of protein and polyphenol species in ready to drink black tea liquor production. *J. Food Process Eng.* 33, 481–505. doi: 10.1111/j.1745-4530.2008.00286.x

Yao, J., Chen, Q., Zhong, G., Cao, W., Yu, A., and Liu, Y. (2014). Immobilization and characterization of tannase from a metagenomic library and its use for removal of tannins from green tea infusion. *J. Microbiol. Biotechnol.* 24, 80–86. doi: 10.4014/jmb.1308.08047

Yao, J., Fan, X. J., Lu, Y., and Liu, Y. H. (2011). Isolation and characterization of a novel tannase from a metagenomic library. *J. Agric. Food Chem.* 59, 3812–3818. doi: 10.1021/jf104394m

Yun, J., and Ryu, S. (2005). Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb. Cell Fact.* 4:8. doi: 10.1186/1475-2859-4-8

Zhu, B., and Panek, J. S. (2001). Methodology based on chiral silanes in the synthesis of polypropionate-derived natural products - Total synthesis of epothilone A. *Eur. J. Org. Chem.* 2001, 1701–1714. doi: 10.1002/1099-0690(200105)2001:9<1701::AID-EJOC1701>3.0.CO;2-#

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 June 2015
doi: 10.3389/fmicb.2015.00548

Glucose-tolerant β-glucosidase retrieved from a Kusaya gravy metagenome

*Taku Uchiyama[1], Katusro Yaoi[1] and Kentaro Miyazaki[1,2]**

[1] *Bioproduction Research Institute, National Institute of Advanced Industrial Science and Technology Tsukuba, Ibaraki, Japan,* [2] *Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan*

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Trevor Carlos Charles, University of Waterloo, Canada
David L. Bernick, University of California, Santa Cruz, USA

***Correspondence:**
Kentaro Miyazaki, Department of Life Science and Biotechnology, Bioproduction Research Institute – National Institute of Advanced Industrial Science and Technology, Tsukuba Central 6, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan
miyazaki-kentaro@aist.go.jp

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *10 April 2015*
Accepted: *19 May 2015*
Published: *16 June 2015*

Citation:
Uchiyama T, Yaoi K and Miyazaki K (2015) Glucose-tolerant β-glucosidase retrieved from a Kusaya gravy metagenome. Front. Microbiol. 6:548. doi: 10.3389/fmicb.2015.00548

β-glucosidases (BGLs) hydrolyze cello-oligosaccharides to glucose and play a crucial role in the enzymatic saccharification of cellulosic biomass. Despite their significance for the production of glucose, most identified BGLs are commonly inhibited by low (∼mM) concentrations of glucose. Therefore, BGLs that are insensitive to glucose inhibition have great biotechnological merit. We applied a metagenomic approach to screen for such rare glucose-tolerant BGLs. A metagenomic library was created in *Escherichia coli* (∼10,000 colonies) and grown on LB agar plates containing 5-bromo-4-chloro-3-indolyl-β-D-glucoside, yielding 828 positive (blue) colonies. These were then arrayed in 96-well plates, grown in LB, and secondarily screened for activity in the presence of 10% (w/v) glucose. Seven glucose-tolerant clones were identified, each of which contained a single *bgl* gene. The genes were classified into two groups, differing by two nucleotides. The deduced amino acid sequences of these genes were identical (452 aa) and found to belong to the glycosyl hydrolase family 1. The recombinant protein (Ks5A7) was overproduced in *E. coli* as a C-terminal 6 × His-tagged protein and purified to apparent homogeneity. The molecular mass of the purified Ks5A7 was determined to be 54 kDa by SDS-PAGE, and 160 kDa by gel filtration analysis. The enzyme was optimally active at 45°C and pH 5.0–6.5 and retained full or 1.5–2-fold enhanced activity in the presence of 0.1–0.5 M glucose. It had a low K_M (78 μM with *p*-nitrophenyl β-D-glucoside; 0.36 mM with cellobiose) and high V_{max} (91 μmol min^{-1} mg^{-1} with *p*-nitrophenyl β-D-glucoside; 155 μmol min^{-1} mg^{-1} with cellobiose) among known glucose-tolerant BGLs and was free from substrate (0.1 M cellobiose) inhibition. The efficient use of Ks5A7 in conjunction with *Trichoderma reesei* cellulases in enzymatic saccharification of alkaline-treated rice straw was demonstrated by increased production of glucose.

Keywords: β-glucosidase, cellulosic biomass, enzymatic saccharification, metagenome, substrate inhibition, product inhibition

Abbreviations: GH1, glycoside hydrolase family 1; pNP *p*-nitrophenol, pNPGlc *p*-nitrophenyl β-D-glucopyranoside; pNPFuc, *p*-nitrophenyl β-D-fucopyranoside; X-glc, 5-bromo-4-chloro-3-indolyl-β-D-glucoside.

Introduction

Cellulose, the most abundant component of biomass on earth, is a linear polymer of D-glucose linked by β-1,4-glucosidic bonds. Because of the increasing demand for energy and the continuous depletion of fossil fuels, the production of bio-energy and bio-based products from cellulosic biomass is one of the biggest challenges in biotechnology. The breakdown of cellulosic biomass to glucose involves physical–chemical treatment followed by enzymatic saccharification of the raw material. The enzymatic process involves the synergistic actions of four classes of enzymes: (i) endo-β-1,4-glucanase (EC 3.2.1.4); (ii) exo-cellobiohydrolase (EC 3.2.1.91); (iii) copper-dependent lytic polysaccharide monooxygenase; and (iv) β-glucosidase (EC 3.2.1.21, BGL). Endo-glucanase and exo-cellobiohydrolase act on cellulose to produce cellobiose, which often inhibits the activities of the enzymes that catalyze its production (Coughlan, 1985; Kadam and Demain, 1989; Watanabe et al., 1992). β-glucosidases (BGLs) act on cellobiose (and cello-oligosaccharides) to produce glucose; this can reduce the inhibitory effect of cellobiose on endo-glucanase and exo-cellobiohydrolase (Xin et al., 1993; Saha et al., 1994). However, most of the microbial BGLs known to date are highly sensitive to glucose (Gueguen et al., 1995; Saha et al., 1995). Furthermore, BGLs are also inhibited by their substrate, cellobiose (Woodward and Wiseman, 1982; Schmid and Wandrey, 1987). Thus, the development of BGLs that are insensitive to glucose and cellobiose inhibition will have a significant impact on the enzymatic saccharification of cellulosic biomass and will accelerate the entire process of cellulose breakdown.

To date, several glucose-tolerant BGLs have been identified in insects (Uchima et al., 2011), fungi (Saha and Bothast, 1996; Yan and Lin, 1997; Riou et al., 1998; Decker et al., 2001; Zanoelo et al., 2004; Souza et al., 2014), bacteria (Pérez-Pons et al., 1995), and metagenomes (Fang et al., 2010; Biver et al., 2014). Recently, we have identified a glucose-tolerant BGL (Td2F2) in a wood compost metagenomic library (Uchiyama et al., 2013). Td2F2 has a unique property in that its activity is not reduced by glucose but is stimulated in the presence of high concentrations of glucose (0.1 M or higher). The basis for this unique property is its high transglycosylation activity. The tolerance to glucose and high transglycosylation activity of Td2F2 will be strongly advantageous when it is used in the enzymatic saccharification of cellulose as well as the enzymatic synthesis of stereo- and regio-specific glycosides.

To identify other potentially useful BGLs, we screened a metagenomic library of Kusaya (a Japanese traditional fermentation product made from fish) gravy as a source for genomes. The library was constructed in *Escherichia coli*, which was first screened for BGL activity in the absence of glucose. Positive clones were then screened in the presence of glucose. As a result of this screen, we successfully obtained a glucose-tolerant BGL, which we named Ks5A7. The gene encoding Ks5A7 was overexpressed in *E. coli*, and the recombinant enzyme was characterized. We applied Ks5A7 to the saccharification of alkaline-treated rice straw, in combination with fungal cellulases from *Trichoderma reesei*, to demonstrate its efficiency for enhancing the production of glucose.

Materials and Methods

Reagents

Restriction endonucleases, DNA ligase, and DNA polymerase were purchased from Takara Bio (Shiga, Japan). The QIAquick Kit was obtained from Qiagen (Hilden, Germany). 5-Bromo-4-chloro-3-indolyl-β-D-glucoside (X-glc) was purchased from Rose Scientific (Edmonton, AB, Canada). p-Nitrophenyl (pNP) α-D-galactopyranoside, pNP α-D-glucopyranoside, and pNP β-D-xylopyranoside were purchased from Nacalai (Kyoto, Japan). pNP α-D-mannopyranoside was purchased from Senn Chemicals (Zürich, Switzerland). The following chemicals were purchased from Sigma (St. Louis, MO, USA): avicel, pNP α-L-arabinofuranoside, pNP α-L-arabinopyranoside, pNP β-L-arabinopyranoside, pNP α-L-fucopyranoside, pNP β-D-fucopyranoside (pNPFuc), pNP β-D-galactopyranoside, pNP β-D-glucopyranoside (pNPGlc), pNP β-D-mannopyranoside, pNP N-acetyl-β-D-glucosaminide, pNP α-L-rhamnopyranoside, pNP α-D-xylopyranoside, and pNP β-D-cellobioside, sophorose, nigerose, maltose, isomaltose, lactose, and salicin. Cello-origosaccharides and laminaribiose were purchased from Seikagaku Kogyo (Tokyo, Japan). Gentiobiose was purchased from Tokyo Chemical Industry (Tokyo, Japan).

Library Construction and Screening for BGLs

Kusaya gravy was sampled at Niijima Island, Tokyo, Japan in May, 2007. The metagenome was purified, fragmented by partial digestion with *Sau*3AI, and ligated into a p18GFP vector at the *Bam*HI site, as described previously (Uchiyama and Watanabe, 2008). *E. coli* DH10B cells were transformed with the ligation mixture and grown at 37°C overnight on LB agar plates containing 100 μg mL^{-1} ampicillin (Amp) to yield ∼380,000 colonies. The colonies were scraped from the plates, mixed well, appropriately diluted, and regrown on LB agar plates containing 100 μg mL^{-1} Amp, 10 μM isopropyl-β-D-thio-galactopyranoside (IPTG), and 20 μg mL^{-1} X-glc. Approximately 10,000 colonies appeared on the plates; colonies that turned blue in color after prolonged incubation at 4°C for 3 weeks were selected and arrayed in a 96-well format.

Screening for Glucose-Tolerant BGLs

Blue *E. coli* colonies arrayed in 96-well plates were grown in 800 μL of LB liquid medium containing 100 μg mL^{-1} Amp, 10 μM IPTG at 37°C overnight with vigorous agitation (1,000 rpm) in a Taitec (Saitama, Japan) MBR-420FL shaker. Cultures were then transferred to three 96-well plates (200 μL each, with the remaining 200 μL reserved for stock), pelleted by centrifugation (3,220 × *g*, 15 min, 4°C), and the supernatant discarded. Cells were resuspended in 0.1 M sodium phosphate buffer, pH 6.0, containing 1 mM pNPGlc and 0 or 10% (w/v) glucose, and incubated at 37°C with agitation (1,000 rpm). After 48 h, cells were pelleted by centrifugation (3,220 × *g*, 15 min,

4°C), and 50 μL of the supernatants were transferred to fresh 96-well plates; 100 μL of 0.1 M Na_2CO_3 was added to each well, and absorbance at 405 nm was read using a Molecular Devices (Sunnyvale, CA, USA) plate reader (VersaMax).

DNA Sequencing and Sequence Data Analysis

A shotgun DNA library was produced using plasmids partially digested with *Alu*I. The products were separated by agarose gel electrophoresis and fragments 1–3 kb in length were gel-purified and cloned into a suicide vector pre-digested with *Sma*I (Miyazaki, 2010). The DNA sequences of the cloned fragments were determined from one end of the vector, flanked by the *Sma*I site, by the Sanger method. A sequence similarity search was performed using BLAST software (Altschul et al., 1997) and the National Center for Biotechnology Information (NCBI) database.

Production and Purification of Recombinant Ks5A7

To remove two *Nde*I sites encoded in the *ks5a7* gene, two rounds of QuikChange-based site-directed mutagenesis (Weiner et al., 1994) were performed using sets of primers xNdeI-1+ and xNdeI-1−, followed by xNdeI-2+ and xNdeI-2− (**Table 1**). After removing the two *Nde*I sites, the *ks5a7* gene was amplified by PCR using forward (Ks5A7Fwd) and reverse (Ks5A7Rev) primers (**Table 1**). The amplicon (1.4 kbp) was gel-purified, digested with *Nde*I and *Xho*I, and cloned into the same sites of the pET29b (+) vector to fuse a 6 × His-tag to the C-terminus of the recombinant protein. The expression plasmid was introduced into *E. coli* Rosetta (DE3) and grown on LB agar plates containing 50 μg mL^{-1} kanamycin and 34 μg mL^{-1} chloramphenicol. A single colony was selected and grown in 1 L of Overnight Express Instant LB Medium (Novagen, Madison, WI, USA) containing 50 μg mL^{-1} kanamycin and 34 μg mL^{-1} chloramphenicol at 30°C with agitation (200 rpm). After 18 h, cells were collected by centrifugation (5,000 × *g*, 10 min, 4°C) and resuspended in 100 mL of BugBuster (Novagen) and Benzonase (Novagen). After gentle agitation at room temperature for 30 min, debris was removed by centrifugation (15,000 × *g*, 20 min, 4°C). The supernatant was then loaded onto a Ni-NTA column (5 mL; Qiagen, Hilden, Germany) pre-equilibrated with 20 mM sodium phosphate buffer (pH 7.4) containing 0.5 M NaCl. After washing the column with 100 mL of 20 mM sodium buffer (pH 7.4) containing 0.5 M NaCl, the column was further washed with 100 mL of 20 mM sodium phosphate buffer (pH 7.4) containing 0.5 M NaCl and 25 mM imidazole. Bound proteins were then eluted with a linear gradient of imidazole from 25 to 500 mM in 20 mM sodium phosphate buffer (pH 7.4) containing 0.5 M NaCl, over a total volume of 100 mL. Active fractions were combined and buffer-exchanged to 20 mM sodium phosphate (pH 7.4) containing 50 mM NaCl using an Amicon Ultra-15. The concentration of Ks5A7 was determined based on the molecular coefficient of 117,035 M^{-1} cm^{-1} at 280 nm. Calculations were performed using the ProtParam tool at http://www.expasy.ch/tools/protparam.html (Gasteiger et al., 2005).

TABLE 1 | Oligonucleotide primers used in this study.

Primer	sequence
xNdeI-1+	5′- ATGATATTGTTCCATAcGTTACTCTTTTTCACTGG -3′
xNdeI-1–	5′- CCAGTGAAAAAGAGTAACgTATGGAACAATATCAT -3′
xNdeI-2+	5′- TTCTTGACTTAAATGATGCATAcGTCTGGTCTGTTTCATT -3′
xNdeI-2–	5′- AATGAAACAGACCAGACgTATGCATCATTTAAGTCAAGAA -3′
Ks5A7Fwd	5′- AAAAACATATGATGAAATTTAATGAAAACTTTGTTTGGGGT -3′
Ks5A7Rev	5′- TTTTTCTCGAGTAGGTTCTCACCATTTTCTTCAATA -3′
E163Q+	5′- AAATACATTATGACATTTAATcAACCTCAGTGCACAATT -3′
E163Q–	5′- AATTGTGCACTGAGGTTgATTAAATGTCATAATGTATTT -3′
E357Q+	5′- ACCTACCTTTTTATATAACTcAAAACGGCCTTGC -3′
E357Q–	5′- GCAAGGCCGTTTTgAGTTATATAAAAAGGTAGGT -3′

Construction, Expression, and Protein Purification of E163Q and E3570Q Variants of Ks5A7

Site-directed mutagenesis was carried out following the QuikChange protocol (Weiner et al., 1994). For E163Q, a set of complementary primers (E163Q+ and E163Q–, **Table 1**) was used. For E357Q, a set of complementary primers (E357Q+ and E3457–, **Table 1**) was used. Gene expression and protein purification were carried out in the same manner as for the wild type.

Molecular Mass

Polyacrylamide gel electrophoresis was performed under denaturing conditions using a DRC XV Pantera gel (7.5–15% [w/v] gradient polyacrylamide) in a Tris-Glycine buffer system containing 0.1% (w/v) sodium dodecylsulfate. Samples were heat-treated at 95°C for 5 min with 2-mercaptoethanol and 0.1% (w/v) sodium dodecylsulfate prior to electrophoresis. Gel filtration analysis was carried out using a GE Healthcare column (Superose 6 10/300 GL, 1 cm × 30 cm) in 20 mM Tris-HCl (pH 7.0) containing 0.2 M NaCl and 10 mM dithiothreitol at a flow rate of 0.5 mL min^{-1}. The molecular weight standards were thyroglobulin (670 kDa), bovine γglobulin (158 kDa), chicken ovalbumin (44 kDa), equine myoglobin (17 kDa), and vitamin B12 (1.35 kDa).

Enzyme Assays

Enzyme activity was routinely assayed in a 85-μL reaction mixture containing McIlvaine buffer (pH 5.5; McIlvaine, 1921), 5 mM pNPGlc, and 1.0 ng μL^{-1} enzyme. After 5 min of incubation at 45°C, the reaction was stopped by incubation at 95°C for 3 min; 85 μL of 0.2 M Na_2CO_3 was added to the mixture, and the levels of liberated *p*-nitrophenol (pNP) were determined at 405 nm using a Molecular Devices plate reader (VersaMax). Optimal reaction temperature and pH were determined by changing the assay temperature or buffers in the presence of 5 mM pNPGlc and 1.0 ng μL^{-1} enzyme. Inhibition of pNPGlc hydrolysis by glucose was tested in a 85-μL reaction mixture containing 5–20 mM pNPGlc, McIlvaine buffer (pH 5.5), 1.0 ng μL^{-1} enzyme, and varied concentrations of glucose (0–0.5 M). Kinetic constants were determined at 45°C from the initial rate of activity. The reaction was performed for 5 min

and stopped by incubation at 95°C for 3 min. For pNPGlc and pNPFuc, the assay was performed in a 85-μL reaction mixture containing McIlvaine buffer (pH 5.5), 0.0156–0.5 mM substrate, and 0.1 ng μL^{-1} enzyme; 85 μL of 0.2 M Na_2CO_3 was added to the mixture, and levels of liberated pNP were determined.

The enzyme activity with respect to oligosaccharide substrates was determined in a 50-μL reaction mixture containing McIlvaine buffer (pH 5.5), 1.0 mg mL^{-1} substrate, and 0.1 ng μL^{-1} enzyme. The reaction was stopped by heating the sample to 98°C for 5 min. The concentration of released glucose was determined using an Invitrogen Amplex Red glucose/glucose oxidase assay kit, according to the manufacturer's instructions. The kinetic constants for cello-oligosaccharides were determined in a 50-μL reaction mixture containing McIlvaine buffer (pH 5.5), 0.0625–4 mM substrate, and 0.1 ng μL^{-1} enzyme. The kinetic constants, K_M and k_{cat}, were calculated by non-linear regression with the Michaelis–Menten equation using GraphPad PRISM Version 6.0 (GraphPad Software).

Saccharification of Alkaline-Treated Rice Straw

Alkaline-treated rice straw was prepared by incubation in 0.5% [w/v] NaOH at 100°C for 5 min as described previously (Kawai et al., 2012), which was purchased from Japan Bioindustry Association. *T. reesei* strain PC-3-7 (ATCC 66589) were purchased from American Type Culture Collection (ATCC). For preparation of crude cellulases from the *T. reesei* strain PC-3-7, the fungus was cultivated on potato dextrose agar and 10^7 conidia were collected and inoculated into 50 mL of basal medium (Kawamori et al., 1986) containing 1 % (w/v) avicel. The inoculum was cultivated for 1 week at 28°C, 220 rpm. After cultivation, the culture was centrifuged at 8,000 × *g* for 20 min at 4°C, and the supernatant was filtered. The resulting filtrate was used as the crude cellulases.

The concentration of the crude cellulases was determined using a Quick Start Bradford Dye Reagent (Bio-Rad Laboratories, Hercles, CA, USA) with bovine γ-globulin as the standard. Saccharification of alkaline-treated rice straw was performed in a hermetically closed 20-mL plastic bottle at 50°C, with shaking at 150 rpm. The reaction medium contained 50 mg mL^{-1} alkaline-treated rice straw, 100 mM sodium acetate buffer pH 5.0, 0.2 mg mL^{-1} sodium azide, and 150 μg mL^{-1} of crude cellulases. BGL (Ks5A7 or Td2F2) was added to a concentration of 5 μg mL^{-1}. After the reaction, the supernatants were boiled for 5 min, and the production of glucose and cellobiose was measured by HPLC following the method described previously (Kawai et al., 2012). Preparation of Td2F2 was as described previously (Uchiyama et al., 2013).

Nucleotide Sequence Accession Numbers

The nucleotide sequence for Ks5A7 has been deposited in GenBank/EMBL/DDBJ under the accession number HV348683.

Results and Discussion

Screening for BGL in a Metagenomic Library of Kusaya Gravy

A metagenomic library was constructed in *E. coli* using Kusaya gravy, a traditional Japanese fermentation food product of dried fish, as a source of the metagenome. The library containing ∼380,000 clones included insert fragments ranging from 5 to 20 kbp in length. A portion of the library (∼10,000 clones) was used to screen for BGL by growing on LB agar plates containing X-glc as a substrate. Although overnight cultivation generated very few positive (i.e., blue) colonies, prolonged incubation at 4°C gradually increased the number of positive colonies, yielding ∼1,000 blue colonies after 3 weeks. The positive colonies were then streaked onto LB agar plates containing X-glc for single isolation, yielding 828 clones in total.

Screening for Glucose-Tolerant BGLs

The clones initially identified as positive were arrayed in a 96-well format. Clones were grown in LB, and whole cells were used to determine activity in the presence and absence of 10% (w/v) glucose. Although the majority of clones exhibited no activity in the presence of 10% (w/v) glucose, seven (5A7, 5B6, 5F2, 6C8, 7F9, 9B4, and 10H11) retained >20% activity relative to the glucose-free condition. DNA sequencing was performed from one end of the plasmids, revealing that three clones (7F9, 9B4, and 10H11) had identical insert fragments.

DNA Sequencing of Glucose-Tolerant BGLs

Plasmids were purified from the five different clones: 5A7, 5B6, 5F2, 6C8, and 7F9. For each plasmid, a total of 96 shotgun clones were analyzed. Although no complete *bgl* gene was obtained from the partially determined nucleotide sequences, the results suggested that the clones carried *bgl* genes with high identity. We then synthesized a set of PCR primers to amplify the *bgl* gene from the five plasmids. All five clones produced a 1.4-kbp amplicon. DNA sequencing of the fragments revealed that the five *bgl* genes could be classified into two groups, differing by only two nucleotide substitutions. The deduced amino acid sequences were identical, and the gene obtained from clone 5A7 was used for subsequent studies.

The *bgl* gene *ks5a7* contained 1,359 bp, with a GC content of 32.3%. The predicted ATG initiation codon was preceded by a possible ribosomal binding site, 5′-AAGAGGA-3′. The deduced amino acid sequence contained 452 amino acids and had a calculated molecular mass of 52,509 Da.

Using BLAST-P[1], we found that Ks5A7 was highly similar to enzymes belonging to the glycoside hydrolase family 1 (GH1) of the carbohydrate-active enzyme classification database (Lombard et al., 2014)[2]. Ks5A7 exhibited the highest identity (57%) with a putative BGL from *Clostridiales bacterium* oral taxon 876 and a 55% identity with a putative BGL from *Clostridium hathewayi* DSM13479. When compared with functionally characterized

[1]http://www.ncbi.nlm.nih.gov/BLAST/
[2]http://www.cazy.org/

BGLs, the Ks5A7 showed the highest (46%) identity with that of *Thermotoga neapolitana* (Yernool et al., 2000; Park et al., 2005).

Overproduction of Ks5A7

Ks5A7 was produced as a C-terminal 6 × His-tagged protein using a pET system (Studier and Moffatt, 1986). Two *E. coli* strains, Rosetta (DE3), and BL21 (DE3), were tested as a host. Approximately 2.5-fold higher activity was obtained from the cell extract prepared from Rosetta (DE3) compared with that from BL21 (DE3). Ks5A7 contained a high rate of rare codons (52 of a total 452 amino acids). Of particular note, all 17 Arg residues were encoded by rare codons: 14 AGA, 2 CGA, and 1 AGG. Because Rosetta (DE3) carries a plasmid containing seven genes for rare tRNA codons, including those for AGA, and AGG, the low production level in BL21 (DE3) might have been improved in Rosetta (DE3) as a result of the supply of rare tRNAs. In terms of temperature, in Rosetta (DE3), the activity was 10-fold higher at 30°C than at 37°C.

Expressed recombinant protein was readily purified to homogeneity using a Ni-NTA column. A large quantity of purified enzyme was recovered, with a typical final yield of 70 mg L^{-1} culture, representing a 30% yield.

General Properties of Ks5A7

Purified recombinant Ks5A7 had a molecular mass of ∼50 kDa according to SDS-PAGE (**Figure 1A**), which is in agreement with the mass calculated from the deduced amino acid sequence (53,573 Da). The molecular mass of the native structure of Ks5A7 was determined by gel filtration column chromatography (**Figure 1B**). Ks5A7 was eluted at the 160 kDa position, suggestive of multimeric states (trimer or tetramer).

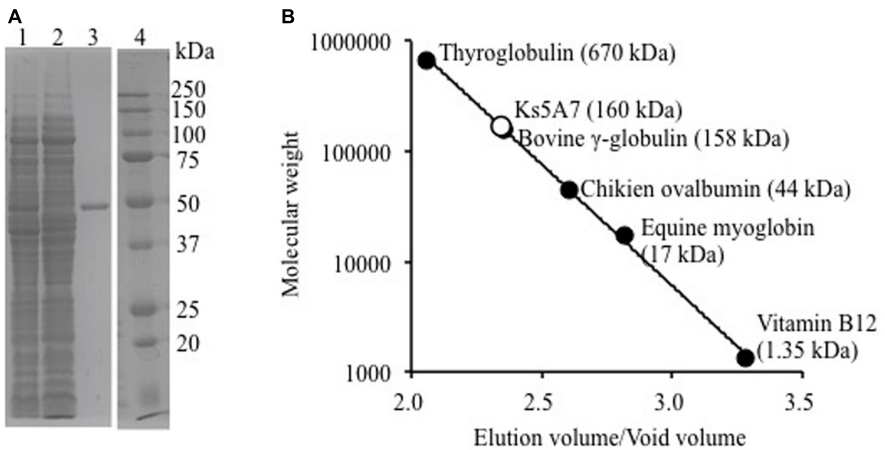

FIGURE 1 | Molecular mass analysis of recombinant Ks5A7. (A) SDS-PAGE. Lane 1, soluble protein fraction; lane 2, flow-through from Ni-NTA column; lane 3, purified Ks5A7; lane 4, molecular markers. Ks5A7 migrated at ∼50 kDa. **(B)** Gel filtration. Symbols: solid circles, molecular mass of protein markers; open circle, Ks5A7. Ks5A7 was eluted at ∼160 kDa.

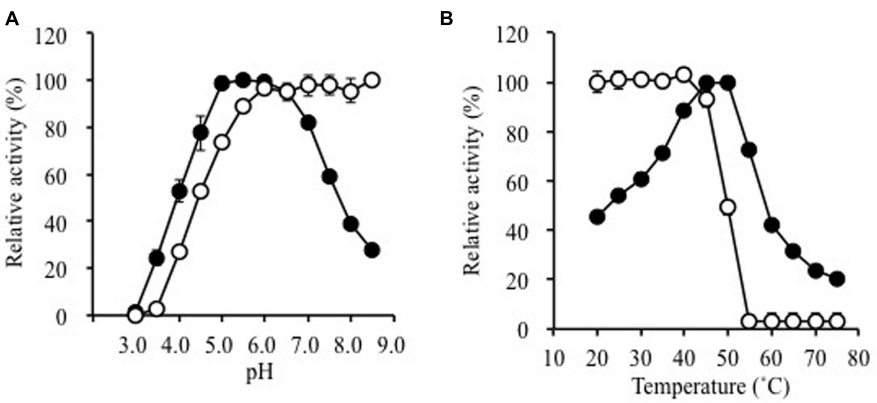

FIGURE 2 | Effects of (A) pH and (B) temperature on the activity (solid circles) and stability (open circles) of purified Ks5A7. With regard to the pH-dependence of stability, the enzyme was incubated for 30 min at various pH values. In terms of the pH-dependence of activity, the enzyme was assayed at various pH values by the standard assay method. To address the temperature-dependence of stability, the enzyme was incubated for 10 min at various temperatures. In terms of the temperature-dependence of activity, the enzyme was assayed at various temperatures by the standard assay method. Error bars, SD. $N = 3$.

The pH-stability and pH-dependence of activity are illustrated in **Figure 2A**. The enzyme was fairly stable at pH 5.5–8.5 (30 min at 25°C). It was optimally active between pH 5.0 and 6.0 (specific activity, 49.1 ± 0.4 μmol min^{-1} mg^{-1}) with ∼80% activity at pH 4.5 and 7.0, respectively. The effects of temperature on stability and activity are shown in **Figure 2B**. The enzyme was inactivated upon incubation at 55°C for 10 min. Maximal activity was observed at 50°C in a 5-min assay (specific activity, 58.4 ± 1.4 μmol min^{-1} mg^{-1}).

On the basis of similarity to the known GH1 family BGLs, it has been inferred that E163 and E357 function as an acid-base catalyst and nucleophile, respectively (Withers et al., 1990; Wang et al., 1995). They were individually substituted to glutamine, and the resultant mutant enzymes were characterized. No activity was observed when 5 mM pNPGlc was used for both enzymes (data not shown), suggesting the same roles for these residues in catalysis as observed in other GH1 BGLs.

Activity with *p*-Nitrophenyl Substrates and Oligosaccharides

The substrate specificity of Ks5A7 was characterized using a fixed concentration (5 mM) of various *p*-nitrophenyl substrates and oligosaccharides. For *p*-nitrophenyl substrates, the enzyme showed the highest activity for pNPFuc, followed by pNPGlc (**Table 2**). Dual pNPFuc and pNPGlc activities have been reported for a BGL enzyme from *Bifidobacterium breve* (Nunoura et al., 1996a,b). However, the activity of *Bifidobacterium* BGL lost 30% of its original activity in the presence of 0.1 M glucose, whereas Ks5A7 displayed 150% activity under the same conditions (see below, **Figure 3A**). Both enzymes belong to the GH1 family but share only 37% of their amino acid sequence identity. Therefore, these two enzymes are distinct, and the basis for the dual pNPFuc/pNPGlu activities remains unknown.

As shown in **Table 2**, Ks5A7 was found to possess enzyme activity for cello-oligosaccharides from cellobiose to cellopentaose. Ks5A7 hydrolyzed a range of β-linked glycosides including β(1,2), β(1,3), and β(1,4) but not β(1,6). No activity was detected for the oligosaccharides with α-linkages.

Kinetic Constants of Ks5A7

The steady-state kinetic constants of Ks5A7 for pNPGlc, pNPFuc, and cello-oligosaccharides are shown in **Table 3**. The K_M for

TABLE 2 | Substrate specificity of the recombinant Ks5A7.

Substrate[a]	Linkage of glycosyl group	Relative activity (%)
Chromogenic substrates		
pNPGlc	β-Glucose	100 ± 9.7[b]
pNPFuc	β-Fucose	164 ± 12
p-Nitrophenyl-β-D-galactopyranoside	β-Galactose	64.5 ± 5.5
p-Nitrophenyl-β-D-xylopyranoside	β-Xylose	1.05 ± 0.0
p-Nitrophenyl-β-D-cellobioside	β-Cellobiose	7.88 ± 0.0
p-Nitrophenyl-β-D-lactopyranoside	β-Lactose	46.8 ± 5.5
p-Nitrophenyl-α-D-glucopyranoside	α-Glucose	2.54 ± 0.0
p-Nitrophenyl-α-L-arabinopyranoside	α-Alabinose	3.31 ± 0.0
Oligosaccharides		
Cellobiose	β(1,4)Glucose	100 ± 8.4[c]
Cellotriose	β(1,4)Glucose	116 ± 7.0
Cellotetraose	β(1,4)Glucose	121 ± 21
Cellopentaose	β(1,4)Glucose	90.1 ± 8.8
Laminaribiose	β(1,3)Glucose	112 ± 0.8
Sophorose	β(1,2)Glucose	65.5 ± 9.7
Salicin	β(1,4)Glucose	59.1 ± 3.8

[a]*No activity was detected with p-nitrophenyl-β-D-mannopyranoside, p-nitrophenyl- N-acetyl-β-D-glucosaminide, p-nitrophenyl-β-L-arabinopyranoside, p-nitrophenyl-α-D-galactopyranoside, p-nitrophenyl-α-D-xylopyranoside, p-nitrophenyl-α-L-fucopyranoside, p-nitrophenyl-α-L-arabinofuranoside, and p-nitrophenyl-α-L-rhamnopyranoside, oligosaccharides, such as gentiobiose, nigerose, maltose, isomaltose, and lactose.*

[b]*The specific activity of Ks5A7 for pNPGlc was 53.9 ± 5.2 μmol min^{-1} mg^{-1}, by measuring the release of pNP.*

[c]*The specific activity of Ks5A7 for cellobiose was 170 ± 20 μmol min^{-1} mg^{-1}, by measuring the release of glucose.*

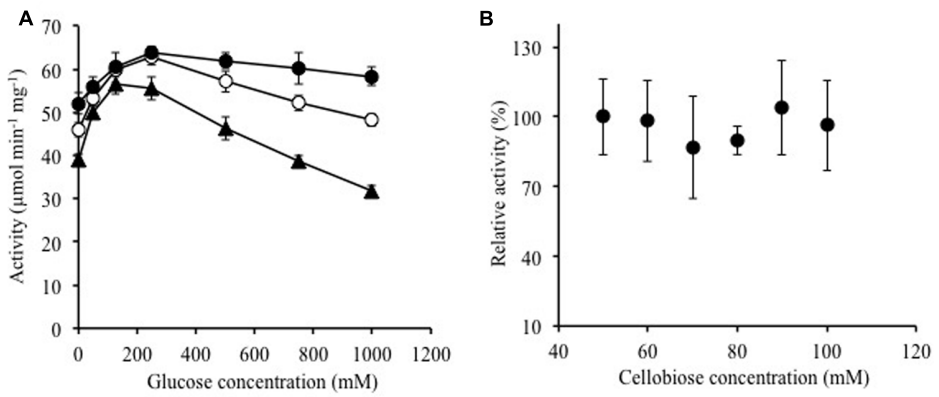

FIGURE 3 | Effects of (A) glucose and (B) cellobiose on enzymatic activity of Ks5A7. (A) Activities were determined in the presence of various concentrations of glucose using 5 (solid triangles), 10 (open circles), or 20 mM of pNPGlc (solid circles) at 45°C. **(B)** Activities were determined in the presence of various concentrations of cellobiose at 45°C. The specific activity for cellobiose was measured by evaluating the glucose concentration. The activity for 50 mM cellobiose was taken to be 100%. The specific activity was 131.9 ± 21 μmol min^{-1} mg^{-1}. Error bars, SD. $N = 3$.

pNPFuc was higher (0.152 mM) than that for pNPGlc, but the V_{max} (137 μmol min^{-1} mg^{-1}) was also higher for pNPFuc, resulting in similar overall catalytic efficiency (k_{cat}/K_M) for the two substrates. Compared with other known glucose-tolerant BGLs (Pérez-Pons et al., 1995; Saha and Bothast, 1996; Yan and Lin, 1997; Riou et al., 1998; Decker et al., 2001; Zanoelo et al., 2004; Fang et al., 2010; Uchima et al., 2011; Uchiyama et al., 2013; Biver et al., 2014; Souza et al., 2014), the K_M of Ks5A7 for pNPGlc was the lowest (0.078 mM) and the V_{max} was relatively high (90.8 μmol min^{-1} mg^{-1}).

TABLE 3 | Steady-state kinetic constants of the recombinant Ks5A7.

Substrate	K_M (mM)	V_{max} (μmol min^{-1} mg^{-1})	k_{cat} (s^{-1})	k_{cat}/K_M (s^{-1} mM^{-1})
Aryl-β-glycosides				
pNPGlc	0.078 ± 0.002	90.8 ± 0.7	81.0 ± 0.6	1045
pNPFuc	0.152 ± 0.004	137 ± 0.2	122 ± 1.2	803
Oligosaccharides				
Cellobiose	0.358 ± 0.055	155 ± 8.3	138 ± 7.4	386
Cellotriose	0.163 ± 0.016	111 ± 3.2	99.1 ± 2.8	610
Cellotetraose	0.160 ± 0.012	114 ± 2.5	101 ± 2.2	634
Cellopentaose	0.132 ± 0.011	112 ± 2.5	99.7 ± 2.2	753

TABLE 4 | Effects of organic solvents, metal ions, and chelating agent on the enzyme activities of the recombinant Td2F2.

Reagent	Concentration	Relative activity (%)
No additive		100 ± 0.9[a]
Ethanol	10% v/v	102 ± 2.5
Ethanol	25% v/v	44.2 ± 2.2
Dimethylsulfoxide	10% v/v	85.9 ± 0.9
Dimethylsulfoxide	25% v/v	43.8 ± 1.2
$AlCl_3$	1 mM	73.6 ± 2.1
$CaCl_2$	1 mM	84.0 ± 0.5
$CoCl_2$	1 mM	85.7 ± 1.2
$CuCl_2$	1 mM	29.6 ± 0.1
$FeCl_3$	1 mM	101 ± 1.4
$MgCl_2$	1 mM	75.2 ± 1.9
$MnCl_2$	1 mM	81.8 ± 2.7
$NiCl_2$	1 mM	88.6 ± 1.6
$ZnCl_2$	1 mM	43.9 ± 0.2
EDTA	10 mM	102 ± 4.0
DTT	10 mM	91.8 ± 1.1

[a]*The activity without an additional regent was taken to be 100% (specific activity 50.7 ± 0.5 μmol min^{-1} mg^{-1}).*

For cello-oligosaccharides, the K_M value was highest with cellobiose as a substrate, and it gradually decreased as the chain length increased, suggesting that the active site include subsites that accommodate the oligosaccharides. The absence of glucose inhibition is presumably because the small glucose molecule cannot efficiently bind to the active site. The V_{max} value was slightly higher with cellobiose than with other cello-oligosaccharides. The overall reaction efficiency was highest with cellopentaose. The time-course analysis of cellopentaose hydrolysis by HPLC revealed that the only products were cellotetraose and glucose, indicating that glucose was liberated from cellopentaose, and confirming the exo-type of activity of Ks5A7 (data not shown). Compared with known glucose-tolerant BGLs (Pérez-Pons et al., 1995; Saha and Bothast, 1996; Riou et al., 1998; Zanoelo et al., 2004; Fang et al., 2010; Uchiyama et al., 2013; Biver et al., 2014; Souza et al., 2014), Ks5A7 had the lowest K_M (0.36 mM) for cellobiose and a relatively high V_{max} (155 μmol min^{-1} mg^{-1}).

Effect of Solvents, Metal Ions, and Chelating and Reducing Agents

The effects of various regents and metal cations were examined (**Table 4**); 10% (v/v) ethanol did not affect the activity, whereas 25% (v/v) ethanol reduced the activity to 44%. The addition of 10% or 25% (v/v) DMSO reduced enzyme activity. Among the metal ions tested (1 mM fixed concentration), significant inactivation was observed with $CuCl_2$ and $ZnCl_2$, whereas more than 70% of the activity remained in the presence of $AlCl_3$, $CaCl_2$, $CoCl_2$, $FeCl_3$, $MgCl_2$, $MnCl_2$, and $NiCl_2$. The chelating agent EDTA (10 mM) did not affect enzyme activity, suggesting that divalent cations are not involved in catalysis. The reducing agent dithiothreitol (10 mM) slightly reduced activity (to 92%), indicating that the seven cysteines in each protein (per subunit) might be involved in catalysis or structural formation.

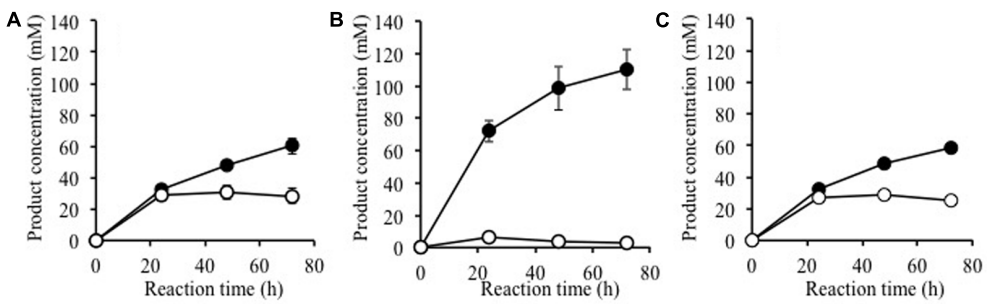

FIGURE 4 | Saccharification of alkaline-treated rice straw by the crude cellulases from *Trichoderma reesei* with β-glucosidase. Production of glucose (solid circles) and cellobiose (open circles) are shown. The reaction was performed with **(A)** no BGL addition, **(B)** plus Ks5A7, and **(C)** plus Td2F2. Error bars, SD. $N = 3$.

Effect of Glucose and Cellobiose on Ks5A7 Activity

Ks5A7 was initially identified as a glucose-tolerant enzyme, but the screening process involved whole cells rather than extracted enzymes. Therefore, we verified that the purified enzyme also showed tolerance to glucose. As shown in **Figure 3A**, no loss of activity was observed in the tested range, 0–0.75 M, at a substrate concentration of 5 mM pNPGlc. At 1.0 M, the activity was reduced to ∼80%. To date, several BGLs that enhance activities in the presence of glucose have been identified (Pérez-Pons et al., 1995; Zanoelo et al., 2004; Fang et al., 2010; Uchima et al., 2011; Uchiyama et al., 2013; Biver et al., 2014; Souza et al., 2014). Similar to that of these enzymes, the activity of Ks5A7 was also enhanced by glucose. In the presence of 250 mM glucose (and 5 mM pNPGlc), the activity was enhanced 1.4-fold, compared with activity in the absence of glucose (**Figure 3A**). At higher concentrations of glucose, however, activity was reduced. This pattern is consistent with the sensitivity to glucose of several other glucose-activated BGLs (Pérez-Pons et al., 1995; Fang et al., 2010; Uchima et al., 2011; Souza et al., 2014).

We recently obtained another glucose-activated BGL, Td2F2, from a wood compost metagenomic library (Uchiyama et al., 2013). In the case of Td2F2, the basis for the enhanced activity in the presence of glucose is due to the strong glycosyltransferase activity (Uchiyama et al., 2013). Taking this into account, we analyzed the reaction products of Ks5A7 after incubation with 5 mM pNPGlc and 250 mM glucose. Glucose was identified as the sole product, suggesting a lack of transglycosylation activity in Ks5A7.

Using cellobiose as a substrate, we investigated substrate inhibition of the enzyme in the tested range, from 50 to 100 mM (**Figure 3B**); no substrate inhibition occurred, at least up to 100 mM cellobiose.

Product inhibition by glucose (Gueguen et al., 1995; Saha et al., 1995) and substrate inhibition by cellobiose (Woodward and Wiseman, 1982; Schmid and Wandrey, 1987) are common major problems for BGLs. Ks5A7 is resistant not only to glucose but also to cellobiose. These unique properties are ideal for cellulosic biomass degradation.

Effect of BGLs on the Enzymatic Saccharification of Alkaline-Treated Rice Straw Hydrolysis

Using alkaline-treated rice straw as a substrate, we investigated whether Ks5A7 (or Td2F2) would be effective for the enzymatic degradation of cellulosic materials. Cellulases from *T. reesei* PC3-7 were used as base enzymes in the reaction, to which a BGL (Ks5A7 or Td2F2) was added (**Figure 4**). Compared with the control (no BGL addition, **Figure 4A**, filled circle), a two fold increase of glucose was observed for Ks5A7 (**Figure 4B**, filled circle), which was much more effective than Td2F2 (**Figure 4C**, filled circle). In addition, virtually no accumulation was observed for cellobiose (**Figure 4B**, open circle). This is probably because Ks5A7 has a higher catalytic efficiency in response to cellobiose than did Td2F2: Ks5A7, K_M, 0.358 mM, and k_{cat}, 155 s^{-1}; Td2F2, K_M; 4.44 mM, kcat; 7.13 s^{-1} (Uchiyama et al., 2013). Td2F2 is the GH1 BGL, which was obtained from the wood compost metagenome and is insensitive to glucose.

Acknowledgments

The authors thank Akiko Rokutani, Shiori Mizuta, Tetsushi Kawai (Japan Bioindustry Association), Noriko Ida (Japan Bioindustry Association), and Yoshinori Kobayashi (Japan Bioindustry Association) for technical assistance. Alkaline treated rice straw was provided by Yoshinori Kobayashi (Japan Bioindustry Association). This work was supported in part by The New Energy and Industrial Technology Development Organization (NEDO) and the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) 26292048 (to KM).

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Biver, S., Stroobants, A., Portetelle, D., and Vandenbol, M. (2014). Two promising alkaline β-glucosidases isolated by functional metagenomics from agricultural soil, including one showing high tolerance towards harsh detergents, oxidants and glucose. *J. Ind. Microbiol. Biotechnol.* 41, 479–488. doi: 10.1007/s10295-014-1400-0

Coughlan, M. P. (1985). The properties of fungal and bacterial cellulases with comment on their production and application. *Biotechnol. Genet. Eng. Rev.* 3, 39–109. doi: 10.1080/02648725.1985.10647809

Decker, C. H., Visser, J., and Schreier, P. (2001). Beta-glucosidase multiplicity from *Aspergillus tubingensis* CBS 643.92: purification and characterization of four beta-glucosidases and their differentiation with respect to substrate specificity, glucose inhibition and acid tolerance. *Appl. Microbiol. Biotechnol.* 55, 157–163. doi: 10.1007/s002530000462

Fang, Z., Fang, W., Liu, J., Hong, Y., Peng, H., Zhang, X., et al. (2010). Cloning and characterization of a beta-glucosidase from marine microbial metagenome with excellent glucose tolerance. *J. Microbiol. Biotechnol.* 20, 1351–1358. doi: 10.4014/jmb.1003.03011

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). "Protein identification and analysis tools on the ExPASy server," in *The Proteomics Protocols Handbook*, ed. J. M. Walker (New York, NY: Humana Press), 571–607.

Gueguen, Y., Chemardin, P., Arnaud, A., and Galzy, P. (1995). Purification and characterization of an intracellular β-glucosidases from *Botrytis cinerea*. *Enzyme Microb. Technol.* 78, 900–906. doi: 10.1016/0141-0229(94)00143-F

Kadam, S. K., and Demain, A. L. (1989). Addition of cloned β-glucosidase enhances the degradation of crystalline cellulose by *Clostridium thermocellum* cellulase complex. *Biochem. Biophys. Res. Commun.* 161, 706–711. doi: 10.1016/0006-291X(89)92657-0

Kawai, T., Nakazawa, H., Ida, N., Okada, H., Tani, S., Sumitani, J., et al. (2012). Analysis of the saccharification capability of high-functional cellulase JN11 for various pretreated biomasses through a comparison with commercially available counterparts. *J. Ind. Microbiol. Biotechnol.* 39, 1741–1749. doi: 10.1007/s10295-012-1195-9

Kawamori, M., Morikawa, Y., and Takasawa, S. (1986). Induction and production of cellulases by L-sorbose in *Trichoderma reesei*. *Appl. Microbiol. Biotechnol.* 24, 449–453. doi: 10.1007/bf00250321

Lombard, V., Golaconda-Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

McIlvaine, T. C. (1921). A buffer solution for colorimetric comparison. *J. Biol. Chem.* 49, 183–186.

Miyazaki, K. (2010). Lethal ccdB gene-based zero-background vector for construction of shotgun libraries. *J. Biosci. Bioeng.* 110, 372–373. doi: 10.1016/j.jbiosc.2010.02.016

Nunoura, N., Ohdan, K., Tanaka, K., Tamaki, H., Yano, T., Inui, M., et al. (1996a). Cloning and nucleotide sequence of the β-D-glucosidase gene from *Bifidobacterium breve* clb, and expression of β-D-glucosidase activity in *Escherichia coli*. *Biosci. Biotechnol. Biochem.* 60, 2011–2018. doi: 10.1271/bbb.60.2011

Nunoura, N., Ohdan, K., Yano, T., Yamamoto, K., and Kumagai, H. (1996b). Purification and characterization of β-D-glucosidase (β-D-fucosidase) from *Bifidobacterium breve* clb acclimated to cellobiose. *Biosci. Biotechnol. Biochem.* 60, 188–193. doi: 10.1271/bbb.60.188

Park, T. H., Choi, K. W., Park, C. S., Lee, S. B., Kang, H. Y., Shon, K. J., et al. (2005). Substrate specificity and transglycosylation catalyzed by a thermostable β-glucosidase from marine hyperthermophile *Thermotoga neapolitana*. *Appl. Microbiol. Biotechnol.* 69, 411–422. doi: 10.1007/s00253-005-0055-1

Pérez-Pons, J. A., Rebordosa, X., and Querol, E. (1995). Properties of a novel glucose-enhanced beta-glucosidae purified from *Streptomyces* sp. *(ATCC*11238). *Biochim. Biophys. Acta* 1251, 145–153. doi: 10.1016/0167-4838(95)00074-5

Riou, C., Salmon, J. M., Vallier, M. J., Günata, Z., and Barre, P. (1998). Purification, characterization, and substrate specificity of a novel highly glucose-tolerant β-glucosidase from *Aspergillus oryzae*. *Appl. Environ. Microbiol.* 64, 3607–3614.

Saha, B. C., and Bothast, R. J. (1996). Production, purification, and characterization of a highly glucose-tolerant novel *β*-glucosidase from *Candida peltata*. *Appl. Environ. Microbiol.* 62, 3165–3170.

Saha, B. C., Freer, S. N., and Bothast, R. J. (1994). Production, purification, and properties of a thermostable β-glucosidase from a color variant strain of *Aureobasidium pullulans*. *Appl. Environ. Microbiol.* 60, 3774–3780.

Saha, B. C., Freer, S. N., and Bothast, R. J. (1995). "Thermostable β-glucosidases," in *Enzymatic Degradation of Insoluble Carbohydrates*, eds J. N. Saddler and M. H. Penner (Washington, DC: American Chemical Society), 197–207.

Schmid, G., and Wandrey, C. (1987). Purification and partial characterization of a cellodextrin glucohydrolase (β-glucosidase) from *Trichoderma reesei* strain QM9414. *Biotechnol. Bioeng.* 30, 571–585. doi: 10.1002/bit.260300415

Souza, F. H. M., Meleiro, L. P., Machado, C. B., Maldonado, R. F., Souza, T. A. C. B., Masui, D. C., et al. (2014). Gene cloning, expression and biochemical characterization of a glucose- and xylose-stimulated β-glucosidase from *Humicola insolens* RP86. *J. Mol. Catal. B Enzym.* 106, 1–10. doi: 10.1016/j.molcatb.2014.04.007

Studier, F. W., and Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* 189, 113–130. doi: 10.1016/0022-2836(86)90385-2

Uchima, C. A., Tokuda, G., Watanabe, H., Kitamoto, K., and Arioka, M. (2011). Heterologous expression and characterization of a glucose-stimulated β-glucosidase from the termite Neotermes koshunensis in *Aspergillus oryzae*. *Appl. Microbiol. Biotechnol.* 89, 1761–1771. doi: 10.1007/s00253-010-2963-y

Uchiyama, T., Miyazaki, K., and Yaoi, K. (2013). Characterization of a novel β-glucosidase from microbial metagenome with strong transglycosylation activity. *J. Biol. Chem.* 288, 18325–18334. doi 10.1074/jbc.M113.471342

Uchiyama, T., and Watanabe, K. (2008). Substrate-induced gene expression (SIGEX) screening of metagenome libraries. *Nat. Protoc.* 3, 1202–12012. doi: 10.1038/nprot.2008.96

Wang, Q., Trimbur, D., Graham, R., Warren, R. A., and Withers, S. G. (1995). Identification of the acid/base catalyst in *Agrobacterium faecalis* beta-glucosidase by kinetic analysis of mutants. *Biochemistry* 34, 14554–14562. doi: 10.1021/bi00044a034

Watanabe, T., Sato, T., Yoshioka, S., Kushijima, T., and Kuwahara, M. (1992). Purification and properties of *Aspergillus niger* β-glucosidase. *J. Biochem.* 209, 651–659.

Weiner, M. P., Costa, G. L., Schoettlin, W., Cline, J., Mathur, E., and Bauer, J. C. (1994). Site-directed mutagenesis of double-stranded DNA by the polymerase chain reaction. *Gene* 151, 119–123. doi: 10.1016/0378-1119(94)90641-6

Withers, S. G., Warren, R. A. J., Street, I. P., Rupitz, K., Kempton, J. B., and Aebersold, R. (1990). Unequivocal demonstration of the involvement of a glutamate residue as a nucleophile in the mechanism of a retaining glycosidase. *J. Am. Chem. Soc.* 112, 5887–5889. doi: 10.1021/ja00171a043

Woodward, J., and Wiseman, A. (1982). Fungal and other β-glucosidases-their properties and applications. *Enzyme Microb. Technol.* 4, 73–79. doi: 10.1016/0141-0229(82)90084-9

Xin, Z., Yinbo, Q., and Peiji, G. (1993). Acceleration of ethanol production from paper mill waste fiber by supplementation with β-glucosidase. *Enzyme Microb. Technol.* 15, 62–65. doi: 10.1016/0141-0229(93)90117-K

Yan, T.-R., and Lin, C.-L. (1997). Purification and characterization of a glucose-toleratnt β-glucosidase from *Aspergillus niger* CCRC 31494. *Biosci. Biotechnol. Biochem.* 61, 965–970. doi: 10.1271/bbb.61.965

Yernool, D. A., McCarthy, J. K., Eveleigh, D. E., and Bok, J. D. (2000). Cloning and characterization of the glucooligosaccharide catabolic pathway β-glucan glucohydrolase and cellobiose phosphorylase in the marine hyperthermophile *Thermotoga neapolitana*. *J. Bacteriol.* 182, 5172–5179. doi: 10.1128/JB.182.18.5172-5179.2000

Zanoelo, F. F., Polizeli-Mde, L., Terenzi, H. F., and Jorge, J. A. (2004). Beta-glucosidae activity from thermophilic fungus *Scytalidium thermophilum* is stimulated by glucose and xylose. *FEMS Microbiol. Lett.* 240, 137–143. doi: 10.1016/j.femsle.2004.09.021

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 October 2015
doi: 10.3389/fmicb.2015.01121

Salt resistance genes revealed by functional metagenomics from brines and moderate-salinity rhizosphere within a hypersaline environment

*Salvador Mirete[1], Merit R. Mora-Ruiz[2], María Lamprecht-Grandío[1], Carolina G. de Figueras[1], Ramon Rosselló-Móra[2] and José E. González-Pastor[1]**

[1] *Laboratory of Molecular Adaptation, Department of Molecular Evolution, Centro de Astrobiología, Consejo Superior de Investigaciones Científicas – Instituto Nacional de Técnica Aeroespacial, Madrid, Spain,* [2] *Marine Microbiology Group, Department of Ecology and Marine Resources, Mediterranean Institute for Advanced Studies, Consejo Superior de Investigaciones Científicas – Universidad de las Islas Baleares, Esporles, Spain*

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
William C. Nelson, University of Southern California, USA
Trevor Carlos Charles, University of Waterloo, Canada
Roy D. Sleator, Cork Institute of Technology, Ireland

***Correspondence:**
José E. González-Pastor
gonzalezpje@cab.inta-csic.es

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *05 June 2015*
Accepted: *28 September 2015*
Published: *13 October 2015*

Citation:
Mirete S, Mora-Ruiz MR, Lamprecht-Grandío M, de Figueras CG, Rosselló-Móra R and González-Pastor JE (2015) Salt resistance genes revealed by functional metagenomics from brines and moderate-salinity rhizosphere within a hypersaline environment. Front. Microbiol. 6:1121. doi: 10.3389/fmicb.2015.01121

Hypersaline environments are considered one of the most extreme habitats on earth and microorganisms have developed diverse molecular mechanisms of adaptation to withstand these conditions. The present study was aimed at identifying novel genes from the microbial communities of a moderate-salinity rhizosphere and brine from the Es Trenc saltern (Mallorca, Spain), which could confer increased salt resistance to *Escherichia coli*. The microbial diversity assessed by pyrosequencing of 16S rRNA gene libraries revealed the presence of communities that are typical in such environments and the remarkable presence of three bacterial groups never revealed as major components of salt brines. Metagenomic libraries from brine and rhizosphere samples, were transferred to the osmosensitive strain *E. coli* MKH13, and screened for salt resistance. Eleven genes that conferred salt resistance were identified, some encoding for well-known proteins previously related to osmoadaptation such as a glycerol transporter and a proton pump, whereas others encoded proteins not previously related to this function in microorganisms such as DNA/RNA helicases, an endonuclease III (Nth) and hypothetical proteins of unknown function. Furthermore, four of the retrieved genes were cloned and expressed in *Bacillus subtilis* and they also conferred salt resistance to this bacterium, broadening the spectrum of bacterial species in which these genes can function. This is the first report of salt resistance genes recovered from metagenomes of a hypersaline environment.

Keywords: functional metagenomics, salt resistance genes, stress response, hypersaline, rhizosphere, brine, saltern, DNA repair

INTRODUCTION

Life under extreme osmotic pressure in the environment represents a challenge for the vast majority of the microorganisms. Hypersaline habitats such as lakes, salt ponds, and sediments associated with marine ecosystems are considered extreme environments constituted by a discontinuous salinity gradient where salt can reach saturation by evaporation processes (Oren, 2002). These salt-enriched habitats constitute appropriate systems to address questions related to

the molecular mechanisms of adaptation to elevated concentrations of NaCl since the native microbial consortia that inhabit these hypersaline environments can grow in the presence of more than 30% (w/v) total salts (Rodriguez-Valera et al., 1985; Antón et al., 2000). Although the predominant salt-adapted organisms belong to halophilic *Archaea* such as the members of the family *Halobacteriaceae*, representatives of *Bacteria* and *Eukarya* can also thrive under these harsh conditions (Oren, 2008).

In general, halophiles adapt to the presence of salt by employing two main strategies to maintain the osmotic balance between the cytoplasm and the surrounding medium: the "salt-in-cytoplasm" strategy and the compatible solute strategy (Galinski, 1995; Sleator and Hill, 2001; Oren, 2008). The 'salt-in' strategy is characterized by increasing the salt concentration inside the cell, leading to significant changes in the enzymatic machinery. These include the over-representation of highly acidic amino acids such as aspartate (Asp), and a low proportion of hydrophobic residues that tend to form coil regions instead of helical structures when compared to non-halophile proteins (Paul et al., 2008; Rhodes et al., 2010). Microorganisms that use this strategy include the bacterium *Salinibacter ruber* and also extremely halophilic *Archaea* such as *Halobacterium* sp. whose proteins are very acidic (Oren, 2008). On the other hand, the compatible solute strategy is phylogenetically more widespread than the "salt-in" strategy and consists of the use of osmoprotectants or compatible solutes that do not interfere with the metabolism of the cell. In an initial phase of osmoadaptation using this strategy, high osmolarity conditions can trigger accumulation of K^+ ions in the cytoplasm, which can eventually lead to salt tolerance as they can serve as intracellular osmoprotectants (Csonka, 1989; Sleator and Hill, 2001). In a secondary response, compatible solutes can act as organic osmoprotectants that are biosynthesized and/or accumulated inside the cell to restore the cell volume and turgor pressure lost during the osmotic stress (Csonka, 1989; Sleator and Hill, 2001). There is a great variety of organic solutes that can act as osmoprotectants, including glycine betaine and glycerol. Some of these solutes are found in specific phylogenetic groups while others are widely distributed in halophilic organisms (Oren, 2008).

The vast majority of the mechanisms of elevated salt resistance and osmoprotection are derived from the knowledge of cultivated microorganisms and their sequenced genomes, thus this information may be biased and may overlook specific strategies of adaptation (Wu et al., 2009). In fact, previous studies using metagenomic sequencing approaches in well-characterized hypersaline environments have revealed novel lineages and genomes from diverse microorganisms without previously cultured representatives (Narasingarao et al., 2012; López-López et al., 2013). Moreover, recent genomic studies on the genus *Halorhodospira* have revealed a combined use of both strategies of salt adaptation (Deole et al., 2013) and through metagenomic analysis an acid-shifted proteome has been described in a hypersaline mat from Guerrero Negro (Kunin et al., 2008). On the basis of these findings, the notion of a correlation between phylogenetic affiliation and the strategy of osmotic adaptation should be revised (Oren, 2013).

Functional metagenomics is a culture independent approach, which is based on the construction of gene libraries using environmental DNA and subsequent functional screening of the resulting clones to search for enzymatic activities. Advantages of this approach include the identification of functional genes during the screening and also that the nucleotide sequences retrieved are not derived from previously sequenced genes, which enables the identification of both novel and known genes (Simon and Daniel, 2009; López-Pérez and Mirete, 2014). Thus, functional metagenomics has recently been used to identify novel genes involved in salt tolerance from microorganisms of a freshwater pond water (Kapardar et al., 2010) and also from the human gut microbiome (Culligan et al., 2012). Nevertheless, to our knowledge a functional metagenomic strategy has not been used to retrieve novel salt resistant genes from microorganisms of hypersaline environments. In this work, we employed this approach to search for salt resistance genes of microorganisms present in two different niches within a solar saltern in the south of Mallorca, Spain: (i) saturated sodium chloride brines, and (ii) moderate-salinity rhizosphere from the halophyte *Arthrocnemum macrostachyum*. To complement the study, the microbial diversity of the brines and the rhizosphere was characterized by amplifying and sequencing the 16S rRNA gene using 454 technology (pyrotagging). The microbial DNA from those samples was also used to construct two small-insert metagenomic libraries which were used to transform the *Escherichia coli* strain MKH13 which is more susceptible to elevated salt concentrations than wild type *E. coli* strains (Haardt et al., 1995). Library screening identified 11 different genes involved in salt resistance, some of which were similar to previously identified genes encoding for proteins conferring salt resistance whereas others encode for proteins that eventually may be related to novel salt resistance mechanisms.

MATERIALS AND METHODS

Bacterial Strains, Media, and Growth Conditions

Escherichia coli DH10B (Invitrogen) and MKH13 [MC4100 Δ*(putPA)101* Δ*(proP)2* Δ*(proU)*; Haardt et al., 1995] strains, and *Bacillus subtilis* PY79 strain (Youngman et al., 1984) were routinely grown in Luria-Bertani (LB) medium (Laboratorios Conda) at 37°C. *E. coli* DH10B was used as a host to maintain and to construct the metagenomic libraries. The growth medium for transformed *E. coli* strains was supplemented with 50 mg ml^{-1} ampicillin (Ap) to maintain the pBluescript SKII (+) plasmid ($pSKII^+$), and 100 mg ml^{-1} spectinomycin (Sp) for transformation of *B. subtilis* cells with the pdr111 plasmid. Screening for salt resistance clones and growth curves were carried out in LB medium supplemented with NaCl (Sigma). LB medium also contains NaCl (0.5%), however, the NaCl concentrations mentioned in this study are referred only to the supplemented NaCl.

For the growth curves, cells were cultured overnight in LB broth or LB broth supplemented with 3% NaCl at 37°C, then diluted to an OD_{600} of 0.01 with or without 3% NaCl and 200 ml was transferred to sterile a 96-well micro-titre plate (Starstedt, Inc., Newton, MA, USA) and grown at 37°C for 50 cycles (49 h). OD_{600} was measured every 60 min by using a microplate reader (Tecan Genios, Mannedorf, Switzerland). Non-inoculated wells served as the blank and their values were subtracted from those obtained in inoculated wells. All experiments were carried out in triplicate and the results for each data point were represented as the mean and SEM determined with OriginPro8 software (OriginLab Corporation, Northampton, MA, USA).

DNA Isolation from Brine and Rhizosphere Samples

Brine and rhizosphere samples used in this study were recovered from the Es Trenc saltern (Mallorca, Spain) in August 2012. Total salinity (%) was determined by refractometry and electric conductivity for brine and rhizosphere samples, respectively, and using three independent replicas. Microbial cells were collected from 400 ml of brine samples by filtration on a 0.22-mm-pore-size membrane filter (Nalgene). The filter was mixed with 5 ml of lysis buffer [100 mM Tris-HCl, 100 mM de EDTA, 100 mM Na_2HPO_4 (pH 8.6) and 1% SDS]. The mix was incubated at 65°C with occasional vortex mixing. Samples were centrifuged at 4500 rpm for 5 min at 4°C, and the supernatants were collected. Then, 1.7 ml of NaCl 5 M and 1.7 ml of 10% CTAB were added to the supernatant and then incubated in a 65°C water bath for 10 min with occasional vortex mixing. An equal volume of phenol-chloroform-isoamyl-alcohol (25:24:1; PCIA) was added and centrifuged at 12000 rpm for 15 min at room temperature. The aqueous layer was transferred to a fresh tube and an equal volume of chloroform was added. The mix was then centrifuged at 12000 rpm for 15 min at room temperature. The aqueous layer was removed and transferred to a fresh tube. To precipitate the DNA, 0.6 volumes of isopropanol were added to each tube, stored at room temperature for 1 h and centrifuged at 12000 rpm for 20 min at room temperature. After decanting the supernatant, the pellet was washed with 1 ml of 70% (vol/vol) EtOH and centrifuged at 12000 rpm for 5 min at room temperature. Finally, the pellet was air dried and resuspended in 200 μl of sterile deionized water.

Rhizosphere samples used in this study were obtained from plants of the species *A. macrostachyum*. These samples were kept in 50-mL tubes containing RNA Later (Sigma) and stored at −80°C. In order to extract DNA, the rhizosphere and the soil adhered to the roots were thawed and aseptically processed with the BIO101 FastDNA Spin kit for soil (Qbiogene) and the FastPrep device following to the manufacturer's recommendations.

Determination of the Community Structure of the Samples

PCR Amplification and 454-Pyrosequencing

16S rRNA gene amplification was performed using bacterial primer pairs GM3 and 630R for *Bacteria* (RB: *Bacteria* in rhizosphere and BB: *Bacteria* in brines), and 21F and 1492R for *Archaea* (RA: *Archaea* in rhizosphere and BA: *Archaea* in brine; Supplementary Table S1) and previously reported conditions (Lane et al., 1985). A five-cycle PCR was performed in a final volume of 25 μL in triplicate to incorporate tags and linker into the amplicon using 1:25 dilution of the original products as templates, and also using the same temperature cycles as for the first PCR. The second PCR was performed using the forward primers GM3-PS (*Bacteria*), 21F-PS (*Archaea*) and the reverse primer 907R-PS (Supplementary Table S1). The products were visualized after electrophoresis in 1% agarose gel run in 1X TAE buffer, at 25 V for 50 min. Two bands were observed, a first of ∼1500 bp and the second of ∼960 bp. The smaller band was excised and eluted using the Zymoclean^{TM} Gel DNA recovery Kit (Zymo Research, Orange, CA, USA) following the manufacturer's instructions. The concentration of the barcoded-amplicons was measured with Mass-Ruler Express forward DNA Ladder Mix (Thermo Scientific). Finally, an equimolar mixture of the amplicons was sent to the sequencing company Macrogen, Inc. (Seoul, Korea). The samples were sequenced using 454 GS-FLX+ Titanium technology. Sequences were submitted in the European Nucleotide Archive (ENA) under the Study Accession Number PRJEB9023 (samples ERS696577–80).

OTU (Operational Taxonomic Unit) Clustering, Phylogenetic Affiliation, and Selection of OPUs (Operational Phylogenetic Units)

Sequences with <300 nucleotides were removed, and low-quality sequences were trimmed with a window size of 25 and average quality score of 25. No ambiguities and mismatches in reads with primer pairs and barcodes were allowed. Chimeras were removed with the application Chimera Uchime. The trimming process was performed using Mothur software (Schloss et al., 2009). The adequate selected sequences were clustered in operational taxonomic units (OTUs) at 99% using the UCLUST tool in QIIME (Caporaso et al., 2010). We consider one OTU each unique cluster of sequences with identities ≥99%. The longest read of each OTU was selected as representative.

Phylogenetic inference was performed using the ARB software package (Ludwig et al., 2004). Sequences were aligned with SINA aligner (Pruesse et al., 2012), using LTPs115 database (Yarza et al., 2010). Alignments were manually inspected and improved, and sequences were added to the non-redundant SILVA REF115 database (Quast et al., 2013) with the ARB parsimony tool to a default tree. The non-type strain closest relative sequences of an acceptable quality were selected and merged with the LTP115 database. The Neighbor-Joining algorithm was used for the final tree reconstruction, with the Jukes-Cantor correction with *Bacteria* and *Archaea* filter depending Domain, using only almost complete sequences of all reference entries. Representative of each OTU were finally added to the reference tree with the parsimony tool. Sequences were grouped in operational phylogenetic units (OPUs; França et al., 2014) based on the visual inspection of the tree. We consider an OPU as the smallest clade containing one or more amplified sequences affiliating together with reference sequences available in the public repositories.

When possible, the OPUs should include a type strain sequence present in the LTP database (Yarza et al., 2010).

Ecological Indexes

Operational phylogenetic units were used to calculate rarefaction curves and the Shannon-Wiener (H'), Chao 1, and Dominance (D) indexes per sample with PAST *v* 3.01 software (Hammer and Harper, 2008).

Construction of Metagenomic Libraries

The construction of metagenomic libraries and their subsequent amplification was accomplished as previously described (Mirete et al., 2007; González-Pastor and Mirete, 2010). Briefly, the metagenomic DNA was partially digested using Sau3AI, and fragments from 1 to 8 kb were collected directly from a 0.8% low-melting-point agarose gel with the QIAquick extraction gel (QIAGEN) for ligation into the dephosphorylated and BamHI-digested pSKII$^+$ vector. DNA (100 ng) excised from the gel was mixed with the vector at a molar ratio of 1:1. Ligation mixtures were incubated overnight at 16°C using T4 DNA ligase (Roche) and used to transform *E. coli* DH10B cells (Invitrogen) by electroporation with a Micropulser (Bio-Rad) according to the manufacturer's instructions.

Screening for Salt Resistance

Recombinant plasmids from the metagenomic libraries constructed in *E. coli* DH10B cells were extracted using the Qiaprep Spin Miniprep kit (Qiagen) and ∼100 ng of vector were used to transform electrocompetent cells of *E. coli* MKH13. Electrocompetent cells of *E. coli* MKH13 were prepared according to Dower et al. (1988). Cells grown to mid-exponential phase (OD 0.6) were harvested by centrifugation and washed three times with a low salt buffer (1 mM Hepes, pH 7.0). Cells were resuspended in cold 10% glycerol and stored at −80°C.

After electroporation of MKH13 cells, $\sim 5 \times 10^4$ transformed cells per amplified library were subsequently screened on LB agar plates supplemented with 50 mg/ml Ap and 3% NaCl, a lethal concentration of salts for MKH13 cells. Plates were then incubated at 37°C for 72 h. To ensure that the resistance phenotype was not due to the presence of chromosomal mutations, the resistant colonies were pooled, their plasmidic DNA was isolated and it was used to transform MKH13 cells, and colonies were selected on LB-Ap plates without 3% NaCl. From each transformation, 100 colonies were patched onto LB-Ap plates containing 3% NaCl. Recombinant plasmids isolated from salt-resistant clones were digested with XhoI and XbaI, to select those which are unique in their restriction patterns.

In silico Analysis of Salt Resistant Clones

The DNA inserts of the plasmids from salt resistant colonies were sequenced on both strands with universal primers M13F and M13R and others for primer walking by using the ABI PRISM dye terminator cycle-sequencing ready-reaction kit (Perkin-Elmer, Waltham, MA, USA) and an ABI PRISM 377 sequencer (Perkin-Elmer), according to the manufacturer's instructions. Sequences were assembled and analyzed with the Editseq and Seqman programs from the DNAStar package. Prediction of potential open reading frames (ORFs) were conducted using ORF Finder and FGENESB (Solovyev and Salamov, 2011), which are available at the NCBI web page[1] and www.softberry.com, respectively. The bacterial code was selected, allowing ATG, CTG, GTG, and TTG as alternative start codons for translation to protein sequences. All the predicted ORFs longer than 90 bp were translated and used as queries in BlastP and their putative function was annotated based on their similarities to protein family domains by using Pfam (protein families) available at the European Bioinformatics Institute (EMBL-EBI[2]). Those sequences with an E value more than 0.001 in the BlastP searches and longer than 300 bp were considered as hypothetical. Transmembrane helices were predicted with TMpred[3].

Cloning of Genes Conferring Salt Resistance

To determine which ORFs were involved in salt resistance in the recombinant plasmids bearing more than one ORF, they were cloned individually in the vector pSKII$^+$. Thus, PCR-amplified fragments containing these genes were digested with XhoI/HindIII and XbaI restriction enzymes and ligated into pSKII$^+$ digested with the same restriction enzymes. The plasmids obtained were used to transform the MKH13 strain, and growth of the resulting clones was compared with that of the original clone carrying the entire environmental DNA fragment. PCR amplification of the ORFs was carried out using the following reaction mixture: 25 ng of plasmid DNA, 500 μM of each of the four dNTPs, 2.5 U of *Pfu* Ultra DNA polymerase (Stratagene) and 100 nM of each forward and reverse primers (described in Supplementary Table S2A, Supporting information) up to a total volume of 50 μl. The PCR amplification program used was as follows: 1 cycle of 5 min at 94°C, 30 cycles of 30 s at 94°C, 30 s at 52°C, 5 min at 72°C and finally 1 cycle of 10 min at 72°C. PCR amplification products were excised from agarose gels and purified using the Qiaquick Extration Gel kit (Qiagen). Purified PCR products were then digested with the appropriate restriction enzymes (Roche) and ligated into pSKII$^+$. To incorporate their native expression sequences (promoters and ribosome binding sites), a region of ∼200 bp located upstream of the start codon was also amplified. Some of the ORFs were truncated or the 5′ region was close to the polylinker sequence of the pSKII$^+$ vector, and they were subcloned in the same orientation as of the original clone. The *E. coli* genes encoding the endonuclease (*nth*) and the RNA helicase (*rhlE*) were amplified by PCR from DNA of the MKH13 strain (primers are described in Supplementary Table S2B) and similarly subcloned in the pSKII$^+$ vector. *E. coli* genomic DNA was isolated using the Wizard Genomic DNA Purification Kit as recommended by the manufacturer (Promega, Madison, WI, USA). The MKH13 strain was transformed with these genes and the growth of the resulting strains was tested by growth experiments carried out on LB-agar supplemented with 3% NaCl.

[1] http://www.ncbi.nlm.nih.gov/gorf/gorf.html

[2] http://pfam.xfam.org/

[3] http://www.ch.embnet.org/software/TMPRED_form.html

To assess the salt resistance in *B. subtilis,* the genes were cloned in plasmid pdr111 using the specific primer listed in Supplementary Table S3. This plasmid was a gift from D. Rudner (Harvard Medical School) and derives from pDR66, thus carrying front and back sequences of the *B. subtilis amyE* gene, which encodes an alpha-amylase. It also contains the hyper-SPANK promoter (Phyperspank), which is inducible by IPTG. The recombinant plasmids were then transferred to *B. subtilis* strain PY79 with selection for Sp resistance. pdr111 is not capable of replication in *B. subtilis*, thus the DNA fragment is inserted in the *amyE* locus in the chromosome, the transformants were screened for the absence of amylase activity on starch plates. Briefly, for transformation of *B. subtilis*, cultures grown overnight on LB broth at 30°C were diluted to OD600 nm of 0.08 in 10 ml of the modified competence medium (MCM) and were incubated at 37°C with agitation (200 rpm; Spizizen, 1958). At the onset of stationary phase (OD 600 nm = 1.5–2), 1 mg of the recombinant plasmids were added to 1 ml of the culture. Then, culture was incubated at least 2 h at 37°C and 200 rpm before plating on LB solid medium containing Sp (100 mg ml^{-1}). Growth curves were carried out as previously described either in the presence or in the absence of 1 mM IPTG.

Elemental Quantification of Na^+ in Resistant Clones

Escherichia coli MKH13 carrying the empty vector and recombinant clones were grown aerobically in LB liquid medium containing 50 mg ml^{-1} Ap at 37°C in a shaking incubator, and growth was monitored as optical density at 600 nm (OD_{600}). NaCl was added at 6% in early stationary phase to the cultures and grown for one additional hour. Cultures were washed four times extensively with ultrapure MiliQ H_2O and centrifugation. Washed pellets were lyophilized, pulverized and subsequently the concentration of Na^+ was measured by inductively coupled plasma spectroscopy-mass spectrometry (ICP-MS) analysis at SIdI (UAM, Madrid). Results were expressed as mg of Na^+ g^{-1} dry weight of cells. One-way ANOVA and Tukey's test were used for statistical analysis with OriginPro8 software (OriginLab Corporation, Northampton, MA, USA).

RESULTS

Microbial Community Structure of the Brine and Rhizosphere Samples

In order to search for genes that could confer increased salt resistance to *E. coli*, we sampled two sites in the hypersaline environment Es Trenc: (i) brine from a crystallizer pond (total salinity of 38.53 ± 0.23%), and (ii) moderate-salinity rhizosphere from the halophyte *A. macrostachyum* (total salinity of 3.28 ± 0.48%). DNA isolated from these samples was used to explore the bacterial and archaeal diversity. 16S rRNA gene sequences were clustered at an identity threshold 99%, resulting in a total of 970 OTUs (Supplementary Table S4) that after the phylogenetic inference produced a total of 226 OPUs, 200 for *Bacteria* and 26 for *Archaea* (**Figure 1**, Supplementary Table S5).

Most bacterial OPUs (187 OPUs*)* were detected only in RB, while BB contained just 13 OPUs, and only two were shared by both samples (OPUs 109 and 144). The sequences were distributed in 16 phyla (**Figure 1A**; Supplementary Table S5). A total of 102 OPUs affiliated with the phylum *Proteobacteria* (47 *Alpha-,* 8 *Beta-,* 30 *Gamma-,* and 17- *Deltaproteobacteria);* 31 with *Actinobacteria*, 27 with *Bacteroidetes* and 17 with *Firmicutes*. The major OPUs in RB were OPU 120 (*Ardenticatenamaritima,* 5.0%), OPU 153 (*Cytophagales,* 3.6%), OPU 125 (*Bacillus halosaccharovorans,* 3.3%), OPU 172 (*Actinobacteria,* 3.0%), OPU 90 (*Sorangiineae,* 2.9%) and, OPU 22 (*Rhodobacteraceae,* 2.4%). In no case one OPU exceeded 5.1% of the total sequences (Supplementary Table S5). On the other hand, the major OPUs in BB were OPU 102 (Uncultured GR-WP33–58, 43.38%, a *Deltaproteobacteria* close to *Myxobacteria*), OPU 143 (Uncultured *Chitinophagaceae,* 12.6%), and OPU 34 (Uncultured *Limimonas*, 12.6%). The latter OPU and the OPU 109 (*Rhodopirellula*) were the unique OPUs present both in RB and BB (Supplementary Table S5).

Sequences affiliated with *Archaea* generated lower diversity yields with 26 OPUs, all them in the *Euryarchaeota* phylum (**Figure 1B**). Most of the OPUs affiliated with *Halobacteriaceae* (90.8% for RA and 100% for BA). *Methanosarcinaceae* and *Methanoregulaceae* were present only in RA with 3.9 and 5.3%, respectively. The most representative in RA sample were OPUs 204 and 205 (*Haladaptatus* sp., 52.6%), OPUs 215 and 216 (*Halopelagius* sp., 10.5%), OPUs 201–203 (*Halococcus* sp., 9.2%), OPU 226 (*Methanolinea mesophila,* 5.3%), and OPU 225 (*Methanosarcina* sp., 3.9%). While, sequences in sample BA were represented principally by OPUs 209–213 (*Halorubrum* sp., 61.2%), OPU 220 (*Haloquadratum* sp., 16.7%), OPUs 221 and 222 (*Haloarcula* sp., 3.8%), OPU 208 (*Halomarina oriensis,* 3.7%), OPU 223 (*Halonotius* sp., 3.7%), and OPU 224 (*Halobacteriaceae,* 3.7%; Supplementary Table S5).

Bacterial diversity (H′) and richness (Chao-1) indexes were higher in RB (4.5 and 221.5, respectively) than in BB (1.8 and 12, respectively; Supplementary Table S4). However, the abundances were more homogeneously distributed in RB than in BB. In accordance Dominance index for RB was the lowest in comparison with all samples (Supplementary Table S4). *Archaea* presented similar values for diversity (2.0), richness (13), and dominance (0.2) in both samples.

Construction of Metagenomic Libraries

In order to search for genes that could confer increased salt resistance to *E. coli*, we screened two metagenomic libraries constructed in the high-copy-number vector $pSKII^+$ with environmental DNA isolated from brine and from rhizosphere samples. Approximately 236,250 (brine) and 192,000 (rhizosphere) recombinant clones were obtained and the libraries were subsequently amplified as described in Experimental procedures. Fragment length polymorphism analysis of 16 random clones per library showed an average insert size of 3 kb as shown in Supplementary Table S6. Overall, ∼1.2 Gb of environmental DNA was cloned within these libraries.

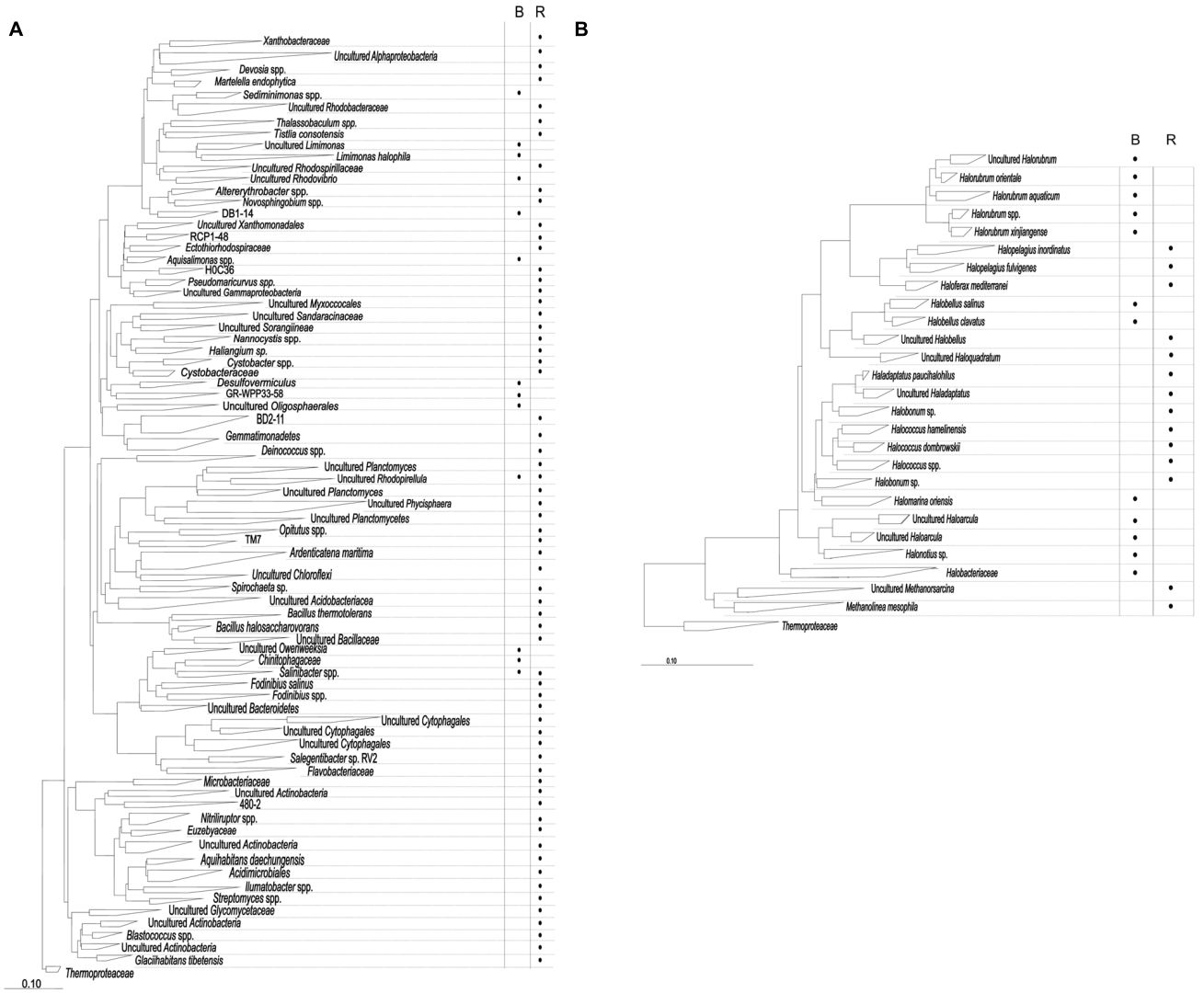

FIGURE 1 | 16S rRNA phylogenetic reconstruction for bacterial (A) and archaeal (B) sequences. The presence of OPUs with abundances >0.5% in each sample type (B = brines; R = rhizosphere) is indicated with a dot. Each OPU results from the phylogenetic inference resulting from the parsimony insertion of representatives of each sequence cluster at 99% identity, each representing independent OTUs.

Screening of the Metagenomic Libraries for NaCl Resistant Clones

Recombinant plasmids from the two metagenomic libraries constructed in *E. coli* DH10B strain were used to transform the osmosensitive *E. coli* MKH13 strain. MKH13 is less salt-resistant than *E. coli* wild type strains, because it carries mutations in the ProP and ProU transport systems involved in the efficient uptake of the osmoprotectant proline betaine (*N*,*N*-dimethyl-L-proline; Haardt et al., 1995). One of the main problems of using *E. coli* as a host for metagenomic libraries is to obtain the appropriate expression of genes from other microorganisms. Thus, the use of the MKH13 strain could favor the selection of genes conferring salt resistance, but poorly expressed in this bacterium. As a result, a total of 101 and 12 salt resistance clones were obtained for brine and rhizosphere samples, respectively. Of these, eight clones containing genes that conferred salt resistance to the host, pSR1–3 from brine and pSR4–8 from rhizosphere (**Table 1**) were found unique in their enzymatic restriction pattern. The strain MKH13 transformed with the recombinant plasmids showed a better growth rate in LB supplemented with 3% NaCl than MKH13 cells transformed with an empty vector (**Figures 2B,D**) whereas no differences in growth rate was observed in the presence of LB medium without supplemented NaCl (**Figures 2A,C**). All the clones were also assayed in the presence of LB supplemented with 4% NaCl and an increase in the growth rate was also observed in clones pSR2, pSR4, and pSR8 (data not shown).

A total of 14 genes were predicted using FGENESB and ORF Finder programs in the sequenced inserts from the eight plasmids (pSR1–pSR8) conferring salt resistance (**Table 1** and **Figure 3**). Sequence analyses of these environmental DNA

TABLE 1 | Description of NaCl-resistant plasmids (pSR1 to pSR8) and their observed sequence similarities.

Plasmid (type of sample)	GenBank accession number no.	Size (bp)	G+C content (%)	ORF[a]	Length (aa)[b]	Closest similar protein (organism_accession number)	Domain[c]	E value	% Identity	No. of TM-helices
pSR1 (brine)	KM603475	2478	70.26	**1***	171	Prolyl-tripeptidyl peptidase precursor (*Candidatus accumulibacter*); EXI70660; 322 aa	B	3,00E-72	102/161 (63%)	0
				2*	621	DNA helicase II (*Salinibacter ruber* DSM 13855); YP_003572414; 1227 aa	B	0.0	464/484 (96%)	0
pSR2 (brine)	KM603476	1509	56.20	**1***	337	Hypothetical protein (*Natrinema pellirubrum*); WP_006182474; 833 aa	A	0.0	291/337 (86%)	1
				2	96	Hypothetical protein (*Halosimplex carlsbadense*); WP_006886095; 168 aa	A	5,00E-51	80/96 (83%)	0
pSR3 (brine)	KM603477	1838	60.50	**1***	392	Probable cell surface glycoprotein (*Natronomonas moolapensis* 8.8.11) YP_007488500	A	7,00E-108	212/382 (55%)	2
				2*	97	IISH7-type transposase (*Natronomonas moolapensis* 8.8.11) YP_007488498.1; 558 aa	A	8,00E-50	85/97 (88%)	0
pSR4 (rhizosphere)	KM603478	920	62.61	**1**	217	Endonuclease III (*Verrucomicrobia bacterium* DG1235); WP_008102102; 229 aa	B	1,00E-126	174/217 (80%)	0
pSR5 (rhizosphere)	KM603479	1800	64.15	**1***	304	Hypothetical protein (*Monosiga brevicollis* MX1); XP_001749465; 1630 aa	E	9,00E-27	96/270 (36%)	1
				2	164	Site-specific recombinase (*Pseudomonas fuscovaginae*); WP_029533036; 197 aa	B	4,00E-24	56/110 (51%)	0
pSR6 (rhizosphere)	KM603480	2341	49.40	1*	168	OmpA/MotB domain-containing protein (*Haliscomenobacter hydrossis* DSM 1100) YP_004451090; 1170 aa	B	5,00E-48	82/168 (49%)	0
				2	214	Glycerol uptake facilitator or related permease (*Methylacidiphilum infernorum* V4); YP_001939268; 209 aa	B	3,00E-38	86/202 (43%)	6
				3*	216	Hypothetical protein (*Paenibacillus daejeonensis*) WP_020618935; 465 aa (putative sulfatase)	B	1,00E-78	117/240 (49%)	0
pSR7 (rhizosphere)	KM603481	1131	54.64	**1***	377	RNA helicase (*Methylophaga thiooxydans*); WP_008290720; 605 aa	B	0.0	323/383 (84%)	0
pSR8 (rhizosphere)	KM603482	426	70.66	**1***	142	Pyrophosphate-energized proton pump (*Ilumatobacter coccineus* YM16–304) YP_007562997; 698 aa	B	1,00E-50	95/142 (67%)	3

[a]*ORFs involved in NaCl resistance are shown in boldface type, and asterisks indicate incomplete ORFs.*
[b]*aa, amino acids.*
[c]*A: Archaea, B: Bacteria, and E: Eukarya.*

fragments revealed the presence of one unique ORF in pSR4, pSR7 and pSR8, two ORFs in pSR1, pSR2, pSR3 and pSR5, and three ORFs in pSR6. The G+C content of these DNA fragments varied from 49.4 to 70.7% indicating their diverse phylogenetic origin. Most of the genes analyzed in this study encoded amino acid sequences similar to bacterial proteins whereas the inserts present in pSR2 and pSR3 may have been retrieved from archaeal organisms due to their similarities with

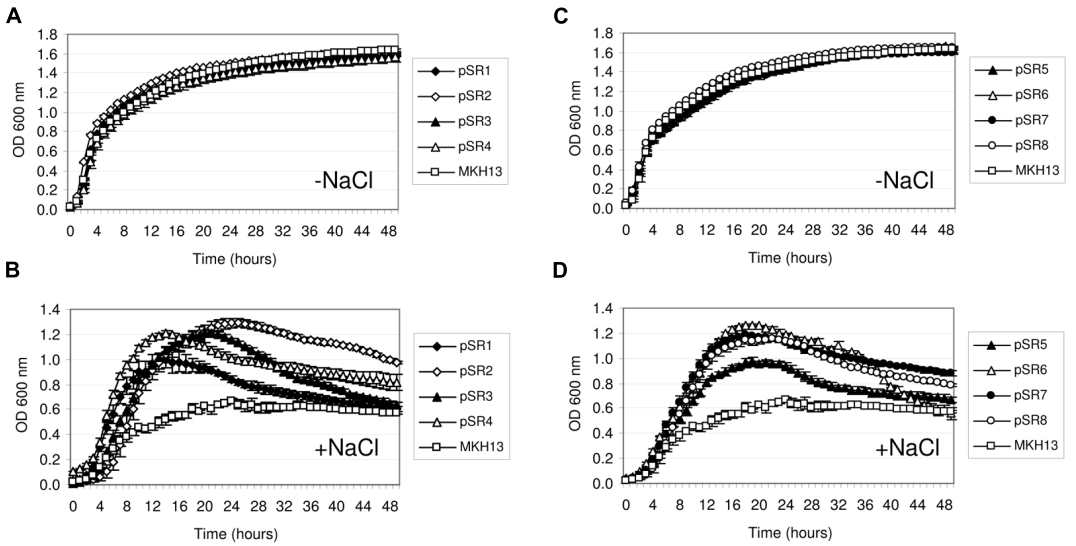

FIGURE 2 | Growth curves of *Escherichia coli* MKH13 cells carrying plasmids with salt resistance genes (pSR1-pSR8) and MKH13-pSKII$^+$ in LB broth and LB broth supplemented with 3% NaCl. Clones pSR1, pSR2, pSR3, pSR4, and MKH13-pSKII$^+$ in LB broth **(A)** and LB broth supplemented with 3% NaCl **(B)**. Clones pSR5, pSR6, pSR7, pSR8, and MKH13-pSKII$^+$ in LB broth **(C)** and LB broth supplemented with 3% NaCl **(D)**.

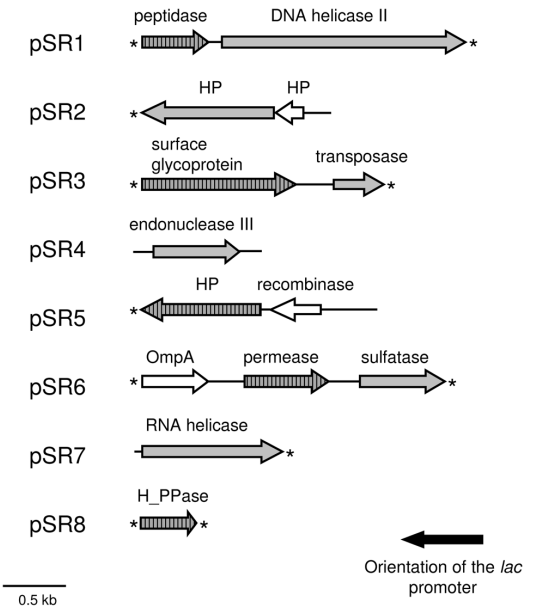

FIGURE 3 | Schematic organization of the ORFs identified in the pSR1-pSR8 plasmids. Arrows denote the location and the transcriptional orientation of the ORFs in the different plasmids. ORFs involved in NaCl resistance are indicated by gray arrows and those whose phenotype was not resistant are shown in white arrows. The presence of predicted transmembrane helices is represented by arrows shaded with vertical bars. Asterisks indicate incomplete ORFs. HP, hypothetical protein.

members of this domain. In addition, BLASTP analyses revealed that pSR5-*orf1* may be from eukaryotic origin whereas pSR5-*orf2* was probably derived from a bacterium related to the *Pseudomonas* genus. This result suggests that pSR5 may be a chimeric clone or that this clone may be derived from a fragment of a mobile element. Alternatively, pSR5-*orf1* may be just an uncommon bacterial gene with the eukaryotic sequence being the closest gene sequenced. BLASTP as well as the protein family domains (Pfam) databases were used to functionally categorize the genes retrieved and showed that pSR1-*orf2* and pSR4-*orf1* encoded proteins related to DNA repair processes such as a DNA helicase II and an endonuclease III, respectively (**Table 1** and Supplementary Table S7). It is also interesting to note that genes related to structural dynamics of nucleic acids were also retrieved, including a IISH7-type transposase encoded by pSR3-*orf2*, a putative site-specific recombinase encoded by pSR5-*orf2* and a putative RNA helicase, particularly a DEAD-box helicase encoded by pSR7-*orf1* (**Table 1**). The deduced amino acid sequence of pSR7-*orf1* contained the five conserved sequence motifs found in members of the DEAD-box helicase family: II or Walker B (VLDEADEM; positions 10–17), III (SAT; positions 43–45), IV (IIFVRT; positions 105–110); V (LVATDVAARGLD; positions 155–166) and VI (YVHRIGRTGRAG; positions 185–196). Putative proteins encoded by pSR3-*orf1*, pSR6-*orf2,* and pSR8-*orf1* were similar to a cell surface glycoprotein, a permease related to glycerol uptake and a proton pump, respectively. These may be related to either transport mechanisms or to membrane components, in agreement with the presence of transmembrane segments predicted in their amino acid sequences (**Table 1**). The protein encoded by pSR6-*orf3* showed homology with choline-sulfatases from *Vibrio* sp., *Cyclobacterium qasimii* and *Clostridiales*. Also, it contained the motif SDHGEFL (positions 71–77), which is highly similar to a peptide signature apparently specific to choline sulfatases SDHGDML (Cregut et al., 2014).

In addition, hypothetical proteins were also found, such as those encoded by pSR2-*orf1,* pSR2-*orf2,* pSR5-*orf1,* and pSR6-*orf3*. In the case of pSR5-*orf1*, Pfam analysis showed that the

encoded protein contained a VWA (von Willebrand factor type A) domain present in some eukaryotes (Supplementary Table S7).

Identification of Genes Conferring NaCl Resistance

The recombinant plasmids pSR4, pSR7, and pSR8 contained a single ORF each, encoding an endonuclease III, a RNA helicase and a proton pump, respectively, which are responsible for the NaCl resistance phenotype (**Table 1**, **Figures 2B,D**). Five recombinant plasmids contained more than one ORF (pSR1, pSR2, pSR3, pSR5, and pSR6) as shown in **Table 1** and **Figure 3**. The DNA insert of pSR1 contains two ORFs: *orf1* encoding a peptidase S9 and *orf2* encoding a DNA helicase II. Clones harboring each one of these ORFs were NaCl resistant since an increase in the growth rate was observed compared to the growth of MKH13-pSKII$^+$cells, and even slightly more pronounced than that of the original clone (Supplementary Figure S1). In the case of the DNA insert from pSR2, two ORFs were identified, both encoding hypothetical proteins. pSR2-*orf1* clearly conferred resistance to NaCl whereas the slight resistance observed in the growth of pSR2-*orf2* (Supplementary Figure S2B) may be explained by its limited growth in LB not supplemented with NaCl (Supplementary Figure S2A). The sequence of the DNA insert of pSR3 plasmid revealed that it contained two ORFs, *orf1* encoded a probable cell surface glycoprotein whereas *orf2* encoded a IISH7-type transposase. These two genes were both involved in the NaCl resistance observed in the original clone as shown in Supplementary Figure S3. In the case of the DNA sequence of pSR5 two ORFs were identified and whose amino acid sequences were similar to a hypothetical protein (*orf1*) and to a recombinase (*orf2*). The increased growth rates observed for these clones revealed that pSR5-*orf1* provided NaCl resistance when compared with that of MKH13-pSKII$^+$, and its growth rate was similar to that of the original clone although slightly delayed (Supplementary Figure S4), whereas the growth rate of the clone harboring pSR5-*orf2* was reduced when compared with that of the control strain in the LB medium supplemented with NaCl (Supplementary Figure S4B). Three ORFs were found in the DNA insert of pSR6, encoding a protein similar to an OmpA (*orf1*), a permease involved in glycerol uptake (*orf2*) and a putative permease (*orf3*). Clones containing *orf2* and *orf3*, but not *orf1*, exhibited higher growth rates than that observed in the control (MKH13 pSKII$^+$) in LB medium supplemented with NaCl, indicating that *orf2* and *orf3* may be responsible for the NaCl resistance observed in the original clone (Supplementary Figure S5).

Assessment of Salt Resistance in the *E. coli* Homologs of Environmental Genes

The discovery of salt-resistance genes related to nucleic acid metabolism has been an interesting finding in this work. Thus, to explore the specificity of these environmental genes in the resistance phenotype, their *E. coli* homologs were cloned and tested for growth in the presence of NaCl. The proteins encoded by pSR4-*orf1* and pSR7-*orf1* were similar to the endonuclease III (Nth, 38.53% identity; 49.54% similarity) and the DEAD-box RNA helicase (RhlE, 31.94% identity; 46.86% similarity) of *E. coli*, respectively. These genes were PCR amplified using genomic DNA from MKH13 cells, digested with either XhoI or HindIII and XbaI and ligated into pSKII$^+$ digested with the same restriction enzymes. The growth on LB supplemented with 3% NaCl of the clones harboring the environmental genes, their *E. coli* homologs and the empty plasmid were compared. As a result, the growth rates of the strain carrying the *nth* gene of *E. coli* and the control strain (MKH13 pSKII$^+$) were similar in contrast with the increased growth rate observed for the clone pSR4 (**Figure 4**), indicating that the environmental endonuclease III but not its *E. coli* homolog specifically conferred salt resistance. The growth of the clone carrying the pSR7 plasmid, which encoded a protein similar to a DEAD-box RNA helicase, and the clone containing the *rhlE* gene of *E. coli* were also compared. As a result, we observed a reduced growth rate of the *rhlE* clone in the presence of LB alone and a prolonged lag phase in the presence of NaCl (**Figure 5**). These results suggest that the RNA helicase of environmental origin may provide a faster adaptation to the presence of NaCl in LB medium than its *E. coli* homolog.

Expression of Salt Resistance Genes in *Bacillus subtilis*

In order to investigate the expression of some of the retrieved environmental genes involved in salt resistance in other hosts than *E. coli*, four of the identified genes were transferred to the

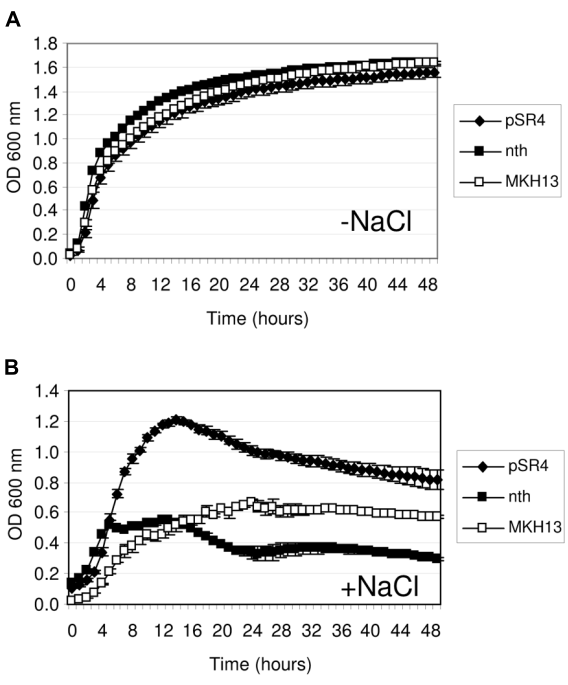

FIGURE 4 | Growth curve of *E. coli* MKH13 cells carrying pSR4, the *E. coli nth* gene, and MKH13-pSKII$^+$ in LB broth (A) and LB broth supplemented with 3% NaCl (B).

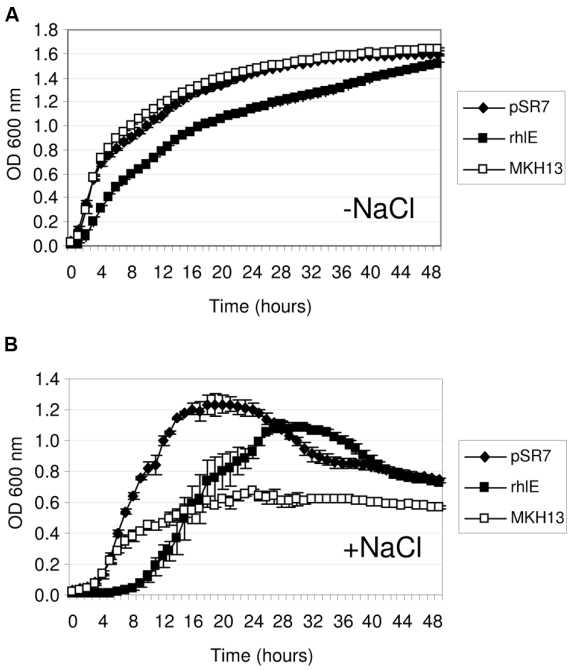

FIGURE 5 | Growth curve of *E. coli* MKH13 cells carrying pSR7, the *E. coli rhlE* gene and MKH13-pSKII$^+$ in LB broth (A) and LB broth supplemented with 3% NaCl (B).

model organism *B. subtilis* (PY79 strain). This bacterium was chosen as a representative of Gram-positive bacteria because it is suitable for genetic manipulation (Earl et al., 2008). PY79 strain exhibited increased resistance to NaCl than *E. coli* MKH13, thus salt concentration was adjusted to 6% in the growth experiments. The genes selected to be expressed in *B. subtilis* were those related to metabolism of nucleic acids (pSR1-*orf2*, pSR4-*orf1,* and pSR7-*orf1*) and also one encoding for a protein similar to a permease (pSR6-*orf2*). These four genes were subcloned into pdr111 vector, under an inducible IPTG promoter, the hyper-SPANK promoter. The resulting constructions were inserted at the *amyE* locus in the *B. subtilis* chromosome. In the growth experiments, bacteria carrying the empty vector inserted in the chromosome were used as negative control. Interestingly, *B. subtilis* transformed with these genes and grown either in the presence or in the absence of IPTG exhibited an increased growth rate in comparison with the negative control, as shown in **Figure 6**. These results indicated that some basal level of expression is occurring when *B. subtillis* was transformed with these environmental genes. From these, all the clones but pSR7-*orf1* showed a slight higher growth rate in the presence of salt in the medium when IPTG was supplemented than those without it, indicating that these genes were induced by IPTG, and properly expressed by *B. subtilis*, conferring resistance to NaCl.

Determination of Cellular Na^+ Content

To assess the extent by which clones pSR1 to pSR8 can accumulate Na^+ ions, the cellular concentration of this element was measured by ICP-MS after 1 h of growing bacterial cells with 6% NaCl (**Figure 7**). From the quantification of Na^+, resistant clones were grouped into two categories according to whether these clones can accumulate more or less sodium. The first group consisted of clones which accumulated more sodium than the control (pSR3, pSR4, and pSR7). This included clones involved in DNA repair such as the endonuclease III encoded by pSR4-*orf1*. The second group showed the same sodium concentration in the cell compared to the control cells (pSR1, pSR2, pSR5, pSR6, and pSR8). This included clones carrying genes related to the modification of DNA such as the DNA helicase II (pSR1-*orf2)* or of unknown function (pSR2-*orf1*). It also included clones with genes that may be involved in osmotic equilibrium such as pSR6 with two genes, pSR1-*orf2* and pSR6-*orf3*, encoding a glycerol permease and a putative sulfatase, respectively and pSR8, with one gene, pSR8-*orf1*, encoding for a proton pump.

Further quantification of the cellular content of Na^+ ions determined by ICP-MS on the pSR6 clone revealed that the recombinant plasmid encoding only the putative permease, pSR6-*orf2*, accumulated significantly more sodium than the original and pSR6-*orf3* clones and also more than MKH13 cells (**Figure 8**).

DISCUSSION

Functional metagenomics allows access to the potential genetic diversity of both cultured and uncultured bacteria present in a particular environment (Handelsman, 2004). Therefore, this approach was used in this study to decipher the molecular mechanisms that may contribute to the overall cellular resistance and by which microbial communities adapt to high salt content. This has been employed in diverse studies aimed to elucidate the mechanisms of adaptation of microbial consortia to a number of extreme conditions such as high nickel and arsenic content, and acidic pH from the acid mine drainage environment of Rio Tinto (Mirete et al., 2007; González-Pastor and Mirete, 2010; Guazzaroni et al., 2013; Morgante et al., 2014). Although functional metagenomics has been applied to screen for genes related to salt resistance in environmental samples from the human gut microbiome (Culligan et al., 2012, 2013), and also from a freshwater pond (Kapardar et al., 2010), to the best of our knowledge this is the first study to report novel salt resistance determinants from microorganisms of a hypersaline environment by using functional screening of metagenomic libraries.

The two samples from which the metagenomes originated exhibited a microbial composition in accordance with the kind of sample (soil or brine) and high salinities. The rhizosphere was very diverse in its bacterial composition with 187 distinct OPUs in accordance with the known complexity of the system (Philippot et al., 2013). The relative abundances of the representatives of each lineage were well-balanced and none exceeded the 5.1% of the total diversity. The composition of the main taxonomic groups were *Alpha-* and

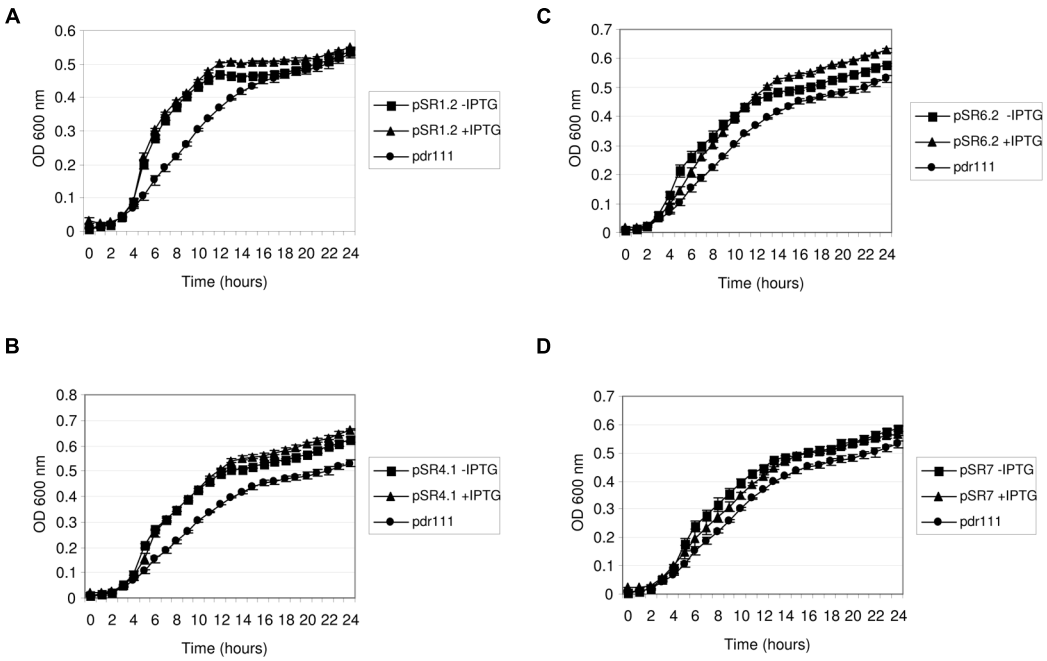

FIGURE 6 | Growth of *Bacillus subtilis* clones in NaCl. *B. subtilis* clones pSR1-*orf2* **(A)**, pSR4-*orf1* **(B)**, pSR6-*orf2* **(C)** and pSR7-*orf1* **(D)** were grown in LB broth supplemented with 6% NaCl in the presence and in the absence of 1mM IPTG. *B. subtilis* strain PY79 with the empty plasmid pdr111inserted in the chromosome was used as negative control.

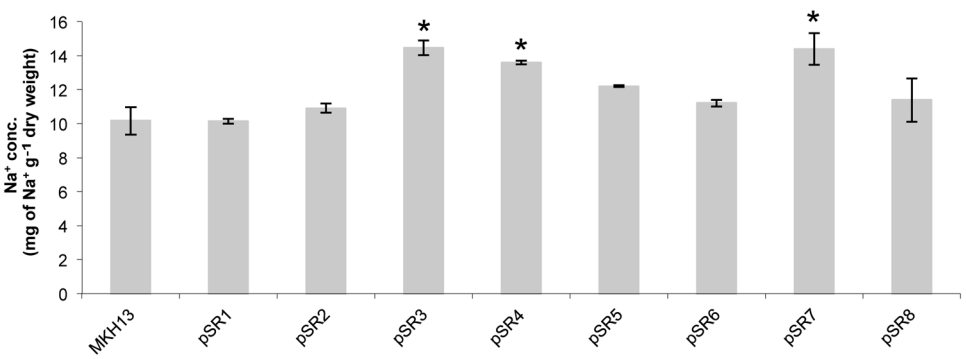

FIGURE 7 | Test for cellular content of Na^+ ion in *E. coli* clones pSR1 to pSR8 and MKH13-pSKII$^+$ after 1 h of growth with 6% NaCl. Values are the averages of two independent ICP-MS measurements. Error bars indicate standard deviation. An asterisk indicates significantly different from control cells as determined by one-way ANOVA followed by Tukey's test ($p < 0.05$).

Gammaproteobacteria and especially deltaproteobacterial which are close relatives to *Myxobacteria*, together with *Actinobacteria, Firmicutes, Bacteroidetes,* and *Gemmatimonadetes* are known to be common inhabitants of rhizosphere soils (Philippot et al., 2013). It is worth noting the relative high abundances of organisms related to *A. maritima*, a *Chloroflexi* representative known as an iron and nitrate reducer (Kawaichi et al., 2013), and *B. halosaccharovorans*, a moderately halophilic *Firmicutes*, both in accordance with the saline conditions of the environment (Mehrshad et al., 2013). The archaeal composition was less complex with only representatives of the *Halobacteriaceae* family in accordance with the high salinity concentrations (Oren, 2008), and representatives of the Rice Cluster I methanogens (*Methanosarcinales* and *Methanomicrobiales*; Conrad et al., 2006) also common in soils and widely distributed. The most remarkable observations were the high abundance (over 50% of the total archaeal diversity) of a close relative of the halobacterial genus *Haladaptatus*, originally isolated from low-salt and sulfide rich environments (Savage et al., 2007); and the methanogenic species *M. mesophila* initially described in rice field soil (Sakai et al., 2012), and member of the Rice Cluster I (Conrad et al., 2006). Altogether the results on the community structure of this soil agree with the fact that the anaerobic hypersaline sediments below the brine crystallizers may be a

source of methane and sulfide (López-López et al., 2010), and these may influence (by diffusion of ions and migration of microorganisms) the surrounding soils from which the plants were sampled.

The microbial composition of the salt brines was remarkable. The archaeal community was only constituted by members of *Halobacteriaceae* and with the genera *Haloquadratum, Halorubrum,* and *Haloarcula* as the most abundant. This structure was in accordance with the known microbiota in brines (Oren, 2008). However, the bacterial composition was remarkably different from what was expected. In general *Salinibacter* representatives have been found to be the major bacterial fraction in brines, in proportions that range from 5 to 30% (Antón et al., 2008). However, despite sequences of this lineage being found in the brines studied here, these constituted a minority (about 5% of the total bacterial diversity). The most conspicuous observation was the detection of three major groups of bacteria not previously observed as major components with ecological relevance in hypersaline habitats. The most represented bacterial lineage affiliated with representatives of the uncultured myxobacterial clade GR-WP33–58. Sequences of this deltaproteobacterial lineage were first detected in deep-sea Antarctic samples (Moreira et al., 2006). However, since its initial detection, similar sequences were retrieved mostly in marine samples (according to the identifiers in the entries from the NCBI). Some sequences of this clade had also been retrieved from hypersaline microbial mats (Harris et al., 2013) and saline soils (Castro-Silva et al., 2013), pointing to that its presence in brines may not be anomalous. The second most relevant proteobacterial group detected, and also in higher sequence abundances than *Salinibacter* were relatives of *Limimonas* (Amoozegar et al., 2013), an extremely halophilic member of *Rhodospirillaceae*. Finally, a third relevant group affiliated with relatives of the *Chitinophagaceae* lineage within *Bacteroidetes*. Similar sequences were detected in the hypersaline Lake Tyrrel in Australia (Podell et al., 2013). Despite the sequences retrieved for the bacterial domain being in accordance with the hypersaline nature of the sample, the lower occurrence of *Salinibacter*, and the prevalence of representatives from the uncultured GR-WP33–58 clade need further investigation as such community structure has not been observed before.

The construction of metagenomic libraries and their subsequent functional screening to search for novel salt resistance genes was considered in this study taking into account the microbial diversity observed in the brine and rhizosphere samples. It is worth to note that the genes identified here and those found in the natural host may not be involved in a similar degree of salt tolerance. In general, a correlation was observed between the putative phylogenetic affiliation of the environmental DNA fragments present in the positive clones and the sample origin (brine or rhizosphere). For example ORFs identified in clones derived from the brine sample (pSR1–pSR3) were similar to those from organisms detected in brine samples such as members of *Salinibacter* and *Halobacteriaceae* whereas ORFs from clones derived from the rhizospheric soil (pSR4–pSR8) were assigned to microorganisms found in this sample

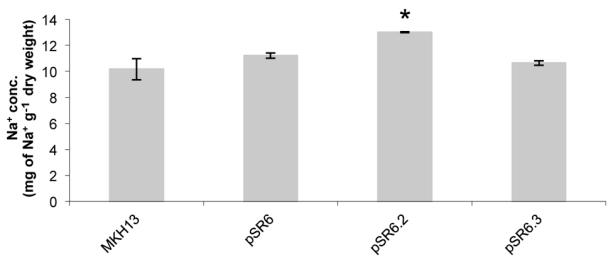

FIGURE 8 | Test for cellular content of Na^+ ion in *E. coli* clones pSR6, pSR6-*orf2*, pSR6-*orf3,* and MKH13-pSKII$^+$ after 1 h of growth with 6% NaCl. Values are the averages of two independent ICP-MS measurements. Error bars indicate standard deviation. An asterisk indicates significantly different from pSR6, pSR6-*orf3* and control cells as determined by one-way ANOVA followed by Tukey's test ($p < 0.05$).

including representatives of *Gammaproteobacteria*, *Firmicutes*, *Verrucomicrobia*, *Bacteroidetes,* and *Actinobacteria*.

In microorganisms, a well-known response to salt stress is the increase in concentration in the cytoplasm of compatible solutes such as glycerol and glycine betaine, in response to an elevated osmolarity in the surrounding medium. The synthesis of these solutes is often energetically less favorable than the uptake from the external environment and thus the accumulation of compatible solutes can inhibit endogenous synthesis (Sleator and Hill, 2001). The finding of pSR6-*orf2,* which encoded a putative glycerol permease, and conferred NaCl resistance not only in *E. coli* MKH13 but also in *B. subtilis*, illustrates the presence of this strategy within the rhizosphere bacterial community. Also, pSR6-*orf3* encoded a putative choline sulfatase, which was responsible for the resistance phenotype observed when it was cloned independently. Choline sulfatases encoded by *betC* genes are necessary to convert choline sulfate into choline and are found in several microorganisms present in rhizospheric environments including *Sinorhizobium meliloti* (Østerås et al., 1998). Although the *betC* gene is absent within the *E. coli* genome, we can assume that the presence of a gene encoding a choline sulfatase may favor the synthesis of glycine betaine from choline since in *E. coli* cells this last conversion can be carried out through two oxidations steps catalyzed by a choline dehydrogenase (BetA) and a glycine betaine aldehyde dehydrogenase (BetB; Østerås et al., 1998; Sleator and Hill, 2001). It is interesting to note that only the clone carrying pSR6-*orf2* accumulated more Na^+ than the control, the original clone pSR6 and pSR6-*orf3*.

In addition, an ORF from pSR8 encoding a proton pumping membrane-bound pyrophosphatase (H^+-PPase) was identified in this study. These proteins have been found in all three domains of life and can confer resistance to cells against diverse abiotic stress such as cold, drought, NaCl and metal cations, probably because the enzyme generates a membrane potential by using PPi (Yoon et al., 2013; Tsai et al., 2014). Membrane-bound pyrophosphatases can require Na^+ for their activity and they can also catalyze the transport of Na^+ outside the cell, as it has been demonstrated in the archaeal PPase from the mesophile *Methanosarcina mazei* and in two bacterial

PPases from the hyperthermophile *Thermotoga maritima* and the moderate thermophile *Moorella thermoacetica* (Malinen et al., 2007). More recently, an integral membrane pyrophosphatase subfamily has been described in diverse bacterial species which has the ability to transport both Na^+ and H^+ outside bacterial cells and which may have evolved from Na-PPases (Luoto et al., 2013). Thus, the membrane-bound pyrophosphatase encoded by pSR8-*orf1*, coupled with Na^+/H^+ antiporters present in *E. coli*, may be playing an important role in the adaptation of bacterial cells to increased salt content (Baykov et al., 2013).

A relevant finding derived from this study is the identification of salt resistance genes related to DNA repair and to structural dynamics of nucleic acids. Examples of these genes are pSR1-*orf2* and pSR7-*orf1*, which encoded a DNA and a DEAD-box RNA helicase, respectively. These genes were also responsible for the NaCl resistance phenotype observed in *B. subtilis*. Interestingly, the environmental RNA helicase encoded by pSR7 showed better adaptation to NaCl than that cloned from *E. coli*. DNA helicases are involved in unwinding double strand DNA and thus play key roles in cellular processes such as recombination, replication, transcription and repair processes whereas RNA helicases are capable of unwinding RNA duplexes and thus participate in ribosome biogenesis, transcription, translation initiation and RNA degradation (Tanner and Linder, 2001; Delagoutte and von Hippel, 2002; Kaberdin and Bläsi, 2013). In bacteria, DEAD-box RNA helicases involved in cold and oxidative stress response have been reported in the cyanobacterium *Anabaena* sp. (Yu and Owttrim, 2000) and in *Clostridium perfringens* (Briolat and Reysset, 2002), respectively. Also, upregulation of both RNA and DNA helicases transcript levels has been observed when *Desulfovibrio vulgaris* was exposed to elevated sodium chloride concentration (Mukhopadhyay et al., 2006). The role played by these helicases may be similar to that observed in other enzymes involved in the molecular conformation of nucleic acids. In plants, these proteins have been shown to be also related to salt stress. For example, the DEAD-box DNA/RNA helicase from pea overexpressed in tobacco conferred increased salt resistance (Sanan-Mishra et al., 2005) and DEAD-box RNA helicases are induced under elevated salt conditions in *Hordeum vulgare* (Nakamura et al., 2004) and in the halophyte *Apocynum venetum* (Liu et al., 2008). In our study, the cells carrying the DEAD-box RNA helicase encoded by pSR7-*orf1* showed more accumulation of Na^+ ions than the control, which was also reported in the leaves of transgenic tobacco plants overexpressing the DEAD-box helicase (Sanan-Mishra et al., 2005). Thus, this protein may be linked to a more specific response to salt stress that may allow the accumulation of Na^+ ions inside the cell. This will be the basis for future studies to clarify the precise molecular mechanism of salt resistance conferred by the DEAD box DNA/RNA helicases.

A resistance phenotype to NaCl was observed in clone pSR4, which encoded a protein similar to an endonuclease III. In *E. coli* this protein is encoded by the *nth* gene and displays DNA glycosylase activity involved in base-excision repair as a cellular defense against a variety of DNA damages caused by desiccation and UV irradiation (Kish and DiRuggiero, 2012). The enzymatic activity of Nth is specific for the repair of oxidized bases in DNA, particularly pyrimidines substrates such as thymine glycol, 5-hydroxycytosine and 5-hydroxyuracil (Dizdaroglu, 2005). Repair of oxidized DNA bases after exposure to elevated doses of gamma radiation has been reported in the extremely halophilic archaeon *Halobacterium salinarum* (Kish et al., 2009) whose genome contains diverse homologs of DNA glycosylases including *nth* homologs (Dassarma et al., 2001). The endonuclease III identified in this study, which also conferred salt resistance in *B. subtilis* (**Figure 6**), was similar to the *E. coli* Nth, however, the latter did not confer salt resistance (**Figure 4**). Although, to the best of our knowledge, the effect of high salt concentrations on DNA modifications *in vivo* has not been described before, our results suggest the possibility of a specific role in repairing DNA lesions produced by NaCl in both *E. coli* and *B. subtilis* cells. Also, in the human gut environment, two genes encoding MazG were found to be involved in salt tolerance, and it was suggested that this protein may play a role in the removal of abnormal nucleotides from nascent DNA strands (Culligan et al., 2012). Diverse DNA repair pathways have been identified to withstand diverse environmental stress associated to hypersaline environments such as ionizing radiation (IR) or desiccation in halophiles (Kish and DiRuggiero, 2012) and also in the rhizosphere-associated bacterium, *Sinorhizobium meliloti* (Humann et al., 2009), which is in agreement with the rhizosphere origin of pSR4-*orf1*.

CONCLUSION

The two different samples from a hypersaline environment (i.e., brine and rhizosphere) studied in this work exhibited a microbial composition that was in agreement with their saline nature. The rhizospheric soil showed a balanced community structure comparable with other such samples. The brine community structure was in agreement with what was expected for the archaeal counterpart, but not for the bacterial composition. Conspicuously, the bacterial diversity was dominated by three lineages never reported as major components of hypersaline habitats, and the expected major key player *Salinibacter* was in a noticeable minority. The use of functional metagenomics allowed the identification of diverse genes conferring salt resistance to *E. coli* and encoding for: (i) well-known proteins involved in osmoadaptation such as a glycerol permease and a proton pump, (ii) proteins related to repair, replication and transcription of nucleic acids such as RNA and DNA helicases and an endonuclease III, and (iii) hypothetical proteins of unknown function. It is worth noting that the environmental endonuclease III and the hypothetical proteins identified here may represent novel mechanisms of osmoadaptation. The link between DNA repair enzymes and stress processes involved in cellular dehydration such as desiccation and UV radiation have been previously described in *Deinococcus radiodurans* (Mattimore and Battista, 1996; Kish and DiRuggiero, 2012). To our knowledge this is the first report to identify a

specific DNA repair gene from a moderate-salinity rhizosphere associated with a hypersaline environment which can provide salt resistance to *E. coli*. Further analysis of these genes will be necessary to elucidate their precise mechanism of action.

ACKNOWLEDGMENTS

We would like to thank Rubén Morón and Margarita Rodríguez for their active interest and assistance in the laboratory. We are also grateful to Dr. Erhard Bremer (Laboratory for Molecular Microbiology, Faculty of Biology, Philipps University of Marburg, Marburg, Germany) for kindly providing *E. coli* MKH13. We also thank Josefa Antón Botella for critical reading of the manuscript. This work was funded by the Spanish Ministry of Science and Innovation (CGL2012-39627-C03/02 and 03); the latter also supported with European Regional Development Fund (FEDER). MM-R Ph.D. is supported by fellowship CVU 265934 of the National Council of Science and Technology (CONACyT), Mexico.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01121

REFERENCES

Amoozegar, M. A., Makhdoumi-Kakhki, A., Ramezani, M., Nikou, M. M., Fazeli, S. A. S., Schumann, P., et al. (2013). *Limimonas halophile* gen. nov., sp., nov., an extremely halophilic bacterium in the family Rhodospirillaceae. *Int. J. Syst. Evol. Microbiol.* 63, 1562–1567.

Antón, J., Peña, A., Santos, F., Martínez-García, M., Schmitt-Kopplin, P., and Rosselló-Móra, R. (2008). Distribution, abundance and diversity of the extremely halophilic bacterium *Salinibacter ruber*. *Saline Syst.* 4:15. doi: 10.1186/1746-1448-4-15

Antón, J., Rosselló-Mora, R., Rodríguez-Valera, F., and Amann, R. (2000). Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Appl. Environ. Microbiol.* 66, 3052–3057. doi: 10.1128/AEM.66.7.3052-3057.2000

Baykov, A. A., Malinen, A. M., Luoto, H. H., and Lahti, R. (2013). Pyrophosphate-fueled Na^+ and H^+ transport in prokaryotes. *Microbiol. Mol. Biol. Rev.* 77, 267–276. doi: 10.1128/mmbr.00003-13

Briolat, V., and Reysset, G. (2002). Identification of the *Clostridium perfringens* genes involved in the adaptive response to oxidative stress. *J. Bacteriol.* 184, 2333–2343. doi: 10.1128/JB.184.9.2333-2343.2002

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Castro-Silva, C., Ruíz-Valdiviezo, V. M., Valenzuela-Encinas, C., Alcántara-Hernández, R. J., Navarro-Noya, Y. E., Vázquez-Núñez, E., et al. (2013). The bacterial community structure in an alkaline saline soil spiked with anthracene. *Electron. J. Biotechnol.* 16, 10–10. doi: 10.2225/vol16-issue5-fulltext-14

Conrad, R., Erkel, C., and Liesack, W. (2006). Rice cluster I methanogens, an important group of *Archaea* producing greenhouse gas in soil. *Curr. Opin. Biotechnol.* 17, 262–267. doi: 10.1016/j.copbio.2006.04.002

Cregut, M., Durand, M.-J., and Thouand, G. (2014). The diversity and functions of choline sulphatases in microorganisms. *Microb. Ecol.* 67, 350–357. doi: 10.1007/s00248-013-0328-7

Csonka, L. N. (1989). Physiological and genetic responses of bacteria to osmotic stress. *Microbiol. Rev.* 53, 121–147.

Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. (2012). Functional metagenomics reveals novel salt tolerance loci from the human gut microbiome. *ISME J.* 6, 1916–1925. doi: 10.1038/ismej.2012.38

Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. (2013). Functional environmental screening of a metagenomic library identifies stlA; a unique salt tolerance locus from the human gut microbiome. *PLoS ONE* 8:e82985. doi: 10.1371/journal.pone.0082985

Dassarma, S., Kennedy, S., Berquist, B., Victor Ng, W., Baliga, N., Spudich, J., et al. (2001). Genomic perspective on the photobiology of *Halobacterium* species NRC-1, a phototrophic, phototactic, and UV-tolerant haloarchaeon. *Photosyn. Res.* 70, 3–17. doi: 10.1023/a:1013879706863

Delagoutte, E., and von Hippel, P. H. (2002). Helicase mechanisms and the coupling of helicases within macromolecular machines part I: structures and properties of isolated helicases. *Q. Rev. Biophys.* 35, 431–478. doi: 10.1017/S0033583502003852

Deole, R., Challacombe, J., Raiford, D. W., and Hoff, W. D. (2013). An extremely halophilic proteobacterium combines a highly acidic proteome with a low cytoplasmic potassium content. *J. Biol. Chem.* 288, 581–588. doi: 10.1074/jbc.M112.420505

Dizdaroglu, M. (2005). Base-excision repair of oxidative DNA damage by DNA glycosylases. *Mutat. Res.* 591, 45–59. doi: 10.1016/j.mrfmmm.2005.01.033

Dower, W. J., Miller, J. F., and Ragsdale, C. W. (1988). High efficiency transformation of *E. coli* by high voltage electroporation. *Nucleic Acids Res.* 16, 6127–6145. doi: 10.1093/nar/16.13.6127

Earl, A. M., Losick, R., and Kolter, R. (2008). Ecology and genomics of *Bacillus subtilis*. *Trends Microbiol.* 16, 269–275. doi: 10.1016/j.tim.2008.03.004

França, L., Lopéz-Lopéz, A., Rosselló-Móra, R., and Costa, M. S. (2014). Microbial diversity and dynamics of a groundwater and a still bottled natural mineral water. *Environ. Microbiol.* 17, 577–593. doi: 10.1111/1462-2920.12430

Galinski, E. A. (1995). Osmoadaptation in bacteria. *Adv. Microb. Physiol.* 37, 273–328. doi: 10.1016/S0065-2911(08)60148-4

González-Pastor, J. E., and Mirete, S. (2010). "Novel metal resistance genes from microorganisms: a functional metagenomic approach," in *Molecular Methods in Metagenomics*, eds R. Daniely and W. Streit (Verleg: Springer).

Guazzaroni, M.-E., Morgante, V., Mirete, S., and González-Pastor, J. E. (2013). Novel acid resistance genes from the metagenome of the Tinto River, an extremely acidic environment. *Environ. Microbiol.* 15, 1088–1102. doi: 10.1111/1462-2920.12021

Haardt, M., Kempf, B., Faatz, E., and Bremer, E. (1995). The osmoprotectant proline betaine is a major substrate for the binding-protein-dependent transport system ProU of *Escherichia coli* K-12. *Mol. Gen. Genet.* 246, 783–796. doi: 10.1007/BF00290728

Hammer, Ø., and Harper, D. A. T. (2008). *Paleontological Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

Harris, J. K., Caporaso, J. G., Walker, J. J., Spear, J. R., Gold, N. J., Robertson, C. E., et al. (2013). Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. *ISME J.* 7, 50–60. doi: 10.1038/ismej.2012.79

Humann, J. L., Ziemkiewicz, H. T., Yurgel, S. N., and Kahn, M. L. (2009). Regulatory and DNA repair genes contribute to the desiccation resistance of *Sinorhizobium meliloti* Rm1021. *Appl. Environ. Microbiol.* 75, 446–453. doi: 10.1128/aem.02207-08

Kaberdin, V. R., and Bläsi, U. (2013). Bacterial helicases in post-transcriptional control. *Biochim. Biophys. Acta* 1829, 878–883. doi: 10.1016/j.bbagrm.2012.12.005

Kapardar, R. K., Ranjan, R., Grover, A., Puri, M., and Sharma, R. (2010). Identification and characterization of genes conferring salt tolerance to *Escherichia coli* from pond water metagenome.

Bioresour. Technol. 101, 3917–3924. doi: 10.1016/j.biortech.2010.01.017

Kawaichi, S., Ito, N., Kamikawa, R., Sugawara, T., Yoshida, T., and Sako, Y. (2013). Ardenticatena maritima gen. nov., sp. nov., a ferric iron- and nitrate-reducing bacterium of the phylum 'Chloroflexi' isolated from an iron-rich coastal hydrothermal field, and description of Ardenticatenia classis nov. *Int. J. Syst. Evol. Microbiol.* 63, 2992–3002. doi: 10.1099/ijs.0.046532-0

Kish, A., and DiRuggiero, J. (2012). "DNA replication and repair in halophiles," in *Advances in Understanding the Biology of Halophilic Microorganisms*, ed. R. Vreeland (Amsterdam: Springer), 163–198.

Kish, A., Kirkali, G., Robinson, C., Rosenblatt, R., Jaruga, P., Dizdaroglu, M., et al. (2009). Salt shield: intracellular salts provide cellular protection against ionizing radiation in the halophilic archaeon. *Halobacterium salinarum* NRC-1. *Environ. Microbiol.* 11, 1066–1078.

Kunin, V., Raes, J., Harris, J. K., Spear, J. R., Walker, J. J., Ivanova, N., et al. (2008). Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. *Mol. Sys. Biol.* 4:198. doi: 10.1038/msb.2008.35

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U.S.A.* 82, 6955–6959. doi: 10.1073/pnas.82.20.6955

Liu, H. H., Liu, J., Fan, S. L., Song, M. Z., Han, X. L., Liu, F., et al. (2008). Molecular cloning and characterization of a salinity stress-induced gene encoding DEAD-box helicase from the halophyte *Apocynum venetum*. *J. Exp. Bot.* 59, 633–644. doi: 10.1093/jxb/erm355

López-López, A., Richter, M., Peña, A., Tamames, J., and Rosselló-Móra, R. (2013). New insights into the archaeal diversity of a hypersaline microbial mat obtained by a metagenomic approach. *Syst. Appl. Microbiol.* 36, 205–214. doi: 10.1016/j.syapm.2012.11.008

López-López, A., Yarza, P., Richter, M., Suárez-Suárez, A., Antón, J., Niemann, H., et al. (2010). Extremely halophilic microbial communities in anaerobic sediments from a solar saltern. *Environ. Microbiol. Rep.* 2, 258–271. doi: 10.1111/j.1758-2229.2009.00108.x

López-Pérez, M., and Mirete, S. (2014). Discovery of novel antibiotic resistance genes through metagenomics. *Recent Adv. DNA Gene Seq.* 8, 15–19. doi: 10.2174/2352092208666141013231244

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, et al. (2004). ARB: a software environment for sequence data. *Nucleic Acids Res.* 32, 1363–1371. doi: 10.1093/nar/gkh293

Luoto, H. H., Baykov, A. A., Lahti, R., and Malinen, A. M. (2013). Membrane-integral pyrophosphatase subfamily capable of translocating both Na^+ and H^+. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1255–1260. doi: 10.1073/pnas.12178 16110

Malinen, A. M., Belogurov, G. A., Baykov, A. A., and Lahti, R. (2007). Na^+-Pyrophosphatase: a novel primary sodium pump. *Biochemistry* 46, 8872–8878. doi: 10.1021/bi700564b

Mattimore, V., and Battista, J. R. (1996). Radioresistance of *Deinococcus radiodurans*: functions necessary to survive ionizing radiation are also necessary to survive prolonged desiccation. *J. Bacteriol.* 178, 633–637.

Mehrshad, M., Amoozegar, M. A., Didari, M., Bagheri, M., Fazeli, S. A. S., Schumann, P., et al. (2013). *Bacillus halosaccharovorans* sp. nov., a moderately halophilic bacterium from a hypersaline lake. *Int. J. Syst. Evol. Microbiol.* 63, 2776–2781. doi: 10.1099/ijs.0.046961-0

Mirete, S., De Figueras, C. G., and Gonzalez-Pastor, J. E. (2007). Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. *Appl. Environ. Microbiol.* 73, 6001–6011. doi: 10.1128/AEM.00048-07

Moreira, D., Rodríguez-Valera, F., and López-García, P. (2006). Metagenomic analysis of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group. *Microbiology* 152, 505–517. doi: 10.1099/mic.0.28254-0

Morgante, V., Mirete, S., González De Figueras, C., Postigo Cacho, M., and González-Pastor, J. E. (2014). Exploring the diversity of arsenic resistance genes from acid mine drainage microorganisms. *Environ. Microbiol.* 17, 1910–1925. doi: 10.1111/1462-2920.12505

Mukhopadhyay, A., He, Z., Alm, E. J., Arkin, A. P., Baidoo, E. E., Borglin, S. C., et al. (2006). Salt stress in *Desulfovibrio vulgaris* Hildenborough: an integrated genomics approach. *J. Bacteriol.* 188, 4068–4078. doi: 10.1128/JB.01921-05

Nakamura, T., Muramoto, Y., Yokota, S., Ueda, A., and Takabe, T. (2004). Structural and transcriptional characterization of a salt-responsive gene encoding putative ATP-dependent RNA helicase in barley. *Plant Sci.* 167, 63–70. doi: 10.1016/j.plantsci.2004.03.001

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., et al. (2012). De novo metagenomic assembly reveals abundant novel major lineage of *Archaea* in hypersaline microbial communities. *ISME J.* 6, 81–93. doi: 10.1038/ismej.2011.78

Oren, A. (2002). *Halophilic Microorganisms and their Environments*. Berlin: Kluwer Academic Publishers.

Oren, A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems* 4:2. doi: 10.1186/1746-1448-4-2

Oren, A. (2013). Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes. *Front. Microbiol.* 4:315. doi: 10.3389/fmicb.2013.00315

Østerås, M., Boncompagni, E., Vincent, N., Poggi, M.-C., and Le Rudulier, D. (1998). Presence of a gene encoding choline sulfatase in *Sinorhizobium meliloti* bet operon: choline-*O*-sulfate is metabolized into glycine betaine. *Proc. Natl. Acad. Sci. U.S.A.* 95, 11394–11399. doi: 10.1073/pnas.95.19.11394

Paul, S., Bag, S. K., Das, S., Harvill, E. T., and Dutta, C. (2008). Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol.* 9:R70. doi: 10.1186/gb-2008-9-4-r70

Philippot, L., Raaijmakers, J. M., Lemanceau, P., and Van Der Putten, W. (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* 11, 789–799. doi: 10.1038/nrmicro3109

Podell, S., Ugalde, J. A., Narasingarao, P., Banfield, J., Heidelberg, K. B., and Allen, E. E. (2013). Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLoS ONE* 8:e61692. doi: 10.1371/journal.pone.0061692

Pruesse, E., Peplies, J., and Glockner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi: 10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rhodes, M. E., Fitz-Gibbon, S. T., Oren, A., and House, C. H. (2010). Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea. *Environ. Microbiol.* 12, 2613–2623. doi: 10.1111/j.1462-2920.2010.02232.x

Rodriguez-Valera, F., Ventosa, A., Juez, G., and Imhoff, J. F. (1985). Variation of environmental features and microbial populations with salt concentrations in a multi-pond saltern. *Microb. Ecol.* 11, 107–115. doi: 10.1007/BF02010483

Sakai, S., Ehara, M., Tseng, I.-C., Yamaguchi, T., Bräuer, S. L., Cadillo-Quiroz, H., et al. (2012). *Methanolinea mesophila* sp. nov., a hydrogenotrophic methanogen isolated from rice field soil, and a proposal of the archaeal family Methanoregulaceae fam. nov. within the order Methanomicrobiales. *Int. J. Syst. Evol. Microbiol.* 62, 1389–1395. doi: 10.1099/ijs.0.035048-0

Sanan-Mishra, N., Pham, X. H., Sopory, S. K., and Tuteja, N. (2005). Pea DNA helicase 45 overexpression in tobacco confers high salinity tolerance without affecting yield. *Proc. Natl. Acad. Sci. U.S.A.* 102, 509–514. doi: 10.1073/pnas.0406485102

Savage, K. N., Krumholz, L. R., Oren, A., and Elshahed, M. S. (2007). *Haladaptatus paucihalophilus* gen. nov., sp., nov., a halophilic archaeon isolated from a low-salt sulfide-rich spring. *Int. J. Syst. Evol. Microbiol.* 57, 19–24. doi: 10.1099/ijs.0.64464-0

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Simon, C., and Daniel, R. (2009). Achievements and new knowledge unraveled by metagenomic approaches. *Appl. Microbiol. Biotechnol.* 85, 265–276. doi: 10.1007/s00253-009-2233-z

Sleator, R. D., and Hill, C. (2001). Bacterial osmoadaptation: the role of osmolytes in bacterial stress and virulence. *FEMS Microbiol. Rev.* 26, 49–71. doi: 10.1111/j.1574-6976.2002.tb00598.x

Solovyev, V., and Salamov, A. (2011). "Automatic annotation of microbial genomes and metagenomic sequences," in *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies*, ed. R. W. Li (New York, NY: Nova Science Publishers), 61–78.

Spizizen, J. (1958). Transformation of biochemically deficient strains of *Bacillus subtilis* by deoxyribonucleate. *Proc. Natl. Acad. Sci. U.S.A.* 44, 1072–1078. doi: 10.1073/pnas.44.10.1072

Tanner, N. K., and Linder, P. (2001). DExD/H box RNA helicases: from generic motors to specific dissociation functions. *Mol. Cell* 8, 251–262. doi: 10.1016/S1097-2765(01)00329-X

Tsai, J.-Y., Kellosalo, J., Sun, Y.-J., and Goldman, A. (2014). Proton/sodium pumping pyrophosphatases: the last of the primary ion pumps. *Curr. Opin. Struct. Biol.* 27, 38–47. doi: 10.1016/j.sbi.2014.03.007

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R. D., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and *Archaea*. *Nature* 462, 1056–1060. doi: 10.1038/nature08656

Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K., Glöckner, F. O., et al. (2010). Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Syst. Appl. Microbiol.* 33, 291–299. doi: 10.1016/j.syapm.2010.08.001

Yoon, H.-S., Kim, S.-Y., and Kim, I.-S. (2013). Stress response of plant H^+-PPase-expressing transgenic *Escherichia coli* and *Saccharomyces cerevisiae*: a potentially useful mechanism for the development of stress-tolerant organisms. *J. Appl. Gen.* 54, 129–133. doi: 10.1007/s13353-012-0117-x

Youngman, P., Perkins, J. B., and Losick, R. (1984). Construction of a cloning site near one end of Tn917 into which foreign DNA may be inserted without affecting transposition in *Bacillus subtilis* or expression of the transposon-borne erm gene. *Plasmid* 12, 1–9. doi: 10.1016/0147-619X(84)90061-1

Yu, E., and Owttrim, G. W. (2000). Characterization of the cold stress-induced cyanobacterial DEAD-box protein CrhC as an RNA helicase. *Nucleic Acids Res.* 28, 3926–3934. doi: 10.1093/nar/28.20.3926

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

PERSPECTIVE
published: 29 October 2015
doi: 10.3389/fmicb.2015.01196

Current and future resources for functional metagenomics

*Kathy N. Lam, Jiujun Cheng, Katja Engel, Josh D. Neufeld and Trevor C. Charles **

Department of Biology, University of Waterloo, Waterloo, ON, Canada

Functional metagenomics is a powerful experimental approach for studying gene function, starting from the extracted DNA of mixed microbial populations. A functional approach relies on the construction and screening of metagenomic libraries—physical libraries that contain DNA cloned from environmental metagenomes. The information obtained from functional metagenomics can help in future annotations of gene function and serve as a complement to sequence-based metagenomics. In this Perspective, we begin by summarizing the technical challenges of constructing metagenomic libraries and emphasize their value as resources. We then discuss libraries constructed using the popular cloning vector, pCC1FOS, and highlight the strengths and shortcomings of this system, alongside possible strategies to maximize existing pCC1FOS-based libraries by screening in diverse hosts. Finally, we discuss the known bias of libraries constructed from human gut and marine water samples, present results that suggest bias may also occur for soil libraries, and consider factors that bias metagenomic libraries in general. We anticipate that discussion of current resources and limitations will advance tools and technologies for functional metagenomics research.

Keywords: functional metagenomics, metagenomic library, cosmid library, fosmid library, pCC1FOS, cloning bias, library bias, RK2

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Kentaro Miyazaki, National Institute of Advanced Industrial Science and Technology, Japan
Alexander Wentzel, SINTEF Materials and Chemistry, Norway

***Correspondence:**
Trevor C. Charles
tcharles@uwaterloo.ca

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *12 August 2015*
Accepted: *14 October 2015*
Published: *29 October 2015*

Citation:
Lam KN, Cheng J, Engel K, Neufeld JD and Charles TC (2015) Current and future resources for functional metagenomics. Front. Microbiol. 6:1196. doi: 10.3389/fmicb.2015.01196

THE CHALLENGES OF CONSTRUCTING LARGE-INSERT METAGENOMIC LIBRARIES

Functional metagenomics involves isolating DNA from microbial communities to study the functions of encoded proteins. It involves cloning DNA fragments, expressing genes in a surrogate host, and screening for enzymatic activities. Using this function-based approach allows for discovery of novel enzymes whose functions would not be predicted based on DNA sequence alone. Information from function-based analyses can then be used to annotate genomes and metagenomes derived solely from sequence-based analyses. Thus, functional metagenomics complements sequence-based metagenomics, analogous to how molecular genetics of model organisms has provided knowledge of gene function that is widely applicable in genomics.

Functional metagenomics begins with the construction of a metagenomic library (**Figure 1A**). Cosmid- or fosmid-based libraries are often preferred due to their large and consistent insert size and high cloning efficiency. DNA is first extracted from the environmental sample of interest, then size-selected, end-repaired, and ligated to a *cos*-based vector, allowing packaging by lambda phage for subsequent transduction of *Escherichia coli* (**Figure 1A**). The resulting library contains relatively large insert DNA, typically 25–40 kb for *cos*-based vectors. With the steps involved, the construction of a metagenomic library can be laborious and time-consuming, requiring a high level of skill at the laboratory bench.

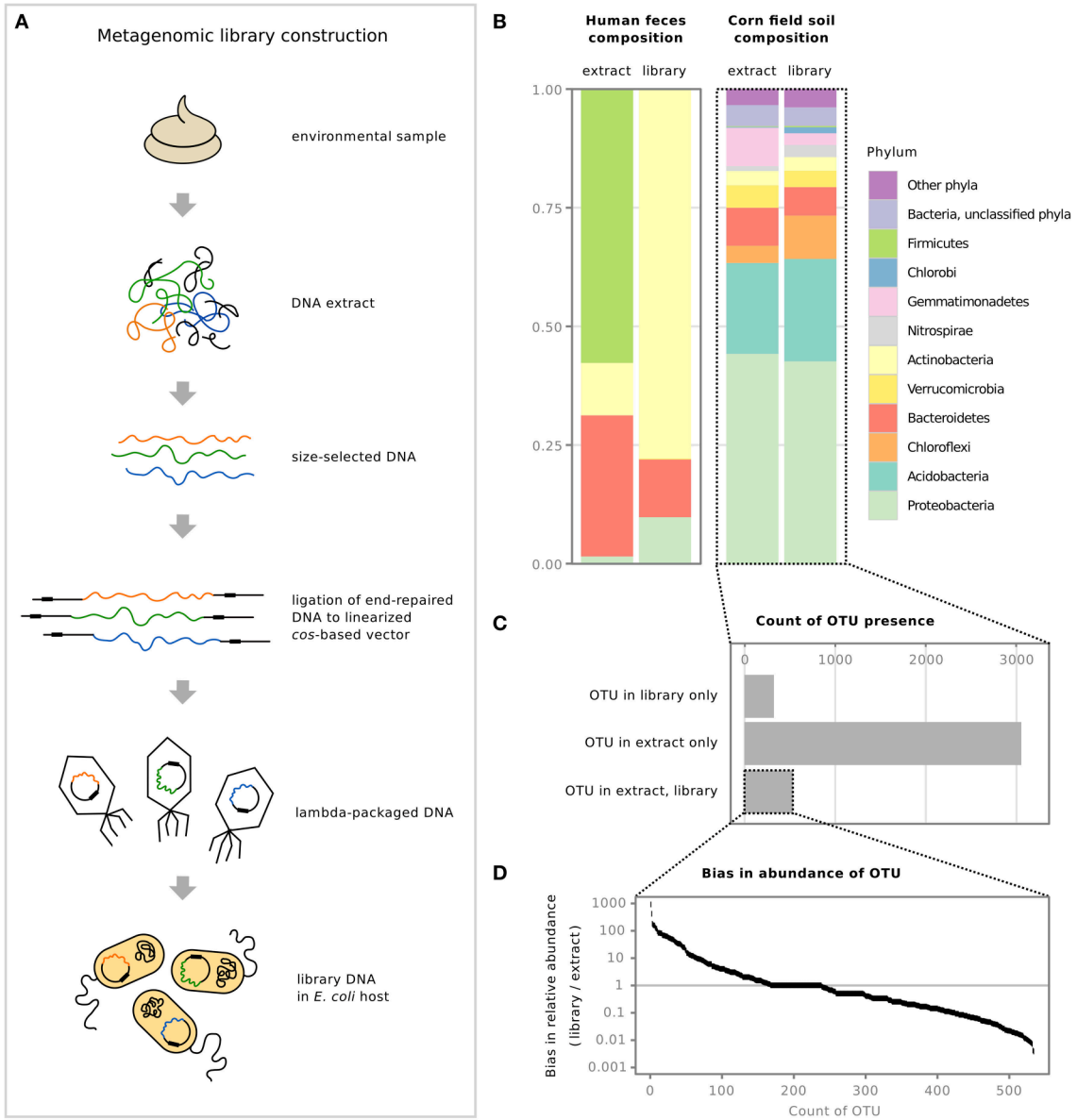

FIGURE 1 | Metagenomic libraries exhibit cloning bias when compared to the original environmental sample. (A) Steps involved in the construction of a metagenomic library, from original environmental sample to the final library in the *E. coli* host (adapted from Lam and Charles, 2015). **(B)** Relative abundance of bacterial phyla from two previously constructed metagenomic libraries, a human fecal library (Lam and Charles, 2015), and a corn field soil library (Cheng et al., 2014), compared to their original sample DNA extracts. **(C)** Number of OTUs identified from corn field soil DNA extract and library, and whether the OTUs were present in the library sample only, the extract sample only, or present in both. **(D)** Examination of cloning bias by comparing the relative abundance of OTUs that were present in both the DNA extract and the cosmid library, shown on a log scale; horizontal line at 1 denotes equal relative abundance in both samples.

There are several technically challenging steps in library construction. First, the extracted DNA must be of sufficient length for efficient packaging into lambda phage heads (Parks and Graham, 1997). Extraction usually employs gentle lysis to avoid shearing DNA (Zhou et al., 1996) but even so it may be difficult to achieve large fragment sizes (Kakirde et al., 2010). We find that starting with crude DNA extracts containing at least ~75 kb fragments leads to high-quality libraries and it is crucial to check the fragment size range by pulsed-field electrophoresis before proceeding. A particularly useful and affordable molecular ladder for pulsed-field gels is self-ligated lambda DNA, which can be easily prepared and results in bands at approximately 50, 100, and 150 kb. A freeze-grinding step prior to extraction (Lee and Hallam, 2009) can substantially improve cell lysis. Although this step may fragment DNA (Brady, 2007), we find it does not hinder library construction, consistent with previous work showing that freeze-grinding results in minimal shearing (Zhou et al., 1996).

Extracts are often contaminated with compounds that co-purify with DNA, requiring additional purification steps that may lead to sample loss. Common contaminants in soil-derived DNA extracts are humic acids, which may interfere with enzymatic reactions (Tebbe and Vahjen, 1993). Non-linear electrophoresis is effective for contaminant removal (Pel et al., 2009) and generates purified and concentrated DNA suitable for PCR or metagenomic analysis (Engel et al., 2012), yet requires specialized equipment. We have found that for library construction, humic acids can simply be allowed to run off the gel during pulsed-field electrophoresis of crude extract for size-selection because they migrate much faster than large DNA fragments. Alternatively, to avoid contaminating the circulating buffer, electrophoresis can be paused after humic acids have formed a front, the part of the gel containing the humic acids excised, and then this region replaced with fresh gel (Cheng et al., 2014). Others have reported that contaminating nucleases are effectively inhibited by treating extracted DNA in an agarose plug with sodium chloride and formamide (Liles et al., 2008).

After the DNA has been size-selected and purified, it must be end-repaired and ligated to a desphosphorylated, blunt-ended vector. To ensure proper size range before ligation, the DNA can be checked for co-migration with the largest band of a lambda-HindIII ladder on an agarose gel (Brady, 2007) or the sample can be run on a pulsed-field gel for a more accurate size assessment. The end-repair is a challenging step because there is no simple way to confirm that ends are indeed blunt following the reaction. We use a small amount of the ligation to transform *E. coli* prior to the costly packaging step; resulting transformants indicate the presence of circular DNA molecules arising from ligation of successfully blunt-ended fragments. Though the ligation conditions may not favor formation of circular molecules, this is our best proxy for successful end-repair.

Other challenges include the sensitivity of packaging extracts and preparation of purified digested and dephosphorylated vector DNA for ligation. Although excellent commercial products are available for both, in-house vector preparation may still be required when specific expression hosts are to be used in functional screening outside the host range of available commercial vectors (Wexler et al., 2005; Craig et al., 2010; Troeschel et al., 2010; Cheng et al., 2014). The culminating step of library construction is the transduction of *E. coli*, and although it is possible to generate many thousands of clones with the first attempt, troubleshooting may be required to increase library size. When transduction results in a disappointingly small number of transductants (zero in the worst case!), it is not easy to determine the cause.

Indeed, metagenomic library construction is in many ways an art that takes time and practice to master. Given the substantial challenges and costs associated with library construction, as well as possible difficulties in obtaining rare environmental samples, a clear corollary is that we ought to find ways to maximize these valuable resources for shared benefit. In particular, collections of metagenomic libraries that can be used in a variety of hosts would be extremely valuable if able to be accessed by the scientific community. We have previously made our libraries publicly available (Neufeld et al., 2011) and we continue to advocate for increased sharing (Charles and Neufeld, 2015). Though there are obvious administrative obstacles, services such as Addgene (Herscovitch et al., 2012) may facilitate these efforts.

MAKING THE MOST OF WHAT WE HAVE: LEVERAGING EXISTING LIBRARIES

Due to the difficulties of library construction, commercial products that aid in generation of libraries are popular. Indeed, one widely used cloning-ready commercial vector is pCC1FOS (Genbank accession EU140751; Epicentre Biotechnologies). In recent years, as functional metagenomics has gained traction, metagenomic libraries from remarkably diverse environments have been constructed using pCC1FOS (**Table 1**). The pCC1FOS vector has several advantages. It carries a chloramphenicol resistance (*cat*) marker that is superior to the common ampicillin resistance (*bla*) marker, obviating the occurrence of satellite colonies associated with beta-lactamase secretion that can be problematic for the dense platings often required for library construction. In addition to an F plasmid *oriV* for single-copy maintenance, pCC1FOS also carries an *oriV* from the RK2 plasmid. The RK2 *oriV* is broad-host-range, conferring replication ability in diverse members of the *Proteobacteria* (Ayres et al., 1993), but requires the *trfA* gene product for replication and results in an estimated 15 copies per cell (Durland and Helinski, 1990). Though *trfA* is not carried by the fosmid, it can be provided in trans; notably, the commercial *E. coli* strain EPI300 (Epicentre Biotechnologies) carries *trfA* under the control of an inducible promoter that is advertised to increase copy number from 1 copy per cell to 10–200 copies. The strain likely possesses a *trfA* copy-up mutant allele under control of *araC-*P_{BAD}, which is induced by L-arabinose (Wild et al., 2002). In the past, we preferred HB101 as a library host due to its receptiveness to transduction, but EPI300 appears to transduce at least as well as, if not better than, HB101. It also has the advantages of being an *endA1* mutant and supporting copy-number inducibility, allowing for less-degraded and higher-yield plasmid preparations.

Despite its popularity, pCC1FOS has some disadvantages that make resulting libraries less versatile than they could be. First, pCC1FOS does not possess an *oriT* that would allow the fosmid to be efficiently transferred by conjugation, mediated by a helper plasmid, to other species or strains that may be more suitable for heterologous expression. To achieve conjugation capabilities, we have added the RK2 *oriT* to pCC1FOS (Lam and Charles, unpublished), as have others (Aakvik et al., 2009; Buck, 2012; Terrón-González et al., 2013). To enable conjugation after library construction has already taken place, others have retrofitted individual pCC1FOS-based clones with an *oriT* (Li et al., 2011; Buck, 2012). These modifications illustrate the need for fosmid and cosmid vector design to include the *oriT* so that duplication of work can be avoided. It is possible that transformation can be used to transfer libraries to other hosts, but only for recipients that are amenable to those techniques and that will not reject DNA that has been synthesized in *E. coli* due to the presence of host restriction-modification systems. In some cases, it will be

TABLE 1 | Examples of metagenomic libraries constructed from diverse environmental samples using cloning vector pCC1FOS/pCC2FOS or derivatives.

Environment	Library vector; screening host, if relevant	References
HOST-ASSOCIATED ENVIRONMENTS		
Bovine rumen	pCC1FOS; *E. coli*	Wang et al., 2013
Elephant feces	pCC1FOS; *E. coli*	Rabausch et al., 2013
Human distal ileum	pCC1FOS; *E. coli*	Cecchini et al., 2013
Human feces	pCC1FOS; *E. coli*	Jones et al., 2008
Human feces (pescatarian)	pCC1FOS; *E. coli*	Tasse et al., 2010
Marine sponge	pCC1FOS	Yung et al., 2009
Termite gut	pCC1FOS, pCC2FOS; *E. coli*	Warnecke et al., 2007; Liu et al., 2011
EXTREME ENVIRONMENTS		
Alaskan soil	pCC1FOS; *E. coli*	Allen et al., 2009
Alaskan floodplain soil	pCC1FOS; *E. coli*	Williamson et al., 2005
Antarctic Pennisula meltwater	pCC1FOS; *E. coli*	Ferrés et al., 2015
Glacial ice	pCC1FOS; *E. coli*	Simon et al., 2009
Hot spring sediment and biofilm	pCT3FK; *E. coli, Thermus thermophilus*	Leis et al., 2015
Hydrothermal fluids	pCC1FOS; *E. coli*	Böhnke and Perner, 2015
MARINE OR FRESHWATER ENVIRONMENTS		
Bog	pCC1FOS; *E. coli*	Sommer et al., 2010
Marine sediment	pRS44; *Pseudomonas fluorescens, Xanthomonas campestris*	Aakvik et al., 2009
Ocean tidal flat sediment	pCC1FOS; *E. coli*	Lee et al., 2006, 2015
Ocean water column	pCC1FOS	DeLong et al., 2006
River sediment	pCC1FOS; *E. coli*	Rabausch et al., 2013
POLLUTED ENVIRONMENTS		
Crude oil-contaminated shore	pMPO579; *E. coli**	Terrón-González et al., 2013
Polluted river	pCC1FOS; *E. coli*	Vercammen et al., 2013
AGRICULTURAL, ENGINEERED, OR OTHER ENVIRONMENTS		
Activated sludge	pCC1FOS, pCC2FOS; *E. coli*	Suenaga et al., 2007; Zhang and Han, 2009
Compost, leaf branch	pCC1FOS; *E. coli*	Sulaiman et al., 2012
Compost, lumber waste	pCT3FK; *E. coli, Thermus thermophilus*	Leis et al., 2015
Compost, wood/plant debris/manure	pCC1FOS; *E. coli*	Ohlhoff et al., 2015
Decomposing leaf litter	pCC1FOS; *E. coli*	Nyyssönen et al., 2013
Orchard soil	pCC1FOS; *E. coli*	Donato et al., 2010
Sugarcane bagasse	pCC1FOS	Mhuantong et al., 2015

Libraries that are based on the commercial pCC1FOS or pCC2FOS vector can be screened in any RK2-compatible host that expresses the trfA gene product required for the broad-host-range RK2 oriV origin of replication.
**modified strains derived from E. coli EPI300 to increase transcription.*

desirable to modify these host strains by deleting the restriction-modification genes.

Given that the broad-host-range *oriV* is used to achieve a higher copy number in EPI300 expressing the *trfA* gene, another disadvantage of pCC1FOS is that *trfA* is not included on the vector. The consequence is that species that would otherwise be able to use the *oriV* cannot replicate pCC1FOS. It is not surprising then that for the vast majority of studies highlighted here (**Table 1**), *E. coli* was used as the screening host. This is a disadvantage for functional metagenomics as different clones can be isolated from the same metagenomic library when different screening hosts are used (Martinez et al., 2004; Craig et al., 2010). We found that using the legume-symbiont *Sinorhizobium meliloti* as a host results in a much greater diversity of clones than *E. coli* when screening our corn field soil metagenomic library for beta-galactosidase activity, though this greater diversity does not appear to be related to phylogenetic distance of the origin of the cloned DNA to the surrogate host (Cheng et al., in preparation). The importance of devising systems that allow for functional screening in diverse expression hosts has been reviewed by others (Uchiyama and Miyazaki, 2009; Taupp et al., 2011; Ekkers et al., 2012; Liebl et al., 2014), but what of the large number of libraries that have already been constructed? Can we make use of them for screening in non-*E. coli* hosts? The libraries listed in **Table 1**, as well as potentially many other metagenomic libraries constructed using pCC1FOS or derivatives, would be accessible to any RK2-compatible host if a copy of the *trfA* gene were also made available. This solution has already been applied: one group inserted the *trfA* gene into the chromosome of the *Gammaproteobacteria* species *Pseudomonas fluorescens* and *Xanthomonas campestris* for screening of libraries constructed using a pCC1FOS derivative (Aakvik et al., 2009). Another group

inserted *araC-P_{BAD}-trfA* into the *E. coli* EL350 chromosome to give copy number inducibility to the lambda Red recombineering strain (Westenberg et al., 2010). The introduction of *trfA* into RK2-compatible species is a straightforward way to expand the range of expression hosts for existing pCC1FOS-based libraries.

An alternative to inserting the *trfA* gene into desired expression hosts is to modify the vector for integration into the host genome, bypassing the requirement for *trfA*. This strategy has been employed to integrate clones into a target locus in the genome of the thermophile *Thermus thermophilus* for functional screening, by modifying pCC1FOS to include a selectable marker as well as regions for homologous recombination (Angelov et al., 2009). In our lab, pCC1FOS was modified to carry ΦC31 *att* sites (Heil and Charles, unpublished) for integrase-mediated site-specific recombination of cloned insert DNA into the genomes of landing pad strains, including *S. meliloti* and *Agrobacterium tumefaciens* (Heil et al., 2012). As a general strategy, however, chromosomal integration is potentially less useful than clone maintenance due to the difficulty in retrieving the integrated DNA for manipulation, including DNA sequence analysis, when non-arrayed (i.e., pooled) libraries have been screened.

KNOWING THE EXTENT OF WHAT WE HAVE: EXAMINING CLONING BIAS

Beyond the practical questions of how to optimize vectors for library construction and how to maximize valuable existing libraries, there is a technical question that we find particularly interesting: how much of the sequence diversity present in original DNA extracts is captured in constructed libraries, and what affects this? Though not so much a concern for functional screens, it is interesting to consider the factors that influence library representativeness; elucidating these factors may lead to development of better strategies for accessing the full potential of environmental metagenomes. We previously used shotgun sequencing to examine bias in a human fecal library (Lam and Charles, 2015) and here we also present the results of 16S rRNA gene sequencing to examine bias in a corn field soil library (Cheng et al., 2014); see Supplementary Material for details. Both libraries were constructed using the RK2-based cosmid pJC8 (Genbank accession KC149513).

The bias discussed here is from comparing DNA extracted from the sample to the final cloned library DNA isolated from *E. coli* (**Figure 1A**). Analysis at the phylum-level showed that although the fecal library differed substantially in the relative abundance of phyla compared to its corresponding extract, the relative abundance of phyla in the corn field soil library seemed similar to its extract (**Figure 1B**). We present these results for the soil library but exercise caution in their interpretation as the majority of 16S rRNA gene sequences from the metagenomic library sample was *E. coli* contamination, despite treating the library cosmid DNA preparation with Plasmid-Safe DNase to remove host genomic DNA prior to PCR. After subtracting *E. coli* host sequences, approximately 30,000 sequences remained to represent the metagenomic library (see Supplementary Material for details). The high level of host contamination could be due to preferential amplification of template during PCR based on differences in DNA conformation: though present in very small quantities, linear DNA may be more efficiently amplified over supercoiled or closed circular plasmid DNA (Chen et al., 2007). This issue of *E. coli* host contamination in 16S rRNA gene analysis needs to be addressed for future examination of bias in metagenomic libraries.

When we examined the soil samples more closely, we found that the similarity of the library and extract at the phylum level does not extend to the "species" level: examination of the individual OTUs in each sample revealed that only a small fraction of OTUs were shared between the library and original sample (**Figure 1C**). Interestingly, our analysis indicated that there were a number of OTUs in the library that were not identified in the extract sample (**Figure 1C**) and although this number is halved when the library data are compared to extract data that have not been rarefied (data not shown), they nevertheless remain, indicating that these OTUs are either extremely rare in the original sample and their DNA is preferentially cloned or that the identification of these OTUs is due to sequencing errors. A further analysis of the OTU fraction that is shared between extract and library samples shows a large range in the bias in relative abundance of each OTU, with some OTUs exhibiting ∼1000-fold overrepresentation and others ∼1000-fold underrepresentation in the library (**Figure 1D**). While there may be concern that 16S rRNA gene profiles of libraries compared to extracts may not provide an accurate comparison of cloned DNA content in general, we have previously shown from analysis of shotgun sequence data that for large-insert RK2 *oriV*-based cosmid libraries, 16 S rRNA gene content tracks well with genomic content (Lam and Charles, 2015). The analysis of the corn field DNA extract and corresponding metagenomic library suggests that though the overall relative abundance of phyla may remain similar, bias is occurring on the level of individual OTUs.

The fact that certain taxa are under- or overrepresented might not pose a barrier to screening, but it may be useful to know what sequences are not likely to be captured in libraries. Several studies that have compared shotgun sequencing of original samples to corresponding metagenomic libraries from marine water (Temperton et al., 2009; Ghai et al., 2010; Danhorn et al., 2012), as well as our own comparative work on feces (Lam and Charles, 2015), have shown that AT-rich sequences are underrepresented in libraries. Our analysis—in which we compared promoter consensus sequences between extract and library samples—lends support to the hypothesis that the bias is related to spurious transcription of metagenomic DNA from AT-rich sequences recognized as σ^{70} promoters in the *E. coli* library host (Lam and Charles, 2015) although other factors may be contributing, such as gene product toxicity (Sorek et al., 2007). Notably, we have shown that DNA fragmentation is not a cause of bias (Lam and Charles, 2015). The specific factors affecting the "clonability" of DNA, and the mechanisms that lead to DNA exclusion, still need to be experimentally determined.

The stability of foreign DNA in *E. coli* is influenced by the vector copy number and, as a result, single-copy fosmids may

be ideal as the library backbone (Kim et al., 1992), although the success of some functional screens may be dependent on a higher gene dose. Plasmid vectors that are not *cos*-based provide an alternative where cloning is substantially less difficult as large-fragment DNA need not be isolated and packaging and transduction are not required; the disadvantages, however, are that a smaller insert size means that larger operons will not be intact, and if the plasmid has a high copy number—true of conventional cloning vectors—this may lead to greater insert instability and exclusion (Lam and Charles, 2015). Other alternatives to fosmid vectors include BACs (Kakirde et al., 2011), which have the ability to capture even larger insert sizes at approximately 100 kb on average (Kakirde et al., 2010), and linear vectors, which may provide exceptional stability (Godiska et al., 2010). However, *cos*-based vectors are likely to remain popular for their advantages: the availability of high-quality commercial packaging extracts, greater efficiency of transduction over transformation, and decreased probability of insert concatemers due to the phage head upper size limit. Though there exists variety in library cloning vectors, further work is required to understand how and to what extent cloning vector choice and strategy impacts library sequence bias.

CONCLUDING REMARKS

Depending on the target activity, functional screens can exhibit a low hit rate (Uchiyama and Miyazaki, 2009) the reasons for which might include barriers at the level of both transcription and translation. Improving *E. coli* as a screening host to address these problems will likely improve future hit rates. Examples include introducing heterologous sigma factors to guide RNA polymerase to otherwise untranscribed regions (Gaida et al., 2015), employing T7 RNA polymerase to help drive transcription (Terrón-González et al., 2013), as well as forming hybrid ribosomes (Kitahara et al., 2012) that may influence expression. Nevertheless, it will be important to move beyond *E. coli* into different screening hosts, particularly for the complementation of mutant phenotypes not possible in *E. coli*. The identification of obstacles to cloning and screening will aid in the development of new tools and technologies for functional metagenomics (Engel et al., 2013), providing us with greater reach in terms of what we are able to gather from functional screens. The refinement of methods will be crucial in bioprospecting for novel enzymes and compounds as well as for the determination of gene function that will guide the development of reliable models of microbial ecosystem functioning.

AUTHOR CONTRIBUTIONS

KL and TC conceived the ideas. JC prepared DNA from the soil-related samples. KE carried out V3 region PCR on the soil-related samples and managed sequencing sample submission. KL analyzed the sequence data, made the figures, performed the literature review, and wrote the paper. TC, JN, JC, and KE revised the manuscript. TC and JN provided reagents and materials. All authors read and approved the manuscript.

FUNDING

Research funding was provided by a Strategic Projects Grant (381646–09) from the Natural Sciences and Engineering Research Council of Canada, by Genome Canada for the project "Microbial Genomics for Biofuels and Co-Products from Biorefining Processes," and by a University of Waterloo CIHR Research Incentive Fund. KL was supported by a CGS-D scholarship from the Canadian Institutes of Health Research.

ACKNOWLEDGMENTS

We are grateful to Brent Seuradge for advice on the AXIOME2 pipeline, Michael J. Lynch for help with 16S rRNA gene analysis, and Michael W. Hall for assistance in AXIOME2 and BIOM-related issues. We acknowledge funding from NSERC (Strategic Projects Grant), Genome Canada and Genome Prairie, and the McMaster-Waterloo Bioinformatics Initiative. KL was supported by a CIHR CGS-D.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01196

REFERENCES

Aakvik, T., Degnes, K. F., Dahlsrud, R., Schmidt, F., Dam, R., Yu, L., et al. (2009). A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species. *FEMS Microbiol. Lett.* 296, 149–158. doi: 10.1111/j.1574-6968.2009.01639.x

Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A., and Handelsman, J. (2009). Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J.* 3, 243–251. doi: 10.1038/ismej.2008.86

Angelov, A., Mientus, M., Liebl, S., and Liebl, W. (2009). A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles. *Syst. Appl. Microbiol.* 32, 177–185. doi: 10.1016/j.syapm.2008.01.003

Ayres, E. K., Thomson, V. J., Merino, G., Balderes, D., and Figurski, D. H. (1993). Precise deletions in large bacterial genomes by vector-mediated excision (VEX): the *trfA* gene of promiscuous plasmid RK2 is essential for replication in several Gram-negative hosts. *J. Mol. Biol.* 230, 174–185. doi: 10.1006/jmbi.1993.1134

Böhnke, S., and Perner, M. (2015). A function-based screen for seeking RubisCO active clones from metagenomes: novel enzymes influencing RubisCO activity. *ISME J.* 9, 735–745. doi: 10.1038/ismej.2014.163

Brady, S. F. (2007). Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* 2, 1297–1305. doi: 10.1038/nprot.2007.195

Buck, J. D. (2012). *Physiological Effects of Heterologous Expression of Proteorhodopsin Photosystems*. Available online at: http://hdl.handle.net/1721.1/71464

Cecchini, D. A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J., et al. (2013). Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. *PLoS ONE* 8:e72766. doi: 10.1371/journal.pone.0072766

Charles, T. C., and Neufeld, J. D. (2015). "Open resource metagenomics," in *Encyclopedia of Metagenomics*, ed K. E. Nelson (New York, NY: Springer), 573–575.

Chen, J., Kadlubar, F. F., and Chen, J. Z. (2007). DNA supercoiling suppresses real-time PCR: a new approach to the quantification of mitochondrial DNA damage and repair. *Nucleic Acids Res.* 35, 1377–1388. doi: 10.1093/nar/gkm010

Cheng, J., Pinnell, L., Engel, K., Neufeld, J. D., and Charles, T. C. (2014). Versatile broad-host-range cosmids for construction of high quality metagenomic libraries. *J. Microbiol. Methods* 99, 27–34. doi: 10.1016/j.mimet.2014.01.015

Craig, J. W., Chang, F.-Y., Kim, J. H., Obiajulu, S. C., and Brady, S. F. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse *Proteobacteria*. *Appl. Environ. Microbiol.* 76, 1633–1641. doi: 10.1128/AEM.02169-09

Danhorn, T., Young, C. R., and DeLong, E. F. (2012). Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis. *ISME J.* 6, 2056–2066. doi: 10.1038/ismej.2012.35

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496–503. doi: 10.1126/science.1120250

Donato, J. J., Moe, L. A., Converse, B. J., Smart, K. D., Berklein, F. C., McManus, P. S., et al. (2010). Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins. *Appl. Environ. Microbiol.* 76, 4396–4401. doi: 10.1128/AEM.01763-09

Durland, R. H., and Helinski, D. R. (1990). Replication of the broad-host-range plasmid RK2: direct measurement of intracellular concentrations of the essential TrfA replication proteins and their effect on plasmid copy number. *J. Bacteriol.* 172, 3849–3858.

Ekkers, D. M., Cretoiu, M. S., Kielak, A. M., and Elsas, J. D. (2012). The great screen anomaly—a new frontier in product discovery through functional metagenomics. *Appl. Microbiol. Biotechnol.* 93, 1005–1020. doi: 10.1007/s00253-011-3804-3

Engel, K., Ashby, D., Brady, S. F., Cowan, D. A., Doemer, J., Edwards, E. A., et al. (2013). Meeting report: 1st international functional metagenomics workshop May 7-8, 2012, St. Jacobs, Ontario, Canada. *Stand. Genomic Sci.* 8, 106–111. doi: 10.4056/sigs.3406845

Engel, K., Pinnell, L., Cheng, J., Charles, T. C., and Neufeld, J. D. (2012). Nonlinear electrophoresis for purification of soil DNA for metagenomics. *J. Microbiol. Methods* 88, 35–40. doi: 10.1016/j.mimet.2011.10.007

Ferrés, I., Amarelle, V., Noya, F., and Fabiano, E. (2015). Construction and screening of a functional metagenomic library to identify novel enzymes produced by Antarctic bacteria. *Adv. Polar Sci.* 26, 96–101. doi: 10.13679/j.advps.2015.1.00096

Gaida, S. M., Sandoval, N. R., Nicolaou, S. A., Chen, Y., Venkataramanan, K. P., and Papoutsakis, E. T. (2015). Expression of heterologous sigma factors enables functional screening of metagenomic and heterologous genomic libraries. *Nat. Commun.* 6:7045. doi: 10.1038/ncomms8045

Ghai, R., Martin-Cuadrado, A.-B., Molto, A. G., Heredia, I. G., Cabrera, R., Martin, J., et al. (2010). Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* 4, 1154–1166. doi: 10.1038/ismej.2010.44

Godiska, R., Mead, D., Dhodda, V., Wu, C., Hochstein, R., Karsi, A., et al. (2010). Linear plasmid vector for cloning of repetitive or unstable sequences in *Escherichia coli*. *Nucleic Acids Res.* 38:e88. doi: 10.1093/nar/gkp1181

Heil, J. R., Cheng, J., and Charles, T. C. (2012). Site-specific bacterial chromosome engineering: ΦC31 integrase mediated cassette exchange (IMCE). *J. Vis. Exp.* e3698. doi: 10.3791/3698. Available online at: http://www.jove.com/video/3698/site-specific-bacterial-chromosome-engineering-c31-integrase-mediated

Herscovitch, M., Perkins, E., Baltus, A., and Fan, M. (2012). Addgene provides an open forum for plasmid sharing. *Nat. Biotechnol.* 30, 316–317. doi: 10.1038/nbt.2177

Jones, B. V., Begley, M., Hill, C., Gahan, C. G. M., and Marchesi, J. R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13580–13585. doi: 10.1073/pnas.0804437105

Kakirde, K. S., Parsley, L. C., and Liles, M. R. (2010). Size does matter: application-driven approaches for soil metagenomics. *Soil Biol. Biochem.* 42, 1911–1923. doi: 10.1016/j.soilbio.2010.07.021

Kakirde, K. S., Wild, J., Godiska, R., Mead, D. A., Wiggins, A. G., Goodman, R. M., et al. (2011). Gram negative shuttle BAC vector for heterologous expression of metagenomic libraries. *Gene* 475, 57–62. doi: 10.1016/j.gene.2010.11.004

Kim, U.-J., Shizuya, H., de Jong, P. J., Birren, B., and Simon, M. I. (1992). Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res.* 20, 1083–1085. doi: 10.1093/nar/20.5.1083

Kitahara, K., Yasutake, Y., and Miyazaki, K. (2012). Mutational robustness of 16S ribosomal RNA, shown by experimental horizontal gene transfer in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19220–19225. doi: 10.1073/pnas.1213609109

Lam, K. N., and Charles, T. C. (2015). Strong spurious transcription likely contributes to DNA insert bias in typical metagenomic clone libraries. *Microbiome* 3:22. doi: 10.1186/s40168-015-0086-5

Lee, D.-H., Choi, S.-L., Rha, E., Kim, S. J., Yeom, S.-J., Moon, J.-H., et al. (2015). A novel psychrophilic alkaline phosphatase from the metagenome of tidal flat sediments. *BMC Biotechnol.* 15:1. doi: 10.1186/s12896-015-0115-2

Lee, M. H., Lee, C. H., Oh, T. K., Song, J. K., and Yoon, J. H. (2006). Isolation and characterization of a novel lipase from a metagenomic library of tidal flat sediments: evidence for a new family of bacterial lipases. *Appl. Environ. Microbiol.* 72, 7406–7409. doi: 10.1128/AEM.01157-06

Lee, S., and Hallam, S. J. (2009). Extraction of high molecular weight genomic DNA from soils and sediments. *J. Vis. Exp.* 33:e1569. doi: 10.3791/1569

Leis, B., Angelov, A., Mientus, M., Li, H., Pham, V. T. T., Lauinger, B., et al. (2015). Identification of novel esterase-active enzymes from hot environments by use of the host bacterium *Thermus thermophilus*. *Front. Microbiol.* 6:275. doi: 10.3389/fmicb.2015.00275

Li, C., Zhang, F., and Kelly, W. L. (2011). Heterologous production of thiostrepton A and biosynthetic engineering of thiostrepton analogs. *Mol. Biosyst.* 7, 82–90. doi: 10.1039/C0MB00129E

Liebl, W., Angelov, A., Juergensen, J., Chow, J., Loeschcke, A., Drepper, T., et al. (2014). Alternative hosts for functional (meta)genome analysis. *Appl. Microbiol. Biotechnol.* 98, 8099–8109. doi: 10.1007/s00253-014-5961-7

Liles, M. R., Williamson, L. L., Rodbumrer, J., Torsvik, V., Goodman, R. M., and Handelsman, J. (2008). Recovery, purification, and cloning of high-molecular-weight DNA from soil microorganisms. *Appl. Environ. Microbiol.* 74, 3302–3305. doi: 10.1128/AEM.02630-07

Liu, N., Yan, X., Zhang, M., Xie, L., Wang, Q., Huang, Y., et al. (2011). Microbiome of fungus-growing termites: a new reservoir for lignocellulase genes. *Appl. Environ. Microbiol.* 77, 48–56. doi: 10.1128/AEM.01521-10

Martinez, A., Kolvek, S. J., Yip, C. L. T., Hopke, J., Brown, K. A., MacNeil, I. A., et al. (2004). Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. *Appl. Environ. Microbiol.* 70, 2452–2463. doi: 10.1128/AEM.70.4.2452-2463.2004

Mhuantong, W., Charoensawan, V., Kanokratana, P., Tangphatsornruang, S., and Champreda, V. (2015). Comparative analysis of sugarcane bagasse metagenome reveals unique and conserved biomass-degrading enzymes among lignocellulolytic microbial communities. *Biotechnol. Biofuels* 8:16. doi: 10.1186/s13068-015-0200-8

Neufeld, J. D., Engel, K., Cheng, J., Moreno-Hagelsieb, G., Rose, D. R., and Charles, T. C. (2011). Open resource metagenomics: a model for sharing metagenomic libraries. *Stand. Genomic Sci.* 5, 203–210. doi: 10.4056/sigs.1974654

Nyyssönen, M., Tran, H. M., Karaoz, U., Weihe, C., Hadi, M. Z., Martiny, J. B. H., et al. (2013). Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Front. Microbiol.* 4:282. doi: 10.3389/fmicb.2013.00282

Ohlhoff, C. W., Kirby, B. M., Van Zyl, L., Mutepfa, D. L. R., Casanueva, A., Huddy, R. J., et al. (2015). An unusual feruloyl esterase belonging to family VIII esterases and displaying a broad substrate range. *J. Mol. Catal. B Enzym.* 118, 79–88. doi: 10.1016/j.molcatb.2015.04.010

Parks, R. J., and Graham, F. L. (1997). A helper-dependent system for adenovirus vector production helps define a lower limit for efficient DNA packaging. *J. Virol.* 71, 3293–3298.

Pel, J., Broemeling, D., Mai, L., Poon, H.-L., Tropini, G., Warren, R. L., et al. (2009). Nonlinear electrophoretic response yields a unique parameter for separation of biomolecules. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14796–14801. doi: 10.1073/pnas.0907402106

Rabausch, U., Juergensen, J., Ilmberger, N., Böhnke, S., Fischer, S., Schubach, B., et al. (2013). Functional screening of metagenome and genome libraries for detection of novel flavonoid-modifying enzymes. *Appl. Environ. Microbiol.* 79, 4551–4563. doi: 10.1128/AEM.01077-13

Simon, C., Herath, J., Rockstroh, S., and Daniel, R. (2009). Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl. Environ. Microbiol.* 75, 2964–2968. doi: 10.1128/AEM.02644-08

Sommer, M. O. A., Church, G. M., and Dantas, G. (2010). A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. *Mol. Syst. Biol.* 6:360. doi: 10.1038/msb.2010.16

Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., and Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318, 1449–1452. doi: 10.1126/science.1147112

Suenaga, H., Ohnuki, T., and Miyazaki, K. (2007). Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environ. Microbiol.* 9, 2289–2297. doi: 10.1111/j.1462-2920.2007.01342.x

Sulaiman, S., Yamato, S., Kanaya, E., Kim, J.-J., Koga, Y., Takano, K., et al. (2012). Isolation of a novel cutinase homolog with polyethylene terephthalate-degrading activity from leaf-branch compost by using a metagenomic approach. *Appl. Environ. Microbiol.* 78, 1556–1562. doi: 10.1128/AEM.06725-11

Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., et al. (2010). Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.* 11, 1605–1612. doi: 10.1101/gr.108332.110

Taupp, M., Mewis, K., and Hallam, S. J. (2011). The art and design of functional metagenomic screens. *Curr. Opin. Biotechnol.* 22, 1–8. doi: 10.1016/j.copbio.2011.02.010

Tebbe, C. C., and Vahjen, W. (1993). Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl. Environ. Microbiol.* 59, 2657–2665.

Temperton, B., Field, D., Oliver, A., Tiwari, B., Mühling, M., Joint, I., et al. (2009). Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *ISME J.* 3, 792–796. doi: 10.1038/ismej.2009.32

Terrón-González, L., Medina, C., Limón-Mortés, M. C., and Santero, E. (2013). Heterologous viral expression systems in fosmid vectors increase the functional analysis potential of metagenomic libraries. *Sci. Rep.* 3:1107. doi: 10.1038/srep01107

Troeschel, S. C., Drepper, T., Leggewie, C., Streit, W. R., and Jaeger, K.-E. (2010). "Novel tools for the functional expression of metagenomic DNA," in *Metagenomics: Methods and Protocols Methods in Molecular Biology*, eds W. R. Streit and R. Daniel (New York, NY: Humana Press), 117–139. doi: 10.1007/978-1-60761-823-2_8

Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* 20, 616–622. doi: 10.1016/j.copbio.2009.09.010

Vercammen, K., Garcia-Armisen, T., Goeders, N., Van Melderen, L., Bodilis, J., and Cornelis, P. (2013). Identification of a metagenomic gene cluster containing a new class A beta-lactamase and toxin-antitoxin systems. *Microbiologyopen* 2, 674–683. doi: 10.1002/mbo3.104

Wang, L., Hatem, A., Catalyurek, U. V., Morrison, M., and Yu, Z. (2013). Metagenomic insights into the carbohydrate-active enzymes carried by the microorganisms adhering to solid digesta in the rumen of cows. *PLoS ONE* 8:e78507. doi: 10.1371/journal.pone.0078507

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565. doi: 10.1038/nature06269

Westenberg, M., Bamps, S., Soedling, H., Hope, I. A., and Dolphin, C. T. (2010). *Escherichia coli* MW005: lambda Red-mediated recombineering and copy-number induction of oriV-equipped constructs in a single host. *BMC Biotechnol.* 10:27. doi: 10.1186/1472-6750-10-27

Wexler, M., Bond, P. L., Richardson, D. J., and Johnston, A. W. B. (2005). A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environ. Microbiol.* 7, 1917–1926. doi: 10.1111/j.1462-2920.2005.00854.x

Wild, J., Hradecna, Z., and Szygbalski, W. (2002). Conditionally amplifiable BACs: switching from single-copy to high-copy vectors and genomic clones. *Genome Res.* 12, 1434–1444. doi: 10.1101/gr.130502

Williamson, L. L., Borlee, B. R., Schloss, P. D., Guan, C., Allen, H. K., and Handelsman, J. (2005). Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor. *Appl. Environ. Microbiol.* 71, 6335–6344. doi: 10.1128/AEM.71.10.6335-6344.2005

Yung, P. Y., Burke, C., Lewis, M., Egan, S., Kjelleberg, S., and Thomas, T. (2009). Phylogenetic screening of a bacterial, metagenomic library using homing endonuclease restriction and marker insertion. *Nucleic Acids Res.* 37:e144. doi: 10.1093/nar/gkp746

Zhang, T.,and Han, W.-J. (2009). Gene cloning and characterization of a novel esterase from activated sludge metagenome. *Microb. Cell Fact.* 8:67. doi: 10.1186/1475-2859-8-67

Zhou, J., Bruns, M. A., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* 62, 316–322.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 05 June 2015
doi: 10.3389/fmicb.2015.00563

Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology

*Lisa Ufarté [1,2,3], Gabrielle Potocki-Veronese [1,2,3] and Élisabeth Laville [1,2,3]**

[1] Université de Toulouse, Institut National des Sciences Appliquées (INSA), Université Paul Sabatier (UPS), Institut National Polytechnique (INP), Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés (LISBP), Toulouse, France, [2] INRA - UMR792 Ingénierie des Systèmes Biologiques et des Procédés, Toulouse, France, [3] CNRS, UMR5504, Toulouse, France

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Marc Strous, University of Calgary, Canada
Lukasz Jaroszewski, Sanford-Burnham Institute for Medical Research, USA

***Correspondence:**
Élisabeth Laville, Equipe de Catalyse et Ingénierie Moléculaire Enzymatiques, Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés, INSA - UMR INRA 792 - UMR CNRS 5504, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France
laville@insa-toulouse.fr

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *17 April 2015*
Accepted: *21 May 2015*
Published: *05 June 2015*

Citation:
Ufarté L, Potocki-Veronese G and Laville É (2015) Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. Front. Microbiol. 6:563. doi: 10.3389/fmicb.2015.00563

The rapid expansion of new sequencing technologies has enabled large-scale functional exploration of numerous microbial ecosystems, by establishing catalogs of functional genes and by comparing their prevalence in various microbiota. However, sequence similarity does not necessarily reflect functional conservation, since just a few modifications in a gene sequence can have a strong impact on the activity and the specificity of the corresponding enzyme or the recognition for a sensor. Similarly, some microorganisms harbor certain identified functions yet do not have the expected related genes in their genome. Finally, there are simply too many protein families whose function is not yet known, even though they are highly abundant in certain ecosystems. In this context, the discovery of new protein functions, using either sequence-based or activity-based approaches, is of crucial importance for the discovery of new enzymes and for improving the quality of annotation in public databases. This paper lists and explores the latest advances in this field, along with the challenges to be addressed, particularly where microfluidic technologies are concerned.

Keywords: metagenomics, discovery of new functions, proteins, high throughput screening, microbial ecosystems, microbial ecology, biotechnologies

Introduction

The implications of the discovery of new protein functions are numerous, from both cognitive and applicative points of view. Firstly, it improves understanding of how microbial ecosystems function, in order to identify biomarkers and levers that will help optimize the services rendered, regardless of the field of application. Next, the discovery of new enzymes and transporters enables expansion of the catalog of functions available for metabolic pathway engineering and synthetic biology. Finally, the identification and characterization of new protein families, whose functions, three-dimensional structure and catalytic mechanism have never been described, furthers understanding of the protein structure/function relationship. This is an essential prerequisite if we are to draw full benefit from these proteins, both for medical applications (for example, designing specific inhibitors) and for relevant integration into biotechnological processes.

Many reviews have been published on functional metagenomics these last 10 years. Many of them focus on the strategies of library creation and on bio-informatic developments (Di Bella et al., 2013;

Ladoukakis et al., 2014), while others describe the various approaches set up to discover novel targets [like therapeutic molecules (Culligan et al., 2014)] for a specific application. In particular several review papers have been written on the numerous activity-based metagenomics studies carried out to find new enzymes for biotechnological applications, without necessarily finding new functions or new protein families (Ferrer et al., 2009; Steele et al., 2009). The present review focuses on all the functional metagenomics approaches, sequence- or activity-based, allowing the discovery of new functions and families from the uncultured fraction of microbial ecosystems, and makes a recent overview on the advances of microfluidics for ultra-fast microbial screening of metagenomes.

Sampling Strategies

The literature describes a wide variety of microbial environments sampled in the search for new enzymes. A large number of studies look at ecosystems with high taxonomic and functional diversity, such as soils or natural aquatic environments that are either undisturbed or exposed to various pollutants (Gilbert et al., 2008; Brennerova et al., 2009; Zanaroli et al., 2010). Extreme environments enable the discovery of enzymes that are naturally adapted to the constraints of certain industrial processes, such as glycoside hydrolases and halotolerant esterases (Ferrer et al., 2005; LeCleir et al., 2007), thermostable lipases (Tirawongsaroj et al., 2008), or even psychrophilic DNA-polymerases (Simon et al., 2009). Other microbial ecosystems, such as anaerobic digesters including both human and/or animal intestinal microbiota and industrial remediation reactors, are naturally specialized in metabolizing certain substrates. These are ideal targets for research into particular functions, such as the degrading activity of lignocellulosic plant biomass (Warnecke et al., 2007; Tasse et al., 2010; Hess et al., 2011; Bastien et al., 2013) or dioxygenases for the degradation of aromatic compounds (Suenaga et al., 2007).

Some studies refer to enrichment steps that occur before sampling, with the aim of increasing the relative abundance of micro-organisms that have the target function. This enrichment can be done by modifying the physical and chemical conditions of the natural environment (van Elsas et al., 2008) or by incorporating the substrate to be metabolized *in vivo* (Hess et al., 2011) or *in vitro*, in reactors (DeAngelis et al., 2010) or mesocosms (Jacquiod et al., 2013). Through stable isotopic probing and cloning of the DNA of micro-organisms able to metabolize a specifically labeled substrate for the creation of metagenome libraries, it is possible to increase the frequency of positive clones by several orders of magnitude (Chen and Murrell, 2010). These approaches require functional and taxonomic controls at the different stages of enrichment, which are often sequential, to prevent the proliferation of populations dependent on the activity of the populations preferred at the outset. These kinds of checks are difficult to do *in vivo*, where there would actually be an increased risk of selecting populations able to metabolize only the degradation products of the initial substrate, to the detriment of those able to attack the more resistant original substrate with its more complex structure.

Functional Screening: New Challenges for the Discovery of Functions

Two complementary approaches can be used to discover new functions and protein families within microbial communities. The first involves the analysis of nucleotide, ribonucleotide or protein sequences, and the other the direct screening of functions before sequencing (**Figure 1**).

The Sequence, Marker of Originality

There have been a number of large-scale random metagenome sequencing projects (Yooseph et al., 2007; Vogel et al., 2009; Gilbert et al., 2010; Qin et al., 2010; Hess et al., 2011) over the past few years, resulting in catalogs listing millions of genes from different ecosystems, the majority of which are recorded in the GOLD[1] (RRID:nif-0000-02918), MG-RAST[2] (RRID:OMICS_01456) and EMBL-EBI[3] (RRID:nlx_72386) metagenomics databases. At the same time, the obstacles inherent to metatranscriptomic sampling (fragility of mRNA, difficulty with extraction from natural environments, separation of other types of RNA) have been removed, opening a window into the functional dynamics of ecosystems according to biotic or abiotic constraints (Saleh-Lakha et al., 2005; Warnecke and Hess, 2009; Schmieder et al., 2012). Metatranscriptomes sequencing has thus enabled the identification of new gene families, such as those found in microbial communities (prokaryotes and/or eukaryotes) expressed specifically in response to variations in the environment (Bailly et al., 2007; Frias-Lopez et al., 2008; Gilbert et al., 2008) and new enzyme sequences belonging to known carbohydrate active enzymes families (Poretsky et al., 2005; Tartar et al., 2009; Damon et al., 2012).

Regardless of the origin of the sequences (DNA or cDNA, with or without prior cloning in an expression host), the advances made with automatic annotation, most notably thanks to the IMG-M (RRID:nif-0000-03010) and MG-RAST (RRID:OMICS_01456) servers (Markowitz et al., 2007; Meyer et al., 2008), now make it possible to quantify and compare the abundance of the main functional families in the target ecosystems (Thomas et al., 2012), identified through comparison of sequences with the general functional databases: KEGG (RRID:nif-0000-21234) (Kanehisa and Goto, 2000), eggNOG (RRID:nif-0000-02789) (Muller et al., 2010), and COG/KOG (RRID:nif-0000-21313) (Tatusov et al., 2003). They also enable research into specific protein families, thanks to motif detection using Pfam (RRID:nlx_72111) (Finn et al., 2010), TIGRFAM (RRID:nif-0000-03560) (Selengut et al., 2007), CDD (RRID:nif-0000-02647) (Marchler-Bauer et al., 2009), Prosite (RRID:nif-0000-03351) (Sigrist et al., 2010), and HMM model construction (*Hidden Markov Models*; Söding, 2005). Other servers can be used to interrogate databases specialized in specific enzymatic families (**Table 1**).

Finally, the performance of methods used to assemble next generation sequencing reads is set to open up access to a plethora of complete genes to feed expert databases,

[1] http://www.genomesonline.org/cgi-bin/GOLD/index
[2] http://metagenomics.anl.gov/
[3] http://www.ebi.ac.uk/metagenomics

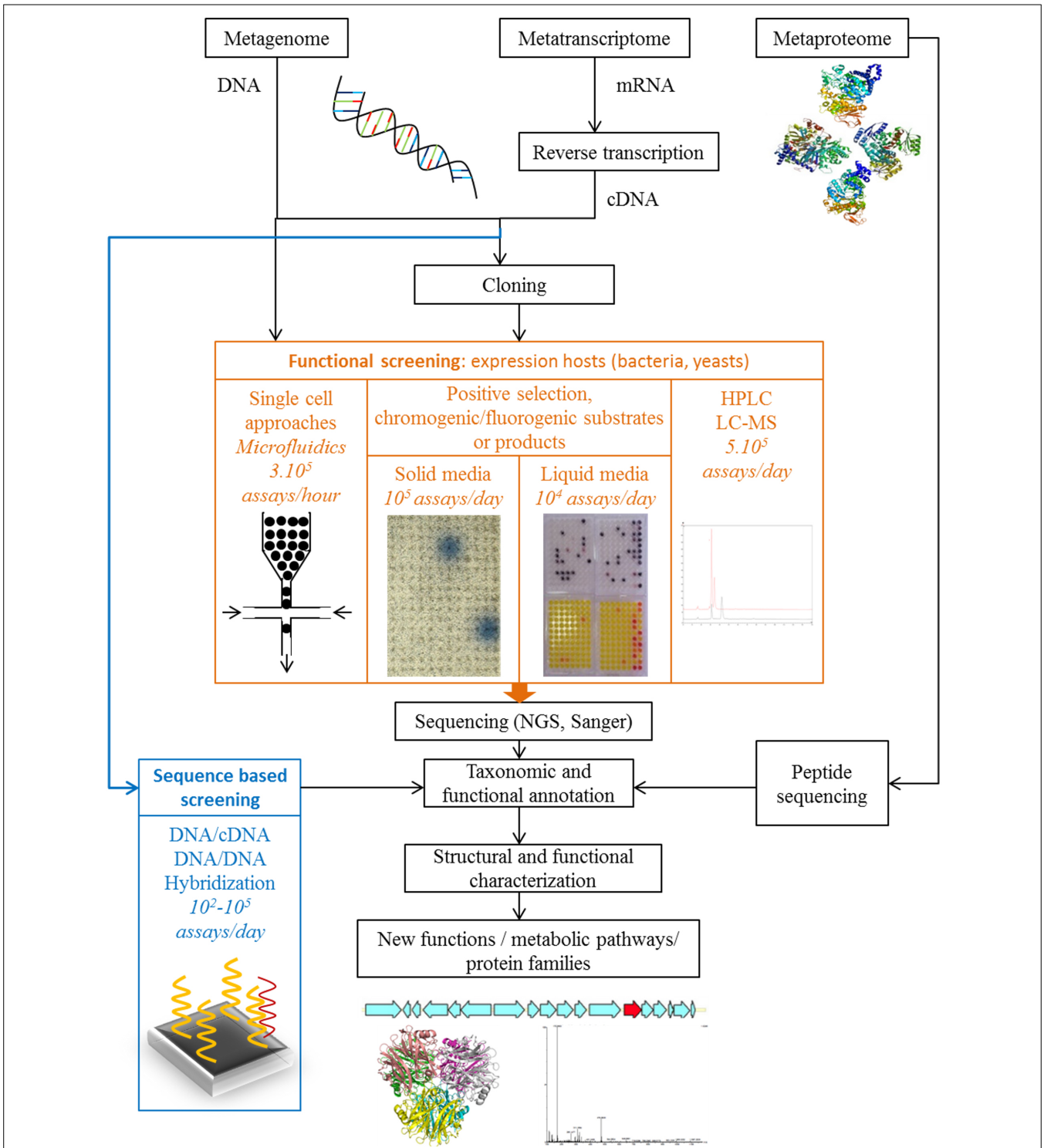

FIGURE 1 | Strategies for the functional exploration of metagenomes, metatranscriptomes and metaproteomics to discover new functions and protein families.

which currently only contain a tiny percentage of genes from uncultivated organisms—less than 1% for the CAZy database (RRID:OMICS_01677), for example—while the majority of metagenomic studies published target ecosystems with a high number of plant polysaccharide degradation activities by carbohydrate active enzymes (André et al., 2014).

Even based on a large majority of truncated genes, metagenomes and metatranscriptomes functional annotation

TABLE 1 | Examples of databases specialized in enzymatic functions of biotechnological interest.

Databases	Enzymes	References
MetaBioME	Enzymes of industrial interest	Sharma et al. (2010)
CAZy (RRID:OMICS_01677)	carbohydrate active enzymes Auxiliary redox enzymes for lignocellulose degradation	Cantarel et al. (2012) Levasseur et al. (2013)
CAT (RRID:OMICS_01676)	carbohydrate active enzymes	Park et al. (2010)
LccED	Laccases	Sirim et al. (2011)
LED (RRID:nif-0000-03084)	Lipases	Pleiss et al. (2000)
MEROPS (RRID:nif-0000-03112)	Proteases	Rawlings et al. (2012)
ThYme	Thioesterases	Cantu et al. (2011)

enables *in silico* estimations of the functional diversity of the ecosystem and identification of the most original sequences within a known protein family. It is then possible to use PCR (Polymerase Chain Reaction) to capture those sequences specifically, and test their function experimentally to assess their applicative value. In this way, the sequencing of the rumen metagenome (268 Gb) enabled identification of 27,755 coding genes for carbohydrate active enzymes, and isolation of 51 active enzymes belonging to known families specifically involved in lignocellulose degradation (Hess et al., 2011).

PCR, and more generally DNA/DNA or DNA/cDNA hybridization, also make it possible to directly capture coding genes for protein families that are abundant and/or expressed in the target ecosystem, but with no need for *a priori* large-scale sequencing. This strategy requires the conception of nucleic acid probes or PCR primers using consensus sequences specific to known protein families. There are plenty of examples of the discovery of enzymes in metagenomes using these approaches, for instance bacterial laccases (Ausec et al., 2011), dioxygenases (Zaprasis et al., 2009), nitrites reductases (Bartossek et al., 2010), hydrogenases (Schmidt et al., 2010), hydrazine oxidoreductases (Li et al., 2010), or chitinases (Hjort et al., 2010) from various ecosystems. The Gene-Targeted-metagenomics approach (Iwai et al., 2009) combines PCR screening and amplicon pyrosequencing to generate primers in an iterative manner and increase the structural diversity of the target protein families, for example the dioxygenases from the microbiota of contaminated soil. Elsewhere, the use of high-density functional microarrays considerably multiplies the number of probes and is therefore a low-cost way of obtaining a snapshot of the abundance and diversity of sequences within specific protein families and even, where the DNA or cDNA has been cloned (He et al., 2010; Weckx et al., 2010), directly capturing targets of interest while rationalizing sequencing. Using a similar strategy, the solution hybrid selection method enables the selection of fragments of coding DNA for specific enzymatic families using 31-mers capture probes. Applied to the capture of cDNA, this method provides access to entire genes which can be then cloned and their activity tested (Bragalini et al., 2014). Solution hybrid selection can therefore be used to explore the taxonomic and functional diversity of all protein families. More especially, this approach opens the way for the selection and characterization of families that are highly represented in a microbiome but whose function remains unknown, in order to further the understanding of ecosystemic functions and discover novel biocatalysts.

Metaproteomics has recently proved its worth in identifying new protein families and/or functions. Paired with genomic, metagenomic and metatranscriptomic data (Erickson et al., 2012), it provides access to excellent biomarkers of the functional state of the ecosystem. Recent developments, such as high-throughput electrospray ionization paired with mass spectrometry, enable full metaproteome analysis after separation of proteins by liquid chromatography. It is thus possible to highlight hundreds of proteins with no associated function and new enzyme families playing a key functional role in the ecosystem (Ram et al., 2005).

This latter example illustrates the need for research and/or experimental proof of function for proteins where the function remains unknown (products of orphan genes or, on the contrary, genes highly prevalent in the microbial realm but that have never been characterized) or poorly annotated. In fact, annotation errors, which are especially common for multi-modular proteins such as carbohydrate active enzymes, are spread at an increasing rate as a result of the explosion in the number of functional genomics and meta-genomic, -transcriptomic and -proteomic projects. New annotation strategies, most notably based on the prediction of the three-dimensional structure of proteins, are also worth exploring (Uchiyama and Miyazaki, 2009). However, at the present time, it is very difficult to predict the specificity of substrate and the mechanism of action (and therefore the function of the protein) on the basis of sequence or even structure, especially where there is no homologue characterized from a structural and functional point of view. Functional screening can address this challenge.

Activity Screening: Speeding up the Discovery of Biotechnology Tools

There are three prerequisites for this approach: (i) the cloning of DNA or cDNA in an expression vector for the creation of, respectively, metagenomic or metatranscriptomic libraries, (ii) heterologous expression of cloned genes in a microbial host, iii) the conception of efficient phenotypic screens to isolate the clones of interest that produce the target activity, also referred to as "hits."

Using this approach, the functions of a protein can be accessed without any prior information on its sequence. It is therefore the only way of identifying novel protein families that have known functions or previously unseen functions (as long as an adequate screen can be developed). Finally, it helps to rationalize sequencing efforts and focus them only on the hits: for example, those that are of biotechnological interest. The expression potential of the selected heterologous host, the size of the DNA inserts and the type of vectors all determine the success of functional screening. Short fragments of metagenomic DNA (smaller than 15 kb, and most often between 2 and 5 kb), or cDNA for the metatranscriptomic libraries, cloned in plasmids under the influence of a strong expression promoter, enable the

overexpression of a single protein, and the easy recovery and sequencing of the hits' DNA (Uchiyama and Miyazaki, 2009). On the other hand, fragments of bacterial DNA measuring between 15 and 40 kb, 25 and 45 kb or even 100 and 200 kb, cloned respectively in cosmids, fosmids or bacterial artificial chromosomes, can be used to explore a functional diversity of several Gb per library and, above all, provide access to operon-type multigene clusters, coding for complete catabolic or anabolic pathways This is of major interest for the discovery of cocktails of synergistic activities that degrade complex substrates such as plant cell walls for biorefineries. This strategy also ensures high reliability for the taxonomic annotation of inserts, and can even be used to identify the mobile elements responsible for the plasticity of the bacterial metagenome, mediated by horizontal gene transfers (Tasse et al., 2010). However, it requires sensitive activity screens, since the target genes are only weakly expressed, controlled by their own native promoters.

Escherichia coli, whose transformation efficiency is exceptionally high, even for fosmids or bacterial artificial chromosomes, remains the host of choice in the immense majority of studies published. The first exhaustive functional screening study of a fosmid library revealed that *E. coli* can be used to express genes from bacteria that are very different from a taxonomical point of view, including a large number of Bacteroidetes and Gram-positive bacteria (Tasse et al., 2010), contrary to what had been predicted by *in silico* detection of expression signals compatible with *E. coli* (Gabor et al., 2004). However, the value of developing shuttle vectors to screen metagenomic libraries in hosts with different expression and secretion potentials, for example *Bacillus*, *Sphingomonas*, *Streptomyces*, *Thermus*, or the α-, β- and γ−proteobacteria (Taupp et al., 2011; Ekkers et al., 2012) must not be underestimated, if we are to unlock the functional potential of varied taxons and increase the sensitivity of screens. Finally, it is still very difficult to get access to the uncultivated fraction of eukaryotic microorganisms, due to the lack of screening hosts with sufficient transformation efficiency for the creation of large clone libraries (and thus the exploration of a vast array of sequences) and compatible with the post-translational modifications required to obtain functional recombinant proteins from eukaryotes. Thus, at the present time, only a few studies have been published on the enzyme activity-based screening of metatranscriptomic libraries (making it possible to do away with introns) of eukaryotes from soil, rumen and the gut of the termite (Bailly et al., 2007; Findley et al., 2011, Sethi et al., 2013).

Regardless of the type of library screened, the functional exploration of hundreds of thousands of clones is required, whereas the hit rate rarely exceeds 6‰ (Duan et al., 2009; Bastien et al., 2013). This requires very high throughput primary screens, in a solid medium before or after the automated organization of libraries in 96- or 384-well micro-plate format, in a liquid medium after enzymatic cell lysis and/or thawing and freezing (Bao et al., 2011), or using UV-inducible auto-lytic vectors (Li et al., 2007). This stage is very often followed by medium or low throughput characterization of the properties of the hits obtained, particularly to assess their biotechnological interest (Tasse et al., 2010).

Two generic strategies, used at throughputs exceeding 400,000 tests per week, have been and continue to be applied widely. Positive selection on a medium containing, for example, substrates to be metabolized as the sole source of carbon, can be used to isolate enzymes (Henne et al., 1999), complete catabolic pathways (Cecchini et al., 2013), or membrane transporters (Majerník et al., 2001). This approach also helps easily identify antibiotic resistant genes (Diaz-Torres et al., 2006). The use of chromogenic (Beloqui et al., 2010; Bastien et al., 2013; Nyyssönen et al., 2013), fluorescent (LeCleir et al., 2007), or opalescent substrates or reagents, such as insoluble polymers or proteins (Mayumi et al., 2008; Waschkowitz et al., 2009), or simply the observation of an original clone phenotype, has already enabled the isolation of several 100 catabolic enzymes, like the numerous hydrolases of very varied taxonomic origin (Simon and Daniel, 2009), some of which were coded by genes that are very abundant in the target ecosystem (Jones et al., 2008; Gloux et al., 2011), but also, although much less frequently, new oxidoreductases (Knietsch et al., 2003). Novel enzymes (laccases, esterases and oxygenases in particular) from microbial communities of very diverse origins (soil, water, activated sludge, digestive tracts) have been highlighted for their capacity to degrade pollutants such as nitriles (Robertson and Steer, 2004), lindane (Boubakri et al., 2006), styrene (Van Hellemond et al., 2007), naphthalene (Ono et al., 2007), aliphatic and aromatic carbohydrates (Uchiyama et al., 2004; Brennerova et al., 2009; Lu et al., 2012), organophosphorus (Kambiranda et al., 2009; Math et al., 2010), or plastic materials (Mayumi et al., 2008).

The discovery of proteins involved in prokaryote-eukaryote interactions (Lakhdari et al., 2010) or anabolic pathways is rarer, since it often requires the development of complex screens and lower throughputs. Nonetheless, a few examples of simple screens, based on the aptitude of metagenomic clones to inhibit the growth of a strain by producing antibacterial activity or to complement an auxotrophic strain for a specific compound, have enabled the identification of new pathways for the synthesis of antimicrobials (Brady and Clardy, 2004) or biotin (Entcheva et al., 2001). Nano-technologies, and in particular the latest developments focused on the medium-throughput screening of libraries obtained by combinatorial protein engineering, enable the design of custom microarrays and covered with one to several 100 specific enzymatic substrates, the processing of which may be followed by fluorescence, chemiluminescence, immunodetection, surface plasmon resonance or mass spectrometry (André et al., 2014). Nanostructure-initiator mass spectrometry technology, combining fluorescence and mass spectrometry, is the first example of a functional metagenomic application for the discovery of anabolic enzymes, namely sialyltransferases (Northen et al., 2008).

The Immense Challenges of Ultra-fast Screening (Figure 2)

Microfluidic technologies are of undeniable interest when it comes to reaching screening rates of a million clones per day. The substrate induced gene-expression screening method has been developed to use fluorescence-activated cell sorting

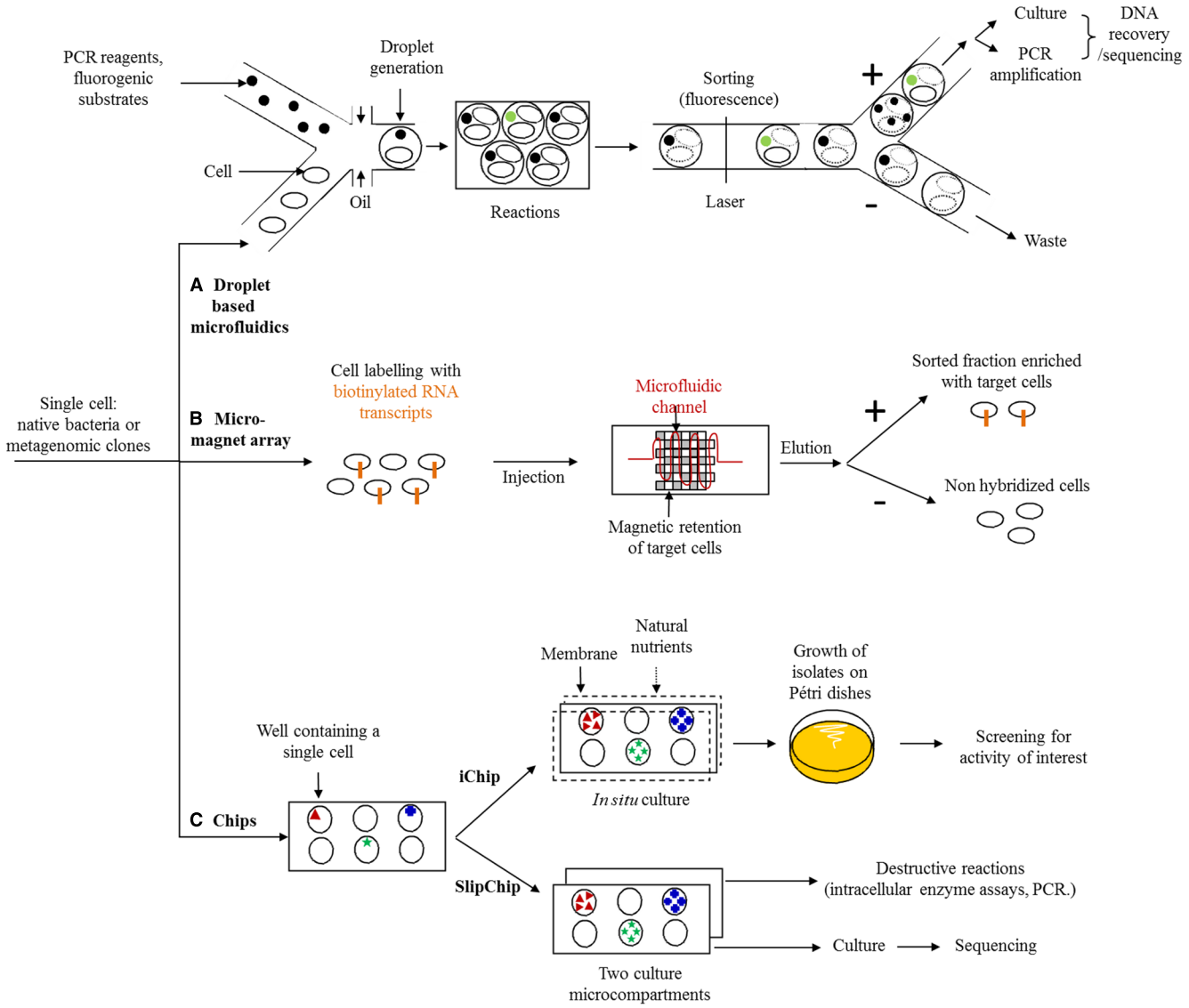

FIGURE 2 | Microfluidic strategies for new enzyme screening. **(A)** Droplet based microfluidics: single cells are encapsulated with probes or fluorogenic substrates to create microdroplets, where reactions happen (substrate degradation, PCR). The hits are sorted using fluorescence detection. Non-lysed cells are cultured and DNA fragments from lysed cells are amplified. Both methods allow the recovery and sequencing of DNA. **(B)** Micro-magnet array: target cells are labeled with biotinylated RNA transcripts probes and injected inside the microchannel. Target cells are captured in the channel thanks to magnetic forces while non-targets cells pass through the device. **(C)** Chips: the chip wells are filled with a single cell. The iChip is covered by membranes, and reintroduced into original environment, where natural nutrients flow through membranes. Colonies are further isolated on Petri dishes to be screened for the activity of interest. The SlipChip is composed of two culture microcompartments which are further separated for destructive and non-destructive assays.

to isolate plasmidic clones containing genes (or fragments of genes) that induce the expression of a fluorescent marker in response to a specific substrate. However, this technique is only suited to small substrates that are non-lethal and internalizable for the host strain (Uchiyama and Watanabe, 2008). Finally, the advances made over the past few years in cellular compartmentalization (Nawy, 2013), selective sorting, based on sequence detection (Pivetal et al., 2014; Lim et al., 2015) or specific metabolites (Kürsten et al., 2014) and the control of reaction kinetics (Mazutis et al., 2009) in microfluidic circuits should allow for a huge acceleration in the discovery of new proteins and metabolic pathways expressed in prokaryotes and eukaryotes in an intercellular, membrane or extracellular manner.

The very first examples of metagenome functional exploration applications have already been used to establish the proof of concept regarding the effectiveness of microfluidics in the discovery of new bioactive molecules and new enzymes. For example, droplet-based microfluidics technology was recently used by the teams of A. Griffiths and A. Drevelle to isolate new

strains producing cellobiohydrolase and cellulase activities at a rate of 300,000 cells sorted per hour, using just a few microliters of reagent, i.e., 250,000 times less than with the conventional technologies mentioned above (Najah et al., 2014). Here, soil bacteria and a fluorescent substrate were co-encapsulated in micro-droplets in order to sort cells on the basis of the extracellular activity only. In fact, the strategy used, which requires the seeding of cells on a defined medium after sorting, is not compatible with the detection of intracellular enzymes, which require a lethal lysis step to convert the substrate. Applying a similar principle, the ultra-rapid sorting of eukaryote cells encapsulated with their substrate now also makes it possible to select yeast clones presenting extracellular enzymatic activities (Sjostrom et al., 2014). This technology should, in the short term, make it possible to explore the functional diversity of uncultivated eukaryotes at a very high throughput, by directly sorting fungal populations or libraries of metatranscriptomic clones. In the latter case, access to the sequence involved in the target activity will be easy, since the libraries are built using hosts whose culture is well managed, with insertion of the metatranscriptomic cDNA fragment into a specific region of the genome. Where sorting is done without cloning of the metagenome or metatranscriptome, only microorganisms capable of growth on a defined medium can be recovered, which hugely limits access to functional diversity.

To increase the proportion of cultivable organisms, Kim Lewis' team recently used the iChip to simultaneously isolate and cultivate soil bacteria thanks to the delivery of nutrients from the original medium, into which the iChip is introduced, via semi-permeable membranes. This method enables an increase in cultivable organisms ranging from 1 to 50%. Using colonies cultivated in the chip, the clones isolated in a Petri dish were screened for the production of antimicrobial compounds (Ling et al., 2015). A novel antibiotic was thus identified, together with its biosynthesis pathway, after sequencing and functional annotation of the complete genome.

It is quite another matter when it comes to selecting, on the basis of intracellular activity, completely uncultivable organisms or metagenomic clones containing DNA inserts of several dozen kbp, which are difficult to amplify using PCR. In this case, to liberate the enzymes in question, we are required to include a cellular lysis step, preventing seeding after sorting. On the other hand, this approach is compatible with the sorting of plasmid clone libraries, where the metagenomic or metatranscriptomic inserts can easily be amplified using PCR, on the basis of just a few dozen lysed cells. For libraries with large DNA inserts, the barriers are now being broken down, most notably thanks to the development of the SlipChips microfluidic approach (Ma et al., 2014), which uses two culture microcompartments, where the content of one can be lysed for the detection of enzymatic activities, for example, and the other is used as a backup replicate for the culture and recovery of subsequent DNA for sequencing. In spite of these recent, highly encouraging developments, the proof of concept has not yet been established for the identification of new functions and intracellular metabolic pathways.

Conclusion

The rapid expansion of meta-omic technologies over the past decade has shed light on the functions of the uncultivated fraction of microbial ecosystems. A huge number of enzymes have been discovered, in particular through experimental approaches to functional metagenomes exploration. Where their performance can be rapidly assessed within the framework of a known process, or where they catalyze new, previously undescribed reactions, many of them have provided new tools for industrial biotechnologies. However, several challenges still need to be addressed to speed up the rate at which new functions are discovered and to make optimal use of the functional diversity that so far remains unexplored. Firstly, while the uncultivated prokaryote fraction of microbial communities is still extensively studied, the functions of the eukaryote fraction are relatively unexplored from an experimental angle, even though they play a fundamental role for numerous ecosystems. Secondly, in the majority of cases, the functions discovered using meta-omic approaches play a catabolic role, mainly involved in the deconstruction of plant biomass or in bioremediation. It is thus necessary to develop functional screens to access anabolic functions and enrich the catalog of reactions available for synthetic biology. Finally, there are very few studies aimed at identifying the role of protein families that are highly prevalent in the target ecosystem but that have not yet been characterized, even though some of them could be considered as biomarkers of the functional state of the microbial community. Indeed, sequence-based functional metagenomic projects continuously highlight many sequences annotated as domains of unknown function in the Pfam database (RRID: nlx_72111) (Bateman et al., 2010; Finn et al., 2014), some with 3D structures solved thanks to structural genomics initiatives, and available in the Protein Data Bank (RRID: nif-0000-00135). With the goal of characterizing these new protein families and identifying previously unseen functions from the selection the most prevalent protein families (those containing the highest number of homologous sequences without any associated function) in the target ecosystem, the integration of structural, biochemical, genomic and meta-omic data is now also possible (Ladevèze et al., 2013). It allows to benefit from the huge amount of long scaffolds now available in sequence databases, and to access the genomic context of the targeted genes in order to facilitate functional assignation. In the next few years, these strategies should enhance our understanding of how microbial ecosystems function and, at the same time, enable greater control over them.

Author Contributions

LU, GPV, EL contributed equally to this work.

Acknowledgments

This research was funded by the Ministry of Education and Research (Ministère de l'Enseignement supérieur et de la Recherche, MESR), the Agence Nationale de la Recherche (Grant Number ANR 2011-Nano 007 03) and the INRA metaprogramme M2E (project Metascreen).

References

André, I., Potocki-Véronèse, G., Barbe, S., Moulis, C., and Remaud-Siméon, M. (2014). CAZyme discovery and design for sweet dreams. *Curr. Opin. Chem. Biol.* 19, 17–24. doi: 10.1016/j.cbpa.2013.11.014

Ausec, L., van Elsas, J. D., and Mandic-Mulec, I. (2011). Two- and three-domain bacterial laccase-like genes are present in drained peat soils. *Soil Biol. Biochem.* 43, 975–983. doi: 10.1016/j.soilbio.2011.01.013

Bailly, J., Fraissinet-Tachet, L., Verner, M.-C., Debaud, J.-C., Lemaire, M., Wésolowski-Louvel, M., et al. (2007). Soil eukaryotic functional diversity, a metatranscriptomic approach. *ISME J.* 1, 632–642. doi: 10.1038/ismej.2007.68

Bao, L., Huang, Q., Chang, L., Zhou, J., and Lu, H. (2011). Screening and characterization of a cellulase with endocellulase and exocellulase activity from yak rumen metagenome. *J. Mol. Catal. B Enzym.* 73, 104–110. doi: 10.1016/j.molcatb.2011.08.006

Bartossek, R., Nicol, G. W., Lanzen, A., Klenk, H.-P., and Schleper, C. (2010). Homologues of nitrite reductases in ammonia-oxidizing archaea: diversity and genomic context. *Environ. Microbiol.* 12, 1075–1088. doi: 10.1111/j.1462-2920.2010.02153.x

Bastien, G., Arnal, G., Bozonnet, S., Laguerre, S., Ferreira, F., Fauré, R., et al. (2013). Mining for hemicellulases in the fungus-growing termite *Pseudacanthotermes militaris* using functional metagenomics. *Biotechnol. Biofuels* 6, 78. doi: 10.1186/1754-6834-6-78

Bateman, A., Coggill, P., and Finn, R. D. (2010). DUFs: families in search of function. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* 66, 1148–1152. doi: 10.1107/S1744309110001685

Beloqui, A., Polaina, J., Vieites, J. M., Reyes-Duarte, D., Torres, R., Golyshina, O. V., et al. (2010). Novel hybrid esterase-haloacid dehalogenase enzyme. *Chembiochem* 11, 1975–1978. doi: 10.1002/cbic.201000258

Boubakri, H., Beuf, M., Simonet, P., and Vogel, T. M. (2006). Development of metagenomic DNA shuffling for the construction of a xenobiotic gene. *Gene* 375, 87–94. doi: 10.1016/j.gene.2006.02.027

Brady, S. F., and Clardy, J. (2004). Palmitoylputrescine, an antibiotic isolated from the heterologous expression of DNA extracted from bromeliad tank water. *J. Nat. Prod.* 67, 1283–1286. doi: 10.1021/np0499766

Bragalini, C., Ribiere, C., Parisot, N., Vallon, L., Prudent, E., Peyretaillade, E., et al. (2014). Solution hybrid selection capture for the recovery of functional full-length eukaryotic cDNAs from complex environmental samples. *DNA Res.* 21, 685–694. doi: 10.1093/dnares/dsu030

Brennerova, M. V., Josefiova, J., Brenner, V., Pieper, D. H., and Junca, H. (2009). Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. *Environ. Microbiol.* 11, 2216–2227. doi: 10.1111/j.1462-2920.2009.01943.x

Cantarel, B. L., Lombard, V., and Henrissat, B. (2012). Complex carbohydrate utilization by the healthy human microbiome. *PLoS ONE* 7:e28742. doi: 10.1371/journal.pone.0028742

Cantu, D. C., Chen, Y., Lemons, M. L., and Reilly, P. J. (2011). ThYme: a database for thioester-active enzymes. *Nucleic Acids Res.* 39, D342–D346. doi: 10.1093/nar/gkq1072

Cecchini, D. A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J., et al. (2013). Functional metagenomics reveals novel pathways of prebiotic breakdown by human gut bacteria. *PLoS ONE* 8:e72766. doi: 10.1371/journal.pone.0072766

Chen, Y., and Murrell, J. C. (2010). When metagenomics meets stable-isotope probing: progress and perspectives. *Trends Microbiol.* 18, 157–163. doi: 10.1016/j.tim.2010.02.002

Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. (2014). Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Virulence* 5, 399–412. doi: 10.4161/viru.27208

Damon, C., Lehembre, F., Oger-Desfeux, C., Luis, P., Ranger, J., Fraissinet-Tachet, L., et al. (2012). Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. *PLoS ONE* 7:e28967. doi: 10.1371/journal.pone.0028967

DeAngelis, K. M., Gladden, J. M., Allgaier, M., D'haeseleer, P., Fortney, J. L., Reddy, A., et al. (2010). Strategies for enhancing the effectiveness of metagenomic-based enzyme discovery in lignocellulolytic microbial communities. *BioEnergy Res.* 3, 146–158. doi: 10.1007/s12155-010-9089-z

Diaz-Torres, M. L., Villedieu, A., Hunt, N., McNab, R., Spratt, D. A., Allan, E., et al. (2006). Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS Microbiol. Lett.* 258, 257–262. doi: 10.1111/j.1574-6968.2006.00221.x

Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P., and Reid, G. (2013). High throughput sequencing methods and analysis for microbiome research. *J. Microbiol. Methods* 95, 401–414. doi: 10.1016/j.mimet.2013.08.011

Duan, C.-J., Xian, L., Zhao, G.-C., Feng, Y., Pang, H., Bai, X.-L., et al. (2009). Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J. Appl. Microbiol.* 107, 245–256. doi: 10.1111/j.1365-2672.2009.04202.x

Ekkers, D. M., Cretoiu, M. S., Kielak, A. M., and van Elsas, J. D. (2012). The great screen anomaly—a new frontier in product discovery through functional metagenomics. *Appl. Microbiol. Biotechnol.* 93, 1005–1020. doi: 10.1007/s00253-011-3804-3

Entcheva, P., Liebl, W., Johann, A., Hartsch, T., and Streit, W. R. (2001). Direct cloning from enrichment cultures, a reliable strategy for isolation of complete operons and genes from microbial consortia. *Appl. Environ. Microbiol.* 67, 89–99. doi: 10.1128/AEM.67.1.89-99.2001

Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* 7:e49138. doi: 10.1371/journal.pone.0049138

Ferrer, M., Beloqui, A., Timmis, K. N., and Golyshin, P. N. (2009). Metagenomics for mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.* 16, 109–123. doi: 10.1159/000142898

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Martins Dos Santos, V. A. P., Yakimov, M. M., et al. (2005). Microbial enzymes mined from the Urania deep-sea hypersaline anoxic basin. *Chem. Biol.* 12, 895–904. doi: 10.1016/j.chembiol.2005.05.020

Findley, S. D., Mormile, M. R., Sommer-Hurley, A., Zhang, X.-C., Tipton, P., Arnett, K., et al. (2011). Activity-based metagenomic screening and biochemical characterization of bovine ruminal protozoan glycoside hydrolases. *Appl. Environ. Microbiol.* 77, 8106–8113. doi: 10.1128/AEM.05925-11

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., et al. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222. doi: 10.1093/nar/gkp985

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., et al. (2008). Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3805–3810. doi: 10.1073/pnas.0708897105

Gabor, E. M., Alkema, W. B. L., and Janssen, D. B. (2004). Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ. Microbiol.* 6, 879–886. doi: 10.1111/j.1462-2920.2004.00640.x

Gilbert, J. A., Field, D., Huang, Y., Edwards, R., Li, W., Gilna, P., et al. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* 3:e3042. doi: 10.1371/journal.pone.0003042

Gilbert, J. A., Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., et al. (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a "Multi-Omic" study of seasonal and diel temporal variation. *PLoS ONE* 5:e15545. doi: 10.1371/journal.pone.0015545

Gloux, K., Berteau, O., El oumami, H., Beguet, F., Leclerc, M., and Dore, J. (2011). A metagenomic β-glucuronidase uncovers a core adaptive function of the human intestinal microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4539–4546. doi: 10.1073/pnas.1000066107

He, S., Kunin, V., Haynes, M., Martin, H. G., Ivanova, N., Rohwer, F., et al. (2010). Metatranscriptomic array analysis of "*Candidatus* Accumulibacter phosphatis"-enriched enhanced biological phosphorus removal sludge: metatranscriptomic array analysis of EBPR sludge. *Environ. Microbiol.* 12, 1205–1217. doi: 10.1111/j.1462-2920.2010.02163.x

Henne, A., Daniel, R., Schmitz, R. A., and Gottschalk, G. (1999). Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl. Environ. Microbiol.* 65, 3901–3907.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331, 463–467. doi: 10.1126/science.1200387

Hjort, K., Bergström, M., Adesina, M. F., Jansson, J. K., Smalla, K., and Sjöling, S. (2010). Chitinase genes revealed and compared in bacterial isolates, DNA extracts and a metagenomic library from a phytopathogen-suppressive soil. *FEMS Microbiol. Ecol.* 71, 197–207. doi: 10.1111/j.1574-6941.2009.00801.x

Iwai, S., Chai, B., Sul, W. J., Cole, J. R., Hashsham, S. A., and Tiedje, J. M. (2009). Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. *ISME J.* 4, 279–285. doi: 10.1038/ismej.2009.104

Jacquiod, S., Franqueville, L., Cécillon, S., Vogel, T. M., and Simonet, P. (2013). Soil bacterial community shifts after chitin enrichment: an integrative metagenomic approach. *PLoS ONE* 8:e79699. doi: 10.1371/journal.pone.0079699

Jones, B. V., Begley, M., Hill, C., Gahan, C. G. M., and Marchesi, J. R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13580–13585. doi: 10.1073/pnas.0804437105

Kambiranda, D. M., Asraful-Islam, S. M., Cho, K. M., Math, R. K., Lee, Y. H., Kim, H., et al. (2009). Expression of esterase gene in yeast for organophosphates biodegradation. *Pestic. Biochem. Physiol.* 94, 15–20. doi: 10.1016/j.pestbp.2009.02.006

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Knietsch, A., Waschkowitz, T., Bowien, S., Henne, A., and Daniel, R. (2003). Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. *Appl. Environ. Microbiol.* 69, 1408–1416. doi: 10.1128/AEM.69.3.1408-1416.2003

Kürsten, D., Kothe, E., Wetzel, K., Bergmann, K., and Köhler, J. M. (2014). Micro-segmented flow and multisensor-technology for microbial activity profiling. *Environ. Sci. Process. Impacts* 16, 2362–2370. doi: 10.1039/C4EM00255E

Ladevèze, S., Tarquis, L., Cecchini, D. A., Bercovici, J., André, I., Topham, C. M., et al. (2013). Role of glycoside phosphorylases in mannose foraging by human gut bacteria. *J. Biol. Chem.* 288, 32370–32383. doi: 10.1074/jbc.M113.483628

Ladoukakis, E., Kolisis, F. N., and Chatziioannou, A. A. (2014). Integrative workflows for metagenomic analysis. *Front. Cell Dev. Biol.* 2:70. doi: 10.3389/fcell.2014.00070

Lakhdari, O., Cultrone, A., Tap, J., Gloux, K., Bernard, F., Ehrlich, S. D., et al. (2010). Functional metagenomics: a high throughput screening method to decipher microbiota-driven NF-κB modulation in the human gut. *PLoS ONE* 5:e13092. doi: 10.1371/journal.pone.0013092

LeCleir, G. R., Buchan, A., Maurer, J., Moran, M. A., and Hollibaugh, J. T. (2007). Comparison of chitinolytic enzymes from an alkaline, hypersaline lake and an estuary. *Environ. Microbiol.* 9, 197–205. doi: 10.1111/j.1462-2920.2006.01128.x

Levasseur, A., Drula, E., Lombard, V., Coutinho, P. M., and Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. *Biotechnol. Biofuels* 6, 41. doi: 10.1186/1754-6834-6-41

Li, M., Hong, Y., Klotz, M. G., and Gu, J.-D. (2010). A comparison of primer sets for detecting 16S rRNA and hydrazine oxidoreductase genes of anaerobic ammonium-oxidizing bacteria in marine sediments. *Appl. Microbiol. Biotechnol.* 86, 781–790. doi: 10.1007/s00253-009-2361-5

Li, S., Xu, L., Hua, H., Ren, C., and Lin, Z. (2007). A set of UV-inducible autolytic vectors for high throughput screening. *J. Biotechnol.* 127, 647–652. doi: 10.1016/j.jbiotec.2006.07.030

Lim, S. W., Tran, T. M., and Abate, A. R. (2015). PCR-activated cell sorting for cultivation-free enrichment and sequencing of rare microbes. *PLoS ONE* 10:e0113549. doi: 10.1371/journal.pone.0113549

Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., et al. (2015). A new antibiotic kills pathogens without detectable resistance. *Nature* 517, 455–459. doi: 10.1038/nature14098

Lu, Z., Deng, Y., Van Nostrand, J. D., He, Z., Voordeckers, J., Zhou, A., et al. (2012). Microbial gene functions enriched in the Deepwater Horizon deep-sea oil plume. *ISME J.* 6, 451–460. doi: 10.1038/ismej.2011.91

Ma, L., Datta, S. S., Karymov, M. A., Pan, Q., Begolo, S., and Ismagilov, R. F. (2014). Individually addressable arrays of replica microbial cultures enabled by splitting SlipChips. *Integr. Biol.* 6, 796–805. doi: 10.1039/C4IB00109E

Majerník, A., Gottschalk, G., and Daniel, R. (2001). Screening of environmental DNA libraries for the presence of genes conferring Na+(Li+)/H+ antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. *J. Bacteriol.* 183, 6645–6653. doi: 10.1128/JB.183.22.6645-6653.2001

Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., et al. (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.* 37, D205–D210. doi: 10.1093/nar/gkn845

Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., et al. (2007). IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 36, D534–D538. doi: 10.1093/nar/gkm869

Math, R. K., Asraful Islam, S. M., Cho, K. M., Hong, S. J., Kim, J. M., Yun, M. G., et al. (2010). Isolation of a novel gene encoding a 3,5,6-trichloro-2-pyridinol degrading enzyme from a cow rumen metagenomic library. *Biodegradation* 21, 565–573. doi: 10.1007/s10532-009-9324-5

Mayumi, D., Akutsu-Shigeno, Y., Uchiyama, H., Nomura, N., and Nakajima-Kambe, T. (2008). Identification and characterization of novel poly (DL-lactic acid) depolymerases from metagenome. *Appl. Microbiol. Biotechnol.* 79, 743–750. doi: 10.1007/s00253-008-1477-3

Mazutis, L., Baret, J.-C., and Griffiths, A. D. (2009). A fast and efficient microfluidic system for highly selective one-to-one droplet fusion. *Lab Chip.* 9, 2665–2672. doi: 10.1039/b903608c

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., et al. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* 38, D190–D195. doi: 10.1093/nar/gkp951

Najah, M., Calbrix, R., Mahendra-Wijaya, I. P., Beneyton, T., Griffiths, A. D., and Drevelle, A. (2014). Droplet-based microfluidics platform for ultra-high-throughput bioprospecting of cellulolytic microorganisms. *Chem. Biol.* 21, 1722–1732. doi: 10.1016/j.chembiol.2014.10.020

Nawy, T. (2013). Lab-On-A-Chip: receptive cells feel the squeeze. *Nat. Methods* 10, 198–198. doi: 10.1038/nmeth.2395

Northen, T. R., Lee, J.-C., Hoang, L., Raymond, J., Hwang, D.-R., Yannone, S. M., et al. (2008). A nanostructure-initiator mass spectrometry-based enzyme activity assay. *Proc. Natl. Acad. Sci. U.S.A.* 105, 3678–3683. doi: 10.1073/pnas.0712332105

Nyyssönen, M., Tran, H. M., Karaoz, U., Weihe, C., Hadi, M. Z., Martiny, J. B. H., et al. (2013). Coupled high-throughput functional screening and next generation sequencing for identification of plant polymer decomposing enzymes in metagenomic libraries. *Front. Microbiol.* 4:282. doi: 10.3389/fmicb.2013.00282

Ono, A., Miyazaki, R., Sota, M., Ohtsubo, Y., Nagata, Y., and Tsuda, M. (2007). Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. *Appl. Microbiol. Biotechnol.* 74, 501–510. doi: 10.1007/s00253-006-0671-4

Park, B. H., Karpinets, T. V., Syed, M. H., Leuze, M. R., and Uberbacher, E. C. (2010). CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* 20, 1574–1584. doi: 10.1093/glycob/cwq106

Pivetal, J., Toru, S., Frenea-Robin, M., Haddour, N., Cecillon, S., Dempsey, N. M., et al. (2014). Selective isolation of bacterial cells within a microfluidic device using magnetic probe-based cell fishing. *Sens. Actuators B Chem.* 195, 581–589. doi: 10.1016/j.snb.2014.01.004

Pleiss, J., Fischer, M., Peiker, M., Thiele, C., and Schmid, R. D. (2000). Lipase engineering database. *J. Mol. Catal. B Enzym.* 10, 491–508. doi: 10.1016/S1381-1177(00)00092-8

Poretsky, R. S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M., et al. (2005). Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* 71, 4121–4126. doi: 10.1128/AEM.71.7.4121-4126.2005

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Ram, R. J., Verberkmoes, N. C., Thelen, M. P., Tyson, G. W., Baker, B. J., Blake, R. C., et al. (2005). Community proteomics of a natural microbial biofilm. *Science* 308, 1915–1920. doi: 10.1126/science. 1109070

Rawlings, N. D., Barrett, A. J., and Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 40, D343–D350. doi: 10.1093/nar/gkr987

Robertson, D. E., and Steer, B. A. (2004). Recent progress in biocatalyst discovery and optimization. *Curr. Opin. Chem. Biol.* 8, 141–149. doi: 10.1016/j.cbpa.2004.02.010

Saleh-Lakha, S., Miller, M., Campbell, R. G., Schneider, K., Elahimanesh, P., Hart, M. M., et al. (2005). Microbial gene expression in soil: methods, applications and challenges. *J. Microbiol. Methods* 63, 1–19. doi: 10.1016/j.mimet.2005.03.007

Schmidt, O., Drake, H. L., and Horn, M. A. (2010). Hitherto unknown [Fe-Fe]-hydrogenase gene diversity in anaerobes and anoxic enrichments from a moderately acidic fen. *Appl. Environ. Microbiol.* 76, 2027–2031. doi: 10.1128/AEM.02895-09

Schmieder, R., Lim, Y. W., and Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* 28, 433–435. doi: 10.1093/bioinformatics/btr669

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., et al. (2007). TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.* 35, D260–D264. doi: 10.1093/nar/gkl1043

Sethi, A., Slack, J. M., Kovaleva, E. S., Buchman, G. W., and Scharf, M. E. (2013). Lignin-associated metagene expression in a lignocellulose-digesting termite. *Insect Biochem. Mol. Biol.* 43, 91–101. doi: 10.1016/j.ibmb.2012.10.001

Sharma, V. K., Kumar, N., Prakash, T., and Taylor, T. D. (2010). MetaBioME: a database to explore commercially useful enzymes in metagenomic datasets. *Nucleic Acids Res.* 38, D468–D472. doi: 10.1093/nar/gkp1001

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 38, D161–D166. doi: 10.1093/nar/gkp885

Simon, C., and Daniel, R. (2009). Achievements and new knowledge unraveled by metagenomic approaches. *Appl. Microbiol. Biotechnol.* 85, 265–276. doi: 10.1007/s00253-009-2233-z

Simon, C., Herath, J., Rockstroh, S., and Daniel, R. (2009). Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Appl. Environ. Microbiol.* 75, 2964–2968. doi: 10.1128/AEM.02644-08

Sirim, D., Wagner, F., Wang, L., Schmid, R. D., and Pleiss, J. (2011). The Laccase Engineering Database: a classification and analysis system for laccases and related multicopper oxidases. *Database* 2011, bar006. doi: 10.1093/database/bar006

Sjostrom, S. L., Bai, Y., Huang, M., Liu, Z., Nielsen, J., Joensson, H. N., et al. (2014). High-throughput screening for industrial enzyme production hosts by droplet microfluidics. *Lab Chip* 14, 806–813. doi: 10.1039/C3LC51202A

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960. doi: 10.1093/bioinformatics/bti125

Suenaga, H., Ohnuki, T., and Miyazaki, K. (2007). Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environ. Microbiol.* 9, 2289–2297. doi: 10.1111/j.1462-2920.2007.01342.x

Steele, H. L., Jaeger, K.-E., Daniel, R., and Streit, W. R. (2009). Advances in recovery of novel biocatalysts from metagenomes. *J. Mol. Microbiol. Biotechnol.* 16, 25–37. doi: 10.1159/000142892

Tartar, A., Wheeler, M. M., Zhou, X., Coy, M. R., Boucias, D. G., and Scharf, M. E. (2009). Parallel metatranscriptome analyses of host and symbiont gene expression in the gut of the termite *Reticulitermes flavipes*. *Biotechnol. Biofuels* 2, 25. doi: 10.1186/1754-6834-2-25

Tasse, L., Bercovici, J., Pizzut-Serin, S., Robe, P., Tap, J., Klopp, C., et al. (2010). Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *Genome Res.* 20, 1605–1612. doi: 10.1101/gr.108332.110

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41. doi: 10.1186/1471-2105-4-41

Taupp, M., Mewis, K., and Hallam, S. J. (2011). The art and design of functional metagenomic screens. *Curr. Opin. Biotechnol.* 22, 465–472. doi: 10.1016/j.copbio.2011.02.010

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics—a guide from sampling to data analysis. *Microb. Inform. Exp.* 2, 3. doi: 10.1186/2042-5783-2-3

Tirawongsaroj, P., Sriprang, R., Harnpicharnchai, P., Thongaram, T., Champreda, V., Tanapongpipat, S., et al. (2008). Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. *J. Biotechnol.* 133, 42–49. doi: 10.1016/j.jbiotec.2007.08.046

Uchiyama, T., Abe, T., Ikemura, T., and Watanabe, K. (2004). Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat. Biotechnol.* 23, 88–93. doi: 10.1038/nbt1048

Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* 20, 616–622. doi: 10.1016/j.copbio.2009.09.010

Uchiyama, T., and Watanabe, K. (2008). Substrate-induced gene expression (SIGEX) screening of metagenome libraries. *Nat. Protoc.* 3, 1202–1212. doi: 10.1038/nprot.2008.96

van Elsas, J. D., Costa, R., Jansson, J., Sjöling, S., Bailey, M., Nalin, R., et al. (2008). The metagenomics of disease-suppressive soils—experiences from the METACONTROL project. *Trends Biotechnol.* 26, 591–601. doi: 10.1016/j.tibtech.2008.07.004

Van Hellemond, E. W., Janssen, D. B., and Fraaije, M. W. (2007). Discovery of a novel styrene monooxygenase originating from the metagenome. *Appl. Environ. Microbiol.* 73, 5832–5839. doi: 10.1128/AEM.02708-06

Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., van Elsas, J. D., et al. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol.* 7, 252–252. doi: 10.1038/nrmicro2119

Warnecke, F., and Hess, M. (2009). A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J. Biotechnol.* 142, 91–95. doi: 10.1016/j.jbiotec.2009.03.022

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565. doi: 10.1038/nature06269

Waschkowitz, T., Rockstroh, S., and Daniel, R. (2009). Isolation and Characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Appl. Environ. Microbiol.* 75, 2506–2516. doi: 10.1128/AEM.02136-08

Weckx, S., Van der Meulen, R., Allemeersch, J., Huys, G., Vandamme, P., Van Hummelen, P., et al. (2010). Community dynamics of bacteria in sourdough fermentations as revealed by their metatranscriptome. *Appl. Environ. Microbiol.* 76, 5402–5408. doi: 10.1128/AEM.00570-10

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5:e16. doi: 10.1371/journal.pbio.0050016

Zanaroli, G., Balloi, A., Negroni, A., Daffonchio, D., Young, L. Y., and Fava, F. (2010). Characterization of the microbial community from the marine sediment of the Venice lagoon capable of reductive dechlorination of coplanar polychlorinated biphenyls (PCBs). *J. Hazard. Mater.* 178, 417–426. doi: 10.1016/j.jhazmat.2010.01.097

Zaprasis, A., Liu, Y.-J., Liu, S.-J., Drake, H. L., and Horn, M. A. (2009). Abundance of novel and diverse tfda-like genes, encoding putative phenoxyalkanoic acid herbicide-degrading dioxygenases, in soil. *Appl. Environ. Microbiol.* 76, 119–128. doi: 10.1128/AEM.01727-09

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

MINI REVIEW
published: 22 September 2015
doi: 10.3389/fmicb.2015.01018

Targeted metagenomics unveils the molecular basis for adaptive evolution of enzymes to their environment

*Hikaru Suenaga**

Bioproduction Research Institute – National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

Microorganisms have a wonderful ability to adapt rapidly to new or altered environmental conditions. Enzymes are the basis of metabolism in all living organisms and, therefore, enzyme adaptation plays a crucial role in the adaptation of microorganisms. Comparisons of homology and parallel beneficial mutations in an enzyme family provide valuable hints of how an enzyme adapted to an ecological system; consequently, a series of enzyme collections is required to investigate enzyme evolution. Targeted metagenomics is a promising tool for the construction of enzyme pools and for studying the adaptive evolution of enzymes. This perspective article presents a summary of targeted metagenomic approaches useful for this purpose.

Keywords: targeted metagenomics, enzyme adaptation, environmental microbiology, directed evolution, high-throughput screening

OPEN ACCESS

Edited by:
Roy D. Sleator,
Cork Institute of Technology, Ireland

Reviewed by:
Suleyman Yildirim,
Istanbul Medipol University, Turkey
Marla Trindade,
University of the Western Cape, South Africa

***Correspondence:**
Hikaru Suenaga,
Bioproduction Research Institute – National Institute of Advanced Industrial Science and Technology (AIST), Central 6, 1-1-1 Higashi, Tsukuba, Japan
suenaga-hikaru@aist.go.jp

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *04 June 2015*
Accepted: *08 September 2015*
Published: *22 September 2015*

Citation:
Suenaga H (2015) Targeted metagenomics unveils the molecular basis for adaptive evolution of enzymes to their environment. Front. Microbiol. 6:1018. doi: 10.3389/fmicb.2015.01018

Introduction

Enzymes are the driving force behind life since they catalyze the biochemical reactions, and hence the metabolism, of all living organisms. Enzymes have evolved and been optimized for the metabolic networks of individual species (Copley, 2012). The pressure of survival at the metabolic level allows organisms to adapt to a changing chemical environment, such as the ability of bacteria to degrade xenobiotic compounds (Portnoy et al., 2011). There are many reports that microbes adapt to changes in their environment by improving their ability to degrade natural or xenobiotic compounds, and degradation enzymes play a crucial role in these adaptation mechanisms (Janssen et al., 2005). Therefore, in order to understand the ability of microorganisms to adapt rapidly to a new environment, it is necessary to understand how enzymes evolve to make this adaptation possible.

Comparison of the sequence and activity of enzymes from the same family but from different organisms indicates that enzymes are derived from a common ancestor and have accumulated mutations that allow them to adapt to environmental pressures. A collection or pool of related enzymes must be studied to understand enzyme evolution. There are two approaches for obtaining these specific enzyme pools: (i) construct the pool by directed evolution in the laboratory or (ii) retrieve the enzymes from the natural environment. Directed evolution, first used 20 years ago, mimics natural evolutionary processes (Stemmer, 1994; Dalby, 2011), allows the artificial evolution of enzymes in the laboratory under controlled selection pressures, and has resulted in the identification of different adaptive mechanisms (Arnold, 2001). Another approach is to isolate enzymes from microorganisms that show a specific enzymatic activity. For example, various homologous genes involved in the degradation of aromatic compounds have repeatedly been

identified in microorganisms isolated from aromatics-contaminated environments (Furukawa et al., 2004; Vilchez-Vargas et al., 2010). These gene collections can also be useful for investigating molecular mechanisms in the adaptive evolution of xenobiotic-degrading enzymes and bacteria in the natural environment. However the majority of microorganisms in natural environments cannot be cultured using readily available technologies (Amann et al., 1995; Quince et al., 2008). This has spurred the development of metagenomics, which allows us to obtain various genes of interest from the entire microbial community (Handelsman, 2004; Shade et al., 2012). Metagenomics is, therefore, a powerful tool for constructing comprehensive gene collections of specific groups of enzymes from microbes in various habitats. This collection is useful for studying the adaptive evolution of enzymes and their host microorganisms.

Two Strategies for Metagenomics

Metagenomics approaches are roughly classified into two groups: (i) whole metagenomics and (ii) targeted metagenomics, and are based on random and selective sequencing strategies, respectively. Many projects based on the random sequencing of microbial domains, such as the bacteria and archaea, and of viruses, have been reported (Thomas et al., 2012; Sharpton, 2014). Although whole metagenomic analyses revealed that microbial communities are well adapted to their geochemical conditions, those analyses provided no definitive evidence for the positive selection of enzymes for key ecological processes under environmental pressures. This lack of evidence is likely due to insufficient sequence data for the target enzyme group (Hemme et al., 2010). Mutations in the genes encoding such key enzymes would provide an adaptive phenotype optimized for a specific niche (Chattopadhyay et al., 2013). Therefore, high-resolution metagenomic sequencing to collect data of sufficient breadth and depth for any particular gene is necessary to verify the adaptive processes of enzymes in their ecosystem. This "targeted metagenomics" approach would be a suitable tool for constructing gene collections of specific groups of enzymes which are useful for studying their adaptive evolution. Previously, we presented a summary of the targeted metagenomics approaches to understanding the composition of gene clusters for key ecological processes in microbial communities (Suenaga, 2012). In this review, we focus on targeted metagenomics studies for surveying the adaptive evolution of enzymes toward environmental changes.

Strategies for Targeted Metagenomics

In a targeted metagenomics approach, a deliberately selected DNA pool is sequenced. The selection process is usually based on (i) sequence-driven screening or (ii) function-driven screening. By focusing efforts on selective sequence analysis, targeted metagenomics can provide broad coverage and extensive redundancy of sequences for targeted genes and reveal specific genome areas directly linked to an ecological function, even at low abundances within a metagenome (Suenaga, 2012). Better sequence coverage of the obtained target metagenomics can be beneficial for genome assembly and subsequent data analysis. Examples of studies on targeted metagenomics are summarized below.

Targeted Metagenomics Based on Sequence-driven Screening

The PCR-based approach has been used extensively to retrieve specific genes from a pool of DNA. Instead of cloning all the extracted DNA, primers are designed specifically against an identified target gene, such as phenol hydroxylase (Futamata et al., 2001), catechol 2,3-dioxygenase (Mesarch et al., 2000), and methane monooxygenase (Henckel et al., 2000). The advantage of using sequence-driven screening is that it uses well-established and high-throughput techniques, such as PCR and hybridization, and can be used for different targets. On the other hand, this approach requires designing DNA probes and primers derived from conserved regions of known gene or protein families. Thus, already-known sequence types will be identified and only a fragment of the main target gene will be amplified. Despite this limitation, combining PCR detection of small conserved regions with genome sequencing/walking at flanking regions makes it possible to obtain the entire gene and thus reconstruct the evolution of the target enzymes in response to alterations in the ecosystem.

Dissimilatory sulfate reduction is a crucial process in the mineralization of organic matter in marine sediments. PCR screening of a metagenomic fosmid library (11,000 clones) using degenerate primers resulted in the identification of three fosmid DNA fragments harboring a core set of essential genes for dissimilatory sulfate reduction; these fragments contained genes associated with the reduction of sulfur intermediates (*dsrAB* gene) and the synthesis of the prosthetic group of dissimilatory sulfate reductase (*aprA* gene; Mussmann et al., 2005). Complete sequence analysis of all fosmid inserts revealed the genomic context of the key enzymes of dissimilatory sulfate reduction as well as novel genes functionally involved in sulfate respiration in their flanking regions. The results support the hypothesis that the set of genes responsible for dissimilatory sulfate reduction was concomitantly transferred in a single event among prokaryotes.

Denitrification is a microbial respiratory process within the nitrogen cycle responsible for the return of fixed nitrogen to the atmosphere. A sequence-driven screening (colony hybridization) of 77,000 clones from a soil metagenomic library led to the identification of positive clones, and subsequent sequencing analysis revealed nine denitrification gene clusters (Ginolhac et al., 2004; Demanèche et al., 2009). This targeted metagenomics study indicated that the gene clusters involved in denitrification were probably subject to shuffling by endogenous gene displacement or by horizontal gene transfer between bacteria.

Targeted Metagenomics Based on Function-driven Screening

Function-driven screening strategies potentially provide a means of revealing undiscovered genes or gene families that cannot be detected by sequence-driven approaches, although this screening is more laborious than sequence-based screening procedures (Ferrer et al., 2005; Fernández-Arrojo et al., 2010).

Nitrilases are important in synthesis and degradation for nitriles which are attractive starting compounds in the synthesis of fine chemicals. However, nitrilase genes are quite rare in bacterial genomes, and fewer than 20 were reported in the scientific and patent literature prior to the application of metagenomics (Podar et al., 2005). A leading metagenome company, Diversa Co. (USA), reported that 651 environmental samples collected worldwide from terrestrial and aquatic microenvironments were used to construct a metagenomics library, allowing identification of 137 new nitrilases by visual observation of *Escherichia coli* cells grown in liquid medium supplemented with nitrile substrate (Robertson et al., 2004). Phylogenetic analysis and enzymatic characterization of these enzymes revealed important correlations between sequence clades and selective properties of three structurally distinct nitrile substrates. Together with other metagenomic surveys for nitrilases (DeSantis et al., 2002; Bayer et al., 2011), the metagenomics approach has helped reveal the ecological distribution and diversity of nitrilases.

Deep-sea areas require that microbial communities adapt to harsh physical conditions, particularly high salinity and high pressure (Daffonchio et al., 2006; Smedile et al., 2013). A set of eight different enzymes was screened for activity from metagenomic fosmid and phage libraries constructed using DNA from five distinct deep-sea environments (Alcaide et al., 2015). The activities of the purified metagenomic proteins were characterized at various temperatures and salt conditions. The results suggested that adaptation to high pressure is linked to high thermal resistance in salt-saturated deep-sea conditions. Therefore, salinity might increase the temperature window for enzyme activity, and possibly microbial growth, in deep-sea habitats.

Extradiol dioxygenases (EDOs) are enzymes that play an important role in the catabolism of aromatic compounds (Sipilä et al., 2008; Brennerova et al., 2009), cleaving the aromatic ring of catechol compounds, which are common intermediates in the aerobic microbial degradation of natural and xenobiotic aromatic compounds (Furukawa et al., 2004). Based on the activity of EDO enzymes, 96,000 fosmid clones were screened, and subsequent sequencing of positive fosmids led to the identification of 43 novel EDO genes (Suenaga et al., 2007, 2009). Using combinations of single nucleotide polymorphisms (SNPs), a possible evolutionary lineage of the EDO genes was constructed (**Figure 1**) and suggested that these genes evolved from a common ancestor (group 1 and 3), then diverged through the accumulation of various nucleotide mutations. Furthermore, investigation of the kinetic properties and thermal stability of the purified EDO enzymes showed an apparent trade-off between activity and stability (**Figure 1**). Bloom et al. (2006) reported that cytochrome P450 BM3 mutants with higher stabilities were more likely to acquire new or improved functions through random mutagenesis. They concluded that protein stability promotes adaptive protein evolution. Similarly, in EDO enzymes, the most thermostable ancestral groups (group 1 and 3) may have evolved toward more active groups (group 2 through group 5 and 6) by sacrificing thermostability. Note that EDO enzymes that had acquired higher activities (group 2 and 5) were more frequently discovered in the retrieved EDO clones, likely reflecting the allele frequencies in the environment.

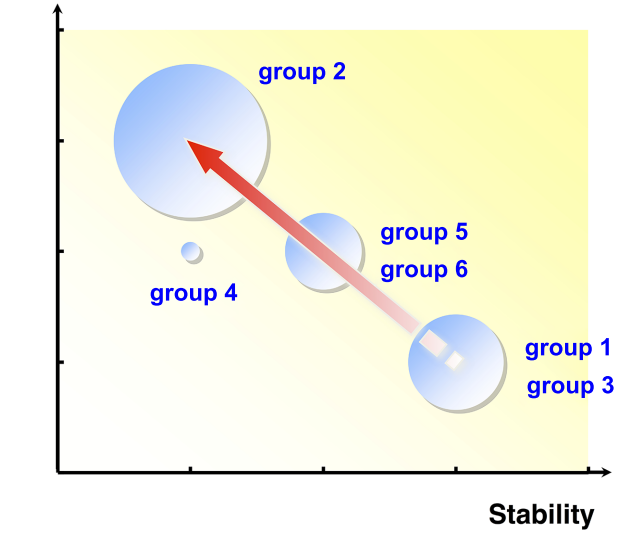

FIGURE 1 | The relationship between activity and thermostability of purified metagenomic extradiol dioxygenase (EDO) enzymes. The size of each circle is proportional to the number of EDO enzymes in the group. The arrow indicates the proposed genetic evolutionary pathway. The thermostable ancestral groups, group 1 and 3, may have adaptively evolved toward the more active group 2 via group 5 and 6 by sacrificing unessential thermostability. EDO enzymes that acquired higher activities (group 2) were more frequently discovered in the retrieved enzyme collection.

The above studies of marine enzymes and EDO enzymes incorporated three-dimensional structural analyses to unveil the molecular mechanisms of enzyme adaptation, but the structural basis for enzyme evolution remains unclear. The amount of data on enzyme diversity made available by metagenomic approaches exceeds our ability to analyze the data based on our current knowledge of protein structure/function.

Future Perspective

In the Section "Introduction", I stated that directed evolution and metagenomics are different approaches for creating enzyme pools that can provide valuable hints on how enzymes adapt to ecological conditions. However, both approaches use the same key technology: high-throughput screening to collect the target enzymes. A variety of high-throughput screening methods have been established in recent years, and continue to develop in

step with new developments in robotics, analytical devices, and visualizing assays. For example, microarray-based technologies coupled with microfluidic devices, cell compartmentalization, flow cytometry, and cell sorting have been proposed as promising new tools (Tracy et al., 2010; Simon and Daniel, 2011; Ekkers et al., 2012; Zhou et al., 2015). These screening systems offer higher levels of quantification and the possibility to detect multiple traits in one assay. Researchers in the two fields can share their wide knowledge of enzymes and up-to-date technologies to analyze enzyme characteristics.

Environmental pressures led to today's diverse enzymes distributed throughout the earth's ecosystems. Therefore, the collection of metagenomic enzyme pools from extreme environments, such as deep-sea hydrothermal vent fields, contaminated sites, and hot springs, is effective for studying the adaptive evolution of enzymes and their host microorganisms. In the near future, by integrating scientific knowledge in environmental microbiology, enzymology, and geology, it will be possible to assemble and use good quality enzyme collections suitable for the analysis of enzyme evolution.

Acknowledgment

This work was performed as part of a project supported by JSPS Grant.

References

Alcaide, M., Stogios, P. J., Lafraya, , Á, Tchigvintsev, A., Flick, R., Bargiela, R., et al. (2015). Pressure adaptation is linked to thermal adaptation in salt-saturated marine habitats. *Environ. Microbiol.* 17, 332–345. doi: 10.1111/1462-2920.12660

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59, 143–169.

Arnold, F. H. (2001). How proteins adapt: lessons from directed evolution. *Trends Biochem. Sci.* 26, 100–106. doi: 10.1101/sqb.2009.74.046

Bayer, S., Birkemeyer, C., and Ballschmiter, M. (2011). A nitrilase from a metagenomic library acts regioselectively on aliphatic dinitriles. *Appl. Microbiol. Biotechnol* 89, 91–98. doi: 10.1007/s00253-010-2831-9

Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5869–5874. doi: 10.1073/pnas.0510098103

Brennerova, M. V., Josefiova, J., Brenner, V., Pieper, D. H., and Junca, H. (2009). Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. *Environ. Microbiol.* 11, 2216–2227. doi: 10.1111/j.1462-2920.2009.01943.x

Chattopadhyay, S., Taub, F., Paul, S., Weissman, S. J., and Sokurenko, E. V. (2013). Microbial variome database: point mutations, adaptive or not, in bacterial core genomes. *Mol. Biol. Evol.* 30, 1465–1470. doi: 10.1093/molbev/mst048

Copley, S. D. (2012). Toward a systems biology perspective on enzyme evolution. *J. Biol. Chem.* 287, 3–10. doi: 10.1074/jbc.R111.254714

Daffonchio, D., Borin, S., Brusa, T., Brusetti, L., van der Wielen, P. W. J. J., Bolhuis, H., et al. (2006). Stratified prokaryote network in the oxic-anoxic transition of a deep-sea halocline. *Nature* 440, 203–207. doi: 10.1038/nature04418

Dalby, P. A. (2011). Strategy and success for the directed evolution of enzymes. *Curr. Opin. Struct. Biol.* 21, 473–480. doi: 10.1016/j.sbi.2011.05.003

Demanèche, S., Philippot, L., David, M. M., Navarro, E., Vogel, T. M., and Simonet, P. (2009). Characterization of denitrification gene clusters of soil bacteria via a metagenomic approach. *Appl. Environ. Microbiol.* 75, 534–537. doi: 10.1128/AEM.01706-08

DeSantis, G., Zhu, Z., Greenberg, W. A., Wong, K., Chaplin, J., Hanson, S. R., et al. (2002). An enzyme library approach to biocatalysis: development of nitrilases for enantioselective production of carboxylic acid derivatives. *J. Am. Chem. Soc.* 124, 9024–9025. doi: 10.1021/ja0259842

Ekkers, D. M., Cretoiu, M. S., Kielak, A. M., and van Elsas, J. D. (2012). The great screen anomaly - a new frontier in product discovery through functional metagenomics. *Appl. Microbiol. Biotechnol.* 93, 1005–1020. doi: 10.1007/s00253-011-3804-3

Fernández-Arrojo, L., Guazzaroni, M.-E., López-Cortés, N., Beloqui, A., and Ferrer, M. (2010). Metagenomic era for biocatalyst identification. *Curr. Opin. Biotechnol.* 21, 725–733. doi: 10.1016/j.copbio.2010.09.006

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Reyes-Duarte, D., Santos, V. A., et al. (2005). Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ. Microbiol.* 7, 1996–2010. doi: 10.1111/j.1462-2920.2005.00920.x

Furukawa, K., Suenaga, H., and Goto, M. (2004). Biphenyl dioxygenases: functional versatilities and directed evolution. *J. Bacteriol.* 186, 5189–5196. doi: 10.1128/JB.186.16.5189

Futamata, H., Harayama, S., and Watanabe, K. (2001). Group-specific monitoring of phenol hydroxylase genes for a functional assessment of phenol-stimulated trichloroethylene bioremediation. *Appl. Environ. Microbiol.* 67, 4671–4677. doi: 10.1128/AEM.67.10.4671-4677.2001

Ginolhac, A., Jarrin, C., Gillet, B., Robe, P., Pujic, P., Tuphile, K., et al. (2004). Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl. Environ. Microbiol.* 70, 5522–5527. doi: 10.1128/AEM.70.9.5522-5527.2004

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MBR.68.4.669

Hemme, C. L., Deng, Y., Gentry, T. J., Fields, M. W., Wu, L., Barua, S., et al. (2010). Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J.* 4, 660–672. doi: 10.1038/ismej.2009.154

Henckel, T., Jäckel, U., Schnell, S., and Conrad, R. (2000). Molecular analyses of novel methanotrophic communities in forest soil that oxidize atmospheric methane molecular analyses of novel methanotrophic communities in forest soil that oxidize atmospheric methane. *Appl. Environ. Microbiol.* 66, 1801–1808. doi: 10.1128/AEM.66.5.1801-1808.2000

Janssen, D. B., Dinkla, I. J. T., Poelarends, G. J., and Terpstra, P. (2005). Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environ. Microbiol.* 7, 1868–1882. doi: 10.1111/j.1462-2920.2005.00966.x

Mesarch, M. B., Nakatsu, C. H., and Nies, L. (2000). Development of catechol 2,3-dioxygenase-specific primers for monitoring bioremediation by competitive quantitative PCR. *Appl. Environ. Microbiol.* 66, 678–683. doi: 10.1128/AEM.66.2.678-683.2000

Mussmann, M., Richter, M., Lombardot, T., Meyerdierks, A., Kuever, J., Kube, M., et al. (2005). Clustered genes related to sulfate respiration in uncultured prokaryotes support the theory of their concomitant horizontal transfer. *J. Bacteriol.* 187, 7126–7137. doi: 10.1128/JB.187.20.7126

Podar, M., Eads, J. R., and Richardson, T. H. (2005). Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. *BMC Evol. Biol.* 5:42. doi: 10.1186/1471-2148-5-42

Portnoy, V. A., Bezdan, D., and Zengler, K. (2011). Adaptive laboratory evolution-harnessing the power of biology for metabolic engineering. *Curr. Opin. Biotechnol.* 22, 590–594. doi: 10.1016/j.copbio.2011.03.007

Quince, C., Curtis, T. P., and Sloan, W. T. (2008). The rational exploration of microbial diversity. *ISME J.* 2, 997–1006. doi: 10.1038/ismej.2008.69

Robertson, D. E., Chaplin, J. A., Desantis, G., Podar, M., Madden, M., Chi, E., et al. (2004). Exploring nitrilase sequence space for enantioselective catalysis exploring nitrilase sequence space for enantioselective catalysis. *Appl. Environ. Mivrobiol.* 70, 2429–2436. doi: 10.1128/AEM.70.4.2429

Shade, A., Hogan, C. S., Klimowicz, A. K., Linske, M., McManus, P. S., and Handelsman, J. (2012). Culturing captures members of the soil rare biosphere. *Environ. Microbiol.* 14, 2247–2252. doi: 10.1111/j.1462-2920.2012.02817.x

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209

Simon, C., and Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* 77, 1153–1161. doi: 10.1128/AEM.02345-10

Sipilä, T. P., Keskinen, A.-K., Åkerman, M.-L., Fortelius, C., Haahtela, K., and Yrjälä, K. (2008). High aromatic ring-cleavage diversity in birch rhizosphere: PAH treatment-specific changes of I.E.3 group extradiol dioxygenases and 16S rRNA bacterial communities in soil. *ISME J.* 2, 968–981. doi: 10.1038/ismej.2008.50

Smedile, F., Messina, E., La Cono, V., Tsoy, O., Monticelli, L. S., Borghini, M., et al. (2013). Metagenomic analysis of hadopelagic microbial assemblages thriving at the deepest part of mediterranean sea. Matapan-Vavilov deep. *Environ. Microbiol.* 15, 167–182. doi: 10.1111/j.1462-2920.2012.02827.x

Stemmer, W. P. C. (1994). DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* 91, 10747–10751. doi: 10.1073/pnas.91.22.10747

Suenaga, H. (2012). Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ. Microbiol.* 14, 13–22. doi: 10.1111/j.1462-2920.2011.02438.x

Suenaga, H., Mizuta, S., and Miyazaki, K. (2009). The molecular basis for adaptive evolution in novel extradiol dioxygenases retrieved from the metagenome. *FEMS Micriobiol. Ecol.* 69, 472–480. doi: 10.1111/j.1574-6941.2009.00719.x

Suenaga, H., Ohnuki, T., and Miyazaki, K. (2007). Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environ. Microbiol.* 9, 2289–2297. doi: 10.1111/j.1462-2920.2007.01342.x

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Inform. Exp.* 2, 1–12. doi: 10.1186/2042-5783-2-3

Tracy, B. P., Gaida, S. M., and Papoutsakis, E. T. (2010). Flow cytometry for bacteria: enabling metabolic engineering, synthetic biology and the elucidation of complex phenotypes. *Curr. Opin. Biotechnol.* 21, 85–99. doi: 10.1016/j.copbio.2010.02.006

Vilchez-Vargas, R., Junca, H., and Pieper, D. H. (2010). Metabolic networks, microbial ecology and 'omics' technologies: towards understanding in situ biodegradation processes. *Environ. Microbiol.* 12, 3089–3104. doi: 10.1111/j.1462-2920.2010.02340.x

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S. G., and Alvarez-cohen, L. (2015). High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *mBio* 6:e2288-14. doi: 10.1128/mBio.02288-14

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 28 August 2015
doi: 10.3389/fmicb.2015.00890

Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates

Marla Trindade[1], Leonardo Joaquim van Zyl[1], José Navarro-Fernández[1,2] and Ahmed Abd Elrazak[1,3]*

[1] Institute for Microbial Biotechnology and Metagenomics, University of the Western Cape, Bellville, South Africa, [2] Centro Regional de Hemodonación, Servicio de Hematología y Oncología Médica, Universidad de Murcia, IMIB-Arrixaca, Murcia, Spain, [3] Botany Department, Faculty of Science, Mansoura University, Mansoura, Egypt

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Brett A. Neilan, The University of New South Wales, Australia
Ute Hentschel, University of Wuerzburg, Germany
Jason Christopher Kwan, University of Wisconsin-Madison, USA

***Correspondence:**
Marla Trindade, Institute for Microbial Biotechnology and Metagenomics, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa ituffin@uwc.ac.za

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *27 April 2015*
Accepted: *17 August 2015*
Published: *28 August 2015*

Citation:
Trindade M, van Zyl LJ, Navarro-Fernández J and Abd Elrazak A (2015) Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. Front. Microbiol. 6:890. doi: 10.3389/fmicb.2015.00890

Microbial natural products exhibit immense structural diversity and complexity and have captured the attention of researchers for several decades. They have been explored for a wide spectrum of applications, most noteworthy being their prominent role in medicine, and their versatility expands to application as drugs for many diseases. Accessing unexplored environments harboring unique microorganisms is expected to yield novel bioactive metabolites with distinguishing functionalities, which can be supplied to the starved pharmaceutical market. For this purpose the oceans have turned out to be an attractive and productive field. Owing to the enormous biodiversity of marine microorganisms, as well as the growing evidence that many metabolites previously isolated from marine invertebrates and algae are actually produced by their associated bacteria, the interest in marine microorganisms has intensified. Since the majority of the microorganisms are uncultured, metagenomic tools are required to exploit the untapped biochemistry. However, after years of employing metagenomics for marine drug discovery, new drugs are vastly under-represented. While a plethora of natural product biosynthetic genes and clusters are reported, only a minor number of potential therapeutic compounds have resulted through functional metagenomic screening. This review explores specific obstacles that have led to the low success rate. In addition to the typical problems encountered with traditional functional metagenomic-based screens for novel biocatalysts, there are enormous limitations which are particular to drug-like metabolites. We also present how targeted and function-guided strategies, employing modern, and multi-disciplinary approaches have yielded some of the most exciting discoveries attributed to uncultured marine bacteria. These discoveries set the stage for progressing the production of drug candidates from uncultured bacteria for pre-clinical and clinical development.

Keywords: uncultured microbes, metagenomics, symbionts, marine natural products, drug discovery, functional screening

Marine Microorganisms as a Novel Source of Natural Products

Natural products remain a major resource for drug production today and during the past 30 years, 70% of antimicrobials and 60% of chemotherapeutics have been developed or analogously synthesized from them (Pomponi, 2001; Grüschow et al., 2011). Traditionally, terrestrial sources have provided the bulk of natural products for drug molecules. However, participation by the major pharmaceutical companies declined in the mid-nineties, largely owing to the high rediscovery rate and decreased number of novel compound identifications (Molinski et al., 2009). In the meantime infectious diseases and multiple drug resistant strains have bloomed, urging scientists to mine for novel drugs in non-terrestrial and unexplored environments. A chemoinformatics study showed that 71% of the marine natural products were not represented in terrestrial natural products, and that 53% have been found only once (Montaser and Luesch, 2011). Complementary studies investigating the distribution of natural products in chemical space has shown clearly that marine natural products have the broadest distribution, covering many drug-relevant areas (Tao et al., 2015). As such, the focus has recently shifted to marine natural product bioprospecting, which has delivered remarkably high hit rates (Gerwick and Moore, 2012; Blunt et al., 2015).

The ocean harbors a number of ecological niches and has proven to be home to more microorganisms than any other environment. Considering that 70% of our planet's surface is covered by the oceans, it is not surprising that certain marine ecosystems harbor much higher biological and chemical diversity than what is found terrestrially. Furthermore, the sedentary lifestyle of many of the organisms necessitates a chemical means of defense, and as such natural products are produced as chemical weapons which have evolved into highly effective inhibitors (Spainhour, 2005). Since the released compounds become rapidly diluted, marine natural products tend to be highly potent in order to be effective (Haefner, 2003). The rich biodiversity contained within the oceans (15 animal phyla exclusive to the oceans) makes them a unique and rich drug discovery reservoir (Leal et al., 2012).

Marine natural product discovery was initially focused on the easily accessible macro-organisms (such as algae, soft corals, and sponges) from which a range of bioactive compounds have been described (Bergmann and Feeney, 1951; McGivern, 2007; Hu et al., 2011; Leal et al., 2012). However, efforts have gradually turned to the smaller forms of life such as bacteria and fungi (Gerwick and Moore, 2012) which constitute a large portion of the marine biomass (Sogin et al., 2006). Considering the enormous number of microbes, their vast metabolic diversity and the rate of mutations during the past 3.5 billion years, it is expected that there are high levels of genetic and phenotypic variation in marine environments (Sogin et al., 2006). Furthermore, marine microorganisms live in a biologically competitive environment with unique, harsh, and fluctuating conditions. They encounter enormous physical and chemical variability including low temperature, high pressure, oligotrophy, high salinity and other competitive environments, and are especially rich in chlorine and bromine elements. Global scale analyses of bacterial diversity identify environment salinity and temperature as the major determinants of microbial community composition, resulting in distinct marine microbiota being selected (Lozupone and Knight, 2007). Biofilm formation is a crucial aspect where cell densities are typically 100–1000 fold higher in a biofilm assemblage than in the surrounding water column (Wahl et al., 2012). Furthermore, the increased competition amongst organisms is thought to be the source of higher production levels of secondary metabolites (Teasdale et al., 2009). In contrast to typical terrestrial environments, marine environments have a very high bacterial diversity at the higher taxonomic levels and a global biogeographical study has shown that there is no more than 12% taxon overlap between bacterial assemblages within and between habitat types (Nemergut et al., 2011). As a result marine microorganisms represent a unique source of genetic information and biosynthetic capacity which translates to huge chemical diversity.

Marine Microbial Natural Products

Marine microorganisms produce a vast variety of secondary metabolites which could be used to supply the starved pharmaceutical market. Microbial natural products have notable potent therapeutic activities, and also often possess the desirable pharmacokinetic properties required for clinical development (Farnet and Zazopoulos, 2005). More than half of the known natural products with anti-microbial, anti-tumor (Bewley and Faulkner, 1998; Feling et al., 2003; Taori et al., 2008; Rath et al., 2011) or anti-viral activity are of bacterial origin (Berdy, 2005). Additional categories include anti-parasitic (Kirst et al., 2002; Abdel-Mageed et al., 2010), anti-nematodal (Donia and Hamann, 2003), anti-inflammation (Strangman, 2007), and neurological (Sudek et al., 2007). Pharmaceutically relevant natural products belong to different chemical classes that differ not only in structure, but also in the mechanisms by which they are synthesized. The molecular classes which become pharmaceuticals tend to be alkaloids, terpenoids, polyketides and small peptides, and a wide range of bioactive properties are observed within each class (Graça et al., 2013). Furthermore, the elucidation of novel hybrid compounds is providing deeper insights into fascinating enzyme assemblies and mechanisms behind the diversity in structure and biological functions observed in these compounds. Some marine derived microbial examples can be found in the following references: alkaloids (Charan et al., 2004; Abdelmohsen et al., 2012); terpenoids (Kuzuyama and Seto, 2003; Cho et al., 2006; Strangman, 2007; Solanki et al., 2008); polyketides (Olano et al., 2009; Harunari et al., 2014), peptides (Pettit et al., 2009; Chopra et al., 2014); and hybrids (Hardt et al., 2000; Feling et al., 2003; Oh et al., 2007; Blunt et al., 2015).

An additional attraction of microbially derived natural products is that they offer an answer to the supply problem, a major bottleneck in the drug discovery pipeline. The progression of many marine natural products with promising

pharmaceutical relevance into clinical phases are halted since the clinical trial stage requires a considerable amount of drug mass; usually kilogram amounts (Tsukimoto et al., 2011). Most pharmaceutically interesting compounds are found in minute amounts, therefore bioprospecting cannot rely on wild-harvesting as it could lead to the extinction of marine species. More economically feasible, environmentally friendly, and sustainable sources of lead compounds are required. Microbial-based production of lead compounds therefore offers a sustainable solution through the use of culturable marine microorganisms (microbial fermentation). Marine bacteria can respond positively during scaling up processes, and can incorporate sustainable chemical processes for faster establishment of a pilot plant for production (Abd Elrazak et al., 2013). The current industrial process for the production of Yondelis, for example, involves the fermentation of *Pseudomonas fluorescens* for the production of the starting material cyanosafracin B, followed by semi-synthesis to generate the final drug (Cuevas et al., 2000). Furthermore, strain intensification and elicitation to improve expression are possible through metabolic engineering, as well as the unlocking of untapped cryptic biosynthetic pathways through heterologous host expression (Li and Neubauer, 2014).

Marine Metagenomics

There is remarkable potential harbored within microorganisms to produce diverse drug-like small molecules for a wide range of applications. The impact and possible success of a single new discovery distinguishes natural products from all other sources of chemical diversity (Farnet and Zazopoulos, 2005). However, traditional culture-based approaches used to identify microbial metabolites likely miss the vast majority of bacterial natural products. Only about 1% of bacteria are cultured *in vitro* and of the approximately 61 bacterial phyla known, 31 lack cultivable representatives (Vartoukian et al., 2010). Seawater bacteria have a 10-fold lower representation of cultured isolates compared to other environments (Amann et al., 1995). Therefore, if the natural products discovered from cultured marine bacteria are an indication of the diversity available, culture-independent approaches are expected to more successfully access the untapped reservoir of chemical diversity and contribute many more novel marine-derived discoveries.

The study of DNA obtained directly from an environmental sample (metagenomics) accesses the collective genomes and bioactive potential of bacterial consortia (Handelsman, 2004). Metagenomics therefore provides a means of exploring novel metabolites from bacteria that are known to be present in marine environments but which remain recalcitrant to culturing (Banik and Brady, 2010). Moreover, metagenomics is particularly attractive for natural product discovery because the genetic information encoding the activities of interest are generally clustered on bacterial genomes, making it possible to clone an entire pathway on an individual or at least a small number of overlapping library clones (Handelsman et al., 1998; Banik and Brady, 2010). Therefore, high throughput metagenomic screening approaches, using both sequence-based and function-based screening, can be employed, in theory, to de-replicate known pathways and compounds and reduce the high degree of redundancy obtained through traditional culture based approaches. Metagenomic screening approaches cover a large range of techniques and are subject to the specifications of the target compound. The particular focus of this review is to evaluate the impact of function-guided strategies as a tool in marine natural product discovery. Specifically, we compare two different functional approaches and their contributions to unlocking the natural product potential harbored in marine microbial genomes.

Classic Functional Metagenomic Screening

In natural product discovery, classical functional screening involves the generation and subsequent screening of metagenomic libraries for the direct detection of the metabolite's properties (e.g., antibacterial, antifungal, antitumor, antiviral activity; Rocha-Martin et al., 2014; **Figure 1**). Using whole cells, the culture supernatant or cell pellet extract, this screening approach has been employed with some success. One of the simplest strategies is to test for growth inhibition against a test microbe in top agar overlay assays. This has led to the characterization of a variety of new antibiotics from soil-derived environments (Brady and Clardy, 2000, 2005; Curtois et al., 2003), but no marine-derived studies have been reported, to our knowledge. A more typical approach to functional screening is to screen for a readily detectable phenotype which is representative of the desired bioactive compound, either through the visual detection of pigment production or the use of chromogenic and fluorogenic enzyme substrates which allow the detection of specific catalytic functions encoded on individual clones when incorporated into the growth medium (Ferrer et al., 2009; Guazzaroni et al., 2015). The antibacterial pigments violacein (Brady et al., 2001), indigo (Lim et al., 2005), and turbomycins (Gillespie et al., 2002) have been isolated from soil metagenomic libraries. Success with marine libraries; however, has not been reported. A number of other function based screens have yielded a range of different bioactive compounds or activities. Although these screens have not been employed in marine library screening, they are worthy of mention because we expect it is only a matter of time before they are reported. An acylhomoserine lactone synthase promoter fused to a *lacZ* reporter has been employed to identify AHL lactonases capable of inhibiting *Pseudomonas aeruginosa* biofilms (Schipper et al., 2009). A phosphopantetheinyl transferase (PPTase)-targeting functional screen has resulted in the efficient recovery of natural product gene clusters from metagenomic libraries (Owen et al., 2012). Non-ribosomal peptide synthetase and polyketide synthase (PKS) enzymes are activated by PPTases, therefore these enzymes are frequently associated with secondary metabolite gene clusters (Osbourn, 2010; Owen et al., 2012). There is a much greater chance of detecting the expression of a single intact gene than an entire biosynthetic operon, therefore focusing on only a single gene target for the recovery of NRPS and PKS gene clusters

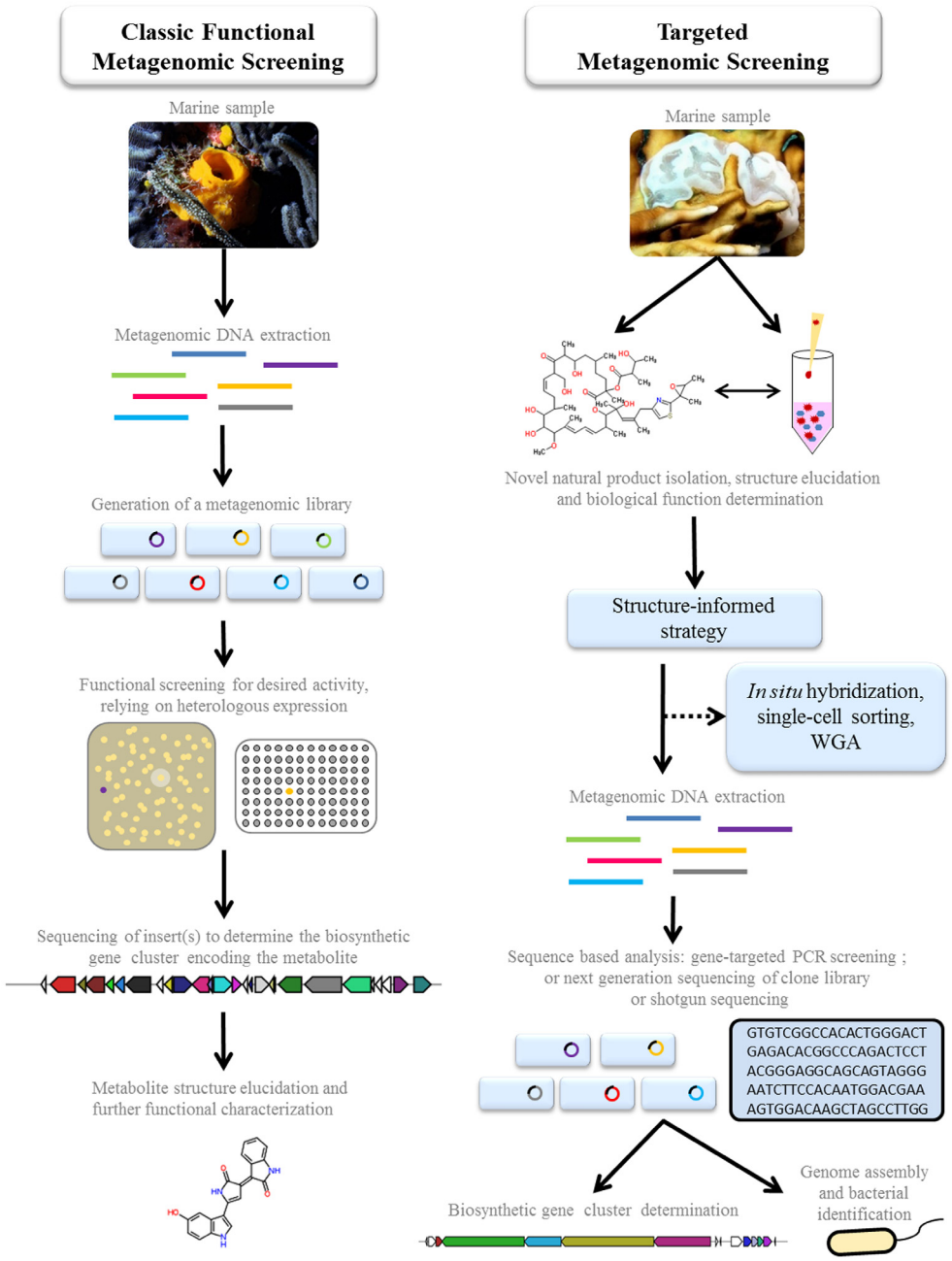

FIGURE 1 | A comparison of two function-driven approaches to employ metagenomics for the discovery and production of pharmaceutically relevant marine natural products. Classic functional metagenomic screening: metagenomic libraries are generated in a suitable host and activity screened in a variety of ways, to detect clones expressing metabolites with potential therapeutic properties. The active clones are sequenced to determine the biosynthetic pathway. For certain classes of secondary metabolites, sequence from overlapping clones may be required to compose the entire pathway. The structure of the expressed metabolite is elucidated, following chemical dereplication and characterization methods. If the metabolite is novel, further functional characterization is conducted to evaluate its therapeutic potential. **Targeted metagenomic screening**: these strategies are guided by traditional chemistry and structure/function-based discoveries in which novel natural products are first isolated and characterized directly from the marine organism or environment. Guided by the chemical classification, a targeted sequence-based analysis can be employed to identify whether the metabolite is microbially encoded, and to subsequently describe the biosynthetic gene cluster. This approach has been employed successfully (detailed in text) when integrated with a number of technologies such as *in situ* hybridization, single-cell sorting, and whole genome amplification (WGA). The sequence-based analysis of the metagenomic DNA can include gene-targeting using degenerate PCR amplification; or next generation sequencing of the clone library or of the metagenomic DNA directly (shotgun). Where sufficient sequence information is assembled, full genome information can be used to describe novel and uncultured bacteria. The elucidation of the genetic clusters provides the foundation for direct production of the pharmaceutical drug and new analogs through metabolic engineering, and opens the possibility to produce the drugs through heterologous expression.

by association increases the chances of identifying "hits" (Owen et al., 2012).

Function-driven screening strategies offer significant advantages to sequence/homology based screening (Tuffin et al., 2009; Kennedy et al., 2010; Suenaga, 2012). This is primarily due to the fact that prior knowledge of the gene sequence for the target activity of interest is not needed, and as a result it is expected that functional screening increases the 'novelty' hit rate. This increases the potential of identifying entirely new classes of genes for both known and novel functions (Sharma and Vakhlu, 2014). Furthermore, the hits obtained represent an "insurance policy"; guaranteed success of expression in the heterologous host, enabling one to screen for particular properties and under specified conditions, as well as facilitating downstream analyses. The dearth of marine natural product discoveries through functional metagenomics is puzzling considering the increased research focus on marine microorganisms over the last decade (Kennedy et al., 2010). We propose two major reasons for this, (i) heterologous expression challenges and (ii) the sequence technology boom.

Challenges Associated with Classic Functional Metagenomics

Natural product discovery, using metagenomics, faces a number of significant challenges and limitations when employing classic functional screening approaches (Kennedy et al., 2010; Li and Neubauer, 2014; Reen et al., 2015). The most well-known are those associated with heterologous gene expression. Gabor et al. (2004) estimated using *in silico* analysis that only 40% of enzymatic activities can be identified by random cloning of environmental DNA in *Escherichia coli*. Many studies have highlighted heterologous expression as an enormous challenge limiting the robustness of metagenomics to fully access metabolic potential (Ferrer et al., 2009; Uchiyama and Miyazaki, 2009; Reen et al., 2015). In natural product discovery, these challenges are augmented for a number of reasons.

(i) Unlike for other biotechnologically important enzymes and activities typically screened in metagenomic studies, such as the glycosyl hydrolases for example, the activities encoded by particularly the PKS and NRPS pathways, require optimal induction conditions of many genes for expression. The enzymatic megacomplexes for dedicated synthesis of their cognate products are encoded by massive gene clusters, some composed of over 20 genes which are distributed between multiple polycistronic transcriptional units (Gao et al., 2010; Osbourn, 2010). Obviously there is a much lower chance of expressing an entire biosynthetic pathway in any given heterologous host than a single active enzyme. Secondary metabolite pathways are regulated by pathway specific proteins as well as global regulatory elements in response to changes in nutrient conditions or environmental signals (Van Wezel and McDowall, 2011). The extremely diverse marine specific factors responsible for unique biochemistries are difficult to replicate in functional screening. For example, it is well-understood that many secondary metabolite pathways expressed in their natural environmental conditions remain silent under laboratory conditions (Montaser and Luesch, 2011), and this is magnified in heterologous systems. The synergies associated with complex symbiotic and competitive interactions cannot easily be incorporated in simple expression systems.

(ii) Even if heterologous expression of a particular pathway is successful, it may not necessarily produce the same compound. Only one isomer may be active and not the other due to the requirement of intermediate compound(s) from the original host or environment (Taylor et al., 2007; Sagar et al., 2010). Furthermore, the absence of a required post-translational modification process, the requirement of *in trans* genetic elements or the fragmentation of previously clustered genes would not allow functional detection (Kwan et al., 2012; Nakabachi et al., 2013). The use and development of alternative bacterial hosts, expression systems, and multi-host shuttle vectors is crucial to overcoming the limitations discussed. The ability to screen using alternative transcriptional machinery and promoter recognition capabilities should broaden the spectrum of gene expression. Recently, in order to achieve good heterologous expression of novel bioactive compounds, the development of marine-derived hosts such as actinomycete, cyanobacteria, and symbiotic fungi was undertaken to optimize heterologous production (Rocha-Martin et al., 2014). The ability to replicate in multiple hosts enables the screening to be conducted in the background of different regulatory and metabolic networks. Furthermore, biosynthetic pathways have also been shown to result in different phenotypes when expressed in different hosts (Craig et al., 2010).

(iii) Owing to the large sizes of the biosynthetic pathways, which routinely approach 100 kb, functional screening of metagenomic libraries for the encoded activity is restricted by the need for the entire cluster to be recovered on a single clone (Kim et al., 2010). Libraries therefore need to be prepared in bacterial artificial chromosomes (BACs), which can be maintained at low copy number and can carry DNA inserts as large as 350 kb (Shizuya and Simon, 1992). However, it is a major technical challenge to preserve the large size of the metagenomic DNA while sufficiently removing impurities that inhibit cloning. In practice, metagenomic BAC libraries only manage 40 kb insert sizes and rarely greater than 70–100 kb (Handelsman et al., 1998; Kakirde et al., 2010). Furthermore, metagenomes representing symbiotic communities associated with marine invertebrates represent hundreds of individual genomes. To adequately represent each one requires massive DNA libraries, in the order of 10^6 clones, to be constructed and screened (Freeman et al., 2012). Therefore, metagenome libraries generally vastly underrepresent the true diversity, which has so far prohibited the realization of a functional metagenomic approach (Fisch et al., 2009).

(iv) Activities which are initially identified and associated with a library clone extract are sometimes lost before the chemical structure can be determined due to strong negative selection in the heterologous system (Curtois et al., 2003).

(v) Microbial-derived compounds often have multiple activities; for example anti-tumor (Abbas et al., 2013; Du et al.,

2013), anti-inflammatory (Chandak et al., 2014), and anti-protozoan (Abdel-Mageed et al., 2010) compounds also display antibacterial activity which may be toxic to the heterologous host. A large proportion of sought-after activities will therefore never be represented in metagenomic libraries. This cannot necessarily be overcome by the use of shuttle vectors because it is in the initial library construction phase that the clones harboring toxic activities will be lost. Ideally metagenomic libraries constructed in shuttle vectors need to be transformed/transfected into the multiple hosts; however, the levels of efficiency required are difficult to generate in non-*E. coli* hosts. Maintaining low copy numbers may enable the host to survive the toxicity; however, it is highly likely that the screening method will not be sensitive enough to detect the active clone.

The Sequence Boom

To overcome some of the challenges associated with functional screening, sequence/homology-based screening has been employed in a number of different ways. It is not the intention of this review to compare function vs. sequence based metagenomic methods; however, a brief review is presented to put into context the need for continued attention to functional metagenomic tools.

Metagenomic DNA or clone libraries can be screened using degenerate PCR primers designed to conserved sequences within biosynthetic gene clusters (Banik and Brady, 2010). The clustering of biosynthetic genes on a contiguous region of DNA makes homology-based screening attractive. The domain architecture of PKSs and NRPSs in most cases mirrors the structure of the assembled metabolite (Piel et al., 2004c). Therefore, the use of degenerate primers is routinely and successfully employed to first detect conserved NRPS and PKS motifs, followed by the recovery of the remainder of the biosynthetic enzymes by association (Moffitt and Neilan, 2001; Dunlap et al., 2007; Bayer et al., 2013). Furthermore, the identification of relatives of known biosynthetic variants could be a strategy to identify or synthesize new structural variants to provide compounds with improved pharmacological properties (Banik and Brady, 2010). However, in some cases up to 99% of the genes detected through PCR screening can represent dominant sequences which are already known and alternative strategies are required to overcome the presence of similar sequences (Piel et al., 2004c; Schirmer et al., 2005; Fieseler et al., 2007; Kennedy et al., 2008; Hochmuth et al., 2010; Siegl and Hentschel, 2010; Pimentel-Elardo et al., 2012; Della Sala et al., 2013, 2014).

Exciting advancements in next generation DNA sequencing and bioinformatics technologies now negates the need to prepare and sequence clone-libraries. Shotgun metagenomic sequencing has made it possible to rapidly identify large biosynthetic gene clusters and subsequent predictions of their chemical structure can be made (Caboche et al., 2008, 2010; Röttig et al., 2011; Medema et al., 2012, 2014; Blin et al., 2013). While purely *in silico* approaches are generally limited to the detection of one or more well-characterized gene cluster classes (Cimermancic et al., 2014), continued developments in bioinformatics pipelines and other technologies are already improving access to diverse and novel secondary metabolite genes and clusters, including providing access to the "rare biosphere" (we refer readers to a number of examples: Li et al., 2010; Sagar et al., 2010; Trindade-Silva et al., 2012; Woodhouse et al., 2013; Cimermancic et al., 2014). Furthermore, the deposition of more functionally curated sequence data in publically available databases should improve the ability to use purely bioinformatics based screening for the identification of novel gene clusters (Tuffin et al., 2009; Suenaga, 2012).

In silico approaches facilitate rapid dereplication of common biosynthesis clusters and thus the prioritization of new chemical scaffolds for experimental characterization. Although, targeted induction in heterologous expression systems has delivered some success from the marine environment (Long et al., 2005; Schmidt et al., 2005; Hochmuth et al., 2010; Rath et al., 2011; Bonet et al., 2015; Li et al., 2015), it is not easily going to deliver compounds with the sought after properties required by the pharmaceutical markets in a high throughput manner, when taking a purely *in silico* discovery route. For example, the *swf* cluster, a new mono-modular type I PKS/FAS (fatty acid synthase) was identified through screening of the *Plakortis simplex* sponge metagenome (Della Sala et al., 2013). The entire pathway was expressed in *E. coli*; however, the production of an associated metabolite was not detected.

Notwithstanding all the difficulties associated with heterologous expression and the inability to conduct this in a high throughput manner, novel sequence will not necessarily result in the pharmaceutically required biological properties. It is currently easier and cheaper to generate vast volumes of gene and genome sequence information than it is to produce the experimental characterizations, and the gap between these is growing rapidly (Prakash and Taylor, 2012; Scholz et al., 2012; Teeling and Glöckner, 2012; Reen et al., 2015).

Targeted Metagenomic Strategies in Marine Discovery

From a pharmaceutical point of view, marine drug discovery necessitates a focus on functionality. Irrespective of the approach employed, obtaining biologically active and pure compounds with the desired activity or properties is the end goal. The ability to achieve this through function-driven screening strategies is, in principle, the golden standard. Given the limitations discussed above this will remain a major challenge. Relative to other environmental biodiscovery efforts, classical functional metagenomic screening of marine sources has yet to contribute significantly to the pharmaceutical industry. However, significant improvements in the chemical and genetic sciences and the integration of these technologies, has resulted in a number of successes which are beginning to drive the development of parallel technologies.

Instead of functionally screening a metagenome clone library, a targeted approach which harnesses prior knowledge of marine natural product diversity, chemistry, and biological activity is

bridging the gap between the accumulation of microbial genetic datasets and functional and ecological relevance (**Figure 1**). In this section we highlight some of the recent discoveries that have employed metagenomic strategies which were guided primarily by initial structural and functional characteristics and associated pharmaceutical interest.

Bryostatins

Bryostatin 1, a polyketide initially detected in 1968 in extracts from the marine bryozoan *Bugula neritina* (Pettit, 1991), raised interest due to its cytotoxic activity against multiple carcinomas, with proteinase kinase C as its molecular target (Mayer et al., 2010). Bryostatin 1 has been tested in over 80 clinical trials for cancer and is also being assessed in Phase I trials as an anti-Alzheimer's drug. Although the *in vivo* activity was initially detected directly from the bryozoan, it was for many years suspected that the compound was produced by a bacterial symbiont since a difference in the types of bryostatins found in *B. neritina* correlated with genetically different bacterial symbionts (Davidson and Haygood, 1999). A particular symbiont in the larvae of the bryozoan was identified and suspected to be the producer, and was proposed as a novel gamma-proteobacterium, '*Candidatus* Endobugula sertula.' Attempts to separate the bacterial cells from the host as a way to confirm *Ca* E. sertula as the producer of the bryostatin were inconclusive, therefore a metagenomic approach was employed (Davidson et al., 2001). Since, bryostatin is a type I polyketide, PKS-based screening was conducted and led to the confirmation that the genes coding for type I PKS complex were derived from the symbiotic population. Further query involving the growth of *B. neritina* colonies after antibiotic treatments and *in situ* hybridization experiments confirmed that "*E. sertula*" was the source of the bryostatins. A cosmid library was prepared from *B. neritina* Mission Bay metagenomic DNA, and was screened by hybridization (Hildebrand et al., 2004) using a β-ketoacyl synthase probe identified previously (Davidson et al., 2001). Several overlapping clones were sequenced revealing the 65 kb *bry* gene cluster (Hildebrand et al., 2004). Probes spanning the *bry* gene cluster were hybridized to '*Candidatus* E. sertula'-enriched DNA to confirm the symbiont as the origin of the gene cluster. Further interrogation in two closely related "*E. sertula*" strains from different host species identified two different gene cluster arrangements (Sudek et al., 2007). In one strain the gene cluster is contiguous, while in the other strain the PKS genes are split from the accessory genes. Due to the difficulties in obtaining sufficient supply of the bryostatins, their clinical application occurred decades after their discovery. Since "*E. sertula*" remains unculturable, heterologous expression of the *bry* gene cluster could be considered for the production of bryostatins in large enough quantities for pharmaceutical development.

ET-743 (Yondelis®)

Anti-cancer activity in extract from the sea squirt *Ecteinascidia turbinata* was identified in 1969; however, it was only in 1984 that the structure of one of the compounds, Ecteinascidin 743 (ET-743), was determined (Rinehart, 2000). ET-743 (Yondelis®) is now an approved anti-cancer agent (Bewley and Faulkner, 1998). Attempts to farm the sea squirt to provide sufficient supply of the compound had limited success, and it is currently generated in suitable quantities for clinical use by a lengthy semi-synthetic process (Cuevas et al., 2000; Rath et al., 2011). The similarity of ET-743 to three other bacterial derived natural products (saframycin A, *Streptomyces lavendulae*; saframycin Mx1, *Myxococcus xanthus*; safracin B, *Pseudomonas fluorescens*; Rath et al., 2011) suggested that ET-743 was produced by a marine bacterial symbiont. Using metagenomic sequencing of total DNA from the microbial consortium associated with the tunicate resulted in the assembly of a 35 kb contig containing 25 genes encoding a NRPS biosynthetic pathway. Rigorous sequence analysis of two large unlinked contigs suggested that '*Candidatus* Endoecteinascidia frumentensis' was the producer of the metabolite. Subsequent metaproteomic analysis confirmed expression of three key biosynthetic proteins. The complete genome of '*Candidatus* Endoecteinascidia frumentensis' was very recently determined, showing an extremely reduced genome (∼631 kb) and evidence of an endosymbiotic lifestyle (Schofield et al., 2015). Having the pathway elucidated provides the foundation for direct production of the drug and new analogs through metabolic engineering (Rath et al., 2011).

Patellazoles

The *Lissoclinum patella* tunicate has garnered interest due to it representing a rich source of potential bioactive drug leads (Kwan et al., 2012; Schmidt et al., 2012). The patellazoles were isolated directly from the tunicate in the late 1980s and characterized as a new family of novel thiazole-containing polyketide metabolites (Corley et al., 1988; Zabriskie et al., 1988). In addition to their chemical novelty, they gained interest due to their potent cytotoxic activity against human cell lines as well as antifungal (*Candida albicans*) activity (Zabriskie et al., 1988). The patellamides, also isolated from this tunicate had already been shown to be produced by the cyanobacterial symbiont, *Prochloron didemni* (Schmidt et al., 2005). Although, *P. didemni* is the major symbiont, *L. patella* harbors a complex microbiome (Donia et al., 2011), and therefore there stood the possibility that the patellazoles were also produced by a symbiont. Due to the multiple acetate units, patellazoles were hypothesized to be produced by a type I PKS pathway, as well as a NRPS module for the incorporation and cyclization of a cysteine unit to generate the thiazole ring (Kwan et al., 2012). Based on this information an exhaustive sequence based screening of a metagenome clone library prepared from the tunic-cloaca habitats was undertaken, but did not locate the biosynthetic pathway. PCR amplification revealed PKS genes from the *trans*-acyltransferase family, consistent with patellazole biosynthesis, in the tiny zooids but not in the bulk tunic. DNA extracted from the zooids fraction was subjected to shotgun sequencing and the assembly thereof resulted in a complete genome which contained a 86 kb *trans*-AT PKS pathway. The predicted biosynthetic model of the encoded pathway was consistent with patellazoles structure, thus strongly supporting the assignment. The assembled genome was considered to belong to an uncultured symbiont, designated as *Candidatus* "Endolissoclinum faulkneri," most closely related to free-living marine α-proteobacteria.

Pederin-led Discovery of the Onnamides

Pederin and mycalamides A and B, encoded by a mixed modular PKS-non-ribosomal peptide synthetase (NRPS) system, are highly active antitumor compounds (Narquizian and Kocienski, 2000). These compounds block mitosis at levels as low as 1 ng/ml by inhibiting protein and DNA synthesis without affecting RNA synthesis (Singh and Yousuf Ali, 2007). They prevent cell division, and have been shown to extend the life of cancerous mice. Consequently they have garnered interest as potential anti-cancer treatments (Kanamitsu and Frank, 1987). These compounds were initially known exclusively from terrestrial *Paederus* and *Paederidus* beetles and after many years of speculation were finally shown to be produced by an uncultured symbiotic *Pseudomonas* associated with the *Paederus fuscipes* beetles using metagenomic approaches (Piel et al., 2004a). Interestingly, these insects use pederin as a chemical weapon against predators and when in contact with human skin cause severe dermatitis (Borroni et al., 1991; Piel et al., 2004a). Metabolites with high structural similarity to pederin were identified in the marine sponges of the order *Lithistida* (Bewley and Faulkner, 1998), many of which exhibit extremely potent antitumor activity and also selectivity against solid tumor cell lines (Cichewicz et al., 2004). A pederin-informed survey of PKS amplicons from the Japanese sponge *Theonella swinhoei* metagenome, a species with exceptionally rich chemistry (Fusetani and Matsunaga, 1993; Bewley and Faulkner, 1998), revealed a wide range of distinct KS sequences (Piel et al., 2004b). Three of these belonged to an evolutionarily distinct enzyme family, the *trans*-acyl-transferase (*trans*-AT) PKSs, and corresponded to onnamide biosynthesis pathways. These *trans*-AT PKSs therefore were expected to encode the pederin-like compounds with antitumor activity produced by the sponges. Further, screening of the metagenomes from other *T. swinhoei* specimens revealed that these *trans*-AT PKSs could only be detected in the sponges which had previously shown to contain pederin-type compounds, while no amplification was obtained for sponges devoid of these compounds. It has now been confirmed that the onnamides (**Figure 2**) are produced by an unculturable symbiont, '*Candidatus* Entotheonella spp.' (Wilson et al., 2014).

Psymberin

Psymberin, a highly cytotoxic and selective antitumor polyketide, has been isolated from a number of different marine sponges (Bielitza and Pietruszka, 2013). It took 11 years and 600 samples for the structure of this compound to be assigned. There is immense interest in this natural product due to its complex architecture, biological properties and scarcity in nature. As with

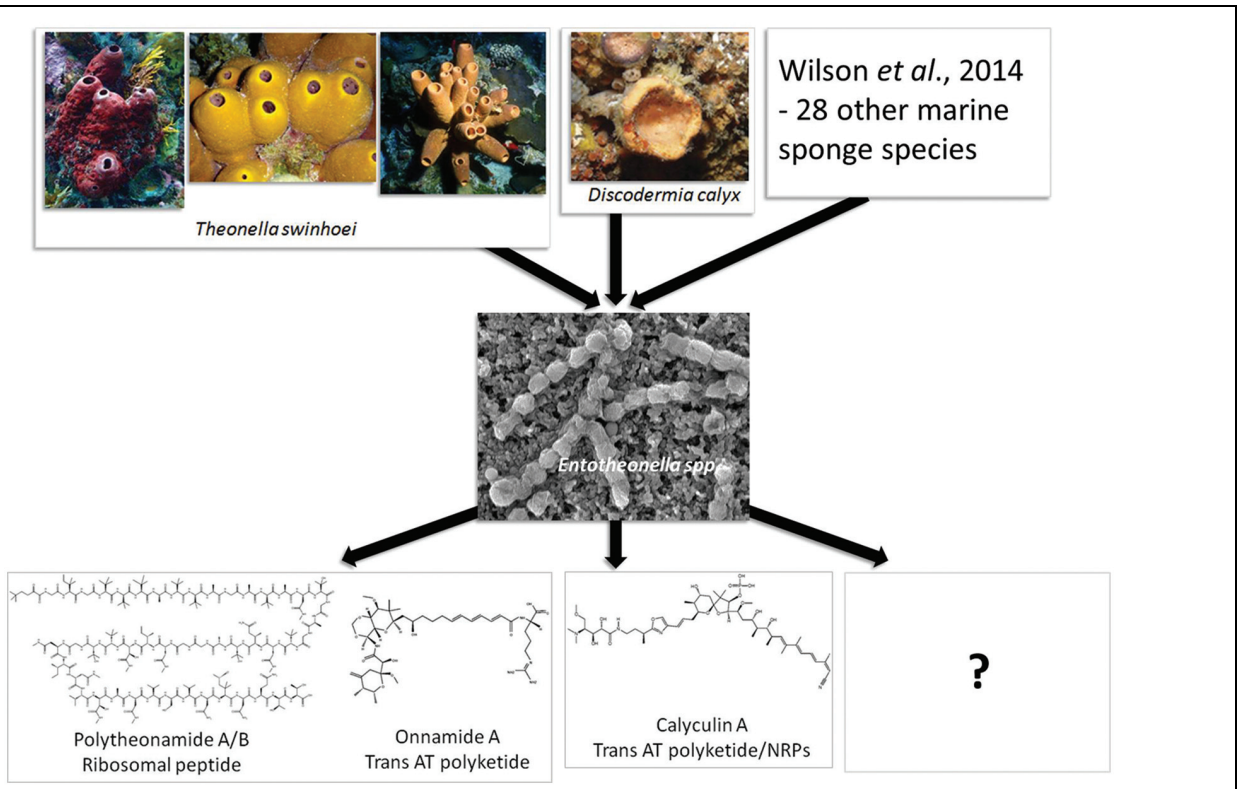

FIGURE 2 | A representation of the ubiquity of "*Entotheonella*" species in taxonomically diverse marine sponges and the notable secondary metabolites they produce. Metabolite structure and function information obtained directly from the *Theonella swinhoei* and *Discodermia calyx* sponges informed a targeted metagenomic approach to identify the biosynthetic pathways encoding these metabolites. This ultimately led to the discovery of "*Entotheonella,*" described as "talented producers" due to their chemically diverse metabolism. The full potential of *Entotheonella* species has yet to be explored. Photos were provided by T. Mori, P. Poppe, and T. Wakimoto.

the onnamides, psymberin is a member of the pederin family, also synthesized by a *trans*-AT PKS (Piel et al., 2004b), but is distinguished from the other pederins due to its excellent cytotoxicity values which exceeds those of the other family members. A *trans*-AT PKS PCR screening approach, as described above for the onnamides, was used to elucidate the complete biosynthetic pathway for psymberin from the *Psammocinia aff. bulbosa* sponge metagenome (Fisch et al., 2009). The genomic region showed typical bacterial architecture, suggesting a bacterial symbiont origin. However, unlike for the pederin and onnamide examples, there were not enough similarities to other bacterial genes to identify the bacterium.

Polytheonamides

The polytheonamides (**Figure 2**) represent another group of exceptionally potent natural product toxins isolated from the *Theonella swinhoei* sponges, and are particularly noteworthy for their size and structural complexity (Hamada et al., 2010). These 48-residue peptides were expected to be products of a non-ribosomal peptide synthetase, since of the 19 different amino acids that constitute these peptides, 13 are non-proteinogenic. Furthermore, the peptides include multiple D-configured and C-methylated residues. However, the size of the NRP biosynthetic machinery required to produce a 48 residue peptide prompted a search for an unusual ribosomal pathway. With the knowledge of the peptide sequence, degenerate PCR primers were designed to the proposed precursor peptide, and used to conduct a semi-nested PCR from a *T. swinhoei* metagenome (Freeman et al., 2012). Sequenced amplicons revealed codon sequences that precisely corresponded to an unprocessed polytheonamide precursor, not only confirming a ribosomal origin, but also suggesting that it is produced by a bacterial endosymbiont. Further screening of the metagenome library revealed the entire 12 gene biosynthetic pathway. Microscopic analysis of *T. swinhoei* (Y chemotype) samples identified a highly enriched population of fluorescent filamentous bacteria showing morphological similarity to the symbiont '*Candidatus* Entotheonella palauensis,' the suspected producer of antifungal peptides isolated from a Palauan *T. swinhoei* chemotype (Schmidt et al., 2000). Using single cell genomics (fluorescence assisted cell sorting and whole genome amplification) combined with pathway specific PCR, the identification of the polytheonamide producer was attributed to an uncultured "*Entotheonella*" spp. (Wilson et al., 2014). Further, screening using onnamide pathway specific markers indicated that the "*Entotheonella*" spp. were the source of both the onnamide and polytheonamide compounds.

Calyculin A

Calyculin A was originally described in 1986 as a major cytotoxic compound isolated from *Discodermia calyx*, a marine sponge of the Theonellidae family (Kato et al., 1986), and is to date associated exclusively with marine sponges (Wakimoto et al., 2014). Calyculin A represents a fairly sophisticated and unique structure whose biosynthesis was reminiscent of a polyketide and non-ribosomal peptide hybrid pathway incorporating some remarkable modification processes. Calyculin-related compounds have been isolated from a number of different sponges which hinted toward a symbiont being responsible for its production (Dumdei et al., 1997; Edrada et al., 2002; Kehraus et al., 2002). However, it was only very recently that the biosynthetic gene cluster was identified through a metagenomic approach. Based on the initial hypothesis that calyculin A was a type I polyketide, a metagenome library of *D. calyx* was sequentially screened by PCR amplification using *trans*-AT-type KS, adenylation domain (NRPS) and HMGS-like motif primers (Wakimoto et al., 2014). Spanning over 150 kb, a gene cluster containing 29 KS and 5 A domains was identified. The collinearity between the order of the modules and the order of the biosynthetic reactions provided strong evidence that the cluster encoded calyculin A synthesis. Using the entire gene cluster as a probe and employing CARD-FISH, as well as laser microdissection, a filamentous bacterium was identified to harbor the calyculin pathway. The 16S rRNA sequence of this bacterium displayed 97% identity to the '*Candidatus* Entotheonella factor' isolated from the *T. swinhoei* sponges.

From Function to Genes to Species

The power of metagenomics to identify novel and pharmaceutically relevant organisms, resulting from first obtaining functional and structural data, has been elegantly represented in the examples discussed above. To demonstrate this further, the "*Entotheonella*" and the '*Candidatus* Endolissoclinum faulkneri' stories are elaborated (**Figure 2**).

Genome sequencing of several of the single cell sorted events in the Wilson et al. (2014) study indicated the presence of two closely related "*Entotheonella*" variants, with 97.6% identical 16S rRNA sequences, and 97% identity to *E. palauensis*. These are only 82% identical to representatives from known bacterial phyla and form a well-separated clade, and therefore have been proposed to represent a new candidate phylum "Tectomicrobia." Both genomes exceed 9 Mb, representing some of the largest known prokaryote genomes. Analysis of the metabolic genes identified over 28 biosynthetic clusters, encoding ribosomal, polyketide and non-ribosomal peptide biosynthesis. Using bioinformatics based predictions, several of the clusters could be assigned to known bioactive peptides isolated from the Japanese *T. swinhoei*, and together with tandem mass spectrometry-based molecular networking a high diversity of previously unknown metabolite families were identified. The combination of these properties is so rare that the new phylum to which these isolates have been assigned is considered the successor to the Actinobacteria, the well-known source of the majority of the world's antibiotics and anticancer agents (Jaspars and Challis, 2014). Screening for the distribution of these talented producers indicated that they are geographically widespread and are symbiotically associated with other sponge types (Wilson et al., 2014; **Figure 2**). The discovery of a calyculin producing "*Entotheonella*" symbiotically associated with *D. calyx* further expands the number of biosynthetic enzymes and chemical scaffolds encompassed by this genus (Wakimoto et al., 2014), but also serves to highlight the differences between the

"*Entotheonella*" populations in different sponges. Attempts to culture the producing symbionts have been unsuccessful. Access to the genome sequences should give important insights to the organism's metabolism, and such clues to their physiology could inform on the development of appropriate culturing strategies. Several uncultured symbionts have been successfully isolated using such genome sequence-guided strategies (Renesto et al., 2003; Bomar et al., 2011).

In the Kwan et al. (2012) study the patellazole encoding *Ca.* E. faulkneri genome assembled to a mere 1.48 Mbp and showed extensive genome reduction characteristics. Unlike other bacteria with streamlined genomes, *Ca.* E. faulkneri has distinguishing features which strongly suggest that it could not exist independently of its host, *L. patella*. Phylogenetic analysis of patellazole-containing and patellazole-negative tunicates provides evidence that the symbiont has coevolved with the tunicate host and would therefore be transmitted vertically. The patellazole pathway is the only secondary metabolite pathway encoded in the *Ca.* E. faulkneri genome, and represents >10% of the coding sequence. The maintenance of such a large pathway in a genome that is so streamlined as to eliminate most functions indicates its importance to the symbiotic relationship. However, the patellazoles are highly toxic to eukaryotic cells and are found in high amounts in *L. patella*, and it is intriguing that the host has apparently adapted to tolerate such high concentrations. Clearly the patellazoles provide important chemical defense to the host which in turn ensures that the symbiont is maintained. Interestingly, *Ca.* E. faulkneri is found sporadically in *L. patella* tunicates. Patellazole-positive and negative *L. patella* collected within 250 m of each other show that *Ca.* E. faulkneri is only associated with the patellazole-positive colonies, and only in the zooids fraction. This is despite patellazole-positive and negative colonies having nearly identical tunicate phylogeny, and containing virtually identical microbial communities, with the exception of the *Ca.* E. faulkneri. The exclusive localization of *Ca.* E. faulkneri in the zooids and only in certain *L. patella* colonies is intriguing. Considering the *L. patella* zooids filter feed and excrete waste into the cloacal cavities, this could perhaps provide some leads of investigation to further understanding the *Ca.* E. faulkneri localization and the symbiotic relationship.

These discoveries raise several fundamental biological questions relating to: symbiont and secondary metabolite evolution, mechanisms of natural product symbiosis, the role of the natural products in imparting a direct competitive advantage to individual members of a bacterial consortium, and how these symbiotic interactions contribute to the ecology of the marine environment. However, what is now clearly appreciated is that the genomes of previously uncultured bacteria harbor an unprecedented richness of novel compound diversity, and await discovery.

Conclusion and Future Prospects

The remarkable exploration of marine organisms and their structurally diverse natural products spans a highly active period of over 40 years (Gerwick and Moore, 2012). With attention turning to marine microorganisms as a source of new natural product chemistry, and the realization that many compounds previously isolated are metabolic products of unculturable microbes, marine metagenomics promises to illuminate new bioactivities and chemistries that were previously unattainable. Despite metagenomics being a relatively young technology, it is globally appreciated that major advances are needed given the challenges that now bottleneck future developments, irrespective of whether functional or sequence guided approaches are to be employed. In order to maximize our ability to harvest marine resources the synergic combination of a number of complementary methodologies and integration of functional and informatics approaches will be required (Reen et al., 2015). The examples presented, employing a targeted and function-guided strategy, demonstrate how metagenomic technologies have advanced several research disciplines and our understanding of microbial genetic and biological diversity and ecology. Armed with information of the chemical structure and biological activity of pharmaceutically relevant compounds, an informed metagenomic strategy, in combination with *in situ* hybridization, single cell-sorting, whole genome amplification, and next generation sequencing, has successfully identified novel biosynthetic gene clusters and novel microbes that produce the metabolites. The path that led from similar compounds being found in organisms as divergent as marine sponges and beetles, to the discovery that microorganisms were the producers, and the role metagenomics played, makes a fascinating story demonstrating a perfect blend of fundamental and applied science, exemplifying the power of employing integrated technologies.

For marine metagenomics to significantly contribute to delivering pharmaceutically relevant compounds, improvements in, and integration of, various approaches and strategies is key. One of the most important hindrances encountered thus far in natural product research is re-isolation of known compounds. Thus chemical and biological de-replication is a crucial step in the process, and applies to metagenomic guided discovery as well, irrespective of the metagenomic approach employed. While sequence-based metagenomic approaches offer the power of discrimination, the expression of the pathways and the functional and biochemical characterization of the encoded products is crucial. Genome data is being produced at a dizzying pace; however, without focusing on heterologous expression challenges and the development of functional screens our capacity to uncover and develop the next generation of pharmaceutically relevant molecules will be limited (Reen et al., 2015). There are two long standing schools of thought on natural products discovery: 'isolate and then test' vs. 'test and then isolate' (Gerwick and Moore, 2012). A parallel can be drawn to employing metagenomic tools to natural product discovery: "sequence and then test" vs. "test and then sequence." This review summarizes some of the most recent marine discoveries through the latter approach, born out of traditional chemistry-guided discovery conducted over several decades. However, to maximize our capacity to mine metagenomes for activities which have yet to be identified, parallel developments in a number of technologies need continuous attention;

including biological assay screening; isolation and separation methods and analytical chemistry techniques. Peptidogenomics represents a recent advancement in high throughput mass spectrometry (MS; Kersten et al., 2011; Bouslimani et al., 2014; Medema et al., 2014). This automated approach iteratively matches the chemotypes of peptide natural products to their biosynthetic gene clusters through *de novo* tandem MS (MSn) and genome-mining (Reen et al., 2015). This constitutes a paradigm shift from the one molecule-per-study approach to drug discovery (Medema et al., 2014), and may be the key to revealing novel marine natural products from metagenomes, for advancement into the drug discovery development pipeline.

There is no doubt that as yet uncultured bacteria are a rich source of novel bioactive molecules with potent therapeutic activity, and these are exciting times to be a researcher in the field.

Acknowledgments

We thank the South African National Research Foundation, The Department of Science and Technology (DST) and the University of the Western Cape for financial support. The authors would like to thank T. Mori, P. Poppe, and T. Wakimoto for the photographic contributions as indicated in the figure legends.

References

Abbas, S. E., Abdel Gawad, N. M., George, R. F., and Akar, Y. A. (2013). Synthesis, antitumor and antibacterial activities of some novel tetrahydrobenzo[4,5]thieno[2,3-*d*]pyrimidine derivatives. *Eur. J. Med. Chem.* 65, 195–204. doi: 10.1016/j.ejmech.2013.04.055

Abd Elrazak, A., Ward, A., and Glassey, J. (2013). Response surface methodology for optimising the culture conditions for eicosapentaenoic acid production by marine bacteria. *J. Ind. Microbiol. Biotechnol.* 40, 477–487. doi: 10.1007/s10295-013-1238-x

Abdel-Mageed, W., Milne, B., Wagner, M., Schumacher, M., Sandor, P., Pathom-aree, W., et al. (2010). Dermacozines, a new phenazine family from deep-sea dermacocci isolated from a Mariane trench sediment. *Org. Biomol. Chem.* 8, 2352–2362. doi: 10.1039/c001445a

Abdelmohsen, U., Szesny, M., Othman, E., Schirmeister, T., Grond, S., Stopper, H., et al. (2012). Antioxidant and anti-protease activities of diazepinomicin from the sponge-associated *Micromonospora* strain RV115. *Mar. Drugs* 10, 2208–2221. doi: 10.3390/md10102208

Amann, R., Ludwig, W., and Schleifer, K. (1995). Phylogenetic identification and in situ detection of individual microbial cells without culturing. *Microbiol. Rev.* 59, 143–169.

Banik, J., and Brady, S. (2010). Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. *Curr. Opin. Microbiol.* 13, 603–609. doi: 10.1016/j.mib.2010.08.012

Bayer, K., Scheuermayer, M., Fieseler, L., and Hentschel, U. (2013). Genomic mining for novel FADH2-dependent halogenases in marine sponge-associated microbial consortia. *Mar. Biotechnol.* 15, 63–72. doi: 10.1007/s10126-012-9455-2

Berdy, J. (2005). Bioactive microbial metabolites—a personal view. *J. Antibiot.* 58, 1–26. doi: 10.1038/ja.2005.1

Bergmann, W., and Feeney, R. (1951). Contributions to the study of marine products. XXXII. The nucleosides of sponges. *J. Org. Chem.* 16, 981–987. doi: 10.1021/jo01146a023

Bewley, C., and Faulkner, D. (1998). Lithistid sponges: star performers or hosts to the stars. *Angew. Chem. Int. Ed.* 37, 2162–2178.

Bielitza, M., and Pietruszka, J. (2013). The psymberin story—biological properties and approaches towards total and analogue syntheses. *Angew. Chem. Int. Ed.* 52, 10960–10985. doi: 10.1002/anie.201301259

Blin, K., Medema, M., Kazempour, D., Fischbach, M., Breitling, R., Takano, E., et al. (2013). antiSMASH 2.0. A versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* 41, 204–212. doi: 10.1093/nar/gkt449

Blunt, J., Copp, B., Keyzers, R., Munro, M., and Prinsep, M. (2015). Marine natural products. *Nat. Prod. Rep.* 32, 116–211. doi: 10.1039/c4np00144c

Bomar, L., Maltz, M., Colston, S., and Graf, J. (2011). Directed culturing of microorganisms using metatranscriptomics. *MBio* 2, e12–e11. doi: 10.1128/mBio.00012-11

Bonet, B., Teufel, R., Crüsemann, M., Ziemert, N., and Moore, B. (2015). Direct capture and heterologous expression of *Salinispora* natural product genes for the biosynthesis of enterocin. *J. Nat. Prod.* 78, 539–542. doi: 10.1021/np500664q

Borroni, G., Brazzelli, V., Rosso, R., and Pavan, M. (1991). Paederusfuscipes dermatitis. A histopathological study. *Am. J. Dermatopathol.* 13, 467–474. doi: 10.1097/00000372-199110000-00007

Bouslimani, A., Sánchez, L., Garg, N., and Dorrestein, P. (2014). Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Prod. Rep.* 31, 718–729. doi: 10.1039/c4np00044g

Brady, S., Chao, C., Handelsman, J., and Clardy, J. (2001). Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org. Lett.* 3, 1981–1984. doi: 10.1021/ol015949k

Brady, S., and Clardy, J. (2000). Long-chain n-acyl amino acid antibiotics isolated from heterologously expressed environmental DNA. *J. Am. Chem. Soc.* 122, 12903–12904. doi: 10.1021/ja002990u

Brady, S., and Clardy, J. (2005). Cloning and heterologous expression of isocyanide biosynthetic genes from environmental DNA. *Angew. Chem. Int. Ed. Engl.* 44, 7063–7065. doi: 10.1002/anie.200501941

Caboche, S., Leclere, V., Pupin, M., Kucherov, G., and Jacques, P. (2010). Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. *J. Bacteriol.* 192, 5143–5150. doi: 10.1128/JB.00315-10

Caboche, S., Pupin, M., Leclere, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.* 36, 326–331. doi: 10.1093/nar/gkm792

Chandak, N., Kumar, P., Kaushik, P., Varshney, P., Sharma, C., Kaushik, D., et al. (2014). Dual evaluation of some novel 2-amino-substituted coumarinylthiazoles as anti-inflammatory–antimicrobial agents and their docking studies with COX-1/COX-2 active sites. *J. Enzyme Inhib. Med. Chem.* 29, 476–484. doi: 10.3109/14756366.2013.805755

Charan, R., Schlingmann, G., Janso, J., Bernan, V., Feng, X., and Carter, G. (2004). Diazepinomicin, a new antimicrobial alkaloid from a marine *Micromonospora* sp. *J. Nat.Prod.* 67, 1431–1433. doi: 10.1021/np040042r

Cho, J., Kwon, H., Williams, P., Jensen, P., and Fenical, W. (2006). Azamerone, a terpenoid phthalazinone from a marine-derived bacterium related to the genus *Streptomyces* (Actinomycetales). *Org. Lett.* 8, 2471–2474. doi: 10.1021/ol060630r

Chopra, L., Singh, G., Choudhary, V., and Sahoo, K. (2014). Sonorensin: an antimicrobial peptide, belonging to the heterocycloanthracin subfamily of bacteriocins, from a new marine isolate, *Bacillus sonorensis* MT93. *Appl. Environ. Microbiol.* 80, 2981–2990. doi: 10.1128/AEM.04259-13

Cichewicz, R., Valeriote, F., and Crews, P. (2004). Psymberin, a potent sponge-derived cytotoxin from *Psammocinia* distantly related to the pederin family. *Org. Lett.* 6, 1951–1954. doi: 10.1021/ol049503q

Cimermancic, P., Medema, M., Claesen, J., Kurita, K., Wieland Brown, L., Mavrommatis, K., et al. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158, 412–421. doi: 10.1016/j.cell.2014.06.034

Corley, D., Moore, R., and Paul, V. (1988). Patellazole B: a novel cytotoxic thiazole-containing macrolide from the marine tunicate *Lissoclinum patella*. *J. Am. Chem. Soc.* 110, 7920–7922. doi: 10.1021/ja00231a078

Craig, J., Chang, F., Kim, J., Obiajulu, S., and Brady, S. (2010). Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl. Environ. Microbiol.* 76, 1633–1641. doi: 10.1128/AEM.02169-09

Cuevas, C., Perez, M., Martin, M., Chicarro, J., Fernández-Rivas, C., Flores, M., et al. (2000). Synthesis of ecteinascidin ET-743 and phthalascidin Pt-650 from cyanosafracin B. *Org. Lett.* 2, 2545–2548. doi: 10.1021/ol0062502

Curtois, S., Cappellano, C., Ball, M., Francou, F., Normand, P., Helynck, G., et al. (2003). Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl. Environ. Microbiol.* 69, 49–55. doi: 10.1128/AEM.69.1.49-55.2003

Davidson, S., Allen, S., Lim, G., Anderson, C., and Haygood, M. (2001). Evidence for the biosynthesis of bryostatins by the bacterial symbiont "Candidatus Endobugula sertula" of the bryozoan *Bugula neritina*. *Appl. Environ. Microbiol.* 67, 4531–4537. doi: 10.1128/AEM.67.10.4531-4537.2001

Davidson, S., and Haygood, M. (1999). Identification of sibling species of the bryozoan *Bugula neritina* that produce different anticancer bryostatins and harbor distinct strains of the bacterial symbiont "Candidatus Endobugula sertula.". *Biol. Bull.* 196, 273–280. doi: 10.2307/1542952

Della Sala, G., Hochmuth, T., Costantino, V., Teta, R., Gerwick, W., Gerwick, L., et al. (2013). Polyketide genes in the marine sponge Plakortis simplex: a new group of mono-modular type I polyketide synthases from sponge symbionts. *Environ. Microbiol. Rep.* 5, 809–818. doi: 10.1111/1758-2229.12081

Della Sala, G., Hochmuth, T., Teta, R., Costantino, V., and Mangoni, A. (2014). Polyketide synthases in the microbiome of the marine sponge Plakortis halichondrioides: a metagenomic update. *Mar. Drugs* 12, 5425–5440. doi: 10.3390/md12115425

Donia, M., Fricke, W., Partensky, F., Cox, J., Elshahawi, S., White, J., et al. (2011). Complex microbiome underlying secondary and primary metabolism in the tunicate-Prochloron symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 108, E1423–E1432. doi: 10.1073/pnas.1111712108

Donia, M., and Hamann, M. (2003). Marine natural products and their potential applications as anti-infective agents. *Lancet Infect. Dis.* 3, 338–348. doi: 10.1016/S1473-3099(03)00655-8

Du, Q., Li, D., Pi, Y., Li, J., Sun, J., Fang, F., et al. (2013). Novel 1,3,4-oxadiazole thioester derivatives targeting thymidylate synthase as dual anticancer/antimicrobial agents. *Biorg. Med. Chem.* 21, 2286–2297. doi: 10.1016/j.bmc.2013.02.008

Dumdei, E., Blunt, J., Munro, M., and Pannell, L. (1997). Isolation of calyculins, calyculinamides, and swinholide H from the New Zealand deep-water marine sponge *Lamellomorpha strongylata*. *J. Org. Chem.* 62, 2636–2639. doi: 10.1021/jo961745j

Dunlap, W., Battershill, C., Liptrot, C., Cobb, R., Bourne, D., Jaspars, M., et al. (2007). Biomedicinals from the phytosymbionts of marine invertebrates: a molecular approach. *Methods* 42, 358–376. doi: 10.1016/j.ymeth.2007.03.001

Edrada, R., Ebel, R., Supriyono, A., Wray, V., Schupp, P., Steube, K., et al. (2002). Swinhoeiamide A, a new highly active calyculin derivative from the marine sponge *Theonella swinhoei*. *J. Nat. Prod.* 65, 1168–1172. doi: 10.1021/np020049d

Farnet, C., and Zazopoulos, E. (2005). "Improving drug discovery from microorganisms," in *Natural Products: Drug Discovery and Therapeutic Medicine*, eds L. Zhang and A. Demain (Totowa, NJ: Humana Press Inc.), 95–106.

Feling, R., Buchanan, G., Mincer, T., Kauffman, C., Jensen, P., and Fenical, W. (2003). Salinosporamide A: a highly cytotoxic proteasome inhibitor from a novel microbial source, a marine bacterium of the new genus Salinospora. *Angew. Chemie. Int. Ed.* 42, 355–357. doi: 10.1002/anie.200390115

Ferrer, M., Beloqui, A., Timmis, K., and Golyshin, P. (2009). Metagenomis fro mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.* 16, 109–123. doi: 10.1159/000142898

Fieseler, L., Hentschel, U., Grozdanov, L., Schirmer, A., Wen, G., Platzer, M., et al. (2007). Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial consortia associated with marine sponges. *Appl. Environ. Microbiol.* 73, 2144–2155. doi: 10.1128/AEM.02260-06

Fisch, K., Gurgui, C., Heycke, N., van der Sar, S., Anderson, S., Webb, V., et al. (2009). Polyketide assembly lines of uncultivated sponge symbionts from structure-based gene targeting. *Nat. Chem. Biol.* 5, 494–501. doi: 10.1038/nchembio.176

Freeman, M., Gurgui, C., Helf, M., Morinaka, B., Uria, A., Oldham, N., et al. (2012). Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. *Science* 338, 387–390. doi: 10.1126/science.1226121

Fusetani, N., and Matsunaga, S. (1993). Bioactive sponge peptides. *Chem. Rev.* 93, 1793–1806. doi: 10.1021/cr00021a007

Gabor, E., Alkema, W., and Janssen, D. (2004). Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ. Microbiol.* 6, 879–886. doi: 10.1111/j.1462-2920.2004.00640.x

Gao, X., Wang, P., and Tang, Y. (2010). Engineered polyketide biosynthesis and biocatalysis in *Escherichia coli*. *Appl. Microbiol. Biotechnol.* 88, 1233–1242. doi: 10.1007/s00253-010-2860-4

Gerwick, W., and Moore, B. (2012). Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem. Biol.* 19, 85–98. doi: 10.1016/j.chembiol.2011.12.014

Gillespie, D., Brady, S., Bettermann, A., Cianciotto, N., Liles, M., Rondon, M., et al. (2002). Isolation of antibiotics turbomycina and B from a metagenomic library of soil microbial DNA. *Appl. Environ. Microbiol.* 68, 4301–4306. doi: 10.1128/AEM.68.9.4301-4306.2002

Graça, A., Bondoso, J., Gaspar, H., Xavier, J., Monteiro, M., de la Cruz, M., et al. (2013). Antimicrobial activity of heterotrophic bacterial communities from the marine sponge *Erylus discophorus* (astrophorida, geodiidae). *PLoS ONE* 8:e78992. doi: 10.1371/journal.pone.0078992

Grüschow, S., Rackham, E., and Goss, R. (2011). Diversity in natural product families is governed by more than enzyme promiscuity alone: establishing control of the pacidamycin portfolio. *Chem. Sci.* 2, 2182–2186. doi: 10.1039/c1sc00378j

Guazzaroni, M.-E., Silva-Rocha, R., and Ward, R. (2015). Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. *Microb. Biotechnol.* 8, 52–64. doi: 10.1111/1751-7915.12146

Haefner, B. (2003). Drugs form the deep: marine natural products as drug candidates. *Drug Discov. Today* 8, 536–544. doi: 10.1016/S1359-6446(03)02713-2

Hamada, N., Matsunaga, S., Fujiwara, M., Fujjita, K., Hirota, H., Schmuck, R., et al. (2010). Solution structure of polytheonamide B, a highly cytotoxic nonribosomal polypeptide from marine sponge. *J. Am. Chem. Soc.* 132, 12941–12945. doi: 10.1021/ja104616z

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249. doi: 10.1016/S1074-5521(98)90108-9

Hardt, I., Jensen, P., and Fenical, W. (2000). Neomarinone, and new cytotoxic marinone derivatives, produced by a marine filamentous bacterium (actinomycetales). *Tetrahedron Lett.* 41, 2073–2076. doi: 10.1016/S0040-4039(00)00117-9

Harunari, E., Imada, C., Igarashi, Y., Fukuda, T., Terahara, T., and Kobayashi, T. (2014). Hyaluromycin, a new hyaluronidase inhibitor of polyketide origin from marine *Streptomyces* sp. *Mar. Drugs* 12, 491–507. doi: 10.3390/md12010491

Hildebrand, M., Waggoner, L., Liu, H., Sudek, S., Allen, S., Anderson, C., et al. (2004). bryA: an unusual modular polyketide synthase gene from the uncultivated bacterial symbiont of the marine bryozoan *Bugula neritina*. *Chem. Biol.* 11, 1543–1552. doi: 10.1016/j.chembiol.2004.08.018

Hochmuth, T., Niederkruger, H., Gernert, C., Siegl, A., Taudien, S., Platzer, M., et al. (2010). Linking chemical and microbial diversity in marine sponges: possible role for poribac- teria as producers of methyl-branched fatty acids. *Chembiochem* 11, 2572–2578. doi: 10.1002/cbic.201000510

Hu, G., Yuan, J., Sun, L., She, Z., Wu, J., Lan, X., et al. (2011). Statistical research on marine natural products based on data obtained between 1985 and 2008. *Mar. Drugs* 9, 514–525. doi: 10.3390/md9040514

Jaspars, M., and Challis, G. (2014). A talented genus. *Nature* 506, 38–39. doi: 10.1038/nature13049

Kakirde, K., Parsley, L., and Liles, M. (2010). size does matter: application-driven approaches for soil metagenomics. *Soil Biol. Biochem.* 42, 1911–1923. doi: 10.1016/j.soilbio.2010.07.021

Kanamitsu, K., and Frank, J. (1987). Paederus, sensulato (Coleoptera: Staphylinidae): natural history and medical importance. *J. Med. Entomol.* 24, 155–191. doi: 10.1093/jmedent/24.2.155

Kato, Y., Fusetani, N., Matsunaga, S., Hashimoto, K., Fujita, S., and Furuya, T. (1986). Calyculin A, a novel antitumor metabolite from the marine sponge *Discodermia calyx*. *J. Am. Chem. Soc.* 108, 2780–2781. doi: 10.1021/ja00270a061

Kehraus, S., König, G., and Wright, A. (2002). A new cytotoxic calyculinamide derivative, geometricin A, from the Australian sponge *Luffariella geometrica*. *J. Nat. Prod.* 65, 1056–1058. doi: 10.1021/np010544u

Kennedy, J., Codling, C., Jones, B., Dobson, A., and Marchesi, J. (2008). Diversity of microbes associated with the marine sponge, *Haliclona simulans*, isolated from Irish waters and identification of polyketide synthase genes from the sponge metagenome. *Environ. Microbiol.* 10, 1888–1902. doi: 10.1111/j.1462-2920.2008.01614.x

Kennedy, J., Flemer, B., Jackson, S., Lejon, D., Morrissey, J., O'Gara, F., et al. (2010). Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. *Mar. Drugs* 8, 608–628. doi: 10.3390/md8030608

Kersten, R., Yang, Y., Xu, Y., Cimermancic, P., Nam, S., Fenical, W., et al. (2011). A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* 7, 794–802. doi: 10.1038/nchembio.684

Kim, J., Feng, Z., Bauer, J., Kallifidas, D., Calle, P., and Brady, S. (2010). Cloning large natural product gene clusters from the environment: piecing environmental DNA gene clusters back together with TAR. *Biopolymers* 93, 833–844. doi: 10.1002/bip.21450

Kirst, H., Creemer, L., Naylor, S., Pugh, P., Snyder, D., Winkle, J., et al. (2002). Evaluation and development of spinosyns to control ectoparasites on cattle and sheep. *Curr. Top. Med. Chem.* 2, 675–699. doi: 10.2174/1568026023393615

Kuzuyama, T., and Seto, H. (2003). Diversity of the biosynthesis of the isoprene units. *Nat. Prod. Rep.* 20, 171–183. doi: 10.1039/b109860h

Kwan, J., Donia, M., Han, A., Hirose, E., Haygood, M., and Schmidt, E. (2012). Genome streamlining and chemical defense in a coral reef symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20655–20660. doi: 10.1073/pnas.1213820109

Leal, M., Puga, J., Serôdio, J., Gomes, N., and Calado, R. (2012). Trends in the discovery of new marine natural products from invertebrates over the last two decades–where and what are we bioprospecting? *PLoS ONE* 7:e30580. doi: 10.1371/journal.pone.0030580

Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P., et al. (2010). Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10430–10435. doi: 10.1073/pnas.0913677107

Li, J., and Neubauer, P. (2014). *Escherichia coli* as a cell factory for heterologous production of nonribosomal peptides and polyketides. *New Biotechnol.* 31, 1–7. doi: 10.1016/j.nbt.2014.03.006

Li, Y., Li, Z., Yamanaka, K., Xu, Y., Zhang, W., Vlamakis, H., et al. (2015). Directed natural product biosynthesis gene cluster capture and expression in the model bacterium *Bacillus subtilis*. *Sci. Rep.* 5, 9383. doi: 10.1038/srep09383

Lim, H., Chung, E., Kim, J., Choi, G., Jang, K., Chung, Y., et al. (2005). Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Appl. Environ. Microbiol.* 2005, 7768–7777. doi: 10.1128/AEM.71.12.7768-7777.2005

Long, P., Dunlap, W., Battershill, C., and Jaspars, M. (2005). Shotgun cloning and heterologous expression of the patellamide gene cluster as a strategy to achieving sustained metabolite production. *Chem. Biochem.* 6, 1760–1765.

Lozupone, C., and Knight, R. (2007). Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11436–11440. doi: 10.1073/pnas.0611525104

Mayer, A. M., Glaser, K. B., Cuevas, C., Jacobs, R. S., Kem, W., Little, R. D., et al. (2010). The odyssey of marine pharmaceuticals: a current pipeline perspective. *Trends Pharmacol. Sci.* 31, 255–265. doi: 10.1016/j.tips.2010.02.005

McGivern, J. (2007). Ziconotide: a review of its pharmacology and use in the treatment of pain. *Neuropsychiatr. Dis. Treat.* 3, 69–85. doi: 10.2147/nedt.2007.3.1.69

Medema, M., Paalvast, Y., Nguyen, D., Melnik, A., Dorrestein, P., Takano, E., et al. (2014). Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.* 10:e1003822. doi: 10.1371/journal.pcbi.1003822

Medema, M., van Raaphorst, R., Takano, E., and Breitling, R. (2012). Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol.* 10, 191–202. doi: 10.1038/nrmicro2717

Moffitt, M., and Neilan, B. (2001). On the presence of peptide synthetase and polyketide synthase genes in the cyanobacterial genus *Nodularia*. *FEMS Microbiol. Lett.* 196, 207–214. doi: 10.1111/j.1574-6968.2001.tb10566.x

Molinski, T., Dalisay, D., Lievens, S., and Saludes, J. (2009). Drug development from marine natural products. *Nat. Rev. Drug Discov.* 8, 69–85. doi: 10.1038/nrd2487

Montaser, R., and Luesch, H. (2011). Marine natural products: a wave of new drugs? *Future Med. Chem.* 3, 1475–1489. doi: 10.4155/fmc.11.118

Nakabachi, A., Ueoka, R., Oshima, K., Teta, R., Mangoni, A., Gurgui, M., et al. (2013). Defensive bacteriome symbiont with a drastically reduced genome. *Curr. Biol.* 23, 1478–1484. doi: 10.1016/j.cub.2013.06.027

Narquizian, R., and Kocienski, P. (2000). "The pederin family of antitumor agents: structures, synthesis and biological activity," in *The Role of Natural Products in Drug Discovery*, eds R. Mulzer and R. Bohlmann (New York, NY: Springer), 25–56.

Nemergut, D., Costello, E., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S., et al. (2011). Global patterns in the biogeography of bacterial taxa. *Environ. Microbiol.* 13, 135–144. doi: 10.1111/j.1462-2920.2010.02315.x

Oh, D.-C., Strangman, W., Kauffman, C., Jensen, P., and Fenical, W. (2007). Thalassospiramide G, a new γ-amino-acid-bearing peptide from the marine bacterium *Thalassospira* sp. *Org. Lett.* 9, 1525–1528. doi: 10.3390/md11030611

Olano, C., Méndez, C., and Salas, J. (2009). Antitumor compounds from marine Actinomycetes. *Mar. Drugs* 7, 210–248. doi: 10.3390/md7020210

Osbourn, A. (2010). Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends Genet.* 26, 449–457. doi: 10.1016/j.tig.2010.07.001

Owen, J., Robins, K., Parachin, N., and Ackerley, D. (2012). A functional screen for recovery of 4'-phosphopantetheinyl transferase and associated natural product biosynthesis genes from metagenome libraries. *Environ. Microbiol.* 14, 1198–1209. doi: 10.1111/j.1462-2920.2012.02699.x

Pettit, G. (1991). The bryostatins. *Prog. Chem. Org. Nat. Prod.* 57, 153–195.

Pettit, G., Knight, J., Herald, D., Pettit, R., Hogan, F., Mukku, V., et al. (2009). Antineoplastic agents. Isolation and structure elucidation of bacillistatins 1 and 2 from a marine *Bacillus silvestris*. *J. Nat. Prod.* 72, 366–371. doi: 10.1021/np800603u

Piel, J., Höfer, I., and Hui, D. (2004a). Evidence for a symbiosis island involved in horizontal acquisition of pederin biosynthetic capabilities by the bacterial symbiont of *Paederus fuscipes* beetles. *J. Bacteriol.* 186, 1280–1286. doi: 10.1128/JB.186.5.1280-1286.2004

Piel, J., Hui, D., Fusetani, N., and Matsunaga, S. (2004b). Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environ. Microbiol.* 6, 921–927. doi: 10.1111/j.1462-2920.2004.00531.x

Piel, J., Hui, D., Wen, G., Butzke, D., Platzer, M., Fusetani, N., et al. (2004c). Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16222–16227. doi: 10.1073/pnas.0405976101

Pimentel-Elardo, S., Grozdanov, L., Proksch, S., and Hentschel, U. (2012). Diversity of nonribosomal peptide synthetase genes in the microbial metagenomes of marine sponges. *Mar. Drugs* 10, 1192–1202. doi: 10.3390/md10061192

Pomponi, S. (2001). The oceans and human health: the discovery and development of marine-derived drugs. *Oceanography* 14, 78–87. doi: 10.5670/oceanog.2001.53

Prakash, T., and Taylor, T. (2012). Functional assignment of metagenomic data: challenges and applications. *Brief. Bioinform.* 13, 711–727. doi: 10.1093/bib/bbs033

Rath, C., Janto, B., Earl, J., Ahmed, A., Hu, F. Z., Hiller, L., et al. (2011). Meta-omic characterization of the marine invertebrate microbial consortium that produces the chemotherapeutic natural product ET-743. *ACS Chem. Biol.* 6, 1244–1256. doi: 10.1021/cb200244t

Reen, F., Gutiérrez-Barranquero, J., Dobson, A., Adams, C., and O'Gara, F. (2015). Emerging concepts promising new horizons for marine biodiscovery and synthetic biology. *Mar. Drugs* 13, 294–2954. doi: 10.3390/md13052924

Renesto, P., Crapouleta, N., Ogata, H., La Scola, B., Vestrisa, G., Claverie, J.-M., et al. (2003). Genome-based design of a cell-free culture medium for *Tropheryma whipplei*. *Lancet* 362, 447–449. doi: 10.1016/S0140-6736(03)14071-8

Rinehart, K. (2000). Antitumor compounds from tunicates. *Med. Res. Rev.* 20, 1–27.

Rocha-Martin, J., Harrington, C., Dobson, A., and O'Gara, F. (2014). Emerging strategies and integrated systems microbiology technologies for

biodiscovery of marine bioactive compounds. *Mar. Drugs* 12, 3516–3559. doi: 10.3390/md12063516

Röttig, M., Medema, M., Blin, K., Weber, T., Rausch, C., and Kohlbacher, O. (2011). NRPSpredictor2. A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* 39, 362–367. doi: 10.1093/nar/gkr323

Sagar, S., Kaur, M., and Minneman, K. (2010). Antiviral lead compounds from marine sponges. *Mar. Drugs* 8, 2619–2638. doi: 10.3390/md8102619

Schipper, C., Hornung, C., Bijtenhoorn, P., Quitschau, M., Grond, S., and Streit, W. (2009). Metagenome-derived clones encoding two novel lactonase family proteins involved in biofilm inhibition in *Pseudomonas aeruginosa*. *Appl. Environ. Microbiol.* 75, 224–233. doi: 10.1128/AEM.01389-08

Schirmer, A., Gadkari, R., Reeves, C., Ibrahim, F., Delong, E., and Hutchinson, C. (2005). Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge *Discodermia dissoluta*. *Society* 71, 4840–4849.

Schmidt, E., Donia, M., McIntosh, J., Fricke, W., and Ravel, J. (2012). Origin and variation of tunicate secondary metabolites. *J. Nat. Prod.* 75, 295–304. doi: 10.1021/np200665k

Schmidt, E., Nelson, J., Rasko, D., Sudek, S., Eisen, J., Haygood, M., et al. (2005). Patellamide A and C biosynthesis by a microcin-like pathway in *Prochloron didemni*, the cyanobacterial symbiont of *Lissoclinum patella*. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7315–7320. doi: 10.1073/pnas.0501424102

Schmidt, E., Obraztsova, A., Davidson, S., Faulkner, D., and Haygood, M. (2000). Identification of the antifungal peptide-containing symbiont of the marine sponge *Theonella swinhoei* as a novel d-proteobacterium, "Candidatus Entotheonella palauensis". *Mar. Biol.* 136, 969–977. doi: 10.1007/s002270000273

Schofield, M. M., Jain, S., Porat, D., Dick, G., and Sherman, D. (2015). Identification and analysis of the bacterial endosymbiont specialized for production of the chemotherapeutic natural product ET-743. *Environ. Microbiol.* doi: 10.1111/1462-2920.12908 [Epub ahead of print].

Scholz, M., Lo, C., and Chain, P. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* 23, 9–15.

Sharma, S., and Vakhlu, J. (2014). "Metagenomics as advanced screening methods for novel microbial metabolite," in *Microbial Biotechnology Progress and Trends*, eds F. D. Harzevili and H. Chen (Boca Raton, FL: CRC Press), 43–62.

Shizuya, H., and Simon, M. (1992). Cloning and stable maintenance of 300 kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. U.S.A.* 89, 8794–8797. doi: 10.1073/pnas.89.18.8794

Siegl, A., and Hentschel, U. (2010). PKS and NRPS gene clusters from microbial symbiont cells of marine sponges by whole genome amplification. *Environ. Microbiol. Rep.* 2, 507–513. doi: 10.1111/j.1758-2229.2009.00057.x

Singh, G., and Yousuf Ali, S. (2007). Paederus dermatitis. *Indian J. Dermatol. Venereol. Leprol.* 73, 13–15. doi: 10.4103/0378-6323.30644

Sogin, M., Morrison, H., Huber, J., Mark Welch, D., Huse, S., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Solanki, R., Khanna, M., and Lal, R. (2008). Bioactive compounds from marine actinomycetes. *Indian J. Microbiol.* 48, 410–431. doi: 10.1007/s12088-008-0052-z

Spainhour, C. (2005). "Natural products," in *Drug Discovery Handbook*, ed. S. Gad (Hoboken, NJ: John Wiley & Sons, Inc.).

Strangman, W. (2007). *Anti-Inflammatory Capabilities of Compounds from Marine Bacteria in a Mouse Model of Allergic Inflammation and Asthma*. Ph.D. thesis, University of California, San Diego, CA.

Sudek, S., Lopanik, N., Waggoner, L., Hildebrand, M., Anderson, C., Liu, H., et al. (2007). Identification of the putative bryostatin polyketide synthase gene cluster from "Candidatus Endobugulasertula", the uncultivated microbial symbiont of the marine bryozoan *Bugula neritina*. *J. Nat. Prod.* 70, 67–74. doi: 10.1021/np060361d

Suenaga, H. (2012). Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ. Microbiol.* 14, 13–22. doi: 10.1111/j.1462-2920.2011.02438.x

Tao, L., Zhu, F., Qin, C., Zhang, C., Chen, S., Zhang, P., et al. (2015). Clustered distribution of natural product leads of drugs in the chemical space as influenced by the privileged target-sites. *Sci. Rep.* 5, 9325. doi: 10.1038/srep09325

Taori, K., Paul, V., and Luesch, H. (2008). Structure and activity of largazole, a potent antiproliferative agent from the Floridian marine cyanobacterium *Symploca* sp. *J. Am. Chem. Soc.* 130, 1806–1807. doi: 10.1021/ja806461e

Taylor, M., Radax, R., Steger, D., and Wagner, M. (2007). Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiol. Mol. Biol. Rev.* 71, 295–307. doi: 10.1128/MMBR.00040-06

Teasdale, M., Liu, J., Wallace, J., Akhlaghi, F., and Rowley, D. (2009). Secondary metabolites produced by the marine bacterium *Halobacillus salinus* that inhibit quorum sensing-controlled phenotypes in gram-negative bacteria. *Appl. Environ. Microbiol.* 75, 567–572. doi: 10.1128/AEM.00632-08

Teeling, H., and Glöckner, F. (2012). Current opportunities and challenges in microbial metagenome analysis-A bioinformatic perspective. *Brief. Bioinform.* 13, 728–742. doi: 10.1093/bib/bbs039

Trindade-Silva, A., Rua, C., Andrade, B., Vicente, A., Silva, G., Berlinck, R., et al. (2012). Polyketide synthase gene diversity within the endemic sponge *Arenosclera brasiliensis* microbiome. *Appl. Environ. Microbiol.* 79, 1598–1605. doi: 10.1128/AEM.03354-12

Tsukimoto, M., Nagaoka, M., Shishido, Y., Fujimoto, J., Nishisaka, F., Matsumoto, S., et al. (2011). Bacterial production of the tunicate-derived antitumor cyclic depsipeptide didemnin B. *J. Nat. Prod.* 74, 2329–2331. doi: 10.1021/np200543z

Tuffin, M., Anderson, D., Heath, C., and Cowan, D. (2009). Metagenomic gene discovery: how far have we moved into novel sequence space? *Biotechnol. J.* 4, 1671–1683. doi: 10.1002/biot.200900235

Uchiyama, T., and Miyazaki, K. (2009). Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr. Opin. Biotechnol.* 20, 616–622. doi: 10.1016/j.copbio.2009.09.010

Van Wezel, G., and McDowall, K. (2011). The regulation of the secondary metabolism of *Streptomyces*: new links and experimental advances. *Nat. Prod. Rep.* 28, 1311–1333. doi: 10.1039/c1np00003a

Vartoukian, S., Palmer, R., and Wade, W. (2010). Strategies for culture of 'unculturable'bacteria". *FEMS Microbiol. Lett.* 309, 1–7. doi: 10.1111/j.1574-6968.2010.02000.x

Wahl, M., Goecke, F., Labes, A., Dobretsov, S., and Weinberger, F. (2012). The second skin: ecological role of epibiotic biofilms on marine organisms. *Front. Microbiol.* 3:292. doi: 10.3389/fmicb.2012.00292

Wakimoto, T., Egami, Y., Nakashima, Y., Wakimoto, Y., Mori, T., Awakawa, T., et al. (2014). Calyculin biogenesis from a pyrophosphate protoxin produced by a sponge symbiont. *Nat. Chem. Biol.* 10, 648–655. doi: 10.1038/nchembio.1573

Wilson, M., Mori, T., Rückert, C., Uria, A. R., Helf, M. J., Takada, K., et al. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* 506, 58–62. doi: 10.1038/nature12959

Woodhouse, J., Fan, L., Brown, M., Thomas, T., and Neilan, B. (2013). Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges. *ISME J.* 7, 1842–1851. doi: 10.1038/ismej.2013.65

Zabriskie, T., Mayne, C., and Ireland, C. (1988). Patellazole C: a novel cytotoxic macrolide from *Lissoclinum patella*. *J. Am. Chem. Soc.* 110, 7919–7920. doi: 10.1021/ja00231a077

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

METHODS
published: 07 July 2015
doi: 10.3389/fmicb.2015.00694

Novel molecular markers for the detection of methanogens and phylogenetic analyses of methanogenic communities

Lukasz Dziewit[1*†], *Adam Pyzik*[2†], *Krzysztof Romaniuk*[1], *Adam Sobczak*[2,3], *Pawel Szczesny*[4,5], *Leszek Lipinski*[2], *Dariusz Bartosik*[1] *and Lukasz Drewniak*[6]

[1] *Department of Bacterial Genetics, Institute of Microbiology, Faculty of Biology, University of Warsaw, Warsaw, Poland,* [2] *Laboratory of RNA Metabolism and Functional Genomics, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland,* [3] *Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland,* [4] *Department of Bioinformatics, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland,* [5] *Department of Systems Biology, Institute of Plant Experimental Biology and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland,* [6] *Laboratory of Environmental Pollution Analysis, Faculty of Biology, University of Warsaw, Warsaw, Poland*

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
James Chong, University of York, UK
Susanna Theroux, Brown University, USA

***Correspondence:**
Lukasz Dziewit, Department of Bacterial Genetics, Institute of Microbiology, Faculty of Biology, University of Warsaw, Miecznikowa 1, Warsaw 02-096, Poland
ldziewit@biol.uw.edu.pl

[†] *These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *07 May 2015*
Accepted: *22 June 2015*
Published: *07 July 2015*

Citation:
Dziewit L, Pyzik A, Romaniuk K, Sobczak A, Szczesny P, Lipinski L, Bartosik D and Drewniak L (2015) Novel molecular markers for the detection of methanogens and phylogenetic analyses of methanogenic communities. Front. Microbiol. 6:694. doi: 10.3389/fmicb.2015.00694

Methanogenic *Archaea* produce approximately one billion tons of methane annually, but their biology remains largely unknown. This is partially due to the large phylogenetic and phenotypic diversity of this group of organisms, which inhabit various anoxic environments including peatlands, freshwater sediments, landfills, anaerobic digesters and the intestinal tracts of ruminants. Research is also hampered by the inability to cultivate methanogenic *Archaea*. Therefore, biodiversity studies have relied on the use of 16S rRNA and *mcrA* [encoding the α subunit of the methyl coenzyme M (methyl-CoM) reductase] genes as molecular markers for the detection and phylogenetic analysis of methanogens. Here, we describe four novel molecular markers that should prove useful in the detailed analysis of methanogenic consortia, with a special focus on methylotrophic methanogens. We have developed and validated sets of degenerate PCR primers for the amplification of genes encoding key enzymes involved in methanogenesis: *mcrB* and *mcrG* (encoding β and γ subunits of the methyl-CoM reductase, involved in the conversion of methyl-CoM to methane), *mtaB* (encoding methanol-5-hydroxybenzimidazolylcobamide Co-methyltransferase, catalyzing the conversion of methanol to methyl-CoM) and *mtbA* (encoding methylated [methylamine-specific corrinoid protein]:coenzyme M methyltransferase, involved in the conversion of mono-, di- and trimethylamine into methyl-CoM). The sensitivity of these primers was verified by high-throughput sequencing of PCR products amplified from DNA isolated from microorganisms present in anaerobic digesters. The selectivity of the markers was analyzed using phylogenetic methods. Our results indicate that the selected markers and the PCR primer sets can be used as specific tools for in-depth diversity analyses of methanogenic consortia.

Keywords: methanogenesis, metagenomics, methanogenic consortia, *mcrB*, *mcrG*, *mtaB*, *mtbA*

Introduction

Methanogenesis is a metabolic process driven by obligate anaerobic *Archaea*. It is responsible for the production of over 90% of methane on Earth (Costa and Leigh, 2014). There are three main methanogenic pathways: (i) hydrogenotrophic methanogenesis using H_2/CO_2 for methane synthesis, (ii) acetoclastic methanogenesis, in which the methyl group from acetate is transferred to tetrahydrosarcinapterin and then to coenzyme M (CoM), and (iii) methylotrophic methanogenesis, using methyl groups from methanol and methylamines (mono-, di-, and trimethylamine) for the production of methyl-coenzyme M (**Figure 1**). The final step in all these pathways is common and involves the conversion of methyl-CoM into methane by methyl-coenzyme M reductase, an enzymatic complex that is present in all methanogens (Borrel et al., 2013) (**Figure 1**).

Methanogenesis is of great importance for biotechnology (e.g., fuel production) and environmental protection (methane emissions contribute to global warming) (Escamilla-Alvaradoa et al., 2012). Therefore, the process has been extensively studied (Gao and Gupta, 2007; Ferry, 2010; Yoon et al., 2013). Consequently, novel species representing particular groups of methanogens are regularly reported (e.g., Dridi et al., 2012; Garcia-Maldonado et al., 2015), and various tools for the genetic and bioinformatic analysis of methanogenic *Archaea* are being developed (e.g., Farkas et al., 2013; Zakrzewski et al., 2013).

Methanogenic *Archaea* form complex consortia which remain largely uncharacterized. Methanogens form close relationships with their syntrophic partners and require very specific environmental conditions for growth, so they have proven very difficult to cultivate in the laboratory (Sekiguchi, 2006; Sakai et al., 2009). Therefore, a number of culture-independent methods have been applied to examine methanogenic consortia: (i) community fingerprinting by denaturing gradient gel electrophoresis—DGGE (Watanabe et al., 2004), (ii) single strand conformation polymorphism—SSCP (Delbes et al., 2001), (iii) terminal restriction fragment length polymorphism—T-RFLP (Akuzawa et al., 2011), (iv) fluorescence *in situ* hybridization—FISH (Diaz et al., 2006), and (v) real-time quantitative PCR—qPCR (Sawayama et al., 2006). However, the most reliable approach for the characterization of methanogenic communities is high-throughput sequencing using either 454 pyrosequencing (e.g., Schlüter et al., 2008; Rademacher et al., 2012; Stolze et al., 2015) or Illumina sequencing technologies (e.g., Caporaso et al., 2011; Zhou et al., 2011; Kuroda et al., 2014; Li et al., 2014).

The most frequently used molecular marker for phylogenetic analyses in metagenomic studies, of methanogenic communities is the 16S rRNA gene. However, low specificity of the oligonucleotide primers employed means that they generate 16S rDNA amplicons for all *Archaea* (not only methanogens) whose DNA is present in the analyzed sample. In the search for a more specific molecular marker for methanogens, the gene encoding the α subunit of the methyl-CoM reductase (*mcrA*) was identified and primers were developed for its amplification (Springer et al., 1995; Lueders et al., 2001; Luton et al., 2002; Friedrich et al., 2005; Yu et al., 2005; Denman et al., 2007; Steinberg and Regan, 2009). Of these, primers designed by Luton et al. (2002), are probably the most extensively used in ecological studies, since they produce the lowest bias in amplifying *mcrA* gene fragments from a wide range of phylogenetically diverse methanogens (e.g., Juottonen et al., 2006).

Several studies have demonstrated that the phylogeny of methanogens based on 16S rDNA and *mcrA* markers is consistent, although greater richness is usually observed using the latter (Luton et al., 2002; Hallam et al., 2003; Bapteste et al., 2005; Nettmann et al., 2008; Borrel et al., 2013). Interestingly, Wilkins and coworkers showed that these two genes produce different taxonomic profiles for samples taken from anaerobic digesters, i.e., environments extremely rich in methanogens (Wilkins et al., 2015). Clearly, the characterization of methanogenic communities requires a systematic approach using reliable molecular markers.

In this study, we have developed a set of degenerate PCR primers for the amplification of genes encoding key enzymes involved in methanogenesis. Some of these represent an alternative to *mcrA* primers commonly used for metagenomic analyses of methanogens. These novel primers amplify fragments of other genes of the *mcr* cluster, i.e., *mcrB* and *mcrG* encoding subunits β and γ of methyl-CoM reductase, respectively. Moreover, we have identified appropriate molecular markers for methylotrophic methanogens, which are probably the least explored group of methanogenic *Archaea*. These primers amplify fragments of the genes *mtaB* (encoding methanol-5-hydroxybenzimidazolylcobamide Co-methyltransferase, which is responsible for the conversion of methanol to methyl-CoM) and *mtbA* (encoding methylated [methylamine-specific corrinoid protein]:coenzyme M methyltransferase involved in the conversion of methylated amines into methyl-CoM). The extended panel of molecular markers provided by these novel primer sets should permit a deeper insight into the complex phylogeny, biology, and evolution of methanogens.

Materials and Methods

Standard Genetic Manipulations

PCR was performed in a Mastercycler (Eppendorf) using Taq DNA polymerase (Qiagen; with supplied buffer), dNTP mixture and appropriate primer pairs [**Table 1** and additionally primer pairs S-D-Arch-0349-a-S-17/S-D-Arch-0786-a-A-20 for amplification of the variable region (V3V4) of the archaeal 16S rRNA gene (Klindworth et al., 2013), and MLf/MLr for *mcrA* gene amplification (Luton et al., 2002)]. PCR products of the methanogenesis-linked genes were purified by gel extraction, cloned using the pGEM®-T Easy Vector System (Promega) and transformed into *E. coli* TG1 (Stratagene) according to a standard procedure (Kushner, 1978). Standard methods were used for the isolation of recombinant plasmid DNA and for common DNA manipulation techniques (Sambrook and Russell, 2001).

Sample Collection

Samples of microbial consortia involved in biogas production were collected from (i) the fermenter tank of an agricultural two-stage biogas plant anaerobic digester (AD) in Miedzyrzec

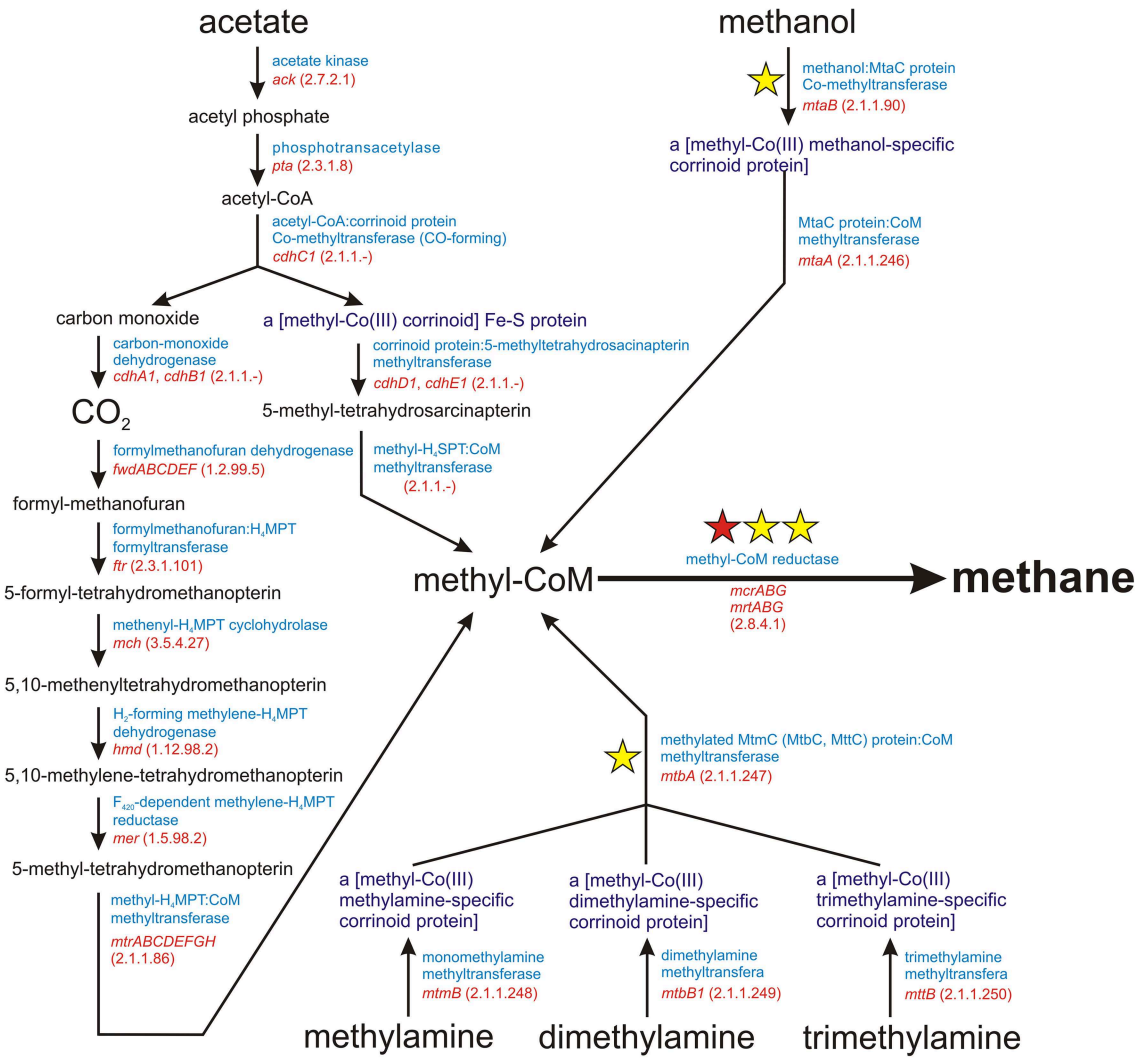

FIGURE 1 | Schematic diagram of the superpathway of methanogenesis. E.C. numbers for particular enzymes are shown in parentheses. The red star indicates the *mcrA* gene encoding subunit α of a methyl-coenzyme M reductase I, which is commonly used as a molecular marker for the detection of methanogens. The yellow stars denote molecular markers developed in this study, for which sets of PCR primers were designed.

Podlaski (Poland) and (ii) an effluent sludge tank from a one-stage wastewater treatment plant anaerobic digester (WD) at MPWIK Pulawy (Poland). In both cases, the samples were centrifuged (8000 g, 4°C, 15 min) and the pellets immediately stored in dry ice prior to DNA extraction.

DNA Extraction and Purification

DNA was isolated from anaerobic digester samples using a modified bead beating protocol. 1 g of pellet material (containing solids and microorganisms) was resuspended in 2 ml of lysis buffer [100 mM Tris-HCl (pH 8.0), 100 mM sodium EDTA (pH 8.0), 100 mM sodium phosphate (pH 8.0), 1.5 M NaCl, 1% (w/v) CTAB] (Zhou et al., 1996). The cells were then disrupted by a 5-step bead beating protocol performed at 1800 rpm (4 × 15 s) and 3200 rpm (1 × 15 s) (MiniBeadBeater 8) using 0.8 g of zirconia/silica beads (ø 0.5 mm, BioSpec). After each round of bead beating the sample was centrifuged (8000 g, 5 min, 4°C), the supernatant retained, and the pellet resuspended in fresh lysis buffer. In addition, after the third round of bead beating, the samples were freeze/thawed five times. The supernatant from each round was extracted with phenol-chloroform-isoamyl alcohol [25:24:1 (vol)]. DNA was then precipitated with one volume of isopropanol, 0.1 volume of 3 M sodium acetate (pH 5.2), recovered by centrifugation at 13,000 g for 20 min, and the pellets washed twice with 70% (v/v) ethanol before resuspending in TE buffer.

The prepared DNA was purified to remove proteins, humic substances, and other impurities by cesium chloride density gradient centrifugation. The concentration and quality of the purified DNA were estimated using a NanoDrop 2000c spectrophotometer (NanoDrop Technologies) and by agarose gel electrophoresis. The applied method yielded highly pure DNA

TABLE 1 | Oligonucleotide primers (specific to *mcrB*, *mcrG*, *mtaB*, and *mtbA* genes) and PCR conditions.

Gene	Name and sequence of the oligonucleotide primer*	PCR product size	PCR conditions**
mcrB	LMCRB: 5′- TWYCARGGHYTVAAYGC -3′ RMCRB: 5′- CCDCCDCCDCCRTARAT -3′	∼392 bp	96°C–30 s; 56°C–30 s; 72°C–40 s 39 cycles
mcrG	LMCRG1: 5′-CAYCCDCCDYTNGADGARATGGA-3′ RMCRG1: 5′-TCRAACATYANWCCRTYYTCRTC-3′	∼356 bp	96°C–30 s; 56°C–30 s; 72°C–35 s 39 cycles
mtaB	LMTAB: 5′- CARGCHAAYACYGCMATGTT -3′ RMTAB: 5′- CYTGDGGRTCYCKGTA -3′	∼436 bp	96°C–30 s; 56°C–30 s; 72°C–40 s 39 cycles
mtbA	LMTBA: 5′- TTCTCCCTTGCMCAGCA -3′ RMTBA: 5′- ACWGGRTCVAGRTTWCC -3′	∼413 bp	96°C–30 s; 55°C–30 s; 72°C–40 s 39 cycles

**IUPAC code: A (adenine), C (cytosine), G (guanine), T (thymine), R (A or G), Y (C or T), W (A or T), K (G or T), M (A or C), D (A or G or T), H (A or C or T), V (A or C or G), N (A or C or G or T).*

***PCR conditions were specified for Taq DNA polymerase (Qiagen). The applicability of other (high fidelity) polymerases [i.e., Phusion High-Fidelity DNA Polymerase (Thermo Scientific) and KAPA polymerase (Kapa Biosystems)] was also confirmed.*

($A_{260}/A_{280} = 1.8$; $A_{260}/A_{230} = 1.9$) suitable for metagenomic analysis.

Library Preparation and Amplicon Sequencing

PCR products were analyzed by electrophoresis on 2% agarose gels (1x TAE buffer) with ethidium bromide staining. The amplified DNA fragments from replicate PCRs were pooled and then purified using Agencourt AMPure XP beads (Beckman Coulter). Approximately 250 ng of each amplicon was used for library preparation with an Illumina TruSeq DNA Sample Preparation Kit according to the manufacturer's protocol, except that the final library amplification was omitted. The libraries were verified using a 2100 Bioanalyzer (Agilent) High-Sensitivity DNA Assay and KAPA Library Quantification Kit for the Illumina. Sequencing of amplicon DNA was performed using an Illumina MiSeq (MiSeq Reagent Kit v2, 500 cycles) with a read length of 250 bp.

Designing Oligonucleotide Primers Specific for *mcrB, mcrG, mtaB,* and *mtbA* Genes

Data from the NCBI database were used to design degenerate primers to amplify *mcrB*, *mcrG*, *mtaB*, and *mtbA* gene fragments. A two-stage design strategy was employed. First, nucleotide sequences of genes annotated as *mcrB*, *mcrG*, *mtaB*, and *mtbA* were retrieved from the NCBI database. These sequences were then used as a query to recover additional gene sequences that were not annotated or were incorrectly annotated. Nucleotide sequences of particular genes were retrieved from genome sequences (completed and drafts) of methanogenic *Archaea* available on Jan 10th 2014. For each gene, multiple sequence alignments were prepared using ClustalW (Chenna et al., 2003) and MEGA6 (Tamura et al., 2013). Conserved regions within the obtained alignments were identified and used in the design of appropriate degenerate primers. Primer pairs with the lowest degree of degeneracy and producing amplicons not exceeding 500 bp were chosen. This size limit was imposed so that both 454 pyrosequencing and Illumina platforms could be used for amplicon sequencing.

In silico PCR with iPCRess (Slater and Birney, 2005) was done on dataset consisting of complete microbial genomes (5274 in total) obtained from NCBI database. We allowed for two mismatches per primer and required that both primers match and the product length is similar (±50 nucleotides) to expected length. The only exception was the set of *mcrG*-specific primers, that required allowance of 4 mismatches, due to bigger length of their sequences.

Bioinformatic Analysis of High-throughput Amplicon Sequencing Data

For each selected protein family a reference set of sequences was assembled from the results of searches of the NCBI NR database with BLAST software (Altschul et al., 1997), using known archaeal members of each family as query sequences and an *E*-value of 0.001 as the threshold. These reference sets were not specifically curated to allow the presence of false positives such as proteins from *Bacteria* or *Eukarya*. We consider them false positives, as the process of methanogenesis is limited only to *Archaea*. A presence of the sequences more similar to bacterial homologs of marker proteins than to archaeal ones would indicate low specificity of the designed primers. We specifically screened for such a cases after phylogenetic placement of reads.

Paired-end reads were merged with FLASH (Magoc and Salzberg, 2011) and then mapped to reference sets using BLASTX, again with an *E*-value of 0.001 as the threshold. Translated sequences were extracted from the BLAST high scoring pairs (HSPs), and reads with no hits, containing stop codons (presumably generated by frameshifts) or sequences shorter than 30 amino acids were discarded. Therefore, for each primer pair, a reference set of known protein sequences was obtained, as well as a set of protein sequences derived from sequenced amplicons. The latter are referred as "inferred peptides" as they correspond to fragments of target proteins. The ratio of number of inferred peptides to number of all merged reads is the measure of primer sensitivity.

Sequences from the reference sets were aligned with MAFFT (Katoh and Standley, 2013) using default options. Based on these alignments, a maximum likelihood phylogenetic tree was constructed for each protein family using FastTree software (Price et al., 2009) with the Gamma20 model. Sequences inferred from reads were then merged with sequences from reference sets for each protein family and aligned with MAFFT as described above. The resulting alignment and the phylogenetic tree of reference sequences were used as the input to the Evolutionary Placement Algorithm, part of the RAxML package (Stamatakis, 2014). The reads were placed on the reference phylogenetic tree using the PROTGAMMAWAG substitution model. Placements were subsequently trimmed with guppy software (Matsen et al.,

2010) using 0.01 as the minimal threshold mass from the leaf to the root. Results underwent guppy "fat" conversion to the PhyloXML file format (Han and Zmasek, 2009) and were then visualized using Archeopteryx software (Han and Zmasek, 2009). The visualization resulted in coloring branches that point to a node or a leaf to which reads were assigned in red. All other branches were colored in black.

Amplicons from 16S rDNA were processed differently. All sequence reads were processed via the NGS analysis pipeline of the SILVA rRNA gene database project (SILVAngs 1.2) (Quast et al., 2013). Using the SILVA Incremental Aligner [SINA SINA v1.2.10 for ARB SVN (revision 21008)] (Pruesse et al., 2012), each read was aligned against the SILVA SSU rRNA SEED and quality controlled (Quast et al., 2013). Reads shorter than 50 aligned nucleotides and reads with more than 2% ambiguities or 2% homopolymers, were excluded from further processing. In addition, putative contaminants and artifacts, and reads with low alignment quality (50 alignment identity, 40 alignment score reported by the SINA), were identified and excluded from downstream analysis.

The classification of each operational taxonomic unit (OTU) reference read was mapped onto all reads that were assigned to the respective OTU. This yielded quantitative information (number of individual reads per taxonomic path), within the limitations of PCR and sequencing technique biases, as well as multiple rRNA operons. Reads without any BLAST hits or those with weak BLAST hits, where the function "(% sequence identity + % alignment coverage)/2" did not exceed the value of 93, remained unclassified.

Raw sequences obtained in this study have been deposited in the SRA (NCBI) database with the accession number PRJNA284604.

Results and Discussion

General Diversity of *Archaea* in Anaerobic Digesters—16S rRNA and *mcrA* Molecular Marker Analyses

In the analyses performed in this study metagenomic DNA was extracted from two samples of microbial consortia involved in biogas production (and therefore reach in methanogens). For the description of the overall diversity of *Archaea* in the analyzed samples, 16S rDNA-specific primers were used (Klindworth et al., 2013). This analysis revealed that methanogens are dominant microorganisms in the studied anaerobic digesters (74% for AD and 95% for WD) and include representatives of four of the seven methanogenic orders (i.e., *Methanosarcinales, Methanomicrobiales, Methanobacteriales*, *Methanomassiliicoccales*). The most abundant methanogens in both digesters were *Methanosarcinales*, represented by the families *Methanosaetaceae* (~38%) and *Methanosarcinaceae* (~18%), followed by *Methanomicrobiaceae* (~20%) of the *Methanomicrobiales* order (**Figure 2**).

Abundant non-methanogenic *Archaea* such as Miscellaneous Crenarchaeotic Group (MCG) (11%) and *Halobacteria* (7%) represented by Deep Sea Euryarchaeotic Group (DSEG) and Deep Sea Hydrothermal Vent Gp 6 (DHVEG-6) were also detected in the AD sample (**Figure 2**). These groups are

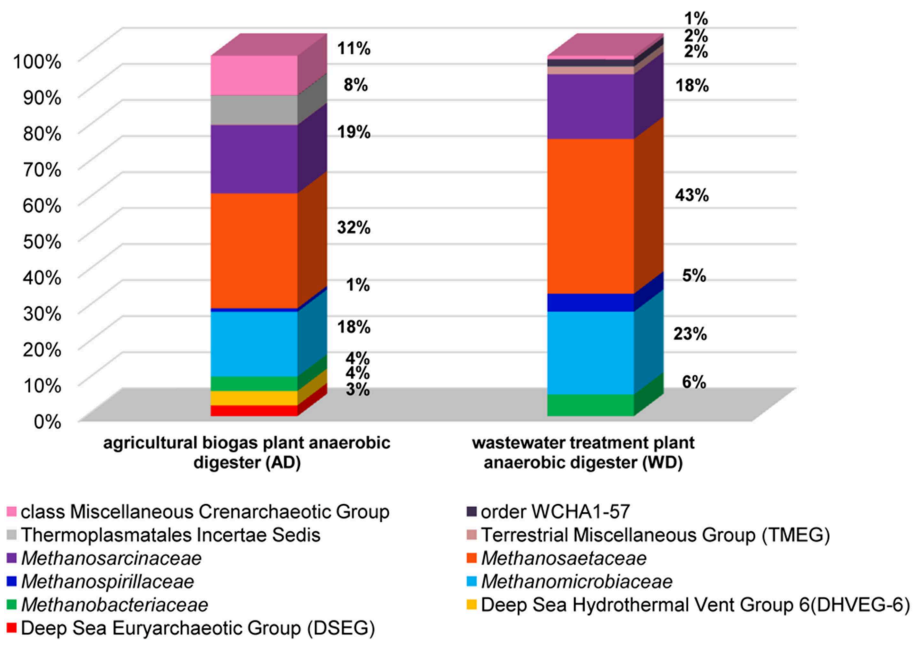

FIGURE 2 | Relative abundance of archaeal OTUs defined using the 16S rRNA gene hyper-variable region V3V4. The bar chart shows the diversity of *Archaea* at the lowest reliable taxonomic level (where possible the default family is denoted in the key). AD, agricultural biogas plant anaerobic digester; WD, wastewater treatment plant anaerobic digester.

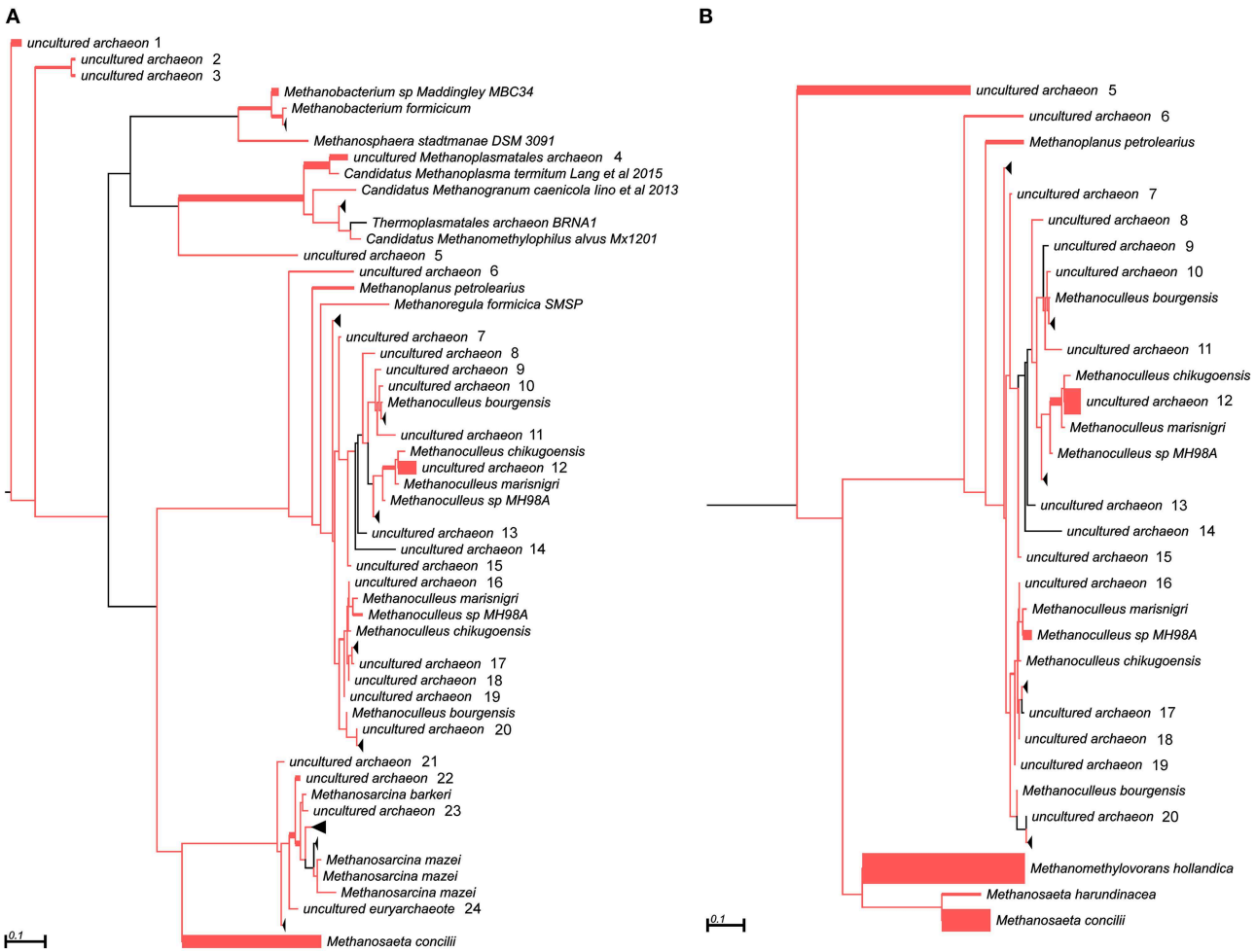

FIGURE 3 | Phylogenetic placement of *mcrA* amplicons from AD (A) and WD (B) samples. The width of the red branches corresponds to the number of unique *mcrA* amplicon sequence reads in that particular branch (this can be either a leaf or node). The collapse of some branches (mapped to uncultured archaeons) to increase the clarity of the trees is indicated by a black triangle. Numbers next to the entries "uncultured archeon" indicate the same microorganism on both trees.

phylogenetically diverse and there is a little knowledge of their ecology and metabolism, however it seems that MCG archaeons are able to ferment wide variety of recalcitrant substrates (Kubo et al., 2012) and DSEG are positively correlated with putative ammonia-oxidizing *Thaumarchaeota* (Restrepo-Ortiz and Casamayor, 2013).

In addition to the 16S rRNA marker, the *mcrA* gene was used for taxonomic profiling of methanogenic communities in both digesters. The *mcrA* gene fragments amplified using primers MLf/MLr (Luton et al., 2002) were sequenced and analyzed. More than half of the sequences (57%) amplified from the AD sample were assigned to uncultured *Archaea*, belonging to the *Methanomassiliicoccales* (23%), *Methanomicrobiales* (13%), *Methanobacteriales* (11%) and *Methanosarcinales* (10%) orders (**Figure 3**), suggesting dominance of hydrogenotrophic methanogens over acetoclastic *Archaea*. The most abundant genera in AD were *Methanobacterium* sp. Maddingley MBC34 (11%) followed by *Methanosaeta concilli* (9%) and *Methanoculleus* spp. (4%) (**Figure 3**). Similarly in WD, the majority of the *mcrA* amplicons were classified as uncultured *Archaea* belonging to orders *Methanomicrobiales* (27%) and *Methanomassiliicoccales* (7%) (**Figure 3**), while at the genus level most of the methanogens were identified as *Methanometylovorans hollandica* (21%), *Methanosaeta concilli* (16%), *Methonoculleus* spp. (12%), or *Methanoplanus petrolearius* (3%) (**Figure 3**).

The results obtained for both marker genes (16S rRNA and *mcrA*) only partially overlapped, probably due to differences in primer affinities and variation in the gene copy numbers. This observation is in agreement with a previous report showing that these two marker genes generate different taxonomic profiles (Wilkins et al., 2015). Therefore, for a greater insight into the structure of methanogenic communities and to verify the obtained results, novel molecular markers specific for other methanogenesis-linked genes were developed.

TABLE 2 | Summary of bioinformatic analysis of sequenced *mcrA*, *mcrB*, *mcrG*, *mtaB,* and *mtbA* amplicons.

Sample*	Number of paired reads	Number of merged reads	Number of inferred peptides**	Primer sensitivity (% of correct product)
mcrA_AD	17,365	12,816	11,931	93
mcrA_WD	9277	4318	2572	59
mcrB_AD	32,094	23,188	21.939	94
mcrB_WD	50,485	40,242	25,035	57
mcrG_AD	42,185	29,330	21,988	74
mcrG_WD	34,945	28,660	18,272	63
mtaB_AD	26,500	20,753	19,163	92
mtaB_WD	36,148	15,293	13,231	86
mtbA_AD	33,770	22,852	10,027	43
mtbA_WD	31,601	19,150	10,961	57

AD, agricultural biogas plant anaerobic digester; WD, wastewater treatment plant anaerobic digester.

**Inferred peptides number denote how many peptides that are sufficiently long and similar to a target protein can be extracted from the reads. Percent of correct product is the ratio between number of peptides and number of reads.*

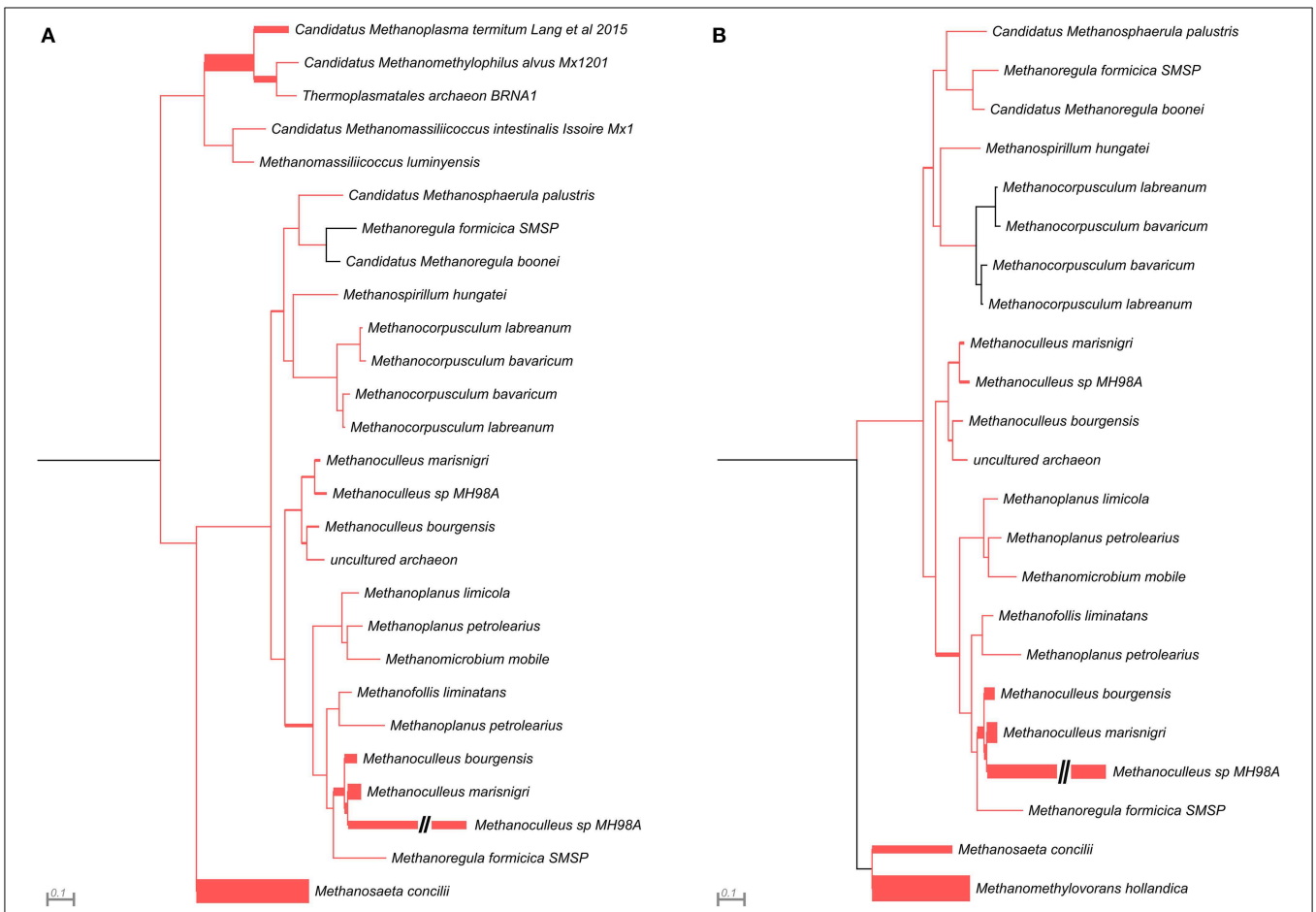

FIGURE 4 | Phylogenetic placement of *mcrB* amplicons from AD (A) and WD (B) samples. The width of the red branches corresponds to the number of unique *mcrB* amplicon sequence reads in that particular branch (this can be either a leaf or node). The leaf for *Methanoculleus* sp. MH98A was shortened, as indicated by two slashes.

Development of *mcrB*-, *mcrG*-, *mtaB*-, and *mtbA*-specific Primers

For the design of degenerate primers specific for the *mcrB*, *mcrG*, *mtaB*, and *mtbA* genes, sequences were retrieved from the NCBI database [36 nucleotide sequences for *mcrB* (**Figure S1**), 61 for *mcrG* (**Figure S2**), 26 for *mtaB* (**Figure S3**) and 13 for *mtbA* (**Figure S4**)]. The *mcrG* gene turned out to be highly variable, which hampered primer design. Therefore, phylogenetic analysis was performed to distinguish conserved clusters among the analyzed *mcrG* genes. Two groups of *mcrG* sequences were distinguished: (i) MCR_G1 (grouping 35 *mcrG* genes of *Methanobacterium* spp., *Methanobrevibacter* spp., *Methanocaldococcus* spp., *Methanococcus* spp., *Methanothermobacter* spp., *Methanothermococcus* spp., *Methanothermus* spp., *Methanotorris* spp., *Methanosphaera* spp.), and (ii) MCR_G2 (grouping 26 *mcrG* genes of *Methanocella* spp., *Methanococcoides* spp., *Methanocorpusculum* spp., *Methanoculleus* spp., *Methanohalobium* spp., *Methanohalophilus* spp., *Methanolobus* spp., *Methanoplanus* spp., *Methanopyrus* spp., *Methanoregula* spp., *Methanosalsum* spp., *Methanosarcina* spp., *Methanospirillum* spp., *Methanosphaerula* spp.) (**Figure S5**). The nucleotide sequences of *mcrG* genes from particular groups were then used to design specific primer pairs.

For the subsequent functional analyses, 28 primers were selected for synthesis, including 6 for *mcrB*, 9 for *mcrG* and *mtaB*, and 4 for *mtbA*. The initial PCRs were performed with all primer pairs and DNA samples from the AD and WD fermenters as templates. The primer pairs giving the strongest amplification products of the expected size were selected for further analysis. The PCR products were cloned in vector pGEM-T Easy and then inserts of five random clones from each experimental set were sequenced using the sequencing primer M13 Reverse. The BLAST analysis of the resulting sequences revealed the specificity of each primer pair. At this stage, all primers designed for amplification of the *mcrG* genes of MCR_G2 group methanogens were rejected due to low specificity. Based on those analyses and taking into account the amplification yield, four primer pairs were selected and the optimal PCR conditions were determined (**Table 1**). Primer pairs specificity was also initially confirmed by *in silico* PCR analysis using 5274 complete microbial genomes (Table S1).

Since the panel of primers developed in this study was designed to be used in the high-throughput amplicon sequencing analysis of methanogenic communities, their selectivity was tested in the high-throughput sequencing experiments.

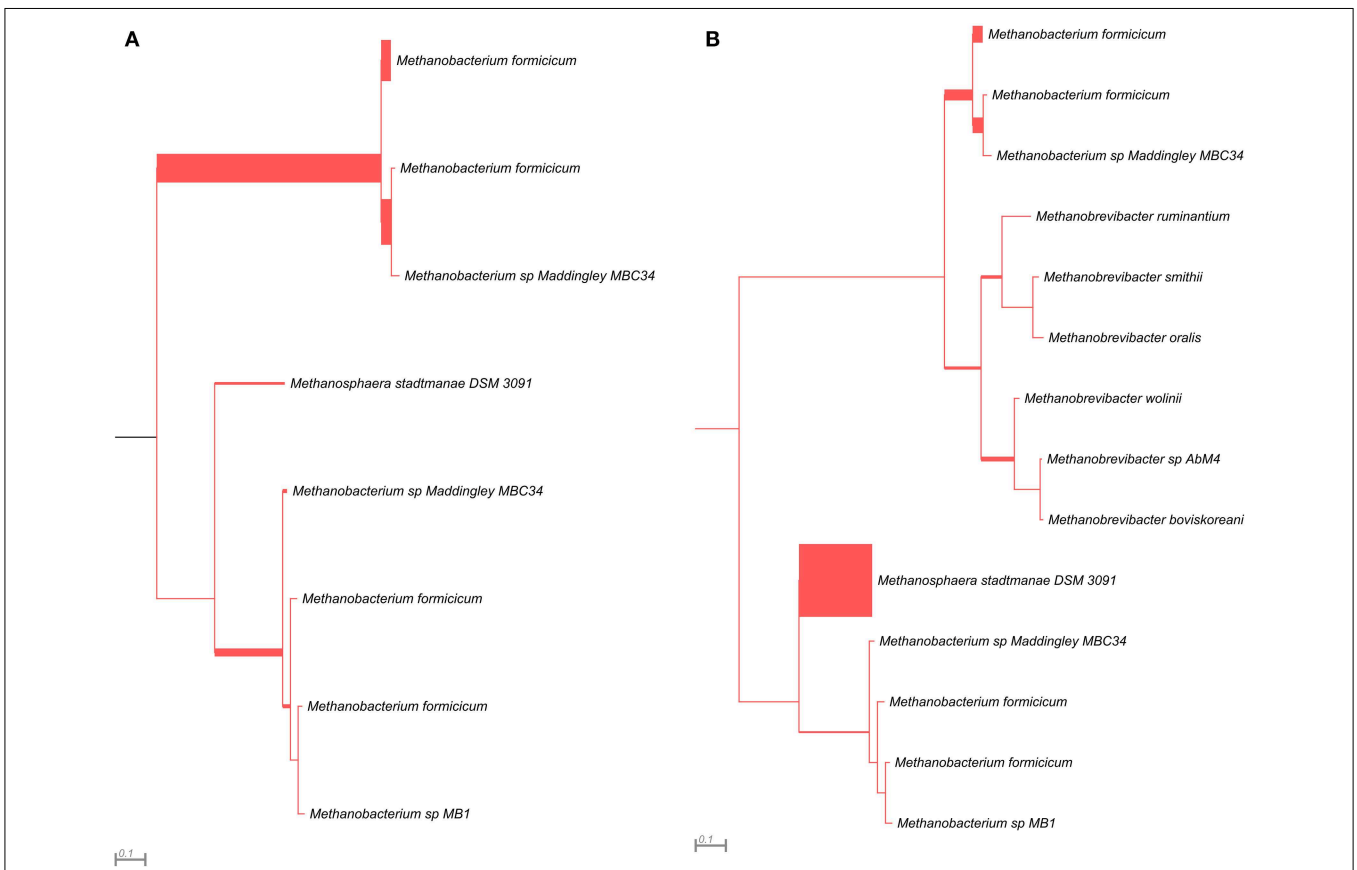

FIGURE 5 | Phylogenetic placement of *mcrG* amplicons from AD (A) and WD (B) samples. The width of the red branches corresponds to the number of unique *mcrG* amplicon sequence reads in that particular branch (this can be either a leaf or node).

Analysis of the Selectivity of the *mcrB*- and *mcrG*-specific Primers

DNA fragments were amplified using the developed primer pairs with template DNAs isolated from the anaerobic reactors AD and WD. The raw sequence data obtained from Illumina sequencing were processed and analyzed (**Table 2**).

This analysis revealed that LMCRB/RMCRB primers, designed to the *mcrB* gene, amplified DNA fragments comprising sequences representing four methanogenic orders: *Methanobacteriales*, *Methanomassiliicoccales*, *Methanomicrobiales*, and *Methanosarcinales* (**Figure 4**). The dominant genus in both digesters was *Methanoculleus* spp. (48% for AD and 67% for WD), with *M. marisnigri* as the most abundant species (37 and 53%, respectively). This finding remains in good agreement with previous observations showing that the predominant order in biogas-producing microbial communities in anaerobic digesters is usually *Methanomicrobiales*, and the most abundant species is hydrogenotrophic *M. marisnigri* (Wirth et al., 2012). Moreover, in AD, 27% of sequences were classified as uncultured *Methanomassiliicoccales* (with 4% described as *Candidatus* Methanoplasma termitum) and 17% as *Methanosaeta concilli*. The second and third most abundant methanogens in WD were *Methanomethylovorans hollandica* (19%) and *Methanosaeta concilli* (6%), respectively (**Figure 4**).

The *mcrG* gene fragments (amplified with primers LMCRG1/RMCRG1) comprised sequences representing five methanogenic orders: *Methanobacteriales*, *Methanococcales*, *Methanomicrobiales*, *Methanomassiliicoccales*, and *Methanosarcinales*. However, representatives of hydrogenotrophic *Methanobacteriales* were absolutely dominant in both digesters (**Figure 5**). The most abundant OTU_{mcrG} in AD was assigned to *Methanobacterium* spp. (97%) (with 7% mapped to *M. formicicum)*, while WD was dominated by *Methanosphaera stadtmanae* (54%) and *Methanobacterium* spp. (39%) (with 28% mapped to *M. formicicum)* and *Methanobrevibacter* spp. (5%) (**Figure 5**).

The above analysis revealed that primers LMCRB/RMCRB are highly specific for *mcrB* genes of methanogens. Therefore, similarly to the commonly employed *mcrA*-specific primers, they may be used for an overall characterization of the taxonomic structure of methanogenic communities. The application of both *mcrA* and *mcrB* molecular markers permits cross-checking and should give a deeper and more detailed insight into the taxonomic structure of various methanogenic communities. It is worth mentioning that the results obtained using the newly developed primers for *mcrB* were partially consistent with those obtained by *mcrA* analysis, and confirmed that the hydrogenotrophic pathway of methane synthesis is employed in the analyzed environments. Moreover, these results demonstrated the importance of the newly described seventh order of methanogenic *Methanomassiliicoccales* (Iino et al., 2013; Borrel et al., 2014) in the analyzed biogas digesters (**Figure 6**, Table S2).

The *mcrG* primers LMCRG1/RMCRG1 permitted the analysis of the minority of methanogenic *Archaea* that were not dominant in *mcrA*/*mcrB* analysis (except *Methanobacterium* for the *mcrA* marker). Therefore, the obtained results were not consistent with those obtained by *mcrA* and *mcrB* analyses. This is the consequence of the fact that the primers LMCRG1/RMCRG1 are specific only for the previously described MCR_G1 group of sequences (**Figure S5**) and their use could generate programmed bias (**Figure 6**, Table S2).

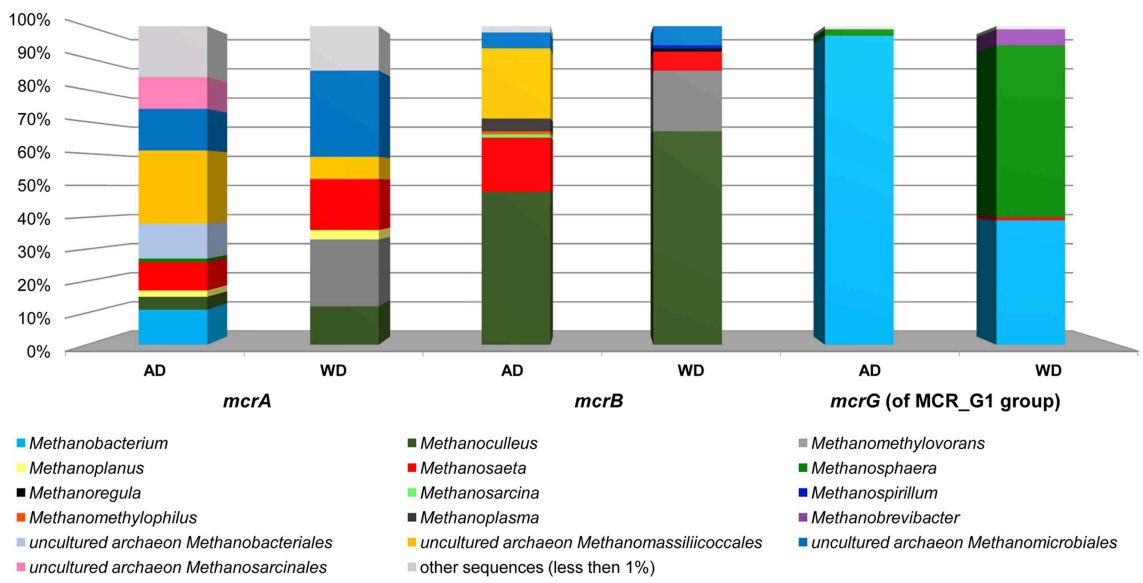

FIGURE 6 | Relative abundance of archaeal OTUs defined using the *mcrA* and *mcrB* and *mcrG* (of MCR_G1 group) gene fragments. The bar chart shows the diversity of *Archaea* at the lowest reliable taxonomic level (mostly genus). AD, agricultural biogas plant anaerobic digester; WD, wastewater treatment plant anaerobic digester.

Analysis of the Selectivity of the *mtaB*- and *mtbA*-specific Primers

In the course of this study, two other marker genes (*mtaB* and *mtbA*) specific for methylotrophic methanogens were selected and primer pairs developed. High-throughput sequencing of amplicons obtained using *mtaB* primers LMTAB/RMTAB detected sequences representing only two orders, *Methanosarcinales* and *Methanobacteriales*. In AD, 76% of sequences were assigned to *Methanosarcina* spp. [including *M. barkeri* (69%) and *M. mazei* (7%)] and 23% to *Methanosphaera stadtmanae*. Reactor WD was dominated by *M. hollandica* (94%), followed by *M. stadtmanae* (6%). In comparison, use of *mtbA*-specific primers LMTBA/RMTBA detected sequences mostly belonging to the *Methanosarcinales*, with two dominating species: *M. barkeri* (99%) in AD and *M. hollandica* (99%) in WD. Single sequences in WD and AD were assigned to *Halobacteriales* and *Methanomassiliicoccales*, respectively.

Sequencing of the *mtaB* and *mtbA* amplicons clearly indicated that in the analyzed digesters, *Methanosarcinales* are mainly responsible for the utilization of methylamines, while the conversion of methanol to methane is additionally performed by *M. stadtmanae* (of *Methanobacteriales*), which is consistent with previous studies (Fricke et al., 2006; Liu and Whitman, 2008).

Conclusions

Four novel molecular markers were designed and tested for the detection and taxonomic analyses of methanogenic communities. Primers specific to the *mcrB* and *mcrG* genes (present in all methanogens), as well as the *mtaB* and *mtbA* genes, characteristic for methylotrophic methanogens, were developed. High-throughput sequencing of the amplicons obtained using these primers revealed their high specificity and indicated that these marker genes could be used for taxonomic profiling of methanogenic consortia.

The *mcrB* and *mcrG* molecular markers increased the resolution of high-throughput amplicon sequencing analyses of methanogenic communities that until now have only been investigated using the *mcrA* gene. The use of *mcrA*, *mcrB*, and *mcrG*, together with the 16S rRNA gene marker, should give a much broader overview of the taxonomic diversity of complex methanogenic communities. In addition, the analysis of two other marker genes (*mtaB* and *mtbA*) can provide an insight into the metabolic potential of the analyzed methanogens, since they permit the detection and analysis of an enigmatic group of methylotrophic methanogens, which are able to produce methane from methanol or methylamines.

Acknowledgments

This work was supported by the National Centre for Research and Development (Poland) grant no. 177481, as well as by the EU European Regional Development Fund, the Operational Program Innovative Economy 2007–2013, agreement POIG.01.01.02-14-054/09-00. Some experiments were carried out with the use of CePT infrastructure financed by the European Union—the European Regional Development Fund [Innovative economy 2007–13, Agreement POIG.02.02.00-14-024/08-00].

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00694

Figure S1 | Alignment of the conserved fragments of the *mcrB* genes of 36 methanogens used in the design of primers LMCRB and RMCRB.

Figure S2 | Alignment of the conserved fragments of the *mcrG* genes (of MCR_G1 group) of 35 methanogens used in the design of primers LMCRG1 and RMCRG1.

Figure S3 | Alignment of the conserved fragments of the *mtaB* genes of 26 methanogens used in the design of primers LMTAB and RMTAB.

Figure S4 | Alignment of the conserved fragments of the *mtbA* genes of 13 methanogens used in the design of primers LMTBA and RMTBA.

Figure S5 | Phylogenetic tree for *mcrG* nucleotide sequences (from NCBI database). The tree was constructed using the maximum-likelihood algorithm. Statistical support for the internal nodes was determined by 1000 bootstrap replicates and values of >50% are shown.

References

Akuzawa, M., Hori, T., Haruta, S., Ueno, Y., Ishii, M., and Igarashi, Y. (2011). Distinctive responses of metabolically active microbiota to acidification in a thermophilic anaerobic digester. *Microb. Ecol.* 61, 595–605. doi: 10.1007/s00248-010-9788-1

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Bapteste, E., Brochier, C., and Boucher, Y. (2005). Higher-level classification of the *Archaea*: evolution of methanogenesis and methanogens. *Archaea* 1, 353–363. doi: 10.1155/2005/859728

Borrel, G., O'Toole, P. W., Harris, H. M., Peyret, P., Brugere, J. F., and Gribaldo, S. (2013). Phylogenomic data support a seventh order of methylotrophic methanogens and provide insights into the evolution of methanogenesis. *Genome Biol. Evol.* 5, 1769–1780. doi: 10.1093/gbe/evt128

Borrel, G., Parisot, N., Harris, H. M., Peyretaillade, E., Gaci, N., Tottey, W., et al. (2014). Comparative genomics highlights the unique biology of *Methanomassiliicoccales*, a *Thermoplasmatales*-related seventh order of methanogenic archaea that encodes pyrrolysine. *BMC Genomics* 15:679. doi: 10.1186/1471-2164-15-679

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4516–4522. doi: 10.1073/pnas.1000080107

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., et al. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31, 3497–3500. doi: 10.1093/nar/gkg500

Costa, K. C., and Leigh, J. A. (2014). Metabolic versatility in methanogens. *Curr. Opin. Biotechnol.* 29, 70–75. doi: 10.1016/j.copbio.2014.02.012

Delbes, C., Moletta, R., and Godon, J. (2001). Bacterial and archaeal 16S rDNA and 16S rRNA dynamics during an acetate crisis in an anaerobic digestor ecosystem. *FEMS Microbiol. Ecol.* 35, 19–26. doi: 10.1016/S0168-6496(00)00107-0

Denman, S. E., Tomkins, N. W., and McSweeney, C. S. (2007). Quantitation and diversity analysis of ruminal methanogenic populations in response to the antimethanogenic compound bromochloromethane. *FEMS Microbiol. Ecol.* 62, 313–322. doi: 10.1111/j.1574-6941.2007.00394.x

Diaz, E. E., Stams, A. J., Amils, R., and Sanz, J. L. (2006). Phenotypic properties and microbial diversity of methanogenic granules from a full-scale upflow anaerobic sludge bed reactor treating brewery wastewater. *Appl. Environ. Microbiol.* 72, 4942–4949. doi: 10.1128/AEM.02985-05

Dridi, B., Fardeau, M. L., Ollivier, B., Raoult, D., and Drancourt, M. (2012). *Methanomassiliicoccus luminyensis* gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* 62, 1902–1907. doi: 10.1099/ijs.0.033712-0

Escamilla-Alvaradoa, C., Ríos-Lealb, E., Ponce-Noyolac, M. T., and Poggi-Varaldo, H. M. (2012). Gas biofuels from solid substrate hydrogenogenic–methanogenic fermentation of the organic fraction of municipal solid waste. *Process Biochem.* 47, 1572–1587. doi: 10.1016/j.procbio.2011.12.006

Farkas, J. A., Picking, J. W., and Santangelo, T. J. (2013). Genetic techniques for the *Archaea*. *Annu. Rev. Genet.* 47, 539–561. doi: 10.1146/annurev-genet-111212-133225

Ferry, J. G. (2010). The chemical biology of methanogenesis. *Planet. Space Sci.* 58, 1775–1783. doi: 10.1016/j.pss.2010.08.014

Fricke, W. F., Seedorf, H., Henne, A., Kruer, M., Liesegang, H., Hedderich, R., et al. (2006). The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. *J. Bacteriol.* 188, 642–658. doi: 10.1128/JB.188.2.642-658.2006

Friedrich, C. G., Bardischewsky, F., Rother, D., Quentmeier, A., and Fischer, J. (2005). Prokaryotic sulfur oxidation. *Curr. Opin. Microbiol.* 8, 253–259. doi: 10.1016/j.mib.2005.04.005

Gao, B., and Gupta, R. S. (2007). Phylogenomic analysis of proteins that are distinctive of *Archaea* and its main subgroups and the origin of methanogenesis. *BMC Genomics* 8:86. doi: 10.1186/1471-2164-8-86

Garcia-Maldonado, J. Q., Bebout, B. M., Everroad, R. C., and Lopez-Cortes, A. (2015). Evidence of novel phylogenetic lineages of methanogenic archaea from hypersaline microbial mats. *Microb. Ecol.* 69, 106–117. doi: 10.1007/s00248-014-0473-7

Hallam, S. J., Girguis, P. R., Preston, C. M., Richardson, P. M., and Delong, E. F. (2003). Identification of methyl coenzyme M reductase A (*mcrA*) genes associated with methane-oxidizing archaea. *Appl. Environ. Microbiol.* 69, 5483–5491. doi: 10.1128/AEM.69.9.5483-5491.2003

Han, M. V., and Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10:356. doi: 10.1186/1471-2105-10-356

Iino, T., Tamaki, H., Tamazawa, S., Ueno, Y., Ohkuma, M., Suzuki, K., et al. (2013). *Candidatus* Methanogranum caenicola: a novel methanogen from the anaerobic digested sludge, and proposal of *Methanomassiliicoccaceae* fam. nov. and *Methanomassiliicoccales* ord. nov., for a methanogenic lineage of the class *Thermoplasmata*. *Microbes Environ.* 28, 244–250. doi: 10.1264/jsme2.ME12189

Juottonen, H., Galand, P. E., and Yrjala, K. (2006). Detection of methanogenic *Archaea* in peat: comparison of PCR primers targeting the *mcrA* gene. *Res. Microbiol.* 157, 914–921. doi: 10.1016/j.resmic.2006.08.006

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41:e1. doi: 10.1093/nar/gks808

Kubo, K., Lloyd, K. G., Biddle, J. F., Amann, R., Teske, A., and Knittel, K. (2012). *Archaea* of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J.* 6, 1949–1965. doi: 10.1038/ismej.2012.37

Kuroda, K., Hatamoto, M., Nakahara, N., Abe, K., Takahashi, M., Araki, N., et al. (2014). Community composition of known and uncultured archaeal lineages in anaerobic or anoxic wastewater treatment sludge. *Microb. Ecol.* 69, 586–596. doi: 10.1007/s00248-014-0525-z

Kushner, S. R. (1978). "An improved method for transformation of *E. coli* with ColE1 derived plasmids," in *Genetic Engineering*, eds H. B. Boyer and S. Nicosia (Amsterdam: Elsevier/North-Holland), 17–23.

Li, Y. F., Nelson, M. C., Chen, P. H., Graf, J., Li, Y., and Yu, Z. (2014). Comparison of the microbial communities in solid-state anaerobic digestion (SS-AD) reactors operated at mesophilic and thermophilic temperatures. *Appl. Microbiol. Biotechnol.* 99, 969–980. doi: 10.1007/s00253-014-6036-5

Liu, Y., and Whitman, W. B. (2008). Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. *Ann. N.Y. Acad. Sci.* 1125, 171–189. doi: 10.1196/annals.1419.019

Lueders, T., Chin, K. J., Conrad, R., and Friedrich, M. (2001). Molecular analyses of methyl-coenzyme M reductase alpha-subunit (*mcrA*) genes in rice field soil and enrichment cultures reveal the methanogenic phenotype of a novel archaeal lineage. *Environ. Microbiol.* 3, 194–204. doi: 10.1046/j.1462-2920.2001.00179.x

Luton, P. E., Wayne, J. M., Sharp, R. J., and Riley, P. W. (2002). The *mcrA* gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology* 148, 3521–3530. doi: 10.1099/00221287-148-11-3521

Magoc, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538. doi: 10.1186/1471-2105-11-538

Nettmann, E., Bergmann, I., Mundt, K., Linke, B., and Klocke, M. (2008). *Archaea* diversity within a commercial biogas plant utilizing herbal biomass determined by 16S rDNA and *mcrA* analysis. *J. Appl. Microbiol.* 105, 1835–1850. doi: 10.1111/j.1365-2672.2008.03949.x

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Pruesse, E., Peplies, J., and Glockner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 28, 1823–1829. doi: 10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rademacher, A., Zakrzewski, M., Schluter, A., Schonberg, M., Szczepanowski, R., Goesmann, A., et al. (2012). Characterization of microbial biofilms in a thermophilic biogas system by high-throughput metagenome sequencing. *FEMS Microbiol. Ecol.* 79, 785–799. doi: 10.1111/j.1574-6941.2011.01265.x

Restrepo-Ortiz, C. X., and Casamayor, E. O. (2013). Environmental distribution of two widespread uncultured freshwater *Euryarchaeota* clades unveiled by specific primers and quantitative PCR. *Environ. Microbiol. Rep.* 5, 861–867. doi: 10.1111/1758-2229.12088

Sakai, S., Imachi, H., Sekiguchi, Y., Tseng, I. C., Ohashi, A., Harada, H., et al. (2009). Cultivation of methanogens under low-hydrogen conditions by using the coculture method. *Appl. Environ. Microbiol.* 75, 4892–4896. doi: 10.1128/AEM.02835-08

Sambrook, J., and Russell, D. W. (2001). *Molecular Cloning: A Laboratory Manual.* New York. NY: Cold Spring Harbor Laboratory Press.

Sawayama, S., Tsukahara, K., and Yagishita, T. (2006). Phylogenetic description of immobilized methanogenic community using real-time PCR in a fixed-bed anaerobic digester. *Bioresour. Technol.* 97, 69–76. doi: 10.1016/j.biortech.2005.02.011

Schlüter, A., Bekel, T., Diaz, N. N., Dondrup, M., Eichenlaub, R., Gartemann, K. H., et al. (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J. Biotechnol.* 136, 77–90. doi: 10.1016/j.jbiotec.2008.05.008

Sekiguchi, Y. (2006). Yet-to-be cultured microorganisms relevant to methane fermentation processes. *Microbes Environ.* 21, 1–15. doi: 10.1264/jsme2.21.1

Slater, G. S., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. doi: 10.1186/1471-2105-6-31

Springer, E., Sachs, M. S., Woese, C. R., and Boone, D. R. (1995). Partial gene sequences for the A subunit of methyl-coenzyme M reductase (*mcrI*) as a

phylogenetic tool for the family *Methanosarcinaceae*. *Int. J. Syst. Bacteriol.* 45, 554–559. doi: 10.1099/00207713-45-3-554

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Steinberg, L. M., and Regan, J. M. (2009). *mcrA*-targeted real-time quantitative PCR method to examine methanogen communities. *Appl. Environ. Microbiol.* 75, 4435–4442. doi: 10.1128/AEM.02858-08

Stolze, Y., Zakrzewski, M., Maus, I., Eikmeyer, F., Jaenicke, S., Rottmann, N., et al. (2015). Comparative metagenomics of biogas-producing microbial communities from production-scale biogas plants operating under wet or dry fermentation conditions. *Biotechnol. Biofuels* 8, 14. doi: 10.1186/s13068-014-0193-8

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Watanabe, T., Asakawa, S., Nakamura, A., Nagaoka, K., and Kimura, M. (2004). DGGE method for analyzing 16S rDNA of methanogenic archaeal community in paddy field soil. *FEMS Microbiol. Lett.* 232, 153–163. doi: 10.1016/S0378-1097(04)00045-X

Wilkins, D., Lu, X. Y., Shen, Z., Chen, J., and Lee, P. K. (2015). Pyrosequencing of *mcrA* and archaeal 16S rRNA genes reveals diversity and substrate preferences of methanogen communities in anaerobic digesters. *Appl. Environ. Microbiol.* 81, 604–613. doi: 10.1128/AEM.02566-14

Wirth, R., Kovacs, E., Maroti, G., Bagi, Z., Rakhely, G., and Kovacs, K. L. (2012). Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. *Biotechnol. Biofuels* 5:41. doi: 10.1186/1754-6834-5-41

Yoon, S. H., Turkarslan, S., Reiss, D. J., Pan, M., Burn, J. A., Costa, K. C., et al. (2013). A systems level predictive model for global gene regulation of methanogenesis in a hydrogenotrophic methanogen. *Genome Res.* 23, 1839–1851. doi: 10.1101/gr.153916.112

Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnol. Bioeng.* 89, 670–679. doi: 10.1002/bit.20347

Zakrzewski, M., Bekel, T., Ander, C., Pühler, A., Rupp, O., Stoye, J., et al. (2013). MetaSAMS—a novel software platform for taxonomic classification, functional annotation and comparative analysis of metagenome datasets. *J. Biotechnol.* 167, 156–165. doi: 10.1016/j.jbiotec.2012.09.013

Zhou, H. W., Li, D. F., Tam, N. F., Jiang, X. T., Zhang, H., Sheng, H. F., et al. (2011). BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J.* 5, 741–749. doi: 10.1038/ismej.2010.160

Zhou, J., Bruns, M. A., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* 62, 316–322.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

MINI REVIEW
published: 09 July 2015
doi: 10.3389/fmicb.2015.00678

Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial)

Adina Howe[1]* *and Patrick S. G. Chain*[2]

[1] GERMS Laboratory, Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA, USA, [2] Bioinformatics and Analytics Team, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Marc Strous, University of Calgary, Canada
Mick Watson, The Roslin Institute, UK

***Correspondence:**
Adina Howe, GERMS Laboratory, Department of Agricultural and Biosystems Engineering, Iowa State University, 3346 Elings Hall, Ames, IA 50011, USA
adina@iastate.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *05 May 2015*
Accepted: *22 June 2015*
Published: *09 July 2015*

Citation:
Howe A and Chain PSG (2015) Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). Front. Microbiol. 6:678. doi: 10.3389/fmicb.2015.00678

Metagenomic investigations hold great promise for informing the genetics, physiology, and ecology of environmental microorganisms. Current challenges for metagenomic analysis are related to our ability to connect the dots between sequencing reads, their population of origin, and their encoding functions. Assembly-based methods reduce dataset size by extending overlapping reads into larger contiguous sequences (contigs), providing contextual information for genetic sequences that does not rely on existing references. These methods, however, tend to be computationally intensive and are again challenged by sequencing errors as well as by genomic repeats While numerous tools have been developed based on these methodological concepts, they present confounding choices and training requirements to metagenomic investigators. To help with accessibility to assembly tools, this review also includes an IPython Notebook metagenomic assembly tutorial. This tutorial has instructions for execution any operating system using Amazon Elastic Cloud Compute and guides users through downloading, assembly, and mapping reads to contigs of a mock microbiome metagenome. Despite its challenges, metagenomic analysis has already revealed novel insights into many environments on Earth. As software, training, and data continue to emerge, metagenomic data access and its discoveries will to grow.

Keywords: metagenomes, assembly, review, challenges, tutorial

Overview

The application of high throughput sequencing technologies for environmental microbiology is arguably as transformative as the invention of the microscope. When we began to *see* previously invisible microorganisms, we discovered the vast number of microbes in our environments. These observations significantly expanded the scope of microbiology as we began to have a better sense of the diversity of organisms outside of what we could grow in the laboratory. Presently, with sequencing technologies, we now *read* the genetic code of microorganisms, assembling microbial genomes without the need to even culture them, and in some cases providing clues as to how to culture them. This accessibility to genes has allowed us to investigate microorganisms and their predicted functional profiles in increasingly complex natural environments through approaches

like metagenomics. In this review, we discuss how sequencing technologies can help us understand microbial communities and the challenges and opportunities involved in analyzing these very large datasets with metagenome assembly.

Metagenomic Assembly

In analyzing microbes using genomics, one of the earliest forms of analysis involved genome assembly. Note that in this review, we use the phrase assembly to refer to *de novo* assembly, or the assembly of contigs without the use of previous references. From even the early days in sequencing, genome assembly has been a revered subspecialty in bioinformatics. Assembly began as an extension of local sequence alignments, where each sequencing read was compared with all other reads, followed by the subsequent assembly of the highest scoring pairs, essentially identifying overlapping sequences for extension into longer contiguous sequences, or contigs. These assemblers were developed for the then-standard Sanger sequencing technology. They were effective at retroactive correction of assembly errors, using the long, accurate Sanger read lengths for decision making with regards to variant calls and conflicts in read mate pairs that indicate possible chimeras or rearrangements (Dear and Staden, 1991; Lawrence et al., 1994; Myers, 1995; Bonfield and Whitwham, 2010).

The advent of next generation sequencing (NGS) technologies changed the type of sequencing data available to microbiologists and also expanded the types of questions that could be asked of sequencing. NGS reads are much cheaper than Sanger reads but are also much shorter in length (e.g., ∼100–250 bp). Assembly of NGS short read data is hampered both by the length of reads and the large number of reads that typically exceed by one or more orders of magnitude the number of reads that would be needed for the same project using Sanger sequencing. While fold coverage necessary for adequate assembly with Sanger data approached 10-fold coverage, with short-read technologies such as Illumina, the fold coverage needed for adequate assembly is generally 100-fold or greater (Sims et al., 2014). The number of read-to-read comparisons and the storing of this information quickly exceed the memory available on even very large memory machines. A series of more memory efficient methods based on *de Bruijn* graphs have been developed to tackle this assembly problem (Pevzner et al., 2001) and reviewed in (Pop, 2009; Miller et al., 2010).

Due to the increased cost-effectiveness, and to a lesser extent, the throughput of the newer, next-generation sequencing platforms, the number of shotgun metagenome projects in the microbiology field has surged. Today, thousands of projects are underway, exploring systems of low complexity, such as acid mine drainage (Tyson et al., 2004), ocean oil spills (Mason et al., 2012), and deep sea hydrothermal vents (Xie et al., 2011), to those of extreme complexity. In complex environments, metagenomes require deep sequencing for assembly; current sequencing efforts (less than 1 Tbp per sample) in soils and sediments resulting in less than half of the reads incorporated into assembled contigs (Luo et al., 2012; Howe et al., 2014) suggest that these environments contain very high diversity. While the specific goals of all these projects vary, most initial questions revolve around the characterization of functional and taxonomic composition. While there have been many recent advances in examining these questions using read-based approaches (Segata et al., 2012; Wood and Salzberg, 2014; Freitas et al., 2015), these are limited to supervised approaches, meaning that a limiting factor is the presence of an available database with appropriate reference genomes. For many of the ecosystems explored using metagenomics, there is a gross lack of high quality reference genomes. Without sufficiently similar references for dominant organisms in a sample, metagenome assembly is an approach that can provide greater insight into the community by delivering longer, contiguous sequences that can subsequently be investigated using more traditional approaches for classification of taxonomy and function. These contigs can sometimes approach the size of an entire genome, possibly linking functional genes to phylogenetic markers and allowing a more comprehensive reconstruction of the metabolic potential of a particular genome (Albertsen et al., 2013; Sharon et al., 2013; Wrighton et al., 2014).

Current Challenges with Metagenome Assemblies

While the throughput of sequencers seems astronomical compared with a decade ago, it can still be difficult to have sufficient sequence representation from the large number of different organisms that can be found in many ecosystems. Due to variable relative abundance of different community members within a population, some genomes may be covered many thousands of times while others are only covered by a handful of sequencing reads or none at all. Some communities may even be sufficiently diverse that no member is represented very highly. Because any assembly of sequence data requires overlaps among reads, assembly of the less dominant members of a community may require additional sequencing.

These considerations, along with the cost, often dictate the level of sequencing effort dedicated to a project. The most prominent sequencing platforms currently used for metagenomes include ones that produces millions to billions of short (<300 bp) reads (e.g., Illumina sequencing platforms). Estimations of community diversity often precede metagenomic sequencing efforts. While these efforts (often using rRNA gene amplicon analysis) can be revealing for community studies by themselves, they can be inaccurate when it comes to strain-level diversification or population heterogeneity. For example, while some dominant rRNA members may be clonal in origin, others rRNA sequences may represent a broader diversity of genotypes.

Another challenge for metagenomic assembly is that despite the improvements in assembly algorithms and the advancement of computer hardware technology, assembly of such abundant, complex data can often overwhelm any given computer's memory constraints. This issue is contributed to by the natural diversity of the community and the variants found within the population and is further exacerbated by sequencing errors that are present (even at very low levels) within the sequencing data.

Strategies for Metagenome Assembly

There are an increasing number of assembly programs focused on the issue of metagenome assembly (Peng et al., 2011; Namiki et al., 2012; Li et al., 2015), most of which are based on *de Bruijn* graph assembly, that involves deconstructing the short reads into ever shorter *k*-mers of length k, finding overlaps of k-1, and traversing through the graph of *k*-mers/overlaps. There are a number of areas where metagenome assembly efforts have focused on improving. Some methods try to address the memory constraints in generating large assembly graphs, generally using a divide and conquer strategy. Other assemblers try to improve the ability to handle minor variants (or sequence errors) within otherwise identical *k*-mers by weighting *k*-mers by frequency or by collapsing paths depending on connectivity (e.g., bifurcating and rejoining paths). Other methods try to tackle some of the many complications that occur with the presence of genomes with high variations in abundance, for example by iterating over a series of different *k*-mer sizes. The length of the *k*-mer defines two things: 1) the overlap size needed among *k*-mers to allow assembly of two *k*-mers, and 2) the size of the repeat that can be resolved by the *k*-mer. Given sufficient coverage, longer *k*-mers will provide a simpler graph and a more robust assembly since repeats smaller than size k will be resolved within the graph. However, for organisms of lower abundance (i.e., genomes of lower coverage), the chance of sequencing overlapping regions (of size *k*) of the genome is also decreased (with longer *k* length), dictating the lower bound of organism abundance that can be assembled.

Because *de Bruijn* graph assembly is based on the smaller *k*-mer lengths and not on full read lengths, the smallest contigs are generally of size k+1, and it is possible to generate contigs from the graph that are not reflected by any read. If this was not already complicated, because of the highly conserved nature of functional features (homologous sequences) within disparate genomes, e.g., multiple copies of rRNA gene sequences, assemblers can generate chimeric contigs at any *k*-mer that is shared among two genomes (or within a genome). After assembly, contigs with minimal or no read coverage can be removed, and some of the chimeras can be resolved using paired-end reads if available. While these and other metagenome assembly issues can be somewhat addressed post-assembly, specialized tools are not yet available that address all of them. An alternative strategy for assembly of metagenomes includes using different algorithms that use reference genomes or genes for more specialized, targeted assembly (Boisvert et al., 2012).

Accessibility to Metagenome Assembly

The challenges that face most scientists when confronted with metagenome assembly appear daunting: a wide array of assembly tools, each with their own strengths and weaknesses, and none ideal for any given metagenomic community of varying diversity, nor tailored to function within any given computational environment. In addition, this can become substantially more complex if using multiple technologies with differing error models, read lengths, and amounts of data since most bioinformatics tools are truly developed for highly specific data types.

Further exacerbating the situation is that most of these tools (especially newer ones) require knowledge of executing a command in a Unix environment. This obstacle, mainly the lack of individuals cross-trained in microbiology and practical bioinformatics is arguably one of the largest facing the field. Knowledge of the specific questions being asked of a sequencing dataset, the opportunities and limitations of an experiment, and the skills to effectively analyze these datasets can ensure that the data and algorithms used are appropriate for the question. While the number of microbiologists with bioinformatics skills is increasing, it is not yet commonplace, and sequencing is increasingly prevalent in most areas of biology and has already been declared democratized by a number of groups (Kumar et al., 2013; Koren et al., 2014; Meijueiro et al., 2014). As evident from the challenges above for metagenome assembly, even within the area of bioinformatics, there can be many subspecialties, each requiring a level of sophistication often beyond the average microbiologist. In an effort to make available some of the skills needed for metagenome analysis, including metagenome assembly, this review includes a tutorial on some of the steps for analyzing a simulated mock metagenome from the Human Microbiome Project.[1] Given the challenges of accessibility to computational resources, this tutorial has been designed for implementation on rentable cloud computing.[2] We also note that there are a number of challenges in metagenomics, and in this review, we focus on challenges facing individuals whose goal is to analyze a community using metagenome assembly. However, it is also important to consider that many other questions can be asked using a metagenome without specifically requiring an assembly (reviewed in, Sharpton, 2014), such as aligning reads to known references (reviewed in (Trapnell and Salzberg, 2009; Li and Homer, 2010; Fonseca et al., 2012) and read-based functional annotations (reviewed in, De Filippo et al., 2012; Prakash and Taylor, 2012).

References

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable *de novo* metagenome assembly and profiling. *Genome Biol.* 13, R122. doi: 10.1186/gb-2012-13-12-r122

Bonfield, J. K., and Whitwham, A. (2010). Gap5–editing the billion fragment sequence assembly. *Bioinformatics* 26, 1699–1703. doi: 10.1093/bioinformatics/btq268

Dear, S., and Staden, R. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res.* 19, 3907–3911. doi: 10.1093/nar/19.14.3907

De Filippo, C., Ramazzotti, M., Fontana, P., and Cavalieri, D. (2012). Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief. Bioinform.* 13, 696–710. doi: 10.1093/bib/bbs070

[1]http://hmpdacc.org/

[2]http://nbviewer.ipython.org/github/germs-lab/frontiers-review-2015/blob/master/frontiers-nb-2015.ipynb

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169–3177. doi: 10.1093/bioinformatics/bts605

Freitas, T. A. K., Li, P. E., Scholz, M. B., and Chain, P. S. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* 43, e69. doi: 10.1093/nar/gkv180

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4904–4909. doi: 10.1073/pnas.1402564111

Koren, S., Treangen, T. J., Hill, C. M., Pop, M., and Phillippy, A. M. (2014). Automated ensemble assembly and validation of microbial genomes. *BMC Bioinform*. 15:126. doi: 10.1101/002469

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet*. 4:237. doi: 10.3389/fgene.2013.00237

Lawrence, C., Honda, S., Parrott, N. W., Flood, T. C., Gu, L., Zhang, L., et al. (1994). The genome reconstruction manager: a software environment for supporting high-throughput DNA sequencing. *Genomics*. 23, 192–201. doi: 10.1006/geno.1994.1477

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* 11, 473–483. doi: 10.1093/bib/bbq015

Luo, C., Tsementzi, D., Kyrpides, N. C., and Konstantinidis, K. T. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* 6, 898–901. doi: 10.1038/ismej.2011.147

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S., Dubinsky, E. A., Fortney, J. L., et al. (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.* 6, 1715–1727. doi: 10.1038/ismej.2012.59

Meijueiro, M. L., Santoyo, F., Ramírez, L., and Pisabarro, A. G. (2014). Transcriptome characteristics of filamentous fungi deduced using high-throughput analytical technologies. *Brief. Funct. Genomics* 13, 440–450. doi: 10.1093/bfgp/elu033

Miller, J. R., Koren, S., and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Presentation. *Genomics* 95, 315–327. doi: 10.1016/j.ygeno.2010.03.001

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.* 2, 275–290. doi: 10.1089/cmb.1995.2.275

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res*. 40, e155. doi: 10.1093/nar/gks678

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2011). Meta-IDBA: a *de Novo* assembler for metagenomic data. *Bioinformatics* 27, i94–i101. doi: 10.1093/bioinformatics/btr216

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A*. 98, 9748–9753. doi: 10.1073/pnas.171285098

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform*. 10, 354–366. doi: 10.1093/bib/bbp026

Prakash, T., and Taylor, T. D. (2012). Functional assignment of metagenomic data: challenges and applications. *Brief. Bioinform*. 13, 711–727. doi: 10.1093/bib/bbs033

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet*. 15, 121–132. doi: 10.1038/nrg3642

Trapnell, C., and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nat. Biotechnol.* 27, 455–457. doi: 10.1038/nbt0509-455

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43. doi: 10.1038/nature02340

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46. doi: 10.1186/gb-2014-15-3-r46

Wrighton, K. C., Castelle, C. J., Wilkins, M. J., Hug, L., Sharon, I., and Thomas, B. C. (2014). Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J.* 8, 1452–1463. doi: 10.1038/ismej.2013.249

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., et al. (2011). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J.* 5, 414–426. doi: 10.1038/ismej.2010.144

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 11 August 2015
doi: 10.3389/fmicb.2015.00805

The green impact: bacterioplankton response toward a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches

Bernd Wemheuer[1†], Franziska Wemheuer[2†], Jacqueline Hollensteiner[1], Frauke-Dorothee Meyer[1], Sonja Voget[1] and Rolf Daniel[1]*

[1] Genomic and Applied Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University Göttingen, Göttingen, Germany, [2] Department for Crop Sciences, Georg-August-University Göttingen, Göttingen, Germany

OPEN ACCESS

Edited by:
Eamonn P. Culligan,
University College Cork, Ireland

Reviewed by:
Byron Crump,
Oregon State University, USA
Marc Strous,
University of Calgary, Canada

***Correspondence:**
Rolf Daniel,
Department of Genomic and Applied Microbiology and Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University Göttingen, Grisebachstr. 8, D-37077 Göttingen, Germany
rdaniel@gwdg.de

[†] These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: 11 May 2015
Accepted: 22 July 2015
Published: 11 August 2015

Citation:
Wemheuer B, Wemheuer F, Hollensteiner J, Meyer F-D, Voget S and Daniel R (2015) The green impact: bacterioplankton response toward a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches. Front. Microbiol. 6:805. doi: 10.3389/fmicb.2015.00805

Phytoplankton blooms exhibit a severe impact on bacterioplankton communities as they change nutrient availabilities and other environmental factors. In the current study, the response of a bacterioplankton community to a *Phaeocystis globosa* spring bloom was investigated in the southern North Sea. For this purpose, water samples were taken inside and reference samples outside of an algal spring bloom. Structural changes of the bacterioplankton community were assessed by amplicon-based analysis of 16S rRNA genes and transcripts generated from environmental DNA and RNA, respectively. Several marine groups responded to bloom presence. The abundance of the *Roseobacter* RCA cluster and the SAR92 clade significantly increased in bloom presence in the total and active fraction of the bacterial community. Functional changes were investigated by direct sequencing of environmental DNA and mRNA. The corresponding datasets comprised more than 500 million sequences across all samples. Metatranscriptomic data sets were mapped on representative genomes of abundant marine groups present in the samples and on assembled metagenomic and metatranscriptomic datasets. Differences in gene expression profiles between non-bloom and bloom samples were recorded. The genome-wide gene expression level of *Planktomarina temperata*, an abundant member of the *Roseobacter* RCA cluster, was higher inside the bloom. Genes that were differently expressed included transposases, which showed increased expression levels inside the bloom. This might contribute to the adaptation of this organism toward environmental stresses through genome reorganization. In addition, several genes affiliated to the SAR92 clade were significantly upregulated inside the bloom including genes encoding for proteins involved in isoleucine and leucine incorporation. Obtained results provide novel insights into compositional and functional variations of marine bacterioplankton communities as response to a phytoplankton bloom.

Keywords: bacterioplankton, metagenomics, metatranscriptomics, algal bloom, functional changes, *Planktomarina temperata*, SAR92

Introduction

Bacteria are major drivers in cycling of nitrogen, carbon, and other elements in marine ecosystems (Azam et al., 1983; Arrigo, 2005; DeLong and Karl, 2005). More than 50% of organic matter produced by phytoplankton is remineralized by marine bacteria (Cole et al., 1988; Karner and Herndl, 1992; Ducklow et al., 1993). Therefore, bacteria play an important role during and after bloom events as large amounts of organic matter are generated by primary production (Azam, 1998).

Recent studies investigating bacterioplankton communities during phytoplankton blooms revealed that community structures and diversity were highly affected (Teeling et al., 2012; Liu et al., 2013; Wemheuer et al., 2014). Observed patterns were correlated to changes of nutrient concentrations and other environmental factors such as water depth or algal species (Fandino et al., 2001; Pinhassi et al., 2004; Grossart et al., 2005; Teeling et al., 2012; Liu et al., 2013; Wemheuer et al., 2014; Gomes et al., 2015). Consequently, understanding the dynamics and interactions between bacterial communities and phytoplankton blooms is crucial to validate the ecological impact of bloom events.

One region with annually recurring spring phytoplankton blooms is the North Sea, a typical coastal shelf sea of the temperate zone. Shelf seas are highly productive due to the continuous nutrient supply by rivers. During the last 40 years, the North Sea and in particular its southern region, the German Bight, underwent high nutrient loading and warming (McQuatters-Gollop et al., 2007; Wiltshire et al., 2008, 2010). Recent studies aimed at understanding bacterial responses to phytoplankton blooms in the North Sea (Alderkamp et al., 2006; Teeling et al., 2012; Wemheuer et al., 2014). A dynamic succession of distinct bacterial clades before, during, and after bloom events in the North Sea was observed in several investigations (Alderkamp et al., 2006; Alonso and Pernthaler, 2006a,b; Teeling et al., 2012). The results indicate that specialized populations occupy ecological niches provided by phytoplankton-derived substrates (Teeling et al., 2012). Klindworth et al. (2014) investigated the diversity and activity of marine bacterioplankton during the same bloom event applying metatranscriptomic techniques. They showed that members of the *Rhodobacteraceae* and SAR92 clade exhibited high metabolic activity levels. However, recent research focused mainly on changes of community structure as response to phytoplankton blooms, but functional changes and their resulting ecological impacts have been rarely studied. In addition, larger comparative metagenomic and metatranscriptomic studies investigating structural and functional changes of the bacterioplankton during the bloom event are lacking.

In a previous study, we investigated structural differences of the active bacterioplankton community as response toward a *Phaeocystis globosa* bloom in the southern North Sea in spring 2010 mainly by 16S rRNA pyrotag sequencing (Wemheuer et al., 2014). This microalgae has a cosmopolitan distribution (Schoemann et al., 2005) and is considered to be responsible for harmful algal blooms (Veldhuis and Wassmann, 2005). These blooms have been observed in many marine environments, including the coast of the eastern English Channel, the southern North Sea and the south coast of China (Schoemann et al., 2005). We found that the phytoplankton spring bloom impacted bacterioplankton community structures and the abundance of certain bacterial groups significantly. For example, the *Roseobacter* RCA cluster and the SAR92 clade were significantly more abundant in the bloom at active community level. For the current study, more than 500 million sequences derived from direct sequencing of environmental DNA and rRNA depleted RNA were added to obtain functional insights into the bloom event. Metatranscriptomic data was mapped on the assembled metagenomic and metatranscriptomic data sets and on the genomes of abundant marine bacteria, e.g., *P. temperata* RCA23, a member of the *Roseobacter* RCA cluster. In addition, 16S rRNA genes and transcripts were studied by pyrotag sequencing to obtain insights into structural dynamics of the total and active bacterioplankton community, respectively. The comprehensive experimental design and method combination of this study sheds new light on ecological roles and functions of single members of the bacterioplankton community and the entire community.

Materials and Methods

Sampling and Sample Preparation

Ten water samples for bacterioplankton analyses were collected in the southern North Sea at nine stations in and outside of a *P. globosa* bloom in May 2010 (**Figure 1**; **Table 1**). Six samples were taken in the presence of a phytoplankton bloom (3a, 3b, and 4) and three in bloom absence (5–15). One sample was taken

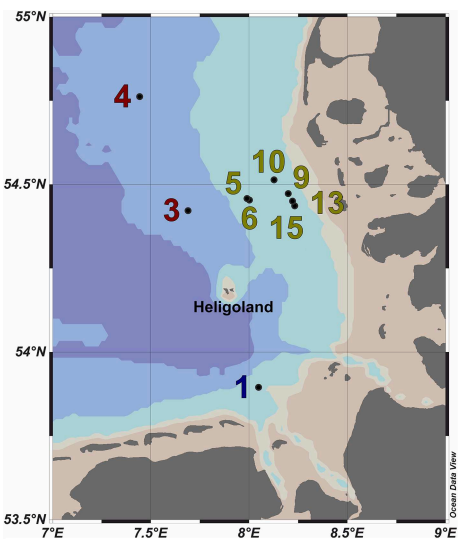

FIGURE 1 | Map of the German Bight showing the locations of the nine sampling stations visited in May 2010. Stations inside the examined phytoplankton bloom are depicted in green; those outside the bloom in red. Station 1 is depicted in blue as it was located in a bloom outside of the area of the examined bloom. Shading of the water masses refers to bottom depth. The map was generated using the Ocean Data View software package [version 4.7.2; Schlitzer, 2015 (http://odv.awi.de/)].

TABLE 1 | Sampling site characteristics.

Sample	Ship station	Origin	Date (mm/dd/yyyy)	Latitude (°N)	Longitude (°E)	Depth (m)	Bottom depth (m)
1	655	River outfall	05/25/2010	53.8955	8.0496	2	15.5
3a	657a	No bloom	05/26/2010	54.4223	7.6833	2	22.1
3b	657b	No bloom	05/26/2010	54.4223	7.6833	12	22.1
4	658	No bloom	05/26/2010	54.7626	7.4463	2	20
5	659	Bloom	05/26/2010	54.4575	7.9893	9	12.5
6	660	Bloom	05/27/2010	54.4542	8.0018	2	12.5
9	664	Bloom	05/28/2010	54.4733	8.1972	2	12
10	665	Bloom	05/28/2010	54.5135	8.128	2	11
13	668	Bloom	05/29/2010	54.4365	8.2328	10	12
15	671	Bloom	05/30/2010	54.449	8.22	2	12

near to a river outfall (1). Stations inside the bloom were located by satellite images and are characterized by their increased chlorophyll content. Note that sample 9 was taken in the bloom area and is considered as a bloom sample despite its relative low chlorophyll content. Sampling and filtration were performed as described previously (Wemheuer et al., 2014). In brief, obtained water samples were initially filtered using a 10 μm nylon net filter and 2.7 μm glass fiber filter. Bacterioplankton was subsequently harvested from a prefiltered 1 l sample on a filter sandwich consisting of a glass fiber and 0.2 μm polycarbonate filter (47 mm diameter). Samples for community analysis were stored at −80°C until further analysis. Several environmental parameters such as chlorophyll *a* (Chl *a*), particulate organic nitrogen (PON), salinity, temperature, and nitrate content were determined as described previously (Wemheuer et al., 2014) (**Table 2**).

Extraction and Purification of Environmental DNA and RNA

Environmental DNA and RNA were co-extracted from the filter sandwich as described by Weinbauer et al. (2002). DNA and RNA were subsequently purified employing the peqGOLD gel extraction kit (Peqlab, Erlangen, Germany) and the RNeasy Mini Kit (Qiagen, Hilden, Germany), respectively, as recommended by the manufacturers. Residual DNA was removed from RNA samples and its absence was confirmed according to Wemheuer et al. (2012).

To assess bacterioplankton community structures, DNA-free RNA was directly converted to cDNA employing the SuperScript® III reverse transcriptase (Invitrogen™, Carlsbad, USA) using a primer specific for the conserved region downstream to variable region 6 of the 16S rRNA (1063r 5′-CTCACGRCACGAGCTGACG-3′). The reaction mixture (20 μl) contained 4 μl of five-fold reaction buffer, 500 μM of each of the four desoxynucleoside triphosphates, 5 mM DTT, 1 μM of the reverse primer, 1 U RiboLock™ RNase Inhibitor (Thermo Fisher Scientific, Schwerte, Germany), 200 U of the reverse transcriptase and approximately 100 ng DNA-free RNA. The reaction was incubated at 55°C for 1 h and subsequently inactivated by incubation at 70°C for 15 min. To remove the RNA in the RNA/DNA hybrids, 2.5 U RNase H (Thermo Fischer Scientific) were added and the reaction incubated at 37° for 15 min followed by inactivation at 65°C for 10 min. Obtained cDNA was subsequently subjected to 16S rRNA gene PCR (as described below). To assess community functions, environmental mRNA was enriched from total RNA using the RiboMinus™ transcriptome isolation kit for Bacteria (Invitrogen™, Carlsbad, USA) with one modification. The initial denaturation of RNA was performed at 70°C for 10 min. RNA was subsequently converted to cDNA employing the SuperScript™ double-stranded cDNA synthesis kit (Invitrogen™) with slight modifications according to Wemheuer et al. (2014). The Göttingen Genomics Laboratory determined the sequences of the extracted DNA and enriched mRNA-derived cDNA using a Roche 454 GS-FLX+ pyrosequencer with titanium chemistry (Roche, Mannheim, Germany) and an Illumina Genome Analyzer IIx (San Diego, USA), respectively (**Table 3**).

Processing and Analysis of Metagenomic and Metatranscriptomic Datasets

Generated metagenomic and metatranscriptomic datasets were initially processed according to Voget et al. (2014). Briefly, fastq files derived from Illumina sequencing were processed employing the Trimmomatic tool version 0.30 (Bolger et al., 2014). Sff files derived from pyrosequencing were converted to fastq files prior to quality filtering. Afterwards, all sequences were combined and assembled at different kmer values (29–109 in 10 bp steps) with Velvet and Metavelvet (Zerbino and Birney, 2008; Namiki et al., 2012). Subsequently, all obtained contigs were joined and resulting sequences were dereplicated employing Usearch version 7.0.190 (Edgar, 2010). Open reading frames (ORFs) were predicted for all remaining contigs using Prodigal version 2.6 (Hyatt et al., 2010). Short contigs (<150 bp) were removed prior to further analysis.

As the metagenomic and metatranscriptromic datasets are likely to contain algal-derived sequences, we subtracted bacterial genes by blast alignment (Camacho et al., 2009) against 15 reference genomes of abundant marine lineages (**Table 4**) obtained from the integrated microbial genomes (IMG) platform (Markowitz et al., 2012). Genomes of abundant phylogenetic groups as found in the 16S rRNA analysis were chosen for this additional filtering step. Only sequences with an *e*-value below 0.001 were used in the subsequent analysis. Remaining ORFs

TABLE 2 | Environmental parameters measured for the 10 water samples.

Sample	Temperature (°C)	Salinity (psu)	Fluorescence (FU)	Transmission (%)	Density (g/l)	Chlorophyll a (µg/L)	Phaeopigment (µg/L)	Suspended particulate matter (µg/L)	Particulate organic carbon (µg/L)	Particulate organic nitrogen (µg/L)	Nitrate (µM)	Nitrite (µM)	Mono-nitrogen oxides (µM)	Phosphate (µM)
RIVER OUTFALL														
1	11.09	30.24	1.21	57.2	1023.1	4.38	2.11	9.8	997.4	152.1	8.46	0.19	8.65	0.15
NON-BLOOM														
3a	9.43	31.42	0.2	87.98	1024.27	1.12	0.25	4.6	291.2	43	7.395	0.25	7.65	0.02
3b	8.18	32.01	0.77	84.83	1024.93	3.37	1.08	7.15	496.4	82.2	5.9	0.27	6.17	0.04
4	9.73	32.71	0.49	81.23	1025.2	2.55	0.37	6.15	290.4	46.9	6.17	0.24	6.41	0.03
BLOOM														
5	10.8	30.64	2.76	60.14	1023.5	11.45	7.03	3.27	1673	213.6	5.03	0.24	5.27	0.1
6	10.83	30.65	1.89	72.78	1023.4	7.34	2.77	11.3	728.2	106.2	9.11	0.42	9.53	0.08
9	10.9	30.76	1.14	87.28	1023.5	2.19	0.6	3.2	367.7	49	4.94	0.24	5.18	0.02
10	11.4	31.11	2.27	74.79	1023.7	6.93	2.1	7.5	737.5	95.3	3.685	0.29	3.98	0.07
13	11.83	31.18	2.8	67.83	1023.7	5.53	3.16	9.91	936.5	123.5	2.085	0.21	2.29	0.1
15	11.7	31.04	NA(*)	76.59	1023.6	5.33	1.58	6.25	624.4	83.9	2.97	0.3	3.27	0.08

(*)Not measured due to fluorometer malfunction.

Although not exhibiting high chlorophyll a values, sample 9 is considered as a bloom sample as it was taken in the bloom area.

were further classified employing UProC version 1.2 in protein mode (Meinicke, 2015).

Metatranscriptomic datasets were mapped on the 15 genomes and on the assembled contigs using Bowtie 2 version 2.2.4 (Langmead and Salzberg, 2012) with one mismatch in the seed and multiple hits reporting enabled for the metagenomic binning. Ribosomal RNA was removed from metatranscriptomic datasets prior to mapping employing SortMeRNA version 2.0 (Kopylova et al., 2012) (**Table 5**). The number of unique sequences per gene was calculated, and the overall mapping result was normalized by the total number of unique reads in the sample. The top 23,884 open reading frames based on number of assigned, normalized reads are provided in Supplementary Table S1.

Results from the genome mapping were additionally normalized by the unique RNA/DNA ratio calculated for each bacterial group in the respective sample (see Campbell and Kirchman, 2013). In detail, the ratio was calculated by dividing the relative abundance of a 16S rRNA transcript by the relative abundance of the corresponding 16S rRNA gene. On the one hand, assuming that the major fraction of the bacterioplankton community is present outside and inside the algal bloom, the abundance at DNA level of single species is mainly linked to cell abundance rather than other factors such as 16S rRNA gene copy number. On the other hand, RNA abundance is correlated with protein synthesis (Blazewicz et al., 2013) and may indirectly serve as approximation for gene expression levels. Therefore, a high RNA/DNA ratio reflects an increased gene expression per cell and vice versa.

Amplification and Sequencing of 16S rRNA

To assess bacterial community structures, the V3–V6 region of the bacterial 16S rRNA was amplified by PCR. The PCR reaction (50 µl) contained 10 µl of five-fold Phusion HF buffer, 200 µM of each of the four desoxynucleoside triphosphates, 1.5 mM $MgCl_2$, 4 µM of each primer, 2.5% DMSO, 2 U of Phusion high fidelity hot start DNA polymerase (Thermo Fisher Scientific), and approximately 50 ng of DNA or 25 ng of cDNA as template. The following thermal cycling scheme was used: initial denaturation at 98°C for 5 min, 25 cycles of denaturation at 98°C for 45 s, annealing at 60°C for 45 s, followed by extension at 72°C for 30 s. The final extension was carried out at 72°C for 5 min. Negative controls were performed by using the reaction mixture without template. The V3–V6 region was amplified with the following set of primers according to Muyzer et al. (1995) containing the Roche 454 pyrosequencing adaptors, keys and one unique MID per sample (underlined): 341f 5′-CCATCTCATCCCTGCGTG TCTCCGAC-TCAG-$(dN)_{10}$-CCTACGGRAGGCAGCAG-3′ and 1063r 5′-CCTATCCCCTGTGTGCCTTGGCAGTC-TCA G-CTCACGRCACGAGCTGACG-3′. Obtained PCR products were controlled for appropriate size and subsequently purified using the peqGOLD gel extraction kit (Peqlab) as recommended by the manufacturer. Three independent PCR reactions were performed per sample, purified by gel extraction, and pooled in equal amounts. Quantification of the PCR products was performed using the Quant-iT dsDNA HS assay kit and a Qubit fluorometer (Invitrogen™) as recommended by the

TABLE 3 | Sequence statistics of the quality trimmed metagenome and metatranscriptome data used in this study.

Sample	Technology	Single run/Paired End/Unpaired (SR/PE/UP)	Type	Sequence number	Average read length	Total basepairs
1*	454 FLX+	SR	gDNA	338,735	252.66	85,583,953
1*	454 FLX+	SR	mRNA	421,864	275.62	116,273,765
1	Illumina GIIA	PE	gDNA	21,186,566	104.44	2,212,675,749
1	Illumina GIIA	UP	gDNA	362,317	83.16	30,128,732
1	Illumina GIIA	SR	mRNA	20,782,683	73.56	1,528,826,535
3a	Illumina GIIA	PE	gDNA	12,578,554	104.29	1,311,791,324
3a	Illumina GIIA	SR	gDNA	10,843,677	104.93	1,137,876,968
3a	Illumina GIIA	UP	gDNA	33,298,482	103.62	3,450,551,533
3a	Illumina GIIA	SR	mRNA	24,527,136	72.90	1,787,928,626
3b	Illumina GIIA	PE	gDNA	19,210,244	105.28	2,022,514,254
3b	Illumina GIIA	SR	gDNA	8,438,275	104.14	87,878,7402
3b	Illumina GIIA	UP	gDNA	9,783,518	101.24	99,052,5636
3b	Illumina GIIA	SR	mRNA	27,810,866	73.14	2,034,036,519
4*	454 FLX+	SR	mRNA	186,132	255.15	47,492,480
4	Illumina GIIA	PE	gDNA	29,884,508	100.60	3,006,238,128
4	Illumina GIIA	UP	gDNA	642,734	75.36	48,433,894
4	Illumina GIIA	SR	mRNA	14,014,013	73.37	1,028,189,524
5*	454 FLX+	SR	gDNA	391,106	252.93	98,922,278
5*	454 FLX+	SR	mRNA	490,182	274.96	134,779,442
5	Illumina GIIA	PE	gDNA	25,109,444	104.71	2,629,205,182
5	Illumina GIIA	UP	gDNA	17,758,425	102.47	1,819,684,124
5	Illumina GIIA	PE	mRNA	35,492	69.05	2,450,563
5	Illumina GIIA	SR	mRNA	11,759,937	72.22	849,301,842
5	Illumina GIIA	UP	mRNA	8,714,379	103.34	900,565,197
6	Illumina GIIA	PE	gDNA	21,372,550	105.26	2,249,604,814
6	Illumina GIIA	UP	gDNA	32,730,493	103.10	3,374,578,797
6	Illumina GIIA	SR	mRNA	14,893,714	72.25	1,076,040,735
9	Illumina GIIA	PE	gDNA	24,735,216	103.57	2,561,948,071
9	Illumina GIIA	UP	gDNA	539,723	82.72	44,644,280
9	Illumina GIIA	PE	mRNA	43,792	70.96	3,107,595
9	Illumina GIIA	SR	mRNA	15,900,375	72.46	1,152,110,018
9	Illumina GIIA	UP	mRNA	9,779,257	104.90	1,025,839,134
10	Illumina GIIA	PE	gDNA	20,787,102	101.19	2,103,509,139
10	Illumina GIIA	UP	gDNA	662,370	81.29	53,843,938
10	Illumina GIIA	PE	mRNA	87,236	71.04	6,197,063
10	Illumina GIIA	SR	mRNA	9,005,445	72.68	654,542,993
10	Illumina GIIA	UP	mRNA	13,299,109	104.04	1,383,633,962
13*	454 FLX+	SR	mRNA	10,273	254.59	2,615,404
13	Illumina GIIA	PE	gDNA	28,055,686	102.21	2,867,561,091
13	Illumina GIIA	UP	gDNA	858,735	81.56	70,037,612
13	Illumina GIIA	SR	mRNA	26,455,179	72.54	1,919,045,434
15	Illumina GIIA	PE	gDNA	22,984,400	101.81	2,340,126,872
15	Illumina GIIA	UP	gDNA	521,111	81.11	42,269,455
15	Illumina GIIA	SR	mRNA	22,592,228	71.65	1,618,686,787
17*	454 FLX+	SR	mRNA	109,911	235.46	25,879,456
Total				563,993,174	93.49153275	52,728,586,300

Only Illumina-derived data not generated in a paired-end run was used in the metatranscriptomic mapping approach. Seqeuncing was performed using a Roche 454™ GS-FLX+ pyrosequencer with titanium chemistry and an Illumina Genome Analyzer IIx, respectively.

**Published under accession number SRA061816.*

TABLE 4 | Genomes retrieved from the Integrated Microbial Genomes (IMG) database.

IMG name	IMG Taxon ID	Genome Size (bp)	Gene Count	Phylum/Proteobacterial class	Marine Group/ Genus
Polaribacter sp. Hel_I_88	2,558,860,973	3,996,527	3552	*Bacteroidetes*	*Polaribacter*
Marivirga tractuosa H-43, DSM 4126	649,633,065	4,516,490	3857	*Bacteroidetes*	*Marivirga*
Gammaproteobacteria bacterium MOLA455	2,590,828,686	2,605,026	2374	*Gammaproteobacteria*	SAR92 clade
Planktomarina temperata RCA23, DSM 22400 (RCA23)	2,548,877,138	32,88,122	3101	*Alphaproteobacteria*	*Roseobacter* RCA
Betaproteobacteria bacterium MOLA814	2,590,828,684	2,859,706	2733	*Betaproteobacteria*	BAL58 marine group
Polaribacter sp. MED152 (re-annotation)	2,606,217,529	2,961,474	2695	*Bacteroidetes*	*Polaribacter*
Methylophilales sp. HTCC2181	639,857,020	1,304,428	1377	*Betaproteobacteria*	OM43 clade
SAR86 cluster bacterium SAR86B	2,597,489,920	1,679,540	1890	*Gammaproteobacteria*	SAR86 clade
Rhodobacterales sp. HTCC2255 (original sequence, contaminants removed)	2,517,572,075	2,224,475	2209	*Alphaproteobacteria*	*Roseobacter* NAC11-7
Formosa sp. AK20	2,531,839,038	3,055,484	2841	*Bacteroidetes*	*Formosa*
Candidatus *Pelagibacter ubique* SAR11 HTCC1062 (re-annotation)	2,606,217,343	1,308,759	1393	*Alphaproteobacteria*	SAR11 clade
Marine gamma proteobacterium sp. HTCC2207 (re-annotation)	2,606,217,324	2,620,870	2388	*Gammaproteobacteria*	SAR92 clade
SAR86 cluster bacterium SAR86A	2,597,489,919	1,245,342	1340	*Gammaproteobacteria*	SAR86 clade
Candidatus *Pelagibacter ubique* SAR11 HTCC1002 (re-annotation)	2,606,217,624	1,327,604	1415	*Alphaproteobacteria*	SAR11 clade
Formosa agariphila KMM 3901	2,585,427,664	4,228,350	3630	*Bacteroidetes*	*Formosa*

TABLE 5 | Sequence statistics of the quality trimmed and rRNA depleted metatranscriptomic data sets used for mapping.

Sample	Number of sequences	Average read length	Total bps	Mapped reads	Mapping rate (%)
1	4,398,462	73.66	3,23,999,530	3,905,801	88.80
3a	2,256,897	73.52	165,918,281	2,004,411	88.81
3b	3,058,087	73.69	225,347,730	2,538,870	83.02
4	1,536,922	72.64	111,646,018	1,238,340	80.57
5	1,675,015	87.74	146,961,679	1,475,942	88.12
6	1,281,015	72.90	93,392,214	1,102,985	86.10
9	8,352,509	84.56	706,301,999	7,416,227	88.79
10	4,976,491	90.96	452,673,412	4,236,320	85.13
13	15,976,940	72.28	1,154,891,997	13,685,288	85.66
15	2,028,509	72.70	147,476,064	1,662,304	81.95
Total	45,540,847	77.48	3,528,608,924	39,266,488	86.22

Depletion was performed with SortMeRNA (Kopylova et al., 2012).

manufacturer. The Göttingen Genomics Laboratory determined the sequences using a Roche GS-FLX++ 454 pyrosequencer with Titanium chemistry (Roche, Mannheim, Germany).

Processing and Analysis of 16S rRNA Datasets

Generated 16S rRNA gene and rRNA datasets were processed as described by Wietz et al. (2015). In brief, sequences were preprocessed with QIIME and subsequently denoised employing Acacia (Bragg et al., 2012). Remaining primer sequences were truncated employing cutadapt (Martin, 2011). To remove chimeras, sequences were first dereplicated and putative chimeras were removed using UCHIME in *de novo* mode and subsequently in reference mode using the most recent SILVA database (SSURef 119 NR) as reference dataset (Edgar et al., 2011; Quast et al., 2013). Processed sequences of all samples were joined and clustered in operational taxonomic units (OTUs) at 3 and 20% genetic dissimilarity according to Wemheuer et al. (2013) employing the UCLUST algorithm with optimal flag (Edgar, 2010). To determine taxonomy, a consensus sequence for each OTU at 97% genetic similarity was classified by BLAST alignment against the Silva SSURef 119 NR database (Camacho et al., 2009). All non-bacterial OTUs were removed. Sequences statistics are shown in **Table 6**. The curated OTU table is provided as Supplemental Table S2. The final Alpha diversity indices were calculated with QIIME as described by Wemheuer et al. (2013) (see **Table 7**).

TABLE 6 | Statistics of the 16S rRNA analysis.

Sample	Before preprocessing		After preprocessing		After denoising		After removal of non-bacterial or chimeric sequences	
	No. of sequences	Average length	No. of sequences	Average length	No. of sequences	Average length	No. of sequences	Average length
DNA								
1	10,692	705.0	10,486	676.1	10,172	673.9	6380	668.5
3a	12,328	704.6	12,155	674.4	11,814	673.7	6611	667.8
3b	13,234	705.6	13,086	674.9	12,771	674.1	7756	669.7
4	8589	710.5	8467	679.3	8283	678.9	4923	676.0
5	11,954	701.4	11,749	671.8	11,313	670.2	6435	661.9
6	8580	700.0	8466	671.3	8153	668.6	4557	658.1
9	8198	703.5	8088	673.6	7947	673.1	4265	666.1
10	5801	691.7	5726	664.0	5523	662.1	2946	645.3
13	9009	709.6	8904	678.9	8751	678.3	4463	672.4
15	3339	703.5	3306	673.4	3234	673.2	1789	668.4
Total	91,724	704.0	90,433	674.2	87,961	673.0	50,125	666.3
RNA								
1	7296	708.3	7178	678.1	6998	676.9	3099	671.5
3a	12,612	710.5	12,457	680.2	12,078	678.2	4806	670.6
3b	6601	695.8	6510	667.6	6240	664.8	2720	649.5
4	12,901	696.0	1268	668.9	1195	668.7	491	657.4
5	11,944	712.5	11,584	682.1	11,297	680.1	4588	673.2
6	14,831	703.7	14,642	674.7	14,443	674.3	5520	663.3
9	10,499	702.4	4945	673.1	4803	671.7	1836	657.9
10	9480	712.8	9336	682.3	9115	680.7	3489	671.8
13	4669	710.8	4515	680.3	4412	678.9	1795	670.5
15	9251	703.2	9044	674.1	8679	671.4	3638	659.8
Total	100,084	705.4	81,479	677.2	79,260	675.6	31,982	666.0

Statistical Analysis

All statistical analyses were conducted employing R [version 3.1.2; R Core Team, 2014 (http://www.R-project.org/]. Possible correlations between phytoplankton bloom presence and richness (number of OTUs) as well Shannon indices, abundance, and gene expression were determined employing the non-parametric Wilcox rank-sum test (Gifford et al., 2013). Correlations were considered as significant with $P \leq 0.05$. Sample 1 was excluded from the statistical analysis because it was taken in another bloom event.

Sequence Data Deposition

Sequence data were deposited in the sequence read archive of the National Center for Biotechnology Information under accession numbers SRA061816 and SRA060677, respectively (for details see **Table 3**).

Results and Discussion

Characteristics of the Samples

In the current survey, we examined structural and functional responses of the bacterioplankton community toward a phytoplankton bloom. Samples for community analysis were taken randomly at different locations and different depths within a *P. globosa* bloom in the German Bight (**Figure 1**, **Table 1**). Six samples were taken in presence of the phytoplankton bloom (samples 5, 6, 9, 10, 13, and 15) and three in bloom absence (samples 3a, 3b, and 4). One sample was taken near the Weser river outfall (sample 1). Salinity ranged from 30.7 to 32.7 psu. Fluorescence was approximately 0.45 and 2.2 mg/m^3 outside and inside the algal bloom, respectively. Temperatures ranged from 8.2 to 11.8°C. All environmental parameters are listed in **Table 2**. Based on our previous analysis, most measured parameters were significantly linked to algal bloom presence (see Wemheuer et al., 2014). Only the suspended particulate matter content (SPM) and the nitrite concentration exhibited no direct correlation to bloom presence.

Bloom Presence Affects Bacterial Community Structures

Total and active bacterioplankton community structures were assessed by pyrosequencing-based analysis of the V3–V6 region of the 16S rRNA amplified from environmental DNA and RNA, respectively. A total of 50,125 and 31,982 high-quality bacterial 16S rRNA sequences were obtained across all 10 samples at DNA and RNA level, respectively (**Table 6**). Calculated rarefaction

TABLE 7 | Alpha diversity indices at 97 and 80% genetic similarity derived from the 16S rRNA analysis.

	Richness		Maximal number of OTUs		Coverage (%)		Chao1		Shannon (H')	
	97%	80%	97%	80%	97%	80%	97%	80%	97%	80%
DNA										
1	143.3	18.5	223.6	20.2	64.1	91.4	333.5	22.9	4.15	2.25
3a	160.2	19.7	252.7	21.0	63.4	93.8	325.1	23.0	4.19	2.17
3b	147.2	18.3	229.9	19.9	64.0	91.8	292.2	22.5	3.80	2.04
4	135	20.3	186.6	21.3	72.4	95.3	326.1	26.2	4.40	2.37
5	185.4	23.1	318.6	25.3	58.2	91.5	443.6	25.3	4.47	2.35
6	177.4	23.5	288.3	24.9	61.5	94.3	423.3	30.8	4.72	2.42
9	150.7	22.4	221.2	23.9	68.1	93.7	378.5	32.4	4.49	2.38
10	146.2	20.2	223.4	21.6	65.4	93.5	360.6	22.6	4.13	2.03
13	156	24.2	252.4	27.3	61.8	88.6	374.8	34.9	4.35	2.14
15	136	18	216.1	19.0	62.9	94.6	379.1	18.0	3.91	2.14
RNA										
1	282.6	28.6	563.7	32.2	50.1	88.9	882.4	40.6	5.62	2.39
3a	252.9	23.1	482.0	24.9	52.5	92.7	816.4	28.8	5.21	2.26
3b	273	25.9	496.7	27.4	55.0	94.4	689.8	32.3	5.24	2.15
4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	267.9	25	547.1	26.2	49.0	95.4	849.3	35.0	5.15	2.38
6	297.8	24.7	634.1	26.0	47.0	95.0	888.3	31.3	5.53	2.52
9	284.1	22	572.1	22.8	49.7	96.4	829.9	23.3	5.35	2.39
10	275.6	23.6	665.3	26.0	41.4	90.7	919.9	27.5	5.13	2.24
13	246.7	29	508.2	32.7	48.5	88.6	660.1	35.0	5.19	2.40
15	276	22	599.8	23.3	46.0	94.5	873.6	27.6	5.38	2.35

OTU, operational taxonomic unit.

curves (data not shown) as well as diversity indices revealed that the majority of the bacterial community was recovered by the surveying effort (**Table 7**).

Classification of the obtained 16S rRNA sequences revealed that *Proteobacteria* and *Bacteroidetes* were the most abundant bacterial phyla across all samples (approximately 78% and 20%, respectively). At higher taxonomic resolution, the majority of the obtained sequences was affiliated to 17 bacterial groups, clades, and genera (**Figure 2**). These groups represented different lineages within the *Alpha-*, *Beta-*, and *Gammaproteobacteria* and the *Bacteroidetes*. These results are in accordance with our previous study (Wemheuer et al., 2014) and recent investigations of bacterial communities in the North Sea (Alderkamp et al., 2006; Sapp et al., 2007; Teeling et al., 2012). However, in our previous study, the number of *Bacteroidetes* was rather low which can be attributed to the differences in primer pairs and variable regions of the 16S rRNA gene used in our previous study.

Alphaproteobacteria accounted for 50% of all sequences with a higher abundance at DNA and RNA level (59 and 41%, respectively). The opposite was recorded for the *Gammaproteobacteria,* which accounted for 16% (DNA) and 31% (RNA), respectively. The increased abundance of the *Gammaproteobacteria* was mainly attributed to the higher abundances of *Pseudospirrilii* and the SAR92 clade. Changes in the abundances of the different alphapoteobacterial taxa were mainly attributed to the overall low abundance of the SAR11 clade at RNA level. A low activity of SAR11 is supported by other studies (West et al., 2008; Lamy et al., 2010; Klindworth et al., 2014). For example, Lamy et al. found an overall low abundance and activity of the SAR11 clade in a *P. globosa* bloom in the eastern English Channel. In another study, Alonso and Pernthaler (2006b) showed that SAR11 is highly abundant but not very active in costal North Sea waters. In addition, West et al. (2008) demonstrated that SAR11 was more abundant at DNA level than at RNA level in the Southern Ocean.

Several groups responded significantly toward algal bloom presence at DNA and/or RNA level, e.g., the abundance of the SAR92 clade was three times higher at RNA level and in bloom presence. This is in accordance with previous studies (Pinhassi et al., 2005; West et al., 2008; Klindworth et al., 2014; Wemheuer et al., 2014). For example, a phytoplankton bloom induced by inorganic nutrient enrichment influenced SAR92 in a mesocosm experiment (Pinhassi et al., 2005). Klindworth et al. (2014) found that members of the *Rhodobacteraceae* and SAR92 clade exhibited high metabolic activity levels during a bloom succession, which indicates their important role during bloom events. In addition, the 16S cDNA estimates for SAR11 were notably lower in the earlier bloom sample. The authors suggest that members of this clade could not profit from the increasing availability of nutrients in the decaying bloom and thus were outcompeted by other clades. This is in line with our study in

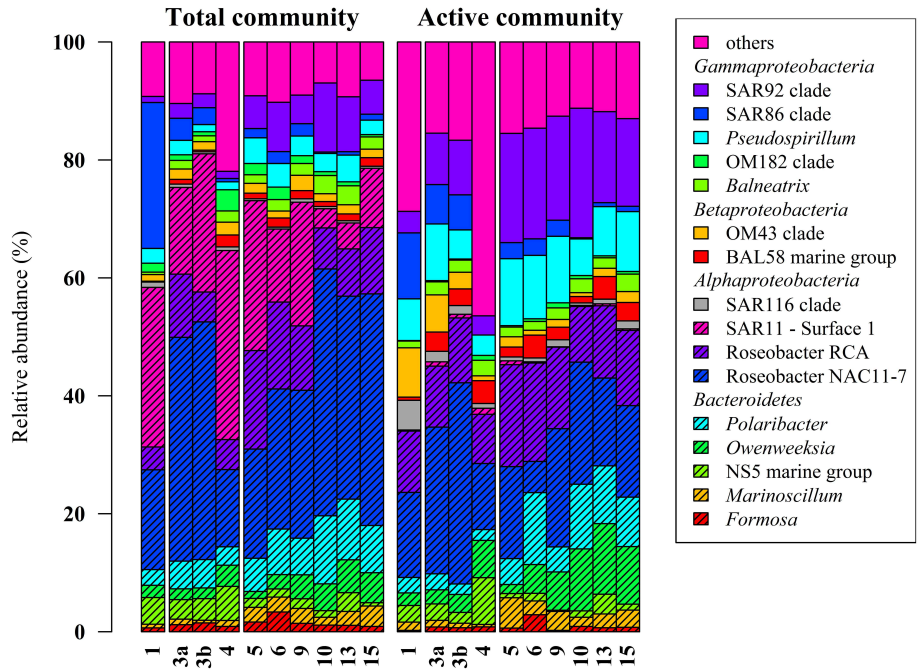

FIGURE 2 | Relative distribution of abundant bacterial lineages in the total (DNA-based) and active (RNA-based) bacterioplankton community at stations outside (1–4) and inside (5–15) the examined phytoplankton bloom. Only groups with an average abundance of more than 1% either at DNA or RNA level are shown. Station 1 is separated from the oher samples because it was located in a bloom outside of the area of the examined bloom.

which we found significant lower abundances of SAR11 in the bloom presence and at RNA level.

Members of the two genera *Marinoscillum* and *Polaribacter* were significantly more abundant in bloom samples both at DNA and RNA level. *Bacteroidetes* are widespread in marine systems and play an important role in organic matter degradation (Gómez-Pereira et al., 2010). The higher abundance of this phylum during the phytoplankton bloom was verified by recent findings (Alderkamp et al., 2006; Lamy et al., 2010; Tada et al., 2012; Teeling et al., 2012). The strongest increase in activity during the senescent stage of a *P. globosa* bloom in the North Sea was observed for *Bacteroidetes* (Alderkamp et al., 2006) A mesocosm experiment targeting bacterial succession patterns during a diatom bloom revealed that *Bacteroidetes* had a relatively high growth potential as the bloom peaked (Tada et al., 2012). The authors suggested that the early development contributed to the initial stage of bloom decomposition. Therefore, this phylum seems to benefit from the conditions provided by the algal bloom and might play an important role in the degradation of phytoplankton-derived organic matter. Klindworth et al. (2014) mapped metatranscriptomic data on assembled and taxonomically classified metagenomic data and found that *Formosa* and *Polaribacter* acted as major algal polymer degraders. A similar conclusion was drawn in a study of a *P. globosa* bloom in the eastern English Channel. Here, members of *Bacteroidetes* group dominated the activities and the abundances during the growth phase of the algae (Lamy et al., 2010).

Bloom Presence Affects Bacterioplankton Gene Expression

After removal of ribosomal RNA, nearly 45 million Illumina reads remained and were used for environmental gene expression analysis. Generated mRNA datasets were initially mapped on the 15 reference genomes belonging to abundant marine genera and lineages. However, only 10% of the sequences mapped to these reference genomes. Most of these sequences were affiliated to the genome of *P. temperata* RCA23 (see Supplementary Table S3). This strain was isolated in the German Wadden Sea (Giebel et al., 2011, 2013), and its genome was recently described (Voget et al., 2014).

Mapping mRNA datasets on genomes is a common approach when analyzing metatranscriptomic data (e.g., Gifford et al., 2013). The advantage of this approach is that community functions can be linked to a certain organism. However, most reads are not included in the analysis because reference genomes for many marine lineages are still missing. Another problem is the data normalization when mapping metatranscriptomic data on genomes. In a transcriptomic approach, all sequences derive from a single organism and data can be normalized by the number of reads mapped. However, in a metatranscriptomic approach, the amount of sequences affiliated to an organism is not only linked to its gene expression but also to the gene expression of all other community members. Thus, a decrease in abundance can be caused either by a lower gene expression of the organism or by an increased expression of other community members.

Here, we mapped the mRNA data on reference genomes representing marine lineages, which were abundant in the 16S rRNA gene analysis. These genomes included data from *P. temperata* RCA23 and HTCC2255, belonging to two *Roseobacter* lineages, the RCA and NAC11-7 clusters. Other reference genomes used were the genomes of HTCC2207 und MOLA455, both members of the SAR92 clade (Stingl et al., 2007; Courties et al., 2014). Mapping the datasets on assembled metagenomic and metatranscriptomic data resulted in an overall alignment rate of more than 86.22% (see **Table 5**). This overall high coverage is higher than in other studies. For example, Kopf et al. (2015) mapped up to 80% of their metatranscriptomic data on the corresponding metagenome. We were able to map our data on approximately 600,000 genes with almost 800,000 different functions.

Most of the genes affiliated to the two members of the SAR92 clade were upregulated which corresponds to their increasing abundance at 16S rRNA transcipt level. For example, one leucyl-tRNA synthetase and three isoleucyl-tRNA synthetases affiliated to MOLA455 were significantly upregulated in the bloom (**Figure 3A**). Morover, two leucine-tRNAs were significantly upregulated in the bloom in *P. temperata* RCA23 (**Figure 3B**). In addition, an isoleucyl-tRNA synthetase affiliated to *P. temperata* RCA23 (**Figure 3A**) and a isoleucyl-tRNA synthetase and a leucyl-tRNA synthetase of *P. temperata* RCA23 (**Figure 3B**) were marginal significantly upregulated in the bloom ($P < 0.1$). This is in line with a study by West et al. (2008). The authors found that the *Roseobacter* groups NAC11-7 and RCA as well as the SAR92 clade were the most important contributors to leucine incorporation during the peak of a naturally iron-fertilized phytoplankton bloom in the Southern Ocean. This result is confirmed by a study about a *P. globosa* bloom in the English Channel (Lamy et al., 2010). Here, *Bacteroidetes* and *Gammaproteobacteria* were the most abundant and active groups during the growth period of the algae. *Gammaproteobacteria* and *Alphaproteobacteri*a dominated by the *Roseobacter* clade accounted for the major part of leucine incorporation after the disappearance of the bloom. In addition, the contributions of different bacterial groups to bulk abundance and leucine incorporation were partly correlated with cell-specific exoproteolytic and exoglucosidic activities and with particulate organic carbon. This indicates some specificity of these bacterial groups with respect to their ecological role in the environment. Interestingly, we identified two beta-glucosidases affiliated to HTCC2207 being expressed only in bloom samples (**Figure 3A**). In the study of Teeling et al. (2012), metagenomic and metaproteomic data indicated the presence of distinct sets of carbohydrate-active enzymes (CAZymes) and transporters, which suggested a positive selection for bacteria with the capacity to decompose phytoplankton biomass. Four HTCC2207 indicator genes have been described to contain cadherin domains involved in complex carbohydrate degradation via cell aggregation and direct binding to cellulose, xylan, and related compounds (Gifford et al., 2013). This might explain the increase of the SAR92 clade as observed in the present study. In addition, the role of *Gammprotaobacteria* during polysaccharide degradation has been recently addressed in a study by Wietz et al. (2015). The authors showed that alginate and agarose degradation covaried with the abundance of different lineages within the *Gammaproteobacteria*.

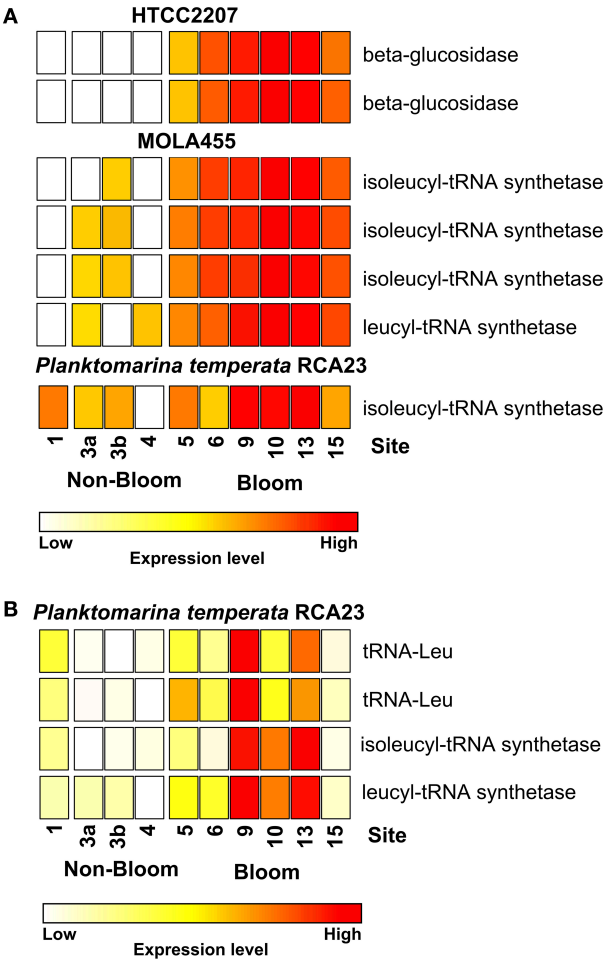

FIGURE 3 | Expression of certain gene involved in leucine/isoleucine incorporation and polymer degradation inside and outside the examined phytoplankton bloom. Data either derives from the mapping on the assembled metagenomic and metatranscriptomic data **(A)** or from the mapping on reference genomes **(B)**. Gene expressions levels are log2 transformed and standardized. To obtain only positive values, normalized gene expression values were increased by one prior to log transformation.

Interestingly, a major fraction of the expressed genes affiliated to SAR92 clade and other bacterial lineages was linked to several genes such as RNA polymerases, heat shock proteins, chaperons, sigma factors, and ribosomal proteins. This is consistent with a study from Klindworth et al. (2014). In this study, the most abundant mRNA transcripts with known functions coded for housekeeping genes including DNA gyrase, elongation factors, and sigma factors. The authors indicated that this reflects differences in nutritional ecological strategies of the dominant bacterial classes.

However, the high number of heat shock and other stress-related genes overexpressed in bloom samples in members of the SAR92 clade might not necessarily reflect its ecological role as

a polymer degrader but might also be caused by a higher stress tolerance toward the rapidly changing conditions during the phytoplankton bloom (Wemheuer et al., 2014). PON, Chl a, and phaeopigments of stations in the bloom area were significantly higher than that outside the bloom area. An increasing pH from 7.9 to 8.7 was observed as a result of CO_2 net fixation into the alga during a phytoplankton bloom (Brussaard et al., 1996). Members of SAR92 clade might benefit from bloom conditions due to their high stress tolerance level rather than filling one of the specialized ecological niches formed during a phytoplankton bloom.

Adaption to Environmental Changes by Higher Genome Plasticity

The overall expression level of *P. temperata* RCA23 was higher inside the bloom (**Figure 4**). Numerous genes encoding for transposases in its genome were highly overexpressed. Transposases are the most abundant and most ubiquitous genes in nature (Aziz et al., 2010). In addition, investigation of the metatranscriptomic bins revealed the presence of several transposases that were affiliated to *P. temperata* and overexpressed in the bloom (**Figure 4**). It has been shown that some bacteria expressed transposases under changing environmental conditions to rearrange genome architecture. For example, up to 81 genes encoded for transposases were upregulated in *Microcystis aeruginosa* relative to the control when grown on urea (Steffen et al., 2014). Genome rearrangements and the resulting genome mosaics have been also found in other members of the *Roseobacter* clade. The genomes of *Octadecabacter arcticus* and *O. antarcticus* are highly different despite their similarity on 16S rRNA gene sequence level and the presence of some unique gene features (Vollmers et al., 2013). This is attributed to genomic rearrangements caused by an unusually high number of transposases in the genomes of both *Octadecabacter* strains. We assume that the recorded overexpression could result in a higher genome plasticity/heterogeneity of this population and thus might be a possible adaptation strategy of *P. temperata* to environmental changes. Moreover, as found in other members of the *Roseobacter* clade, it might be one of the key features of this group explaining its high abundance in marine ecosystems and its ability to adapt to various marine niches. However, comparative genome studies are missing because only one genome of the genus *Planktomarina* is currently available. Consequently, this issue cannot been fully answered yet.

Conclusions

Active bacterial communities in the North Sea are dominated by only a few marine groups such as the *Roseobacter* RCA cluster. Some of these lineages responded significantly toward the *P. globosa* bloom investigated in this study. For example, the SAR92 clade was three times more abundant at active bacterial community level and in bloom presence. The metatranscriptomic approach revealed that these groups are not dominated by well-studied isolates or type species as only 10% of all metatranscriptomic sequences mapped on reference genomes. Therefore, *in situ* experiments employing available isolates do not necessarily reflect environmental conditions and, thus, only provide limited information on the ecological role of the studied isolates. However, mapping these reads on assembled metagenomic and metatranscriptomic sequences led to an overall mapping rate of more than 85% demonstrating the power of this combined approach. The functional analysis performed in this study provides insights into gene expression patterns of the abundant community members. The high abundance of the SAR92 clade, which is supposed to be involved in polymer-degradation during and after the bloom, is attributed to a higher stress tolerance indicated by the high number of

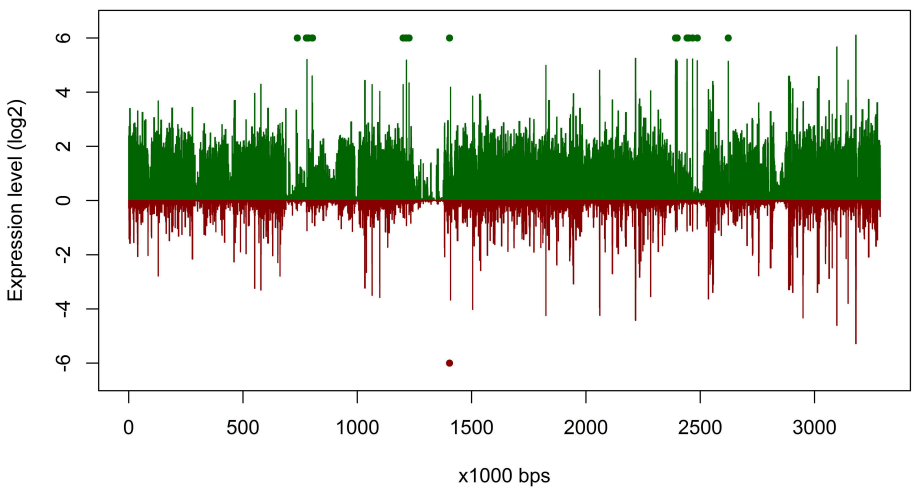

FIGURE 4 | Gene expression of *Planktomarina temperata* RCA 23 inside and outside the examined phytoplankton bloom. Gene expressions are log2 transformed mean values from the three non-bloom and six bloom stations, respectively. To obtain only positive values, normalized gene expression values were increased by one prior to log transformation. Expression inside the bloom is depicted in green, outside the bloom in red. Green/red dots mark the position of transposases, which were highly upregulated.

heat shock expressed in the bloom. Although the number of field studies targeting the active bacterial community either by metatranscriptomic or metaproteomic approaches has been increased over the past years, the complex dynamics of marine environments are still largely unexplored. This study provides a deep insight into structural and functional responses of the bacterioplankton community toward a phytoplankton bloom. Therefore, it paved the way for a better understanding of the complex dynamics of marine bacteria and their interactions with the surrounding environment.

Author Contributions

RD and BW conceived and designed the experiments; BW, FW, JH, SV, and FM performed the experiments and analyzed the data; BW, FW, and RD wrote the paper; all authors reviewed, edited, and approved the manuscript.

Acknowledgments

We thank the crew of the research vessel Heincke for their valuable support during the sampling campaign. We are grateful to Peter Meinicke and Heiko Liesegang for the help during data analysis. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) as part of the collaborative research center TRR51 and the Alfred Wegener Institute under grant number AWI-HE327_00. Additionally, we acknowledge support by DFG and the Open Access Publication Funds of the Göttingen University.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00805

References

Alderkamp, A. C., Sintes, E., and Herndl, G. J. (2006). Abundance and activity of major groups of prokaryotic plankton in the coastal North Sea during spring and summer. *Aquat. Microb. Ecol.* 45, 237–246. doi: 10.3354/ame045237

Alonso, C., and Pernthaler, J. (2006a). Concentration-dependent patterns of leucine incorporation by coastal picoplankton. *Appl. Environ. Microbiol.* 72, 2141–2147. doi: 10.1128/AEM.72.3.2141-2147.2006

Alonso, C., and Pernthaler, J. (2006b). *Roseobacter* and SAR11 dominate microbial glucose uptake in coastal North Sea waters. *Environ. Microbiol.* 8, 2022–2030. doi: 10.1111/j.1462-2920.2006.01082.x

Arrigo, K. R. (2005). Marine microorganisms and global nutrient cycles. *Nature* 437, 349–355. doi: 10.1038/nature04159

Azam, F. (1998). Microbial control of oceanic carbon flux: the plot thickens. *Science* 280, 694–696. doi: 10.1126/science.280.5364.694

Azam, F., Fenchel, T., Field, J. G., Gray, J. S., Meyer-Reil, L. A., and Thingstad, F. (1983). The ecological role of water-column microbes in the sea. *Mar. Ecol. Prog. Ser.* 10, 257–263. doi: 10.3354/meps010257

Aziz, R. K., Breitbart, M., and Edwards, R. A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* 38, 4207–4217. doi: 10.1093/nar/gkq140

Blazewicz, S. J., Barnard, R. L., Daly, R. A., and Firestone, M. K. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME J.* 7, 2061–2068. doi: 10.1038/ismej.2013.102

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P., and Tyson, G. W. (2012). Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nat. Methods* 9, 425–426. doi: 10.1038/nmeth.1990

Brussaard, C. P. D., Gast, G. J., van Duyl, F. C., and Riegman, R. (1996). Impact of phytoplankton bloom magnitude on a pelagic microbial food web. *Mar. Ecol. Prog. Ser.* 144, 211–221. doi: 10.3354/meps144211

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421

Campbell, B. J., and Kirchman, D. L. (2013). Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. *ISME J.* 7, 210–220. doi: 10.1038/ismej.2012.93

Cole, J. J., Findlay, S., and Pace, M. L. (1988). Bacterial production in fresh and saltwater ecosystems: a cross-system overview. *Mar. Ecol. Prog. Ser.* 43, 1–10. doi: 10.3354/meps043001

Courties, A., Riedel, T., Jarek, M., Papadatou, M., Intertaglia, L., Lebaron, P., et al. (2014). Draft genome sequence of the gammaproteobacterial strain MOLA455, a representative of a ubiquitous proteorhodopsin-producing group in the ocean. *Genome Announc.* 2, e01203–e01213. doi: 10.1128/genomea.01203-13

DeLong, E. F., and Karl, D. M. (2005). Genomic perspectives in microbial oceanography. *Nature* 437, 336–342. doi: 10.1038/nature04157

Ducklow, H., Kirchman, D. L., Quinby, H. L., Carlson, C. A., and Dam, H. G. (1993). Stocks and dynamics of bacterioplankton carbon during the spring bloom in the eastern North Atlantic Ocean. *Deep Sea Res. II* 40, 245–263. doi: 10.1016/0967-0645(93)90016-G

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Fandino, L. B., Riemann, L., Steward, G. F., Long, R. A., and Azam, F. (2001). Variations in bacterial community structure during a dinoflagellate bloom analyzed by DGGE and 16S rDNA sequencing. *Aquat. Microb. Ecol.* 23, 119–130. doi: 10.3354/ame023119

Giebel, H.-A., Kalhoefer, D., Gahl-Janssen, R., Choo, Y.-J., Lee, K., Cho, J.-C., et al. (2013). *Planktomarina temperata* gen. nov., sp. nov., belonging to the globally distributed RCA cluster of the marine Roseobacter clade, isolated from the German Wadden Sea. *Int. J. Syst. Evol. Microbiol.* 63, 4207–4217. doi: 10.1099/ijs.0.053249-0

Giebel, H.-A., Kalhoefer, D., Lemke, A., Thole, S., Gahl-Janssen, R., Simon, M., et al. (2011). Distribution of *Roseobacter* RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. *ISME J.* 5, 8–19. doi: 10.1038/ismej.2010.87

Gifford, S. M., Sharma, S., Booth, M., and Moran, M. A. (2013). Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J.* 7, 281–298. doi: 10.1038/ismej.2012.96

Gomes, A., Gasol, J. M., Estrada, M., Franco-Vidal, L., Díaz-Pérez, L., Ferrera, I., et al. (2015). Heterotrophic bacterial responses to the winter–spring phytoplankton bloom in open waters of the NW Mediterranean. *Deep Sea Res. I* 96, 59–68. doi: 10.1016/j.dsr.2014.11.007

Gómez-Pereira, P. R., Fuchs, B. M., Alonso, C., Oliver, M. J., van Beusekom, J. E., and Amann, R. (2010). Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean. *ISME J.* 4, 472–487. doi: 10.1038/ismej.2009.142

Grossart, H. P., Levold, F., Allgaier, M., Simon, M., and Brinkhoff, T. (2005). Marine diatom species harbour distinct bacterial communities. *Environ. Microbiol.* 7, 860–873. doi: 10.1111/j.1462-2920.2005.00759.x

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L., J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Karner, M., and Herndl, G. (1992). Extracellular enzymatic activity and secondary production in free-living and marine-snow-associated bacteria. *Mar. Biol.* 113, 341–347.

Klindworth, A., Mann, A. J., Huang, S., Wichels, A., Quast, C., Waldmann, J., et al. (2014). Diversity and activity of marine bacterioplankton during a diatom bloom in the North Sea assessed by total RNA and pyrotag sequencing. *Mar. Genomics* 18(Pt B), 185–192. doi: 10.1016/j.margen.2014.08.007

Kopf, A., Kostadinov, I., Wichels, A., Quast, C., and Glöckner, F. O. (2015). Metatranscriptome of marine bacterioplankton during winter time in the North Sea assessed by total RNA sequencing. *Mar. Genomics* 19, 45–46. doi: 10.1016/j.margen.2014.11.001

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611

Lamy, D., Obernosterer, I., Laghdass, M., Artigas, L. F., Breton, E., Grattepanche, J. D., et al. (2010). Temporal changes of major bacterial groups and bacterial heterotrophic activity during a *Phaeocystis globosa* bloom in the eastern English Channel. *Aquat. Microb. Ecol.* 58, 95–107. doi: 10.3354/ame01359

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Liu, M., Dong, Y., Zhang, W., Sun, J., Zhou, F., Ren, J., et al. (2013). Diversity of bacterial community during spring phytoplankton blooms in the central Yellow Sea. *Can. J. Microbiol.* 59, 324–332. doi: 10.1139/cjm-2012-0735

Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi: 10.1093/nar/gkr1044

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

McQuatters-Gollop, A., Raitsos, D. E., Edwards, M., Pradhan, Y., Mee, L. D., Lavender, S. J. (2007). A long-term chlorophyll data set reveals regime shift in North Sea phytoplankton biomass unconnected to nutrient trends. *Limnol. Oceanogr.* 52, 635–648. doi: 10.4319/lo.2007.52.2.0635

Meinicke, P. (2015). UProC: tools for ultra-fast protein domain classification. *Bioinformatics* 31, 1382–1388. doi: 10.1093/bioinformatics/btu843

Muyzer, G., Teske, A., Wirsen, C. O., and Jannasch, H. W. (1995). Phylogenetic relationships of *Thiomicrospira* species and their identification in deep-sea hydrothermal vent samples by denaturing gradient gel electrophoresis of 16S rDNA fragments. *Arch. Microbiol.* 164, 165–172. doi: 10.1007/BF02529967

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155. doi: 10.1093/nar/gks678

Pinhassi, J., Sala, M. M., Havskum, H., Peters, F., Guadayol, O., and Malits, A. (2004). Changes in bacterioplankton composition under different phytoplankton regimens. *Appl. Environ. Microbiol.* 70, 6753–6766. doi: 10.1128/AEM.70.11.6753-6766.2004

Pinhassi, J., Simó, R., González, J. M., Vila, M., Alonso-Sáez, L., Kiene, R. P., et al. (2005). Dimethylsulfoniopropionate turnover is linked to the composition and dynamics of the bacterioplankton assemblage during a microcosm phytoplankton bloom. *Appl. Environ. Microbiol.* 71, 7650–7660. doi: 10.1128/AEM.71.12.7650-7660.2005

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Sapp, M., Wichels, A., Wiltshire, K. H., and Gerdts, G. (2007). Bacterial community dynamics during the winter–spring transition in the North Sea. *FEMS Microbiol. Ecol.* 59, 622–637. doi: 10.1111/j.1574-6941.2006.00238.x

Schlitzer, R. (2015). *Ocean Data View*. Available online at: http://odv.awi.de

Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V., and Lancelot, C. (2005). *Phaeocystis* blooms in the global ocean and their controlling mechanisms: a review. *J. Sea Res.* 53, 43–66. doi: 10.1016/j.seares.2004.01.008

Steffen, M. M., Dearth, S. P., Dill, B. D., Li, Z., Larsen, K. M., Campagna, S. R., et al. (2014). Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *ISME J.* 8, 2080–2092. doi: 10.1038/ismej.2014.78

Stingl, U., Desiderio, R. A., Cho, J. C., Vergin, K. L., and Giovannoni, S. J. (2007). The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin. *Appl. Environ. Microbiol.* 73, 2290–2296. doi: 10.1128/AEM.02559-06

Tada, Y., Taniguchi, A., Sato-Takabe, Y., and Hamasaki, K. (2012). Growth and succession patterns of major phylogenetic groups of marine bacteria during a mesocosm diatom bloom. *J. Oceanogr.* 68, 509–519. doi: 10.1007/s10872-012-0114-z

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., et al. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* 336, 608–611. doi: 10.1126/science.1218344

Veldhuis, M. J. W., and Wassmann, P. (2005). Bloom dynamics and biological control of a high biomass HAB species in European coastal waters: a *Phaeocystis* case study. *Harmful Algae* 4, 805–809. doi: 10.1016/j.hal.2004.12.004

Voget, S., Wemheuer, B., Brinkhoff, T., Vollmers, J., Dietrich, S., Giebel, H.-A., et al. (2014). Adaptation of an abundant *Roseobacter* RCA organism to pelagic systems revealed by genomic and transcriptomic analyses. *ISME J.* 9, 371–384. doi: 10.1038/ismej.2014

Vollmers, J., Voget, S., Dietrich, S., Gollnow, K., Smits, M., Meyer, K., et al. (2013). Poles apart: arctic and antarctic *Octadecabacter* strains share high genome plasticity and a new type of Xanthorhodopsin. *PLoS ONE* 8:e63422. doi: 10.1371/journal.pone.0063422

Weinbauer, M. G., Fritz, I., Wenderoth, D. F., and Höfle M. G. (2002). Simultaneous extraction from bacterioplankton of total RNA and DNA suitable for quantitative structure and function analyses. *Appl. Environ. Microbiol.* 68, 1082–1087. doi: 10.1128/AEM.68.3.1082-1087.2002

Wemheuer, B., Güllert, S., Billerbeck, S., Giebel, H.-A., Voget, S., Simon, M., et al. (2014). Impact of a phytoplankton bloom on the diversity of the active bacterial community in the southern North Sea as revealed by metatranscriptomic approaches. *FEMS Microbiol. Ecol.* 87, 378–389. doi: 10.1111/1574-6941.12230

Wemheuer, B., Wemheuer, F., and Daniel, R. (2012). RNA-based assessment of diversity and composition of active archaeal communities in the German Bight. *Archaea* 2012:695826. doi: 10.1155/2012/695826

Wemheuer, B., Taube, R., Akyol, P., Wemheuer, F., and Daniel, R. (2013). Microbial diversity and biochemical potential encoded by thermal spring metagenomes derived from the Kamchatka Peninsula. *Archaea* 2013, 13. doi: 10.1155/2013/136714

West, N. J., Obernosterer, I., Zemb, O., and Lebaron, P. (2008). Major differences of bacterial diversity and activity inside and outside of a natural iron-fertilized phytoplankton bloom in the Southern Ocean. *Environ. Microbiol.* 10, 738–756. doi: 10.1111/j.1462-2920.2007.01497.x

Wietz, M., Wemheuer, B., Simon, H., Giebel, H.-A., Seibt, M. A., Daniel, R., et al. (2015). Bacterial community dynamics during polysaccharide degradation at contrasting sites in the Southern and Atlantic Oceans. *Environ. Microbiol.* doi: 10.1111/1462-2920.12842. [Epub ahead of print].

Wiltshire, K. H., Kraberg, A., Bartsch, I., Boersma, M., Franke, H.-D., Freund, J., et al. (2010). Helgoland roads, North Sea: 45 years of change. *Estuaries Coasts* 33, 295–310. doi: 10.1007/s12237-009-9228-y

Wiltshire, K. H., Malzahn, A. M., Wirtz, K., Janisch, S., Mangelsdorf, P., and Manly, B. F. J. (2008). Resilience of North Sea phytoplankton spring bloom dynamics: an analysis of long-term data at Helgoland Roads. *Limnol. Oceanogr.* 53, 1294–1302. doi: 10.4319/lo.2008.53.4.1294

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 November 2015
doi: 10.3389/fmicb.2015.01236

Metagenome and Metatranscriptome Revealed a Highly Active and Intensive Sulfur Cycle in an Oil-Immersed Hydrothermal Chimney in Guaymas Basin

*Ying He[1,2], Xiaoyuan Feng[1], Jing Fang[1], Yu Zhang[1,2,3] and Xiang Xiao[1,2,3]**

[1] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, [2] State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai, China, [3] Institute of Oceanology, Shanghai Jiao Tong University, Shanghai, China

The hydrothermal vent system is a typical chemosynthetic ecosystem in which microorganisms play essential roles in the geobiochemical cycling. Although it has been well-recognized that the inorganic sulfur compounds are abundant and actively converted through chemosynthetic pathways, the sulfur budget in a hydrothermal vent is poorly characterized due to the complexity of microbial sulfur cycling resulting from the numerous parties involved in the processes. In this study, we performed an integrated metagenomic and metatranscriptomic analysis on a chimney sample from Guaymas Basin to achieve a comprehensive study of each sulfur metabolic pathway and its hosting microorganisms and constructed the microbial sulfur cycle that occurs in the site. Our results clearly illustrated the stratified sulfur oxidation and sulfate reduction at the chimney wall. Besides, sulfur metabolizing is closely interacting with carbon cycles, especially the hydrocarbon degradation process in Guaymas Basin. This work supports that the internal sulfur cycling is intensive and the net sulfur budget is low in the hydrothermal ecosystem.

Keywords: hydrothermal vent, metagenomics, metatranscriptomics, sulfur cycle, carbon cycle

OPEN ACCESS

Edited by:
Roy D. Sleator,
Cork Institute of Technology, Ireland

Reviewed by:
Huiluo Cao,
The University of Hong Kong, Hong Kong
Pat G. Casey,
University College Cork, Ireland

***Correspondence:**
Xiang Xiao
xoxiang@sjtu.edu.cn

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *12 May 2015*
Accepted: *26 October 2015*
Published: *10 November 2015*

Citation:
He Y, Feng X, Fang J, Zhang Y and Xiao X (2015) Metagenome and Metatranscriptome Revealed a Highly Active and Intensive Sulfur Cycle in an Oil-Immersed Hydrothermal Chimney in Guaymas Basin. Front. Microbiol. 6:1236. doi: 10.3389/fmicb.2015.01236

INTRODUCTION

Hydrothermal vents are often discovered in ocean ridges where hydrothermal fluid is emitted after the hydrothermal circulation and alteration of seawater entrained through geothermally heated subseafloor basalt (Von Damm, 1990). The deep-sea hydrothermal vent fluid is commonly characterized by its high temperature, varied salinity, enriched metallic elements, and particularly high contents of reduced chemicals, such as H_2, CH_4, and H_2S (Jannasch and Mottl, 1985). A thermodynamic non-equilibrium is created when the hydrothermal vent fluid encounters sea water that is cold and at a rather high oxidative state, which allows various abiotic and biotic reactions occur. Thus, the hydrothermal vent system is a typical chemosynthetic ecosystem in which microorganisms play essential roles in the generation, consumption, and modification of energy available in the environment (Reysenbach and Shock, 2002).

In the hydrothermal vent ecosystem, almost all types of inorganic sulfur compounds (e.g., S^{2-}, S, $S_2O_2^{2-}$, SO_2, $S_2O_3^{2-}$, and SO_4^{2-}) are abundant and actively converted through chemosynthetic pathways to provide energy and thus sustain the microbial population in the ecosystem (Nakagawa et al., 2005). For example, in the Lost City hydrothermal field, the dominant *Thiomicrospira*-like group, which consists of sulfur-oxidizing chemolithoautotrophs, was observed in the carbonate chimney (Brazelton and Baross, 2010). In the Lau Basin hydrothermal vent field, sulfur-oxidizing Alphaproteobacteria, Gammaproteobacteria, and Epsilonproteobacteria have been suggested to be dominant in the exterior chimney, whereas putative sulfur-reducing Deltaproteobacteria are dominant in the interior of the chimney (Sylvan et al., 2013). In the Guaymas Basin hydrothermal vent field, sulfate-reducing microorganisms, e.g., Desulfobacterales, have been detected and are hypothesized to be involved in the anaerobic methane-oxidation process (Biddle et al., 2012). Moreover, the sulfur cycling is alternated by the chemical reactions that occur during the emitting and growth of the hydrothermal vent. Reduced sulfur compounds are extremely sensitive to oxidants and easily precipitated with metal ions to form chimney or nodule structures (Orcutt et al., 2011). Moreover, shifts in temperature and fluid composition have been observed during the life span of a hydrothermal vent. For example, at 9°N East Pacific Rise, Bio9 vent fluids were 368°C in 1991, increased to an estimated temperature greater than or equal to 388°C after a second volcanic event in 1992, and thereafter declined over the next similar to 2 years reaching a temperature of 365°C in December 1993 (Fornari et al., 1998). The hydrogen concentration in the hydrothermal plum in the NE Lau Basin dropped from 14843 nM in 2008 to 4410 nM in 2010 then further to 7 nM in 2012 (Baumberger et al., 2014). As a result, environmental fluctuations may be induced between sulfate- and sulfur-reducing archaea and contribute to the diverse roles of these microorganisms in the ecosystem (Teske et al., 2014). Therefore, a better understanding of sulfur cycling is essential for describing the geobiochemistry and providing hints to identify the life status of a hydrothermal vent ecosystem.

Due to the complexity of microbial sulfur cycling resulting from the numerous parties involved in the process, the sulfur budget in a hydrothermal vent is poorly characterized. To date, most studies have focused on the abundance and diversity of sulfur oxidizers and sulfate reducers in environmental samples through a metagenomic approach (Nakagawa et al., 2005). The exception is the study conducted by Anantharaman et al. (2013), who combined metatranscriptomic and metagenomic analyses of a hydrothermal plume sample and demonstrated the novel metabolic potentials of the SUP05 group of uncultured sulfur-oxidizing Gammaproteobacteria. However, this finding is based on the near-complete genomes of two SUP05 populations, and the information is restricted to this particular group of sulfur oxidizers (Anantharaman et al., 2013). The in-depth mining of the metatranscriptomic data remains too scarce to allow construction of the entire sulfur cycle and thus further illustrate the interactions of this process with the biological cycling of C, N, and O elements.

The Guaymas Basin in the Gulf of California is a young marginal rift basin characterized by the active hot venting of reduced sulfur compounds and the rapid deposition of organic-rich sediments. These features make the sulfur cycle in this ecosystem particularly intensive and closely interact with the carbon cycle, including hydrocarbon degradation (Bergmann et al., 2011). Thus, this sampling site is ideal for illustrating all of the possible microbial sulfur metabolic pathways and to evaluate the maximal biomass contribution of sulfur-metabolizing microorganism to the hydrothermal vent ecosystem. In this study, we performed an integrated metagenomic and metatranscriptomic analysis on a chimney sample from Guaymas Basin to achieve a comprehensive study of each sulfur metabolic pathway and its hosting microorganisms and constructed the microbial sulfur cycle that occurs in the site.

RESULTS

Composition of the Microbial Community

The composition and function of this microbial community were assessed at both the DNA and RNA levels to estimate the community metabolic potential and activity, respectively. The metagenome and metatranscriptome sequencing resulted in 199,903,215 and 1,885,022,958 bp clean sequences, respectively (**Table 1**). The metagenome raw reads were assembled into 49,055 contigs with an average length of 544 bp. In total, 5,417,253 reads (26.2%) from the metatranscriptome were mapped onto metagenomic contigs for quantification of the gene transcripts. 222 and 690,059 16S rRNA gene fragments were identified from the metagenome and metatranscriptome, respectively. The class-level taxonomic compositions of the metagenome and metatranscriptome revealed obvious differences in the presence and the activity of microbes in this community (**Table 1**). At the DNA level (**Figure 1A**), Archaeoglobi were found to be the most abundant, with 24.0% of the sequences assigned, and followed by Deltaproteobacteria (23.6%) and

TABLE 1 | Summary of the metagenome and metatranscriptome.

	Metagenome	Metatranscriptome
Size of raw reads (bp)	199,903,215	1,885,022,958
Total no. of raw reads	512,830	20,714,538
Size of assembled contigs (bp)	26,703,275	–
Total assembled contigs	49,055	–
Average contig length (bp)	544	–
Average GC content of assembled contigs (%)	43	
Total no. of genes encoding in the contigs	53,034	–
Total no. of metatranscriptomic reads mapped to the metagenome	–	5,417,253
Total no. of 16S rRNA sequences	222	690,059

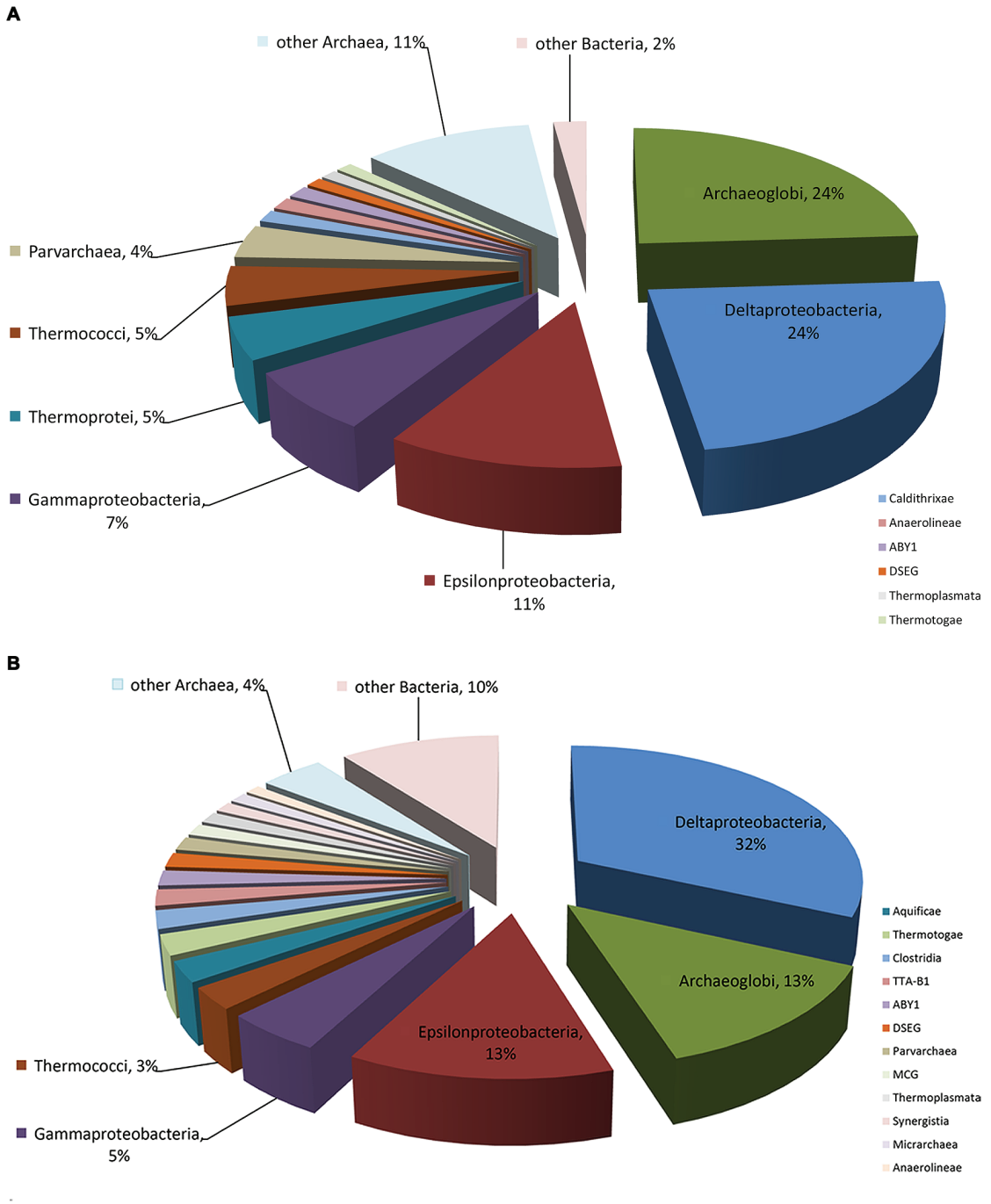

FIGURE 1 | Microbial composition of the enriched AOM-SR community. Detailed information is displayed in **Table 1**. **(A)** Percentage of the microbial community determined from the 16S rRNA gene sequences retrieved from the metagenome. **(B)** Percentage of the microbial community determined from the 16S rRNA gene sequences retrieved from the metatranscriptome.

Epsilonproteobacteria (11.3%). At the RNA level (**Figure 1B**), the same dominant groups were found: Deltaproteobacteria (31.8%), Archaeoglobi (13.3%), and Epsilonproteobacteria (12.8%). As reported previously (He et al., 2013), 53,034 gene features were predicted and then followed by manual examination and 19,491 gene features (36.8%) were considered to have expressions determined by transcriptomic reads mapping (see Materials and Methods). A total of 8929 (45.3%) and 4628 (23.7%) of all of the expressed genes were assigned (based on the BLAST results as described in Section "Materials and Methods") to Bacteria and Archaea, respectively, and the remaining sequences were not assigned to any category. Among

the 13,557 expressed genes with taxonomic information, 2135 (15.7%) were from the highly abundant Archaeoglobi, which is consistent with the results from the 16S rRNA gene analysis. Although the assignment of bacterial genes could not be resolved well at the family level, the dominance of Deltaproteobacteria and Epsilonproteobacteria was still observed. As the archaeal cells typically have fewer copies of the 16S rRNA gene compared with bacterial cells, the proportion of active Archaeoglobi in this community was underestimated. Nevertheless, the predominant active players in this microbial community were Deltaproteobacteria, Archaeoglobi, and Epsilonproteobacteria.

The *de novo* assembly of metagenomic reads and binning by tetranucleotide signatures (Dick et al., 2009) identified three genomic bins (Supplementary Figure S1 and Supplementary Table S1). These three bins (herewith denoted bin20, bin21, and bin22) were assigned based on their phylogenomic marker genes to *Desulfobacteraceae*, Desulfovibrionales and *Archaeoglobus*. The identified genes in the obtained bins ranged from 486 to 1224. The genome completeness was estimated to range from ∼10 to 34%, based on single-copy gene estimation (Supplementary Table S1). These three genomic bins will improve the taxonomic assignment of the expressed genes and the reconstruction of the metabolic pathways.

Sulfur Metabolism

The genes involved in the oxidation of reduced sulfur (ORS) are sulfide quinone oxidoreductase (*sqr*), which mediates the oxidation of sulfide (HS^-) to elemental sulfur (S^0), the Sox enzyme complex (*soxABXYZ*), which is responsible for the oxidation of thiosulfate ($S_2O_3^{2-}$) to elemental sulfur, the reverse dissimilatory sulfite reductase complex (*rdsr*), which is responsible for the oxidation of elemental sulfur to sulfite (SO_3^{2-}), and adenosine 5′-phosphosulfate reductase (*apr*) and sulfate adenylyltransferase (*sat*) for oxidation of sulfite to sulfate (SO_4^{2-}; Anantharaman et al., 2013). Conversely, the genes associated with the dissimilatory sulfate reduction (DSR) pathway (Fritz et al., 2002) are *sat*, *apr,* and sulfite reductase (*dsr*). The repertoire of genes associated with the ORS and DSR pathways were found to be expressed in this community (**Table 2**). Both *apr* and *dsr* were found at high expression levels in bin21 and bin22, confirming their active presence in SRB and *Archaeoglobus*. The *sqr* gene, key gene in the ORS pathway, is found present and active in Epsilonproteobacteria, of which the most highly expressed representative was classified into *Sulfurimonas* (**Figure 2**) that is one of the most abundant sulfur-oxidizing bacteria found in hydrothermal vent chimneys (Cao et al., 2014). The *sox* genes were not identified in either the metagenome or metatranscriptome (**Table 2**). In Epsilonproteobacteria, the proposed microorganism in the present study to perform the ORS pathway, *sat* gene was found to exhibit high and medium expression levels (**Table 2**). However, either *aprAB* or *dsrAB* was identified in the metagenome or metatranscriptome. This finding may be due to the fact that the 454-based metagenomes are still with low coverage and unable to present all the important functional genes. In

TABLE 2 | Genes identified in the sulfur metabolic pathway in the microbial community.

Gene name	Abbrevations	Deltaproteobacteria			Archaeoglobales			Epsilonproteobacteria		
		Assigned taxonomy*		FPKM#	Assigned taxonomy*		FPKM#	Assigned taxonomy*		FPKM#
		Bin	BLAST		Bin	BLAST		Bin	BLAST	
Sulfate adenylyltransferase	*Sat*	–	Deltaproteobacteria	42.94	–	–	–	–	Epsilonproteobacteria	2.86
Adenylyl-sulfate reductase, subunit A	*aprA*	bin21	Desulfovibrionales	2502.51	bin22	*Archaeoglobus fulgidus*	761.68	–	–	–
Adenylyl-sulfate reductase, subunit B	*aprB*	bin21	Desulfovibrionales	235.17	–	*Archaeoglobus*	53.53	–	–	–
Sulfite reductase alpha subunit	*dsrA*	bin21	Deltaproteobacteria	221.81	bin22	*Archaeoglobus*	1914.15	–	–	–
Sulfite reductase beta subunit	*dsrB*	bin21	Deltaproteobacteria	1621.06			–	–	–	–
Sulfide:quinone reductase	*Sqr*	–	–	–	–	–	–	–	Epsilonproteobacteria	137.30

**The taxonomy assignments were determined by two methods, as described in Section "Materials and Methods." The binning index is explained in Supplementary Table S1. #FPKM is based on the maximal expression value of the annotated genes.*

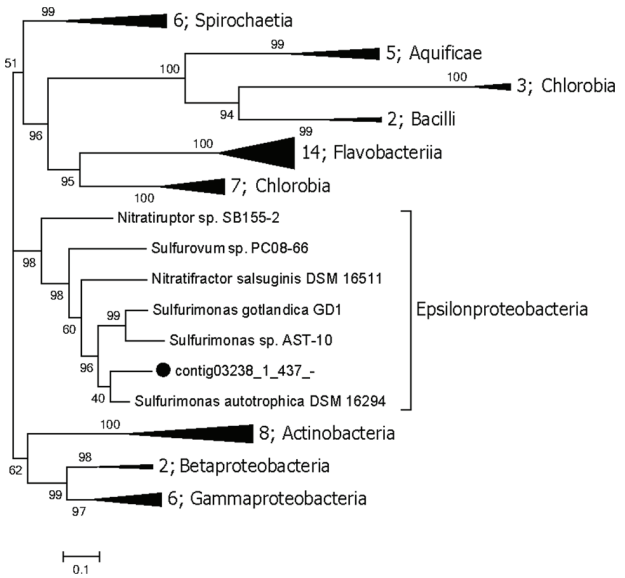

FIGURE 2 | Maximum likelihood phylogenetic tree of the *sqr* genes. The bootstrap values are based on 1000 replicates, and the percentages are shown at the nodes. The genes identified in this study are highlighted with black dots. Numbers of genomes in each collapsed clade are displayed before the clade name.

Deltaproteobacteria and Archaeoglobales, which were proposed to conduct the DSR pathway in this study, *aprAB* and *dsrA* genes were found to be highly expressed in both of these two taxonomic groups, whereas *sat* and *dsrB* genes were found only in Deltaproteobacteria. The phylogenies of *aprA* and *dsrA* further confirmed their assignment to Deltaproteobacteria (Supplementary Figures S2A,B). In a previous study, the *aprA* with the highest abundance was assigned to the genus *Desulfobulbus* (Cao et al., 2014). In our study, the *aprA* gene with the highest expression was assigned to *Desulfovibrio* (Supplementary Figure S2A). To summarize, the taxonomic assignment and expression of key genes in the sulfur cycle suggest that both the ORS and DSR pathways are highly active in this oil-immersed microbial community, and the energy generated by the sulfur metabolism supports the dominant and active group (**Figure 3**).

Because there are no metatranscriptome published for any hydrothermal vent chimneys, we compared the expression patterns of the sulfur-metabolizing genes in this metatranscriptome to those in the available metatranscriptome of a plum sample that was also collected from Guaymas Basin (Lesniewski et al., 2012). As shown in **Figure 4**, sulfur metabolizing (including oxidation and reduction) genes were among the most abundant genes found in the metatranscriptome, and a significant difference (p-value < 0.001) in the expression profiles of sulfur metabolizing genes was observed between the chimney and the plume metatranscriptome. Therefore, the sulfur-metabolizing genes were highly abundant and expressed in this GB chimney sample, and displayed significantly higher expression pattern than those of a hydrothermal vent plume sample from Guaymas Basin.

Carbon Metabolism

In this study, the complete WL pathway was identified in Archaeoglobales with high expression levels (Supplementary Table S2). The CBB cycle was not identified. The genes involved in the complete rTCA cycle were found to be actively present in both Deltaproteobacteria and Epsilonproteobacteria that dominated this chimney microbial community (**Table 3**). The key gene in the rTCA cycle, ATP-citrate lyase (*acl*), identified in this study to exhibit the highest expression was from Epsilonproteobacteria and exhibited the highest similarity to *Sulfurovum*, a novel sulfur-, nitrate-, and thiosulfate-reducing and strictly anaerobic chemolithoautotroph bacterium isolated from a deep-sea hydrothermal vent chimney at the Central Indian Ridge (Mino et al., 2014). In this study, the key enzyme for the utilization of acetate, acetyl-CoA synthetase (*acd/acs*), was found to be expressed and was assigned to sulfate-reducing bacteria (SRB; bin21 as shown in **Table 3**). In addition, the rTCA cycle and WL pathway were found to be the main pathways for carbon fixation by the dominant Bacteria and Archaea, respectively. This result suggests that, in combination with sulfur metabolism, autotrophic carbon fixation may play an important role in the survival and dominance of these species in the community. Moreover, as shown in Supplementary Table S3, genes involved in the flagellar assembly process were found to be actively present in Desulfovibrionales (bin21). The active role of the flagellar system in SRB may facilitate the movement toward electron donors and nutrients that occurs under the highly fluctuating conditions resulting from eruptions of hydrothermal vents. SRB have been reported to have the potential to anaerobically oxidize diverse hydrocarbons, such as alkanes, in Guaymas Basin sediments and chimney samples (Rueter et al., 1994). In this study, the activity and expression level of the presumably key gene in fumarate addition, a process through which alkanes are added to the double bond of fumarate based on the activity of alkylsuccinate synthase (*ass*), was checked. The *ass* genes were found to be highly active in this community, as determined through their expression level, and the most highly expressed hits were from *Desulfoglaeba alkanexedens* (Agrawal and Gieg, 2013), a typical sulfate-reducing and alkane-oxidizing bacterium (Supplementary Table S4). Moreover, the enzymes required for the degradation of a variety of organic compounds, such as hydrocarbons, fatty acids, chitins and proteins, have been detected in both the metagenome and metatranscriptome (Supplementary Table S5). Despite their important roles in carbon and global sulfur cycle, the energy metabolism of SRB remains poorly understood. After taxonomic assignment (see Materials and Methods), cyctochrome c (*cytC*), formate dehydrogenase (*fdh*), F-type ATPase (*atp*), NADH-quinone oxidoreductase (*nuo*), electron transport complex protein (*rnf*) and hydrogenases, such as Ni/Fe-hydrogenase I (*hyaAB*) and hydrogenase nickel incorporation and accessory protein (*hypA* and *hypB*), were found with expressions and assigned to SRB (Supplementary Table S6). The presence of hydrogenases and *fdh*

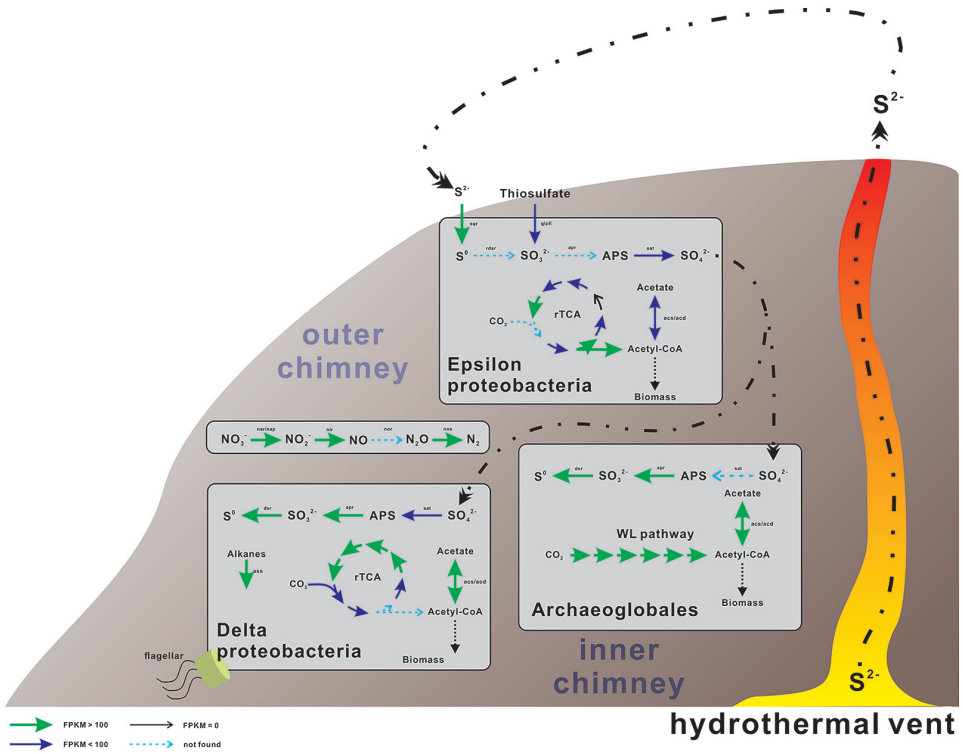

FIGURE 3 | Proposed metabolic and active pathways in this chimney community. The expression and presence of the genes are indicated by the color and thickness of the arrows, as shown in the bottom-right side bar. For genes with multiple hits, only the genes with the highest expression value (FPKM, expected fragments per kilobase of transcript per million fragments mapped) are displayed and discussed. The processes conducted by Archaea are shown on the left, whereas those conducted by bacteria are presented on the right. Detailed information of these genes is displayed in **Tables 2–4**.

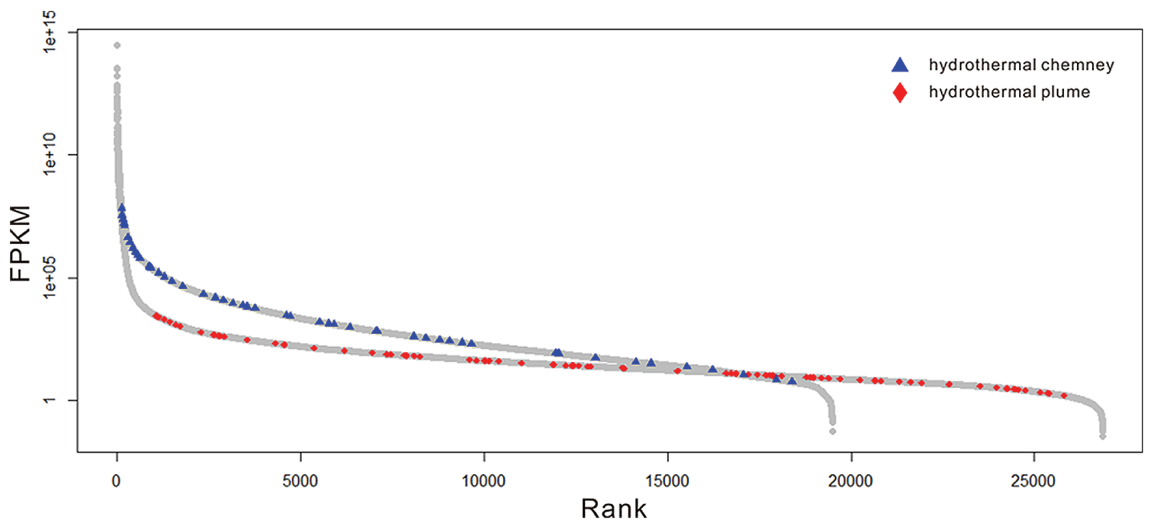

FIGURE 4 | Expression profiles of the genes in the metatranscriptome. The gray circles indicate the genes of the whole community. The blue triangles and red diamonds represent those related to sulfur oxidation and reduction in this metatranscriptome and a previous study (Lesniewski et al., 2012), respectively. The relative abundance of the gene transcripts was normalized to the length of the gene fragment and the total number of all of the transcripts.

may suggest that H_2 or formate and play important roles in the flow of electrons during sulfate reduction. As shown above, the sulfur cycle in this community was particularly intensive and closely interacted with the carbon cycle, including carbon fixation and hydrocarbon degradation, to sustain the primary production in this ecosystem.

TABLE 3 | Genes identified in the rTCA pathway in Delta- and Epsilonproteobacteria species.

Gene name	Abbrevations	Assigned taxonomy*		FPKM#
		Bin	BLAST	
Malate dehydrogenase	*mdh*	–	Bacteria	424.81
Fumarate hydratase subunit alpha	*fumA*	bin21	Desulfovibrionales	355.77
Fumarate hydratase subunit beta	*fumB*	–	Bacteria	135.08
Fumarate hydratase, class II	*fumC*	–	Bacteria	0.00
Fumarate reductase, flavoprotein subunit	*frdA*	–	Desulfovibrionales	1242.93
Fumarate reductase, iron–sulfur subunit	*frdB*	–	Epsilonproteobacteria	7.09
Succinyl-CoA synthetase	*sucC*		Bacteria	57.34
Succinyl-CoA synthetase alpha subunit	*sucD*	–	Desulfobacterales	724.00
2-Oxoglutarate ferredoxin oxidoreductase subunit alpha	*korA*	–	Deltaproteobacteria	65.56
2-Oxoglutarate ferredoxin oxidoreductase subunit beta	*korB*	–	Epsilonproteobacteria	232.91
2-Oxoglutarate ferredoxin oxidoreductase subunit delta	*korD*	–	–	–
2-Oxoglutarate ferredoxin oxidoreductase subunit gamma	*korC*	–	Bacteria	71.01
Isocitrate dehydrogenase	*icdA*	–	Bacteria	93.51
Isocitrate dehydrogenase (NAD+)	*IDH3*	–	Bacteria	0.00
2-Methylisocitrate dehydratase	*acnB*	–	Proteobacteria	20.34
Aconitate hydratase	*acnA*	bin21	Desulfovibrionales	69.50
Aconitate hydratase 2	*acnB*	–	Proteobacteria	20.34
ATP-citrate lyase alpha-subunit	*aclA*	–	Epsilonproteobacteria	238.10
ATP-citrate lyase beta-subunit	*aclB*	–	–	–
Pyruvate ferredoxin oxidoreductase alpha subunit	*porA*	–	Epsilonproteobacteria	9.90
Pyruvate ferredoxin oxidoreductase beta subunit	*porB*	–	Bacteria	59.23
Pyruvate ferredoxin oxidoreductase delta subunit	*porD*	–	Bacteria	2.66
Pyruvate ferredoxin oxidoreductase gamma subunit	*porG*	–	Epsilonproteobacteria	44.25
ADP-forming acetyl-CoA synthetase	*acd*	bin21	–	136.79
Acetate kinase	*ack*	–	*Thermotogaceae*	22.44
Phosphate acetyltransferase	*pta*	–	*Caldisericaceae*	21.14
				424.81

**The taxonomy assignments were determined by two methods, as described in Section "Materials and Methods." The binning index is explained in Supplementary Table S1. #FPKM is based on the maximal expression value of the annotated genes.*

Nitrogen Metabolism

The key genes involved in the nitrogen metabolism were found, and some of these were found to be actively expressed (**Table 4**). Many Bacteria and Archaea have the potential to perform denitrification (Philippot, 2002), and numerous organic and inorganic compounds can be used as electron donors for denitrification. The genes involved in denitrification, including *nar* (nitrate reductase), *nap* (nitrate reductase), *nir* (nitrite reductase), *nor* (nitric oxide reductase) and *nosZ*, were found to be present in the metagenome. The *narG* gene was assigned to *Beggiatoa*, a nitrate-respiring and sulfide-oxidizing bacterium that has been found to dominate microbial mats in hydrothermal sediments in the Guaymas Basin (Winkel et al., 2014). *narJ* was found to be expressed in Alteromonadales, whereas *napA* and *napB* were found to be expressed in Epsilonproteobacteria. To summarize, a complete set of denitrification genes were found in the bacterial community of the chimney, though some of them were found at low expression levels (**Table 4**). Based on this observation, we propose that nitrogen denitrification present in this community is most likely mediated by Gammaproteobacteria and Epsilonproteobacteria, with electrons generated by the ORS pathway.

DISCUSSION

Since the discovery of the deep-sea hydrothermal ecosystem in 1977, it has been proposed that hydrogen sulfide-oxidizing chemoautotrophs may potentially sustain the primary production in these ecosystems (Kvenvolden et al., 1995), where hydrogen sulfide or sulfide is primarily supplied via the high temperatures of seawater-rock interactions in the subseafloor hydrothermal reaction zones (Jannasch and Mottl, 1985). The chemical and microbial oxidation and reduction reactions of sulfur compounds probably establish the overall sulfur metabolism in the ecosystem (Yamamoto and Takai, 2011). There is no doubt that the sulfur cycle is one of the most important microbial chemosynthetic pathways in the microbial habitats of hydrothermal vents, but few studies have attempted to characterize the process, particularly at the function and activity levels. To date, the mechanism through which a microbial community in hydrothermal fields can be fueled by sulfate metabolism remains unclear. In particular, metagenomic approaches have not been widely applied in studies of energy generation by the microbial sulfur cycle in hydrothermal systems. In this study, a combined metagenomic

TABLE 4 | Genes identified in the nitrogen metabolic pathway in the microbial community.

Gene name	Abbrevations	Assigned taxonomy*		FPKM#
		Bin	BLAST	
Nitrate reductase alpha subunit	*narG*	–	Thiotrichales	49.68
Nitrate reductase beta subunit	*narH*	–	–	2360.80
Nitrate reductase gamma subunit	*narI*	–	Bacteria	380.02
Nitrate reductase delta subunit	*narJ*	–	Alteromonadales	7.17
Periplasmic nitrate reductase NapA	*napA*	–	Epsilonproteobacteria	98.48
Cytochrome c-type protein NapB	*napB*	–	Epsilonproteobacteria	4.68
Nitrite reductase (NO-forming)	*nirK*	–	–	–
Nitrite reductase (NO-forming)	*nirS*	–	–	–
Nitric oxide reductase subunit B	*norB*	–	Epsilonproteobacteria	0.00
Nitric oxide reductase subunit C	*norC*	–	–	––
Nitrous-oxide reductase	*nosZ*	–	Proteobacteria	284.50

**The taxonomy assignments were determined by two methods, as described in Section "Materials and Methods." The binning index is explained in Supplementary Table S1. #FPKM is based on the maximal expression value of the annotated genes.*

and metatranscriptomic study of a chimney in the Guaymas Basin provides insight into the complete sulfur cycle based on the results from not only the genomic but also the expression analysis, the combination of which has not been previously used for the analysis of a deep-sea hydrothermal vent chimney sample.

The accumulation of hydrogen sulfides at the outer chimney promoted the coupling of sulfide oxidation to the electron acceptors present in the nearby marine water, including oxygen and nitrate, as supported by the retrieval of the functional and expressed genes described herein (**Tables 2–4** and **Figure 3**). These findings suggest that the coupling between sulfur oxidation and denitrification may fuel some N-metabolizing microorganisms at the sulfide-enriched outer chimney. As proposed in this study, the microorganisms involved in this process were Epsilonproteobacteria as the sulfur-oxidizing bacteria, and Gammaproteobacteria and Epsilonproteobacteria as potential denitrifiers. The other sulfur-metabolizing group, namely sulfate-reducing prokaryotes, may use hydrogen and/or dissolved organic matter as electron donors, as hydrogenases and key genes for the degradation of organic compounds have been identified in this study (Supplementary Tables S5 and S6).

Carbon fixation pathways other than the Calvin–Benson–Bassham (CBB) cycle have been found to exhibit a notable contribution to carbon fixation, mostly at deep-sea hydrothermal vents (Campbell and Cary, 2004). The rTCA cycle was found to be highly expressed in the dominant Delta- and Epsilonproteobacteria. The key enzyme for the utilization of acetate was also identified to be expressed in this study (**Table 3**). Generally, the rTCA cycle appears to be dominant in habitats with a temperature ranging from 20 to 90°C, whereas the CBB cycle and the Wood-Ljungdahl (WL) pathway may be the principal pathways at temperatures lower than 20°C and greater than 90°C, respectively (Hugler and Sievert, 2011). In the present sample, the CBB cycle was not found present, which is consistent with the fact that this sample was collected from a high-temperature condition (He et al., 2013). In addition, the enzymes for the degradation of a variety of organic compounds, such as hydrocarbons, fatty acids, chitins and proteins, have been detected at both DNA and RNA level (Supplementary Table S5). Together, all of these organic compounds may be the carbon source for this microbial community.

In this scenario, both autotrophic and heterotrophic SRB could inhabit the inner chimney (**Figure 3**), where sulfate reduction is coupled to carbon fixation and hydrocarbon oxidation. Based on the expression levels of key genes in rTCA (**Table 3**) and alkane degradation (Supplementary Table S4), hydrocarbon degradation might contribute substantially to the linking of S and C cycle at inner layer chimney. In another word, heterotrophic SRB, commonly found at vent systems, may be the major player in coordinating and influencing the S and C cycle. Compared the expression of key genes in sulfur metabolizing and the rest processes (**Figure 4**), the reduced sulfur would be quickly and intensively oxidized to fuel the community, where sulfate-reducing microbes were found dominated. The composition of the sulfate-reducing community was determined by the way that microbes perform carbon metabolism. In our sample, heterotrophic SRB was found prevalent with their capabilities in hydrocarbon degradation. This finding may improve our understanding on the structure, function, and interaction within microbial community in hydrothermal vent.

Meta-omics based approaches have the advantages in studying the entire microbial community without pure cultures or prior knowledge on the sample. Functional omics approaches, such as transcriptome and proteome, could further confirm the metabolic potential at the active level. More efforts will be spent on quantification and comparison of these function omics datasets. Together with *in situ* carbon stable isotope measurement, and lipid type and diversity analysis, the activity, rate and interaction of key process in a given environmental condition could be accessed and estimated.

MATERIALS AND METHODS

Sample Collection and Processing

The sample 4558-6 under investigation was collected from the outer layer of a black-smoker chimney in the Guaymas Basin and was previously described through a metagenome-based study (He et al., 2013). The sample was fixed with RNAlater (Sigma-Aldrich, Munich, Germany) and stored at −80°C prior to DNA and RNA extraction. DNA isolation was conducted as described previously (Wang et al., 2013). Metagenome pyrosequencing was performed using a 454 Life Sciences GS FLX system with a practical limit of 400 bp. RNA was isolated with a RNA isolation kit (Omega Bio-Tek, Doraville, GA, USA) following the user's manual provided by the manufacturer. RNA samples were treated with DNAse (Thermo) for 45 min at 37°C, and then used as a template for PCR to detect undigested DNA. The mRNA fraction was enriched through the enzymatic digestion of rRNA molecules (mRNA-ONLY Prokaryotic mRNA Isolation kit, Epicentre Biotechnologies, Madison, WI, USA) followed by the subtractive hybridization of rRNA with capture oligonucleotides (Ambion MICROBExpress kit, Life Technologies, Gaithersburg, MD, USA). The mRNA isolates were first amplified (MessageAmp II-Bacteria kit, Ambion, Life Technologies) and then reversely transcribed into complementary DNA. Afterward, the cDNA was directly sequenced using the Illumina (BGI-Shenzhen, China) Hiseq2000 platform (2*90 bp pair-end) for metatranscriptome analysis.

Metagenome Assembly and Annotation

The reads obtained through metagenome sequencing were assembled and annotated as previously described (He et al., 2013). Briefly, low quality sequencing reads were trimmed in Geneious 6.04 (Biomatters Ltd.) and technical replicates were removed with cd-hit (at 96% sequence identity; Fu et al., 2012). After removing short reads (<100 bp), the remaining reads were assembled with Velvet (Zerbino and Birney, 2008). Coding regions of the metagenomic assembly were predicted using FragGeneScan (Rho et al., 2010) and then BLASTed (Altschul et al., 1997; $1e^{-5}$) against an NCBI non-redundant (NR) protein database. The 16S rRNA genes were picked using Sortmerna and BLASTed against GreenGene database (e-value < $1e^{-5}$) respectively. For functional annotation, sequences with matches to the COG (Tatusov et al., 2003), Pfam (Finn et al., 2014), and KEGG (Ogata et al., 1999) databases were retrieved to establish the functional categories and reconstruct the metabolic pathways. The genes of interest, such as transposases, were subjected to manual checkup, and spurious annotations (putative, like-, similar to) were excluded from further analysis.

Taxonomic Assignment

Two different methods were applied to assess the taxonomic information. First, the assembled metagenomic sequences was binned using the tetranucleotide frequencies in emergent self-organizing maps (ESOMs; Dick et al., 2009) with a window size of 8 kbp, a sliding window size of 4 kbp, and the minimum fragment size of 2 kbp. Complete genomic sequences of 20 species were used as references (designated as bin1–20), these microorganism were listed as following: *Acinetobacter pittii* ANC 4052, *Alteromonas macleodii* str. 'Deep ecotype,' *Candidatus Pelagibacter ubique* HTCC1062, uncultured marine crenarchaeote E37-7F, Marine group II euryarchaeote SCGC AAA288-C18, Marine Group II euryarchaeote SCGC AB-629-J06, uncultured marine group II euryarchaeote (marine metagenome), Marine Group III euryarchaeote SCGC AAA007-O11, Marine Group III euryarchaeote SCGC AAA288-E19, *Marinobacter nanhaiticus* D15-8W, *Methylobacter tundripaludum* SV96, *Methylophaga aminisulfidivorans* MP, *Methylotenera mobilis* JLW8, *Nitrosopumilus maritimus* SCM1, *Candidatus Nitrospira defluvii*, *Planctopirus limnophila* DSM 3776, *Pseudomonas denitrificans* ATCC 13867, *Candidatus Ruthia magnifica* str. Cm (Calyptogena magnifica), SAR324 cluster bacterium SCGC AAA240-J09 and SAR86 cluster bacterium SAR86E. After binning, the completeness and taxonomic classification of the genomes within bins were then estimated by counting and BLASTing universal single-copy genes as previously described (Rinke et al., 2013). Alternatively, each predicted sequence feature in the metagenome and metatranscriptome was assigned to a certain taxon if at least 75% of the BLAST hits of this query were from that specific taxon. A BLAST search of all of the reads against the non-redundant protein database in NR was performed. All of the hits obtained from the BLAST searches were retained, and their taxonomic affiliations were determined using MEGAN (Huson et al., 2007) with bit-score values of 100. The taxonomic compositions of each predicted gene feature was then visualized using MEGAN.

Metatranscriptome Mapping and Transcript Quantification

The raw shotgun sequencing metatranscriptomic reads obtained by Illumina pair-end sequencing were dereplicated (100% identity over 100% lengths) and trimmed using sickle[1]. The dereplicated, trimmed, and paired-end Illumina reads were then mapped to the metagenome using Bowtie (Langmead and Salzberg, 2012) with the default parameters. The unique mapped reads were selected, and FPKM (expected fragments per kilobase of transcript per million fragments mapped) was used to estimate the expression level of each gene using a script downloaded from GitHub[2].

Estimation of the Completeness of Genomic Bins

The complete genome sizes of the genomic bins were estimated based on an analysis of conserved single-copy genes (CSCGs) as described by Lloyd et al. (2013). In total, we were able to collect 162 and 139 universal CSCGs for the archaea and bacteria genomes, as in the previous study (Rinke et al., 2013). The ratios between the numbers of CSCGs present in the metagenome and the number of total CSCGs were then used to estimate the size of each genome bin.

[1]https://github.com/najoshi/sickle

[2]https://github.com/minillinim/sam2FPKG

Comparative Analysis

The expression patterns of the sulfur-metabolizing genes in this metatranscriptome were compared to those in the metatranscriptome of a plum sample from Guaymas Basin (Lesniewski et al., 2012). Comparisons between two metatranscriptomes were conducted using the Mann–Whitney *U*-test. The gene expression profiles were compared between two samples using the normalized rank from 0 to 1 in each respective sample as the input. A difference was considered significant if the *p*-value was lower than 0.001.

Construction of a Phylogenetic Tree

The predicted sequence features were checked across multiple annotation databases and then aligned with ClustalW (Larkin et al., 2007), and any gaps were removed manually. To construct functional gene phylogenies, the aligned sequences were analyzed by maximum likelihood-based FastTree (Price et al., 2010) using the Jones–Taylor–Thornton (JTT) with CAT approximation.

Metabolic Pathway Identification

The gene products were searched for similarity against the KEGG database. A match was counted if the similarity search resulted in an expectation e-value below $1e^{-5}$. All of the occurring KO (KEGG Orthology) numbers were mapped against the KEGG pathway functional hierarchies and the COG database. For genes with multiple hits, only the genes with the highest expression value (FPKM) are displayed in the figures and tables and further discussed in the text.

Data Availability

The metatranscriptome sequences are available on NCBI as SRX1008212. The assembled sequence was uploaded to IMG with a project ID Ga0072503.

ACKNOWLEDGMENTS

We thank Anna-Louise Reysenbach for providing the chance to attend the expedition, and all the crew members from AT-26 cruise. This work was supported by National High Technology Research and Development Program of China (Grant No. 2012AA092103), China Ocean Mineral Resources R & D Association (Grant No.DY125-22-04 and DY125-15-T-04).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01236

REFERENCES

Agrawal, A., and Gieg, L. M. (2013). In situ detection of anaerobic alkane metabolites in subsurface environments. *Front. Microbiol.* 4:140. doi: 10.3389/fmicb.2013.00140

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Anantharaman, K., Breier, J. A., Sheik, C. S., and Dick, G. J. (2013). Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 110, 330–335. doi: 10.1073/pnas.1215340110

Baumberger, T., Lilley, M. D., Resing, J. A., Lupton, J. E., Baker, E. T., Butterfield, D. A., et al. (2014). Understanding a submarine eruption through time series hydrothermal plume sampling of dissolved and particulate constituents: West Mata, 2008–2012. *Geochem. Geophys. Geosyst.* 15, 4631–4650. doi: 10.1002/2014GC005460

Bergmann, F. D., Selesi, D., and Meckenstock, R. U. (2011). Identification of new enzymes potentially involved in anaerobic naphthalene degradation by the sulfate-reducing enrichment culture N47. *Arch. Microbiol.* 193, 241–250. doi: 10.1007/s00203-010-0667-4

Biddle, J. F., Cardman, Z., Mendlovitz, H., Albert, D. B., Lloyd, K. G., Boetius, A., et al. (2012). Anaerobic oxidation of methane at different temperature regimes in Guaymas Basin hydrothermal sediments. *ISME J.* 6, 1018–1031. doi: 10.1038/ismej.2011.164

Brazelton, W. J., and Baross, J. A. (2010). Metagenomic comparison of two *Thiomicrospira* lineages inhabiting contrasting deep-sea hydrothermal environments. *PLoS ONE* 5:e13530. doi: 10.1371/journal.pone.0013530

Campbell, B. J., and Cary, S. C. (2004). Abundance of reverse tricarboxylic acid cycle genes in free-living microorganisms at deep-sea hydrothermal vents. *Appl. Environ. Microbiol.* 70, 6282–6289. doi: 10.1128/AEM.70.10.6282-6289.2004

Cao, H., Wang, Y., Lee, O. O., Zeng, X., Shao, Z., and Qian, P. Y. (2014). Microbial sulfur cycle in two hydrothermal chimneys on the Southwest Indian Ridge. *MBio* 5, e980-13. doi: 10.1128/mBio.00980-13

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10, R85. doi: 10.1186/gb-2009-10-8-r85

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Fornari, D. J., Shank, T. M., Von Damm, K. L., Gregg, T. K. P., Lilley, M. D., Levai, G., et al. (1998). Time-series temperature measurements at high-temperature hydrothermal vents, East Pacific Rise 9 °49'-51'N: monitoring a crustal cracking event. *Earth Planet. Sci. Lett.* 160, 419–430. doi: 10.1016/S0012-821X(98)00101-0

Fritz, G., Roth, A., Schiffer, A., Buchert, T., Bourenkov, G., Bartunik, H. D., et al. (2002). Structure of adenylylsulfate reductase from the hyperthermophilic *Archaeoglobus fulgidus* at 1.6-A resolution. *Proc. Natl. Acad. Sci. U.S.A.* 99, 1836–1841. doi: 10.1073/pnas.042664399

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

He, Y., Xiao, X., and Wang, F. (2013). Metagenome reveals potential microbial degradation of hydrocarbon coupled with sulfate reduction in an oil-immersed chimney from Guaymas Basin. *Front. Microbiol.* 4:148. doi: 10.3389/fmicb.2013.00148

Hugler, M., and Sievert, S. M. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Ann. Rev. Mar. Sci.* 3, 261–289. doi: 10.1146/annurev-marine-120709-142712

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Jannasch, H. W., and Mottl, M. J. (1985). Geomicrobiology of deep-sea hydrothermal vents. *Science* 229, 717–725. doi: 10.1126/science.229.4715.717

Kvenvolden, K. A., Hostettler, F. D., Carlson, P. R., Rapp, J. B., Threlkeld, C. N., and Warden, A. (1995). Ubiquitous tar balls with a california-source signature on the shorelines of prince william sound, alaska. *Environ. Sci. Technol.* 29, 2684–2694. doi: 10.1021/es00010a033

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lesniewski, R. A., Jain, S., Anantharaman, K., Schloss, P. D., and Dick, G. J. (2012). The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J.* 6, 2257–2268. doi: 10.1038/ismej.2012.63

Lloyd, K. G., Schreiber, L., Petersen, D. G., Kjeldsen, K. U., Lever, M. A., Steen, A. D., et al. (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* 496, 215–218. doi: 10.1038/nature12033

Mino, S., Kudo, H., Arai, T., Sawabe, T., Takai, K., and Nakagawa, S. (2014). *Sulfurovum aggregans* sp. nov., a hydrogen-oxidizing, thiosulfate-reducing chemolithoautotroph within the Epsilonproteobacteria isolated from a deep-sea hydrothermal vent chimney, and an emended description of the genus *Sulfurovum*. *Int. J. Syst. Evol. Microbiol.* 64, 3195–3201. doi: 10.1099/ijs.0.065094-0

Nakagawa, S., Takai, K., Inagaki, F., Hirayama, H., Nunoura, T., Horikoshi, K., et al. (2005). Distribution, phylogenetic diversity and physiological characteristics of epsilon-*Proteobacteria* in a deep-sea hydrothermal field. *Environ. Microbiol.* 7, 1619–1632. doi: 10.1111/j.1462-2920.2005.00856.x

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34. doi: 10.1093/nar/27.1.29

Orcutt, B. N., Sylvan, J. B., Knab, N. J., and Edwards, K. J. (2011). Microbial ecology of the dark ocean above, at, and below the seafloor. *Microbiol. Mol. Biol. Rev.* 75, 361–422. doi: 10.1128/MMBR.00039-10

Philippot, L. (2002). Denitrifying genes in bacterial and Archaeal genomes. *Biochim. Biophys. Acta* 1577, 355–376. doi: 10.1016/S0167-4781(02)00420-7

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Reysenbach, A. L., and Shock, E. (2002). Merging genomes with geochemistry in hydrothermal ecosystems. *Science* 296, 1077–1082. doi: 10.1126/science.1072483

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38, e191. doi: 10.1093/nar/gkq747

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi: 10.1038/nature12352

Rueter, P., Rabus, R., Wilkes, H., Aeckersberg, F., Rainey, F. A., Jannasch, H. W., et al. (1994). Anaerobic oxidation of hydrocarbons in crude oil by new types of sulphate-reducing bacteria. *Nature* 372, 455–458. doi: 10.1038/372455a0

Sylvan, J. B., Sia, T. Y., Haddad, A. G., Briscoe, L. J., Toner, B. M., Girguis, P. R., et al. (2013). Low temperature geomicrobiology follows host rock composition along a geochemical gradient in lau basin. *Front. Microbiol.* 4:61. doi: 10.3389/fmicb.2013.00061

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41. doi: 10.1186/1471-2105-4-41

Teske, A., Callaghan, A. V., and Larowe, D. E. (2014). Biosphere frontiers of subsurface life in the sedimented hydrothermal system of Guaymas Basin. *Front. Microbiol.* 5:362. doi: 10.3389/fmicb.2014.00362

Von Damm, K. L. (1990). Seafloor hydrothermal activity: black smoker chemistry and chimneys. *Annu. Rev. Earth Planet. Sci.* 18, 173–204. doi: 10.1111/gbi.12086

Wang, J., Shen, J., Wu, Y., Tu, C., Soininen, J., Stegen, J. C., et al. (2013). Phylogenetic beta diversity in bacterial assemblages across ecosystems: deterministic versus stochastic processes. *ISME J.* 7, 1310–1321. doi: 10.1038/ismej.2013.30

Winkel, M., De Beer, D., Lavik, G., Peplies, J., and Mussmann, M. (2014). Close association of active nitrifiers with *Beggiatoa* mats covering deep-sea hydrothermal sediments. *Environ. Microbiol.* 16, 1612–1626. doi: 10.1111/1462-2920.12316

Yamamoto, M., and Takai, K. (2011). Sulfur metabolisms in epsilon– and gamma-*Proteobacteria* in deep-sea hydrothermal fields. *Front. Microbiol.* 2:192. doi: 10.3389/fmicb.2011.00192

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 November 2015
doi: 10.3389/fmicb.2015.01270

Degradation Network Reconstruction in Uric Acid and Ammonium Amendments in Oil-Degrading Marine Microcosms Guided by Metagenomic Data

Rafael Bargiela[1], Christoph Gertler[2†], Mirko Magagnini[3], Francesca Mapelli[4], Jianwei Chen[5], Daniele Daffonchio[4,6], Peter N. Golyshin[2] and Manuel Ferrer[1]**

[1] Systems Biotechnology, Department of Biocatalysis, Institute of Catalysis, Consejo Superior de Investigaciones Científicas, Madrid, Spain, [2] School of Biological Sciences, Bangor University, Bangor, UK, [3] EcoTechSystems Ltd., Ancona, Italy, [4] Department of Food, Environmental and Nutritional Sciences, University of Milan, Milan, Italy, [5] Beijing Genomics Institute, Shenzhen, China, [6] Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Biostimulation with different nitrogen sources is often regarded as a strategy of choice in combating oil spills in marine environments. Such environments are typically depleted in nitrogen, therefore limiting the balanced microbial utilization of carbon-rich petroleum constituents. It is fundamental, yet only scarcely accounted for, to analyze the catabolic consequences of application of biostimulants. Here, we examined such alterations in enrichment microcosms using sediments from chronically crude oil-contaminated marine sediment at Ancona harbor (Italy) amended with natural fertilizer, uric acid (UA), or ammonium (AMM). We applied the web-based AromaDeg resource using as query Illumina HiSeq meta-sequences (UA: 27,893 open reading frames; AMM: 32,180) to identify potential catabolic differences. A total of 45 (for UA) and 65 (AMM) gene sequences encoding key catabolic enzymes matched AromaDeg, and their participation in aromatic degradation reactions could be unambiguously suggested. Genomic signatures for the degradation of aromatics such as 2-chlorobenzoate, indole-3-acetate, biphenyl, gentisate, quinoline and phenanthrene were common for both microcosms. However, those for the degradation of orcinol, ibuprofen, phenylpropionate, homoprotocatechuate and benzene (in UA) and 4-aminobenzene-sulfonate, *p*-cumate, dibenzofuran and phthalate (in AMM), were selectively enriched. Experimental validation was conducted and good agreement with predictions was observed. This suggests certain discrepancies in action of these biostimulants on the genomic content of the initial microbial community for the catabolism of petroleum constituents or aromatics pollutants. In both cases, the emerging microbial communities were phylogenetically highly similar and were composed by very same proteobacterial families. However, examination of taxonomic assignments further revealed different catabolic pathway organization at the organismal level, which should be considered for designing oil spill mitigation strategies in the sea.

Keywords: ammonium, biostimulation, crude oil degradation, enrichment, Mediterranean Sea, metagenomics, microcosm, uric acid

OPEN ACCESS

Edited by:
Eamonn P. Culligan, Cork Institute of Technology, Ireland

Reviewed by:
Romy Chakraborty, Lawrence Berkeley National Lab, USA
Efthymios Ladoukakis, National Technical University of Athens, Greece

****Correspondence:***
Manuel Ferrer
mferrer@icp.csic.es;
Peter N. Golyshin
p.golyshin@bangor.ac.uk

†Present address:
Christoph Gertler, Friedrich Loeffler Institute– Federal Research Institute for Animal Health, Institute for Novel and Emerging Infectious Diseases, 17493 Greifswald, Germany

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *06 July 2015*
Accepted: *30 October 2015*
Published: *24 November 2015*

Citation:
Bargiela R, Gertler C, Magagnini M, Mapelli F, Chen J, Daffonchio D, Golyshin PN and Ferrer M (2015) Degradation Network Reconstruction in Uric Acid and Ammonium Amendments in Oil-Degrading Marine Microcosms Guided by Metagenomic Data. Front. Microbiol. 6:1270. doi: 10.3389/fmicb.2015.01270

INTRODUCTION

Oil pollution still is a global problem (Yakimov et al., 2007; Bargiela et al., 2015). At present, in many sea regions containment and recovery of oil using booms and skimmers is the method of choice for oil spill first responders (Walther, 2014). Especially in the open sea, the use of dispersants in combination with biostimulation and bioaugmentation agents based on non-toxic, natural low cost formulations, is encouraged, although the majority of tests have been performed at lab-scale (Das and Chandran, 2010; Nikolopoulou and Kalogerakis, 2010; Alvarez et al., 2011; Nikolopoulou et al., 2013). In marine systems, the low concentration of nitrogen, phosphorous, and oxygen, together with their low bioavailability are main factors limiting the degradation of carbon-rich hydrophobic compounds (Howarth and Marino, 2006; Venosa et al., 2010; Ly et al., 2014). Attempts have been made to use different nitrogen sources to promote the growth and selection of different microbial strains with greater catabolic capacity for combating oil spills compared to natural attenuation (Teramoto et al., 2009; Venosa et al., 2010). However, crude oil biodegradation requires about 0.04 g of nitrogen per gram of oil (Atlas, 1981) which makes the choice of nitrogen source pivotal for the whole treatment. Recent data highlighted the possible link between N cycling processes and hydrocarbon degradation in marine sediments (Scott et al., 2014). Therefore, it is essential to select appropriated N-containing biostimulants.

The sources of nitrogen for the degradation tests – mostly performed at lab-scale and in minor occasions at field-scale – included nitrate, ammonium (AMM), urea, uric acid (UA), amino acids and the hydrophobic substance lecithin (Garcia-Blanco et al., 2007; Li et al., 2007; Martínez-Pascual et al., 2010; Venosa et al., 2010; Nikolopoulou et al., 2013; Mohseni-Bandpi et al., 2014). Slow-release nitrogen (AMM-based) fertilizers have also been successfully used for growth stimulation in microbial oil remediation (Miyasaka et al., 2006; Teramoto et al., 2009; Reis et al., 2013). However, AMM has been proved ineffective in treatment of real oil spill due to co-precipitation with phosphates in seawater. In a recent study, we have shown that biodegradable natural fertilizers like UA can be used as cost-efficient biostimulant for enhancing bacterial growth in polluted sediments (Gertler et al., 2015). Each nitrogen source has its advantages and disadvantages, yet overall results have shown that the microbial populations were initially different from those found in the absence of biostimulants and that the degradation efficiency generally increased. It is therefore critical to establish how the whole microbial biodegradation network is affected and whether different pollutants are preferentially degraded as a consequence of amendments of biostimulants.

In an early work using the recently developed AromaDeg analysis (Duarte et al., 2014) and a meta-network graphical approach, we reconstructed the catabolic networks associated to microbial communities in a number of chronically polluted sites (Bargiela et al., 2015). The approach focuses on the usage of metagenomic data, which directly leads to a network that included catabolic reactions associated to genes encoding enzymes annotated in the genomes of the community organisms. We found key catabolic variations associated to changes in community structure and environmental constraints (Bargiela et al., 2015). In this work, this approach was applied to draft the catabolic networks of two different enrichment microcosms set up with sediments from chronically crude oil-contaminated marine sediments from Ancona harbor (Italy) and the natural fertilizer UA or AMM as nitrogen sources (Gertler et al., 2015). Ancona harbor is very close to the urban area and hosts a multi-purpose port receiving cruise liners, passenger ferries, commercial liners and fishing boats. A minor part of the related airborne pollutants is due to the vessels calling at the port while the main contribution comes from road traffic and other human activities. Furthermore, sediments in Ancona harbor are heavily contaminated due to its role as a major ferry terminal and industrial port on the Adriatic Sea. We hypothesize that the microbial community shifts previously observed after addition of UA and AMM (Gertler et al., 2015) may have an influence in the selection of certain catabolic pathways. Potential protein-coding genes ($\geq$20 amino acids long) obtained by direct Illumina HiSeq sequencing of DNA material of the corresponding microcosms (Gertler et al., 2015) constituted the input information in our study.

MATERIALS AND METHODS

Study Site, Microcosm Set-up and Sequence Accession Numbers

The starting point of this study were the meta-sequences previously obtained by direct sequencing from two microcosm sets created using sediment samples from the harbor of Ancona (Italy; 43°37′N, 13°30′15″E), as described previously (Gertler et al., 2015). Both microcosm setups were identical in size, composition, incubation, sampling regime and nutrient concentration with exception of the type of nitrogen source applied. Either AMM or UA were supplied in equimolar amounts of nitrogen. Briefly, one-liter Erlenmeyer flasks (duplicates) were filled with 150 g of sand (Sigma–Aldrich, St. Louis, MO, USA), sterilized and spiked with 10 mL of sterile filtered Arabian light crude oil. One gram of sediment from the sampling site was mixed into the oil-spiked sand as the inoculum. Three hundred milliliters of modified ONR7a medium (Dyksterhouse et al., 1995) (omitting AMM chloride and disodium hydrogen phosphate) was added. We added 5 mL of Arabian light crude oil, which based upon average literature values for density and molecular weight equals about 300 mM of C (Wang et al., 2003), 5 mM of NH_4Cl and 0.5 mM of Na_2HPO_4 resulting in a molar N/P ratio of approximately 10:1. For UA treatment microcosm, 0.21 g (1.25 mmol = 5 mmol N) of UA was provided as nitrogen source while the AMM treatment microcosms were each supplied with 2.5 mL of a 2 M AMM chloride solution (5 mmol; pH 7.8). Both treatments also contained 2.5 mL of a 0.2 M disodium hydrogen phosphate solution (0.5 mmol; pH 7.8). Excess amounts of crude oil were added to compensate for the 35% carbon losses due to evaporation of volatile hydrocarbons over the course of the experiment. Including losses due to evaporation, the C/N/P ratio was approximately 400:10:1. Control treatments

were set up: (i) a negative control contained only sterile sand and ONR7a; (ii) two further controls contained sand, ONR7a, crude oil and either UA or AMM chloride solution but no sediment sample; and (iii) one control contained oil, sterile sand, ONR7a medium and a sediment sample, but no additional nitrogen source or phosphorus source was provided. No significant growth was detected under tested control conditions. Under the given assay conditions, the utilization of UA as carbon source is minimal, as the amount of carbon introduced by UA into the microcosms was disproportionately low in contrast to the residual carbon in the sediment and the carbon introduced in form of oil. Briefly, we added 300 mmols of carbon in form of oil and only 6.25 mmols of carbon in form of UA. In addition, the molar ratio C/N in the system (between 10:1 and 40:1, depending UA or AMM was added) implies there was excess of carbon in the medium and thus the growth was limited by N.

The resulting microbial communities from microcosms were destructively sampled after 21 days of incubation at 20°C, the isolated DNA subjected to the paired-end sequencing (Illumina HiSeq 2000) at Beijing Genomics Institute (BGI; China), and gene calling performed as described (Gertler et al., 2015). Taxonomic affiliations of potential protein-coding genes were predicted as described previously (Guazzaroni et al., 2013; Bargiela et al., 2015).

The meta-sequences are available at the National Center for Biotechnology Information (NCBI) with the IDs PRJNA222664 [for MGS-ANC(UA)] and PRJNA222663 [for MGS-ANC(AMM)]. The Whole Genome Shotgun projects are also available at DDBJ/EMBL/GenBank under the accession numbers AZIH00000000 [for MGS-ANC(UA)] and AZIK00000000 [for MGS-ANC(AMM)]. All original non-chimeric 16S small subunit rRNA hypervariable tag 454 sequences were archived at the EBI European Read Archive under accession number PRJEB5322. Note that the samples were named based on the code 'MGS', which refers to MetaGenome Source, followed by a short name indicating the origin of the sample and the nitrogen source, as follows: MGS-ANC(AMM) (the harbor of Ancona and AMM as nitrogen source); MGS-ANC(UA) (the harbor of Ancona and UA as nitrogen source).

Biodegradation Network Reconstruction: Scripts and Commands for Graphics

The web-based AromaDeg resource (Duarte et al., 2014) was used for catabolic network reconstruction. AromaDeg is a web-based resource with an up-to-date and manually curated database that includes an associated query system which exploits phylogenomic analysis of the degradation of aromatic compounds. This database addresses systematic errors produced by standard methods of protein function prediction by improving the accuracy of functional classification of key genes, particularly those encoding proteins of aromatic compounds' degradation. In brief, each query sequence from a genome or metagenome [MGS-ANC(AMM) and MGS-ANC(UA), in this study] that matches a given protein family of AromaDeg is associated with an experimentally validated catabolic enzyme performing an aromatic compound degradation reaction. Individual reactions, and thus the corresponding substrate pollutants and intermediate degradation products, can be linked to reconstruct catabolic networks. We have recently designed an in-house script allowing the automatic reconstruction of such networks in a graphical format, which was used in present work. The script allows visualization and comparison of the abundance levels of genes encoding catabolic enzymes assigned to distinct degradation reactions as well as substrates or intermediates possibly degraded by distinct microbial communities. The complete workflow, including the scripts and commands used for catabolic network reconstruction has recently been reported (Bargiela et al., 2015).

Note that the sequence material used in the present investigation for biodegradation network reconstruction was based upon single biological microcosm replicate to preserve maximum coverage and sequencing depth as well as for other technical reasons, as described previously (Gertler et al., 2015). For each of the metagenome datasets the rarefaction curves of the observed species were estimated to analyze the species sampling coverage, and found that the rarefaction curves indicate closeness to saturation in each of the samples (Gertler et al., 2015). Therefore, with a single run of paired-end Illumina sequencing we determined populations that really represent the actual state of the microbial community in the microcosms and that biases were not introduced due to differences in microbial coverage. Whether or not more replicates may introduce some differences in the present study was not examined. However, because of the low standard deviation in the cultures (also checked for the representativeness of the microcosm by 16S small subunit rRNA hypervariable tag 454 sequences fingerprinting; Gertler et al., 2015) and the fact that sampled 16S rRNA diversity indicated closeness to saturation, we considered that the presented data are valid. Note that experimental validations (see Experimental Validations of Predicted Biodegradation Capacities) were performed in triplicates (with appropriated standard deviations), on the basis of which metagenome-based predictions were confirmed. Therefore, we considered that the differences at the taxonomic, gene content levels and catabolic capacities herein presented are most likely due to actual biological variability and are not random.

Experimental Validations of Predicted Biodegradation Capacities

The ability of each of the microcosms to grow on pollutants expected to be degraded, was confirmed as follows. First, UA and AMM microcosms (in triplicates) were obtained as described above but omitting Arabian light crude oil; instead, a mix of pollutants containing naphthalene, 2,3-dihydroxybiphenyl, benzene, *p*-cumate, orcinol, 2-chlorobenzoate, phthalate and phenylpropionate, all from Fluka-Aldrich-Sigma Chemical Co. (St. Louis, MO, USA), was added at a final concentration of 2 ppm each. These pollutants were selected on the basis of existing analytical methods to quantify their concentrations (Bargiela et al., 2015). Control cultures without the addition of sediments but with chemicals and cultures plus sediments but without the addition of chemicals were set up.

The extent of degradation in test and control samples was quantified as follows. Briefly, bacterial cells (from 300 ml culture)

were separated by centrifugation at 13,000 *g* at room temperature for 10 min. After supernatant separation, bacterial pellet was used for methanol extraction by adding 1.2 mL of cold (−80°C) high-performance liquid chromatography (HPLC)-grade methanol. The samples were then vortex-mixed (for 10 s) and sonicated for 30 s (in a Sonicator® 3000; Misonix) at 15 W in an ice cooler (−20°C). This protocol was repeated twice more with a 5-min storage at −20°C between each cycle, and the final pellet was removed following centrifugation at 12,000 *g* for 10 min at 4°C. Methanol solution was stored at −80°C in 20-mL penicillin vials until they were analyzed by mass spectrometry and different and complementary separation techniques, namely liquid chromatography electrospray ionization quadrupole time-of-flight mass spectrometry (LC-ESI-QTOF-MS) in positive and negative mode, and gas chromatography-mass spectrometry (GC-MS), as described previously (Bargiela et al., 2015). The abundance levels of mass signatures of tested pollutants and key degradation intermediates, namely, salicylate, gentisate, catechol, benzoate and protocatechuate, were used as indicator of the presence of the corresponding enzymes encoded by catabolic genes.

RESULTS AND DISCUSSION

Bacterial Community Structures in Microcosms

A graphical approach recently described (Bargiela et al., 2015) was applied to draft the catabolic networks of two different oil-degrading marine microcosms. They were obtained from Ancona harbor sediments which were applied in a series of two enrichment microcosms, where AMM or UA were supplied to introduce equivalent amounts of nitrogen. Using partial 16S rRNA gene sequences obtained in the non-assembled Illumina reads through a metagenomic approach, it was firstly found a relatively high degree of similarity in the emerging communities (Gertler et al., 2015). Proteobacteria were the most abundant (AMM: 74.5%; UA: 74.2%, total sequences), in agreement with the fact that this bacterial group is the most abundant in other chronically crude oil-contaminated marine sediments within the Mediterranean Sea (Bargiela et al., 2015). Noticeably, all proteobacterial families were found in both microcosms (for details see **Table 1**). However, differences in the abundance of some community members could be observed on the basis of corresponding read frequency. As an example, the percentage of members of the *Rhodobacteraceae* and *Enterobacteriaceae* was elevated in microcosms supplied with AMM (18.2% AMM vs. 0.8% in UA and 5.6% in AMM vs. 3.2% in UA, correspondingly). Conversely, lower percentages of members of the *Alteromonadaceae* (9.6%/19.2%), *Halomonadaceae* (5.6%/7.8%), *Moraxellaceae* (0.5%/7.9%) and *Flavobacteriaceae* (1.8%/5.7%) could be detected in the AMM-supplemented microcosm in comparison to UA-based microcosms. At a genus level, 55 out of 57 identified proteobacterial taxa were common in both communities. However, enrichments containing AMM were characterized by higher percentages (referred to total reads) of Alphaproteobacteria, such as *Roseovarius* sp. (1.4% in AMM vs. 0.1% in UA), *Ruegeria* spp. (1.1%/0.1%) and *Sulfitobacter* sp. (1.5%/0.1%), and some Gammaproteobacteria such as *Vibrio* sp. (2.4%/1.1%). In stark contrast to this, the UA-based enrichments showed significantly elevated percentages of members of the Firmicutes (7.4% in UA enrichments/4.9% in AMM enrichments) and Gammaproteobacteria, such as *Aeromonas* spp. (1.5%/0.5%) and *Pseudoalteromonas* sp. (1.9%/0.4%). Highly elevated percentages in UA enrichments were observed for the genera *Acinetobacter* (0.9% in UA enrichments/0.1% in AMM enrichments), *Halomonas* (6.1%/4.2%), *Marinobacter* (16.8%/6.3%) and *Psychrobacter* (6.8%/0.2%). A direct comparison of percentages of potentially oil degrading microbial genera in both microcosms showed a higher percentage of *Acinetobacter* sp. (0.9%/0.1%), *Idiomarina* sp. (0.8%/0.3%), *Oleiphilus* sp. (0.2%/0.03%) and *Marinobacter* sp. (16.8%/6.3%) but lower percentages of *Alcanivorax* sp. (3.9%/4.9%) and *Thalassolituus* sp. (0.04/0.8%) in the UA treatments (Gertler et al., 2015).

TABLE 1 | Relative abundance of microbial families within the AMM and UA microcosms.

Family or phylum[1]	Relative abundance (%) based on 16S small subunit rRNA data[1]	
	MGS-ANC(AMM)	MGS-ANC(UA)
Pseudomonadaceae	15,24	12,98
Alcanivoraceae	4,99	3,94
Halomonadaceae	5,20	7,80
Enterobacteriaceae	5,62	3,15
Vibrionaceae	4,01	2,26
Aeromonadaceae	0,90	3,11
Alteromonadaceae	9,57	19,18
Chromatiaceae	1,06	0,85
Idiomarinaceae	0,36	0,81
Legionellaceae	0,23	0,91
Methylococcaceae	0,54	1,05
Moraxellaceae	0,45	7,89
Oceanospirillaceae	4,26	4,15
Pseudoalteromonadaceae	0,40	2,18
Shewanellaceae	2,78	2,59
Rhodobacteraceae	18,20	0,81
Comamonadaceae	1,64	0,54
Flavobacteriaceae	1,75	5,69
Actinobacteria	1,81	2,56
Firmicutes	4,86	7,42
Others	16,12	10,29

Results are based on the analysis of the partial 16S ribosomal RNA (rRNA) gene sequences extracted from non-assembled DNA sequences obtained by paired-end Illumina HiSeq 2000 sequencing.

[1]*Only lineages with abundance of reads >1% are shown. Data from Gertler et al. (2015).*

Biodegradation Networks

As we were interested in obtaining networks that emphasized the catabolic differences within both microcosms, we selected a metagenomic approach to query the presumptive degradation capacities associated to both microcosms. The identification

depends heavily on gene abundance and, despite the fact that a substantial faction of less abundant DNA in metagenomes remains undiscovered, the identified catabolic genes are assumed to represent the dominant presumptive pathways in each system. A rarefaction curve of the observed species for both samples to analyze species sampling coverage indicated closeness to saturation in each of the two microcosms (Gertler et al., 2015). In combination with the fact that both samples were sequenced to a similar extent (24,752,834 bp for AMM and 19,364,101 bp for UA; Gertler et al., 2015), this suggests that biases during the comparative analysis within the metagenomes were not introduced due to differences in microbial and sequence coverage.

Using as a query the 27,893 (for UA) and 32,180 (for AMM) potential protein-coding genes (for ≥20 amino acids-long polypeptides) (Gertler et al., 2015), we identified respectively a total of 45 (or 0.16% relative abundance in UA referred to the total number of protein-coding genes) and 65 (or 0.20% relative abundance in AMM) genes encoding catabolic enzymes with matches in AromaDeg (Duarte et al., 2014). This suggests that the biostimulants did not have much influence on the relative abundance of catabolic genes. However, significant differences can be observed when examining the diversity of genes encoding catabolic enzymes assigned to different families (**Figure 1**). The amount of genes encoding Rieske non-heme iron oxygenases and extradiol dioxygenases (EXDO) of the cupin superfamily increased 2- and 4-fold, respectively, and proved more abundant in the AMM microcosm in comparison to those in the UA microcosm (**Figure 1**).

The differences in family shifts may have an influence on degradation capacities provided by microorganisms in AMM and UA microcosms. To assess this, the presumptive aromatic degradation reactions and the substrate pollutants or intermediates possibly degraded by each of the two communities were predicted, and the corresponding degradation networks constructed (**Figure 2**). For that we used the AromaDeg web system that allows identifying catabolic genes and appropriated scripts and commands for graphics (for details see Biodegradation Network Reconstruction: Scripts and Commands for Graphics). Unambiguous reaction specificities could be detected for 35 (in UA) and 48 (in AMM) catabolic genes and were considered in the degradation network (**Figure 2**). However, no clear specificities could be assigned to 4 (in UA) and 11 (in AMM) Rieske oxygenases and 12 (six in UA and six in AMM) EXDO, which subsequently were not considered in the network. As shown in **Figure 2**, on the basis of the presence of genes encoding catabolic genes involved in particular transformations, the potential degradation of nine intermediates involved in the degradation of six key pollutants (2-chlorobenzoate, indole-3-acetate, biphenyl, gentisate, quinoline

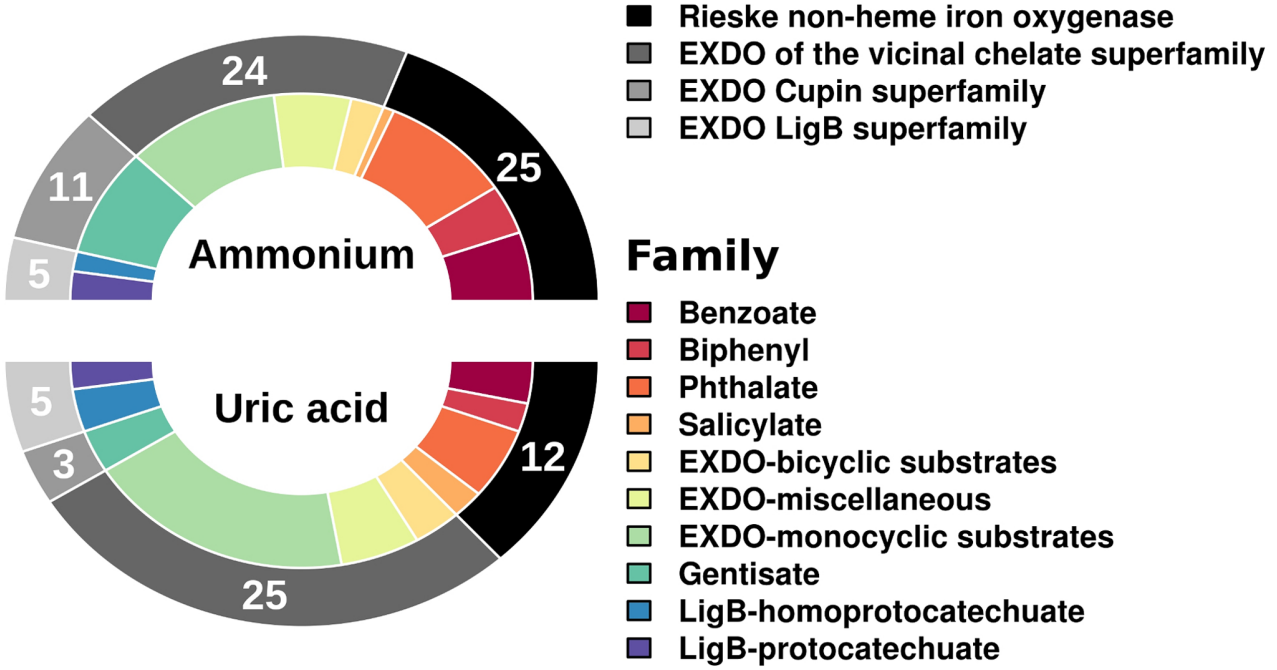

FIGURE 1 | Number and diversity of sequences of gene families encoding key catabolic enzymes involved in the degradation of aromatic pollutants. Catabolic genes were identified as follows. Briefly, predicted open reading frames (ORFs) in the metagenomic DNA sequences were filtered by sequence homology (>50%) and minimum alignment length (> 50 amino acids) according to their similarity to the AromaDeg sequences of key aromatic catabolic gene families (and sub-families) involved in the degradation of aromatic pollutants (Duarte et al., 2014). After a manual check, a final list of gene sequences encoding enzymes potentially involved in degradation was prepared.

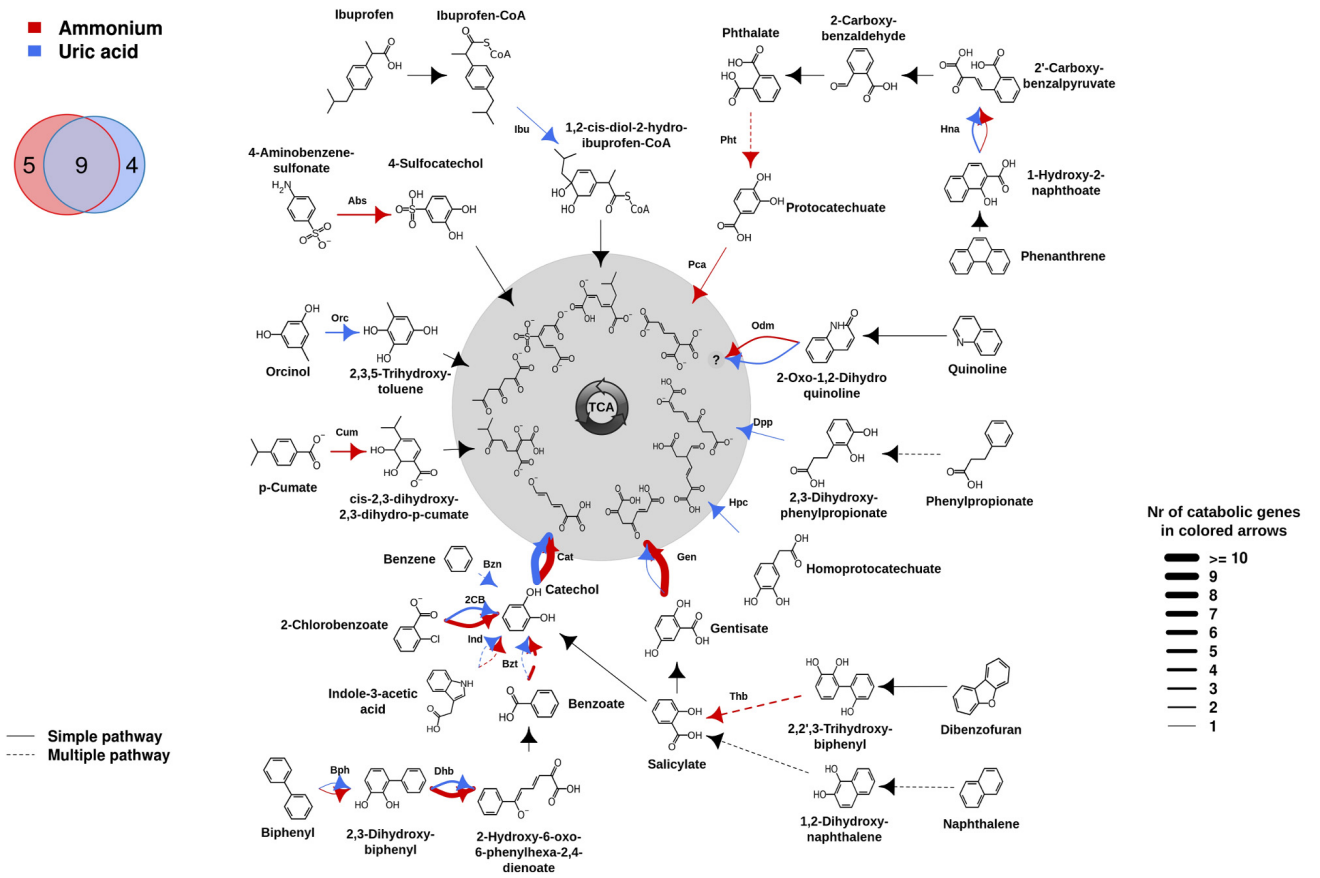

FIGURE 2 | Potential aromatic catabolic networks in the AMM and UA microcosms (see color code). The biodegradation network reconstruction was performed as described in "Materials and Methods." Briefly, catabolic genes were identified as described in the **Figure 1** legend. For network reconstruction, each sequence subsequently was assigned to a metabolic substrate as well as a product (as defined by Duarte et al., 2014) with an assigned code. The putative substrates and products processed in the sample were connected, creating a metabolic network using appropriate scripts and commands (for details, see Bargiela et al., 2015). The number of each catabolic gene assigned to degradation reactions, is represented by the thickness of the lines in the figure and the complete list of substrates possibly degraded by the communities are summarized. Common and microcosm-specific initial pollutants or intermediates for which presumptive degradation signatures were identified are specifically indicated in the Venn diagram. Solid lines represent single step reactions while dotted lines represent degradation steps where multiple reactions are involved (for details see Bargiela et al., 2015). Codes for proteins encoded by genes as follows: Abs, 4-aminobenzenesulfonate 3,4-dioxygenase; Bph, biphenyl dioxygenase; Bzn, benzene dioxygenase; Bzt, benzoate dioxygenase; Cat, catechol 2,3-dioxygenase; 2CB, 2-chlorobenzoate dioxygenase; Cum, *p*-cumate dioxygenase; Dhb, 2,3-Dihydroxybiphenyl dioxygenase; Dpp, 2,3-dihydroxyphenylpropionate dioxygenase; Gen, gentisate dioxygenase; Hna, 1-hydroxy-2-naphthoate dioxygenase; Hpc, homoprotocatechuate 2,3-dioxygenase; Ibu, ibuprofen-CoA dioxygenase; Ind, Rieske oxygenase involved in indole acetic acid degradation; Odm, 2-oxo-1,2-dihydroxyquinoline monooxygenase; Orc, orcinol hydroxylase; Pca, protocatechuate 3,4-dioxygenase; Pht, phthalate 4,5-dioxygenase; Thb, 2,2′,3-trihydroxybiphenyl dioxygenase.

and phenanthrene) was found to be common for both microcosms. They include the transformation of biphenyl by Bph, 2,3-dihydroxybiphenyl by Dhb, benzoate by Bzt, indole-3-acetate by Ind, catechol by Cat, gentisate by Gen, 2-oxo-1,2-dihydroquinoline by Odm, 1-hydroxy-2-naphthoate by Hna, and 2-chlorobenzoate by 2-chlorobenzoate dioxygenase (2CB). Within them, genes encoding Cat were most abundant in both communities (UA: 17; AMM: 14), in agreement with the fact that catechol is the central intermediate for most cyclic aerobic hydrocarbons degradation (Pérez-Pantoja et al., 2009; Vilchez-Vargas et al., 2013). Gentisate and benzoate/2-chlorobenzoate may be most likely preferentially degraded by microorganisms in the AMM microcosm (10 Gen and 4 Bzt/2CB) in comparison to the UA microcosm (1 Gen and 1 Bzt/2CB). Genomic signatures for the degradation of orcinol (or 3,5-dihydroxytoluene) by Orc, phenylpropionate by Dpp, homoprotocatechuate by Hpc, and benzene by Bzn, were only found in the UA microcosm. The potential degradation of ibuprofen by Ibu, although not being a constituent of the crude oil but possibly originated from bilge water from the cruise lines or urban run-off, was also identified in UA microcosm. In stark contrast, the degradation of 4-aminobenzene-sulfonate by Abs, *p*-cumate by Cum, dibenzofuran by Thb, phthalate by Pht and protocatechuate by Pca, was characteristic for the AMM microcosm.

Note that within all pollutants predicted as being potentially degraded by bacteria inhibiting Ancona port (**Figure 2**), independently whether they are enriched with AMM or UA, only the potential degradation of

ibuprofen and 4-aminobenzene-sulfonate was not found associated to bacteria from other chronically crude oil-contaminated sites in oil-polluted sites along the coastlines of the Mediterranean Sea (Bargiela et al., 2015). This suggests that the pollution type and pollutant diversity in Ancona port, which receives chemicals such as alkyl benzene sulfonate detergents and drugs coming from human activities (Martínez-Pascual et al., 2010; Paíga et al., 2013), may have supported the presence of ibuprofen- and sulfonate benzene-growing bacteria. Such bacteria may be further stimulated by either the addition of UA or AMM, respectively.

Experimental Analysis of Catabolic Capacities in AMM and UA Microcosms

Experimental validation assays were conducted to prove the extent of agreement with metagenomic-based predictions. For that, AMM and UA enrichment cultures were set up in triplicates as described in Section "Experimental Validations of Predicted Biodegradation Capacities," in which instead of Arabian light crude oil as the carbon source (used for the initial microcosms), naphthalene, 2,3-dihydroxybiphenyl, benzene, *p*-cumate, orcinol, 2-chlorobenzoate, phthalate and phenylpropionate (2 ppm each) were used. The capacity to degrade other pollutants predicted as potential substrates such as ibuprofen, phenanthrene, dibenzofuran, indole-3-acetic acid, 4-aminobenzene-sulfonate and quinoline, could not be experimentally proved because no analytical procedures could be designed for their analysis in the pollutant mix.

Samples were taken at 21 days of incubation at 20°C. Fingerprinting by LC-ESI-QTOF-MS and GC-MS was used to confirm the degradation of the initial substrates as well as the existence of degradation intermediates in both cultures. A careful inspection of the mass signatures confirmed the lowering in the abundance level of naphthalene, 2,3-dihydroxybiphenyl, and 2-chlorobenzoate, and the presence of catechol, salicylate, gentisate, and benzoate in both microcosms (**Figure 3**). This demonstrates that the naphthalene-to-salicylate-to-gentisate, 2,3-dihydroxybiphenyl-to-benzoate-to-catechol, and 2-chlorobenzoate-to-catechol degradation pathways occurred or were active in both microcosms. Note that the lower abundance level of gentisate in AMM microcosm may correlate with the 10-fold overabundance of genes encoding Gen enzymes in AMM as compared to UA; this may decrease the pool of gentisate in the microcosm when growing in naphthalene. We further found a decreased level of *p*-cumate only associated to the AMM enrichment. Phthalate degradation mostly associated to the AMM microcosm, as confirmed by the higher extend of phthalate degradation by meaning of its residual percentage at the end of the assay (21 ± 1.8% in AMM vs. 92.2 ± 2.3% in UA) and the 22.2-fold higher abundance of protocatechuate in AMM compared to UA assays. In addition, decreased level of orcinol, benzene and phenylpropionate associated only to UA enrichments (**Figure 3**). Accordingly, the benzene-to-catechol, orcinol-, and phenylpropionate-degradation pathways occurred or were active in the UA microcosm, while *p*-cumate degradation mostly occurred in the AMM enrichments.

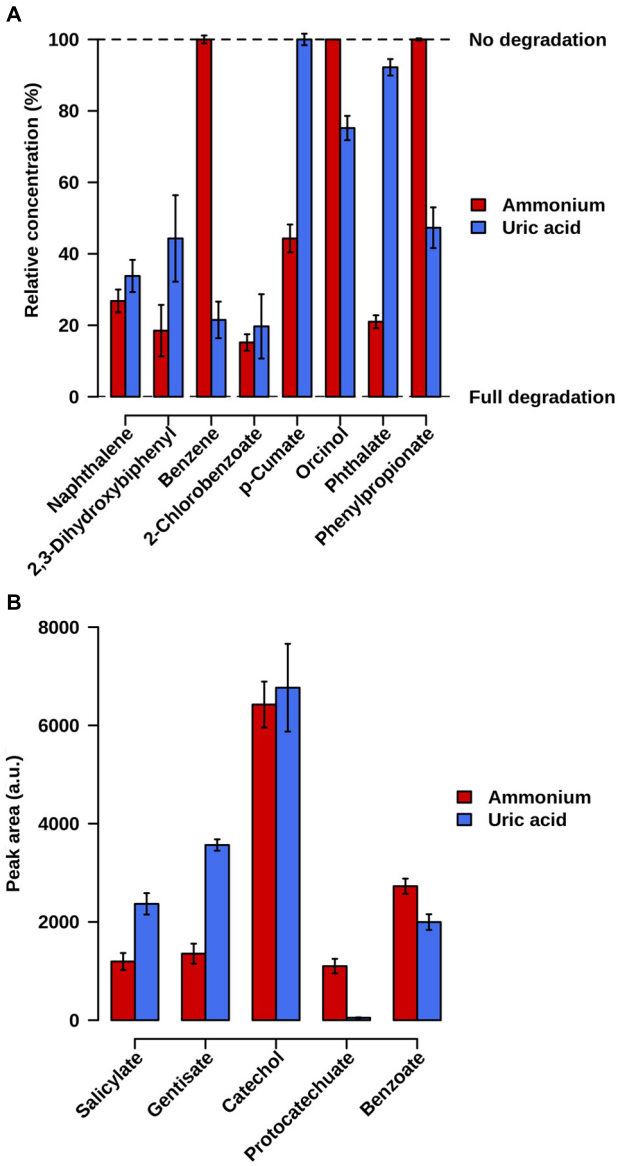

FIGURE 3 | Relative abundance level of initial substrate pollutants (A) and key chemical intermediates (B), in AMM and UA microcosms containing naphthalene, 2,3-dihydroxybiphenyl, benzene, *p*-cumate, orcinol, 2-chlorobenzoate, phthalate and phenylpropionate (2 ppm each) as carbon source. (A) The remaining relative concentration of the initial pollutants used to set up enrichment cultures is shown; 100%, no degradation of initial substrate pollutant; 0%, total degradation (absence of pollutant). **(B)** Values represent the peak area of degradation intermediates in arbitrary units (a.u.). The values were calculated, in triplicate microcosms, by comparing the presence and abundance level after 21-days of the microcosm at 20°C experiment compared to the initial point and after considering the controls assays. Standard deviations (SD) are shown.

The identification of degrading capacities on microcosms depends heavily on enrichment conditions (including cultivation time frame) and bacteria and protein abundance. While these

drawbacks are known, the experimental data presented above (**Figure 3**) fully confirmed our sequence-based predictions (**Figure 2**) for the degradation of all eight pollutants tested in each of the two amendments. This suggests that the differences herein predicted in UA and AMM microcosms (**Figure 2**) are due to real biological differences and not random. Uncertainty remains only for phthalate degradation in UA microcosm: experimental analysis demonstrated the slight degradation of this chemical (**Figure 3**), which was not predicted by sequence analysis (**Figure 2**).

Phylogenetic Identities of Catabolic Genes

We further attempted to analyze the contributions of particular sets of microbes to the entire reconstructed catabolic network, where multiple proteins from multiple organisms may contribute to organic pollutants' decomposition.

As the community structure of the two enrichment cultures was well-characterized (Gertler et al., 2015), the taxonomic affiliations of the catabolic genes identified could be unambiguously established at the family and phylum level. For that, we used tools recently published that provide a high level of confidence (Guazzaroni et al., 2013; Bargiela et al., 2015). **Figure 4** shows the contribution of members assigned to the different bacterial families and phyla in both microcosms to pollutant degradation. They included populations closely related to members of *Aeromonadaceae, Alcanivoracaceae, Alteromonadaceae, Halomonadaceae, Oceanospirillaceae, Piscirickettsiaceae, Pseudomonadaceae, Rhodobacteraceae, Vibrionaceae,* and *Xhantomonadaceae*, as well as to a lesser extent for the phyla Actinobacteria and Firmicutes. These comprise bacterial groups well known for their oil biodegrading capabilities (Yakimov et al., 2007; Jin et al., 2012; Guazzaroni et al., 2013). A further careful examination of the data presented

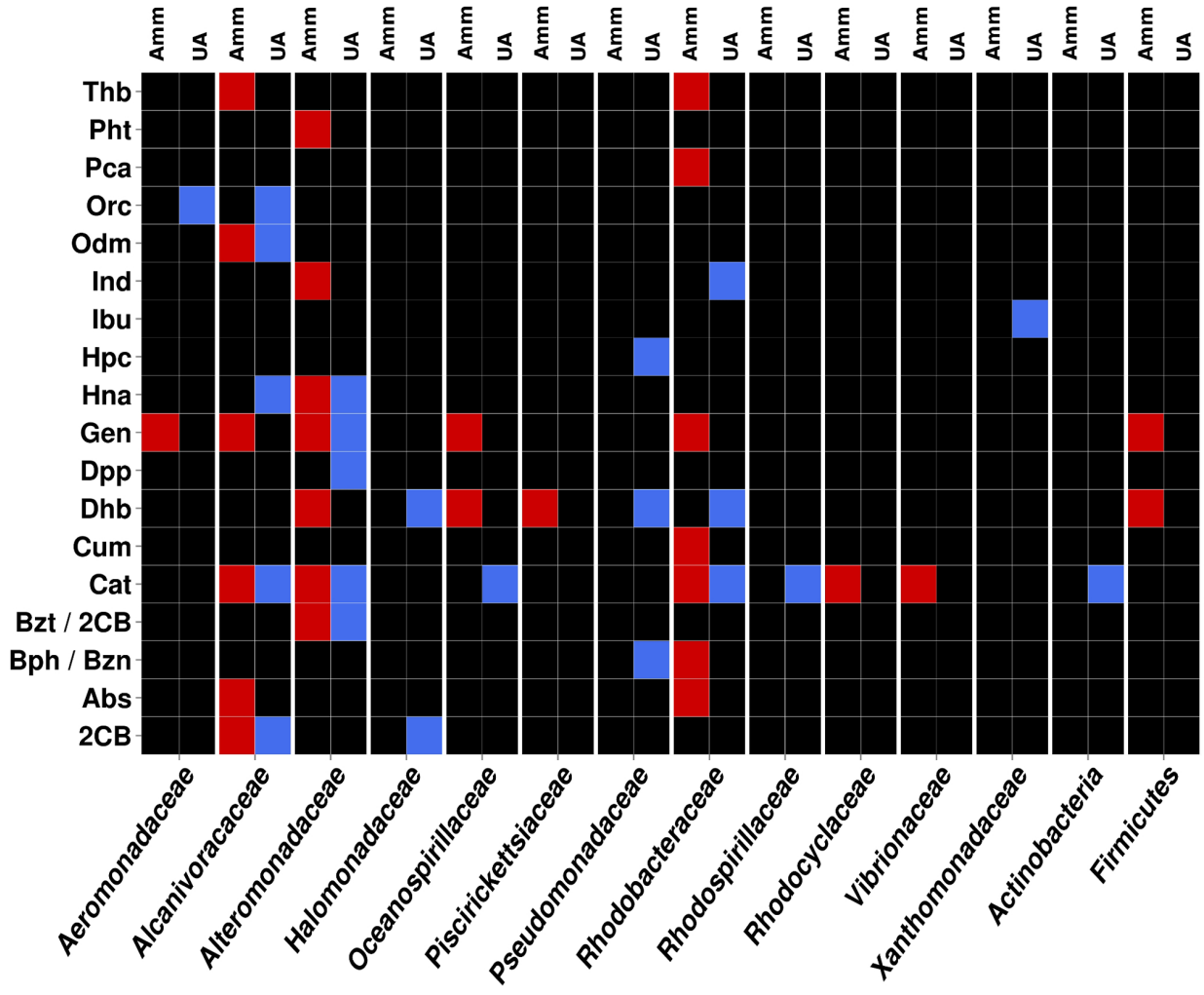

FIGURE 4 | Heat map showing the contribution of the most relevant bacterial members of AMM and UA microcosm to the degradation network in Figure 2. Contributions of each of the distinct members with unambiguous taxonomic assignation per each of the catabolic gene classes found to constitute the AMM and UA communities are differentiated by a color code. The color indicates the presence of a catabolic gene independently of the abundance level. Gene names/codes are identical to those presented in **Figure 2**.

in **Figure 4** clearly leads to the occurrence of a different pathway organization at organism level for the catabolism of 18 different pollutants predicted to be degraded.

As can be seen in **Figure 4**, members of *Alcanivoracaceae, Alteromonadaceae,* and *Rhodobacteraceae* were the major contributors to the networks. They contribute, in combination, to the degradation of 16 out of 18 pollutants predicted in the catabolic network, including dibenzofuran, phenanthrene, indolacetic acid, biphenyl, *p*-cumate, 2-chlorobenzoate, phenylpropionate, aminobenzenesulfonate and gentisate. This is in agreement with the fact that they were among the most abundant members in the established microcosms based on 16S rRNA (**Table 1** and **Figure 5**). Interestingly, *Pseudomonadaceae* which was the second most abundant microbial clade at the level of 16S rRNA in both microcosms (**Table 1** and **Figure 5**), did not contribute to the degradation network in AMM but it does in the UA microcosm (**Figure 4**), where it supports the biphenyl-to-benzoate and homoprotocatechuate degradation.

As shown in **Figure 4**, among the common degradation capacities, a number of observations can be made. First, the degradation of indole acetate by Ind was supported by members of *Alteromonadaceae* in AMM and *Rhodobacteraceae* in UA, which suggests a catabolic replacement. This was also observed for the degradation of biphenyl and benzene (by Bph/Bzn), most likely supported by members of the *Pseudomonadaceae* in UA but *Rhodobacteraceae* in AMM. We identified members of five proteobacterial families (*Aeromonadaceae, Alcanivoracaceae, Alteromonadaceae, Oceanospirillaceae*, and *Rhodobacteraceae*) and of the Firmicutes phylum as key groups for the degradation of gentisate (by Gen) in AMM. By contrast, only members of *Alteromonadaceae* were predicted to support gentisate catabolism in UA. In agreement with this it has been found that AMM promotes the growth of such multiple marine bacteria with the ability to utilize naphthalene (the precursor of gentisate) as a sole carbon in enrichment cultures (Hedlund et al., 1999). Also, the increased abundance of bacteria of the Firmicutes phylum has been demonstrated during bio-stimulation with ammonia (Guazzaroni et al., 2013). The naphthalene-to-gentisate catabolism by bacteria of the family *Alteromonadaceae* has also been found during microcosm assays using seawater and sediment samples obtained after an oil spill along the Korean shoreline without AMM addition (Jin et al., 2012); this agrees with the enrichment of gentisate catabolism by bacteria of this family in UA microcosm.

Multiple bacteria also contributed to the degradation of catechol (by Cat), with members of *Alcanivoracaceae, Alteromonadaceae,* and *Rhodobacteraceae* being common in both treatments. These bacterial groups are known for their capacity to degrade aromatics and haloaromatics to catechol, which can be further catabolised (Brusa et al., 2001; Antunes et al., 2011). Members of the Actinobacteria phyum and *Oceanospirillaceae* family contributed to catechol catabolism exclusively in the UA microcosm, whereas those of *Vibrionaceae* family did so in the AMM treatment. Note that, in accordance with the fact that *cat* genes are the most abundantly present (**Figure 2**) in both microcosms, the number of bacterial groups

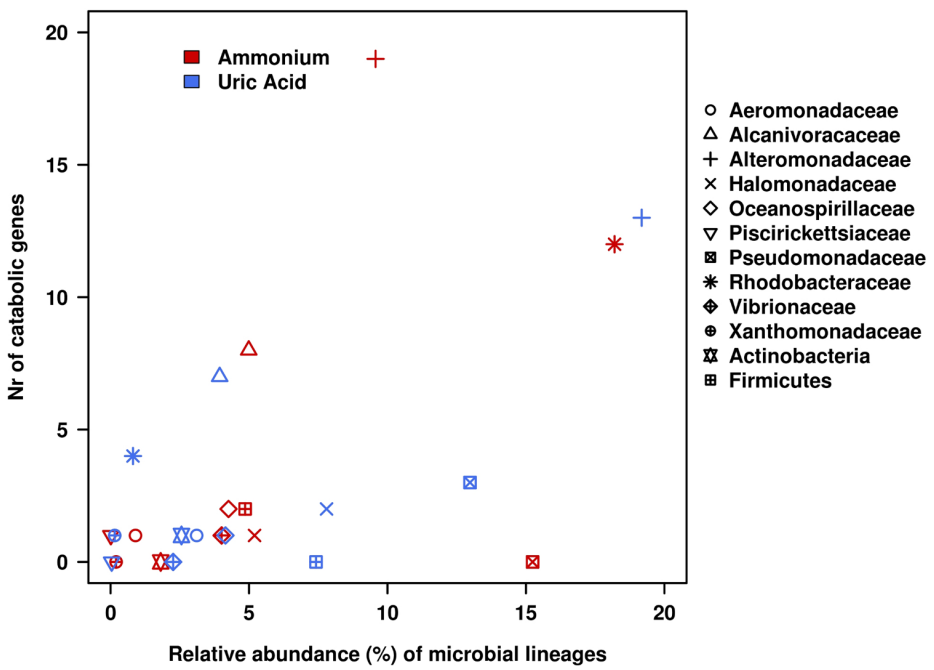

FIGURE 5 | Relative contributions of bacterial lineages in terms of catabolic genes and 16S rRNA gene within the AMM and UA microcosms. Catabolic genes were identified by using the web-based AromaDeg resource (Duarte et al., 2014) and their taxonomic assignation at the level of the family or higher was performed as reported (Guazzaroni et al., 2013; Bargiela et al., 2015). The relative abundance of microbial lineages was based on the analysis of 16S small subunit rRNA partial sequences obtained in the non-assembled Illumina reads (Gertler et al., 2015).

involved in its catabolism was also the highest (8 in total; **Figure 4**). Therefore, a number of bacterial groups within the microcosms exhibit also partial catabolism redundancy.

Interestingly, we noticed that bacteria from the *Halomonadaceae* family contributed also to degradation of aromatics, particularly, 2-chlorobenzoate (through 2CB) and biphenyl (through Dhb) in the UA microcosm (**Figure 4**). This suggests that halomonads not only participate in the conversion of UA to AMM, which further stimulated growth of hydrocarbonoclastic bacteria (Gertler et al., 2015), but also play specific roles in degradation as herein suggested. This agrees with the fact that bacteria from the genus *Halomonas* are capable of degrading chlorobenzoates (de la Haba et al., 2011) and aromatics compounds such as benzoate and catechol (Piubeli et al., 2012), that are intermediate products of biphenyl and 2-chlorobenzoate degradation.

CONCLUSION

Here, we report that different biostimulants applied in chronically polluted sediments have caused significant alteration in degradation capacities, while having no major effect on the taxonomic composition of microbial communities at the level of the family or higher. Experimental validation was conducted for at least eight of the predicted catabolic capacities, and good agreement with metagenomics-based predictions was observed. On the other hand, the metagenomics-guided metabolic reconstruction allowed us to refine the assignment of roles of community members in the utilization of multiple substrates and found different pathway organization at organism level. For example, while biphenyl degradation by Bph, DhB, and Bzt enzymes seems to be carried out by bacteria of *Pseudomonadaceae*, *Halomonadaceae,* and *Rhodobacteraceae* in UA, those of *Alteromonadaceae, Oceanospirillaceae, Picirickettsiaecae,* and Firmicutes may be involved in an alternative pathway in AMM. This demonstrates that different microbial members within microcosms obtained with different nitrogen sources may exhibit partial functional redundancy, and thus, may have a high level of common catabolic capacities. The present investigation provides an estimation of such common and distinct degrading capacities. Indeed, herein we found that 50% of the predicted degradation capacities were common for microorganisms in AMM and UA microcosms (**Figure 2**). However, according to the microbial biodegradation networks herein reconstructed, we also found that the two different biostimulants investigated, UA and AMM, have also changed substrate utilization capacities and preferences, which must be considered for the design of petroleum bioremediation techniques. This was demonstrated by showing that UA enriched for bacteria with the capability of degrading pollutants otherwise not degraded, or possibly degraded at low level, by those stimulated by the addition of AMM, and vice versa.

Therefore, the results of this study show that smart formulations based on the application of multiple nitrogen sources, rather than commonly used single sources (mostly AMM), for example, may increase the efficiency of the biological removal of the widest diversity of aromatic pollutants and could be essential to support effective biodegradation strategies in response to an oil spill incident or in response to chronical pollution. Thus, as herein demonstrated, the utilization of both AMM and UA in conjunction will have a double aim. In one side, AMM may most likely enhance the bio-stimulation of bacterial populations capable of degrading heavy oil components such as naphthalene, phenathrene and dibenzofuran, as well as sulfonated-benzenes and substituted benzoate derivatives such as p-cumate (**Figure 2**). In other side, UA will promote the growth of bacteria most active against benzene, orcinol-, ibuprofen- and phenyl-propionate (**Figure 2**). This will provoke a significant increase in multiple aromatics consumption in polluted areas. Having said that, this work seems to introduce a promising way for future oil-based contamination handling techniques. In this context, it would be very interesting to test the overall cleaning capacity (if any) on a real oil-contaminated marine sample. For that, also another point will be to use the combination of the UA and AMM, which was herein not presented in microcosm assays. It would be interesting to see their combinatory effect in the overall degradation capacity and taxonomic distribution of the microbial niche depending also on their ratio, so to find optimal nitrogen-containing formulations in real scenarios.

We would like to stress the attention to the fact that similarities regarding microbial community composition in the AMM microcosm from Ancona port with those reported in enrichments from surface water bodies at other Mediterranean sites, were found (Gertler et al., 2012). However, a similar comparison with the results from UA microcosm cannot be established because the limited information available. In fact, to the best of our knowledge, there have been only three studies that thoroughly investigated the use of UA in bioremediation trials (Koren et al., 2003; Knezevich et al., 2006; Nikolopoulou et al., 2013). Those studies, however, did not use UA in comparison to other nitrogen sources such as AMM, both in respect to their effect in microcosm population structures and catabolic preferences. Accordingly, herein we reported first evidences linking UA to catabolic preferences at the bacterial level, in comparison to the commonly use nitrogen source AMM.

ACKNOWLEDGMENTS

This research was supported by the European Community Projects MAGICPAH (FP7-KBBE-2009-245226), ULIXES (FP7-KBBE-2010-266473) and KILL-SPILL (FP7-KBBE-2012-312139). We thank EU Horizon 2020 Program for the support of the Project INMARE H2020-BG-2014-2634486. This work was further funded by grants BIO2011-25012, PCIN-2014-107 and BIO2014-54494-R from the Spanish Ministry of Economy and Competitiveness. The authors gratefully acknowledge the financial support provided by the European Regional Development Fund (ERDF). The present investigation was also funded by the Spanish Ministry of Economy and Competitiveness within the ERA NET IB2, grant number ERA-IB-14-030. FM was supported by Università degli Studi di Milano, European Social Fund (FSE) and Regione Lombardia (contract "Dote Ricerca").

REFERENCES

Alvarez, V. M., Marques, J. M., Korenblum, E., and Seldin, L. (2011). Comparative bioremediation of crude oil-amended tropical soil microcosms by natural attenuation, bioaugmentation, or bioenrichment. *Appl. Environ. Soil Sci.* 2011, 156320. doi: 10.1155/2011/156320

Antunes, A., Ngugi, D. K., and Stingl, U. (2011). Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environ. Microbiol. Rep.* 3, 416–433. doi: 10.1111/j.1758-2229.2011.00264.x

Atlas, R. M. (1981). Microbial degradation of petroleum hydrocarbons: an environmental perspective. *Microbiol. Rev.* 45, 180–209.

Bargiela, R., Mapelli, F., Rojo, D., Chouaia, B., Tornés, J., Borin, S., et al. (2015). Bacterial population and biodegradation potential in chronically crude oil-contaminated marine sediments are strongly linked to temperature. *Sci. Rep.* 5, 11651. doi: 10.1038/srep11651

Brusa, T., Borin, S., Ferrari, F., Sorlini, C., Corselli, C., and Daffonchio, D. (2001). Aromatic hydrocarbon degradation patterns and catechol 2,3-dioxygenase genes in microbial cultures from deep anoxic hypersaline lakes in the eastern Mediterranean sea. *Microbiol. Res.* 156, 49–58. doi: 10.1078/0944-5013-00075

Das, N., and Chandran, P. (2010). Microbial degradation of petroleum hydrocarbon contaminants: an overview. *Biotechnol. Res. Int.* 2011, 941810. doi: 10.4061/2011/941810

de la Haba, R. R., Sánchez-Porro, C., and Ventosa, A. (2011). "Taxonomy, phylogeny, and biotechnological interest of the family Halomonadaceae," in *Halophiles and Hypersaline Environments: Current Research and Future Trends*, eds A. Ventosa, A. Oren, and Y. Ma (Heidelberg: Springer), 27–64.

Duarte, M., Jauregui, R., Vilchez-Vargas, R., Junca, H., and Pieper, D. H. (2014). AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database (Oxford)* 2014:bau118. doi: 10.1093/database/bau118

Dyksterhouse, S. E., Gray, J. P., Herwig, R. P., Lara, J. C., and Staley, J. T. (1995). *Cycloclasticus pugetii* gen. nov., sp. nov., an aromatic hydrocarbon-degrading bacterium from marine sediments. *Int. J. Syst. Bacteriol.* 45, 116–123. doi: 10.1099/00207713-45-1-116

Garcia-Blanco, S., Venosa, A. D., Suidan, M. T., Lee, K., Cobanli, S., and Haines, J. R. (2007). Biostimulation for the treatment of an oil-contaminated coastal salt marsh. *Biodegradation* 18, 1–15. doi: 10.1007/s10532-005-9029-3

Gertler, C., Bargiela, R., Mapelli, F., Han, X., Chen, J., Hai, T., et al. (2015). Conversion of uric acid into ammonium in oil-degrading marine microbial communities: a possible role of Halomonads. *Microb. Ecol.* 70, 724–740. doi: 10.1007/s00248-015-0606-7

Gertler, C., Näther, D. J., Cappello, S., Gerdts, G., Quilliam, R. S., Yakimov, M. M., et al. (2012). Composition and dynamics of biostimulated indigenous oil-degrading microbial consortia from the Irish, North and Mediterranean Seas: a mesocosm study. *FEMS Microbiol. Ecol.* 81, 520–536. doi: 10.1111/j.1574-6941.2012.01377.x

Guazzaroni, M. E., Herbst, F. A., Lores, I., Tamames, J., Peláez, A. I., López-Cortés, N., et al. (2013). Metaproteogenomic insights beyond bacterial response to naphthalene exposure and bio-stimulation. *ISME J.* 7, 122–136. doi: 10.1038/ismej.2012.82

Hedlund, B. P., Geiselbrecht, A. D., Bair, T. J., and Staley, J. T. (1999). Polycyclic aromatic hydrocarbon degradation by a new marine bacterium, *Neptunomonas naphthovorans* gen. nov., sp. nov. *Appl. Environ. Microbiol.* 65, 251–259.

Howarth, R. W., and Marino, R. (2006). Nitrogen as the limiting nutrient for eutrophication in coastal marine ecosystems: evolving views over three decades. *Limnol. Oceanogr.* 51, 364–376. doi: 10.4319/lo.2006.51.1_part_2.0364

Jin, H. M., Kim, J. M., Lee, H. J., Madsen, E. L., and Jeon, C. O. (2012). Alteromonas as a key agent of polycyclic aromatic hydrocarbon biodegradation in crude oil-contaminated coastal sediment. *Environ. Sci. Technol.* 46, 7731–7740. doi: 10.1021/es3018545

Knezevich, V., Koren, O., Ron, E. Z., and Rosenberg, E. (2006). Petroleum bioremediation in seawater using guano as the fertilizer. *Bioremediat. J.* 10, 83–91. doi: 10.1080/10889860600939492

Koren, O., Knezevic, V., Ron, E. Z., and Rosenberg, E. (2003). Petroleum pollution bioremediation using water-insoluble uric acid as the nitrogen source. *Appl. Environ. Microbiol.* 69, 6337–6339. doi: 10.1128/AEM.69.10.6337-6339.2003

Li, H., Zhao, Q., Boufadel, M. C., and Venosa, A. D. (2007). A universal nutrient application strategy for the bioremediation of oil-polluted beaches. *Mar. Pollut. Bull.* 54, 1146–1161. doi: 10.1016/j.marpolbul.2007.04.015

Ly, J., Philippart, C. J. M., and Kromkamp, J. C. (2014). Phosphorus limitation during a phytoplankton spring bloom in the western Dutch Wadden Sea. *J. Sea Res.* 88, 109–120.

Martínez-Pascual, E., Jiménez, N., Vidal-Gavilan, G., Viñas, M., and Solanas, A. M. (2010). Chemical and microbial community analysis during aerobic biostimulation assays of non-sulfonated alkyl-benzene-contaminated groundwater. *Appl. Microbiol. Biotechnol.* 88, 985–995. doi: 10.1007/s00253-010-2816-8

Miyasaka, T., Asami, H., and Watanabe, K. (2006). Impacts of bioremediation schemes on bacterial population in naphthalene-contaminated marine sediments. *Biodegradation* 17, 227–235. doi: 10.1007/s10532-005-5018-9

Mohseni-Bandpi, A., Esrafili, A., Nasseri, S., Ashmagh, F. R., Jorfi, S., and Ja'fari, M. (2014). Effectiveness of biostimulation through nutrient content on the bioremediation of phenanthrene contaminated soil. *J. Environ. Health Sci. Eng.* 12, 143. doi: 10.1186/s40201-014-0143-1

Nikolopoulou, M., and Kalogerakis, N. (2010). "Biostimulation strategies for enhanced bioremediation of marine oil spills including chronic pollution," in *Handbook of Hydrocarbon and Lipid Microbiology*, ed. K. N. Timmis (Berlin: Springer-Verlag), 2521–2529.

Nikolopoulou, M., Pasadakis, N., and Kalogerakis, N. (2013). Evaluation of autochthonous bioaugmentation and biostimulation during microcosm-simulated oil spills. *Mar. Pollut. Bull.* 72, 165–173. doi: 10.1016/j.marpolbul.2013.04.007

Paíga, P., Santos, L. H., Amorim, C. G., Araújo, A. N., Montenegro, M. C., Pena, A., et al. (2013). Pilot monitoring study of ibuprofen in surface waters of north of Portugal. *Environ. Sci. Pollut. Res. Int.* 20, 2410–2420. doi: 10.1007/s11356-012-1128-1

Pérez-Pantoja, D., Donoso, R., Junca, H., Gonzalez, B., and Pieper, D. H. (2009). "Phylogenomics of aerobic bacterial degradation of aromatics," in *Handbook of Hydrocarbon and Lipid Microbiology*, ed. K. N. Timmis (Berlin: Springer-Verlag), 1356–1397.

Piubeli, F., Grossman, M. J., Fantinatti-Garboggini, F., and Durrant, L. R. (2012). Identification and characterization of aromatic degrading halomonasin hypersaline produced water and cod reduction by bioremediation by the indigenous microbial population using nutrient addition. *Chem. Eng. Trans.* 27, 385–390.

Reis, E. A., Rocha-Leão, M. H., and Leite, S. G. (2013). Slow-release nutrient capsules for microorganism stimulation in oil remediation. *Appl. Biochem. Biotechnol.* 169, 1241–1249. doi: 10.1007/s12010-012-0022-0

Scott, N. M., Hess, M., Bouskill, N. J., Mason, O. U., Jansson, J. K., and Gilbert, J. A. (2014). The microbial nitrogen cycling potential is impacted by polyaromatic hydrocarbon pollution of marine sediments. *Front. Microbiol.* 5:108. doi: 10.3389/fmicb.2014.00108

Teramoto, M., Suzuki, M., Okazaki, F., Hatmanti, A., and Harayama, S. (2009). Oceanobacter-related bacteria are important for the degradation of petroleum aliphatic hydrocarbons in the tropical marine environment. *Microbiology* 155, 3362–3370. doi: 10.1099/mic.0.030411-0

Venosa, A. D., Campo, P., and Suidan, M. T. (2010). Biodegradability of lingering crude oil 19 years after the Exxon Valdez oil spill. *Environ. Sci. Technol.* 44, 7613–7621. doi: 10.1021/es101042h

Vilchez-Vargas, R., Geffers, R., Suárez-Diez, M., Conte, I., Waliczek, A., Kaser, V. S., et al. (2013). Analysis of the microbial gene landscape and transcriptome for aromatic pollutants and alkane degradation using a novel internally calibrated microarray system. *Environ. Microbiol.* 15, 1016–1039. doi: 10.1111/j.1462-2920.2012.02752.x

Walther, H. R. III. (2014). *Clean Up Techniques used for Coastal Oil Spills: An Analysis of Spills Occurring in Santa Barbara, California, Prince William sound, Alaska, the Sea of Japan and the Gulf Coast*. Ph.D. thesis, Environmental Management, University of San Francisco, San Francisco, CA.

Wang, Z., Hollebone, B. P., Fingas, M., Fieldhouse, B., Sigouin, L., Landriault, M., et al. (2003). *Characteristics of Spilled Oils, Fuels, and Petroleum Products: 1. Composition and Properties of Selected Oils*. Research Triangle Park, NC: United States Environmental Protection Agency, National Exposure Research Laboratory, EPA/600/R-03/072.

Yakimov, M. M., Timmis, K. N., and Golyshin, P. N. (2007). Obligate oil-degrading marine bacteria. *Curr. Opin. Biotechnol.* 18, 257–266. doi: 10.1016/j.copbio.2007.04.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 July 2015
doi: 10.3389/fmicb.2015.00696

Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins

*Karyna Rosario, Ryan O. Schenck, Rachel C. Harbeitner, Stephanie N. Lawler and Mya Breitbart**

College of Marine Science, University of South Florida, St. Petersburg, FL, USA

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Kenneth Stedman, Portland State University, USA
Purificacion Lopez-Garcia, Centre National de la Recherche Scientifique, France

***Correspondence:**
Mya Breitbart, College of Marine Science, University of South Florida, 140 7th Avenue South, St. Petersburg, FL 33701, USA
mya@usf.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *27 April 2015*
Accepted: *23 June 2015*
Published: *10 July 2015*

Citation:
Rosario K, Schenck RO, Harbeitner RC, Lawler SN and Breitbart M (2015) Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. Front. Microbiol. 6:696. doi: 10.3389/fmicb.2015.00696

Viral metagenomics has recently revealed the ubiquitous and diverse nature of single-stranded DNA (ssDNA) viruses that encode a conserved replication initiator protein (Rep) in the marine environment. Although eukaryotic circular Rep-encoding ssDNA (CRESS-DNA) viruses were originally thought to only infect plants and vertebrates, recent studies have identified these viruses in a number of invertebrates. To further explore CRESS-DNA viruses in the marine environment, this study surveyed CRESS-DNA viruses in various marine invertebrate species. A total of 27 novel CRESS-DNA genomes, with Reps that share less than 60.1% identity with previously reported viruses, were recovered from 21 invertebrate species, mainly crustaceans. Phylogenetic analysis based on the Rep revealed a novel clade of CRESS-DNA viruses that included approximately one third of the marine invertebrate associated viruses identified here and whose members may represent a novel family. Investigation of putative capsid proteins (Cap) encoded within the eukaryotic CRESS-DNA viral genomes from this study and those in GenBank demonstrated conserved patterns of predicted intrinsically disordered regions (IDRs), which can be used to complement similarity-based searches to identify divergent structural proteins within novel genomes. Overall, this study expands our knowledge of CRESS-DNA viruses associated with invertebrates and explores a new tool to evaluate divergent structural proteins encoded by these viruses.

Keywords: single-stranded DNA virus, CRESS-DNA virus, circular DNA virus, intrinsically disordered proteins (IDPs), intrinsically disordered regions (IDRs), marine invertebrate, crustaceans

Introduction

Viral metagenomics, or shotgun sequencing of total nucleic acids from purified virus particles, enables examination of viral communities without prior knowledge of the viruses present, thus resulting in an unprecedented view of viral diversity (Breitbart et al., 2002; Edwards and Rohwer, 2005; Angly et al., 2006). This technique has uncovered many novel viral types and extended the environmental distribution of known viral groups (Delwart, 2007; Rosario and Breitbart, 2011).

In particular, the incorporation of rolling circle amplification (RCA) into viral metagenomic studies has unearthed a high diversity and wide distribution of eukaryotic viruses with circular, single-stranded DNA (ssDNA) genomes that encode a conserved replication initiator protein (Rep; Delwart and Li, 2012; Rosario et al., 2012a). Before the metagenomics era, eukaryotic circular Rep-encoding ssDNA (CRESS-DNA) viruses were only known in agricultural and medical fields since they are known plant (*Geminiviridae* and *Nanoviridae*) and vertebrate (*Circoviridae*) pathogens. However, over the past decade metagenomic approaches have revealed the ubiquitous nature of eukaryotic CRESS-DNA viruses, with reports from various environments, including deep-sea vents (Yoshida et al., 2013), Antarctic lakes and ponds (López-Bueno et al., 2009; Zawar-Reza et al., 2014), wastewater (Rosario et al., 2009b; Roux et al., 2013; Kraberger et al., 2015; Phan et al., 2015), freshwater lakes (Roux et al., 2012, 2013), oceans (Rosario et al., 2009a; Labonte and Suttle, 2013; Roux et al., 2013), hot springs (Diemer and Stedman, 2012), the near-surface atmosphere (Whon et al., 2012; Roux et al., 2013), and soils (Kim et al., 2008; Reavy et al., 2015). Novel CRESS-DNA viruses have also been discovered from fecal samples of a variety of vertebrates (Blinkova et al., 2010; Li et al., 2010a,b; Phan et al., 2011; Ge et al., 2012; Ng et al., 2012; Sachsenroder et al., 2012; van den Brand et al., 2012; Cheung et al., 2013, 2014; Sikorski et al., 2013a; Garigliany et al., 2014; Lian et al., 2014; Smits et al., 2014; Zhang et al., 2014; Sasaki et al., 2015). Notably, CRESS-DNA viruses similar to circoviruses, which were previously thought to only infect vertebrates, have now been identified in a myriad of invertebrates, including insects (Ng et al., 2011; Rosario et al., 2011, 2012b; Dayaram et al., 2013; Padilla-Rodriguez et al., 2013; Pham et al., 2013a,b; Garigliany et al., 2015), crustaceans (Dunlap et al., 2013; Hewson et al., 2013a,b; Ng et al., 2013; Pham et al., 2014), cnidarians (Soffer et al., 2014), and gastropods (Dayaram et al., 2015a), suggesting that CRESS-DNA viruses may be prevalent amongst unexplored taxa.

Well-studied viruses from the *Circoviridae*, *Nanoviridae*, and *Geminiviridae* families demonstrate the rapid evolutionary potential of CRESS-DNA viruses due to high nucleotide substitution rates (Duffy et al., 2008; Duffy and Holmes, 2009) as well as mechanistic predispositions to recombination (Lefeuvre et al., 2009; Martin et al., 2011). These characteristics, combined with the high level of recently reported diversity, highlight the need to continually revisit taxonomic classification of this viral group to add new species, genera and/or families. However, this task is complicated by the fact that many of the CRESS-DNA virus genomes exhibit novel genome architectures, only share similarities to the highly conserved Rep of known viruses, and have similarities to viruses belonging to multiple different taxonomic groups (Rosario et al., 2012a; Roux et al., 2013). In addition, the definitive hosts for many of these CRESS-DNA viruses remain unknown, hindering their classification according to traditional standards.

CRESS-DNA viruses are characterized by small genomes (~1.7–3 kb) that contain 2–6 protein-encoding genes. The smallest monopartite CRESS-DNA viruses, members of the *Circoviridae* family, exhibit only two major open reading frames (ORFs), which encode a Rep and a capsid protein (Cap). Many of the novel eukaryotic CRESS-DNA viral genomes obtained from environmental samples or individual organisms through either metagenomic sequencing or degenerate PCR (herein referred to as "metagenomic CRESS-DNA viruses") exhibit similarities to circoviruses and have been referred to as 'circo-like' viruses. Although many of the metagenomic circo-like virus genomes are highly divergent, these surveys have uncovered a novel CRESS-DNA viral group, the proposed Cyclovirus genus (Li et al., 2010a). Cycloviruses, which form a sister group to the *Circovirus* genus within the family *Circoviridae*, have been identified from both vertebrates (Li et al., 2010a; Smits et al., 2013; Tan Le et al., 2013; Garigliany et al., 2014; Zhang et al., 2014) and invertebrates (Rosario et al., 2011, 2012b; Dayaram et al., 2013, 2014, 2015b; Padilla-Rodriguez et al., 2013).

Similarities to circoviruses are mainly based on the Rep whereas the second major ORF in novel circo-like metagenomic CRESS-DNA viruses generally does not have any significant matches in the database but is assumed to encode for a structural protein based on the genomic architecture of known circoviruses. In lieu of significant matches to known structural proteins in the GenBank database, it is important to investigate putative novel Caps in CRESS-DNA viruses to provide evidence regarding their structural function. A potential avenue to identify conserved patterns in highly divergent structural proteins, such as those observed in novel metagenomic CRESS-DNA viruses, is to investigate the presence of predicted intrinsically disordered regions (IDRs). IDRs are regions within a protein that lack a rigid or fixed (i.e., ordered) structure, allowing a protein to exist in different states depending on the substrate with which it is interacting (Dunker et al., 2001; Brown et al., 2011). Research examining IDRs within viral proteomes has revealed that smaller viral genomes, such as those of CRESS-DNA viruses, contain a higher proportion of predicted disordered residues than larger viruses (Xue et al., 2012, 2014; Pushker et al., 2013). Therefore it has been suggested that small viruses may exploit IDRs to encode multifunctional proteins (Xue et al., 2012, 2014; Pushker et al., 2013). Since structural proteins in several viral families commonly contain IDRs (Chen et al., 2006; Goh et al., 2008a,b; Chang et al., 2009; Jensen et al., 2011), the presence of similar patterns of predicted disorder amongst unidentified CRESS-DNA proteins may provide one line of evidence for these proteins representing putative Caps.

To contribute to efforts exploring the diversity of CRESS-DNA viruses in invertebrates, this study investigated various marine invertebrate species for the presence of these viruses. A total of 27 novel CRESS-DNA genomes were recovered from 21 invertebrate species, expanding the known diversity of CRESS-DNA viruses associated with marine organisms and providing the first evidence of viruses associated with some under-sampled taxa. The well-conserved Rep of CRESS-DNA viruses was used to explore the relationships between these novel viruses and previously reported eukaryotic CRESS-DNA viruses in GenBank, including metagenomic CRESS-DNA viruses. In addition, the non-Rep-encoding ORFs (i.e., putative Caps) within these genomes were investigated for IDRs. Disorder prediction methods suggest that CRESS-DNA viral Caps exhibit conserved

patterns of predicted disorder, which can be used to complement similarity-based searches to identify structural proteins within novel CRESS-DNA viral genomes.

Materials and Methods

Sample Processing and Genome Discovery

CRESS-DNA viruses were investigated in a variety of marine invertebrate species that were collected as samples of opportunity (**Table 1** and Supplementary Table S1). Specimens were identified with the highest degree of taxonomic resolution possible based on morphology. Whole organisms or tissue sections were serially rinsed three times using sterile SM Buffer [0.1 M NaCl, 50 mM Tris-HCl (pH 7.5), 10 mM $MgSO_4$]. Viral particles were partially purified from each specimen prior to DNA extraction. For this purpose, samples were homogenized in one of two ways depending on the size of the specimen. Smaller organisms or dissected tissues that could be placed in a 1.5 ml microcentrifuge tube were homogenized in 1 ml of sterile SM Buffer through bead-beating using 1.0 mm sterile glass beads in a bead beater (Biospec Products). Homogenates were then centrifuged at 6000 × *g* for 6 min. Larger organisms or tissues of dissected organisms, such as muscle or gonads, were placed in a gentleMACS^{TM} M tube (Miltenyl Biotec) containing 3 ml of sterile SM buffer. Samples were then homogenized using a gentleMACS dissociator (Miltenyl Biotec) followed by centrifugation at 6000 × *g* for 9 min. The supernatant from both homogenization methods was filtered through a 0.45 μm Sterivex filter (Millipore) and nucleic acids were extracted from 200 μl of filtrate using the QIAmp MinElute Virus Spin Kit (Qiagen).

DNA extracts were amplified through RCA using the illustra TempliPhi Amplification kit (GE Healthcare) to enrich for small circular templates (Kim et al., 2008; Kim and Bae, 2011). RCA-amplified DNA was digested with a suite of FastDigest restriction enzymes (Life Technologies; BamHI, EcoRV, PdmI, HindIII, KpnI, PstI, XhoI, SmaI, BgiII, EcoRI, XbaI, and NcoI) following manufacturer's instructions in separate reactions to obtain complete, unit-length genomes for downstream cloning and sequencing. Restriction enzyme digested products were resolved on an agarose gel and bands ranging in size from 1000 to 4000 bp were excised and cleaned using the Zymoclean Gel DNA Recovery Kit (Zymo Research). Products resulting from blunt-cutting enzyme digestions were cloned using the CloneJET PCR Cloning kit (Life Technologies), whereas products containing sticky ends were cloned using pGEM-3Zf(+) vectors (Promega) pre-digested with the appropriate enzyme. All clones were commercially Sanger sequenced using vector primers and genomes exhibiting significant similarities to eukaryotic CRESS-DNA viruses were completed through primer walking.

Genome Annotation

Genomes were assembled using Sequencher 4.1.4 (Gene Codes Corporation). Putative ORFs >100 amino acids were identified and annotated using SeqBuilder version 11.2.1 (Lasergene). Partial genes or genes that seemed interrupted were analyzed for potential introns using GENSCAN (Burge and Karlin, 1997). The potential origin of replication (*ori*) for each genome was identified by locating a canonical nonanucleotide motif (NANTATTAC; Rosario et al., 2012a) and confirming predicted stem-loop structures using Mfold with constraints applied to prevent hairpin formation within the nonanucleotide motif and a folding temperature set at 17°C (Zuker, 2003). Final annotated genomes have been deposited to GenBank with accession numbers KR528543–KR528569.

Database Sequences and Sequence Analysis

To conduct sequence comparisons, members of the *Circovirus* genus, as well as complete eukaryotic CRESS-DNA viral genomes obtained from environmental samples or individual organisms through either metagenomic sequencing or degenerate PCR (herein referred to as "metagenomic CRESS-DNA viruses") were retrieved from GenBank. Since the Rep is the only conserved protein among CRESS-DNA viruses (Ilyina and Koonin, 1992; Rosario et al., 2012a) this protein was used to compare the different genomes. Rep pairwise identities were calculated using SDT v1.2 (Muhire et al., 2014) and summarized using heat maps generated in R (R Core Team, 2014). A maximum likelihood (ML) phylogenetic tree based on Rep amino acid sequences was also constructed. For this purpose, alignments were performed in MEGA 6.06 (Tamura et al., 2013) using the MUSCLE algorithm (Edgar, 2004) and manually edited. Sequences were inspected for the presence of conserved amino acid motifs that have been shown to play a role in rolling circle replication (RCR) of eukaryotic CRESS-DNA viruses, including three RCR and three superfamily 3 (SF3) helicase motifs (Gorbalenya et al., 1990; Ilyina and Koonin, 1992; Gorbalenya and Koonin, 1993; Rosario et al., 2012a). Although all the recently reported CRESS-DNA viruses are included in the heatmap, only sequences exhibiting all six motifs are included in the phylogenetic analysis. In addition, divergent regions that were poorly aligned, as shown by a high percentage of gaps, were removed from the alignment (Supplementary Data Sheet 1). Since the *Nanoviridae* and *Geminiviridae* are also CRESS-DNA viral families that are evolutionarily related to the *Circoviridae* (Ilyina and Koonin, 1992; Rosario et al., 2012a), select representatives of these families were included in the phylogenetic analysis. The ML phylogenetic tree was inferred using PHYML (Guindon et al., 2010) implementing the best substitution model (rtRev+I+G+F; Dimmic et al., 2002) according to ProtTest (Abascal et al., 2005). Branch support was assessed using the approximate likelihood ratio test (aLRT) SH-like method (Anisimova and Gascuel, 2006).

Intrinsically Disordered Region (IDR) Analysis of Putative Capsid Proteins

To determine if the non-Rep-encoding ORFs from the CRESS-DNA viral genomes presented here (n = 25), circoviruses (n = 15), and metagenomic CRESS-DNA viruses (n = 259; including 37 cycloviruses) represent putative Caps, these proteins were evaluated for IDRs. Disordered protein regions were predicted using the DisProt VL3 disorder predictor (Obradovic et al., 2003; Sickmeier et al., 2007). This artificial neural network utilizes an ensemble of feed forward neural networks with 20 attributes (18 amino acid frequencies, average flexibility,

TABLE 1 | CRESS-DNA genomes identified in this study, the organism they were obtained from, and genome details (acronym, genome length, nonanucleotide motif, genome type, and ORFs identified).

Genome[1]	Organism	Tissue type	Genome (bp)	Genomic architecture	Nonanucleotide[2]	Cap[3]	Rep
P. diogenes Giant Hermit Crab aCV(I0004A)	*Petrochirus diogenes*	Abdomen	1815	Type V	TAGTATTAC	X*	X
Palaemonete sp. Common Grass Shrimp aCV (I0006H)	*Palaemonete* sp.	Hepatopancreas	2257	Type II	TAGTATTAC	X*	X
Aiptasia sp. Sea Anemone aCV (I0007C2)	*Aiptasia* sp.	Whole organism	1901	Type I	CATTATTAC	X	X
Aiptasia sp. Sea Anemone aCV (I0007C3)	*Aiptasia* sp.	Whole organism	1942	Type I	CATTATTAC	X	X
L. variegatus Variable Sea Urchin aCV (I0021)	*Lytechinus variegatus*	Gonads	2167	Type III	GACTATTAC*	X*	X
Didemnum sp. Sea Squirt aCV (I0026A4)	*Didemnum* sp.	Whole organism	2061	Type IV	CAGTATTAC	X	X
Didemnum sp. Sea Squirt aCV (I0026A7)	*Didemnum* sp.	Whole organism	2143	Type I	CAGTATTAC	X*	X
Littorina sp. Snail aCV (I0041)	*Littorina* sp.	Whole organism	2237	Type II	CAGTATTAC	X	X
C. ornatus Ornate Blue Crab aCV (I0054)	*Callinectes ornatus*	Gonads	1241	Type I	CAGTATTAC	X	X
C. sapidus Atlantic Blue Crab aCV (I0056)	*Callinectes sapidus*	Gonads	1876	Type I	CAGTATTAC	X	X
P. intermedius Brackish Grass Shrimp aCV (I0059)	*Palaemonetes intermedius*	Whole organism	2293	Type I	CAGTATTAC	X*	X
F. duorarum Pink Shrimp aCV (I0066)	*Farfantepenaeus duorarum*	Whole organism	1799	Type I	CAGTATTAC	X	X
F. duorarum Pink Shrimp aCV (I0069)	*Farfantepenaeus duorarum*	Whole organism	1966	Type I	CAGTATTAC	X*	X
Marine Snail aCV (I0084)	Marine Snail	Whole organism	2305	Type I	TAGTATTAC	X*	X
Hermit Crab aCV (I0085A4)	Hermit Crab	Abdomen	2291	Type I	TAGTATTAC	X*	X
Hermit Crab aCV (I0085A5)	Hermit Crab	Abdomen	2291	Type I	TAGTATTAC	X*	X
Hermit Crab aCG (I0085b)	Hermit Crab	Abdomen	1063	Type VII	CAGTATTAC		X
Fiddler Crab aCV (I0086a)	Fiddler Crab	Gonads and claw muscle	1635	Type II	GATTATTAC	X	X
Fiddler Crab aCV (I0086b)	Fiddler Crab	Gonads and claw muscle	1511	Type V	AAGTATTAC	X	X
P. kadiakensis Mississippi Grass Shrimp aCV (I0099)	*Palaemonetes kadiakensis*	Whole organism	1895	N/A	None	X*	X
Gammarus sp. Amphipod aCV (I0153)	*Gammarus* sp.	Whole organism	1999	Type I	TAGTATTAC	X*	X
Mytilus sp. Clam aCV (I0169)	*Mytilus* sp.	Whole organism	1894	Type I	TAGTATTAC	X	X
Calanoida sp. Copepod aCV (I0298)	*Calanoida* sp.	Whole organism	2469	Type II	TAGTATTAC	X	X
A. melana Sponge aCG (I0307)	*Artemia melana*	Tissue segment	1826	Type VII	TAGTATTAC		X
P. pacifica Coral aCV (I0345)	*Primnoa pacifica*	Polyps	1240	N/A	None	X*	X
P. placomus Coral aCV (I0351)	*Paramuricea placomus*	Polyps	2292	Type II	TAGTATTAC	X*	X
S. brevirostris Brown Rock Shrimp aCV (I0722)	*Sicyonia brevirostris*	Gonads	1600	Type V	TAATATTAC*	X	X

[1]*Genome names contain abbreviation aCV for associated circular virus or aCG for associated circular genome. ID within parentheses corresponds to ID used throughout the paper.*
[2]*Nonanucleotide motif sequences that were not identified within a stem-loop structure are denoted with an asterisk (*).*
[3]*Non-Rep encoding ORFs were identified as putative capsid proteins based on BLAST results. However, many non-Rep-encoding ORFs did not exhibit any significant matches (marked with an asterisk*).*

and sequence complexity; Obradovic et al., 2003). Disorder disposition scores above a 0.5 threshold indicate intrinsic disorder. Counts and statistical analysis for the fraction of disorder- and order-promoting amino acid residues was conducted using R with the "seqinr" package (Charif and Lobry, 2007).

Results

A total of 27 CRESS-DNA genomes were recovered from 21 marine invertebrates (**Table 1**). Most of the recovered genomes (66.7%) were identified from *Crustacea*, mainly from the order *Decapoda*. Recovered genomes ranged in size from 1063 to

2469 nt and exhibited a variety of genome architectures. Of the 27 genomes identified, 23 exhibited a common putative *ori* marked by a conserved nonanucleotide motif (NANTATTAC) at the apex of a predicted stem-loop structure (**Table 1**). The remaining four genomes lacked a stem-loop structure ($n = 2$) or a stem-loop structure and a nonanucleotide motif ($n = 2$). Genomes lacking the canonical nonanucleotide motif could not be assigned to any genome type; therefore only 25 genomes were assigned to genomic architecture types previously described by Rosario et al. (2012a) (**Figure 1**). The predominant genomic architecture observed was Type I ($n = 13$), which is typical of members of the *Circovirus* genus. However, other genomic architectures were observed including Types II ($n = 5$), III ($n = 1$), IV ($n = 1$), V ($n = 3$), and VII ($n = 2$) (**Figure 1**). It is important to note that genomes exhibiting a Type VII genome architecture only exhibit a single major ORF encoding a Rep. This type of architecture is observed in genomic components of multipartite viruses from the *Nanoviridae* family and satellite DNA molecules that require helper viruses for encapsidation (Gronenborn, 2004; Briddon and Stanley, 2006). Therefore genomes exhibiting only a single major ORF may represent partial genomes of multipartite viruses or non-viral mobile genetic elements such as plasmids (Rosario et al., 2012a).

The majority of the CRESS-DNA viruses detected in marine invertebrates were most similar to viral sequences identified through metagenomic surveys of marine samples (Supplementary Table S1). However, one of genomes, *Lytechinus variegatus* variable sea urchin associated circular virus_I0021, was most similar to plant viruses from the *Geminiviridae* family. Most of the viral genomes had database similarities for the Rep; except for *Sicyonia brevirostris* brown rock shrimp associated circular virus_I0722, which only had similarities for the putative Cap (Supplementary Table S1). Similar to several previously described CRESS-DNA viruses (Li et al., 2010a; Rosario et al., 2012b; van den Brand et al., 2012; Sikorski et al., 2013b; Du et al., 2014; Ng et al., 2014; Dayaram et al., 2015a,b; Kraberger et al., 2015), three viral genomes (*Artemia melana* sponge associated circular virus_I0307, *Didemnum* sp. sea squirt associated circular virus_I0026_A7, and *Palaemonetes kadiakensis* Mississippi grass shrimp associated circular virus_I0099) exhibited Reps interrupted by introns (Supplementary Table S1).

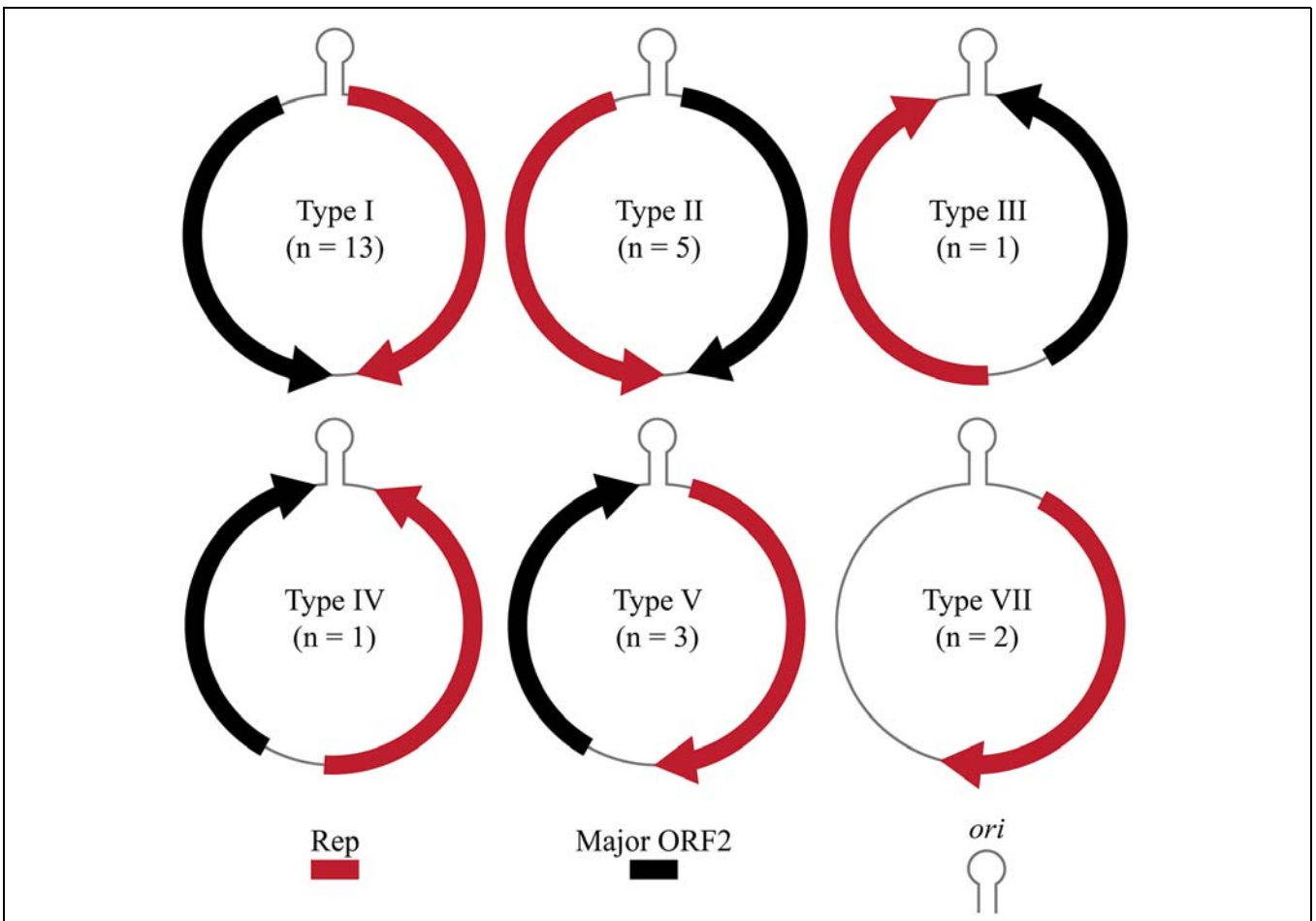

FIGURE 1 | Genome types of novel CRESS-DNA genomes identified in this study (Rosario et al., 2012a). Genome schematics illustrate a major ORF encoding the replication initiator protein (Rep), putative origin of replication (*ori*) marked by stem-loop structure, and a second major ORF.

Pairwise identities indicate that the CRESS-DNA viruses detected in marine invertebrates share less than 60.1% sequence identity (average sequence identity = 26.04%) with previously identified Reps from CRESS-DNA viruses in GenBank, indicating that these viruses represent novel species (**Figure 2**). Twenty-one of the 27 recovered Reps contained all six conserved RCR and helicase motifs (see Materials and Methods) and were used for phylogenetic analysis. Analysis of these Reps with representative

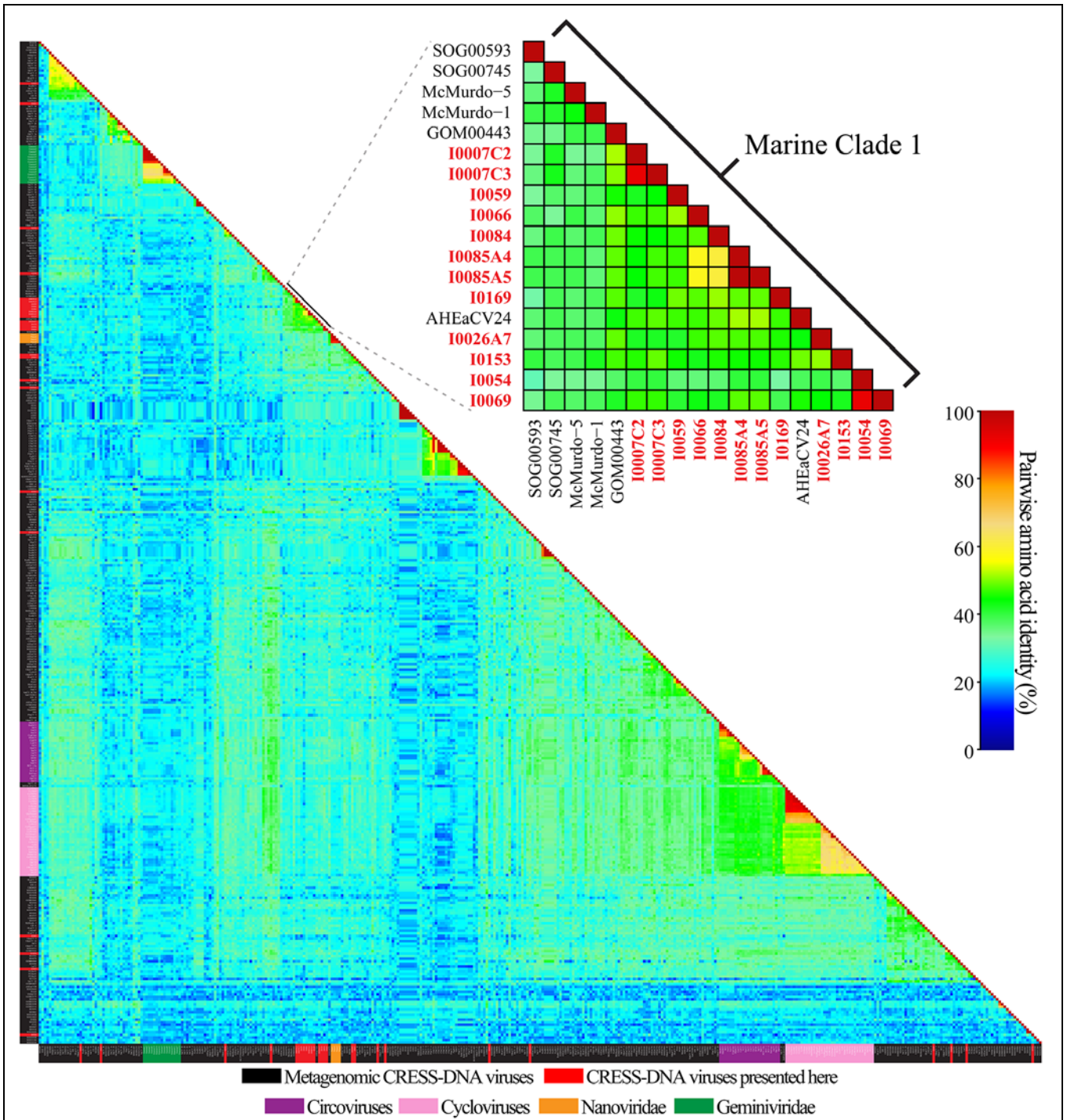

FIGURE 2 | Graphical representation of pairwise amino acid identities of the replication initiator proteins (Rep) from CRESS-DNA genomes from this study, metagenomic CRESS-DNA viruses, cycloviruses, circoviruses, and select members of the *Nanoviridae and Geminiviridae* families. Reps identified from this study within the Marine Clade 1 are in red font. Description of acronyms and the matrix used to generate the heatmap can be found in Supplementary Tables S2 and S3, respectively.

CRESS-DNA viral Reps from GenBank, including available metagenomic CRESS-DNA viral Reps, show that most of the sequences from marine invertebrate associated viruses detected here are more closely related to circo-like viruses recovered through metagenomic surveys of the marine environment than to previously defined CRESS-DNA viral groups (**Figure 3**). Eleven of the 21 Reps from marine invertebrate associated viruses do not form distinct clusters with each other or any known sequences (**Figure 3**). However, ten of the Reps form a well-supported clade that also includes sequences detected

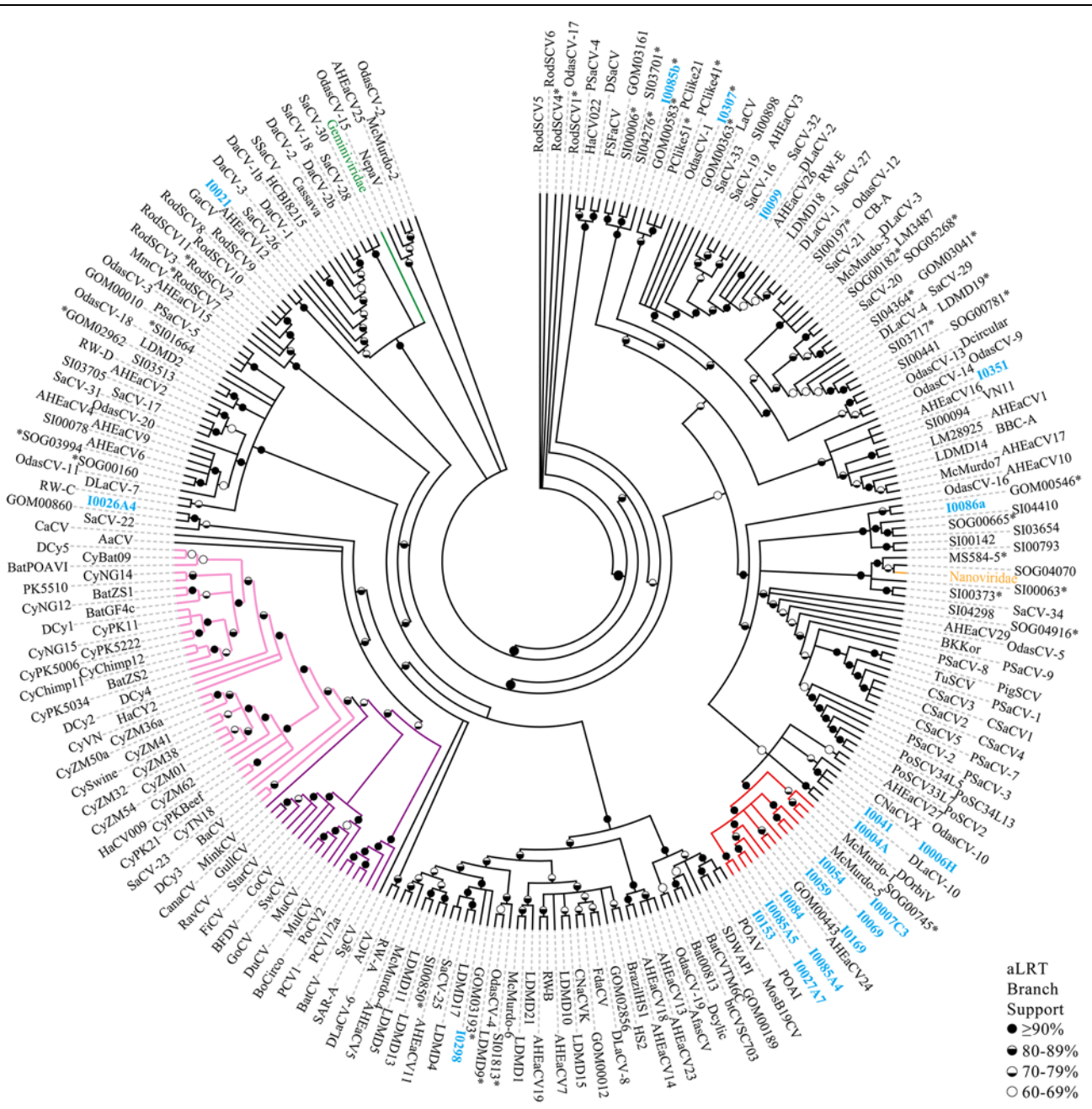

FIGURE 3 | Multifurcation maximum likelihood phylogenetic reconstruction based on the Reps of CRESS-DNA genomes recovered here, metagenomic CRESS-DNA viruses, cycloviruses, circoviruses, and representative members of the *Nanoviridae and Geminiviridae* families. Reps obtained from CRESS-DNA genomes obtained in this study are highlighted in blue font. Branches are colored for the different CRESS-DNA viral groups including the Marine Clade 1 (red), circoviruses (purple), cycloviruses (pink), nanoviruses (orange), and geminiviruses (green). Representative nanoviruses ($n = 4$) and geminiviruses ($n = 15$) have been condensed into their family names. Reps from genomes exhibiting a single ORF are highlighted using an asterisk (*). Branches with less than 60% aLRT branch support have been collapsed. Description of acronyms used can be found in Supplementary Table S4.

in the Gulf of Mexico (GOM00443; JX904231.1), Straight of Georgia (JX904106.1), McMurdo Ice Shelf (YP_009047125.1; YP_009047137.1), and a semi-enclosed shallow estuary (Avon-Heathcote Estuary associated circular virus 24; AJP36460.1). Pairwise identity scores indicate that all members of this clade, named Marine Clade 1 for the purposes of this study, share more than 32.7% identity, with an average pairwise identity score of 47.2% (**Figure 2**). Members of the Marine Clade 1 seem to be more closely related to members of the *Nanoviridae* (31.95% average pairwise identity) than any other known CRESS-DNA viral group; however, members of this clade exhibit different genomic architectures compared to these plant viruses. CRESS-DNA viral genomes from the Marine Clade 1 encode two major ORFs in an ambisense organization (i.e., Type I architecture), which is similar to members of the *Circoviridae*, rather than the single ORF, Type VII genome organization observed in genomic components from the *Nanoviridae*.

Capsid Analysis

Only half of the CRESS-DNA viral genomes described here contained an ORF that had significant BLASTX matches (e-value < 0.001; amino acid identities ranging from 26–54%) to proteins annotated as putative Caps in GenBank (**Table 1**). Furthermore, most of the matches in the database were to putative CRESS-DNA viral Caps detected through metagenomic surveys, which are not supported by biochemical data and have not necessarily been well curated. Therefore, alternative methods were explored to investigate non-Rep-encoding ORFs (i.e., putative Caps) found in CRESS-DNA viral genomes.

The majority of metagenomic CRESS-DNA viruses reported from marine invertebrates in this study and in GenBank are most similar to previously described circoviruses. Therefore, the predicted IDP profiles of well-characterized members of the *Circovirus* genus were examined in an effort to identify conserved patterns in structural proteins encoded by these viruses. These circovirus IDP profiles were then compared against profiles observed in cycloviruses (the proposed sister group to the circoviruses, which exhibit conserved features and share high identities with circoviruses) and other metagenomic CRESS-DNA viruses.

The DisProt VL3 disorder prediction analysis revealed that Caps encoded by members of the *Circovirus* genus ($n = 15$) exhibit one of two protein disorder profiles, distinguished here as Type A or Type B, based on the first 125 amino acids of these proteins (**Figure 4A**). Type A Caps exhibit IDP profiles that are predicted to have the highest degree of disorder closest to the N-terminus (i.e., amino acid residues 1–50) before the profile tapers to a structured region with variable predicted disorder. Type A Caps exhibit significant enrichment for amino acid residues that promote disorder (R, K, E, P, S, Q, and A) within the first 50 residues relative to amino acid residues 51–125 (ANOVA with *post hoc* Tukey's HSD; $p < 0.05$) and a depletion of order promoting amino acid residues (W, C, F, I, Y, V, L, and N) within the first 25 residues relative to amino acid residues 26–125 (ANOVA with *post hoc* Tukey's HSD; $p < 0.05$; **Figure 4B**). On the other hand, Type B Caps exhibit IDP profiles that peak in predicted disorder between amino acid residues 26–75. Type B Caps show an enrichment of disorder promoting residues between residue positions 26 through 75, whereas there is a depletion of predicted order promoting residues in this region compared to residues 1–25 and 76–125 (**Figure 4B**). Beyond 125 amino acids, IDP profiles exhibited more structured regions for both Types A and B Caps, with no distinguishable predicted disorder pattern (**Figure 4A**).

The overwhelming majority of Caps from the *Circovirus* genus (86.7%) exhibited Type A IDP profiles; however, two avian circoviruses, Finch circovirus (YP_803551.1) and Beak and feather disease virus (NP_047277.1), had Type B IDP profiles (**Table 2** and Supplementary Table S5). Similarly, 97.3% of cyclovirus putative Caps ($n = 37$) exhibited Type A IDP profiles. Comparison of IDP profiles showed that a majority of metagenomic CRESS-DNA viruses also contained patterns of increased predicted disorder at the N-terminus of the putative Cap, consistent with the *Circoviridae*. Interestingly, Type B IDP profiles were more prevalent among putative Caps from metagenomic CRESS-DNA viral genomes in GenBank (10.8%; $n = 222$) and the novel genomes reported in this study (56%; $n = 25$). Notably, 7 of the 10 viruses found in the Marine Clade 1 described here exhibit Type B Caps. Among the total 299 CRESS-DNA genome sequences analyzed, most putative Caps exhibit Type A IDP profiles (69.9%), followed by Type B (13%). Notably, most of the putative Caps lacking a significant match in the database exhibited one these profiles.

Discussion

Metagenomic studies have revealed a prodigious amount of diversity in eukaryotic CRESS-DNA viruses in the marine environment (Rosario et al., 2009a; Rosario and Breitbart, 2011; Labonte and Suttle, 2013; McDaniel et al., 2014). However, few studies have isolated these viruses directly from organisms. Building upon recent studies suggesting that CRESS-DNA viruses are associated with marine invertebrates (Dunlap et al., 2013; Hewson et al., 2013a,b; Ng et al., 2013; Pham et al., 2014; Soffer et al., 2014; Dayaram et al., 2015a), this study investigated a variety of marine invertebrates, including under sampled taxa, for the presence of these viruses. Viral genomes presented here were primarily recovered from *Crustacea,* suggesting that this subphylum harbors a rich diversity of CRESS-DNA viruses. This is consistent with previous research that identified CRESS-DNA viruses in copepods (Dunlap et al., 2013), which are the most abundant members of mesozooplankton (Kleppel et al., 1996), as well as different species of shrimp (Ng et al., 2013; Pham et al., 2014), which comprise some of the world's most important food sources (Goss et al., 2000; Paezosuna, 2003). In addition, this is the first study to report viruses associated with marine snails, anemones, sea squirts, and several crab species. Although a definitive host for these viruses cannot be assigned with the present data, this study reveals the need for further examination of viruses associated with common marine invertebrates and experiments to determine their potential impact, if any, on the ecology of these organisms. The grouping of the invertebrate-associated CRESS-DNA viruses reported here with metagenomic

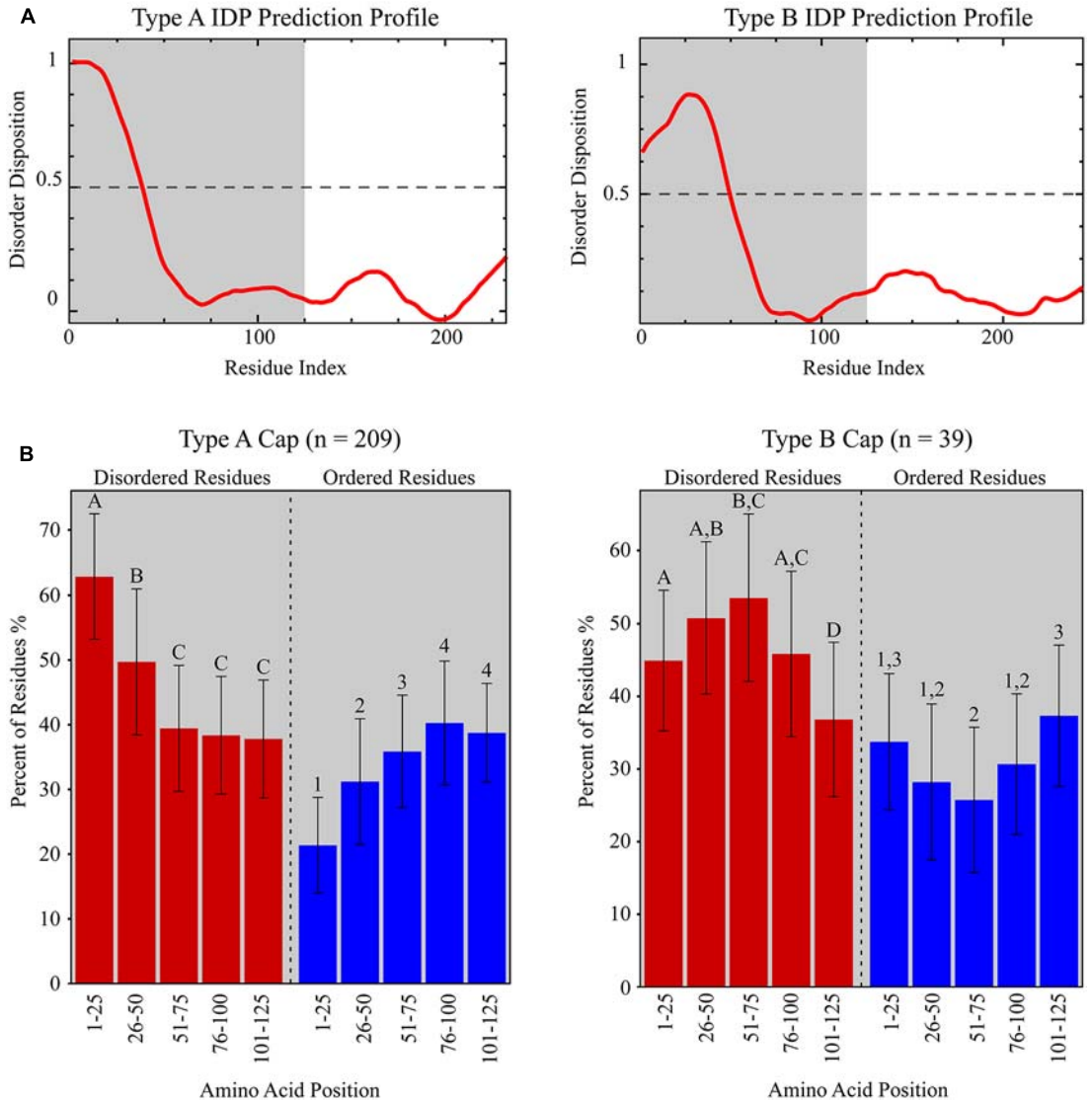

FIGURE 4 | (A) Representative IDP prediction profiles for Type A and Type B capsid proteins (Caps) from the Disprot VL3 predictor. Type A and Type B IDP prediction profiles are based on the Porcine circovirus 2 Cap (NP_937957.1) and the Beak and feather disease virus Cap (NP_047277.1), respectively. The grey shaded area represents the amino acid residue interval used in **(B)**. **(B)** Graphs showing the fraction of predicted disordered (red bars) and ordered (blue bars) residues within discrete amino acid intervals for Type A and Type B Caps identified from all CRESS-DNA viral genomes analyzed in this study. Significantly different amino acid intervals for each Cap type are distinguished using letters ("A", "B", "C", "D" for statistics based on percentage of predicted disordered residues) or numbers ("1", "2", "3", "4" for statistics based on percentage of predicted ordered residues; ANOVA with *post hoc* Tukey's HSD; $p < 0.05$). Note that the percentage of predicted disordered and ordered residues does not add to 100% due to the presence of residues that are not considered either disordered or ordered (i.e., H, M, T, and D).

CRESS-DNA viruses implies that marine invertebrates may serve as hosts for many of the sequences obtained from marine environments.

The marine invertebrate associated CRESS-DNA viruses identified here are only distantly related to known members of the *Circoviridae* and may represent novel groups. Approximately one third of the novel sequences reported here belong to the Marine Clade 1, whose members share an average pairwise identity of 47.2%. Members of this viral clade share an average pairwise identity score of 27.5% with members of the *Circoviridae*, whose members (genus *Circovirus* and proposed genus Cyclovirus) share 48.9% average pairwise identity. Although members of the Marine Clade 1 share slightly higher average pairwise identity with the *Nanoviridae* (31.2%), their genome architecture is clearly distinct from these plant-infecting viruses. Therefore, genomic architectures and comparative Rep analyses suggest that members of the Marine Clade 1 may represent a novel CRESS-DNA viral family.

TABLE 2 | Intrinsically disordered protein (IDP) profile types identified in non-Rep encoding ORFs of CRESS-DNA viruses.

Group	Total sequences	IDP Cap type		
		Type A	Type B	No type
Circoviruses	15	86.7%	13.3%	0.0%
Cycloviruses	37	97.3%	0.0%	2.7%
Metagenomic CRESS-DNA viruses	222	67.6%	10.8%	21.6%
This study	25	40.0%	56.0%	4.0%
Total	299	69.9%	13.0%	17.1%

The highly conserved Rep enables its straightforward identification through similarity-based searches; however, there is currently no reliable method for characterizing highly divergent putative Caps for metagenomic CRESS-DNA viruses. Since many of the novel metagenomic CRESS-DNA viruses are most similar to members of the *Circoviridae*, which only contain two major ORFs encoding a Rep and Cap, the putative Cap is often assigned simply based on the conserved genome architectures exhibited by this group.

This study investigated the IDP profiles of all available circo-like CRESS-DNA viruses to evaluate if putative Caps exhibit conserved patterns that could be used to identify this structural protein even in the absence of significant similarities in the database. The Cap of Porcine circovirus 2 represents a Type A IDP profile and that of Beak and feather disease virus represents a Type B IDP profile. Since the non-Rep-encoding ORF for both of these circoviruses have been shown to be structural (Nawagitgul et al., 2000; Patterson et al., 2013), this provides evidence that both the Type A and Type B IDP profiles represent a Cap. These Cap IDP profiles may be driven by the arginine and/or lysine rich region at the N-terminus of the Cap (Niagro et al., 1998), as both of these amino acids are considered disorder-promoting residues by the DisProt VL3 neural network. In addition to characterizing IDP profiles of circo-like CRESS-DNA viruses, analysis of select *Geminiviridae* and *Nanoviridae* Caps demonstrated that these viruses also exhibit Type A and Type B IDP profiles (Supplementary Table S5). Although further research into these plant virus families is needed, these findings suggest that the IDP patterns identified here may be conserved across Caps from the different families of eukaryotic CRESS-DNA viruses.

Thirteen of the eukaryotic CRESS-DNA viruses presented here had a non-Rep-encoding ORF without any database similarities, which were characterized as a putative Cap based on IDP profiles. Likewise, hypothetical proteins from 32 metagenomic CRESS-DNA viruses were identified as putative Caps using this method (Supplementary Table S5). While the Caps in the database were dominated by Type A IDP profiles, the majority of the new marine invertebrate associated genomes presented here exhibited Type B IDP profiles. In addition, 50 of the CRESS-DNA genomes analyzed here (17.1%; $n = 299$), including the *Primnoa pacifica* coral associated circular virus I0345 identified here, contained a non-Rep-encoding ORF that did not exhibit either the Type A or Type B profile. While it is possible that other IDP profiles representative of novel Caps exist, caution should be used in annotating these ORFs as putative Caps without supporting evidence. Finally, while examining metagenomic sequences annotated as CRESS-DNA viruses in GenBank, numerous genomes were identified that only contained a single ORF, which encoded a Rep. These sequences (Supplementary Table S5), along with the two Type VII genomes found in this study, most likely represent partial viral genomes [i.e., a single component of a multipartite virus (Gutierrez, 1999; Gronenborn, 2004)], satellite DNA molecules (Briddon and Stanley, 2006), or non-viral mobile genetic elements (Rosario et al., 2012a). Genomes exhibiting a single ORF cannot be distinguished phylogenetically from complete viral genomes based on the Rep (**Figure 3**). Therefore, it is important to investigate complete genomes of CRESS-DNA viruses rather than partial sequences.

The IDP analysis has interesting implications for understanding the evolutionary pressures acting upon the Rep and Cap of CRESS-DNA viruses, which include the smallest known eukaryotic viral pathogens. Small viruses exhibit a higher proportion of predicted disordered residues than larger viruses and may exploit IDRs to encode multifunctional proteins (Xue et al., 2012, 2014; Pushker et al., 2013). Rep proteins encoded by CRESS-DNA viruses exhibited low disposition for predicted disorder promoting amino acid residues or an inconsistency in predicted disorder patterns (data not shown), while the Caps consistently exhibited profiles with increased predicted disorder at the N-terminus, suggesting that the high proportion of predicted disordered regions in these small viruses may be driven by the Cap. IDRs have a tendency to evolve more rapidly than structured regions (Brown et al., 2002, 2011; Chen et al., 2006; Bellay et al., 2011; Nilsson et al., 2011; van der Lee et al., 2014); consequently, IDRs may hinder our ability to perform phylogenetic reconstructions based on the Cap. Although we are unable to perform reliable Cap alignments, the ability to classify these proteins within CRESS-DNA virus genomes due to conserved predicted disorder profiles reveals that these viruses exhibit regions in which disorder is conserved despite rapidly evolving amino acids (i.e., flexible disorder; van der Lee et al., 2014).

Although the functional significance of predicted IDP profiles detected in this study has yet to be determined, the identification of conserved IDP profiles may prove useful to identify divergent structural proteins encoded by CRESS-DNA viruses. The identification of a given IDP profile (Type A or B) for a putative ORF in a genomic context may allow the recognition of novel CRESS-DNA viral structural proteins that cannot be identified by standard BLAST searches. The IDP profile analysis needs to be complemented by other genomic features that are characteristic of CRESS-DNA viruses, including the presence of a Rep exhibiting RCR and helicase motifs and a putative *ori* marked by a conserved nonanucleotide motif (NANTATTAC) at the apex of a stem-loop structure. Future work needs to evaluate if the high proportion of IDRs observed in CRESS-DNA viruses and other small viruses is indeed mainly driven by structural proteins.

If this observation is validated, IDP profile analysis of hypothetical proteins may provide a reliable tool to identify structural proteins encoded by small viruses.

Acknowledgments

We acknowledge Ian Hewson, Renee Bishop-Pierce, Christina Kellogg, Robert W. Thacker, Stan Rice, Sandra Gilchrist, Brandan Cole, Brittany Hall, Ernst Peebles, Ralph Kitzmiller, Scott Burghart, and Elise Pickett for sample donations. We thank Bin Xue for his guidance in the intrinsically disordered protein analysis. This work was funded through grant DEB-1239976 from the National Science Foundation's Assembling the Tree of Life Program to KR and MB.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00696

References

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104–2105. doi: 10.1093/bioinformatics/bti263

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi: 10.1371/journal.pbio.0040368

Anisimova, M., and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst. Biol.* 55, 539–552. doi: 10.1080/10635150600755453

Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B. J., et al. (2011). Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* 12, R14. doi: 10.1186/gb-2011-12-2-r14

Blinkova, O., Victoria, J., Li, Y., Keele, B. F., Sanz, C., Ndjango, J. B., et al. (2010). Novel circular DNA viruses in stool samples of wild-living chimpanzees. *J. Gen. Virol.* 91, 74–86. doi: 10.1099/vir.0.015446-0

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Briddon, R. W., and Stanley, J. (2006). Subviral agents associated with plant single-stranded DNA viruses. *Virology* 344, 198–210. doi: 10.1016/j.virol.2005.09.042

Brown, C. J., Johnson, A. K., Dunker, A. K., and Daughdrill, G. W. (2011). Evolution and disorder. *Curr. Opin. Struct. Biol.* 21, 441–446. doi: 10.1016/j.sbi.2011.02.005

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., et al. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* 55, 104–110. doi: 10.1007/s00239-001-2309-6

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Chang, C. K., Hsu, Y. L., Chang, Y. H., Chao, F. A., Wu, M. C., Huang, Y. S., et al. (2009). Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J. Virol.* 83, 2255–2264. doi: 10.1128/JVI.02001-08

Charif, D., and Lobry, J. R. (2007). *Seqin{R} 1.0-2: a Contributed Package to the {R} Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis*. New York: Springer Verlag.

Chen, J. W., Romero, P., Uversky, V. N., and Dunker, A. K. (2006). Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J. Proteome Res.* 5, 879–887. doi: 10.1021/pr060048x

Cheung, A. K., Ng, T. F., Lager, K. M., Alt, D. P., Delwart, E. L., and Pogranichniy, R. M. (2014). Unique circovirus-like genome detected in pig feces. *Genome Announc.* 2:e00251-14. doi: 10.1128/genomeA.00251-14

Cheung, A. K., Ng, T. F., Lager, K. M., Bayles, D. O., Alt, D. P., Delwart, E. L., et al. (2013). A divergent clade of circular single-stranded DNA viruses from pig feces. *Arch. Virol.* 158, 2157–2162. doi: 10.1007/s00705-013-1701-z

Dayaram, A., Galatowitsch, M., Harding, J. S., Arguello-Astorga, G. R., and Varsani, A. (2014). Novel circular DNA viruses identified in *Procordulia grayi* and *Xanthocnemis zealandica* larvae using metagenomic approaches. *Infect. Genet. Evol.* 22, 134–141. doi: 10.1016/j.meegid.2014.01.013

Dayaram, A., Goldstien, S., Arguello-Astorga, G. R., Zawar-Reza, P., Gomez, C., Harding, J. S., et al. (2015a). Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infect. Genet. Evol.* 31, 284–295. doi: 10.1016/j.meegid.2015.02.010

Dayaram, A., Potter, K. A., Pailes, R., Marinov, M., Rosenstein, D. D., and Varsani, A. (2015b). Identification of diverse circular single-stranded DNA viruses in adult dragonflies and damselflies (Insecta: Odonata) of Arizona and Oklahoma, USA. *Infect. Genet. Evol.* 30, 278–287. doi: 10.1016/j.meegid.2014.12.037

Dayaram, A., Potter, K. A., Moline, A. B., Rosenstein, D. D., Marinov, M., Thomas, J. E., et al. (2013). High global diversity of cycloviruses amongst dragonflies. *J. Gen. Virol.* 94, 1827–1840. doi: 10.1099/vir.0.052654-0

Delwart, E. L. (2007). Viral metagenomics. *Rev. Med. Virol.* 17, 115–131. doi: 10.1002/rmv.532

Delwart, E., and Li, L. (2012). Rapidly expanding genetic diversity and host range of the *Circoviridae* viral family and other Rep encoding small circular ssDNA genomes. *Virus Res.* 164, 114–121. doi: 10.1016/j.virusres.2011.11.021

Diemer, G. S., and Stedman, K. M. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Bio. Dir.* 7, 1–14. doi: 10.1186/1745-6150-7-13

Dimmic, M. W., Rest, J. S., Mindell, D. P., and Goldstein, R. A. (2002). rtRev: An amino acid substition matrix for inference of retrovirus and reverse transcriptase phylogeny. *J. Mol. Evol.* 55, 65–73. doi: 10.1007/s00239-001-2304-y

Du, Z., Tang, Y., Zhang, S., She, X., Lan, G., Varsani, A., et al. (2014). Identification and molecular characterization of a single-stranded circular DNA virus with similarities to *Sclerotinia sclerotiorum* hypovirulence-associated DNA virus 1. *Arch. Virol.* 159, 1527–1531. doi: 10.1007/s00705-013-1890-5

Duffy, S., and Holmes, E. C. (2009). Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J. Gen. Virol.* 90, 1539–1547. doi: 10.1099/vir.0.009266-0

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* 9, 267–276. doi: 10.1038/nrg2323

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Jeong, S. O., et al. (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* 19, 26–59. doi: 10.1016/S1093-3263(00)00138-8

Dunlap, D. S., Ng, T. F., Rosario, K., Barbosa, J. G., Greco, A. M., Breitbart, M., et al. (2013). Molecular and microscopic evidence of viruses in marine copepods. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1375–1380. doi: 10.1073/pnas.1216595110

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edwards, R. A., and Rohwer, F. (2005). Viral metagenomics. *Nat. Rev. Microbiol.* 3, 504–510. doi: 10.1038/nrmicro1163

Garigliany, M. M., Borstler, J., Jost, H., Badusche, M., Desmecht, D., Schmidt-Chanasit, J., et al. (2015). Characterization of a novel circo-like virus in *Aedes vexans* mosquitoes from Germany: evidence for a new genus within the family Circoviridae. *J. Gen. Virol.* 96, 915–920. doi: 10.1099/vir.0.000036

Garigliany, M. M., Hagen, R. M., Frickmann, H., May, J., Schwarz, N. G., Perse, A., et al. (2014). Cyclovirus CyCV-VN species distribution is not limited to Vietnam and extends to Africa. *Sci. Rep.* 4, 7552. doi: 10.1038/srep07552

Ge, X., Li, Y., Yang, X., Zhang, H., Zhou, P., Zhang, Y., et al. (2012). Metagenomic analysis of viruses from bat fecal samples reveals many novel

viruses in insectivorous bats in China. *J. Virol.* 86, 4620–4630. doi: 10.1128/JVI.06671-11

Goh, G. K., Dunker, A. K., and Uversky, V. N. (2008a). Protein intrinsic disorder toolbox for comparative analysis of viral proteins. *BMC Genomics* 9(Suppl. 2):S4. doi: 10.1186/1471-2164-9-S2-S4

Goh, G. K., Dunker, A. K., and Uversky, V. N. (2008b). A comparative analysis of viral matrix proteins using disorder predictors. *Virol. J.* 5, 126. doi: 10.1186/1743-422X-5-126

Gorbalenya, A. E., and Koonin, E. V. (1993). Helicases: amino acid sequence comparisons and structure-function relationships. *Curr. Opin. Struct. Biol.* 3, 419–429. doi: 10.1016/S0959-440X(05)80116-2

Gorbalenya, A. E., Koonin, E. V., and Wolf, Y. I. (1990). A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. *FEBS Lett.* 262, 145–148. doi: 10.1016/0014-5793(90)80175-I

Goss, J., Burch, D., and Rickson, R. E. (2000). Agri-food restructuring and third world transnationals: Thailand, the CP Group and the global shrimp industry. *World Dev.* 28, 513–530. doi: 10.1016/S0305-750X(99)00140-0

Gronenborn, B. (2004). Nanoviruses: genome organisation and protein function. *Vet. Microbiol.* 98, 103–109. doi: 10.1016/j.vetmic.2003.10.015

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Gutierrez, C. (1999). Geminivirus DNA replication. *Cell. Mol. Life Sci.* 56, 313–329. doi: 10.1007/s000180050433

Hewson, I., Eaglesham, J. B., Höök, T. O., Labarre, B. A., Sepúlveda, M. S., Thompson, P. D., et al. (2013a). Investigation of viruses in *Diporeia* spp. from the Laurentian Great Lakes and Owasco Lake as potential stressors of declining populations. *J. Great Lakes Res.* 39, 499–506. doi: 10.1016/j.jglr.2013.06.006

Hewson, I., Ng, G., Li, W., Labarre, B. A., Aguirre, I., Barbosa, J. G., et al. (2013b). Metagenomic identification, seasonal dynamics, and potential transmission mechanisms of a *Daphnia*-associated single-stranded DNA virus in two temperate lakes. *Limnol. Oceanogr.* 58, 1605–1620. doi: 10.4319/lo.2013.58.5.1605

Ilyina, T. V., and Koonin, E. V. (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res.* 20, 3279–3285. doi: 10.1093/nar/20.13.3279

Jensen, M. R., Communie, G., Ribeiro, E. A. Jr., Martinez, N., Desfosses, A., Salmon, L., et al. (2011). Intrinsic disorder in measles virus nucleocapsids. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9839–9844. doi: 10.1073/pnas.1103270108

Kim, K. H., and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77, 7663–7668. doi: 10.1128/AEM.00289-11

Kim, K. H., Chang, H. W., Nam, Y. D., Roh, S. W., Kim, M. S., Sung, Y., et al. (2008). Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl. Environ. Microbiol.* 74, 5975–5985. doi: 10.1128/AEM.01275-08

Kleppel, G. S., Burkart, C. A., Carter, K., and Tomas, C. (1996). Diets of calanoid copepods on the West Florida continental shelf: relationships between food concentration, food composition and feeding activity. *Mar. Biol.* 127, 209–217. doi: 10.1007/BF00942105

Kraberger, S., Arguello-Astorga, G. R., Greenfield, L. G., Galilee, C., Law, D., Martin, D. P., et al. (2015). Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. *Infect. Genet. Evol.* 31, 73–86. doi: 10.1016/j.meegid.2015.01.001

Labonte, J. M., and Suttle, C. A. (2013). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* 7, 2169–2177. doi: 10.1038/ismej.2013.110

Lefeuvre, P., Lett, J. M., Varsani, A., and Martin, D. P. (2009). Widely conserved recombination patterns among single-stranded DNA viruses. *J. Virol.* 83, 2697–2707. doi: 10.1128/JVI.02152-08

Li, L., Kapoor, A., Slikas, B., Bamidele, O. S., Wang, C., Shaukat, S., et al. (2010a). Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. *J. Virol.* 84, 1674–1682. doi: 10.1128/JVI.02109-09

Li, L., Victoria, J. G., Wang, C., Jones, M., Fellers, G. M., Kunz, T. H., et al. (2010b). Bat guano virome: predominance of dietary viruses from insects and plants plus novel mammalian viruses. *J. Virol.* 84, 6955–6965. doi: 10.1128/JVI.00501-10

Lian, H., Liu, Y., Li, N., Wang, Y., Zhang, S., and Hu, R. (2014). Novel circovirus from mink, China. *Emerging Infect. Dis.* 20, 1548–1550. doi: 10.3201/eid2009.140015

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A., and Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* 326, 858–861. doi: 10.1126/science.1179287

Martin, D. P., Biagini, P., Lefeuvre, P., Golden, M., Roumagnac, P., and Varsani, A. (2011). Recombination in eukaryotic single stranded DNA viruses. *Viruses* 3, 1699–1738. doi: 10.3390/v3091699

McDaniel, L. D., Rosario, K., Breitbart, M., and Paul, J. H. (2014). Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environ. Microbiol.* 16, 570–585. doi: 10.1111/1462-2920.12184

Muhire, B. M., Varsani, A., and Martin, D. P. (2014). SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE* 9:e108277. doi: 10.1371/journal.pone.0108277

Nawagitgul, P., Morozov, I., Bolin, S. R., Harms, P. A., Sorden, S. D., and Paul, P. S. (2000). Open reading frame 2 of porcine circovirus type 2 encodes a major capsid protein. *J. Gen. Virol.* 81, 2281–2287. doi: 10.1099/0022-1317-81-9-2281

Ng, T. F., Alavandi, S., Varsani, A., Burghart, S., and Breitbart, M. (2013). Metagenomic identification of a nodavirus and a circular ssDNA virus in semi-purified viral nucleic acids from the hepatopancreas of healthy *Farfantepenaeus duorarum* shrimp. *Dis. Aquat. Org.* 105, 237–242. doi: 10.3354/dao02628

Ng, T. F., Chen, L. F., Zhou, Y., Shapiro, B., Stiller, M., Heintzman, P. D., et al. (2014). Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16842–16847. doi: 10.1073/pnas.1410429111

Ng, T. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., et al. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* 86, 12161–12175. doi: 10.1128/JVI.00869-12

Ng, T. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., et al. (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* 6:e20579. doi: 10.1371/journal.pone.0020579

Niagro, F. D., Forsthoefel, A. N., Lawther, R. P., Kamalanathan, L., Ritchie, B. W., Latimer, K. S., et al. (1998). Beak and feather disease virus and porcine circovirus genomes: intermediates between the geminiviruses and plant circoviruses. *Arch. Virol.* 143, 1723–1744. doi: 10.1007/s007050050412

Nilsson, J., Grahn, M., and Wright, A. P. (2011). Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol.* 12, R65. doi: 10.1186/gb-2011-12-7-r65

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., and Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins* 53(Suppl. 6), 566–572. doi: 10.1002/prot.10532

Padilla-Rodriguez, M., Rosario, K., and Breitbart, M. (2013). Novel cyclovirus discovered in the Florida woods cockroach *Eurycotis floridana* (Walker). *Arch. Virol.* 158, 1389–1392. doi: 10.1007/s00705-013-1606-x

Paezosuna, F. (2003). Shrimp aquaculture development and the environment in the Gulf of California ecoregion. *Mar. Pollut. Bull.* 46, 806–815. doi: 10.1016/S0025-326X(03)00107-3

Patterson, E. I., Swarbrick, C. M., Roman, N., Forwood, J. K., and Raidal, S. R. (2013). Differential expression of two isolates of beak and feather disease virus capsid protein in *Escherichia coli*. *J. Virol. Methods* 189, 118–124. doi: 10.1016/j.jviromet.2013.01.020

Pham, H. T., Bergoin, M., and Tijssen, P. (2013a). Acheta domesticus volvovirus, a novel single-stranded circular DNA virus of the house cricket. *Genome Announc.* 1:e00079-13. doi: 10.1128/genomeA.00079-13

Pham, H. T., Iwao, H., Bergoin, M., and Tijssen, P. (2013b). New volvovirus isolates from *Acheta domesticus* (Japan) and *Gryllus assimilis* (United States). *Genome Announc.* 1:e00328-13. doi: 10.1128/genomeA.00328-13

Pham, H. T., Yu, Q., Boisvert, M., Van, H. T., Bergoin, M., and Tijssen, P. (2014). A circo-like virus isolated from *Penaeus monodon* shrimps. *Genome Announc.* 2:e01172-13. doi: 10.1128/genomeA.01172-13

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L., and Delwart, E. L. (2011). The fecal flora of wild rodents. *PLoS Pathog.* 7:e1002218. doi: 10.1371/journal.ppat.1002218

Phan, T. G., Mori, D., Deng, X., Rajindrajith, S., Ranawaka, U., Fan Ng, T. F., et al. (2015). Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. *Virology* 482, 98–104. doi: 10.1016/j.virol.2015.03.011

Pushker, R., Mooney, C., Davey, N. E., Jacque, J. M., and Shields, D. C. (2013). Marked variability in the extent of protein disorder within and between viral families. *PLoS ONE* 8:e60724. doi: 10.1371/journal.pone.0060724

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reavy, B., Swanson, M. M., Cock, P., Dawson, L., Freitag, T. E., Singh, B. K., et al. (2015). Distinct circular ssDNA viruses exist in different soil types. *Appl. Environ. Microbiol.* 81, 3934–3945. doi: 10.1128/AEM.03878-14

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Rosario, K., Duffy, S., and Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J. Gen. Virol.* 90, 2418–2424. doi: 10.1099/vir.0.012955-0

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y., and Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* 11, 2806–2820. doi: 10.1111/j.1462-2920.2009.01964.x

Rosario, K., Duffy, S., and Breitbart, M. (2012a). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch. Virol.* 157, 1851–1871. doi: 10.1007/s00705-012-1391-y

Rosario, K., Dayaram, A., Marinov, M., Ware, J., Kraberger, S., Stainton, D., et al. (2012b). Diverse circular single-stranded DNA viruses discovered in dragonflies (Odonata: Epiprocta). *J. Gen. Virol.* 93, 2668–2681. doi: 10.1099/vir.0.045948-0

Rosario, K., Marinov, M., Stainton, D., Kraberger, S., Wiltshire, E. J., Collings, D. A., et al. (2011). Dragonfly cyclovirus, a novel single-stranded DNA virus discovered in dragonflies (Odonata: Anisoptera). *J. Gen. Virol.* 92, 1302–1308. doi: 10.1099/vir.0.030338-0

Roux, S., Enault, F., Bronner, G., Vaulot, D., Forterre, P., and Krupovic, M. (2013). Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat. Commun.* 4, 2700. doi: 10.1038/ncomms3700

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* 7:e33641. doi: 10.1371/journal.pone.0033641

Sachsenroder, J., Twardziok, S., Hammerl, J. A., Janczyk, P., Wrede, P., Hertwig, S., et al. (2012). Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing. *PLoS ONE* 7:e34631. doi: 10.1371/journal.pone.0034631

Sasaki, M., Orba, Y., Ueno, K., Ishii, A., Moonga, L., Hang'ombe, B. M., et al. (2015). Metagenomic analysis of the shrew enteric virome reveals novel viruses related to human stool-associated viruses. *J. Gen. Virol.* 96, 440–452. doi: 10.1099/vir.0.071209-0

Sickmeier, M., Hamilton, J. A., Legall, T., Vacic, V., Cortese, M. S., Tantos, A., et al. (2007). DisProt: the database of disordered proteins. *Nucleic Acids Res.* 35, D786–D793. doi: 10.1093/nar/gkl893

Sikorski, A., Dayaram, A., and Varsani, A. (2013a). Identification of a novel circular DNA virus in New Zealand fur seal (*Arctocephalus forsteri*) fecal matter. *Genome Announc.* 1:e00558-13. doi: 10.1128/genomeA.00558-13

Sikorski, A., Massaro, M., Kraberger, S., Young, L. M., Smalley, D., Martin, D. P., et al. (2013b). Novel myco-like DNA viruses discovered in the faecal matter of various animals. *Virus Res.* 177, 209–216. doi: 10.1016/j.virusres.2013.08.008

Smits, S. L., Schapendonk, C. M., Van Beek, J., Vennema, H., Schurch, A. C., Schipper, D., et al. (2014). New viruses in idiopathic human diarrhea cases, the Netherlands. *Emerging Infect. Dis.* 20, 1218–1222. doi: 10.3201/eid2007.140190

Smits, S. L., Zijlstra, E. E., Van Hellemond, J. J., Schapendonk, C. M., Bodewes, R., Schurch, A. C., et al. (2013). Novel cyclovirus in human cerebrospinal fluid, Malawi, 2010-2011. *Emerging Infect. Dis.* 19, 1511. doi: 10.3201/eid1909.130404

Soffer, N., Brandt, M. E., Correa, A. M., Smith, T. B., and Thurber, R. V. (2014). Potential role of viruses in white plague coral disease. *ISME J.* 8, 271–283. doi: 10.1038/ismej.2013.137

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tan Le, V., Van Doorn, H. R., Nghia, H. D., Chau, T. T., Tu Le, T. P., De Vries, M., et al. (2013). Identification of a new cyclovirus in cerebrospinal fluid of patients with acute central nervous system infections. *mBio* 4:e00231-13. doi: 10.1128/mbio.00231-13

van den Brand, J. M., Van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D., et al. (2012). Metagenomic analysis of the viral flora of pine marten and European badger feces. *J. Virol.* 86, 2360–2365. doi: 10.1128/JVI.06373-11

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., et al. (2014). Classification of intrinsically disordered regions and proteins. *Chem. Rev.* 114, 6589–6631. doi: 10.1021/cr400525m

Whon, T. W., Kim, M. S., Roh, S. W., Shin, N. R., Lee, H. W., and Bae, J. W. (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J. Virol.* 86, 8221–8231. doi: 10.1128/JVI.00293-12

Xue, B., Blocquel, D., Habchi, J., Uversky, A. V., Kurgan, L., Uversky, V. N., et al. (2014). Structural disorder in viral proteins. *Chem. Rev.* 114, 6880–6911. doi: 10.1021/cr4005692

Xue, B., Dunker, A. K., and Uversky, V. N. (2012). Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* 30, 137–149. doi: 10.1080/07391102.2012.675145

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T., and Takai, K. (2013). Metagenomic analysis of viral communities in (hado)pelagic sediments. *PLoS ONE* 8:e57271. doi: 10.1371/journal.pone.0057271

Zawar-Reza, P., Arguello-Astorga, G. R., Kraberger, S., Julian, L., Stainton, D., Broady, P. A., et al. (2014). Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infect. Genet. Evol.* 26, 132–138. doi: 10.1016/j.meegid.2014.05.018

Zhang, W., Li, L., Deng, X., Kapusinszky, B., Pesavento, P. A., and Delwart, E. (2014). Faecal virome of cats in an animal shelter. *J. Gen. Virol.* 95, 2553–2564. doi: 10.1099/vir.0.069674-0

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415. doi: 10.1093/nar/gkg595

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 01 September 2015
doi: 10.3389/fmicb.2015.00896

Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota

Guanhui Bao[1†], Mingjie Wang[1†], Thomas G. Doak[2,3] and Yuzhen Ye[1]*

[1] School of Informatics and Computing, Indiana University, Bloomington, IN, USA, [2] Department of Biology, Indiana University, Bloomington, IN, USA, [3] National Center for Genome Analysis Support, Indiana University, Bloomington, IN, USA

OPEN ACCESS

Edited by:
Roy D. Sleator,
Cork Institute of Technology, Ireland

Reviewed by:
Suleyman Yildirim,
Istanbul Medipol University International School of Medicine, Turkey
Joseph Wade,
New York State Department of Health, USA

***Correspondence:**
Yuzhen Ye,
School of Informatics and Computing, Indiana University, 150 South Woodlawn Avenue, Bloomington, IN 47405, USA
yye@indiana.edu

[†]These authors have contributed equally to this work.

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *28 April 2015*
Accepted: *17 August 2015*
Published: *01 September 2015*

Citation:
Bao G, Wang M, Doak TG and Ye Y (2015) Strand-specific community RNA-seq reveals prevalent and dynamic antisense transcription in human gut microbiota. Front. Microbiol. 6:896. doi: 10.3389/fmicb.2015.00896

Metagenomics and other meta-omics approaches (including metatranscriptomics) provide insights into the composition and function of microbial communities living in different environments or animal hosts. Metatranscriptomics research provides an unprecedented opportunity to examine gene regulation for many microbial species simultaneously, and more importantly, for the majority that are unculturable microbial species, in their natural environments (or hosts). Current analyses of metatranscriptomic datasets focus on the detection of gene expression levels and the study of the relationship between changes of gene expression and changes of environment. As a demonstration of utilizing metatranscriptomics beyond these common analyses, we developed a computational and statistical procedure to analyze the antisense transcripts in strand-specific metatranscriptomic datasets. Antisense RNAs encoded on the DNA strand opposite a gene's CDS have the potential to form extensive base-pairing interactions with the corresponding sense RNA, and can have important regulatory functions. Most studies of antisense RNAs in bacteria are rather recent, are mostly based on transcriptome analysis, and have been applied mainly to single bacterial species. Application of our approaches to human gut-associated metatranscriptomic datasets allowed us to survey antisense transcription for a large number of bacterial species associated with human beings. The ratio of protein coding genes with antisense transcription ranges from 0 to 35.8% (median = 10.0%) among 47 species. Our results show that antisense transcription is dynamic, varying between human individuals. Functional enrichment analysis revealed a preference of certain gene functions for antisense transcription, and transposase genes are among the most prominent ones (but we also observed antisense transcription in bacterial house-keeping genes).

Keywords: metatranscriptome, metagenome, antisense RNA, human gut microbiota, transposases

Introduction

Advances in sequencing technology have catalyzed the development of metagenomics, which has revolutionized many fields in the study of microbial organisms. Metagenomics has been applied to study microbial communities sampled from various environments and animal hosts (including humans). Several large-scale efforts worth mentioning are the early global ocean surveys

(Nealson and Venter, 2007; Rusch et al., 2007), and more recent MetaHit (Qin et al., 2010) and the NIH Human Microbiome Project (HMP; Human Microbiome Project Consortium, 2012a,b; thanks to which the composition of the human microbiome is now well-studied). The research emphasis now has shifted toward elucidating the functionality and regulatory mechanisms of the microbial communities using other meta-omics approaches, including metatranscriptomics and metaproteomics.

Metatranscriptomics research is creating an unprecedented opportunity to gain knowledge about gene regulation for many microbial species simultaneously, and more importantly, for the vast majority of uncultured microbial species in their natural environments (or hosts). In addition to elucidating functional characteristics of microbial communities, metatranscriptomic data provides information vital for accurate annotations of genes and their regulation in their community—complementary to metagenomic sequencing. Metatranscriptomic data indicate which of the genes encoded in a metagenome are actually transcribed, and which metabolic pathways are active (and the level of their activities), on the basis of their transcripts within a microbial community under various environmental conditions.

Current analyses of metatranscriptomic datasets have largely been limited to the detection of gene expression levels and the relationship between gene expression (and functions and pathways involved) and changes in environmental conditions (de Menezes et al., 2012; Leimena et al., 2013; Franzosa et al., 2014; Jorth et al., 2014; Coolen and Orsi, 2015). However, metatranscriptomics datasets contain rich information, which can be utilized to address important questions, when powered with appropriate computational and statistical approaches. For example, antisense RNAs (asRNAs; Jens and Wolfgang, 2011), which are encoded on the DNA strand opposite to a protein coding (sense) gene transcript (so may play important regulatory roles by forming extensive base-pairing interactions with the corresponding sense RNA), can be revealed by strand-specific metatranscriptomic sequences.

In a standard metatranscriptomic study (using the RNA-seq protocol), total RNA is isolated from the sample, ribosomal RNAs are removed to enrich for mRNA, which is then reverse transcribed into cDNA and subjected to DNA sequencing, using next generation sequencing (NGS) platforms (Giannoukos et al., 2012). It is important to remove the ribosomal RNAs during the process, otherwise the majority of reads from a metatranscriptomic project are rRNA (He et al., 2010). Early metatranscriptomic methods lacked strand specificity, limiting the application of metagenomic datasets in elucidating some regulatory mechanisms in bacteria. However, Giannoukos et al. (2012) presented a protocol for metatranscriptomic analysis of bacterial communities that accommodates both intact and fragmented RNA and combines efficient rRNA removal with strand-specific RNA-seq. Currently, only a handful of such metatranscriptomic datasets are available (and metaproteomic datasets are even scarcer), but we envision a flood of strand-specific RNA-seq metatranscriptomic data in the near future, as experimental techniques mature (Giannoukos et al., 2012; Franzosa et al., 2014).

Antisense RNAs encoded on the DNA strand opposite a gene have the potential to form extensive base-pairing interactions with the corresponding sense RNA (Thomason and Storz, 2010). Unlike other—smaller—regulatory RNAs in bacteria, antisense RNAs range in size from 10 to 1000s of nucleotides, complementary to part of a gene, a complete gene or a group of genes in an operon (Beiter et al., 2009). Although antisense RNAs were first observed in bacteria in the early 1980s (Lacatena and Cesareni, 1981) and their regulatory roles were defined in model systems (Green et al., 1986), most studies of antisense RNAs in bacteria are rather recent. Many antisense RNAs were identified using genome-wide searches for sRNAs and from transcriptome analysis, and have been studied mainly for single bacterial species. The numbers of antisense RNAs reported for different bacteria vary extensively, but 100s have been suggested in some species (Thomason and Storz, 2010). For example, 1,005 antisense RNAs (22% of all genes) were reported for *Escherichia coli* (Dornenburg et al., 2010). Massive antisense transcription was observed for *Synechocystis* PCC6803, with 26.8% of its genes reported to have antisense transcription (Mitschke et al., 2011), and genome-wide antisense transcription was observed in *Helicobacter pylori* (Sharma et al., 2010). Many species have less antisense transcription: for example, only 1.3% of the genes in *Staphylococcus aureus* were reported to have antisense transcription (Beaume et al., 2010). Thomason and Storz (2010) noted in their review that the existence of antisense RNAs was not tested for in many studies.

The availability of human-associated strand-specific metatranscriptomics datasets allows us to examine the antisense transcriptions for a large number of microbial species growing in their natural communities. In this paper we developed computational and statistical approaches to identify antisense transcripts from human gut-associated microbial species and study their dynamics among different human individuals.

Materials and Methods

Dataset

We used the human gut-associated strand-specific metatranscriptomic data from (Franzosa et al., 2014); the datasets were downloaded from the SRA website (SRA accession: SRR769395-SRR769540). In total, we analyzed eight sets of metatranscriptomic datasets; each set contains three metatranscriptomic datasets derived from the same human individual, but prepared using three different methods of sample preservation (frozen, ethanol-fixed, or RNAlater-fixed; Franzosa et al., 2014). The eight individuals are X310763260 (abbreviated as X1), X311245214 (X2), X316192082 (X3), X316701492 (X4), X317690558 (X5), X317802115 (X6), X317822438 (X7), and X319146421 (X8).

Bacterial reference genomes (including the genomic sequences and gene annotations) were downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/genomes/bacteria/). We focused on 116 reference genomes (covering 47 species), which were reported as the main species found in stool samples (Franzosa et al., 2014). For some analyses, including the function

enrichment analysis, we selected a representative strain for each species with multiple strains, to limit the biases that may be introduced by the uneven sampling of the species. See **Data Sheet 1** for the list of 116 strains, and the list of 47 representative strains and the basic information about the genome (e.g., the number of genes found in each genome).

Identification of Sense and Antisense Reads

Raw reads were trimmed with Trimmomatic 0.33 (Bolger et al., 2014) to remove adapter sequences and low quality reads and the trimmed reads were mapped to the 116 bacterial genomes with Bowtie 2 (Langmead and Salzberg, 2012). For simplicity, we call a read that maps to the sense strand a *sense read*, and a read mapped to the antisense strand of a gene an *antisense read*. We used featureCounts twice on the same dataset with the strand setting reversed (-s 1 and then -s 2) to annotate sense and antisense reads (Liao et al., 2014): featureCounts counts mapped reads for genomic features including genes, promoters, gene bodies, and chromosomal locations (given in an input annotation file) and outputs the number of reads assigned to each feature.

We summarize the antisense transcription at both read and gene levels. For each species, we computed two ratios: the *ratio of antisense reads* (out of all reads that can be mapped to the protein coding genes in this species), and the *ratio of genes with antisense transcription* (see below for the detection of genes with antisense transcription using sequencing data).

Detection of Antisense Expression by a Binomial Test

Artifacts introduced by cDNA synthesis and amplification are known problems for antisense RNA detection (Thomason and Storz, 2010), so even for a gene with no actual antisense transcription, we may find reads suggesting antisense transcripts (i.e., the strandedness of RNA-seq data is <100%). To overcome this problem, we use binomial testing to detect genes with antisense transcripts that are unlikely to be the results of such artifacts: let p be the probability of having reads from the antisense strand of a gene, even though there is no real antisense transcription from the gene. A total of c reads are sequenced from the gene (c is approximated as the number of reads that can be mapped to the gene), among which m reads represent antisense transcript. The null hypothesis is that there is no antisense transcription from this gene. We use the binomial test in R (binom.test) to calculate the probability of having c antisense reads (the number of successes) out of m trials (a total of m reads) with a success rate of p. If the probability is low ($\leq$ 0.05 according to one-tailed binomial test), we consider that the gene has antisense transcription (the alternative hypothesis).

Since p (the success rate) is usually unknown for metatranscriptomic datasets (but it was shown that most library treatments in RNA-seq have a strandedness of >95% Sigurgeirsson et al., 2014), we use the lowest ratio of antisense reads from individual bacterial species present in the microbial communities to approximate the p (considering that the strandedness of the RNA-seq will be at least this good). For the human-gut metatranscriptomics datasets we tested, p is 0.01. Using this probability of success, we identified significant antisense transcription for different bacterial species using binomial tests. We also checked which species recruited the most RNA-seq reads (to their protein coding genes), as compared to other species in the eight individuals; their ratios of antisense reads are: 0.0233 and 0.0312, for *Methanobrevibacter smithii* ATCC 35061 in X2 (individual 2) and X8, respectively; 0.0481, 0.0626, and 0.0347 for *Parabacteroides distasonis* ATCC 8503 in X1, X4, and X7, respectively; 0.0296 for *Ruminococcus bromii* in X3; and 0.0078 and 0.0167 for *B. vulgatus* ATCC 8482 in X5 and X6, respectively. Seven out of these eight ratios are <5% (two are close to 1%), consistent with the reported strandedness of most stranded library methods in RNA-seq (>95%; Sigurgeirsson et al., 2014). Thus, we believe that 5% (i.e., strandedness of 95%) is a generous estimate of p for the data sets we used, and we also used this p to provide a more conservative estimate of the genes with antisense transcription in the data sets we analyzed for comparison purposes.

Functional Enrichment Analysis of Genes with Antisense Expression

Functional enrichment analysis was conducted using two different tests for Clusters of Orthologous Groups (COG; Tatusov et al., 1997). We used the representative set of strains (47 in total) for this analysis, and gene annotations for their genomes were downloaded from the NCBI ftp website.

A one-tailed binomial test with Benjamini-Hochberg (BH) false discovery rate (FDR) correction ($q \leq 0.05$) was first used to determine if a COG was significantly enriched in the set of genes with antisense expression. The frequency of a COG among all the COGs for a bacterial genome was used as the hypothesized probability of success for the test. In the subset of genes detected to have antisense expression, the number of occurrences of a COG is considered the number of successes, and the total number of detected genes with antisense expression was used as the number of trials. To ensure the binomial test was conducted in a sufficiently large sample, we only tested genomes with $\geq$ 30 genes with antisense expression. For example, 71 out of 2,204 protein coding genes from *Bacteroides salanitronis* DSM 18170 were detected to have antisense transcription, and 10 out 33 genes that belong to COG4974L were detected to have antisense transcription. Here the number of successes, the number of trials, and the probability of success are 10, 71, and 0.0032 (33/2204), respectively. By the binomial test, the p-value was computed to be 1.14e-07, which was then corrected for multiple testing. This resulted in a q-value of 6.25e-06, indicating a significant enrichment of COG4974L among genes with antisense expression in this species.

For the enrichment analysis, we noted the binomial test with FDR correction penalized heavily for COGs with few genes. Therefore, we also investigated the association between COG family and antisense expression by a one-tailed Fisher's exact test with BH FDR correction ($q \leq 0.05$). For the example above (*B. salanitronis* DSM 18170), the 2 by 2 contingency table is [(10, 23), (61, 2110)] and the q-value

was calculated to be 1.78e-06, also indicating the enrichment of COG4974L in genes with antisense transcription in the genome.

Results

Sample Preservation Method Matters for the Detection of Antisense Transcription from Metatranscriptomic Sequences

Franzosa et al. (2014) used three different methods for preserving samples (frozen, ethanol-fixed, or RNAlater-fixed) for metatranscriptomics sequencing. They showed that measurements of microbial species, gene, and gene transcript composition within self-collected samples were consistent across sampling methods (Franzosa et al., 2014). We first asked if this consistency applied to antisense transcription.

We aligned the eight sets of stool metatranscriptome data against the bacterial reference genomes reported as the main species found in stool samples (Franzosa et al., 2014). For each sample handling method, we computed a profile of antisense transcription, in which a number represents the ratio of genes with antisense transcription in one species in one human individual. To limit the bias that may be introduced by uneven sampling of the strains and species with few RNA-seq reads, we only used one strain for each species, and only kept the ratios calculated for species with at least 100 genes supported by RNA-seq reads in an individual prepared by all three experimental methods (see Human Gut-Associated Microbial Organisms have a Wide Range of Antisense Transcription). In total 196 ratios for each handling method were included for the analysis. Our results show that all three sample-handling approaches result in highly correlated profiles of antisense transcription, with the frozen samples and the RNAlater-fixed samples sharing the most similar profiles (Pearson's $r = 0.84$; two tailed p-value $< 2.2e\text{-}16$) and RNAlater-fixed samples and ethanol-fixed samples sharing the least similarity (Pearson's $r = 0.71$; two tailed p-value $< 2.2e\text{-}16$). However, differences in the profiles are also obvious, as shown in the comparison between the profiles from ethanol-fixed samples and frozen samples (**Figure 1**).

A two-way ANOVA test of antisense transcriptions of all eight individuals by the three different experimental approaches showed that the handling method has the strongest effect on the antisense transcription ($F = 7.05$, p-value $= 0.001$), followed by the individuals ($F = 2.88$, p-value $= 0.007$), and the interaction between handling methods and individuals ($F = 1.89$, p-value $= 0.03$). A Turkey HSD test further revealed significant differences between frozen samples and ethanol-fixed (p-value $= 0.0043$), and between RNAlater-fixed samples and ethanol-fixed (p-value $= 0.0068$; but not between frozen and ethanol-fixed samples). This result suggests that we be cautious with results based on metatranscriptomic datasets

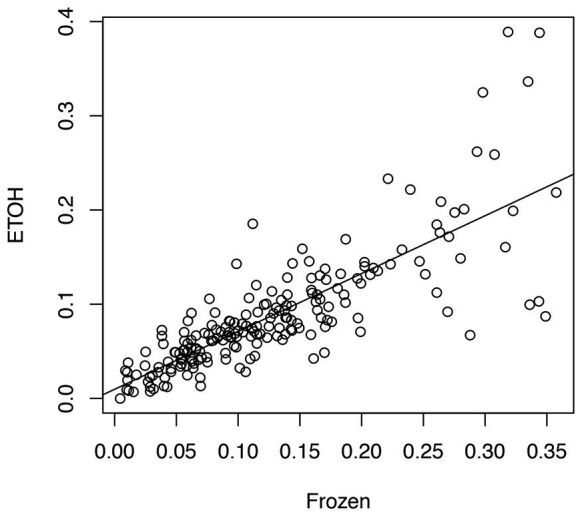

FIGURE 1 | Impacts of different experimental protocols on the profiling of antisense transcription. The figure shows the correlation (and differences) between the profile of antisense transcription (i.e., ratios of genes with antisense transcription in different species across individuals) for frozen samples and the profile of antisense transcription for ethanol-fixed samples (Pearson's $r = 0.71$; two tailed p-value $< 2.2e\text{-}16$). In this plot, the ratios of genes with antisense transcription for the different species based on the frozen samples (Frozen) and ethanol-fixed samples (ETOH) are plotted in the x-axis and y-axis, respectively.

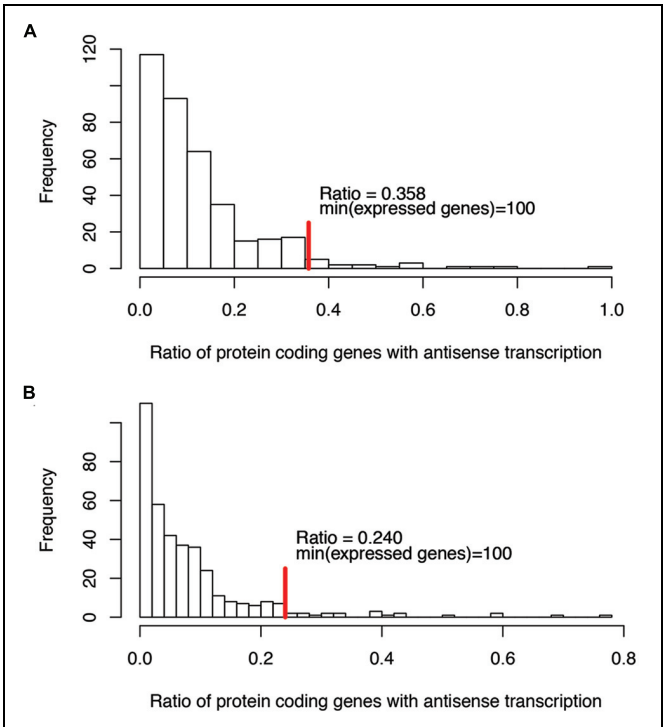

FIGURE 2 | Histograms of the ratios of genes with antisense transcription (over all genes with detectable transcription). Binomial tests were used to determine if a gene has antisense transcription or not, with a success rate of 1% **(A)**, and 5% **(B)**, respectively. The red vertical lines indicate the maximum ratios of genes with antisense transcription, for the genomes in which at least 100 genes have detectable transcription (with RNA-seq reads support).

derived from differently preserved samples (although high correlations were observed among these different approaches as shown in **Figure 1**). In addition, the previous publication reported that ethanol-fixed and RNAlater-fixed approaches can cause some artifacts in some functional genes (Franzosa et al., 2014). Considering both, we used the metatranscriptomics datasets generated from frozen samples for all our below analyses.

Human Gut-Associated Microbial Organisms have a Wide Range of Antisense Transcription

We detected antisense transcription for most of the species we tested. For each species, we computed the ratio of antisense reads (over total reads mapped to the species) and the ratio of genes with antisense transcript (over all genes with detectable transcription; see Materials and Methods). We used datasets derived from all eight individuals (and the ratios for the same species are most likely different in different datasets). The ratios of antisense reads and genes with antisense transcription for all the 116 bacterial strains (covering 47 species) across the samples (from eight individuals), along with other details (such as the total number of mapped reads, antisense reads, etc.), are listed in **Data Sheets 1** and **2** in the Supplementary Material.

For ratios of genes with antisense transcription, we noticed that some species have extremely high ratios (see the long tails in **Figure 2**; we only considered one strain for each species to reduce the bias that may be introduced by multiple strains belonging to the same species for the histograms), and without exception, all these species have few expressed genes (e.g., with <100 of their genes having detectable transcription). Considering that species with few supporting RNA-seq reads tend to be influenced heavily by potential artifacts (due to ambiguous reads mapping, bad gene annotations, etc.), we only considered species with at least 100 of their genes supported by RNA-seq reads, to infer the range of genes with antisense transcription. The ratio of protein coding genes with antisense transcription ranges from 0 to 35.8% (median = 10.0%; **Figure 2A**), based on the binomial tests using a success rate of 1%; the range drops, to between 0 and 24.0% (median = 6.3%; **Figure 2B**) when the more generous estimate of the success rate (5%, indicating a 95% strandedness of the RNA-seq experiments) was used for the binomial testing. In the following, results are based on binomial testing using p of 1%, unless stated otherwise.

Ratios of antisense reads (over total reads mapped to protein coding genes) are generally smaller than ratios of genes with antisense transcription. Similar to the inference of ratios of genes, only species with at least 100 of their genes supported by RNA-seq reads in a dataset were used to infer the range of ratios of antisense reads. **Figure 3** shows the boxplot for the ratios of reads mapped to the antisense strands of protein coding genes: the 95% confidence interval is 0.35–16.3% and the median is 2.5%. The boxplot revealed a few ratios that are significantly higher than the

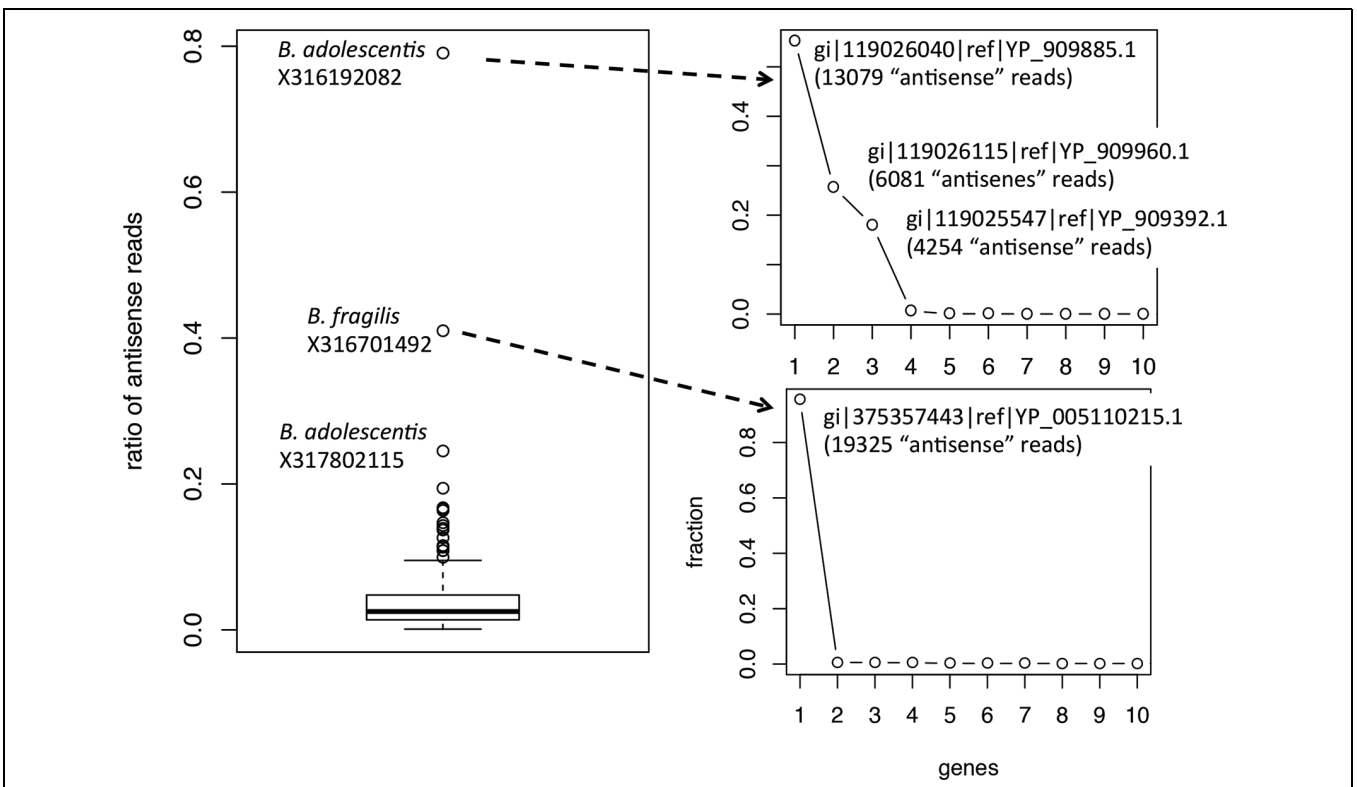

FIGURE 3 | Boxplot of the ratios of antisense reads (left). The plots on the right show the contribution of individual genes to the total antisense reads for the two outliers *B. adolescentis* and *B. fragilis*. Most of the antisense reads came from three and one genes (likely misannotations) for *B. adolescentis* and *B. fragilis*, respectively.

remaining: including the ratio for *B. adolescentis* in individual X316192082, the ratio for *B. fragilis* in individual X316701492, and the ratio for *B. adolescentis* in individual X317802115. As shown in **Figure 3**, for these outliers, most of the "antisense" reads are from a few putative genes (three genes in *B. adolescentis*; and one in *B. fragilis*) that have recruited large numbers of RNA-seq reads; all are hypothetical protein coding genes encoding small proteins without detailed functional annotation (except for gene gi| 119026115|ref| YP_909960.1 in *B. adolescentis*, which was annotated as a DEAD helicase in NCBI annotation; however, searching this protein against the Pfam database revealed no hits). We suspect that these genes are likely ncRNA genes, instead of protein coding genes, and therefore these few large ratios of antisense reads need to be interpreted with caution.

Not surprisingly, most of the strains we tested recruited many more sense than antisense reads, and tend to have more genes with sense transcription than genes with antisense transcription (such as *B. vulgatus* ATCC 8482, as shown in **Figures 4A,B**, *Parabacteroides distasonis* ATCC 8503 as shown in **Figures 4C,D**, and *M. smithii* ATCC 35061 as shown in **Figures 4E,F**). Our results are consistent with a previous study (Franzosa et al., 2014), showing that *M. smithii* is abundant and

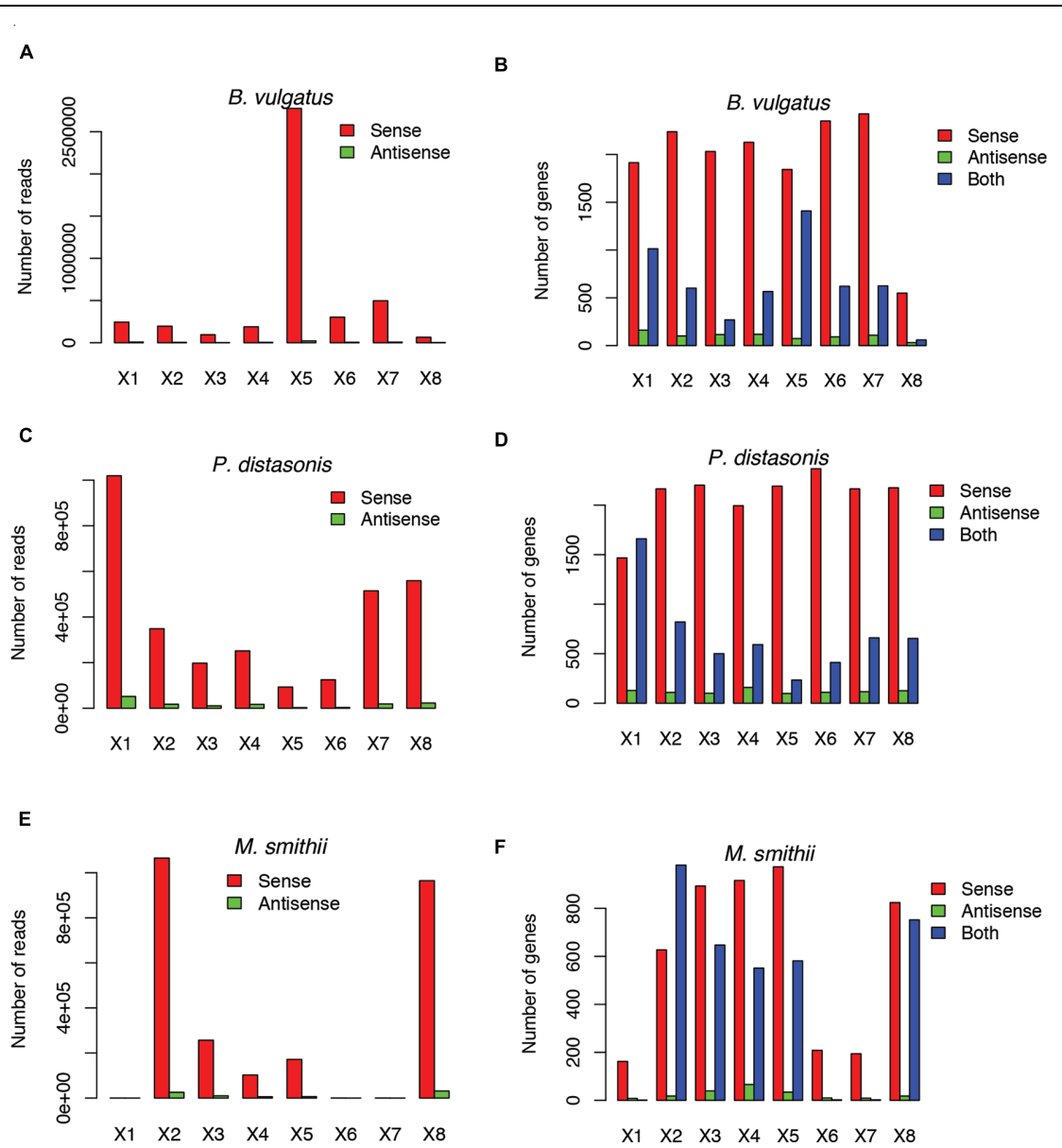

FIGURE 4 | Example species with antisense transcription in different individuals. Three species are shown: *Bacteroides vulgatus* ATCC 8482 **(A,B)**, *Parabacteroides distasonis* ATCC 8503 **(C,D)** and *Methanobrevibacter smithii* ATCC 35061 **(E,F)**. **(A,C,E)** Shows the numbers of sense and antisense reads in these three species, and **(B,D,F)** show the number of genes with sense transcription only (Sense), antisense transcription only (Antisense), and both (Both). X1–X8 indicate the eight individuals.

highly transcriptionally active (supported by huge numbers of RNA-seq reads) in five of the eight individuals (**Figures 4E,F**). But for these species, individual genes may still have significant antisense transcription or even have antisense transcription only; for examples, **Figure 5A** shows the read coverage plot for an operon in *B. vulgatus* (the operon information was extracted from the Database of Prokaryotic Operons; Mao et al., 2009), showing that all four genes in this operon have both sense and antisense transcription; and **Figure 5B** shows that gene BVU_3334 (which encodes for a putative transcriptional regulator) only has antisense reads.

Different species of the same genus showed various ratios of antisense transcripts. **Figure 6** shows the ratio of genes with antisense transcription in different species of *Streptococcus* (one of the dominant genera in human gut microbiota) across the eight human individuals. Overall, *Streptococcus* species have relatively low antisense transcription: the median of the ratios of antisense reads is 1.1% and the median of the ratios of genes with antisense transcription is 4.4%. *S. mutans* and *S. parasanguinis* have the lowest ratio of genes with antisense transcription; other *Staphylococcus* species seem to have higher antisense transcription, but the ratios vary greatly across different individuals. Similar trends are observed in a plot that shows the ratios of antisense reads for these species (Supplementary Figure S1).

Genes with either Sense- or Antisense-Dominating Transcription are Typically Highly Expressed

We can roughly group genes into three categories: genes with mostly sense transcripts, genes with mostly antisense transcripts, and genes in between, based on their sense and antisense transcription. We define $d = (\#\text{sense reads} - \#\text{antisense reads})/(\#\text{sense} + \#\text{antisense reads})$, so that genes with mostly sense transcripts have d that is close to 1, while genes with mostly antisense transcripts have d that is close to -1. **Figure 7** shows the plot of gene expression levels versus the d ratios, using expressed genes from 23 species (each having at least 100 genes with detectable expression), based on the RNA-seq dataset of individual 1 (X310763260; see Supplementary Figure S2 for the plot using all 47 strain; only one strain was included for each species). We used FPKM (Fragments Per Kilobase of transcript per Million mapped reads; Garber et al., 2011) to quantify the gene expression levels, to normalize read counts by the gene length and sequencing depth of the RNA-seq experiments. The number of mapped reads for a dataset was computed as the total number of reads that can be mapped to one of the 116 strains. The plot reveals a "U" shape, indicating that genes with either sense- or antisense-dominated transcription are typically highly expressed, while genes in between have relatively low gene expression. This

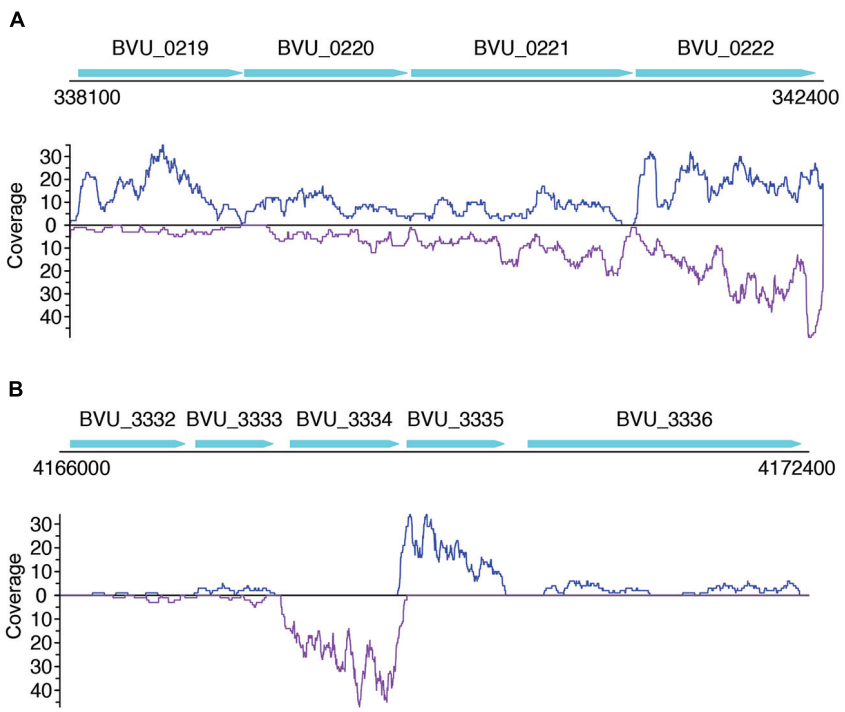

FIGURE 5 | Read coverage plots for example genes in *B. vulgatus*. Genes are represented as arrows in the plots, and the read coverage curves are shown below the genes, with the coverage for sense and antisense reads shown in blue and purple, respectively. **(A)** Read coverage plot for an operon with four genes, shown as cyan arrows on the top: BVU_0219 is a putative aldo/keto reductase, BVU_0220 is a hypothetical protein, BVU_0221 is a putative fucose permease, and BVU_0222 is a putative sorbitol dehydrogenase. **(B)** Read coverage plot for BVU_3334 (and its neighboring genes): BVU_3334 is a putative transcriptional regulator, BVU_3333 is similar to a fructose-6-phosphate aldolase from *E. coli*, BVU_3332 is a putative ABC transporter permease, BVU_3335 is a hypotentical protein, and BVU_3336 is a putative glycosyl transferase.

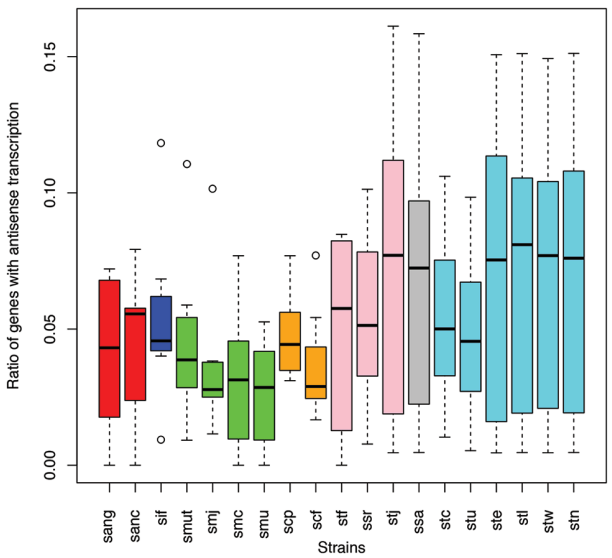

FIGURE 6 | Different *Streptococcus* species have different levels of antisense transcripts. The y-axis shows the ratio of genes with antisense transcription. The x-axis shows the different species; sang: *S. anginosus* C1051, sanc: *S. anginosus* C238, sif: *S. infantarius* CJ18, smut: *S. mutans* GS5, smj: *S. mutans* LJ23, smc: *S. mutans* NN2025, smu: *S. mutans* UA159, scp: *S. parasanguinis* ATCC 15912, scf: *S. parasanguinis* FW213, stf: *S. salivarius* 57 I, ssr: *S. salivarius* CCHSS3, stj: *S. salivarius* JIM8777, ssa: *S. sanguinis* SK36, stc: *S. thermophilus* CNRZ1066, stu: *S. thermophilus* JIM 8232, ste: *S. thermophilus* LMD 9, stl: *S. thermophilus* LMG 18311, stw: *S. thermophilus* MN ZLW 002, stn: *S. thermophilus* ND03. The boxplots for the different strains of the same species are shown in the same color.

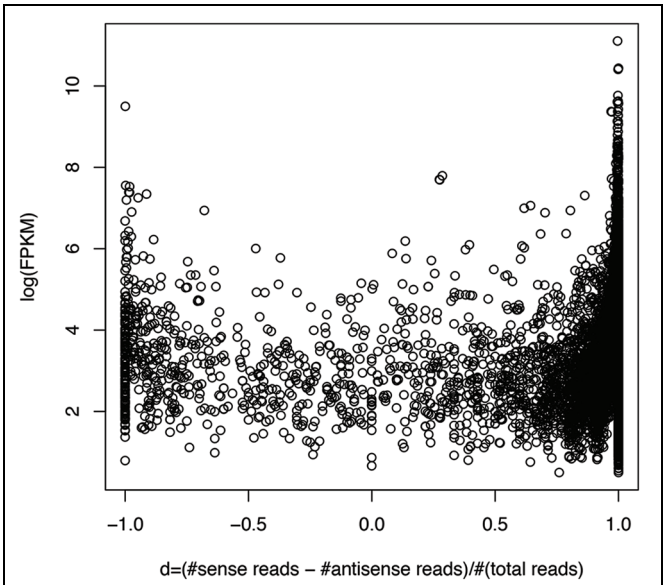

FIGURE 7 | Highly expressed genes tend to be dominated by sense transcription or antisense transcription. Each circle represents a gene. The y-axis shows the gene expression (*log*(FPKM)) and the x-axis shows *d*, which is close to 1 for genes with mostly sense transcription, and −1 with mostly antisense transcription. RNA-seq data from individual 1 (X310763260) was used for this plot. To limit the bias that may be introduced by rare species or genes with low expression levels, we only included the genes (3,689 in total) each supported by at least 20 RNA-seq reads, and are from the species (23 in total) each having at least 100 genes with RNA-seq reads support. See Supplementary Figure S2 for the plot using the original gene expression data, involving 30,493 genes from all 47 strains, one for each species.

correlation is confirmed by a statistical test: the Spearman's correlation coefficient between log(FPKM) and |d| for the genes (each recruited at least 20 RNA-seq reads) shown in **Figure 7** (excluding the genes with d ratios of 1 or −1) is 0.57 (p-value < 2.2e-16). Similar results can be observed using an unfiltered dataset from this individual (Spearman's $r = 0.69$, p-value < 2.2e-16), and RNA-seq datasets from other individuals.

We note that a large fraction of genes have either sense transcription only (which is not surprising), or antisense transcription only. For example, for the dataset X310763260 used in **Figure 7** and Supplementary Figure S2, a total of 6,119 protein coding genes (out of 30,493; 20.1%) have antisense transcription according to the binomial testing (success rate = 1%); among which, 1,877 genes only have antisense transcription. We expect this large ratio (1,877/6,119 = 30.7%) of genes with antisense transcription can be only partially contributed by bad gene annotations (which, however, will be difficult to quantitatively estimate without further experimental proofs). But there are still 430 genes if we only included genes at least 600 bp long (longer genes are more likely to be correctly predicted), with at least three RNA-seq reads mapped to their antisense strands (but no reads mapped to sense strands). The gene BVU_3334 in *B. vulgatus* ATCC 8482 mentioned above (**Figure 5B**) is one of such genes: a total of 257 reads were mapped to its antisense strand, but none to the sense strand.

Dynamic Antisense Transcription in Human Population

Antisense transcription varies between human individuals. For example, as shown in **Figure 6** (and Supplementary Figure S1), the prevalence of antisense transcription in different *Streptococcus* species varies across human individuals. In addition, the actual genes that have antisense transcripts vary greatly: most of the genes with antisense transcription are only found to have antisense transcription in one or only a few individuals (**Figures 8A–C**). For example, a total of 1,535 protein coding genes (out of 4,067; 38%) in *B. vulgatus* ATCC 8482 are found to have antisense transcription in at least one of the eight individuals; however, only two genes are common in all individuals, while 666 genes are found in only one of the individuals (**Figure 8B**). *M. smithii* ATCC 35061 is an exception (**Figure 8D**): many of its genes with antisense transcription are common among the individuals. A total of 792 protein coding genes (out of 1,793 genes; 44%) are found to have antisense reads in at least one of the eight individuals, and 236 of these genes have antisense reads in five individuals (note that *M. smithii* was found to be abundant only in five out of the eight individuals; see **Figures 4E,F**).

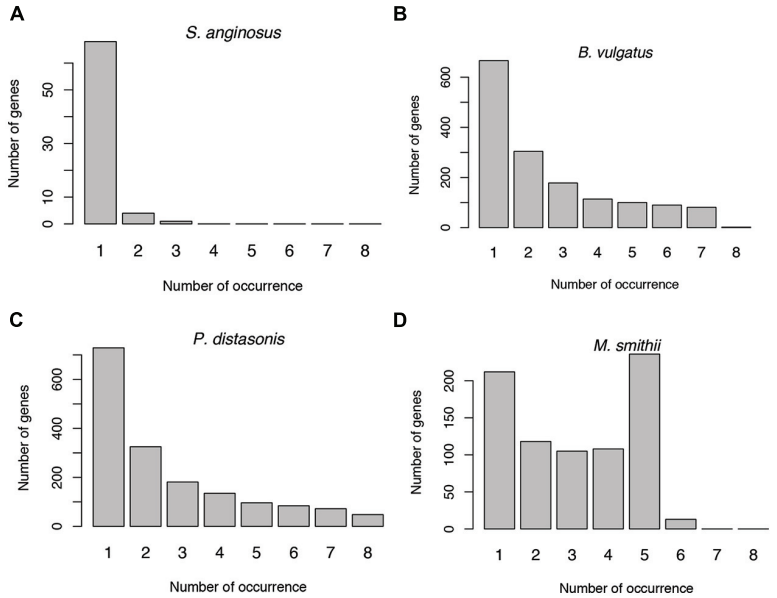

FIGURE 8 | Sharing of genes with antisense transcription among human individuals. Genes associated with *Streptococcus anginosus* C238 **(A)**, *Bacteroides vulgatus* ATCC 8482 **(B)** and *Parabacteroides distasonis* ATCC 8503 **(C)** tend to be unique to different individuals, while genes associated with *Methanobrevibacter smithii* ATCC 35061 tend to be shared by individuals **(D)**. The numbers below the bars indicate the number of individuals sharing the genes with antisense transcription, with 1 indicating the number of genes unique to one individual, and 2–8 for genes shared by two individuals, and then increasing numbers of individuals.

TABLE 1 | Clusters of Orthologous Groups (COG) functions that are enriched in the genes with antisense transcription in 47 strains (q-value < = 0.05 by Fisher's exact test with FDR correction).

COG ID	Cat[$]	Strains[#]	Function description	q-value*
COG3842	E	1	ABC-type Fe^{3+}/spermidine/putrescine transport systems, ATPase components	0.032
COG0493	E, R	1	NADPH-dependent glutamate synthase beta chain or related oxidoreductase	0.032
COG2226	H	1	Ubiquinone/menaquinone biosynthesis C-methylase UbiE	0.029
COG0568	K	1	DNA-directed RNA polymerase	0.0099
COG0583	K	1	DNA-binding transcriptional regulator, LysR family	0.018
COG1961	L	1	Site-specific DNA recombinase related to the DNA invertase Pin	0.033
COG4974	L	2	Site-specific recombinase XerD	2.09e-06; 0.032
COG1178	P	1	ABC-type Fe^{3+} transport system, permease component	0.032
COG2059	P	1	Chromate transport protein ChrA	0.032
COG0628	R	1	Predicted PurR-regulated permease PerM	0.032
COG0534	V	2	Na^{+}-driven multidrug efflux pump	0.014; 0.018
COG2801	X	1	Transposase InsO and inactivated derivatives	0.018
COG2826	X	2	Transposase and inactivated derivatives, IS30 family	7.57e-07; 0.012
COG3293	X	1	Transposase	0.0068
COG3328	X	1	Transposase (or an inactivated derivative)	0.012
COG3378	X	1	Phage- or plasmid-associated DNA primase	0.029
COG3415	X	1	Transposase	0.029
COG3464	X	1	Transposase	0.029
COG3666	X	1	Transposase	0.042

[$]*Functional categories (check the caption in* ***Figure 9*** *for the description of the categories);* [#]*Number of strains with detected antisense expression for the corresponding function; *All q-values will be listed if a function is detected to be enriched in multiple species.*

Functions Enriched in Genes with Antisense Transcription

We used two different statistical tests to detect if genes encoding certain functions tend to have antisense transcription: **Table 1** lists the COG functions that are enriched in genes with antisense transcription based on the Fisher's exact test with BH FDR correction. The binomial tests gave consistent results but with fewer COGs detected to be enriched (see Supplementary Table S1

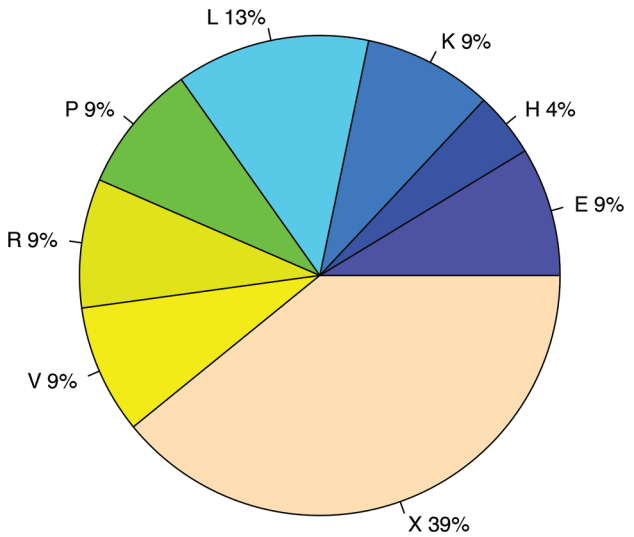

FIGURE 9 | COG functional categories enriched in genes with antisense transcription. The functional categories include E: Amino acid transport and metabolism; H: Coenzyme transport and metabolism; K: Transcription; L: Replication, recombination and repair; P: Inorganic ion transport and metabolism; R: General function prediction only; V: Defense mechanisms; and X: Mobilome, prophages, transposons.

for details). **Figure 9** summarizes the COG functional categories enriched in the genes (associated with the 47 species we tested; only one strain was selected for each species) that have observed antisense transcription. The most significant category is X (mobilome, prophages, and transposons), which has eight COG functions that are significantly enriched in genes with antisense transcription. The next category L, replication, recombination and repair, contains two enriched COG functions (COG1961 and COG4974). Transposases are among the genes frequently identified to have antisense transcription in previous studies: RNA-OUT of the transposon Tn10 (one of the first discovered antisense RNAs), was found to repress transposition by reducing transposase levels (Simons and Kleckner, 1983); and in a study of non-coding RNAs in the archaeon *Sulfolobus solfataricus* (Tang et al., 2005), the most prominent group of antisense RNAs was found to be transcribed in the opposite orientation to the transposase genes encoded by insertion elements (the authors of the paper hypothesized that these antisense RNAs regulate transposition of insertion elements by inhibiting expression of the transposase mRNA). We also identified other functions that are enriched in genes with antisense transcription, which may provide clues to the regulation of these genes.

Discussion

Strand-specific RNA-seq is a powerful tool for transcript discovery, genome annotation and expression profiling (Levin et al., 2010). In eukaryotes, 1000s of RNAs antisense to protein-coding genes have been revealed via high-throughput sequencing analyses (Berretta and Morillon, 2009). In contrast, few reports have identified antisense to protein-coding genes in bacteria, but previous studies have demonstrated that antisense RNAs can regulate expression of their corresponding genes in bacteria (Brantl, 2007). Although several studies have shown that antisense transcription may be widespread in bacteria, a global analysis of antisense transcripts using strand-specific information has only been reported for several model, cultured strains (Passalacqua et al., 2012; Behrens et al., 2014; Siegel et al., 2014). We describe a computational and statistical procedure to derive antisense transcripts from metatranscriptome data of microbial communities. With this method, we survey the antisense RNAs on a much broader scale than conventional methods, which have focused on single species.

Due to the fact that the strandedness is not 100% for RNA-seq experiments, it is necessary to have a way to correct for the artifacts. We proposed to use a binomial test to determine if a gene is likely to have antisense transcriptions, or the antisense reads are more likely artifacts. It helped to remove some of the artifacts. However, we note that this approach will underestimate the ratio of genes with antisense transcription for the species with few RNA-seq reads. This also indicates that when we compare the ratios of genes with antisense transcription for different species, we need to be cautious about the interpretation in comparing the results.

Mapping reads to bacterial genomes has been difficult due to the existence of closely related species in a microbial community and the limited availability of reference genomes (so the actual species might not be presented by the reference genomes; Wang et al., 2012). We acknowledge there is a potential problem with the assignment of sequencing reads to individual genomes due to the ambiguity of mapping. However, the conclusion we drew based on genes should be robust (the sense strand of a gene in one species is likely to be the sense strand as well for its homologs in related species). Also analysis at the pan-genome level or even genus level may be worth pursuing in the future, which may provide insights into the antisense transcription from different angles.

We note that there are other artifacts that may also have impacts on the analysis of antisense transcriptions and the interpretation of the results. For examples, genomic-DNA contamination may result in the detection of artificial antisense transcriptions (Haas et al., 2012). The different genome sizes for the species in a community, and different gene lengths will complicate the analysis of gene expression (Garber et al., 2011). Gene annotations for most of the genomes contain mistakes, and there are complicated gene structures (such as overlapping genes) that are difficult to be considered for antisense transcription analysis. Finally, for metatranscriptomic studies, the RNA-seq data reflects the compound output of the gene expression and the species abundance, making the interpretation of the results less straightforward.

Antisense transcription can be important for the regulation of some functions, such as transposase genes. One interesting example is the *Bacteroides uniformis* mobilizable transposon NBU1. All of its 10 genes have antisense expression in one

individual, and in other individuals also have higher antisense expression for this strain. This result suggests that in most individuals, the inactivation of this transposon by antisense RNAs serves an important regulatory role for its transposition. A further observation is that for a given bacterial species, the set of genes with antisense transcripts varies between human host, suggesting that environmental differences between hosts is leading to antisense-dependent regulatory responses by the resident bacteria.

Genes that have exclusively antisense transcripts are clearly "off"; depending on the efficacy of antisense suppression of sense translation, all genes with >50% antisense may be turned off, for example. And it is not surprising that many genes in a genome are turned off under a given set of conditions. A gene that is repressed for sense expression will naturally show a higher level of antisense expression, even if this is background noise. The question then becomes: are the antisense transcripts we observed actually a mechanism to specifically suppress expression, especially for genes with the highest levels of antisense expression. Strong antisense transcription was detected for the *opa* genes coding for adhesins and invasins, which may have regulatory functions in pathogenic *Neisseria* (Remmele et al., 2014). In the case of transposons, we know that antisense transcripts are a specific mechanism to maintain very low, or episodic, expression (Brantl, 2007). If at least some genes are being regulated by their antisense transcripts, it is no surprise that the levels will vary between different environments, i.e., individuals.

Author Contributions

GB and MW carried out the analysis and drafted the manuscript. TD participated in the analysis and helped to draft the manuscript. YY conceived the study, participated in its design and coordination, participated in the analysis, and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgment

This work was supported by NIH grant R01AI108888 and NSF grant DBI-0845685.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00896

Data Sheet 1 | An Excel file with three spreadsheets listing the species included in the analyses, and details of their ratios of antisense reads and genes with antisense transcription in different individuals.

Data Sheet 2 | An Excel file with eight spreadsheets listing detailed information about antisense and sense reads for individual protein coding genes in different genomes, across eight samples.

References

Beaume, M., Hernandez, D., Farinelli, L., Deluen, C., Linder, P., Gaspin, C., et al. (2010). Cartography of methicillin-resistant *S. aureus* transcripts: detection, orientation and temporal expression during growth phase and stress conditions. *PLoS ONE* 5:e10725. doi: 10.1371/journal.pone.0010725

Behrens, S., Widder, S., Mannala, G. K., Qing, X. X., Madhugiri, R., Kefer, N., et al. (2014). Ultra deep sequencing of *Listeria monocytogenes* sRNA transcriptome revealed new antisense RNAs. *PLoS ONE* 9:e83979. doi: 10.1371/journal.pone.0083979

Beiter, T., Reich, E., Williams, R. W., and Simon, P. (2009). Antisense transcription: a critical look in both directions. *Cell. Mol. Life Sci.* 66, 94–112. doi: 10.1007/s00018-008-8381-y

Berretta, J., and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* 10, 973–982. doi: 10.1038/embor.2009.181

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Brantl, S. (2007). Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr. Opin. Microbiol.* 10, 102–109. doi: 10.1016/j.mib.2007.03.012

Coolen, M. J., and Orsi, W. D. (2015). The transcriptional response of microbial communities in thawing Alaskan permafrost soils. *Front. Microbiol.* 6:197. doi: 10.3389/fmicb.2015.00197

de Menezes, A., Clipson, N., and Doyle, E. (2012). Comparative metatranscriptomics reveals widespread community responses during phenanthrene degradation in soil. *Environ. Microbiol.* 14, 2577–2588. doi: 10.1111/j.1462-2920.2012.02781.x

Dornenburg, J. E., Devita, A. M., Palumbo, M. J., and Wade, J. T. (2010). Widespread antisense transcription in *Escherichia coli*. *MBio* 1, e00024–e00110. doi: 10.1128/mBio.00024-10

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2329–E2338. doi: 10.1073/pnas.1319284111

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* 8, 469–477. doi: 10.1038/nmeth.1613

Giannoukos, G., Ciulla, D. M., Huang, K., Haas, B. J., Izard, J., Levin, J. Z., et al. (2012). Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* 13, R23. doi: 10.1186/gb-2012-13-3-r23

Green, P. J., Pines, O., and Inouye, M. (1986). The role of antisense RNA in gene regulation. *Annu. Rev. Biochem.* 55, 569–597. doi: 10.1146/annurev.bi.55.070186.003033

Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W., and Livny, J. (2012). How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes? *BMC Genomics* 13:734. doi: 10.1186/1471-2164-13-734

He, S., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., et al. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat. Methods* 7, 807–812. doi: 10.1038/nmeth.1507

Human Microbiome Project Consortium. (2012a). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Human Microbiome Project Consortium. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jens, G., and Wolfgang, R. H. (2011). cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* 75, 286–300. doi: 10.1128/MMBR.00032-10

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., and Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *MBio* 5, e01012–e01014. doi: 10.1128/mBio.01012-14

Lacatena, R. M., and Cesareni, G. (1981). Base pairing of RNA I with its complementary sequence in the primer precursor inhibits ColE1 replication. *Nature* 294, 623–626. doi: 10.1038/294623a0

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leimena, M. M., Ramiro-Garcia, J., Davids, M., Van Den Bogert, B., Smidt, H., Smid, E. J., et al. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics* 14:530. doi: 10.1186/1471-2164-14-530

Levin, J. Z., Yassour, M., Adiconis, X. A., Nusbaum, C., Thompson, D. A., Friedman, N., et al. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* 7, 709–715. doi: 10.1038/nmeth.1491

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Mao, F., Dam, P., Chou, J., Olman, V., and Xu, Y. (2009). DOOR: a database for prokaryotic operons. *Nucleic Acids Res.* 37, D459–D463. doi: 10.1093/nar/gkn757

Mitschke, J., Georg, J., Scholz, I., Sharma, C. M., Dienst, D., Bantscheff, J., et al. (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc. Natl. Acad. Sci. U.S.A.* 108, 2124–2129. doi: 10.1073/pnas.1015154108

Nealson, K. H., and Venter, J. C. (2007). Metagenomics and the global ocean survey: what's in it for us, and why should we care? *ISME J.* 1, 185–187. doi: 10.1038/ismej.2007.43

Passalacqua, K. D., Varadarajan, A., Weist, C., Ondov, B. D., Byrd, B., Read, T. D., et al. (2012). Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. *PLoS ONE* 7:e43350. doi: 10.1371/journal.pone.0043350

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Remmele, C. W., Xian, Y., Albrecht, M., Faulstich, M., Fraunholz, M., Heinrichs, E., et al. (2014). Transcriptional landscape and essential genes of *Neisseria gonorrhoeae*. *Nucleic Acids Res.* 42, 10579–10595. doi: 10.1093/nar/gku762

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., et al. (2007). The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5:e77. doi: 10.1371/journal.pbio.0050077

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., et al. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464, 250–255. doi: 10.1038/nature08756

Siegel, T. N., Hon, C. C., Zhang, Q., Lopez-Rubio, J. J., Scheidig-Benatar, C., Martins, R. M., et al. (2014). Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. *BMC Genomics* 15:150. doi: 10.1186/1471-2164-15-150

Sigurgeirsson, B., Emanuelsson, O., and Lundeberg, J. (2014). Analysis of stranded information using an automated procedure for strand specific RNA sequencing. *BMC Genomics* 15:631. doi: 10.1186/1471-2164-15-631

Simons, R. W., and Kleckner, N. (1983). Translational control of IS10 transposition. *Cell* 34, 683–691. doi: 10.1016/0092-8674(83)90401-4

Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., et al. (2005). Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* 55, 469–481. doi: 10.1111/j.1365-2958.2004.04428.x

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631–637. doi: 10.1126/science.278.5338.631

Thomason, M. K., and Storz, G. (2010). Bacterial antisense RNAs: how many are there, and what are they doing? *Annu. Rev. Genet.* 44, 167–188. doi: 10.1146/annurev-genet-102209-163523

Wang, M., Ye, Y., and Tang, H. (2012). A de Bruijn graph approach to the quantification of closely-related genomes in a microbial community. *J. Comput. Biol.* 19, 814–825. doi: 10.1089/cmb.2012.0058

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

REVIEW
published: 06 October 2015
doi: 10.3389/fmicb.2015.01050

Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches

José E. Belizário and Mauro Napolitano*

Department of Pharmacology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil

The human body is the residence of a large number of commensal (non-pathogenic) and pathogenic microbial species that have co-evolved with the human genome, adaptive immune system, and diet. With recent advances in DNA-based technologies, we initiated the exploration of bacterial gene functions and their role in human health. The main goal of the human microbiome project is to characterize the abundance, diversity and functionality of the genes present in all microorganisms that permanently live in different sites of the human body. The gut microbiota expresses over 3.3 million bacterial genes, while the human genome expresses only 20 thousand genes. Microbe gene-products exert pivotal functions via the regulation of food digestion and immune system development. Studies are confirming that manipulation of non-pathogenic bacterial strains in the host can stimulate the recovery of the immune response to pathogenic bacteria causing diseases. Different approaches, including the use of nutraceutics (prebiotics and probiotics) as well as phages engineered with CRISPR/Cas systems and quorum sensing systems have been developed as new therapies for controlling dysbiosis (alterations in microbial community) and common diseases (e.g., diabetes and obesity). The designing and production of pharmaceuticals based on our own body's microbiome is an emerging field and is rapidly growing to be fully explored in the near future. This review provides an outlook on recent findings on the human microbiomes, their impact on health and diseases, and on the development of targeted therapies.

Keywords: microbiome, metagenomics, phage therapy, CRISPR/Cas system, quorum sensing, pharmacomicrobiomics

OPEN ACCESS

Edited by:
Eric Altermann, AgResearch Ltd, New Zealand

Reviewed by:
M Andrea Azcarate-Peril, University of North Carolina at Chapel Hill, USA
Olivia McAuliffe, Teagasc, Ireland

****Correspondence:***
José E. Belizário, Department of Pharmacology, Institute of Biomedical Sciences, University of São Paulo, Avenida Lineu Prestes, 1524, CEP 05508-900, São Paulo, SP, Brazil
jebeliza@usp.br

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *12 May 2015*
Accepted: *14 September 2015*
Published: *06 October 2015*

Citation:
Belizário JE and Napolitano M (2015) Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. Front. Microbiol. 6:1050. doi: 10.3389/fmicb.2015.01050

Introduction

The evolution of *Homo sapiens* is linked to a mutualistic partnership with the human gut microbiota. The human genome is part of a collective genome of complex commensal, symbiotic, and pathogenic microbial community that colonizes the human body. Our microbiome includes not only bacteria, but also viruses, protozoans, and fungi (Backhed et al., 2012). Bacteria are a vast group of living organisms considered a domain of life in themselves (Woese et al., 1990). They are classified using DNA-based tests, morphologically and biochemically based on cell wall type, cell shape, oxygen requirements, endospore production, motility, and energy requirements. Hans Christian Gram (1850–1938), a Danish scientist, discovered that the presence of high levels of peptidoglycan (50–90%) produced a dark violet color, while low levels (<10%) resulted in reddish/pinkish colors, which are the respective staining of Gram-positive and Gram-negative bacteria. The Gram-negative cell wall is also characterized by the presence of lipopolysaccharides

(LPSs). Based on their capacity to produce energy in presence or absence of oxygen, bacteria can also be classified as aerobic, anaerobic or facultative anaerobic. In addition to the generation of ATP via aerobic or anaerobic respiration, bacteria can also produce energy via fermentation. Facultative anaerobic bacteria are able to generate ATP with or without oxygen, while obligated anaerobic bacteria do not tolerate it and only survive in anaerobiosis. *Lactobacillus, Staphylococcus,* and *Escherichia coli* are examples of facultative anaerobic bacteria. *Bacteroides*, on the other hand, are obligated anaerobic species. In inflamed tissues, the enterocytes produce reactive oxygen species (ROS) and kill anaerobic bacteria increasing the abundance of aerobic and facultative species.

Bacteria are classified phylogenetically based on the analysis of nucleotide sequences of small subunit ribosomal RNA operons, mainly variable regions of the bacterial specific ribosomal RNA, 16S rRNA (Woese, 1987; Woese et al., 1990; Marchesi et al., 1998). Currently, the Bacteria domain is divided into many phyla; however, the majority of microbes forming the human microbiota can be assigned to four major phyla: Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria (Zoetendal et al., 2008; Arumugam et al., 2011; Segata et al., 2012). Firmicutes and Bacteroidetes represent more than 90% of the relative abundance of the gut microbiome (Arumugam et al., 2011; Segata et al., 2012). Firmicutes are a diverse phylum composed mainly of the Bacilli and Clostridia classes. They are Gram-positive, anaerobic (Clostridia) and obligate or facultative aerobes (Bacilli) characterized by a low GC content. Bacteria of *Clostridium* species produce endospores in order to survive to adverse (aerobic) conditions (Paredes-Sabja et al., 2014). The phylum Bacteroidetes is composed of Gram-negative, non-spore forming anaerobic bacteria that tolerate the presence of oxygen but cannot use it for growth. Actinobacteria (e.g., *Bifidobacterium)* are Gram-positive, multiple branching rods, non-motile, non-spore-forming, and anaerobic bacteria. Proteobacteria (e.g., *Escherichia*, *Klebsiella, Enterobacter*) are aerobic or facultative anaerobic, Gram-negative, non-spore-forming rod bacteria, which inhabit the intestinal tract of all vertebrates.

Recent survey studies on the variation of human microbiomes concluded that European individuals could be classified in up to three enterotypes based on 16S rRNA gene data and functional metagenome (whole genome shotgun) data (Arumugam et al., 2011; Koren et al., 2013). An enterotype refers to the relative abundance of specific bacterial taxa within the gut microbiomes of humans. The functional metagenome of each enterotype revealed differences in the proportions of genes involved in carbohydrate versus protein metabolism, which is consistent with diets of different populations (Arumugam et al., 2011; Koren et al., 2013). People differ by species composition, distribution, diversity, and numbers of bacteria (Yatsunenko et al., 2012). The dietary habits are the critical contributing factor. Diversity (microbiome variation and complexity) increases from birth and reaches its highest point in early adulthood, thereafter declining with old age. However, larger longitudinal studies that include more populations, such as South Americans, Indians and Africans need to be done to identify the actual structure and biological impact of the distinct human microbiomes. These studies may also reveal how evolution of life-styles modulated ancestral and modern human microbiomes. Here we will present and discuss recent advances of microbiome studies and the strategies for the development of innovative pharmaceuticals based on emerging population and individual microbiota genomic information.

Metagenomics

The recent development of next generation sequencing (NGS) technologies such as 454, Solexa/Illumina, Ion Torrent and Ion Proton sequencers and the parallel expansion of powerful bioinformatics programs made possible the genomic analysis of over 1,000 prokaryotic and 100 eukaryotic organisms, including over 1,200 complete human genomes (Flintoft, 2012; Belizario, 2013). Metagenomics is a biotechnological approach to study genomic sequences of uncultivated microbes directly from their natural sources (Wooley et al., 2010). This allows the simultaneous analysis of microbial diversity connecting it to specific functions in different environments, such as soil, marine environments, and human body habitats (Ley et al., 2008; Robinson et al., 2010; Culligan et al., 2014). Using these novel methods, scientists have provided evidence for the existence of more than one thousand microorganism species living in our body (Arumugam et al., 2011; Segata et al., 2012) and an estimation of 10^7 to 10^9 different species of bacteria living on earth (Curtis et al., 2002). More important, the metagenomics approach has the potential to uncover entirely novel genes, gene families, and their encoded proteins, which might be of biotechnological and pharmaceutical relevance.

Currently several international projects aimed at the characterization of the human microbiota are being carried out. The Human Microbiome Project (HMP) is a research initiative of the National Institute of Health (NIH) in the United States, which aims to characterize the microbial communities found in several different sites of the human body (Turnbaugh et al., 2007; Backhed et al., 2012; Human Microbiome Project, 2012a,b). MetaHIT (Metagenomics of the Human Intestinal Tract) is a project financed by the European Commission and is under management of a consortium of 13 European partners from academia and the industry. The International Human Microbiome Consortium (IHMC) is composed of European, Canadian, Chinese, and US scientific institutions[1].

A simple molecular approach to explore the microbial diversity is based on the analysis of variable regions of 16S rRNA gene using "universal" primers which are complementary to highly conserved sequences among the homologous 16S rRNA genes (Marchesi et al., 1998; Culligan et al., 2014). These genes contain nine hypervariable regions (V1–V9) whose sequence diversity is appropriated for characterizing bacterial community compositions in complex samples (Guo et al., 2013; Jiang et al., 2014; Montassier et al., 2014). DNA sequences obtained with this approach can be mapped onto a reference set of known bacterial genomes. For this purpose useful bioinformatics tools

[1]http://www.human-microbiome.org/

and databases are available. For example, the SILVA database[2] is a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data that helps determine an optimal alignment for the different sequence regions (Pruesse et al., 2007). First released in 1995, The Ribosomal Database Project (RDP) is another database that provides high quality alignments of archaeal and bacterial 16S rRNA sequences as well as fungal 28S rRNA sequences (Maidak et al., 1996). The microbial profiling and phylogenetic clustering of microbiomes of the American and European population have been already deposited and are free for consultation[3,4]. The HMP projects and other independent projects have been generating an enormous amount of metagenomic data and the assemblies of microbiome data is being undertaken by the Genomes OnLine Database[5]. The data management system for cataloging and continuous monitoring of worldwide sequencing projects contains data from over 4000 metagenome sequencing projects, in which more than 1500 are aimed at the characterization of host associated metagenomes (Human Microbiome Jumpstart Reference Strains et al., 2010; Fodor et al., 2012; Reddy et al., 2015).

The first release of the HMP database included microbiome data of nasal passages, the oral cavity, skin, gastrointestinal tract, and urogenital tract (Human Microbiome Project, 2012a,b). **Figure 1** schematically summarizes the data of these studies. The results of over 690 human microbiomes have shown that the majority of bacteria of the gut microbiome belongs to four phyla: Firmicutes, Bacteroidetes, Actinobacteria, and Proteobacteria (Human Microbiome Project, 2012a,b). Only a fraction of microbes identified so far have been successfully cultured, and thousands are yet to be fully sequenced for a deeper taxonomic resolution (strains and subspecies) and functional analysis at the genomic level (Qin et al., 2010; Robinson et al., 2010; Abubucker et al., 2012; Flintoft, 2012; Zhou et al., 2013).

The metagenome wide association studies in development in many countries are promising in predicting new diagnostic and prognostic tools for numerous human disorders. The results of these studies will dramatically increase our knowledge of diseases linked to microbial composition (Qin et al., 2010; Clemente et al., 2012; Flintoft, 2012; Gevers et al., 2012). In order to better understand the host-gene-microbial interactions and the role of non-pathogenic and pathogenic strains in large populations, we need to compare microbiome profiles across multiple body sites and microbiome datasets under environmentally controlled normal and disease conditions. In the following sections, we will provide a synthesis of the recent studies on the human microbiomes identified in some major body sites.

Gut Microbiome

The HMPs have shown that the human gut harbors one of the most complex and abundant ecosystems colonized by more than 100 trillion microorganisms (Human Microbiome Project, 2012a,b). In adults, the majority of the bacteria found in the gut belong to two bacterial phyla, the gram-negative Bacteroidetes and the gram-positive, Firmicutes; and the others represented at subdominant levels are the Actinobacteria, Fusobacteria, and Verrucomicrobia phyla, but this varies dramatically among individuals (Eckburg et al., 2005; Zoetendal et al., 2008; Arumugam et al., 2011; Backhed et al., 2012; Segata et al., 2012). For instance, the most abundant genera from the Bacteroidetes phylum are *Bacteroides* and *Prevotella species*, which represent 80% of all Bacteroidetes in fecal samples. Nonetheless, many of the taxa numerically underrepresented and less-abundant bacterial species exert fundamental functions at a particular location in the gut. To better define these different microbial colonization and microbiota structure in different cohorts arose the concept of 'enterotype clusters' that allow the classification of each individual based on the relative abundance of specific bacterial taxa in fecal samples, and their microbial metabolic and functional pathways (Arumugam et al., 2011; Backhed et al., 2012; Koren et al., 2013). The results of metagenomic sequencing of fecal samples from European, American, and Japanese subjects confirmed the three robust clusters dominated by *Bacteroides* (enterotype 1), *Prevotella* (enterotype 2), and *Ruminococcus* (enterotype 3), each one characterized by specific taxonomic composition and relative abundance of metabolic pathways. For example, enterotype 1 was enriched in biosynthesis of biotin, riboflavin pantothenate and ascorbate; enterotype 2 in biosynthesis of thiamine and folate. Enterotype 3 showed high abundance of genes involved in haem biosynthesis pathway. Although in one of these studies (Arumugam et al., 2011) it was confirmed that a set of 12 genes correlated with age and a set of three functional modules with the body mass index (BMI), further studies will be required to determine if specific microbiome and/or enterotype is associated with gender, BMI, health status, diet, and age of individuals (Arumugam et al., 2011; Backhed et al., 2012; Koren et al., 2013).

Although stable over long periods, the composition and functions of the gut microbiome may be influenced by a number of factors including genetics, mode of delivery, age, diet, geographic location, and medical treatments (Clemente et al., 2012; Brown et al., 2013). The intestinal microbiota is acquired in the postnatal periods of time, consisting of a wide variety of bacteria that plays different functions in the human host, including nutrient absorption, protection against pathogens, and modulation of the immune system (Brown et al., 2013). The gut is an anaerobic environment in which indigenous species have co-evolved with the host. The aerobic pathogenic species cannot invade and colonize it; however, anaerobic and facultative pathogenic species can invade it causing diseases. High diversity defines healthy human gut microbiomes, whereas reduction in diversity may be associated with dysbiosis (Manichanh et al., 2006; Backhed et al., 2012). Dysbiosis refers to an imbalance in the microbiome structure that results from an abnormal ratio of commensal and pathogenic bacterial species. Many studies have suggested a possible direct relationship between dysbiosis and inflammatory and metabolic diseases such as is inflammatory bowel diseases (IBD) including colitis and Crohn's disease (CD), obesity and cancer (Clemente et al., 2012; Sartor and Mazmanian, 2012; Brown et al., 2013). However, investigation of such a

[2]http://www.arb-silva.de

[3]http://www.metahit.eu/

[4]http://www.hmpdacc.org/

[5]http://www.genomesonline.org

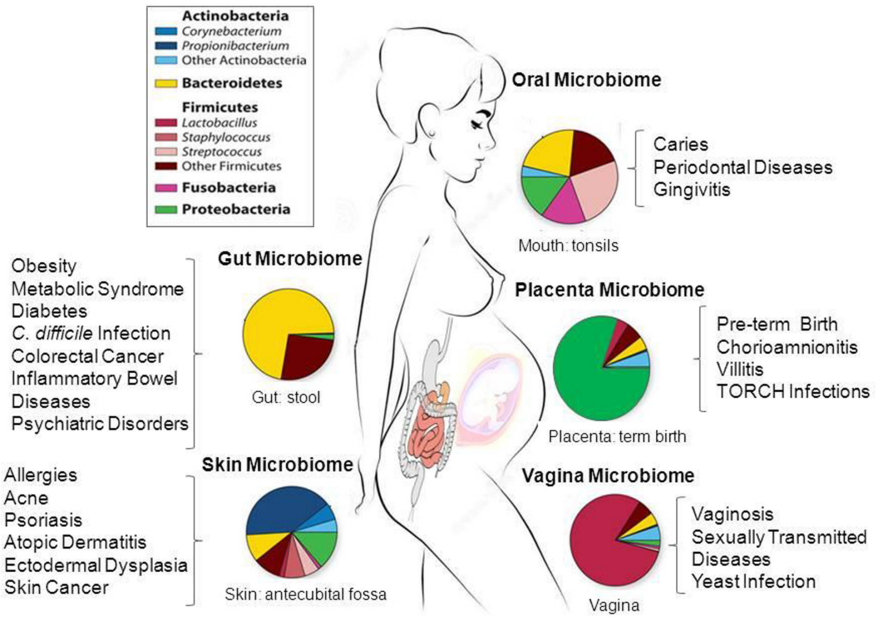

FIGURE 1 | Taxonomic distribution, prevalence and abundance of microbial taxa that inhabit healthy human body sites as defined in the human microbiome projects (HMP). The colored rectangles denote phylum/class and genera. Clinical studies of the microbiome will help to elucidate the link between microbes and the promotion of a large number of diseases and pathological conditions as shown in the figure. The images were adapted from NIH HMP (http://www.hmpdacc.org/) and National Human Genome Research Institute (https://www.genome.gov/). TORCH, Toxoplasmosis, Oher infections (coxsackievirus, HIV, syphilis, etc), Rubella, Cytomegalovirus, Herpes simplex.

complex ecosystem is difficult and it is still not easy to define how shifts in microbial composition and member abundance can lead to diseases. Induction of some IBD has been linked to a reduction of Firmicutes and Bacteroidetes and an expansion of Proteobacteria. For example, *Faecalibacterium prausnitzii*, a prominent member of Clostridium group IV (Firmicutes), protective and anti-inflammatory commensal bacterium, is frequently reduced in CD patients (Sokol et al., 2008; Sartor and Mazmanian, 2012). Despite these advances, it should be noted that microbiota composition varies between different locations in the gastrointestinal tract (Eckburg et al., 2005; Zoetendal et al., 2008; Arumugam et al., 2011; Cucchiara et al., 2012; Segata et al., 2012; Lepage et al., 2013). Most studies in the literature have explored only fecal microbiota. Fecal samples contain between 1,000 and 1,150 bacterial species, and up to 55% are uncultivable and thus uncharacterized (Zoetendal et al., 2008; Qin et al., 2010; Segata et al., 2012; Zhou et al., 2014). Our knowledge is especially limited when it comes to the other parts of the GI tract, a potential source of uncharacterized microbial species, which is largely due to sampling constraints.

The dysregulation of the intestinal immune system can also trigger microbial dysbiosis (Clemente et al., 2012; Sartor and Mazmanian, 2012; Brown et al., 2013). Many different inflammatory diseases are characterized by mutations or loss of some innate response genes in lymphoid tissues, Paneth cells, smaller Peyer's patches and mesenteric lymph nodes (Clemente et al., 2012; Frantz et al., 2012; Sartor and Mazmanian, 2012). The growth of microbiota communities is under control of distinct subfamilies of host genes encoding antimicrobial peptides (AMPs). AMPs are the most ancient component of the innate host response against bacterial infections (Guani-Guerra et al., 2010; Ostaff et al., 2013). When bacteria colonize a given human habitat, the expression of AMPs, including α and β defensins and cathelicidins, is upregulated in order to limit the spreading of bacteria. The equilibrium between the immune system and immunoregulatory functions of bacteria appears to be a delicate balance in which the loss of a specific species can lead to an overreaction or suppression of the innate immune system (Round and Mazmanian, 2009; Clemente et al., 2012; Sartor and Mazmanian, 2012; Brown et al., 2013). Intestinal epithelial cells (IECs) form a physical and immunological barrier that separate luminal bacteria from underlying immune cells in the intestinal mucosa. IECs and hematopoietic cells express a variety of receptors called pattern recognition receptors (PRRs) that mediate the interactions between the immune system and the commensal microbiota (Frantz et al., 2012). Toll-like receptors (TLRs) and nuclear oligomerization domain-like receptors (NLRs) are examples of PRR that recognize unique microbial molecules named microbe-associated molecular patterns (MAMPs) including lipopolysaccharides (LPS), lipid A, peptidoglycans, flagella, and microbial RNA/DNA. These receptors activate inflammasomes and thereby the production of cytokines TNF-α and IL-1β (Brown et al., 2013; Sangiuliano et al., 2014). Myeloid differentiation factor MyD88 is an adaptor protein that is essential for TLRs signaling and host-microbial interactions and tissue homeostasis (Sangiuliano et al., 2014). Mice lacking MyD88 in IECS (IEC-Myd88 −/− mice) display intestinal barrier

disruption, deficiency in the production of pro-inflammatory cytokines and AMPs and overgrowth of several enteric bacterial pathogens (Frantz et al., 2012). It will be important to understand when and how dysbiosis and genetic defects in mucosa-IECs and innate regulatory mechanisms can lead to development of infectious or inflammatory diseases.

Microorganisms synthesize a wide range of low-molecular weight signaling molecules (metabolites), many of which are similar to metabolites produced by human cells (Wikoff et al., 2009). The maintenance of a stable, fermentative gut microbiota requires diets rich in whole plant foods particularly high in dietary fibers and polyphenols (Zoetendal et al., 2008). Under anaerobic conditions, species belonging to the *Bacteroides* genus, and to the Clostridiaceae and Lactobacillaceae families, produce short-chain fatty acids (SCFAs). Acetate (with two carbons), propionate (with three carbons), and butyrate (with four carbons) are SCFA used by the epithelial cells of the colon (colonocytes) and act as a major player in maintenance of gut homeostasis (Meijer et al., 2010). SCFAs induce the secretion of glucagon-like peptide (GLP-1) and peptide YY (PYY), which increase nutrient absorption from the intestinal lumen. This is a key process in controlling mucosal proliferation, differentiation and maintenance of mucosal integrity (Round and Mazmanian, 2009). Individuals colonized by bacteria of the genera *Faecalibacterium, Bifidobacterium, Lactobacillus, Coprococcus, and Methanobrevibacter* have significantly less of a tendency to develop obesity-related diseases like type-2-diabetes and ischemic cardiovascular disorders (Ley et al., 2006; Le Chatelier et al., 2013). These species are characterized by high production of lactate, propionate and butyrate as well as higher hydrogen production rates, which are known to inhibit biofilm formation and activity of pathogens, including *Staphylococcus aureus*, in the gut (Le Chatelier et al., 2013). Genetic and diet-induced mouse models of obesity have shown that the Bacteroidetes/Firmicutes ratio is decreased in obese animals compared to non-obese animals, which is consistent with what has been observed in human obese subjects (Ley et al., 2005, 2006; Le Chatelier et al., 2013; Verdam et al., 2013). However, controversies exist regarding the human data on gut microbiota composition in relation to obesity (Turnbaugh et al., 2009; De Filippo et al., 2010; Verdam et al., 2013). The intestinal microbiota changes in obese mice may increase the intestinal permeability and inflammation locally and in adipose tissues (Cani and Delzenne, 2011; Kootte et al., 2012). As discussed in recent studies, the microbial-derived LPS released through circulation may promote low-grade inflammatory process and metabolic disturbances related to obesity, such as insulin resistance and type-2 diabetes (Cani and Delzenne, 2011; Kootte et al., 2012). Despite our currently incomplete understanding of the mechanisms, there have been high expectations that targeted changes in microbiota by the rational use of prebiotics and probiotics might abolish metabolic alterations associated with obesity (Cani and Delzenne, 2011; Kootte et al., 2012).

Vagina Microbiome

The first study based on pyrosequencing of barcoded 16S rRNA genes of vaginal microbiota performed on samples from North-American women revealed the inherent differences within and between women in different ethnic groups (Ravel et al., 2011). The vaginal microbial composition from three vaginal sites (mid-vagina, cervix, and introitus) has been compared to the buccal mucosa and the perianal region in recent studies (Fettweis et al., 2012; Romero et al., 2014; Vaginal Microbiome consortium[6]). These studies have shown that the vagina possesses over 200 phylotypes and that the most predominant belong to the phyla Firmicutes, Bacteroidetes, Actinobacteria, and Fusobacteria (Ravel et al., 2011; Romero et al., 2014). The vagina has low pH due to secretion of lactic acid and hydrogen peroxide by *Lactobacillus* sp. If *Lactobacillus* decreases under the effects of antibiotics, *Gardnerella vaginalis* and *Peptostreptococcus anaerobius*, *Prevotella* sp., *Mobiluncus* sp. *Sneathia*, *Atopobium vaginae*, *Ureaplasma*, *Mycoplasma*, and numerous fastidious or uncultivated anaerobes can cause bacterial vaginosis (BV). BV is an ecological disorder of the vaginal microbiota that affects millions of women annually, and is associated with numerous adverse health outcomes including preterm birth and acquisition of sexually transmitted infections, e.g., HIV, *Neisseria gonorrhoeae*, *Chlamydia trachomatis*, and HSV-2 (Kenyon et al., 2013). *Lactobacillus* morphotypes predominate in normal grade 1. BVs grade 3 and higher are characterized by a reduced number of lactobacilli and increased diversity, especially high concentration of Gram-negative bacteria and coccobacillus (e.g., *G. vaginalis* and *G. mobiluncus)* and *Peptostreptococcus* (Delaney and Onderdonk, 2001). The results of microbiome studies of the vagina are showing different patterns and imbalances in bacterial communities associated with BVs, as well as those associated with non-infectious pathological states that predict increased risk for infertility, spontaneous abortion, and preterm birth.

Oral Microbiome

Advances in microbiological diagnostic techniques have shown the complex interaction between the oral microbiota and the host (Segata et al., 2012; Jiang et al., 2014; Perez-Chaparro et al., 2014). Bacteria, fungi, archaea, viruses, and protozoa are part of the oral microbiome. The HMP investigated bacterial communities in nine intraoral sites: buccal mucosa, hard palate, keratinized gingiva, palatine tonsils, saliva, sub-and supra gingival plaque, throat, and tongue dorsum (Human Microbiome Project, 2012a,b). Over 300 genera, belonging to more than 20 bacterial phyla were identified (Zhou et al., 2013; Jiang et al., 2014). However, only a limited number of species find proper conditions to colonize the root canal system (Zhou et al., 2013; Perez-Chaparro et al., 2014). The microbiota of periodontitis or caries is usually complex consisting of Gram-negative anaerobic bacteria such as *Porphyromonas gingivalis, Treponema denticola, Prevotella intermedia, Tannerella forsythia,* and *Agregatibacter actinomycetemcomitans* (Mason et al., 2013; Jiang et al., 2014). Most early data on the endodontic microbiota were obtained by culture-based method and it is likely that not-yet-cultivable and unknown species of bacteria play a role in oral microbial shift toward a disease (Mason et al., 2013; Zaura et al., 2014). As expected, deep DNA sequencing data revealed a

[6]http://vmc.vcu.edu/

larger number of taxa involved in endodontic infections. Species of phyla Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes, Synergistetes, and *Candidatus Saccharibacteria* were more frequently found. All these studies on bacterial diversity in endodontic infections revealed high inter-subject variability, indicating the need for further studies using homogenous diagnosis criteria in a significant number of healthy subjects (Mason et al., 2013; Jiang et al., 2014; Perez-Chaparro et al., 2014).

Skin Microbiome

The skin is the human body's largest organ, colonized by over 100 microbial phylotypes, most of which are harmless or even beneficial to their host (Rosenthal et al., 2011; Ladizinski et al., 2014; Zhou et al., 2014). Phylotypes, microbial abundance and diversity differ in relation to skin color, race, and geographic location (Grice et al., 2009; Rosenthal et al., 2011). Colonization is influenced by the ecology and the epidermis layers of the skin surface. Therefore it is highly variable depending on topographical location, endogenous host factors and exogenous environmental factors. The Actinobacteria phylum is the most abundant on the skin. Gram-positive *Staphylococcus epidermidis* and *Propionibacterium acnes* are predominant on human epithelia and in sebaceous follicles, respectively. *Propionibacterium acnes* colonizes healthy pores and is responsible for the production of SCFAs and thiopeptides, which inhibit the growth of *Staphylococcus aureus* and *Streptococcus pyogenes*. However, depending on the host's immune system, the overgrowth and clogging of pores allow subsequent colonization of *S. epidermidis* and *Staphylococcus aureus.* Atopic dermatitis is one chronic inflammatory condition of the skin that occurs in many children and adults (Grice et al., 2009). *Staphylococcus* sp. *Corynebacterium* sp. and the fungi *Candida* sp, and *Malassezia* sp. are also frequently associated with a number of skin diseases, including atopic dermatitis and abnormal flaking and itching of the scalp (Grice et al., 2009).

The skin microbiota is under autonomous control of the local cutaneous immune system, thus it is independent of the systemic immune response which is modulated by the gut microbiota (Naik et al., 2012). The major innate mechanism of antimicrobial defense on the skin consists of AMPs, for example defensins, cathelicidin LL-37 and dermcidin (Guani-Guerra et al., 2010). These peptides are emerging as important tools in the control of skin pathogenic bacteria as well as bacteria involved in diseases of the lung and gastrointestinal tract. Many AMPs bind to the phospholipid membrane surfaces, forming ion-channels and pores causing leakage and cell death (Guani-Guerra et al., 2010; Ostaff et al., 2013). However, their specific immunomodulatory roles in innate immune defense against bacterial and viral infection remain poorly understood (Ostaff et al., 2013; Wang, 2014). An enhanced understanding of the skin microbiome is necessary to gain insight into AMPs and innate response in human skin disorders. The cutaneous inflammatory disorders such as atopic dermatitis, psoriasis, eczema, and primary immunodeficiency syndromes have been associated with dysbiosis in the cutaneous microbiota. The skin commensals promote effector T cell response, via their capacity to control the NF-κB signaling and the production of cytokines TNF-α and IL-1β (Hooper et al., 2012). The binding of the skin microbiota components to TLRs or NLRs allows a sustainable homeostasis toward innate and adaptive immunity within a complex epithelial barrier throughout distinct topographical skin sites.

Placenta Microbiome

Historically, the fetus and intrauterine environment were considered sterile. However, the first profile of microbes in healthy term pregnancies identified a unique microbiome niche in normal placenta, composed of non-pathogenic commensal microbiota from the Firmicutes, Tenericutes, Proteobacteria, Bacteroidetes, and Fusobacteria phyla (Aagaard et al., 2014). This study describes the microbial communities of 320 placental specimens and, despite the expected differences between individuals, the taxonomic classification of the placental microbiome bears most similarity to the non-pregnant oral microbiome, in particular to those associated with tongue, tonsils, and gingival plaques. One predominant species was *Fusobacterium nucleatum*, a Gram-negative oral anaerobe. *E. coli* was also found in placenta; however, it is not present in the oral microbiome (Aagaard et al., 2014). The authors suggested a possible hematological spread of oral microbiome during early vascularization and placentation. The pathways related with the metabolism of cofactors and vitamins were the most abundant among placental functional gene profiles, which is different from the metabolic pathways found in other body sites (Aagaard et al., 2014).

The balance of the different microbe species in and on the human body changes throughout life and particularly in different stages of pregnancy (Qin et al., 2010; Human Microbiome Project, 2012a,b). It is well known that preterm delivery (<37 weeks) causes substantial neonatal mortality and morbidity (DiGiulio et al., 2008). Placentas from normal deliveries and preterm deliveries contained different populations of microbial species (Groer et al., 2014). The gram-negative bacillus *Durkholderia* was associated with preterm delivery and the gram-positive, rod-shaped, facultative anaerobic bacteria *Paenibacillus* with term delivery (Aagaard et al., 2014). Consistent with other studies, an enrichment in Streptococci, *Acinetobacter* and *Klebsiella* was also demonstrated in women with history of antenatal infection (Aagaard et al., 2014).

The presence of different microbes in amniotic fluid, umbilical cord blood, meconium (first stool), placental and fetal membranes suggested the existence of various routes and mechanisms by which bacteria from different microbiota translocate to placenta and babies (DiGiulio et al., 2008). Studies in mice have demonstrated the placental transmission from mother's oral microbiota (Fardini et al., 2010). Many of these organisms are transmitted to babies during nursing. Babies born vaginally have more diverse gut microbial communities similar to their mother's vaginal microbiota, while microbiomes of babies delivered by Cesarean section are similar to skin microbiota (Dominguez-Bello et al., 2010). The lack of exposure to maternal vaginal microbiome might explain why cesarean section babies are at greater risk of developing type 1 diabetes, celiac disease, asthma, and obesity (DiGiulio et al., 2008; Dominguez-Bello et al., 2010). Breastfed babies' microbiome is

enriched with *Lactobacillus* and *Bifidobacterium* species whereas microbiome of babies fed with formula/solid food are enriched with Enterococci, Enterobacteria, Bacteroides, Clostridia, and Streptococci (Guaraldi and Salvatori, 2012; Palmer et al., 2012; Thompson et al., 2015). The transition from breast milk to solid foods is associated with acquisition of a more adulthood-like microbiome; however, infectious diseases, antibiotic use and the characteristics of the diet can interfere with babies' microbiota composition (Thompson et al., 2015). Together, these findings emphasize the need for further studies on placental microbiome for elucidating more mechanisms to be explored in the prevention and treatment of babies from preterm birth and other diseases.

Microbiota-based Pharmaceuticals

Metagenomics has proven to be a powerful tool in determining the diversity and abundance of microbes in the human body. The microbiome databases have been explored as sources of interesting targets to drug development (Cani and Delzenne, 2011; Collison et al., 2012; Haiser and Turnbaugh, 2012; Carr et al., 2013; Wallace and Redinbo, 2013). Therapeutic interventions in the microbiome can be directed against molecular entities, such as essential and antibiotic resistance genes to quorum sensing systems components used to control microbial networking behaviors, including the chemical communication and production of virulence factors (Collison et al., 2012). In the next sections, we will present and discuss strategies to discover novel antimicrobial targets as well as dietary interventions and microbial modification genetic tools to eliminate pathogenic microorganisms and to control dysbiosis.

Targeting Essential Genes

Searching of essential genes for bacterial growth and viability is the first step for identifying potential drug targets (Wallace and Redinbo, 2013). Computational analyses can provide candidate targets in microbial community of pharmacological significance for controlling bacterial species involved in chronic diseases, metabolic, and cardiovascular diseases as well as drug metabolism (Collison et al., 2012). The metagenomic databases are critical for constructing gene and protein networks and an initial framework for drug target screening (Collison et al., 2012; Carr et al., 2013; Manor and Borenstein, 2015). Several bioinformatics approaches have been used to identify microbial gene essentiality and putative new classes and functions for unique microbial genes in the metagenomic databases. HUMAnN is a program for metagenomic functional reconstruction to directly associate community functions with habitat and host phenotype. This program has been used to compare functional diversity and organismal ecology in the human microbiome (Abubucker et al., 2012). About 20% of all genes in a strain are essential and this has gained interest in drug discovery research (Christen et al., 2011). *In vitro* transposition and genetic transformation of the wild-type bacteria using a transposon library is a reliable experimental approach to uncover gene essentiality (van Opijnen et al., 2009). ESSENTIALS is another software for rapid analysis of high throughput transposon insertion sequencing data and discovery of essential genes (Zomer et al., 2012).

The majority of unique targets found in microbes' genomes are genes responsible for the metabolism of carbohydrates, amino acids, xenobiotics, methanogenesis, and the biosynthesis of vitamins and isoprenoids. These genes are either non-homologous or orthologous to those encompassed in human genome. Vitamin biosynthetic pathways constitute a major source of potential drug targets. Most bacteria synthesize thiamine *de novo*, whereas humans depend on dietary uptake. Folic acid (vitamin B9) is an indispensable cofactor, which plays a key role in the methylation cycle and in DNA biosynthesis. Enzymes of the folate biosynthesis pathway, for example, dihydrofolate reductase, have been an attractive pharmaceutical targets for inhibiting folate synthesis. Sulfanilamide and trimethoprim are examples of effective antimicrobials used in a broad range of infectious diseases. Niacin (vitamin B3) participates in the biosynthesis of nicotinamide adenine dinucleotide (NAD^+), a coenzyme essential in electron transport reactions in cell metabolism processes. Bacterial NAD+ kinases have been explored as targets for inhibiting bacterial growth. Methionine is not synthesized *de novo* in humans, and is supplied by diet. In contrast, most bacteria need to synthesize methionine to survive. *S*-adenosylmethionine synthetase, a key enzyme in methionine biosynthesis, is one drug target whose great potential has been explored against various pathogens. New drugs, for example platensimycin and platencin, that inhibit the microbial fatty acid synthesis (FAS) pathway by targeting key FAS enzymes have been successfully developed (Parsons et al., 2014). A recent survey identified 127 orthologous groups conserved in both human and human commensal gut microflora that are not suitable targets for drug development. However among these, the 20 aminoacyl-tRNA synthetases (aaRSs), which encode essential enzymes for protein synthesis, can be used since bacterial and eukaryotic AaRS have different specificity for tRNAs (Ochsner et al., 2007; Mobegi et al., 2014). These are only few examples of attractive targets for drug development; however, metagenomic data will open new frontiers for discovery of essential genes.

Targeting Antibiotic Resistance Genes

The structure of the microbial community is maintained by specific microbial communication, cell signaling through cell-to-cell contact, metabolic interactions, and quorum sensing (Wright, 2010). Species within a bacterial community are either susceptible or resistant to epithelial innate AMPs and/or chemical antibiotics (Seo et al., 2010; Wozniak and Waldor, 2010; Sommer and Dantas, 2011). Bacterial genomes acquired resistance and metabolic genes from mobile genetic elements (MGE), including conjugative transposons, also called integrative conjugative elements (ICE), which are horizontally transferred by bacteriophages and plasmids (Wozniak and Waldor, 2010). Antibiotic resistance genes encoded in microbial genomes include multidrug efflux transporters, tetracycline resistance genes, vancomycin resistance genes, and beta-lactamases. In addition, a number of microbial genes and products, including bacteriocins, lysins, holins, restriction/modification

endonuclease systems, and other virulence factors contribute to resistance to antibiotics (Dawid et al., 2007; Seo et al., 2010; Wozniak and Waldor, 2010; Smillie et al., 2011). Targeted (PCR-based) and functional metagenomic approaches have been used to track the presence of resistance genes or their families in different ecosystems (Mullany, 2014). A method to specifically trap plasmids containing antibiotic resistance genes called transposon-aided capture (TRACA) has been developed (Jones and Marchesi, 2007; Mullany, 2014). In this method, the plasmids are tagged with transposons that contain a selectable marker and a replication origin, which facilitate acquisition of plasmids from the human gut metagenomic DNA extracts, and subsequent maintenance and selection in an *E. coli* host.

Most of the antibiotics used to fight bacterial infections today are derived from soil microbes. Penicillin, the first true antibiotic, came from the soil fungus *Penicillium* (Kardos and Demain, 2011). To investigate the role of soil microbiota as a reservoir of genes encoding antibiotic resistance in the metagenomic data set, the ORFs found on contigs and on unassembled reads were compared with 3,000 known antibiotic resistance genes (Forsberg et al., 2014). It was concluded that most of the identified soil bacteria resistance genes were not typically close to known human pathogen resistance genes, suggesting little sharing between soil and gut bacterial species. A study on the microbiome of uncontacted Amerindians, members of a Yanomami isolated village living in the Amazon region has revealed the highest diversity of bacteria and genetic functions in fecal, oral, and skin bacterial microbiome ever reported compared with the US group (Clemente et al., 2015). Despite their isolation and no known exposure to commercial antibiotics, they carry functional antibiotic resistance genes with over >95% amino acid identity to those that confer resistance to semisynthetic and synthetic antibiotic monobactam and ceftazidime (Clemente et al., 2015). This finding provided important insights into how westernization impacts on the heritability of the microbiome among populations (Yatsunenko et al., 2012). There is evidence suggesting that exposure to microbes from animal gut microbiomes and within our indoor spaces (house, office, schools, cars, etc.) may become new sources for antibiotics and antibiotic resistance genes to human populations (Wright, 2010; Sommer and Dantas, 2011; Forslund et al., 2013). These discoveries emphasize the importance of continued functional investigations on antibiotic resistance reservoirs in metagenomic data from isolated ancestral and modern populations with a given disease.

Targeting Quorum Sensing Systems

The term "Quorum Sensing" (QS) indicates systems used by bacteria to communicate with each other in order to synchronize their gene expression activities and behave in unison as a group (Miller and Bassler, 2001; Waters and Bassler, 2005; Hense and Schuster, 2015). This mechanism controls the synthesis of secreted products, disease-causing virulence factors, and many metabolites, including bacterial antibiotics that target competing bacteria, and substances that suppress the immune system (Miller and Bassler, 2001; Waters and Bassler, 2005). Thus, an alternative to killing or inhibiting growth of pathogenic bacteria is targeting these key regulatory systems (Finch et al., 1998; Defoirdt et al., 2010). Metagenomic studies have identified the genetic and phenotypic diversity of quorum-sensing systems that co-evolved with pathogenic species (Joelsson et al., 2006; Kimura, 2014). QS system was first described in marine bacteria *Vibrio harveyi* and *V. fischeri,* which use LuxI and LuxR proteins to control the expression of the luciferase enzyme for emitting luminesce upon reaching a critical mass or "quorum" (Nealson and Hastings, 1979). These bacteria secrete in the extracellular environment a small molecule, an acylated homoserine lactone (AHL), called autoinducer 1 (AI-1), to communicate with members of the same species (intraspecific communication; Miller and Bassler, 2001; Waters and Bassler, 2005; Ng and Bassler, 2009). After its discovery in marine bacteria, QS systems have been identified in more than 70 different bacterial species, including *Streptococcus pneumoniae*, *Bacillus subtilis*, and *Staphylococcus aureus* (Miller and Bassler, 2001; Waters and Bassler, 2005; Ng and Bassler, 2009). The QS systems control not only bioluminescence, but also other cooperative processes such as sporulation, conjugation, nutrient acquisition, biofilm formation, bio-corrosion, and antibiotics and toxins (Waters and Bassler, 2005; Kimura, 2014; Hense and Schuster, 2015). Remarkably, bacteria not only can communicate with members of the same species, but they are also able to sense the presence of different species in a community (interspecific communication). This interspecific communication is performed using a second type of autoinducer (AI-2). Thus, while each bacterial species has its own AI-1 to talk intraspecifically, AI-2 is common to all Gram-negative and Gram-positive bacteria. In fact AI-2 is not a single molecule but rather it refers to a group of molecules belonging to the family of interconverting furanones derived from 4,5-dihydroxy-2,3-pentanedione (DPD), whose biosynthesis is under the control of the enzyme LuxS (Xavier and Bassler, 2003). Development of novel compounds able to disrupt QS mechanisms has been carried out in recent years. For example QS quenching enzymes like lactonases and acylases are able to degrade acylated homoserine lactone (Dong and Zhang, 2005). A series of compounds, named halogenated furanones produced by many microbial species, mostly belonging to the Proteobacteria, can interfere with AHL and AI-2 QS pathways in Gram-negative and Gram-positive bacteria (Manefield et al., 2002; Rasko et al., 2008; Kayumov et al., 2014). Identification of the chemical signals, receptors, target genes, and mechanisms of signal transduction involved in quorum sensing are essential to our understanding how bacterial cell-cell communication may be used in preventing colonization by pathogenic bacteria. More data from metagenomic and metabolomics studies will help to decode the bacterial cross-talk and microbiome-immune system interplay, and particularly, distinctive regulatory mechanisms.

Targeting Dysbiosis

Fecal Transplantation

Antibiotics have been used to treat infectious diseases over the past century. However, it is clear that antibiotic treatment can render individuals more susceptible to infections (Dethlefsen et al., 2008; Forslund et al., 2013). High doses and frequent

use of antibiotics can disrupt and destabilize the normal bowel microbiota, predisposing patients to develop *Clostridium difficile* infections. Up to 35% of these patients develop a chronic recurrent pattern of disease. Fecal bacteriotherapy is the transplantation of liquid suspension of stool from a donor (usually a family member) and has been used successfully in severe cases of recurrent *C. difficile* relapse (Gough et al., 2011; Rupnik, 2015). However, many problems exist with this therapy since it can increase the risks of transmitting other pathogens (Brandt and Reddy, 2011).

Fecal transplantation studies in mice showed that transferring the microbiota from lean and fat mice to germ-free mice induces greater weight gain in those receiving the microbiota from fat donors (Ley et al., 2006). The discovery of the link between lean-associated microbiome has opened new possibility of using transplanted microbiota to treat metabolic disorders in humans.

Probiotics and Prebiotics

Probiotics are defined as live microorganisms that ultimately improve the balance of the intestinal flora, thus fostering healthy gut functions through a healthy gut microbiome (reviewed in Gareau et al., 2010; Whelan and Quigley, 2013). There are several *in vitro* assays to validate the actual *in vivo* efficacy of probiotic microorganisms, which include specific biological criteria, such as resistance to low gastric pH and capacity to reach the intestines alive to exert beneficial effects on the human body (Papadimitriou et al., 2015). Probiotic microorganisms are mainly lactic acid-producing bacteria of *Lactobacillus* and *Bifidobacterium* genera. Other microorganisms, such as the yeast *Saccharomyces boulardii* and the bacteria *E. coli* Nissle 1917, *Streptococcus thermophilus*, *F. parausnitzii* and *Bacillus polyfermenticus* have also been investigated. The beneficial therapeutic effects and mechanisms of action of Lactobacilli and bifidobacteria in patients with gastrointestinal diseases have long been demonstrated (Ng et al., 2009). These probiotics can prevent or ameliorate clinical symptoms of irritable bowel syndrome, inflammatory and necrotizing enterocolitis and acute diarrhea (Ng et al., 2009; Gareau et al., 2010; Whelan and Quigley, 2013). It was found that they could regulate the balance of intestinal microbiota by physically blocking the adhesion of pathogenic species onto epithelial cells. This is directly mediated by means of increases in the production of a mucosal barrier by goblet epithelial cells (Etzold et al., 2014). In addition, they can regulate epithelial permeability by enhancing the formation of tight-junctions between cells (Ng et al., 2009). Their immune-modulatory effects are associated with a decrease in the production of pro-inflammatory cytokines, as well as the microbial peptides bacteriocins (Ng et al., 2009; Whelan and Quigley, 2013).

The use of probiotics is not limited to gastrointestinal disorders. Studies evaluating their application in dermatology, urology and dentistry have been increasing (Vuotto et al., 2014). *Bifidobacterium bifidum* has been used in the prevention and treatment of infantile eczema. Intra-vaginal administration of *Lactobacillus rhamnosus GR-1 and L. fermentum* RC-14 were shown to have a positive effect on the prevention of recurrent BV and candidiasis (Anukam et al., 2006; Vuotto et al., 2014). Consumption of probiotics can be effective in the prevention of dental caries and periodontal diseases (Pandey et al., 2015). The continuous consumption of Yakult's *L. casei* strain Shirota (LcS), one of the most popular probiotics, in adequate amounts, may reduce the risk of cancers by modulating immune function (Ishikawa et al., 2005). Finally, the treatment of obese mice with *Bifidobacterium infantis* was shown to reduce the production of pro-inflammatory cytokines and white adipose tissue weight (Cani et al., 2007). The effect of the endogenous host microbiota on obesity and beneficial role of probiotics including *L. rhamnosus* and *gasseri* and *Bifidobacterium lactis* in the treatment of adiposity and obesity has been reviewed elsewhere (Mekkes et al., 2014). This is a new area under intense investigation.

Prebiotics are functional food ingredients that can change the composition and/or the activity of the colonic flora (Roberfroid, 2000; Roberfroid, 2007; Brownawell et al., 2012). The dietary supplementation with prebiotics can promote the growth of beneficial bacteria such as lactobacilli and bifidobacteria strains (Roberfroid, 2000, 2007). Poorly digestible carbohydrates (fibers), such as resistant starch, non-starch polysaccharides (e.g., celluloses, hemicelluloses, pectins, and gums), oligosaccharides and polyphenols are resistant to gastric acidity, gastrointestinal absorption, and non-digestible by hydrolysis by mammalian enzymes. Colonic bacteria through carbohydrate hydrolyzing enzymes and fermentation produce hydrogen, methane, carbon dioxide, and SCFA, which can affect host energy levels and gut hormone regulation (Slavin, 2013). The most commonly used prebiotics are fructo-oligosaccharides (FOS) and trans-galacto-oligosaccharides (TOS), for example inulin (Roberfroid, 2000). However, not all dietary carbohydrates are prebiotics (Roberfroid, 2007). Mixtures of probiotic and prebiotic ingredients have been used to selectively stimulate growth or activity of health-promoting bacteria. In conclusion, it appears that the therapeutic use of pro- and prebiotics will find more applications in the near future when large-scale clinical trials and metagenomic surveys will determine which microbes are active, which are damaged, and which may respond to a given prebiotic, probiotic or synbiotic (synergic association of probiotic and prebiotic) at the genomic level (Nagata et al., 2011).

Phage Therapy and CRISPRs

Phage therapy consists of using bacterial viruses bacteriophages, (also known as phages) as antimicrobial agents (Sulakvelidze et al., 2001; Abedon, 2014). Bacteriophages attach to specific receptors present in the host membrane and then inject their genetic material into the bacterium. Viral proteins are then synthetized using the host's translational machinery. Phage infection can result in lysis, lysogeny or resistance. Lytic bacteriophages induce host cell death and breakdown in order to spread the infection whereas lysogenic (or temperate) phages insert their genome into the host DNA. Resistance may be acquired during replicative cycles by gene transposition or recombination.

Phage therapy can potentially have beneficial impact on human microbiomes and host health (Koskella and Meaden, 2013). The host specificity greatly limits the types of bacteria that will enter into contact with a particular phage, therefore

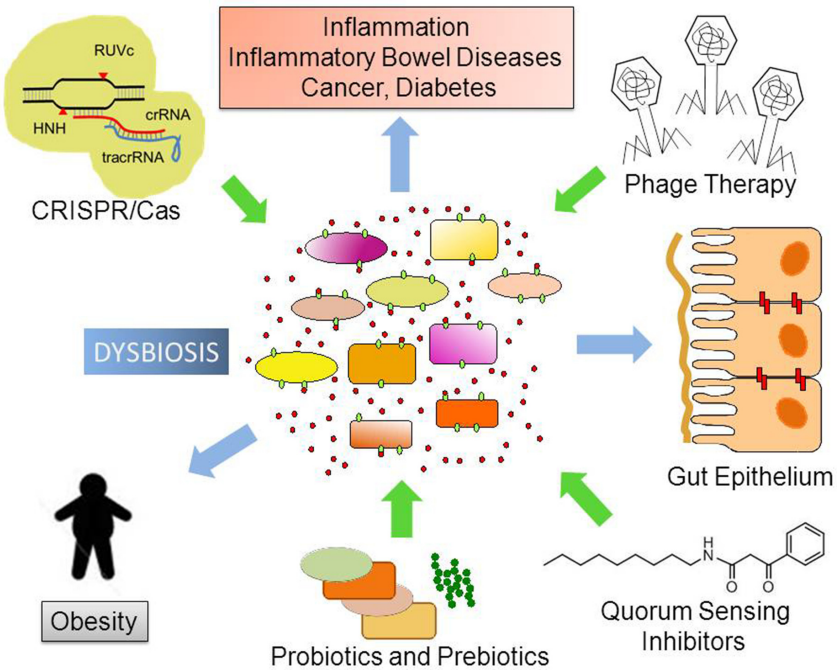

FIGURE 2 | Possible future therapeutic approaches to control dysbiosis. Disturbances in the ecological community of commensal, symbiotic, and pathogenic microorganisms may favor dysbiosis. This leads to increased bacterial translocation and/or release of microorganism-associated molecular patterns (MAMPs), which activate Toll-like receptors (TLRs) in several cell types. Local and body-wide immune system activities promote inflammation, which ultimately leads to chronic diseases. Novel therapeutic approaches like phage therapy, disruption of Quorum Sensing and the use of the biotechnological tool CRISPR/Cas9 to edit microbial genomes have the potential to target specific bacterial taxa thus helping to re-establish homeostasis and microbiome balance.

avoiding the elimination of non-pathogenic species (Koskella and Meaden, 2013). However in order to choose a specific phage to use as a therapeutic agent, it is necessary to know the pathogen causing a given disease. When this is not the case, the use of a cocktail of different species of phages would broaden the range of action but could also have a possible negative effect on the microbial communities (Chan et al., 2013). The synergistic use of phages and low dose of antibiotics, a strategy named Phage-Antibiotic Synergy (PAS), could be useful in certain clinical situations (Comeau et al., 2007).

Bacteria have evolved various mechanisms of defense against phage infections, which act at different levels. In fact they can prevent phage attachment by mutation/loss of membrane receptors or block phage DNA entry with the aid of specific membrane proteins. Furthermore, Bacteria and Archaea developed an intrinsic innate immunity mechanism, which allows them to remember phage infection by capturing short DNA sequences from phage genetic material. These viral sequences are integrated as spacer sequences into their own chromosome, specifically into an array of repeated sequences called Clustered Regularly Interspaced Short Palindromic Repeats or CRISPR, with the help of the proteins encoded by Cas (CRISPR-associated) family of genes (Garneau et al., 2010; van der Oost et al., 2014).

CRISPR loci consist of short (~24–48 nucleotides) repeats separated by similarly sized, unique spacers found in genomes of Archaea (~90%), and Bacteria (~40%) (Garneau et al., 2010; van der Oost et al., 2014). Cas genes encode a large and heterogeneous family of proteins with functional domains typical of nucleases, helicases, polymerases, and polynucleotide-binding proteins. Upon invasion, the host organism samples and integrates in its genome short fragments of the foreign DNA, called protospacers, thus creating immunity against that particular infective agent. The protospacer is flanked by the repeated regions, and transcribed with them into a CRISPR-RNA (crRNA), which guides specific nucleases to a target DNA containing regions complementary to the protospacer. Upon recognition, nucleases cleave invasive DNA preventing it to replicate and blocking infection. Three different types and eleven subtypes of CRIPSR/Cas system can be classified based on their Cas protein repertoire and mechanisms of action (Plagens et al., 2015). A detailed list of type I, II, and III CRISPR-Cas systems is also available at the CRISPRdb website[7]. Type I systems are characterized by the molecular machinery named a Cascade complex (CRISPR-associated complex for antiviral defense) which displays nickase and exonuclease activities. Type III systems are characterized by the presence of Cas10 (the signature protein) and associated proteins. The systems are subclassifed as type III-A (CSM) and type III-B (CMR), depending on their specificity for DNA or RNA targets. In addition, types I and III share a variable number of repeat associated mysterious protein (RAMP) subunits (Rouillon et al., 2013; Plagens et al., 2015).

[7]http://crispr.u-psud.fr/

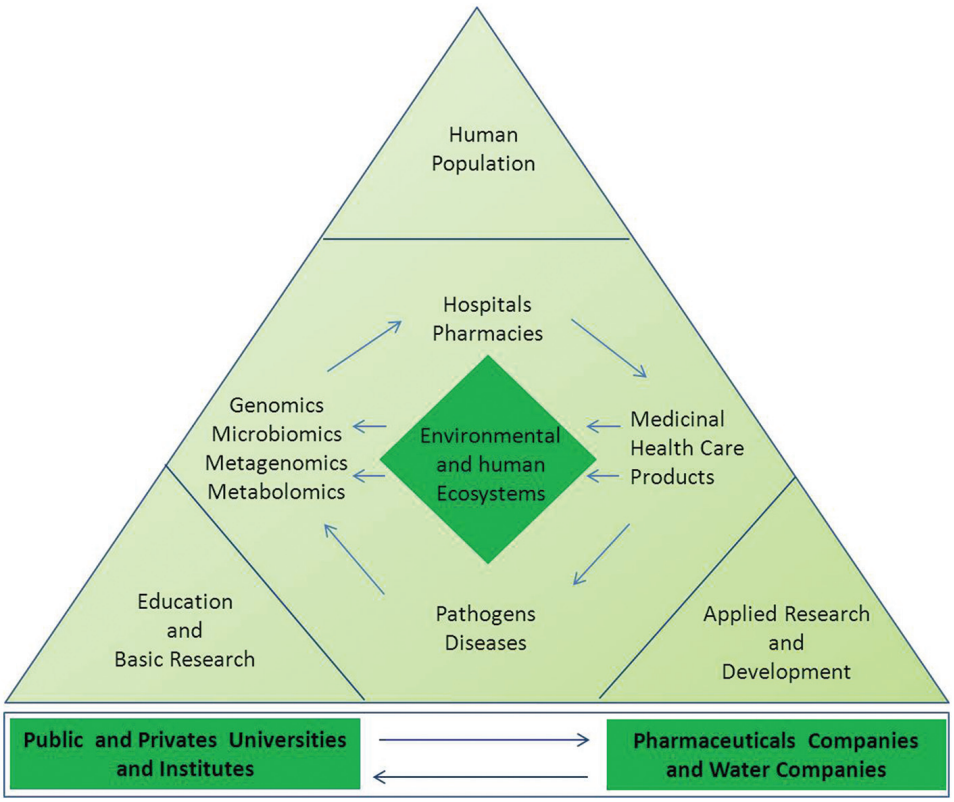

FIGURE 3 | Actions and molecular approaches aiming to protect the environmental and human microbial ecosystems. The measurements of ecological, phylometagenomic, and microbial metabolic variations in the microbiomes require a specialized and complex set of knowledge. Collaboration between universities, research entities, non-governmental organizations (NGO), and the pharmaceutical industry professionals are key for evaluating both biological and pharmaceutical impacts in the ecosystems and elucidating the mechanism-of-action of new compounds in the host and its microbiomes. The utility of metagenomic functional reconstruction for direct association of community functions with habitat and host phenotype will be critical for proper study designs and production of greener pharmaceutical products for future personalized medicine.

Type II is the simplest CRISPR-Cas system that is characterized by the presence of dsDNA endonuclease Cas9 and the transactivating CRISPR-RNA (tracrRNA). The tracrRNA anneals with the invariable regions of mature crRNA creating RNA heterodimers which, in turn, forms a nucleoprotein complex with Cas9, guiding it to the target DNA. Cas9 recognizes and binds to a specific 5′-NGG-3′ motif, called protospacer adjacent motif (PAM). Then the complex searches for a sequence complementary to the spacer portion of crRNA. Cas9 contains two nuclease domains, namely RuvC and HNH, and produces a double strand break in the target. Subsequently, cleaved DNA becomes a substrate of the bacterial DNA repair mechanisms, either non-homologous end joining (NHEJ) or homologous recombination (HR). NHEJ is an imperfect repair system and may cause insertion or deletion (indels) of base pairs, as well as single nucleotide polymorphisms (SNPs). However, high-fidelity HR repair may occur if a sequence complementary to the cleaved fragment is provided. The relative simplicity of the mechanism of action and the peculiarities of Cas9 make the CRISPR/Cas9 system an ideal tool for a vast assortment of procedures, particularly for genomic editing (reviewed in Ma et al., 2014; Selle and Barrangou, 2015; Xiao-Jie et al., 2015). A considerable amount of work in this field has been already done in different organisms, especially eukaryotes, using engineered versions of CRISPR/Cas9. On the other hand, despite its enormous potential, manipulation of bacterial genomes by CRISPR/Cas9 has so far been scarcely executed (Selle and Barrangou, 2015). CRISPR/Cas9 can be used to selectively deplete a given bacterial community of a particular harmful strain or species (Vercoe et al., 2013; Gomaa et al., 2014; Yosef et al., 2015). It has been shown that there is an inverse correlation between the presence of CRISPR loci and acquired antibiotic resistance in *Enterococcus fecalis* (Palmer and Gilmore, 2010), indicating that the use of antibiotics may increase the ability of bacteria to acquire drug resistance-encoding plasmids. CRISPR/Cas9 system can be used to introduce specific mutations into essential, antibiotic resistance, and virulence genes. It has been already shown that by providing *in trans* a DNA (linear or plasmid) homologous to the target sequence, it is possible to introduce very specific mutations to the desired target (Marraffini and Sontheimer, 2008; Jiang et al., 2013; Yosef et al., 2015). Also CRISPR/Cas9 has the potential to directly modulate the expression of particular genes. An engineered version of Cas9 lacking the nuclease

activity but still retaining its binding capacity (dCas9) has already been created to repress bacterial transcription by binding to promoter regions or within a ORF, thus blocking transcriptional initiation and elongation, respectively. dCas9 can also be fused to regulatory domains in order to switch on/off the expression of specific genes (Bikard et al., 2013; Qi et al., 2013). In the near future the engineering of commensal bacteria with improved properties using a CRISPR/Cas system may constitute an effective vaccination tool in public health for prevention of diseases. However, despite great advances, still much work needs to be done in order to improve target specificity and delivering efficiency.

A more complete perspective on how phage therapy and CRISPR/Cas9 systems can be employed to combat pathogenic species within our bodies, especially antibiotic-resistant bacterial pathogens needs the expansion of *in vitro*, *ex vivo,* and *in silico* approaches (Fritz et al., 2013). Several publicly available methods for hit-specific retrieval of protospacers in the reference microbiomes have already been developed (Bi et al., 2012). Over 123,003 protospacers have been predicted based on 690 phage genomes (Zhang et al., 2013). The functional exploration of pathogen-specific bacteriophages and gene therapy depends on development of relevant animal models including transgenic and bacteria-free animals (Fritz et al., 2013). Finally, we will need to confirm the results in the proof-of-concept in well-designed clinical trials. In **Figure 2**, we graphically summarize some of the pharmacological approaches discussed in this article.

Ecopharmacology

To assess the interaction of the human body with pharmaceuticals, we need to understand the complex relationship between ecology, physiology, and pharmacology (Rahman et al., 2007; Flintoft, 2012; Haiser and Turnbaugh, 2012). From pharmacogenomic studies it is clear that sequence variations in drug target proteins, drug-metabolizing enzymes, and drug transporters can alter drug efficacy, produce side effects, causing variable drug responses in individual patients (Wilson and Nicholson, 2009). Microorganisms participate in a very wide range of biotransformations, including hydrolysis, and processing of glutathione conjugates of xenobiotics excreted in the bile (Johnson et al., 2012). Hence, the determination of the genetic variability of human microbiomes has potential to predict the efficacy, bioavailability and individual response variability in drug therapy.

Finally, further studies are needed to elucidate whether the vast number of functional microbiota gene-products exerts unknown off-target effects and how they can negatively or positively affect drug responses. These are the major research challenges for exploring the potential of metagenomics to better understand microbial ecology and to translate the molecular and genomic data into pharmacomicrobiomics (Saad et al., 2012). According to this new ecological paradigm, competency in knowledge, skills, and attitudes as well as integrated environmental conscience and social responsibility are essential for professionals who will in the future create and develop a new generation of green and sustainable pharmaceutical products, as shown in **Figure 3**.

Conclusion and Perspectives

Recent advances in microbiome sequencing projects revealed the high complexity of microbial communities in various human body sites. They have confirmed the critical roles of the human-microbiota ecosystems in health-promoting or disease-causing processes. These studies have highlighted the unexpected and wide-ranging consequences of eliminating certain bacteria living in our body.

While the natural variation of the human microbiota has yet to be fully determined, the annotation and analyses of a large number of human microbiomes have shown that the presence or absence of specific microbial species categorizes human individuals based on enterotypes. It is likely that cultivated and uncultivated microbes will contribute to discovering new fundamental biomarkers for specific human disorders and that they may become better discriminatory tools than human-based ones.

Changes in the stability and dynamic of numerous microbial communities have been associated with several diseases, including type II diabetes, obesity, fatty liver disease, irritable bowel syndrome, and IBDs and even certain cancers. However, further studies need to be done in order to confirm whether low bacterial diversity increases the chances to develop such diseases and metabolic perturbations.

The use of antibiotics compromises genome defense and increases the ability to acquire antibiotic resistance. Prebiotics, probiotics, synbiotics, phage therapy, quorum sensing systems, and CRISPR/Cas systems have been proposed as tools to control and modulate microbial communities. Engineering of pathogen-specific bacteriophages and production of pharmaceuticals based on our own body's microbiome will be possible and fully explored in the near future. The use of novel pharmaceuticals and nutraceuticals to modulate microbial colonization and development of a healthy gut microbial community in early childhood will support healthy adult human body functions and prevent the occurrence of several diseases.

Author's Contribution

The authors conducted the literature review process, grading, and categorizing criteria, and quality of selected articles. The authors read and approved the final manuscript.

Acknowledgments

We thank colleagues of Institute of Biomedical Sciences of the University of São Paulo for insights and productive discussions. This work was supported by grants from Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP, proc. 2015/1177-8, 2015/18647-6) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Aagaard, K., Ma, J., Antony, K. M., Ganu, R., Petrosino, J., and Versalovic, J. (2014). The placenta harbors a unique microbiome. *Sci. Transl. Med.* 6:237ra265. doi: 10.1126/scitranslmed.3008599

Abedon, S. T. (2014). Phage therapy: eco-physiological pharmacology. *Scientifica (Cairo)* 2014:581639. doi: 10.1155/2014/581639

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Anukam, K. C., Osazuwa, E., Osemene, G. I., Ehigiagbe, F., Bruce, A. W., and Reid, G. (2006). Clinical study comparing probiotic *Lactobacillus* GR-1 and RC-14 with metronidazole vaginal gel to treat symptomatic bacterial vaginosis. *Microbes Infect.* 8, 2772–2776. doi: 10.1016/j.micinf.2006.08.008

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Backhed, F., Fraser, C. M., Ringel, Y., Sanders, M. E., Sartor, R. B., Sherman, P. M., et al. (2012). Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell Host Microbe* 12, 611–622. doi: 10.1016/j.chom.2012.10.012

Belizario, J. E. (2013). The humankind genome: from genetic diversity to the origin of human diseases. *Genome* 56, 705–716. doi: 10.1139/gen-2013-0125

Bi, D., Xu, Z., Harrison, E. M., Tai, C., Wei, Y., He, X., et al. (2012). ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res.* 40, D621–D626. doi: 10.1093/nar/gkr846

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L. A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* 41, 7429–7437. doi: 10.1093/nar/gkt520

Brandt, L. J., and Reddy, S. S. (2011). Fecal microbiota transplantation for recurrent *Clostridium difficile* infection. *J. Clin. Gastroenterol.* 45(Suppl.), S159–S167. doi: 10.1097/MCG.0b013e318222e603

Brown, C. T., Sharon, I., Thomas, B. C., Castelle, C. J., Morowitz, M. J., and Banfield, J. F. (2013). Genome resolved analysis of a premature infant gut microbial community reveals a *Varibaculum cambriense* genome and a shift towards fermentation-based metabolism during the third week of life. *Microbiome* 1:30. doi: 10.1186/2049-2618-1-30

Brownawell, A. M., Caers, W., Gibson, G. R., Kendall, C. W., Lewis, K. D., Ringel, Y., et al. (2012). Prebiotics and the health benefits of fiber: current regulatory status, future research, and goals. *J. Nutr.* 142, 962–974. doi: 10.3945/jn.112.158147

Cani, P. D., and Delzenne, N. M. (2011). The gut microbiome as therapeutic target. *Pharmacol. Ther.* 130, 202–212. doi: 10.1016/j.pharmthera.2011.01.012

Cani, P. D., Neyrinck, A. M., Fava, F., Knauf, C., Burcelin, R. G., Tuohy, K. M., et al. (2007). Selective increases of bifidobacteria in gut microflora improve high-fat-diet-induced diabetes in mice through a mechanism associated with endotoxaemia. *Diabetologia* 50, 2374–2383. doi: 10.1007/s00125-007-0791-0

Carr, R., Shen-Orr, S. S., and Borenstein, E. (2013). Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS Comput. Biol.* 9:e1003292. doi: 10.1371/journal.pcbi.1003292

Chan, B. K., Abedon, S. T., and Loc-Carrillo, C. (2013). Phage cocktails and the future of phage therapy. *Future Microbiol* 8, 769–783. doi: 10.2217/fmb.13.47

Christen, B., Abeliuk, E., Collier, J. M., Kalogeraki, V. S., Passarelli, B., Coller, J. A., et al. (2011). The essential genome of a bacterium. *Mol. Syst. Biol.* 7:528. doi: 10.1038/msb.2011.58

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., et al. (2015). The microbiome of uncontacted Amerindians. *Sci. Adv.* 1:e1500183. doi: 10.1126/sciadv.1500183

Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell* 148, 1258–1270. doi: 10.1016/j.cell.2012.01.035

Collison, M., Hirt, R. P., Wipat, A., Nakjang, S., Sanseau, P., and Brown, J. R. (2012). Data mining the human gut microbiota for therapeutic targets. *Brief. Bioinform.* 13, 751–768. doi: 10.1093/bib/bbs002

Comeau, A. M., Tetart, F., Trojet, S. N., Prere, M. F., and Krisch, H. M. (2007). Phage-antibiotic synergy (PAS): beta-lactam and quinolone antibiotics stimulate virulent phage growth. *PLoS ONE* 2:e799. doi: 10.1371/journal.pone.0000799

Cucchiara, S., Stronati, L., and Aloi, M. (2012). Interactions between intestinal microbiota and innate immune system in pediatric inflammatory bowel disease. *J. Clin. Gastroenterol.* 46(Suppl.), S64–S66. doi: 10.1097/MCG.0b013e31826a857f

Culligan, E. P., Sleator, R. D., Marchesi, J. R., and Hill, C. (2014). Metagenomics and novel gene discovery: promise and potential for novel therapeutics. *Virulence* 5, 399–412. doi: 10.4161/viru.27208

Curtis, T. P., Sloan, W. T., and Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10494–10499. doi: 10.1073/pnas.142680199

Dawid, S., Roche, A. M., and Weiser, J. N. (2007). The blp bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both in vitro and in vivo. *Infect. Immun.* 75, 443–451. doi: 10.1128/IAI.01775-05

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

Defoirdt, T., Boon, N., and Bossier, P. (2010). Can bacteria evolve resistance to quorum sensing disruption? *PLoS Pathog.* 6:e1000989. doi: 10.1371/journal.ppat.1000989

Delaney, M. L., and Onderdonk, A. B. (2001). Nugent score related to vaginal culture in pregnant women. *Obstet. Gynecol.* 98, 79–84. doi: 10.1016/S0029-7844(01)01402-8

Dethlefsen, L., Huse, S., Sogin, M. L., and Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* 6:e280. doi: 10.1371/journal.pbio.0060280

DiGiulio, D. B., Romero, R., Amogan, H. P., Kusanovic, J. P., Bik, E. M., Gotsch, F., et al. (2008). Microbial prevalence, diversity and abundance in amniotic fluid during preterm labor: a molecular and culture-based investigation. *PLoS ONE* 3:e3056. doi: 10.1371/journal.pone.0003056

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11971–11975. doi: 10.1073/pnas.1002601107

Dong, Y. H., and Zhang, L. H. (2005). Quorum sensing and quorum-quenching enzymes. *J. Microbiol.* 43, 101–109.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Etzold, S., Kober, O. I., Mackenzie, D. A., Tailford, L. E., Gunning, A. P., Walshaw, J., et al. (2014). Structural basis for adaptation of lactobacilli to gastrointestinal mucus. *Environ. Microbiol.* 16, 888–903. doi: 10.1111/1462-2920.12377

Fardini, Y., Chung, P., Dumm, R., Joshi, N., and Han, Y. W. (2010). Transmission of diverse oral bacteria to murine placenta: evidence for the oral microbiome as a potential source of intrauterine infection. *Infect. Immun.* 78, 1789–1796. doi: 10.1128/IAI.01395-09

Fettweis, J. M., Serrano, M. G., Sheth, N. U., Mayer, C. M., Glascock, A. L., Brooks, J. P., et al. (2012). Species-level classification of the vaginal microbiome. *BMC Genomics* 13(Suppl. 8):S17. doi: 10.1186/1471-2164-13-S8-S17

Finch, R. G., Pritchard, D. I., Bycroft, B. W., Williams, P., and Stewart, G. S. (1998). Quorum sensing: a novel target for anti-infective therapy. *J. Antimicrob. Chemother.* 42, 569–571. doi: 10.1093/jac/42.5.569

Flintoft, L. (2012). Disease genomics: associations go metagenome-wide. *Nat. Rev. Genet.* 13, 756–757. doi: 10.1038/nrg3347

Fodor, A. A., DeSantis, T. Z., Wylie, K. M., Badger, J. H., Ye, Y., Hepburn, T., et al. (2012). The "most wanted" taxa from the human microbiome for whole genome sequencing. *PLoS ONE* 7:e41294. doi: 10.1371/journal.pone.0041294

Forsberg, K. J., Patel, S., Gibson, M. K., Lauber, C. L., Knight, R., Fierer, N., et al. (2014). Bacterial phylogeny structures soil resistomes across habitats. *Nature* 509, 612–616. doi: 10.1038/nature13377

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, A., et al. (2013). Country-specific antibiotic use practices impact the human gut resistome. *Genome Res.* 23, 1163–1169. doi: 10.1101/gr.155465.113

Frantz, A. L., Rogier, E. W., Weber, C. R., Shen, L., Cohen, D. A., Fenton, L. A., et al. (2012). Targeted deletion of MyD88 in intestinal epithelial cells results in compromised antibacterial immunity associated with downregulation of polymeric immunoglobulin receptor, mucin-2, and antibacterial peptides. *Mucosal Immunol.* 5, 501–512. doi: 10.1038/mi.2012.23

Fritz, J. V., Desai, M. S., Shah, P., Schneider, J. G., and Wilmes, P. (2013). From meta-omics to causality: experimental models for human microbiome research. *Microbiome* 1:14. doi: 10.1186/2049-2618-1-14

Gareau, M. G., Sherman, P. M., and Walker, W. A. (2010). Probiotics and the gut microbiota in intestinal health and disease. *Nat. Rev. Gastroenterol. Hepatol.* 7, 503–514. doi: 10.1038/nrgastro.2010.117

Garneau, J. E., Dupuis, M. E., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., et al. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468, 67–71. doi: 10.1038/nature09523

Gevers, D., Pop, M., Schloss, P. D., and Huttenhower, C. (2012). Bioinformatics for the human microbiome project. *PLoS Comput. Biol.* 8:e1002779. doi: 10.1371/journal.pcbi.1002779

Gomaa, A. A., Klumpe, H. E., Luo, M. L., Selle, K., Barrangou, R., and Beisel, C. L. (2014). Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *MBio* 5:e928-13. doi: 10.1128/mBio.00928-13

Gough, E., Shaikh, H., and Manges, A. R. (2011). Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent *Clostridium difficile* infection. *Clin. Infect. Dis.* 53, 994–1002. doi: 10.1093/cid/cir632

Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., et al. (2009). Topographical and temporal diversity of the human skin microbiome. *Science* 324, 1190–1192. doi: 10.1126/science.1171700

Groer, M. W., Luciano, A. A., Dishaw, L. J., Ashmeade, T. L., Miller, E., and Gilbert, J. A. (2014). Development of the preterm infant gut microbiome: a research priority. *Microbiome* 2:38. doi: 10.1186/2049-2618-2-38

Guani-Guerra, E., Santos-Mendoza, T., Lugo-Reyes, S. O., and Teran, L. M. (2010). Antimicrobial peptides: general overview and clinical implications in human health and disease. *Clin. Immunol.* 135, 1–11. doi: 10.1016/j.clim.2009.12.004

Guaraldi, F., and Salvatori, G. (2012). Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front. Cell Infect. Microbiol.* 2:94. doi: 10.3389/fcimb.2012.00094

Guo, F., Ju, F., Cai, L., and Zhang, T. (2013). Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PLoS ONE* 8:e76185. doi: 10.1371/journal.pone.0076185

Haiser, H. J., and Turnbaugh, P. J. (2012). Is it time for a metagenomic basis of therapeutics? *Science* 336, 1253–1255. doi: 10.1126/science.1224396

Hense, B. A., and Schuster, M. (2015). Core principles of bacterial autoinducer systems. *Microbiol. Mol. Biol. Rev.* 79, 153–169. doi: 10.1128/MMBR.00024-14

Hooper, L. V., Littman, D. R., and Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science* 336, 1268–1273. doi: 10.1126/science.1223490

Human Microbiome Jumpstart Reference Strains, C., Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., et al. (2010). A catalog of reference genomes from the human microbiome. *Science* 328, 994–999. doi: 10.1126/science.1183605

Human Microbiome Project, C. (2012a). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Human Microbiome Project, C. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Ishikawa, H., Akedo, I., Otani, T., Suzuki, T., Nakamura, T., Takeyama, I., et al. (2005). Randomized trial of dietary fiber and *Lactobacillus casei* administration for prevention of colorectal tumors. *Int. J. Cancer* 116, 762–767. doi: 10.1002/ijc.21115

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L. A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* 31, 233–239. doi: 10.1038/nbt.2508

Jiang, W., Ling, Z., Lin, X., Chen, Y., Zhang, J., Yu, J., et al. (2014). Pyrosequencing analysis of oral microbiota shifting in various caries states in childhood. *Microb. Ecol.* 67, 962–969. doi: 10.1007/s00248-014-0372-y

Joelsson, A., Liu, Z., and Zhu, J. (2006). Genetic and phenotypic diversity of quorum-sensing systems in clinical and environmental isolates of *Vibrio cholerae*. *Infect. Immun.* 74, 1141–1147. doi: 10.1128/IAI.74.2.1141-1147.2006

Johnson, C. H., Patterson, A. D., Idle, J. R., and Gonzalez, F. J. (2012). Xenobiotic metabolomics: major impact on the metabolome. *Annu. Rev. Pharmacol. Toxicol.* 52, 37–56. doi: 10.1146/annurev-pharmtox-010611-134748

Jones, B. V., and Marchesi, J. R. (2007). Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods* 4, 55–61. doi: 10.1038/nmeth964

Kardos, N., and Demain, A. L. (2011). Penicillin: the medicine with the greatest impact on therapeutic outcomes. *Appl. Microbiol. Biotechnol.* 92, 677–687. doi: 10.1007/s00253-011-3587-6

Kayumov, A. R., Khakimullina, E. N., Sharafutdinov, I. S., Trizna, E. Y., Latypova, L. Z., Thi Lien, H., et al. (2014). Inhibition of biofilm formation in *Bacillus subtilis* by new halogenated furanones. *J. Antibiot. (Tokyo)* 68, 297–301. doi: 10.1038/ja.2014.143

Kenyon, C., Colebunders, R., and Crucitti, T. (2013). The global epidemiology of bacterial vaginosis: a systematic review. *Am. J. Obstet. Gynecol.* 209, 505–523. doi: 10.1016/j.ajog.2013.05.006

Kimura, N. (2014). Metagenomic approaches to understanding phylogenetic diversity in quorum sensing. *Virulence* 5, 433–442. doi: 10.4161/viru.27850

Kootte, R. S., Vrieze, A., Holleman, F., Dallinga-Thie, G. M., Zoetendal, E. G., de Vos, W. M., et al. (2012). The therapeutic potential of manipulating gut microbiota in obesity and type 2 diabetes mellitus. *Diabetes Obes. Metab.* 14, 112–120. doi: 10.1111/j.1463-1326.2011.01483.x

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., et al. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* 9:e1002863. doi: 10.1371/journal.pcbi.1002863

Koskella, B., and Meaden, S. (2013). Understanding bacteriophage specificity in natural microbial communities. *Viruses* 5, 806–823. doi: 10.3390/v5030806

Ladizinski, B., McLean, R., Lee, K. C., Elpern, D. J., and Eron, L. (2014). The human skin microbiome. *Int. J. Dermatol.* 53, 1177–1179. doi: 10.1111/ijd.12609

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541–546. doi: 10.1038/nature12506

Lepage, P., Leclerc, M. C., Joossens, M., Mondot, S., Blottiere, H. M., Raes, J., et al. (2013). A metagenomic insight into our gut's microbiome. *Gut* 62, 146–158. doi: 10.1136/gutjnl-2011-301805

Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11070–11075. doi: 10.1073/pnas.0504978102

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., et al. (2008). Evolution of mammals and their gut microbes. *Science* 320, 1647–1651. doi: 10.1126/science.1155725

Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444, 1022–1023. doi: 10.1038/4441022a

Ma, Y., Zhang, L., and Huang, X. (2014). Genome modification by CRISPR/Cas9. *FEBS J.* 281, 5186–5193. doi: 10.1111/febs.13110

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., and Woese, C. R. (1996). The ribosomal database project (RDP). *Nucleic Acids Res.* 24, 82–85. doi: 10.1093/nar/24.1.82

Manefield, M., Rasmussen, T. B., Henzter, M., Andersen, J. B., Steinberg, P., Kjelleberg, S., et al. (2002). Halogenated furanones inhibit quorum sensing through accelerated LuxR turnover. *Microbiology* 148(Pt 4), 1119–1127. doi: 10.1099/00221287-148-4-1119

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., et al. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* 55, 205–211. doi: 10.1136/gut.2005.073817

Manor, O., and Borenstein, E. (2015). MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol.* 16:53. doi: 10.1186/s13059-015-0610-8

Marchesi, J. R., Sato, T., Weightman, A. J., Martin, T. A., Fry, J. C., Hiom, S. J., et al. (1998). Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Appl Environ. Microbiol.* 64, 795–799.

Marraffini, L. A., and Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322, 1843–1845. doi: 10.1126/science.1165771

Mason, M. R., Nagaraja, H. N., Camerlengo, T., Joshi, V., and Kumar, P. S. (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PLoS ONE* 8:e77287. doi: 10.1371/journal.pone.0077287

Meijer, K., de Vos, P., and Priebe, M. G. (2010). Butyrate and other short-chain fatty acids as modulators of immunity: what relevance for health? *Curr. Opin. Clin. Nutr. Metab. Care* 13, 715–721. doi: 10.1097/MCO.0b013e32833eebe5

Mekkes, M. C., Weenen, T. C., Brummer, R. J., and Claassen, E. (2014). The development of probiotic treatment in obesity: a review. *Benef. Microbes* 5, 19–28. doi: 10.3920/BM2012.0069

Miller, M. B., and Bassler, B. L. (2001). Quorum sensing in bacteria. *Annu. Rev. Microbiol.* 55, 165–199. doi: 10.1146/annurev.micro.55.1.165

Mobegi, F. M., van Hijum, S. A., Burghout, P., Bootsma, H. J., de Vries, S. P., van der Gaast-de Jongh, C. E., et al. (2014). From microbial gene essentiality to novel antimicrobial drug targets. *BMC Genomics* 15:958. doi: 10.1186/1471-2164-15-958

Montassier, E., Batard, E., Massart, S., Gastinne, T., Carton, T., Caillon, J., et al. (2014). 16S rRNA gene pyrosequencing reveals shift in patient faecal microbiota during high-dose chemotherapy as conditioning regimen for bone marrow transplantation. *Microb. Ecol.* 67, 690–699. doi: 10.1007/s00248-013-0355-4

Mullany, P. (2014). Functional metagenomics for the investigation of antibiotic resistance. *Virulence* 5, 443–447. doi: 10.4161/viru.28196

Nagata, S., Asahara, T., Ohta, T., Yamada, T., Kondo, S., Bian, L., et al. (2011). Effect of the continuous intake of probiotic-fermented milk containing *Lactobacillus casei* strain Shirota on fever in a mass outbreak of norovirus gastroenteritis and the faecal microflora in a health service facility for the aged. *Br. J. Nutr.* 106, 549–556. doi: 10.1017/S000711451100064X

Naik, S., Bouladoux, N., Wilhelm, C., Molloy, M. J., Salcedo, R., Kastenmuller, W., et al. (2012). Compartmentalized control of skin immunity by resident commensals. *Science* 337, 1115–1119. doi: 10.1126/science.1225152

Nealson, K. H., and Hastings, J. W. (1979). Bacterial bioluminescence: its control and ecological significance. *Microbiol. Rev.* 43, 496–518.

Ng, S. C., Hart, A. L., Kamm, M. A., Stagg, A. J., and Knight, S. C. (2009). Mechanisms of action of probiotics: recent advances. *Inflamm. Bowel Dis.* 15, 300–310. doi: 10.1002/ibd.20602

Ng, W. L., and Bassler, B. L. (2009). Bacterial quorum-sensing network architectures. *Annu. Rev. Genet.* 43, 197–222. doi: 10.1146/annurev-genet-102108-134304

Ochsner, U. A., Sun, X., Jarvis, T., Critchley, I., and Janjic, N. (2007). Aminoacyl-tRNA synthetases: essential and still promising targets for new anti-infective agents. *Expert Opin. Investig. Drugs* 16, 573–593. doi: 10.1517/13543784.16.5.573

Ostaff, M. J., Stange, E. F., and Wehkamp, J. (2013). Antimicrobial peptides and gut microbiota in homeostasis and pathology. *EMBO Mol. Med.* 5, 1465–1483. doi: 10.1002/emmm.201201773

Palmer, D. J., Metcalfe, J., and Prescott, S. L. (2012). Preventing disease in the 21st century: the importance of maternal and early infant diet and nutrition. *J. Allergy Clin. Immunol.* 130, 733–734. doi: 10.1016/j.jaci.2012.06.038

Palmer, K. L., and Gilmore, M. S. (2010). Multidrug-resistant enterococci lack CRISPR-cas. *mBio* 1:e00227-10. doi: 10.1128/mBio.00227-10

Pandey, V., Berwal, V., Solanki, N., and Malik, N. S. (2015). Probiotics: healthy bugs and nourishing elements of diet. *J. Int. Soc. Prev. Community Dent.* 5, 81–87. doi: 10.4103/2231-0762.155726

Papadimitriou, K., Zoumpopoulou, G., Foligne, B., Alexandraki, V., Kazou, M., Pot, B., et al. (2015). Discovering probiotic microorganisms: *in vitro*, *in vivo*, genetic and omics approaches. *Front. Microbiol.* 6:58. doi: 10.3389/fmicb.2015.00058

Paredes-Sabja, D., Shen, A., and Sorg, J. A. (2014). *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. *Trends Microbiol.* 22, 406–416. doi: 10.1016/j.tim.2014.04.003

Parsons, J. B., Broussard, T. C., Bose, J. L., Rosch, J. W., Jackson, P., Subramanian, C., et al. (2014). Identification of a two-component fatty acid kinase responsible for host fatty acid incorporation by *Staphylococcus aureus*. *Proc. Natl. Acad. Sci. U.S.A.* 111, 10532–10537. doi: 10.1073/pnas.1408797111

Perez-Chaparro, P. J., Goncalves, C., Figueiredo, L. C., Faveri, M., Lobao, E., Tamashiro, N., et al. (2014). Newly identified pathogens associated with periodontitis: a systematic review. *J. Dent. Res.* 93, 846–858. doi: 10.1177/0022034514542468

Plagens, A., Richter, H., Charpentier, E., and Randau, L. (2015). DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiol. Rev.* 3, 442–463. doi: 10.1093/femsre/fuv019

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196. doi: 10.1093/nar/gkm864

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., et al. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173–1183. doi: 10.1016/j.cell.2013.02.022

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Rahman, S. Z., Khan, R. A., Gupta, V., and Uddin, M. (2007). Pharmacoenvironmentology–a component of pharmacovigilance. *Environ. Health* 6:20. doi: 10.1186/1476-069X-6-20

Rasko, D. A., Moreira, C. G., Li de, R., Reading, N. C., Ritchie, J. M., Waldor, M. K., et al. (2008). Targeting QseC signaling and virulence for antibiotic development. *Science* 321, 1078–1080. doi: 10.1126/science.1160354

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., et al. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4680–4687. doi: 10.1073/pnas.1002611107

Reddy, T. B., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., et al. (2015). The genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.* 43, D1099–D1106. doi: 10.1093/nar/gku950

Roberfroid, M. (2007). Prebiotics: the concept revisited. *J. Nutr.* 137(3 Suppl. 2), 830S–837S.

Roberfroid, M. B. (2000). Prebiotics and probiotics: are they functional foods? *Am. J. Clin. Nutr.* 71(6 Suppl.), 1682S–1687S.

Robinson, C. J., Bohannan, B. J., and Young, V. B. (2010). From structure to function: the ecology of host-associated microbial communities. *Microbiol. Mol. Biol. Rev.* 74, 453–476. doi: 10.1128/MMBR.00014-10

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome* 2:4. doi: 10.1186/2049-2618-2-4

Rosenthal, M., Goldberg, D., Aiello, A., Larson, E., and Foxman, B. (2011). Skin microbiota: microbial community structure and its potential association with health and disease. *Infect. Genet. Evol.* 11, 839–848. doi: 10.1016/j.meegid.2011.03.022

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., et al. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. *Mol. Cell.* 52, 124–134. doi: 10.1016/j.molcel.2013.08.020

Round, J. L., and Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev. Immunol.* 9, 313–323. doi: 10.1038/nri2515

Rupnik, M. (2015). Toward a true bacteriotherapy for *Clostridium difficile* infection. *N. Engl. J. Med.* 372, 1566–1568. doi: 10.1056/NEJMcibr1500270

Saad, R., Rizkallah, M. R., and Aziz, R. K. (2012). Gut Pharmacomicrobiomics: the tip of an iceberg of complex interactions between drugs and gut-associated microbes. *Gut Pathog.* 4:16. doi: 10.1186/1757-4749-4-16

Sangiuliano, B., Perez, N. M., Moreira, D. F., and Belizario, J. E. (2014). Cell death-associated molecular-pattern molecules: inflammatory signaling and control. *Mediators Inflamm.* 2014:821043. doi: 10.1155/2014/821043

Sartor, R. B., and Mazmanian, S. K. (2012). Intestinal microbes in inflammatory bowel diseases. *Am. J. Gastroenterol. Suppl.* 1, 15–21. doi: 10.1038/ajgsup.2012.4

Segata, N., Haake, S. K., Mannon, P., Lemon, K. P., Waldron, L., Gevers, D., et al. (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* 13:R42. doi: 10.1186/gb-2012-13-6-r42

Selle, K., and Barrangou, R. (2015). Harnessing CRISPR-Cas systems for bacterial genome editing. *Trends Microbiol.* 23, 225–232. doi: 10.1016/j.tim.2015.01.008

Seo, H. S., Xiong, Y. Q., Mitchell, J., Seepersaud, R., Bayer, A. S., and Sullam, P. M. (2010). Bacteriophage lysin mediates the binding of *Streptococcus mitis* to

human platelets through interaction with fibrinogen. *PLoS Pathog.* 6:e1001047. doi: 10.1371/journal.ppat.1001047

Slavin, J. (2013). Fiber and prebiotics: mechanisms and health benefits. *Nutrients* 5, 1417–1435. doi: 10.3390/nu5041417

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A., and Alm, E. J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480, 241–244. doi: 10.1038/nature10571

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermudez-Humaran, L. G., Gratadoux, J. J., et al. (2008). *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16731–16736. doi: 10.1073/pnas.0804812105

Sommer, M. O., and Dantas, G. (2011). Antibiotics and the resistant microbiome. *Curr. Opin. Microbiol.* 14, 556–563. doi: 10.1016/j.mib.2011.07.005

Sulakvelidze, A., Alavidze, Z., and Morris, J. G. (2001). Bacteriophage therapy. *Antimicrob. Agents Chemother.* 45, 649–659. doi: 10.1128/AAC.45.3.649-659.2001

Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L., and Azcarate-Peril, M. A. (2015). Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Front. Cell Infect. Microbiol.* 5:3. doi: 10.3389/fcimb.2015.00003

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449, 804–810. doi: 10.1038/nature06244

van der Oost, J., Westra, E. R., Jackson, R. N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* 12, 479–492. doi: 10.1038/nrmicro3279

van Opijnen, T., Bodi, K. L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* 6, 767–772. doi: 10.1038/nmeth.1377

Vercoe, R. B., Chang, J. T., Dy, R. L., Taylor, C., Gristwood, T., Clulow, J. S., et al. (2013). Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet.* 9:e1003454. doi: 10.1371/journal.pgen.1003454

Verdam, F. J., Fuentes, S., de Jonge, C., Zoetendal, E. G., Erbil, R., Greve, J. W., et al. (2013). Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity (Silver Spring)* 21, E607–E615. doi: 10.1002/oby.20466

Vuotto, C., Longo, F., and Donelli, G. (2014). Probiotics to counteract biofilm-associated infections: promising and conflicting data. *Int. J. Oral Sci.* 6, 189–194. doi: 10.1038/ijos.2014.52

Wallace, B. D., and Redinbo, M. R. (2013). The human microbiome is a source of therapeutic drug targets. *Curr. Opin. Chem. Biol.* 17, 379–384. doi: 10.1016/j.cbpa.2013.04.011

Wang, G. (2014). Human antimicrobial peptides and proteins. *Pharmaceuticals (Basel)* 7, 545–594. doi: 10.3390/ph7050545

Waters, C. M., and Bassler, B. L. (2005). Quorum sensing: cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.* 21, 319–346. doi: 10.1146/annurev.cellbio.21.012704.131001

Whelan, K., and Quigley, E. M. (2013). Probiotics in the management of irritable bowel syndrome and inflammatory bowel disease. *Curr. Opin. Gastroenterol.* 29, 184–189. doi: 10.1097/MOG.0b013e32835d7bba

Wikoff, W. R., Anfora, A. T., Liu, J., Schultz, P. G., Lesley, S. A., Peters, E. C., et al. (2009). Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3698–3703. doi: 10.1073/pnas.0812874106

Wilson, I. D., and Nicholson, J. K. (2009). The role of gut microbiota in drug response. *Curr. Pharm. Des.* 15, 1519–1523. doi: 10.2174/138161209788168173

Woese, C. R. (1987). Bacterial evolution. *Microbiol. Rev.* 51, 221–271.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.* 87, 4576–4579. doi: 10.1073/pnas.87.12.4576

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput. Biol.* 6:e1000667. doi: 10.1371/journal.pcbi.1000667

Wozniak, R. A., and Waldor, M. K. (2010). Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat. Rev. Microbiol.* 8, 552–563. doi: 10.1038/nrmicro2382

Wright, G. D. (2010). Antibiotic resistance in the environment: a link to the clinic? *Curr. Opin. Microbiol.* 13, 589–594. doi: 10.1016/j.mib.2010.08.005

Xavier, K. B., and Bassler, B. L. (2003). LuxS quorum sensing: more than just a numbers game. *Curr. Opin. Microbiol.* 6, 191–197. doi: 10.1016/S1369-5274(03)00028-6

Xiao-Jie, L., Hui-Ying, X., Zun-Ping, K., Jin-Lian, C., and Li-Juan, J. (2015). CRISPR-Cas9: a new and promising player in gene therapy. *J. Med. Genet.* 52, 289–296. doi: 10.1136/jmedgenet-2014-102968

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Yosef, I., Manor, M., Kiro, R., and Qimron, U. (2015). Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7267–7272. doi: 10.1073/pnas.1500107112

Zaura, E., Nicu, E. A., Krom, B. P., and Keijser, B. J. (2014). Acquiring and maintaining a normal oral microbiome: current perspective. *Front. Cell Infect. Microbiol.* 4:85. doi: 10.3389/fcimb.2014.00085

Zhang, Q., Rho, M., Tang, H., Doak, T. G., and Ye, Y. (2013). CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* 14, R40. doi: 10.1186/gb-2013-14-4-r40

Zhou, Y., Gao, H., Mihindukulasuriya, K. A., La Rosa, P. S., Wylie, K. M., Vishnivetskaya, T., et al. (2013). Biogeography of the ecosystems of the healthy human body. *Genome Biol.* 14:R1. doi: 10.1186/gb-2013-14-1-r1

Zhou, Y., Mihindukulasuriya, K. A., Gao, H., La Rosa, P. S., Wylie, K. M., Martin, J. C., et al. (2014). Exploration of bacterial community classes in major human habitats. *Genome Biol.* 15, R66. doi: 10.1186/gb-2014-15-5-r66

Zoetendal, E. G., Rajilic-Stojanovic, M., and de Vos, W. M. (2008). High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* 57, 1605–1615. doi: 10.1136/gut.2007.133603

Zomer, A., Burghout, P., Bootsma, H. J., Hermans, P. W., and van Hijum, S. A. (2012). ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PLoS ONE* 7:e43012. doi: 10.1371/journal.pone.0043012

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 11 August 2015
doi: 10.3389/fmicb.2015.00824

Characterization of the gut microbiota of Kawasaki disease patients by metagenomic analysis

Akiko Kinumaki[1,2], *Tsuyoshi Sekizuka*[2], *Hiromichi Hamada*[3], *Kengo Kato*[2], *Akifumi Yamashita*[2] *and Makoto Kuroda*[2]*

[1] Department of Pediatrics, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Japan, [2] Laboratory of Bacterial Genomics, Pathogen Genomics Center, National Institute of Infectious Diseases, Shinjuku-ku, Japan, [3] Department of Pediatrics, Faculty of Medicine, Yachiyo Medical Center, Tokyo Women's Medical University, Yachiyo, Japan

Kawasaki disease (KD) is an acute febrile illness of early childhood. Previous reports have suggested that genetic disease susceptibility factors, together with a triggering infectious agent, could be involved in KD pathogenesis; however, the precise etiology of this disease remains unknown. Additionally, previous culture-based studies have suggested a possible role of intestinal microbiota in KD pathogenesis. In this study, we performed metagenomic analysis to comprehensively assess the longitudinal variation in the intestinal microbiota of 28 KD patients. Several notable bacterial genera were commonly extracted during the acute phase, whereas a relative increase in the number of *Ruminococcus* bacteria was observed during the non-acute phase of KD. The metagenomic analysis results based on bacterial species classification suggested that the number of sequencing reads with similarity to five *Streptococcus* spp. (*S. pneumonia, pseudopneumoniae, oralis, gordonii,* and *sanguinis*), in addition to patient-derived *Streptococcus* isolates, markedly increased during the acute phase in most patients. *Streptococci* include a variety of pathogenic bacteria and probiotic bacteria that promote human health; therefore, this further species discrimination could comprehensively illuminate the KD-associated microbiota. The findings of this study suggest that KD-related *Streptococci* might be involved in the pathogenesis of this disease.

Keywords: Kawasaki disease, gut microbiota, metagenomic analysis, *Streptococcus*, mitis group

OPEN ACCESS

Edited by:
Roy D. Sleator,
Cork Institute of Technology, Ireland

Reviewed by:
Suleyman Yildirim,
Istanbul Medipol University International School of Medicine, Turkey
Michael S. Allen,
University of North Texas Health Science Center, USA

***Correspondence:**
Makoto Kuroda,
Pathogen Genomics Center, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan
makokuro@nih.go.jp

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *11 March 2015*
Accepted: *27 July 2015*
Published: *11 August 2015*

Citation:
Kinumaki A, Sekizuka T, Hamada H, Kato K, Yamashita A and Kuroda M (2015) Characterization of the gut microbiota of Kawasaki disease patients by metagenomic analysis. Front. Microbiol. 6:824. doi: 10.3389/fmicb.2015.00824

Introduction

Kawasaki disease (KD) is an acute febrile illness of early childhood. The principal pathology is systemic vasculitis with coronary artery involvement, and KD is the leading cause of acquired heart disease in developed countries. It was originally described by Dr. Tomisaku Kawasaki in 1967 (Kawasaki, 1967), and it is known to occur worldwide in children of all races. However, as the etiology of KD remains unknown, no specific biological markers for diagnostic testing

Abbreviations: BLAST, Basic Local Alignment Search Tool; BWA-SW, Burrows-Wheeler Aligner's Smith-Waterman Alignment; CFU, colony forming unit; EDTA, ethylenediaminetetraacetic acid; FASTQ, a text-based format for storing both a biological sequence (usually a nucleotide sequence) and its corresponding quality scores; KD, Kawasaki disease; LPS, lipopolysaccharide; LEfSe, linear discriminant analysis (LDA) coupled with effect size measurements; LDA, linear discriminant analysis; MEGAN, MEtaGenome Analyzer; PCA, Principal component analysis; PERMANOVA, Permutational Multivariate Analysis Of Variance; TCR, T-cell receptor.

have been characterized to date. The diagnosis of KD is based on the following six clinical features: fever lasting for at least 5 days, changes in the extremities, polymorphous exanthem, bilateral conjunctival injection without exudate, changes in the lips and oral cavity, and cervical lymphadenopathy (Newburger et al., 2004). Although the simultaneous intravenous infusion of gamma globulin and aspirin is effective in reducing systemic inflammation and preventing coronary artery involvement, coronary abnormalities still develop in ∼5% of affected children, and some patients show no response to this therapy (Newburger et al., 1991).

The annual incidence of KD is increasing rapidly in Japan, with 239.6/100,000 children under the age of 5 years affected in 2010. This incidence is by far the highest rate worldwide (Nakamura et al., 2012), and the risk of KD in siblings of affected children is significantly higher than that in the general population (Fujita et al., 1989). The annual incidence rates are also relatively high in other East Asian countries (with 113.1/100,000 children under the age of 5 years affected in Korea and 69/100,000 in Taiwan) but are low in Europe and North America (with 4.9–15.2/100,000 children under the age of 5 years affected in European countries and 19–26.2/100,000 in North American countries) (Uehara and Belay, 2012). A higher rate of KD has been reported in Hawaiian children of Japanese descent compared with those of European descent (Holman et al., 2010), suggesting the importance of genetic factors in disease susceptibility.

Epidemiological studies have shown that the age-specific incidence rate of KD is the highest among children aged 6–11 months and that 88.4% of KD patients are less than 1 year of age (Uehara and Belay, 2012). Interestingly, a seasonal variation in the number of affected KD patients has been observed (Nakamura et al., 2008). These findings suggest that an infectious agent may trigger this disease; however, its etiology remains unknown. A GWAS of KD in Japanese patients has revealed susceptibility loci related to immune disorders and a human leukocyte antigen; such extensive studies will facilitate characterization of the pathogenesis and pathophysiology (Onouchi, 2012; Onouchi et al., 2012).

Previous reports have suggested that an elevation in lipopolysaccharide (LPS, endotoxin)-binding neutrophils or plasma proteins (Takeshita et al., 1999, 2002b), antibody reactivity against mycobacterial heat-shock protein (HSP65) in convalescent sera (Yokota et al., 1993), and unique TCR Vβ expansion by certain superantigens (SAg) in KD patients (Abe et al., 1992; Yoshioka et al., 1999) might be involved in KD pathogenesis. Case reports have suggested that these factors are attributed to the presence of secondary infections with various pathogens, including *Streptococcus pyogenes*, *Staphylococcus aureus*, *Mycoplasma pneumoniae*, *Chlamydia pneumoniae*, *Klebsiella pneumoniae*, adenovirus, Epstein-Barr virus, parvovirus B19, herpesvirus 6, parainfluenza virus, measles, rotavirus, dengue virus, varicella zoster virus, cytomegalovirus, and influenza virus (Johnson and Azimi, 1985; Catalano-Pons et al., 2005; Wang et al., 2005; Joshi et al., 2011; Principi et al., 2013). In animal models, exposure to the *Lactobacillus* bacterial cell wall (Duong et al., 2003), immunization with bacillus Calmette-Guérin (BCG) (Nakamura et al., 2007), or exposure to the *Candida albicans* water-soluble fraction (Nagi-Miura et al., 2004; Ohno, 2004) has been shown to induce vasculitis and coronary arteritis. These observations further suggest that infectious agents promote the onset of KD.

The intestinal microbiota constitutes a vast ecosystem with a crucial role in establishing the mucosal immune system, and the intestinal microbiota of healthy adults is considered to be inter-individually variable and intra-individually stable over long time periods (Eckburg et al., 2005; Jakobsson et al., 2010; Arumugam et al., 2011; Jalanka-Tuovinen et al., 2011). By contrast, the intestinal microbiota of infants is different from that of adults, with intestinal microbiota succession being affected by breast or formula feeding, weaning, diet, and unexpected life events, including infection and antibiotic treatment (Stark and Lee, 1982; Palmer et al., 2007; De Filippo et al., 2010; Koenig et al., 2011; Morotomi et al., 2011). The pathogenesis of KD has been suggested to involve a hyperimmune reaction in children who are genetically susceptible to variations in the normal flora; these variations are induced by environmental factors (Lee et al., 2007).

The intestinal microbiota of KD patients is characterized by a lack of *Lactobacilli* during the acute phase (Takeshita et al., 2002a) and the presence of HSP60-producing Gram-negative microbes (genera *Acinetobacter*, *Enterobacter*, *Neisseria*, and *Veillonella*) and Gram-positive cocci (genera *Streptococcus* and *Staphylococcus*) with the ability to induce Vβ2 T cell expansion (Nagata et al., 2009). However, these studies on the intestinal profiles of KD patients were performed using culture-based methods.

Metagenomic analyses can reveal both the bacterial and viral compositions of the intestinal microbiota; thus, metagenomics can be used to identify potential pathogens in infectious diseases of unknown etiology (Kuroda et al., 2012). For instance, a metagenomic approach has revealed the presence of *Streptococcus* spp. in lymph node specimens of a KD patient, highlighting the possible role of these bacteria in KD (Katano et al., 2012).

In this study, a comparative metagenomic approach was used to characterize the differential microbiota compositions of KD patients by studying individual clinical specimens in a longitudinal manner. No study to date has performed longitudinal analysis of the microbial microbiota compositions of KD patients using a metagenomic approach. Indeed, although previous studies have suggested a possible role of the intestinal microbiota in the pathogenesis of KD, they have relied only on culture-based methods for microbial detection (Takeshita et al., 2002a; Nagata et al., 2009). We therefore performed metagenomic analysis using a non-culture-based method to expand upon these results.

Materials and Methods

Clinical Specimens Used for Comparative Metagenomic Analysis

For the KD patient group, fecal samples were obtained at the time of admission (the acute phase), at the time of discharge

(the convalescent phase), and at 4–6 months after the onset of KD (the non-acute phase). The study protocol was approved by the institutional medical ethics committee of the University of Tokyo, Tokyo Women's Medical University and the National Institute of Infectious Diseases in Japan (Approval No. 295), and it was conducted according to the Declaration of Helsinki Principles. Written informed consent was obtained from the parents of all children for publication of their individual details and accompanying images in this manuscript. The consent form is held by the authors' institution and is available for review.

DNA Extraction from Fecal Samples

Total DNA extraction was performed using a QIAamp® DNA Stool Mini Kit (QIAGEN, Tokyo, Japan) according to the manufacturer's instructions. To increase the recovery of bacterial DNA, particularly from Gram-positive bacteria, pretreatment with lytic enzymes was performed prior to extraction using the stool kit. Briefly, 100 mg of fecal sample was suspended in 10 mL of Tris-EDTA buffer (pH 7.5), and 50 μL of 100 mg/mL lysozyme type VI purified from chicken egg white (MPBIO, Derby, UK) and 50 μL of 1 mg/mL purified achromopeptidase (Wako, Osaka, Japan) were added. The solution was incubated at 37°C for 1 h with mixing, 0.12 g of sodium dodecyl sulfate (final conc. 1%) was added, and the suspension was mixed until it became clear. Next, 100 μL of 20 mg/mL proteinase K (Wako) was added, followed by incubation at 55°C for 1 h with mixing. The cell lysate was then subjected to ethanol precipitation. The precipitant was dissolved in 1.6 mL of ASL buffer from the stool kit and subsequently purified using a QIAamp® DNA Stool Mini Kit (QIAGEN).

DNA Library Preparation for Metagenomic Analysis and Short-read DNA Sequencing

A DNA library was prepared using a Nextera™ DNA Sample Prep Kit (Illumina-compatible, EPICENTRE Biotechnologies, Madison, WI, USA), and DNA clusters were generated on a slide using a Cluster Generation Kit (version 2) with an Illumina cluster station (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. The general procedure described in the standard protocol (Illumina) was performed to obtain standard $\sim 1.0 \times 10^7$ short reads for 1 lane. All of the sequencing runs for generating 126-mers were performed with a Genome Analyzer IIx using an Illumina Sequencing Kit (http://www.illumina.com/systems/retired_gaiix/gaiix-kits.html). Fluorescence images were analyzed using Illumina base-calling pipeline (version 1.4.0) to obtain FASTQ-formatted sequence data. The short-read sequences have been deposited in DNA Data Bank of Japan (DDBJ; accession numbers: DRA000895 and DRA001171). All of the obtained DNA sequencing reads were aligned to a reference human genomic sequence using BWA-SW read-mapping software (Li and Durbin, 2010), with quality trimming to remove low-quality reads. The remaining sequence reads were subjected to a megaBLAST search against a nucleotide database. The results of this search were analyzed and visualized using MEGAN version 4.62.3 (Huson et al., 2011), with a minimum support of 1 hit and a minimum score of 150.

Principal Component Analysis (PCA) and PERMANOVA Analysis

The sequenced reads were assigned to a taxonomic hierarchy using MEGAN software following a megaBLAST homology search. The raw read counts were normalized by the total number of reads, and then PCA was performed using the R "prcomp" and "plot" functions. Permutational multivariate analysis of variance (PERMANOVA) with "ADONIS" was performed using 10,000 × permutations and the "bray" method with R's vegan package (Anderson, 2001).

Linear Discriminant Analysis (LDA) Coupled With Effect Size Measurements (LEfSe)

A metagenomic biomarker discovery approach, LEfSe, was used to identify the microbial components whose sequences were more abundant in the fecal samples of the KD patients during the acute phase than in those of the KD patients during the non-acute phase and the controls. For LEfSe, Kruskal–Wallis and pairwise Wilcoxon tests are performed, followed by LDA to assess the effect size of each differentially abundant taxon (Segata et al., 2011). In this study, a p-value of <0.05 was considered significant for both statistical methods. Bacteria with markedly increased numbers were defined as those with an LDA score ($\log_{10}$) of over 2. Less than 0.01% of the total bacterial reads, corresponding with $\leq 10^7$ CFU/g feces, were omitted from further analysis because of low and unreliable read counts, although significant LDA scores were observed in LEfSe.

Isolation of *Streptococcus* spp. and Species Determination Based on 16S-rRNA Gene

Cultivation of *Streptococcus* spp. was performed using phenylethyl alcohol agar with 5% sheep blood or chocolate agar under anaerobic conditions at 37°C for 48 h. The bacterial species present were determined by performing 16S-rRNA gene sequencing using the bacterial forward primer Bac27F (5′-AGAGTTTGGATCMTGGCTCAG-3′) and the universal reverse primer Univ1492R (5′-CGGTTACCTTGTTACGACTT-3′) (Eden et al., 1991). The obtained sequences were searched against SILVA ribosomal RNA gene database to identify the bacterial species (Quast et al., 2013).

Whole-Genome and Phylogenetic Analyses of Identified *Streptococcus* spp.

A draft genome sequence was obtained by whole-genome sequencing using MiSeq with a NEXTERA XT library preparation kit (Illumina), followed by *de novo* assembly with A5-MiSeq pipeline (Tritt et al., 2012). The resulting scaffolds were annotated using RAST server (Aziz et al., 2008). Maximum likelihood phylogenetic analysis of *Streptococcus* 16S-rDNA was performed using MEGA 6.0 with 1000 bootstrap iterations (Tamura et al., 2013).

Minimum Inhibitory Concentration (MIC) Testing

MIC testing was performed using an Etest (bioMerieux, France) on Muller-Hinton agar (Difco, Augsburg), according to CLSI guidelines (CLSI, 2013).

Results

KD Patients Included in Comparative Metagenomic Analysis

This study evaluated 28 KD patients (15 males and 13 females, aged 1–114 mo; median of 25 mo). All of these patients were enrolled within 4 days of the onset of illness, with day 1 defined as the first day of fever, and they all met the diagnostic criteria for KD established by the American Heart Association (Newburger et al., 2004). All of the KD patients in the study received intravenous gamma globulin (2 g/kg) and aspirin (30–50 mg/kg/day). One male patient (patient P2) had a persistent fever despite receiving these therapies and was administered additional intravenous gamma globulin (1 g/kg) and prednisolone sodium succinate (2 mg/kg/day). This patient had transient dilatation of the coronary artery, whereas the other 27 patients showed no evidence of cardiac abnormalities.

In this study, the time of admission was defined as the acute phase, while 4–6 months after the onset of KD was considered the non-acute phase. The profiles of the participants, including the age, sex, concomitant symptoms and empirical antimicrobial treatment received, are shown in **Table 1**.

Gut Microbiota Analysis Comparing the Acute and Non-acute Phases in KD Patients

A total of 56 samples (28 samples each for the acute and non-acute phases) were collected, including two samples from each KD patient (**Figure 1** and **Table 1**). Extracted DNA was subjected to metagenomic sequencing using an Illumina GAIIx next-generation DNA sequencer, and more than 10 million short 126-mer reads were obtained for each specimen. The short reads were classified at the family level of bacteria, with a threshold megaBLAST homology score of $= 150$. Principal component analysis (PCA) was performed to elucidate the variations between the acute and non-acute phases of KD. The results suggested that the gut microbiota was more variable during the acute phase than during the non-acute phase based on family-level taxonomy (**Figure 2A**). PERMANOVA with 10,000 $\times$ permutations revealed significant dissimilarity of the bacterial communities at the family level between the acute and non-acute

TABLE 1 | Kawasaki disease patient information.

KD patient	Age	Sex	Concomitant symptoms	Antimicrobial treatment during the acute phase
P1	3 y 11 m	M	Vomiting, cough	CMZ
P2	2 y 08 m	M	Diarrhea, vomiting, cough	CFPN-PI/CTX
P3	3 y 00 m	F	Cough	CMZ
P4	2 y 01 m	M	Diarrhea	CFPN-PI/CMZ
P5	1 y 03 m	F	–	AZM/CMZ
P6	2 y 03 m	F	Diarrhea	ABPC/CMZ
P7	3 y 04 m	F	Diarrhea, vomiting	CFPN-PI/CMZ
P8	0 y 08 m	F	Diarrhea, vomiting	ABPC-CVA/CMZ
P9	2 y 06 m	M	Vomiting, abdominal pain	CPDX-PR
P10	5 y 06 m	M	Cough	–
P11	7 y 05 m	F	Diarrhea	CAM
P12	2 y 10 m	F	Neck stiffness	–
P13	0 y 03 m	M	–	CCL
P14	2 y 02 m	M	Rhinorrhea	–
P15	1 y 06 m	F	Cough	CFPN-PI
P16	3 y 10 m	M	–	CDTR-PI
P17	0 y 10 m	F	–	–
P18	1 y 06 m	M	–	CFDN
P19	1 y 04 m	M	Cough	CDTR-PI/FOM
P20	2 y 00 m	F	–	–
P21	4 y 04 m	F	Cough, diarrhea	CFPN-PI/TFLX
P22	2 y 09 m	F	–	AMPC
P23	6 y 01 m	M	–	CFPN-PI
P24	3 y 00 m	M	–	CDTR-PI
P25	1 y 11 m	M	–	CFPN-PI/CAM
P26	9 y 06 m	F	–	CDTR-PI
P27	2 y 11 m	M	–	–
P28	0 y 04 m	M	Vomiting	CDTR-PI

y, year; m, month. M, male; F, female. ABPC, ampicillin; ABPC-CVA, ampicillin-clavulanic acid; AMPC, amoxicillin; AZM, azithromycin; CAM, clarithromycin; CCL, cefaclor; CDTR-PI, cefditoren, pivoxil; CFDN, cefdinir; CFPN-PI, cefcapene pivoxil; CMZ, cefmetazole; CPDX-PR, cefpodoxime proxetil; CTX, cefotaxime; FOM, fosfomycin; TFLX, tosufloxacin.

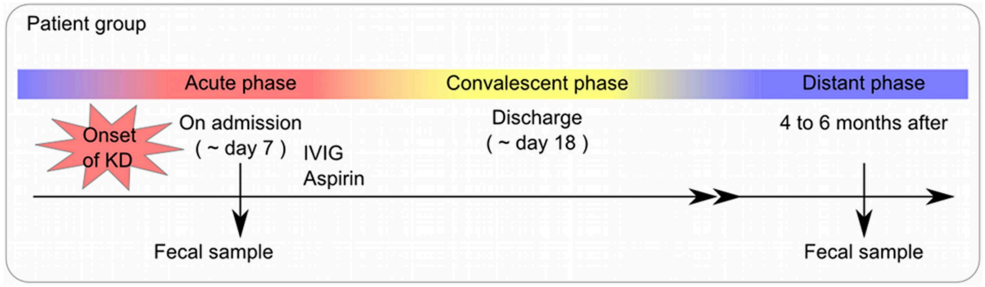

FIGURE 1 | The protocol for collection of clinical specimens.

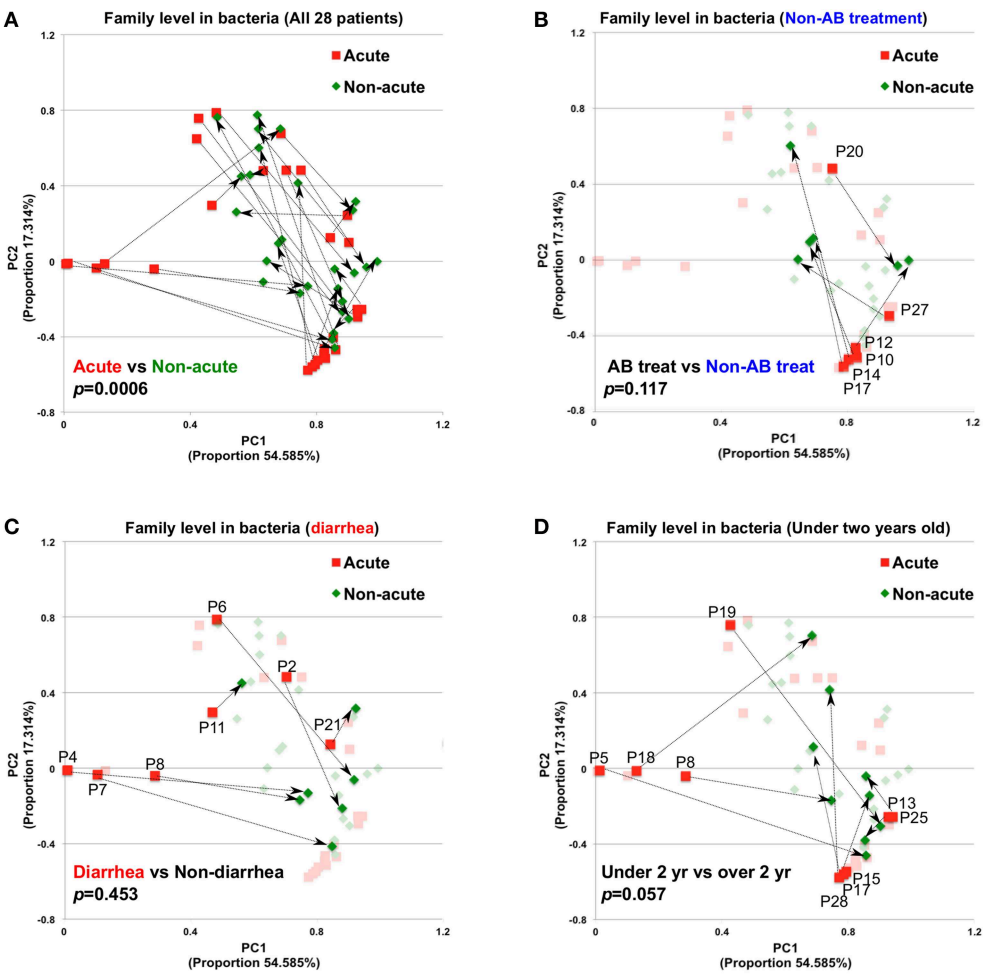

FIGURE 2 | Principal component analysis of gut microbiota compositions during the acute and non-acute phases of KD. (A) Relative abundance was estimated at the family taxonomic level (megaBLAST homology score threshold: ≥150). The arrows indicate the corresponding pair for each patient for the acute and non-acute phases ($n = 28$). Some PCA plot components were highlighted based on the indicated patient-related information, group of non-antimicrobial treatment **(B)**, diarrhea symptom **(C)**, and under two years old **(D)**. To investigate the significance between the tested groups, PERMANOVA with "ADONIS" was performed using 10,000 × permutations, in addition to the "bray" method, with R's vegan package.

phases of KD (F-test = 3.7307, p = 0.0006) (**Figure 2A**). The components were highlighted based on the antimicrobial treatment, occurrence of diarrhea, and age group, indicating that such variations during the acute phase might be associated with patient-related factors. Further, the no antimicrobial treatment group during the acute phase had was clustered in the lower right area of the PCA plot (**Figure 2B**), whereas the diarrhea-positive group during the acute phase exhibited a relatively

scattered cluster in the PCA plot (**Figure 2C**); however, both groups during the non-acute phase were clustered together at the relative center of the plot (**Figures 2B,C**). The patients in the antimicrobial treatment group did not always exhibit diarrhea symptoms (8 diarrhea/22 antimicrobial treatment), and PERMANOVA indicated no significant differences between the two subject groups, suggesting that the gut microbiota was not affected during the acute phase of KD, regardless of the presence of antimicrobial treatment or diarrhea. The age factor showed a possible association with gut microbiota composition because the *p*-value was close to 0.05 when comparing subjects who were less than 2 years of age with those who were over 2 years of age (**Figure 2D**).

To determine the variations in gut microbiota composition between the acute and non-acute phases, linear discriminant analysis (LDA) coupled with effect size measurements (LEfSe) was applied to determine which taxa were enriched in the different groups according to metagenomic analysis (see detailed parameters in Materials and Methods). LEfSe determines which features (organisms, clades, operational taxonomic units, genes, or functions) are most likely to explain differences between classes by coupling standard tests for statistical significance (between the acute and non-acute phases in this study) with additional tests of biological consistency and effect relevance (Segata et al., 2011). The obtained metagenomic reads were classified at the genus level (**Figure 3**). Although LEfSe revealed that the genera *Rothia* and *Staphylococcus* were the most abundant during the acute phase, this dominance was not observed in all of the patients (**Figure 3**). Regardless, relatively increased numbers of *Ruminococcus*, *Blautia*, *Faecalibacterium*, and *Roseburia* bacteria were observed during the non-acute phase of KD (**Figure 3**), indicating that these genera was possibly related to remission in KD patients.

The above genus classifications did not reveal which common features were most likely to explain the differences between the acute and non-acute phases in the tested KD patients ($n = 28$). Because both pathogenic and non-pathogenic species may be included within one bacterial genus, we speculated that genus classifications would not reveal certain potential pathogens involved in KD pathogenesis; thus, further taxonomic

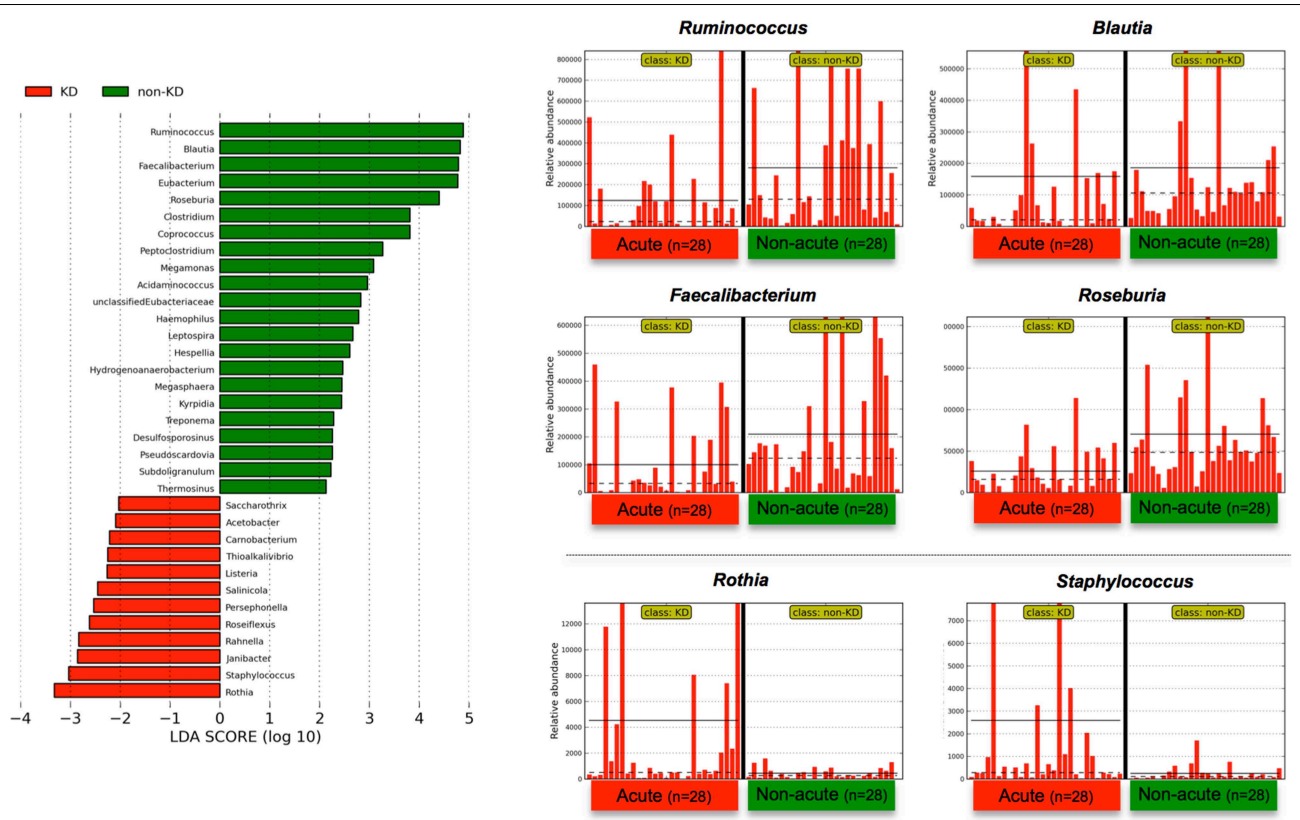

FIGURE 3 | LEfSe at the genus level. Linear discriminant analysis (LDA) combined with effect size measurements (LEfSe) revealed a list of features that enable discrimination between the acute and non-acute phases in the fecal samples. A *p*-value of <0.05 and a score ≥ 2.0 were considered significant in Kruskal–Wallis and pairwise Wilcoxon tests, respectively. The horizontal straight line in the panel indicates the group means, and the dotted line indicates the group medians. The genus *Ruminococcus* was identified as the most predominant during the non-acute phase, and the number of detected reads for all 28 patients was plotted in bar form in the upper-right panel. The straight line indicates the group means, and the dotted line indicates the group medians. Less than 0.01% of the total bacterial reads, corresponding to $\leq 10^7$ CFU/g feces, were omitted from further analysis because of low and unreliable read counts, although significant LDA scores were observed (some bacteria are not shown due to insufficient amounts of reads).

classifications at the species level may allow for effective detection of KD-related pathogens (**Figure 4**). *Roseburia* spp. were relatively abundant during the non-acute phase and could be identified at the genus level (**Figure 3**), whereas *Streptococcus* spp. were predominantly identified during the acute phase, suggesting that some *Streptococcus* spp., including *S. pneumoniae, S. oralis* and other strains, are candidate KD-related pathogens (**Figure 4**). *Staphylococcus hyicus* might have been misidentified due to an insufficient amount of reads (less than 0.01% of the population; <1000 of the assigned reads), although the LDA score was significant.

Indeed, *Streptococcus* spp. were highly abundant in the gut microbiotas of some of the KD patients; for example, 77% of the bacterial reads of one KD patient (the acute phase in P7) were from *Streptococcus*. However, the above-mentioned *S. pneumoniae* and *oralis* species were not cultured from feces during the acute phase of KD under conventional aerobic cultivation on a phenylethyl alcohol agar plate with 5% sheep blood in screening for *Streptococcus* spp., while anaerobic cultivation on chocolate agar resulted in positive *Streptococcus* colonies. In fact, fifty colonies of *Streptococcus* spp. were isolated from P7-feces on chocolate agar under anaerobic conditions with incubation at 37°C for 16 h, and then 16S-rDNA sequencing was performed to determine the bacterial species present. The results suggested that seven *Streptococcus* spp. were unique isolates (P7-Anaero4, P7-Anaero13, P7-Anaero24, P7-Anaero25, P7-Anaero36, P7-Anaero39, and P7-Anaero45) (**Figure 5C**). Using the draft genome sequences of seven P7-*Streptococcus* isolates, including publicly available *Streptococcus* spp. complete genomes (23 species), the metagenomic short reads of all 56 fecal samples were classified at the species level by a megaBLAST search and LEfSe. The results suggested that the amounts of the six P7-feces-related *Streptococcus* isolates (P7-Anaero4, 13, 24, 36, 39, and 45, but not P7-Anaero25) and the five detected *Streptococcus* spp. (*S. pneumonia, pseudopneumoniae, oralis, gordonii,* and *sanguinis*) were apparently increased during the acute phase in most of the KD patients, including P7, whereas *S. pasteurianus* was increased during the non-acute phase (**Figures 5A,B**). Intriguingly, the top 4 most abundant positive isolates were P7-feces-related *Streptococcus* spp. (P7-Anaero4, 45, 24, and 36) rather than defined pathogenic *Streptococcus* species, and all positively detected *Streptococcus* spp. were classified within a taxonomic lineage closely related to *S. oralis* or *pneumonia* (**Figure 5C**). All P7-feces-related *Streptococcus* isolates showed susceptibility to most antimicrobial agents, including cephem, indicating that the detection of abundant P7-feces-related isolates was most likely not correlated with antimicrobial selection.

Based on species-level identification, the first and second highest hits with identical BLAST scores constituted 17.9% of all of the short reads, which were mostly homologous to rRNA genes; thus, 82.1% of the short reads could be assigned to a unique top hit. Although the obtained 126-mer short reads might not have been sufficiently long for correct species assignments, the BLAST search results suggested that the above-mentioned P7-related *Streptococcus* groups were highly abundant during the acute phase of KD, in contrast with other pathogenic *Streptococcus* spp., such as *S. pyogenes, dysgalactiae, mutans,* and *pasteurianus*. The significant detection of unique isolates in KD patients implies a possible association of KD with uncharacterized *Streptococcus* spp.

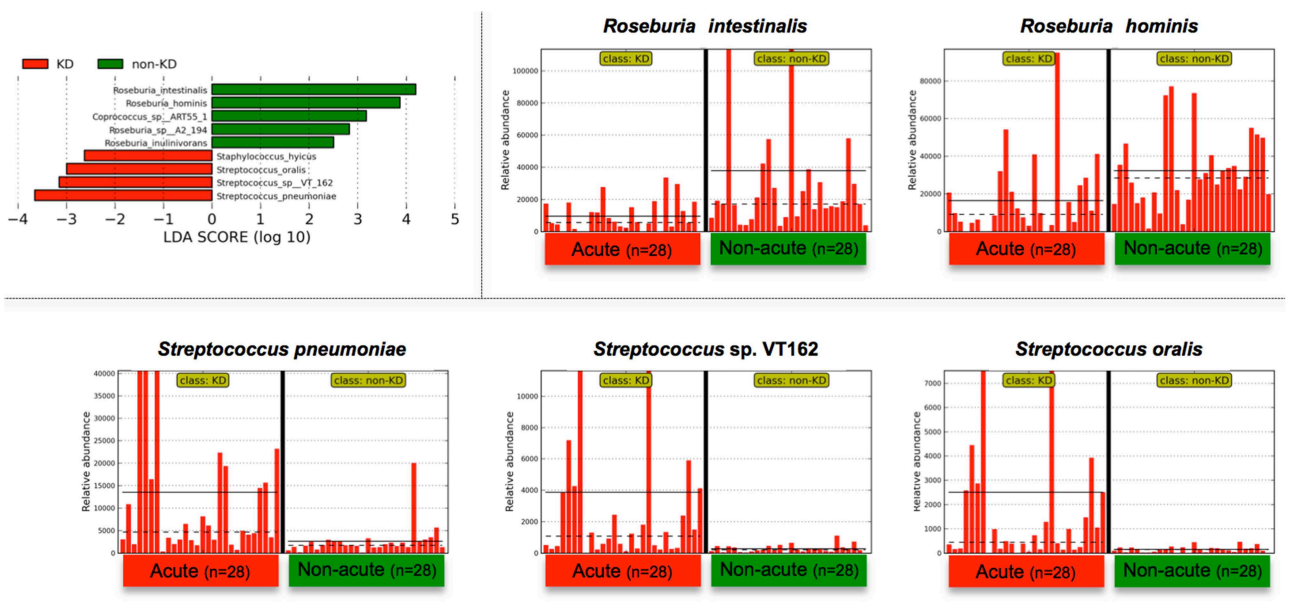

FIGURE 4 | LEfSe at the species level. LEfSe was performed as described in **Figure 3**. *Roseburia* spp. were relatively abundant during the non-acute phase and could be detected at the genus level (**Figure 3**), whereas *Streptococcus* spp. were found to be predominant during the acute phase. Less than 0.01% of the total bacterial reads, corresponding to $\leq 10^7$ CFU/g feces, were omitted from further analysis because of low and unreliable read counts, although significant LDA scores were observed (*Staphylococcus hyicus* is not shown due to insufficient amounts of reads).

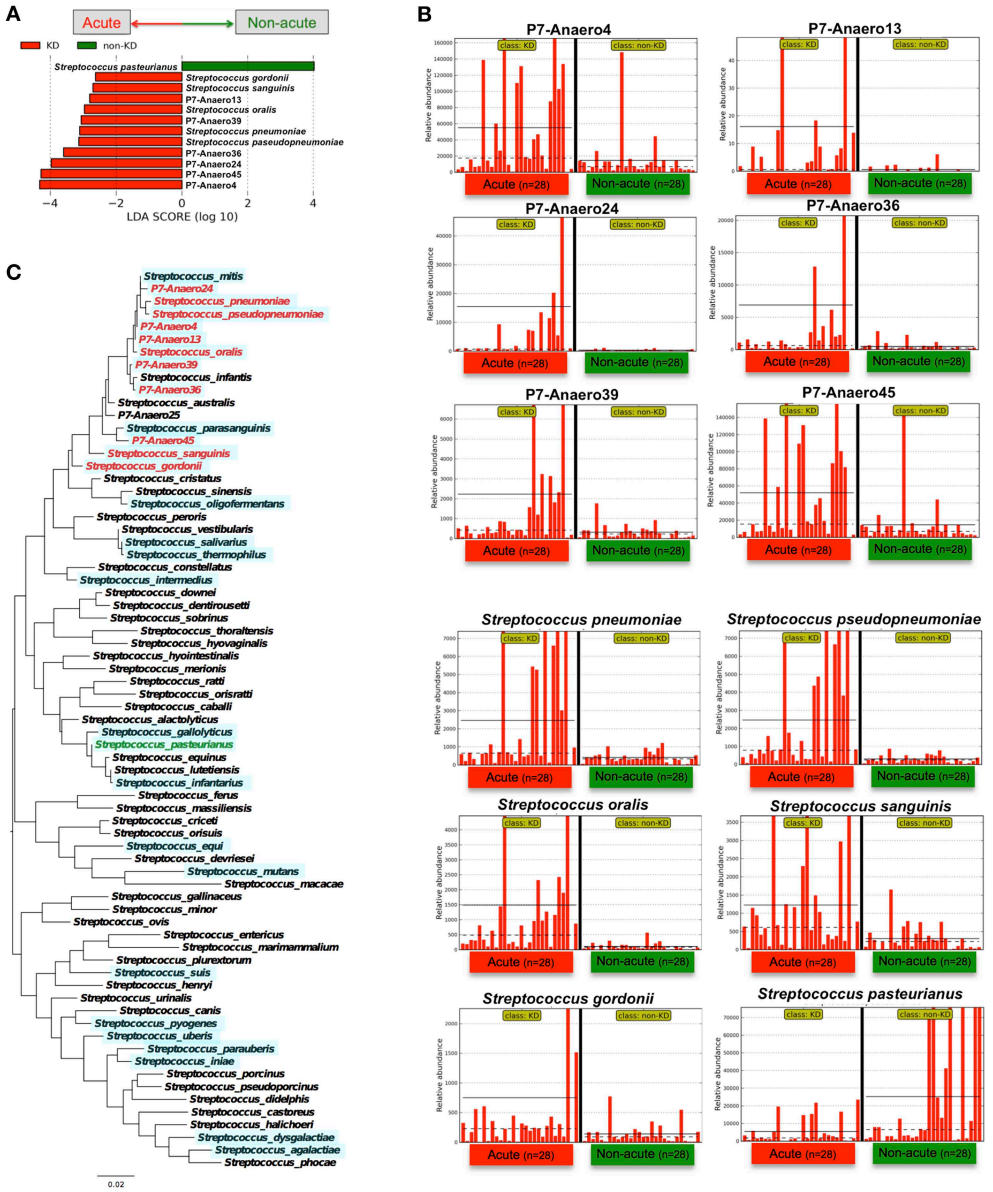

FIGURE 5 | LEfSe of P7-related *Streptococcus* isolates performed using *Streptococcus* genome database. (A) The type of *Streptococcus* spp. markedly differed between the acute and non-acute phases of KD. **(B)** The relative abundance of detected reads for the patients ($n = 28$) was plotted for each *Streptococcus* spp. The horizontal straight line indicates the group means, and the dotted line indicates the group medians in the panel. **(C)** Maximum likelihood phylogenetic analysis of *Streptococcus* 16S-rDNA. The *Streptococcus* species in red and green are those with increased abundance during the acute and non-acute phases of KD, respectively. The complete and draft genome sequences used for the megaBLAST search are highlighted with a light blue background.

Discussion

Various bacterial and viral agents have been reported to be associated with KD pathogenesis (Johnson and Azimi, 1985; Catalano-Pons et al., 2005; Wang et al., 2005; Joshi et al., 2011), but these speculations have been controversial (Wang et al., 2005). Colonization by normal microbiota variants have been suggested to induce a dysregulation in the immune systems of children with a pre-existing genetic defect in immune maturation, leading to a hyperimmune reaction and the development of KD (Lee et al., 2007). In this study, the possible pathogens detected in the KD patients varied for each individual patient; thus, every identified pathogen represented a potential candidate. Regarding virus species, human adenovirus (HAdV) species F was detected in one out of twenty-eight of the patients, despite the absence of gastrointestinal manifestations in that patient. Thus, HAdV was not commonly detected, and no sequences from either other viruses or previously reported pathogens were detected in any of the other KD samples.

Although the gut microbiota markedly differed at the genus level between the acute and non-acute phases of KD (**Figure 3**), we speculated that classification at the species level might be appropriate for identifying disease-associated bacteria because a genus includes species that have varying effects on human health [e.g., *S. pyogenes* infections include acute rheumatic fever, pharyngitis, impetigo and streptococcal toxic shock syndrome (STSS); *S. pneumoniae* causes many types of pneumococcal infections other than pneumonia; and *S. mutans* is a significant contributor to tooth decay in the human oral cavity].

The findings of this study suggested that notable *Streptococcus* spp. in the mitis group, including *S. pneumonia, pseudopneumoniae, mitis, oralis, gordonii,* and *sanguinis,* were highly abundant in the fecal samples during the acute phase (**Figures 4**, **5**); therefore, members of the mitis group of *Streptococci* could be present in the bacterial flora of KD patients. The mitis group comprises agents that contribute to oral biofilms, dental plaques, and infective endocarditis, disease processes that involve bacteria-bacteria and bacteria-host interactions (Whatmore et al., 2000). To further elucidate the association between *Streptococcus* spp. and KD in this study, we isolated a unique *Streptococcus* spp. (**Figure 5C**) and then performed whole-genome sequencing and a megaBLAST homology search. The results revealed a significant abundance of KD-derived *Streptococcus* isolates during the acute phase of the disease (**Figure 5**). Intriguingly, a recent paper has reported metagenomic analysis of the human gut microbiome in liver cirrhosis patients, suggesting that oral commensals, including *Streptococcus* spp., invade the gut in patients with liver cirrhosis (Qin et al., 2014) and implying that uncharacterized *Streptococcus* spp. could be potential biomarkers/pathogens for diseases with unknown etiologies.

A SAg hypothesis for KD on the etiology remains inconclusive, the involvement of single or multiple SAgs on T-cell Vβ repertoires has been speculated for the KD pathogenesis (Matsubara and Fukaya, 2007). A number of studies have found primarily Vβ2 expansion (Abe et al., 1992; Leung et al., 1995; Yoshioka et al., 1999) linking to the Vβ2 specific SAg such as toxic shock syndrome toxin-1 (TSST-1) and SpeC (Nur-Ur Rahman et al., 2011), although there is no direct evidence to suggest SAg involvement. STSS is significantly more frequent in group A ß-hemolytic streptococcal (GAS) patients than in groups B, C, and G streptococcal patients. GAS produces a multitude of surface-bound and secreted virulence factors causing resistance to phagocytosis, complement deposition, antibody opsonization, and neutrophil killing mechanisms, leading to overactive immune response and subverting host innate immune defenses (Walker et al., 2014). Although the isolation of SAg-positive *Streptococcus* from KD patients has been reported (Nagata et al., 2009; Leahy et al., 2012), group A ß-hemolytic streptococcal (GAS) might not contribute to the pathogenesis of this disease because a rapid antigen detection test (RADT) and proper antibiotic treatment prevents GAS pharyngitis during the initial episodes of acute rheumatic fever (Gerber et al., 2009). Because the KD patients in this study (**Table 1**) were empirically treated with antibiotics during the early stages of the disease, the results may reflect the effects of the antibiotic therapy. To address this issue, a full comparison of KD patients who have and have not received antibiotic therapy should be performed in a large, controlled study. It is also possible that antibiotic therapy has an adverse effect on the pathogenesis of KD. Further investigation of the role of *Streptococcus* spp. in the pathogenesis of KD is therefore merited.

Our previous metagenomic approach indicated that *Streptococcus* spp. were present in the lymph node specimen of one KD patient, highlighting the possible role of these bacteria in KD (Katano et al., 2012). To identify the SAg homologs, all short reads and coding sequences of the 7 isolates (P7-Anaero4, 13, 24, 25, 36, 39, and 45) were subjected to a PSI-BLAST homology search against "superantigen, staphylococcal/streptococcal toxin, bacterial (IPR013307)" orthologous proteins; however, no significant match has been found thus far (data not shown). In addition, some sequences from each sample were classified as "Not assigned" in metagenomic analysis, and new pathogenic agents remain to be characterized for some of these unidentified sequences.

The gut microbiota in the non-acute phase of KD (the distant phase) was similar in each patient, and the genera *Ruminococcus*, *Roseburia* and *Faecalibacterium* were predominant (**Figure 3**). In previous reports, an observed elevation in LPS-binding neutrophils or plasma proteins has been observed, suggesting that LPS infusion followed by disruption in intestinal mucosal barrier function might be involved in the pathogenesis of this disease (Takeshita et al., 1999, 2002b); therefore, a well-balanced commensal gut microbiota contributes to the mucosal barrier function of the intestine. Prebiotics, probiotics, and combination synbiotics modulate the balance of the intestinal microbiota and may help to prevent the onset of KD to improve patient prognosis (Bosscher et al., 2009).

The microbiota of KD patients was comprehensively analyzed in this study. Our findings suggest that markedly increased amounts of *Streptococcus* spp. are present in the gut microbiotas of acute-phase KD patients and that this difference in microbiota composition might be related to KD pathogenesis.

Author Contributions

AK and HH collected clinical specimens from the KD patients. TS, KK, and AY performed the metagenomic sequencing and statistical and bioinformatics analyses. AK and MK participated in the design of the study, performed statistical analysis, and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by grants from the NPO Japan Kawasaki Disease Research Center in 2010/2011 and from the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C) in 2011 (23590527)/2014 (26460542). The funders had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

References

Abe, J., Kotzin, B. L., Jujo, K., Melish, M. E., Glode, M. P., Kohsaka, T., et al. (1992). Selective expansion of T cells expressing T-cell receptor variable regions V beta 2 and V beta 8 in Kawasaki disease. *Proc. Natl. Acad. Sci. U.S.A.* 89, 4066–4070. doi: 10.1073/pnas.89.9.4066

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Aust. Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bosscher, D., Breynaert, A., Pieters, L., and Hermans, N. (2009). Food-based strategies to modulate the composition of the intestinal microbiota and their associated health effects. *J. Physiol. Pharmacol.* 60(Suppl. 6), 5–11.

Catalano-Pons, C., Quartier, P., Leruez-Ville, M., Kaguelidou, F., Gendrel, D., Lenoir, G., et al. (2005). Primary cytomegalovirus infection, atypical Kawasaki disease, and coronary aneurysms in 2 infants. *Clin. Infect. Dis.* 41, e53–56. doi: 10.1086/432578

CLSI. (2013). *Clinical and Laboratory Standards Institute. 2013. Performance Standards for Antimicrobial Susceptibility Testing; Twenty-third Informational Supplement. CLSI Document M100-S23*. Wayne, PA: Clinical and Laboratory Standards Institute.

De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., et al. (2010). Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14691–14696. doi: 10.1073/pnas.1005963107

Duong, T. T., Silverman, E. D., Bissessar, M. V., and Yeung, R. S. (2003). Superantigenic activity is responsible for induction of coronary arteritis in mice: an animal model of Kawasaki disease. *Int. Immunol.* 15, 79–89. doi: 10.1093/intimm/dxg007

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638. doi: 10.1126/science.1110591

Eden, P. A., Schmidt, T. M., Blakemore, R. P., and Pace, N. R. (1991). Phylogenetic analysis of Aquaspirillum magnetotacticum using polymerase chain reaction-amplified 16S rRNA-specific DNA. *Int. J. Syst. Bacteriol.* 41, 324–325. doi: 10.1099/00207713-41-2-324

Fujita, Y., Nakamura, Y., Sakata, K., Hara, N., Kobayashi, M., Nagai, M., et al. (1989). Kawasaki disease in families. *Pediatrics* 84, 666–669.

Gerber, M. A., Baltimore, R. S., Eaton, C. B., Gewitz, M., Rowley, A. H., Shulman, S. T., et al. (2009). Prevention of rheumatic fever and diagnosis and treatment of acute Streptococcal pharyngitis: a scientific statement from the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee of the Council on Cardiovascular Disease in the Young, the Interdisciplinary Council on Functional Genomics and Translational Biology, and the Interdisciplinary Council on Quality of Care and Outcomes Research: endorsed by the American Academy of Pediatrics. *Circulation* 119, 1541–1551. doi: 10.1161/CIRCULATIONAHA.109.191959

Holman, R. C., Christensen, K. Y., Belay, E. D., Steiner, C. A., Effler, P. V., Miyamura, J., et al. (2010). Racial/ethnic differences in the incidence of Kawasaki syndrome among children in Hawaii. *Hawaii Med. J.* 69, 194–197.

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Jakobsson, H. E., Jernberg, C., Andersson, A. F., Sjölund-Karlsson, M., Jansson, J. K., and Engstrand, L. (2010). Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS ONE* 5:e9836. doi: 10.1371/journal.pone.0009836

Jalanka-Tuovinen, J., Salonen, A., Nikkilä, J., Immonen, O., Kekkonen, R., Lahti, L., et al. (2011). Intestinal microbiota in healthy adults: temporal analysis reveals individual and common core and relation to intestinal symptoms. *PLoS ONE* 6:e23035. doi: 10.1371/journal.pone.0023035

Johnson, D., and Azimi, P. (1985). Kawasaki disease associated with Klebsiella pneumoniae bacteremia and parainfluenza type 3 virus infection. *Pediatr. Infect. Dis.* 4, 100. doi: 10.1097/00006454-198501000-00024

Joshi, A. V., Jones, K. D., Buckley, A. M., Coren, M. E., and Kampmann, B. (2011). Kawasaki disease coincident with influenza A H1N1/09 infection. *Pediatr. Int.* 53, e1–e2. doi: 10.1111/j.1442-200X.2010.03280.x

Katano, H., Sato, S., Sekizuka, T., Kinumaki, A., Fukumoto, H., Sato, Y., et al. (2012). Pathogenic characterization of a cervical lymph node derived from a patient with Kawasaki disease. *Int. J. Clin. Exp. Pathol.* 5, 814–823.

Kawasaki, T. (1967). [Acute febrile mucocutaneous syndrome with lymphoid involvement with specific desquamation of the fingers and toes in children]. *Arerugi* 16, 178–222.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4578–4585. doi: 10.1073/pnas.1000081107

Kuroda, M., Sekizuka, T., Shinya, F., Takeuchi, F., Kanno, T., Sata, T., et al. (2012). Detection of a possible bioterrorism agent, Francisella sp., in a clinical specimen by use of next-generation direct DNA sequencing. *J. Clin. Microbiol.* 50, 1810–1812. doi: 10.1128/JCM.06715-11

Leahy, T. R., Cohen, E., and Allen, U. D. (2012). Incomplete Kawasaki disease associated with complicated *Streptococcus pyogenes* pneumonia: a case report. *Can J Infect Dis Med Microbiol* 23, 137–139.

Lee, K. Y., Han, J. W., and Lee, J. S. (2007). Kawasaki disease may be a hyperimmune reaction of genetically susceptible children to variants of normal environmental flora. *Med. Hypothes.* 69, 642–651. doi: 10.1016/j.mehy.2006.12.051

Leung, D. Y., Meissner, C., Fulton, D., and Schlievert, P. M. (1995). The potential role of bacterial superantigens in the pathogenesis of Kawasaki syndrome. *J. Clin. Immunol.* 15, 11S-17S. doi: 10.1007/BF01540888

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Matsubara, K., and Fukaya, T. (2007). The role of superantigens of group A Streptococcus and *Staphylococcus aureus* in Kawasaki disease. *Curr. Opin. Infect. Dis.* 20, 298–303. doi: 10.1097/QCO.0b013e3280964d8c

Morotomi, N., Fukuda, K., Nakano, M., Ichihara, S., Oono, T., Yamazaki, T., et al. (2011). Evaluation of intestinal microbiotas of healthy Japanese adults and effect of antibiotics using the 16S ribosomal RNA gene based clone library method. *Biol. Pharmaceut. Bull.* 34, 1011–1020. doi: 10.1248/bpb.34.1011

Nagata, S., Yamashiro, Y., Ohtsuka, Y., Shimizu, T., Sakurai, Y., Misawa, S., et al. (2009). Heat shock proteins and superantigenic properties of bacteria from the gastrointestinal tract of patients with Kawasaki disease. *Immunology* 128, 511–520. doi: 10.1111/j.1365-2567.2009.03135.x

Nagi-Miura, N., Shingo, Y., Adachi, Y., Ishida-Okawara, A., Oharaseki, T., Takahashi, K., et al. (2004). Induction of coronary arteritis with administration of CAWS (*Candida albicans* water-soluble fraction) depending on mouse strains. *Immunopharmacol. Immunotoxicol.* 26, 527–543. doi: 10.1081/IPH-200042295

Nakamura, T., Yamamura, J., Sato, H., Kakinuma, H., and Takahashi, H. (2007). Vasculitis induced by immunization with Bacillus Calmette-Guerin followed by atypical mycobacterium antigen: a new mouse model for Kawasaki disease. *FEMS Immunol. Med. Microbiol.* 49, 391–397. doi: 10.1111/j.1574-695X.2007.00217.x

Nakamura, Y., Yashiro, M., Uehara, R., Oki, I., Watanabe, M., and Yanagawa, H. (2008). Monthly observation of the number of patients with Kawasaki disease and its incidence rates in Japan: chronological and geographical observation from nationwide surveys. *J. Epidemiol.* 18, 273–279. doi: 10.2188/jea.JE2008030

Nakamura, Y., Yashiro, M., Uehara, R., Sadakane, A., Tsuboi, S., Aoyama, Y., et al. (2012). Epidemiologic features of Kawasaki disease in Japan: results of the 2009-2010 nationwide survey. *J. Epidemiol.* 22, 216–221. doi: 10.2188/jea.JE20110126

Newburger, J. W., Takahashi, M., Beiser, A. S., Burns, J. C., Bastian, J., Chung, K. J., et al. (1991). A single intravenous infusion of gamma globulin as compared with four infusions in the treatment of acute Kawasaki syndrome. *N. Eng. J. med.* 324, 1633–1639. doi: 10.1056/NEJM199106063242305

Newburger, J. W., Takahashi, M., Gerber, M. A., Gewitz, M. H., Tani, L. Y., Burns, J. C., et al. (2004). Diagnosis, treatment, and long-term management of Kawasaki disease: a statement for health professionals from the Committee on Rheumatic Fever, Endocarditis, and Kawasaki Disease, Council on Cardiovascular Disease in the Young, American Heart Association. *Pediatrics* 114, 1708–1733. doi: 10.1542/peds.2004-2182

Nur-Ur Rahman, A. K., Bonsor, D. A., Herfst, C. A., Pollard, F., Peirce, M., Wyatt, A. W., et al. (2011). The T cell receptor beta-chain second complementarity determining region loop (CDR2beta governs T cell activation and Vbeta specificity by bacterial superantigens. *J. Biol. Chem.* 286, 4871–4881. doi: 10.1074/jbc.M110.189068

Ohno, N. (2004). Murine model of Kawasaki disease induced by mannoprotein-beta-glucan complex, CAWS, obtained from *Candida albicans*. *Jpn. J. Infect. Dis.* 57, S9–S10.

Onouchi, Y. (2012). Genetics of Kawasaki disease: what we know and don't know. *Circ. J.* 76, 1581–1586. doi: 10.1253/circj.CJ-12-0568

Onouchi, Y., Ozaki, K., Burns, J. C., Shimizu, C., Terai, M., Hamada, H., et al. (2012). A genome-wide association study identifies three new risk loci for Kawasaki disease. *Nat. Genet.* 44, 517–521. doi: 10.1038/ng.2220

Palmer, C., Bik, E. M., Digiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol.* 5:e177. doi: 10.1371/journal.pbio.0050177

Principi, N., Rigante, D., and Esposito, S. (2013). The role of infection in Kawasaki syndrome. *J. Infect.* 67, 1–10. doi: 10.1016/j.jinf.2013.04.004

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., et al. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature* 513, 59–64. doi: 10.1038/nature13568

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Stark, P. L., and Lee, A. (1982). The microbial ecology of the large bowel of breast-fed and formula-fed infants during the first year of life. *J. Med. Microbiol.* 15, 189–203. doi: 10.1099/00222615-15-2-189

Takeshita, S., Kobayashi, I., Kawamura, Y., Tokutomi, T., and Sekine, I. (2002a). Characteristic profile of intestinal microflora in Kawasaki disease. *Acta Paediatr.* 91, 783–788. doi: 10.1111/j.1651-2227.2002.tb03327.x

Takeshita, S., Nakatani, K., Kawase, H., Seki, S., Yamamoto, M., Sekine, I., et al. (1999). The role of bacterial lipopolysaccharide-bound neutrophils in the pathogenesis of Kawasaki disease. *J. Infect. Dis.* 179, 508–512. doi: 10.1086/314600

Takeshita, S., Tsujimoto, H., Kawase, H., Kawamura, Y., and Sekine, I. (2002b). Increased levels of lipopolysaccharide binding protein in plasma in children with kawasaki disease. *Clin. Diagn. Lab. Immunol.* 9, 205–206. doi: 10.1128/cdli.9.1.205-206.2002

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304

Uehara, R., and Belay, E. D. (2012). Epidemiology of Kawasaki disease in Asia, Europe, and the United States. *J. Epidemiol.* 22, 79–85. doi: 10.2188/jea.JE20110131

Walker, M. J., Barnett, T. C., Mcarthur, J. D., Cole, J. N., Gillen, C. M., Henningham, A., et al. (2014). Disease manifestations and pathogenic mechanisms of group a Streptococcus. *Clin. Microbiol. Rev.* 27, 264–301. doi: 10.1128/CMR.00101-13

Wang, C. L., Wu, Y. T., Liu, C. A., Kuo, H. C., and Yang, K. D. (2005). Kawasaki disease: infection, immunity and genetics. *Pediatr. Infect. Dis. J.* 24, 998–1004. doi: 10.1097/01.inf.0000183786.70519.fa

Whatmore, A. M., Efstratiou, A., Pickerill, A. P., Broughton, K., Woodard, G., Sturgeon, D., et al. (2000). Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "Atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. *Infect. Immun.* 68, 1374–1382. doi: 10.1128/IAI.68.3.1374-1382.2000

Yokota, S., Tsubaki, K., Kuriyama, T., Shimizu, H., Ibe, M., Mitsuda, T., et al. (1993). Presence in Kawasaki disease of antibodies to mycobacterial heat-shock protein HSP65 and autoantibodies to epitopes of human HSP65 cognate antigen. *Clin. Immunol. Immunopathol.* 67, 163–170. doi: 10.1006/clin.1993.1060

Yoshioka, T., Matsutani, T., Iwagami, S., Toyosaki-Maeda, T., Yutsudo, T., Tsuruta, Y., et al. (1999). Polyclonal expansion of TCRBV2- and TCRBV6-bearing T cells in patients with Kawasaki disease. *Immunology* 96, 465–472. doi: 10.1046/j.1365-2567.1999.00695.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

MINI REVIEW
published: 19 May 2016
doi: 10.3389/fmicb.2016.00712

Tracking Strains in the Microbiome: Insights from Metagenomics and Models

Ilana L. Brito[1,2] *and Eric J. Alm*[1,2]*

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA, [2] Center for Microbiome, Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA

Transmission usually refers to the movement of pathogenic organisms. Yet, commensal microbes that inhabit the human body also move between individuals and environments. Surprisingly little is known about the transmission of these endogenous microbes, despite increasing realizations of their importance for human health. The health impacts arising from the transmission of commensal bacteria range widely, from the prevention of autoimmune disorders to the spread of antibiotic resistance genes. Despite this importance, there are outstanding basic questions: what is the fraction of the microbiome that is transmissible? What are the primary mechanisms of transmission? Which organisms are the most highly transmissible? Higher resolution genomic data is required to accurately link microbial sources (such as environmental reservoirs or other individuals) with sinks (such as a single person's microbiome). New computational advances enable strain-level resolution of organisms from shotgun metagenomic data, allowing the transmission of strains to be followed over time and after discrete exposure events. Here, we highlight the latest techniques that reveal strain-level resolution from raw metagenomic reads and new studies that are tracking strains across people and environments. We also propose how models of pathogenic transmission may be applied to study the movement of commensals between microbial communities.

Keywords: microbiome, metagenomics, models, biological, strain diversity, genotyping techniques, bacterial genomics

OPEN ACCESS

Edited by:
Eamonn P. Culligan, Cork Institute of Technology, Ireland

Reviewed by:
C. Titus Brown, Michigan State University, USA
Jonathan Badger, National Cancer Institute, USA

***Correspondence:**
Eric J. Alm
ejalm@mit.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *02 November 2015*
Accepted: *28 April 2016*
Published: *19 May 2016*

Citation:
Brito IL and Alm EJ (2016) Tracking Strains in the Microbiome: Insights from Metagenomics and Models. Front. Microbiol. 7:712. doi: 10.3389/fmicb.2016.00712

Since the dawn of germ theory, epidemiology has focused on pathogens, their transmission routes and the consequences of their dispersal. Only recently have we fully appreciated the diverse roles of the thousands of microbial species that inhabit the human body. It is therefore sensible to broaden our questions about transmission dynamics and transmission routes to encompass the full range of commensal organisms. Recently, it has been suggested that diseases associated with dysbioses, such as Crohn's disease, rheumatoid arthritis and multiple sclerosis, may be transmissible (reviewed in Faith et al., 2015). There is also mounting evidence that the passive transmission of commensal bacteria may carry health benefits: in preventing obesity (Mueller et al., 2015), autoimmune disease (Olszak et al., 2012), and even certain cancers (Chen and Blaser, 2007; Hung and Wong, 2009). New therapeutics involve intentionally transmitting entire gut communities to treat recurrent *Clostridium difficile* infections (Kassam et al., 2013), and may ultimately be used to treat a wider array of conditions. Despite advances in DNA sequencing that have enabled

wide-scale characterizations of a large variety of microbial communities, little is known about how non-pathogenic microbes move between people and places.

For instance, we do not know what portion of the microbiome is transmissible. Research has instead focused on what *can* colonize, i.e., determining what factors impact colonization (Sonnenburg et al., 2005; Vaishnava et al., 2008; Goodman et al., 2009; Cullen et al., 2015), rather than what *does* colonize after exposure. What role does the transfer of organisms play in shaping either daily or punctuated shifts in our microbiomes? Our ability to answer these question currently relies on data from 16S marker gene surveys which can resolve differences between species. In some cases, coarse species-level data is sufficient to observe commensal transmission within the microbiome. In the gut, microbes associated with cured meat and cheese appear after ingestion (David et al., 2014a), and exogenous organisms repopulate the gut after acute gastrointestinal illness (David et al., 2014b). Likewise, contact with inanimate objects results in the transmission of commensals from our skin to proximal environments (Costello et al., 2009; Fierer et al., 2010; Lax et al., 2014). Perhaps unsurprisingly, infants are initially colonized by their mothers' skin and vaginal flora depending on birth method (Dominguez-Bello et al., 2010), with potentially long-term consequences for the infant (Munyaka et al., 2014). These studies suggest that we can begin to distinguish between exposure, transient and long-term colonization.

In addition to dynamics, by sampling broadly, we can further determine the routes of transmission among commensal organisms. Of the transmission routes that pathogens exploit—vertical, airborne, sexual, vector-borne, food-based, water-borne or healthcare-associated transmission—which ones are relevant to commensals? Many studies have surveyed the microbes present in each of these sources, but less research has focused on measuring human exposures and examining the dynamics of colonization. This will be easiest in cases involving discrete exposure events, but transmission may alternatively be fluid, that is to say that microbes are continually circulated within our proximal environments. Understanding these dynamics will assist future public health and environmental efforts to promote the spread of beneficial bacteria, while thwarting those that contribute to dysbioses. Measuring these impacts will undoubtedly benefit from higher resolution, strain-level distinctions, made possible by metagenomic whole microbiome shotgun sequencing.

DETERMINING TRANSMISSION ROUTES OF HUMAN-ASSOCIATED MICROBIOTA

In 1994, a gastroenterologist was brought to trial for intentionally infecting his girlfriend with HIV-1 virus carried by one of his patients. In order to prove the source of the girlfriend's infection, evidence was sought in the phylogenies of the virus's reverse transcriptase and envelope glycoprotein genes. Virus recovered from her blood was nested within a clade of the patient's, and 28 additional HIV patients from the area were all outgroups to this clade (Metzker et al., 2002). Only the less mutagenic RT sequences were adequate in showing that the strain present in the girlfriend was derived from the patient's HIV infection. This case is a good illustration of the evidence needed to establish transmission: the phylogeny of a gene that captures nested relationships, comprehensive sampling of potential sources to improve the likelihood of observing a direct transmission link, an organism that has an intermediate level of within-host evolution, and a putative transmission mechanism or discrete transmission event. While a transmission link may be impossible to prove conclusively from genomic data alone, these choices impact confidence in determining the timing and directionality of microbial transmission (reviewed in Pybus and Rambaut, 2009; Romero-Severson et al., 2014; **Figure 1**).

Can molecular epidemiology approaches, typically performed on one species alone, be applied to the diverse communities typical of the human microbiome? Although bacteria mutate less frequently than viral genomes, molecular epidemiology approaches have had some success in inferring the transmission of bacterial pathogens. For example, this was done in the case of the 2001 release of *Bacillus anthracis* in the mail system (Jernigan et al., 2002), as well as in reconstructing the transmission networks of several bacterial outbreaks (reviewed in Gardy et al., 2011; Snitkin et al., 2012; Fricke and Rasko, 2014; Gilchrist et al., 2015). More recently, they have been applied to identify strains of two endogenous human gut bacteria, *Methanobrevibacter smithii* and *Bacteroides thetaiotaomicron* shared between sets of twins (Faith et al., 2013).

Finding signals of transmission within metagenomic data may be made easier if there is more evolutionary divergence between samples. In the absence of high mutation rates, long-term carriage can result in greater within-host evolution, making it easier to reconstruct phylogenies. *Helicobacter pylori*, *Mycobacterium tuberculosis* and *Burkholderia dolosa*, a long-term infection associated with cystic fibrosis, are several bacteria that have accumulated an adequate number of mutations to track transmission across individuals (Falush et al., 2001; Gardy et al., 2011; Lieberman et al., 2011). Evidence that many commensal microbes have long-term residence in the gut and skin, (Faith et al., 2013; Schloissnig et al., 2013; David et al., 2014b; Oh et al., 2014), possibly dating back to birth (Dominguez-Bello et al., 2010), lends credence to applying molecular epidemiology approaches to a range of bacterial species in the human microbiome.

To attain the genomic resolution necessary to infer transmission, these studies have all relied on whole genome sequencing of cultured isolates. Applying this method to the greater variety of bacteria in the human microbiome would have limited scalability and would be restricted to culturable organisms. Single-cell techniques offer a way to circumvent culture limitations and the problems associated with genotyping strains that arise from short-read sequencing (discussed below). These can be technically challenging and costly, as hundreds of single-cell genomes per individual sample would be required to capture the diversity of strains of multiple species that are routine sampled using untargeted metagenomic sequencing. Rather, with short-read metagenomic sequencing, genomes of

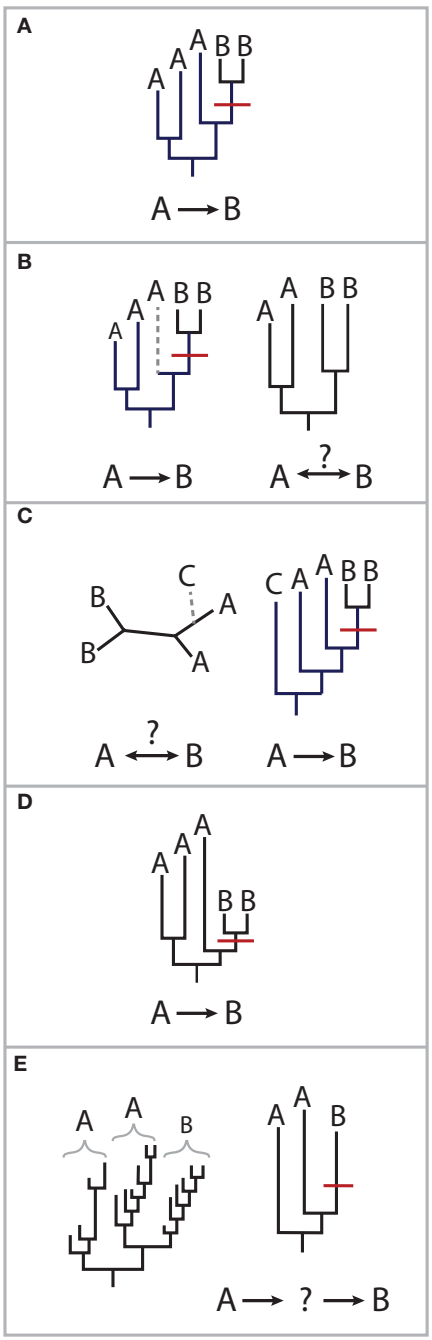

FIGURE 1 | Scenarios for molecular epidemiology approaches. (A) Nesting of one individuals' strain lineages within another's supports transmission from the host carrying the ancestral strain to the host carrying the more recently diverged strain, as shown here of a putative transmission event (shown in red) from person A to person B. **(B)** The loss of lineages can affect our ability to determine directionality. Given the same phylogeny in **(A)**, without the gray lineages, it is unclear which person's strains are ancestral. This can occur due to the choice of gene or characterizing fewer strains in an individual than what is present. **(C)** An outgroup helps distinguish transmission direction. Without lineage **(C)**, it is unclear whether **(A)** transmitted strains to **(B)** or vice versa. The inclusion of appropriate control samples can help reduce the likelihood of indirect transmission from an intermediate host or

(Continued)

FIGURE 1 | Continued
environmental source. In the 1994 case involving HIV, controls were chosen from HIV-infected individuals in the same geography, although not necessarily with the same risk factors (i.e., drug use, sexuality, hemophilia; Metzker et al., 2002). **(D)** Phylogenetic distances may not reflect the timing of transmission. An organism's rate of evolution may depend on factors specific to the individual, such as immunity, diet or genetics, which create different host selective pressures. **(E)** The rate of evolution of the marker gene is important to detect putative direct transmission. Long-term carriage of a microbe with high rates of evolution may result in long branch-lengths, upon which it becomes more difficult to exclude the possibility of indirect transmission.

many species may be acquired from a single sample, providing the raw data to infer transmission networks.

Comprehensive, metagenomic data is inherently more complex because it involves sequencing all bacterial, viral, and eukaryotic (including human) DNA present in a sample simultaneously, and the linkage of reads to each particular genome is lost during this process. To make sense of a diverse set of metagenomic reads, sequences must be aligned to reference genomes or *de novo* assembled draft genomes. Previous efforts to identify organisms this way have had mixed results: only 67% of culture-positive samples for Shiga-toxinogenic *E. coli* O104:H4 were identified by alignments to a *de novo* assembled genome of this organism (Loman et al., 2013). Disentangling genotypes down to the strain-level may be more complicated than this example for several reasons: genotyping strains from many species requires adequate coverage of each species, which may be hard to attain with the highly uneven distribution of species in a sample; individuals typically carry a handful of closely related strains within a species (Faith et al., 2013; Schloissnig et al., 2013; Oh et al., 2014); recombination may occur between closely related strains (Falush et al., 2001); and transmitted organisms are likely to resemble organisms already present in the gut (David et al., 2014b; Krebes et al., 2014). Yet, in order to get closer to proving transmission, we need an organismal resolution more fine-grained than species. The challenge will be to unambiguously genotype strains present within each individual.

ACHIEVING STRAIN-LEVEL ACCURACY

Metagenomic data is more appropriate for strain-calling than 16S rRNA amplicon data. The main reason is that metagenomic sequencing requires relatively few rounds of DNA amplification, compared to 16S amplicon sequencing, thus reducing the chance that PCR and sequencing errors are mistaken as genuine single nucleotide polymorphisms (SNPs). Although there are various computational methods available to address this issue with 16S amplicons, they usually carry the unintended consequence of a loss of resolution (Edgar et al., 2011; Quince et al., 2011; Schloss et al., 2011; Bokulich et al., 2013; Preheim et al., 2013). There is a cost to attaining higher resolution data. The main challenge in defining strains from short-read sequencing is that SNP frequencies in the genome that can be used to distinguish between recently diverged strains do not appear more than

once per 100–250 bp, which is the typical read length of ubiquitous high-throughput short-read sequencers. Therefore, metagenomic sequencing requires far more reads per sample to attain adequate coverage and depth of a genome required for phasing and distinguishing between strains. Also, rather than using standard analytical pipelines that exist for 16S, such as QIIME (Caporaso et al., 2010), there are no universally accepted methods for strain-level characterization from metagenomic data.

There have been several proposed strain-calling methods (**Table 1**), though most of these methods stop short of actually genotyping strains and instead focus on shared genomic features across samples, with the exception of ConStrains method which results in strain genotypes and their abundances (Luo et al., 2015). These methods generally rely on aligning reads to reference genomes, although this may be insufficient for unique samples for which reference genomes do not yet exist. Several methods overcome this limitation, enabling *de novo* assembly of genomes across metagenomic samples (Boisvert et al., 2012; Pell et al., 2012; Howe et al., 2014; Cleary et al., 2015). The Latent Strain Analysis method (Cleary et al., 2015) is notable because species of very low abundance (as low as 0.00001% in one case) distributed across many samples can be successfully assembled.

Both assembly- and alignment-based methods for genotyping strains require high depth and even coverage of each genome or DNA segment being analyzed. This is easily attainable for bacteria-rich samples such as the gut, where the predominance of bacteria results in relatively little human DNA. Conversely, in bacteria-poor environments that may be important for the study of transmission, such as the skin, a large fraction of the DNA sequenced, upwards of 90%, is from human cells (Human Microbiome Project Consortium, 2012). A greater amount of sequencing is therefore required to achieve adequate coverage of bacterial genomes. Additionally, the right-skewed abundance distributions of bacteria in some human body sites, such as the gut, contributes to this problem, such that large increases in sequencing depths are required to adequately cover lowly abundant organisms (Ni et al., 2013; Wendl et al., 2013). Since the costs associated with increased sequencing may soon cease to be a limiting factor and out-of-bag computational methods will become available, strain-level analysis may become as commonplace as marker gene analysis is today.

Newer sequencing approaches that produce longer read lengths may alleviate the need for such high sequencing depth and may allow for strain comparisons that utilize larger genomic regions than outlined in **Table 1** or even full genomes. The minION, made by Oxford Nanopore Technologies, has provided strain-level data in outbreak settings, specifically of Ebola (Quick et al., 2015) and *Salmonella* enterica serovar Enteritidis (Quick et al., 2016) that was used for transmission mapping. It has yet to be used to simultaneously examine the transmission of the numerous members of complex bacterial communities. Other experimental alternatives achieve synthetic long read lengths by manipulating amplification protocols to provide additional linkage information. For example, single kb-length molecules can each be sorted into a well, sheared, identically barcoded, and later assembled into one high fidelity scaffold (Kuleshov et al., 2014). Although this approach is lower throughput, it has been has been used together with short-read sequencing to improve scaffolding of highly-fragmented assemblies that can arise from *de novo* sequencing (Sharon et al., 2015). Proximity ligation is another experimental manipulation that uses Hi-C sequencing, i.e., intra-genome crosslinking, to link read-pairs arising from a single DNA molecule and has also been successfully used to genotypes strains within complex microbiome samples (Beitel et al., 2014; Burton et al., 2014). Although these technologies have been used on a very limited number of samples, they hold tremendous promise for achieving high confidence genotypes required to deconvolve chains of microbial transmission in complex communities.

FRONTIERS OF MICROBIAL TRANSMISSION STUDIES IN HEALTH AND THE ENVIRONMENT

We are now in an age where it is possible to engineer the microbiome to achieve therapeutic outcomes and modify our environments. Live bacterial therapeutics are already being used to treat *Clostridium difficile* infections (Kassam et al., 2013; Olle, 2013), and bioengineered therapeutics are on the horizon. Synthetic strains could be modified for a variety of applications within the human body, for enzyme replacement, disease

TABLE 1 | Methods for strain characterization from metagenomic data.

DNA regions	Considerations
SNPs within core or species-specific genes (Schloissnig et al., 2013; Oh et al., 2014; Ahn et al., 2015; Luo et al., 2015)	Methods either resolve genotypes or examine co-occurrences of SNPs. Genes may have different rates of evolution. Alignments may be difficult in the presence of closely related species.
Non-overlapping 1 kb windows (Franzosa et al., 2015)	Windows may contain a mix of horizontally transferred and core genomes. Limited phylogenetic analysis.
Copy-number variations of genes (Greenblum et al., 2015)	Rates of mutation may be harder to estimate.
Junctions of horizontally transferred regions and core genome (Raveh-Sadka et al., 2015)	Co-occurrence of transferred regions may change rapidly. Assembly may be difficult at repetitive regions common at HGT junctions. HGT may obscure phylogenetic patterns useful for inferring transmission.
CRISPR spacer comparisons (Raveh-Sadka et al., 2015)	Rates of spacer acquisition may be harder to estimate. Identifying source of mobile element may be difficult.

prevention, and diagnostic capabilities; or in the environment, for hazardous material remediation, pest control, and drought prevention. High confidence strain-tracking will be essential to gauge the dispersal of artificially introduced organisms. A handful of studies are beginning to track microbial strains, for example, after intentional inoculation. These include monitoring the infant gut microbiome throughout its development (Sharon et al., 2013); examining the donor and recipients of fecal microbiome transplants; and examining transmission in close-knit agrarian communities as part of the Fiji Community Microbiome Project (www.FijiCOMP.org).

Beyond characterizing strains within isolated samples, longitudinal strain-level data would allow us to approach the question posed earlier in this review: how does transmission impact daily or punctuated shifts in our microbiomes? While it may be straightforward to measure the impacts of transmission after a discrete event, in cases where transmission is continuous between source and sink, estimating rates of dispersal and transfer will be nontrivial. Mathematical models originally intended to capture animal movements or pathogen transmission may be adapted to account for the strain dynamics within diverse microbial communities. Metapopulation models, for example, describe environmental niches as "islands" between which organisms can migrate (Levins, 1969; Hanski, 1998). In the simplest of such models, unoccupied islands become occupied by the influx of bacteria from occupied islands, and extinction events in occupied islands may leave them unoccupied (**Figure 2A**). In the case of the human microbiome, these "islands" could be different individuals or body sites (Costello et al., 2012). Ecological disease models are similar to metapopulation models, though rather than colonizing islands, individuals are infected (**Figure 2B**). They differ in that individuals may transition from susceptible (S) to infected (I) classes, but may also transition to recovered classes (R) where they are no longer susceptible (Anderson and May, 1979). These SIR models come in a wide range of flavors and can be deterministic, stochastic, agent-based or spatially explicit, but they generally monitor the status of infected or uninfected units. Although infection will differ than colonization, these models provide analytical frameworks to start testing transmission rates and mechanisms.

Alternatively, there are models which account for the abundances of organisms within individuals or across a landscape, rather than their mere presence. Within-host pathogen models build upon the SIR model framework and track the abundances of a small number of strains resulting from mutation and local selection, as from immune pressure (Grenfell et al., 2004; Mideo et al., 2008; **Figure 2B**). Within-host and population-based SIR models can be nested as these dynamics may interact at different levels (reviewed in Mideo et al., 2008). Environmental fate-and-transport models similarly model pathogen abundances across landscape features and can incorporate environmental conditions that impact dispersal (reviewed in Brookes et al., 2004; Benham et al., 2006; **Figure 2C**). Fate-and-transport models may also be linked to SIR models to quantify bacterial exposures (Eisenberg et al., 2002). There is ample opportunity to apply

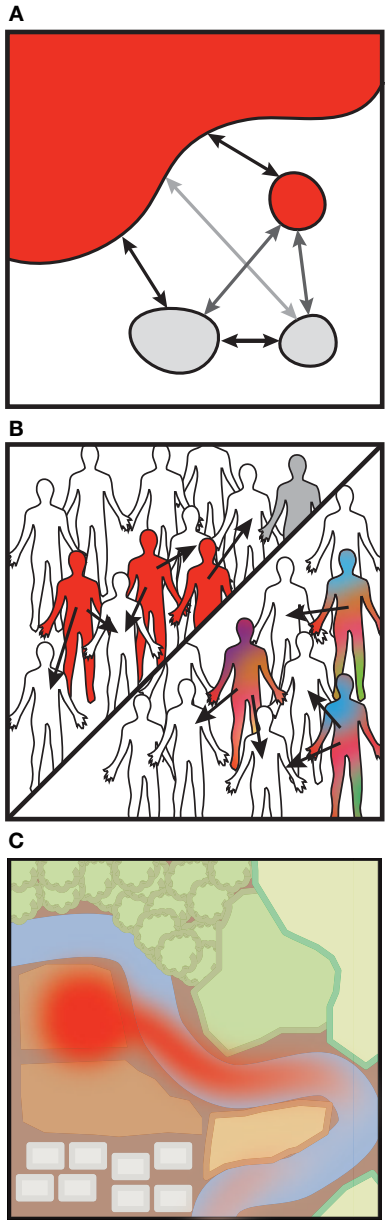

FIGURE 2 | Modeling bacterial transmission. (A) Metapopulation models. Change in island occupancy, by a microbe perhaps, is modeled as a function of migration (m) and an extinction rate (e). Other considerations such as a distance-based probability of infection may modify m.

$$\frac{dP}{dt} = mP\,(1 - P) - eP$$

(B) Susceptible-Infected-Resistant (SIR) models (with or without strain dynamics). Susceptible (S) individuals may become infected (I) and can recover and become immune. SIR models are similar to metapopulation models in that infection rate (β) is akin to migration between islands, as recovery (γ) is akin to extinction in the metapopulation model. Variations may include demographic processes, infection processes (latency, carriage), and alternative hosts or vectors.

$$\frac{dS}{dt} = -\beta SI$$

(Continued)

FIGURE 2 | Continued

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

SIR models that incorporate within-host evolution of specific strains typically are nested models that account for individuals' infection composition.
(C) Landscape fate-and-transport (F&T) models. F&T models estimate microbial abundances rather than a dichotomous infection status. The models stem from traditional advection-dispersion equations. Landscape features such as the surface porosity or water flow can be incorporated.

$$\frac{\partial C}{\partial x} = D\frac{\partial^2 C}{\partial^2 x} - v\frac{\partial C}{\partial x}$$

these techniques toward understanding microbiome-related transmission.

How can microbiome data be incorporated into transmission models? First, models designed for one microbial organism must be adapted to account for many. Parameterizing such models may be challenging given the broad differences in transmission observed between even closely related strains (Lee et al., 2013). Second, models of microbial communities may need to account for microbial interactions. Models of multiple pathogens show that complex dynamics can result from pathogen interactions (Rohani et al., 2003), and there are examples to suggest that this will be true for commensal organisms as well (David et al., 2014b; Hsiao et al., 2014; Seedorf et al., 2014). Lastly, we will also need to transform such models to accommodate compositional data. SIR models of more than one pathogen typically assume that measurements of each pathogen are independent (Rohani et al., 2003). Whereas counting microbes is technically challenging, microbial community measurements often reflect relative abundances of bacteria rather than absolute abundances. Although there are some methods that can escape this limitation (Friedman and Alm, 2012; Kurtz et al., 2015), we still lack principled methods to normalize time series compositional data. Figuring out how to incorporate multiple species into models of microbial transmission will be challenging but is a next logical step in our understanding of these communities.

In the near future, we predict that strain-tracking will become increasingly important, whether for epidemiology, forensics, environmental monitoring, or diagnostics. Metagenomics is currently the most straightforward and affordable data that can be used to track strains, and will likely be the primary source of those data in the near term. Despite the widespread availability of metagenomic sequencing, off-the-shelf methods to identify and evaluate the distribution of strains are still needed. In time, refinements will be made to determine what study design, sample preparation and sequencing depth are needed to substantiate claims of specific transmission chains. When that time comes, we may be able to quantify the role of commensal transmission in Crohn's disease, autoimmune disease, obesity and other microbiome-linked pathologies.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

We would like to thank the Neil and Anna Rasmussen Foundation for their support.

REFERENCES

Ahn, T.-H., Chai, J., and Pan, C. (2015). Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 31, 170–177. doi: 10.1093/bioinformatics/btu641

Anderson, R. M., and May, R. M. (1979). Population biology of infectious diseases: part I. *Nature* 280, 361–367.

Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., et al. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ.* 2:e415. doi: 10.7717/peerj.415

Benham, B. L., Baffaut, C., Zeckoski, R. W., Mankin, K. R., Pachepsky, Y. A., Sadeghi, A. M., et al. (2006). Modeling bacteria fate and transport in watersheds to support TMDLs. *Trans. ASABE* 49, 987–1002. doi: 10.13031/2013.21739

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13, R122. doi: 10.1186/gb-2012-13-12-r122

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Brookes, J. D., Antenucci, J., Hipsey, M., Burch, M. D., Ashbolt, N. J., and Ferguson, C. (2004). Fate and transport of pathogens in lakes and reservoirs. *Environ. Int.* 30, 741–759. doi: 10.1016/j.envint.2003.11.006

Burton, J. N., Liachko, I., Dunham, M. J., and Shendure, J. (2014). Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3* 4, 1339–1346. doi: 10.1534/g3.114.011825

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Chen, Y., and Blaser, M. J. (2007). Inverse associations of *Helicobacter pylori* with asthma and allergy. *Arch. Intern. Med.* 167, 821–827. doi: 10.1001/archinte.167.8.821

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., et al. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33, 1053–1060. doi: 10.1038/nbt.3329

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326, 1694–1697. doi: 10.1126/science.1177486

Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M., and Relman, D. A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* 336, 1255–1262. doi: 10.1126/science.1224203

Cullen, T. W., Schofield, W. B., Barry, N. A., Putnam, E. E., Rundell, E. A., Trent, M. S., et al. (2015). Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* 347, 170–175. doi: 10.1126/science.1260580

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014a). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15, R89. doi: 10.1186/gb-2014-15-7-r89

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014b). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11971–11975. doi: 10.1073/pnas.1002601107

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Eisenberg, J. N. S., Brookhart, M. A., Rice, G., Brown, M., and Colford, J. M. Jr., (2002). Disease transmission models for public health decision making: analysis of epidemic and endemic conditions caused by waterborne pathogens. *Environ. Health Perspect.* 110, 783–790. doi: 10.1289/ehp.02110783

Faith, J. J., Colombel, J.-F., and Gordon, J. I. (2015). Identifying strains that contribute to complex diseases through the study of microbial inheritance. *Proc. Natl. Acad. Sci. U.S.A.* 112, 633–640. doi: 10.1073/pnas.1418781112

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science* 341:1237439. doi: 10.1126/science.1237439

Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M., et al. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc. Natl. Acad. Sci. U.S.A.* 98, 15056–15061. doi: 10.1073/pnas.251396098

Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K., and Knight, R. (2010). Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6477–6481. doi: 10.1073/pnas.1000162107

Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J. M., et al. (2015). Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci. U.S.A.* 112, E2930–E2938. doi: 10.1073/pnas.1423854112

Fricke, W. F., and Rasko, D. A. (2014). Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat. Rev. Genet.* 15, 49–55. doi: 10.1038/nrg3624

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., et al. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* 364, 730–739. doi: 10.1056/NEJMoa1003176

Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A. Jr., and Hewlett, E. L. (2015). Whole-genome sequencing in outbreak analysis. *Clin. Microbiol. Rev.* 28, 541–563. doi: 10.1128/CMR.00075-13

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., et al. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* 6, 279–289. doi: 10.1016/j.chom.2009.08.003

Greenblum, S., Carr, R., and Borenstein, E. (2015). Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160, 583–594. doi: 10.1016/j.cell.2014.12.038

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., et al. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303, 327–332. doi: 10.1126/science.1090727

Hanski, I. (1998). Metapopulation dynamics. *Nature* 396, 41–49. doi: 10.1038/23876

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4904–4909. doi: 10.1073/pnas.1402564111

Hsiao, A., Ahmed, A. M. S., Subramanian, S., Griffin, N. W., Drewry, L. L., Petri, W. A., et al. (2014). Members of the human gut microbiota involved in recovery from Vibrio cholerae infection. *Nature* 515, 423–426. doi: 10.1038/nature13738

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Hung, I. F. N., and Wong, B. C. Y. (2009). Assessing the risks and benefits of treating *Helicobacter pylori* infection. *Ther. Adv. Gastroenterol.* 2, 141–147. doi: 10.1177/1756283X08100279

Jernigan, D. B., Raghunathan, P. L., Bell, B. P., Brechner, R., Bresnitz, E. A., Butler, J. C., et al. (2002). Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. *Emerging Infect. Dis.* 8, 1019–1028. doi: 10.3201/eid0810.020353

Kassam, Z., Lee, C. H., Yuan, Y., and Hunt, R. H. (2013). Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis. *Am. J. Gastroenterol.* 108, 500–508. doi: 10.1038/ajg.2013.59

Krebes, J., Didelot, X., Kennemann, L., and Suerbaum, S. (2014). Bidirectional genomic exchange between *Helicobacter pylori* strains from a family in Coventry, United Kingdom. *Int. J. Med. Microbiol.* 304, 1135–1146. doi: 10.1016/j.ijmm.2014.08.007

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., et al. (2014). Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol.* 32, 261–266. doi: 10.1038/nbt.2833

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.* 15, 237–240. doi: 10.1093/besa/15.3.237

Lax, S., Smith, D. P., Hampton-Marcell, J., Owens, S. M., Handley, K. M., Scott, N. M., et al. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* 345, 1048–1052. doi: 10.1126/science.1254529

Lee, S. M., Donaldson, G. P., Mikulski, Z., Boyajian, S., Ley, K., and Mazmanian, S. K. (2013). Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* 501, 426–429. doi: 10.1038/nature12447

Lieberman, T. D., Michel, J.-B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M. R., et al. (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat. Genet.* 43, 1275–1280. doi: 10.1038/ng.997

Loman, N. J., Constantinidou, C., Christner, M., Rohde, H., Chan, J. Z.-M., Quick, J., et al. (2013). A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 309, 1502–1510. doi: 10.1001/jama.2013.3231

Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., and Gevers, D. (2015). ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* 33, 1045–1052. doi: 10.1038/nbt.3319

Metzker, M. L., Mindell, D. P., Liu, X.-M., Ptak, R. G., Gibbs, R. A., and Hillis, D. M. (2002). Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14292–14297. doi: 10.1073/pnas.222522599

Mideo, N., Alizon, S., and Day, T. (2008). Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends Ecol. Evol.* 23, 511–517. doi: 10.1016/j.tree.2008.05.009

Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z., and Dominguez-Bello, M. G. (2015). The infant microbiome development: mom matters. *Trends Mol. Med.* 21, 109–117. doi: 10.1016/j.molmed.2014.12.002

Munyaka, P. M., Khafipour, E., and Ghia, J.-E. (2014). External influence of early childhood establishment of gut microbiota and subsequent health implications. *Front. Pediatr.* 2:109. doi: 10.3389/fped.2014.00109

Ni, J., Yan, Q., and Yu, Y. (2013). How much metagenomic sequencing is enough to achieve a given goal? *Sci. Rep.* 3, 1968. doi: 10.1038/srep01968

Oh, J., Byrd, A. L., Deming, C., Conlan, S., NISC Comparative Sequencing Program, Kong, H. H., et al. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59–64. doi: 10.1038/nature13786

Olle, B. (2013). Medicines from microbiota. *Nat. Biotechnol.* 31, 309–315. doi: 10.1038/nbt.2548

Olszak, T., An, D., Zeissig, S., Vera, M. P., Richter, J., Franke, A., et al. (2012). Microbial exposure during early life has persistent effects on natural killer T cell function. *Science* 336, 489–493. doi: 10.1126/science.1219328

Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J. M., and Brown, C. T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13272–13277. doi: 10.1073/pnas.1121464109

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., and Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the

operational taxonomic unit. *Appl. Environ. Microbiol.* 79, 6593–6603. doi: 10.1128/AEM.00342-13

Pybus, O. G., and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* 10, 540–550. doi: 10.1038/nrg2583

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol.* 16:114. doi: 10.1186/s13059-015-0677-2

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232. doi: 10.1038/nature16996

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12:38. doi: 10.1186/1471-2105-12-38

Raveh-Sadka, T., Thomas, B. C., Singh, A., Firek, B., Brooks, B., Castelle, C. J., et al. (2015). Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. *Elife* 4, 1–25. doi: 10.7554/eLife.05477

Rohani, P., Green, C. J., Mantilla-Beniers, N. B., and Grenfell, B. T. (2003). Ecological interference between fatal diseases. *Nature* 422, 885–888. doi: 10.1038/nature01542

Romero-Severson, E., Skar, H., Bulla, I., Albert, J., and Leitner, T. (2014). Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Mol. Biol. Evol.* 31, 2472–2482. doi: 10.1093/molbev/msu179

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature* 493, 45–50. doi: 10.1038/nature11711

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* 6:e27310. doi: 10.1371/journal.pone.0027310

Seedorf, H., Griffin, N. W., Ridaura, V. K., Reyes, A., Cheng, J., Rey, F. E., et al. (2014). Bacteria from diverse habitats colonize and compete in the mouse gut. *Cell* 159, 253–266. doi: 10.1016/j.cell.2014.09.008

Sharon, I., Kertesz, M., Hug, L. A., Pushkarev, D., Blauwkamp, T. A., Castelle, C. J., et al. (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res.* 25, 534–543. doi: 10.1101/gr.183012.114

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi: 10.1101/gr.142315.112

Snitkin, E. S., Zelazny, A. M., Thomas, P. J., Stock, F., NISC Comparative Sequencing Program Group, Henderson, D. K., et al. (2012). Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Sci. Transl. Med.* 4:148ra116. doi: 10.1126/scitranslmed.3004129

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C.-H., Westover, B. P., Weatherford, J., et al. (2005). Glycan foraging *in vivo* by an intestine-adapted bacterial symbiont. *Science* 307, 1955–1959. doi: 10.1126/science.1109051

Vaishnava, S., Behrendt, C. L., Ismail, A. S., Eckmann, L., and Hooper, L. V. (2008). Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20858–20863. doi: 10.1073/pnas.0808723105

Wendl, M. C., Kota, K., Weinstock, G. M., and Mitreva, M. (2013). Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *J. Math. Biol.* 67, 1141–1161. doi: 10.1007/s00285-012-0586-x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

OPINION
published: 04 August 2015
doi: 10.3389/fmicb.2015.00697

Pawnobiome: manipulation of the hologenome within one host generation and beyond

Jameson D. Voss[1]*, *Juan C. Leon*[1], *Nikhil V. Dhurandhar*[2] and *Frank T. Robb*[3]

[1] United States Air Force School of Aerospace Medicine, Epidemiology Consult Service, Wright Patterson AFB, OH, USA, [2] Department of Nutritional Sciences, Texas Tech University, Lubbock, TX, USA, [3] Department of Microbiology and Immunology, University of Maryland, Baltimore, MD, USA

Keywords: evolution, microbiome, hologenome, microbiota, pawnobe

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Yiorgos Apidianakis, University of Cyprus, Cyprus
Emiliano J. Salvucci, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
Eugene Rosenberg, Tel Aviv University, Israel

***Correspondence:**
Jameson D. Voss, jameson.voss@us.af.mil

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *20 April 2015*
Accepted: *26 June 2015*
Published: *04 August 2015*

Citation:
Voss JD, Leon JC, Dhurandhar NV and Robb FT (2015) Pawnobiome: manipulation of the hologenome within one host generation and beyond. Front. Microbiol. 6:697. doi: 10.3389/fmicb.2015.00697

Metaorganisms and Hologenome Theory of Evolution

A metaorganism is a collection of interacting organisms where the sum is not the same as the simple addition of the individual isolated parts (Relman, 2008; Webster, 2014). In fact, the gut, nasal, and lung microbiome all influence human phenotype (Redinbo, 2014). More specifically, the human gut microbiome has been linked to brain activity and to behavior (Collins et al., 2012). Similarly, *Bacillus amyloliquefaciens* supplementation improves feed conversion in chickens comparable to antibiotic growth promoters, by increasing villus height and crypt depth throughout the small intestine (Lei et al., 2015). It is apparent that phenotypic changes in the metaorganism influence the entire commensal unit.

The hologenome theory of evolution (HTE) asserts that a unit of selection is the holobiont which includes both the host and all its associated microbiota combined (Zilber-Rosenberg and Rosenberg, 2008). Some established components of the microbiota form a stable connection with the host; so, the entire holobiont is selected simultaneously with each passing host generation. The evolutionary fate of the holobiont unit is further linked with reliable vertical transmission of the microbiota whenever the host produces offspring.

The HTE is an important step forward in considering the evolutionary relevance of the wild-type microbiota, but it is not meant to characterize opportunities in deliberately manipulating and selecting microbes. Additionally, some microbes do not fit well within the HTE because they do not reliably transmit vertically, or they only influence host phenotype transiently. For instance, in one study, a yogurt probiotic altered bacterial carbohydrate metabolism markers without altering the species composition of the fecal bacteria (McNulty et al., 2011; Sanders et al., 2013). Similarly, fecal transplant appears promising for diabetes treatment, but thus far, has only been shown to improve insulin sensitivity temporarily (Vrieze et al., 2013). While a majority of the gut microbes in humans are stable day to day (Lozupone et al., 2012), only 60% of strains are durable beyond 5 years (Faith et al., 2013). These observations are the basis for our concept of the "Pawnobiome," defined as the subset of the microbiome that is purposefully managed for manipulation of the host phenotype, which includes individual microbes named "pawnobes."

Characteristics of the Pawnobiome

As we are defining it, the pawnobiome exists at a border between a stable relationship with the host and an unstable one. If a microbe is in a stable symbiosis which cannot be manipulated independently from the host, it is not a part of the pawnobiome. In other words, the pawnobiome is at the critical interface between temporary and permanent residence in the

hologenome. By allowing both phenotype conservation and innovation, this criticality is likely an important factor determining evolvability of the hologenome (Torres-Sosa et al., 2012). Unlike the microbiota in the HTE, the pawnobiome is not strictly dependent upon a particular host's survival or generation time and can evolve independently and more rapidly than the host. The pawnobiome theory of evolution (PTE) is that as artificial evolution occurs within the pawnobiome, the host phenotype can be substantially altered within a single host generation.

Further, the pawnobes are also genetically adaptable. Gut bacteria, for instance, are known to be hypermutable *in vitro* (LeClerc et al., 1996; Lee et al., 2008) and an experimental *Escherichia coli* model progressed through potentiation, actualization, and refinement over 33,000 generations to gain a novel metabolic function (Blount et al., 2012). With *in vivo* observations, gut bacteria have shown substantial adaptability over a wide range of timescales (Quercia et al., 2014). In addition to criticality and modifiability, a third important feature of the pawnobes is transmissibility. For instance, if a pawnobe enhanced fitness of the host, widespread horizontal transmission could be possible [exemplified by stool donor banks (Burns et al., 2015)].

The term "pawn" has many connotations, but characteristics of criticality, transmissibility and adaptability are particularly relevant to the present theory. The term "pawn" reflects exchange of goods with lending and trade. The analogy can be extended to pawn shops, which exist at the border between regulated commercial exchange and unregulated barter. Further, a broker can reappraise, repackage, and combine with other goods to alter the value based on observable characteristics and transmit to a new owner (i.e., host). Similarly, these features of criticality, transmissibility, and modifiability are also seen in the game of chess. Here, pawns are the strategically important, least glorious pieces that make up the front line, buffering the more valuable pieces that one can control with the pieces one cannot (i.e., criticality). Pawns are captured in a gambit (i.e., transmissibility). Finally, if they survive and advance to the end of the board, they can be upgraded (i.e., modified) to a more functional piece during a single game.

Ultimately, the prefix "pawn" carries the connotation of strategic domestication. The term "domesticated microbiome," is similar with "pawnobiome," but we argue it is less precise because it suggests every microbe interacting with the host is domesticated, which would be unusual at the present time since undomesticated microbiota are still predominant.

Opportunities in Pawnobe Selection

In 1859, Darwin and Bynum devoted the first chapter of "On the Origin of Species" to the variation in domesticated plants and animals. The concluding remarks of chapter 1 are a remarkable foresight into the current groundbreaking developments that are revealing the impact of artificial selection of commensal microbial species. He wrote, "…the most important point of all, is, that the animal or plant should be so highly useful to man, or so much valued by him, that the closest attention should be paid to even the slightest deviation in…each individual."

Darwin described these insights on domesticated species separately from his observations on wildlife; similarly, the products of artificial selection in both kingdom Animalia (i.e., livestock) and Plantae (i.e., cultigens) have a separate name to describe their domestication as we are now also proposing for microbes (i.e., pawnobes).

Within a single host, the microbiome is a large (>100 fold more microbial genetic material than the human genome) and diverse population, (Ezenwa et al., 2012; Lozupone et al., 2012) which creates extraordinary opportunity for artificial selection. The pawnobiome population size can be amplified even more with a large number of hosts. For instance, some skin microbes appear to act as insect repellants (Ezenwa et al., 2012). After using the "closest attention" and selecting the most repellant microbes once, these microbes could be re-challenged and iteratively selected in a large number of hosts to repel medically important insect vectors.

The PTE proposes some elements of the microbiome are modifiable over a short time scale even if others are more difficult to change. Maximizing the utility in medicine, agriculture, and basic science will require new methods (e.g., trans-species artificial selection) to help optimize the pawnobiome.

The utility of the pawnobiome concept is experimentally testable. Because the transmissibility characteristic of pawnobes is not necessarily limited by species or other phylogenetic boundaries, a biocontained murine model could be used for multiple phenotypic traits that can be assessed in mice. Fortunately, mice are an ideal species to test artificial selection using fecal-oral transmission since they can be raised in sterile conditions, producing so-called gnotobiotic mice that are effective models for culturing the human gut microbiome (Goodman et al., 2011), and because they naturally engage in coprophagy (Ridaura et al., 2013). For instance, after administering a commercially successful chicken probiotic such as *B. amyloliquefaciens* (Lei et al., 2015) to gnotobiotic mice, frequent serial passage of the stool of mice with the highest feed conversion could continue until an optimized probiotic was isolated and sequenced.

Another application would create an optimized gut microbiome to resist the metabolic consequences of consumption of a high calorie diet by sedentary individuals. Already, observations in a human trial have identified so called, "super donors" who appear to provide a notably larger metabolic benefit to others upon stool transplant (Udayappan et al., 2014). We propose serial artificial selection could continue after identifying successful transplants. So, to begin the process, stool from a "super donor" could seed a large population of genetically homogenous mice eating a metabolically unhealthy diet (Ridaura et al., 2013). Then, after assessing target metabolic parameters (e.g., body weight, blood glucose, lipids, etc.) at frequent intervals, the stool from the leanest and otherwise healthiest mice could be redistributed to all other mice. Once metabolic parameters were optimized, the final product could be analyzed using community sequencing and metabolomics (Marcobal et al., 2013), and reference data from the Human

Microbiome Project (Peterson et al., 2009). The entire stool, promising components, or individual products could proceed for further testing in animals and finally humans.

Aside from artificial selection, supplementation with probiotics (McNulty et al., 2011; Biagi et al., 2013; Sanders et al., 2013; Hulston et al., 2015), prebiotics (Biagi et al., 2013; Holscher et al., 2015), antibiotics (Cho et al., 2012), and dysbiotics (e.g., emulsifiers) (Chassaing et al., 2015) could alter host phenotype by changing the pawnobiome.

Pawnobiome Selection Limitations and Insights

While enhanced pawnobe selection holds promise, there are at least three limiting factors: observability, attribution, and permanence. Assessment of positive traits for selection requires detectable variation between otherwise genetically homogenous individuals. Additionally, the etiology of the observable variation needs to be attributable to the pawnobes that can be transmitted with the chosen method (i.e., fecal-oral). Another potential threat is permanence (i.e., undesirable chronic effects are not identifiable with short term selection of a transient phenotype). In humans, researchers use screening criteria to select donors without chronic disease for fecal transplant (Vrieze et al., 2013); such a technique could also be used to eliminate highly successful pathogens such as *Pseudomonas aeruginosa* (Markou and Apidianakis, 2014). Even if serial passage in mice led to the emergence of a pathogen, any short term desirable changes in phenotype could be evaluated to try to replicate the effect with a non-infectious vehicle.

Interestingly, the infectious disease risk in pawnobe artificial selection may be an under-recognized threat in other forms of artificial selection (e.g., antibiotic use, agricultural selection). While pawnobe selection occurs over smaller timescales, aggressive selection could pose the same risk for creating emerging pathogens (or releasing control of commensals that are conditional pathogens, like *Clostridium difficile*) over larger timescales. Furthermore, aggressive artificial selection among one species (e.g., a livestock species) could select microbiota that transmit traits by horizontal gene transfer (or other mechanisms) to multiple species simultaneously. Human gut bacteria are transferable between species (Sun et al., 2015), not surprising since humans share microbiota with their cohabiting dogs (Song et al., 2013; Udayappan et al., 2014).

Livestock breeds have undergone aggressive selection for metabolic characteristics that are commercially favorable. There is genetic evidence of selection for fat deposition in sheep (Moradi et al., 2012), feed efficiency in cattle (Bovine HapMap Consortium, 2009), and metabolic regulation in chickens (Rubin et al., 2010). Recently, a chicken breed that was originally commercialized in 1957 and another breed commercialized in 2005 were both raised simultaneously under the same management with the same food (Zuidhof et al., 2014). Under the same regime, the 2005 chicken breed weighed four times as much as the 1957 breed (Zuidhof et al., 2014). Such divergent phenotypes over a short period evidences aggressive recent selection of the host genome, and per the HTE, there should have been corresponding selection of the microbiota that influenced host phenotype in the same direction.

If there was a trans-species effect from aggressive selection in one species it might be detected in the body weight of interacting species. In fact, all animals with historical body weight records have gained weight in mid-life over the past several decades (Klimentidis et al., 2011). Several microbes associated with livestock are known to cause obesity in animals or are associated with human obesity. For instance, gut bacteria in the genus *Lachnospiraceae* are associated with cattle rumination and antibiotic weight gain in several species (Cho et al., 2012; Meehan and Beiko, 2014). Likewise, Adenoviruses (e.g., SMAM1, Ad-36), (Dhurandhar et al., 1992, 1997, 2001; Shang et al., 2014) *Toxoplasma gondii* (Carter, 2013; Reeves et al., 2013), and transmissible spongiform encephalopathies [i.e., Kuru, (Collinge et al., 2008) Creutzfeldt-Jakob Disease, (Manuelidis et al., 2009) Bovine Spongiform Encephalopathy (Strom et al., 2014), and scrapie agents (Kim et al., 1987)] are associated with obesity. Thus, shedding or acquisition of pawnobe species could provide new insights into infectious disease dynamics (Colizza and Vespignani, 2008) and other biological variation.

Conclusions

In summary, the pawnobiome theory describes commensal microbiota that impact host phenotype but can be independently selected. Like other evolutionary developments in genetics and microbiota, pawnobe studies could be applied to agriculture (Thrall et al., 2011). Additionally, pawnobiome host interactions may provide insights for biological theories [e.g., autocenosis and democenosis (Savinov, 2011) symbiogenesis (Mereschkovsky, 1909; Kozo-Polyansky, 1924), synergistic selection (Corning and Szathmáry, 2015), teleonomy (Corning, 2014), endophyte studies (Taghavi et al., 2009), and the hygiene hypothesis (Strachan, 2000)]. Purposeful and cautious artificial selection could have broad ranging applications within biotechnology, health care, and evolutionary biology. Over time, new technologies and methods for strategic selection of the pawnobiome could accelerate this utility.

Funding

The work was partially supported by the Air Force Office of Scientific Research by Grants AFOSR 03-S-28900 and 9550-10-1-0272 and by the NASA Exobiology Program (FTR).

Acknowledgments

We thank the reviewers for making important intellectual contributions. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Air Force, the Department of Defense, or the U.S. Government. Distribution A: Approved for public release; distribution is unlimited. Case Number: 88ABW-2015-1629, 31 Mar 2015.

References

Biagi, E., Candela, M., Turroni, S., Garagnani, P., Franceschi, C., and Brigidi, P. (2013). Ageing and gut microbes: perspectives for health maintenance and longevity. *Pharmacol. Res.* 69, 11–20. doi: 10.1016/j.phrs.2012.10.005

Blount, Z. D., Barrick, J. E., Davidson, C. J., and Lenski, R. E. (2012). Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature* 489, 513–518. doi: 10.1038/nature11514

Bovine HapMap Consortium. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324, 528–532. doi: 10.1126/science.1167936

Burns, L. J., Dubois, N., Smith, M. B., Mendolia, G. M., Burgess, J., Edelstein, C., et al. (2015). 499 Donor recruitment and eligibility for fecal microbiota transplantation: results from an international public stool bank. *Gastroenterology* 148, S-96–S-97. doi: 10.1016/s0016-5085(15)30331-0

Carter, C. J. (2013). Toxoplasmosis and polygenic disease susceptibility genes: extensive toxoplasma gondii host/pathogen Interactome enrichment in nine psychiatric or neurological disorders. *J. Pathog.* 2013:965046. doi: 10.1155/2013/965046

Chassaing, B., Koren, O., Goodrich, J. K., Poole, A. C., Srinivasan, S., Ley, R. E., et al. (2015). Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome. *Nature* 519, 92–96. doi: 10.1038/nature14232

Cho, I., Yamanishi, S., Cox, L., Methé, B. A., Zavadil, J., Li, K., et al. (2012). Antibiotics in early life alter the murine colonic microbiome and adiposity. *Nature* 488, 621–626. doi: 10.1038/nature11400

Colizza, V., and Vespignani, A. (2008). Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. *J. Theor. Biol.* 251, 450–467. doi: 10.1016/j.jtbi.2007.11.028

Collinge, J., Whitfield, J., McKintosh, E., Frosh, A., Mead, S., Hill, A. F., et al. (2008). A clinical study of kuru patients with long incubation periods at the end of the epidemic in Papua New Guinea. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363, 3725–3739. doi: 10.1098/rstb.2008.0068

Collins, S. M., Surette, M., and Bercik, P. (2012). The interplay between the intestinal microbiota and the brain. *Nat. Rev. Microbiol.* 10, 735–742. doi: 10.1038/nrmicro2876

Corning, P. A. (2014). Evolution 'on purpose': how behaviour has shaped the evolutionary process. *Biol. J. Linnean Soc.* 112, 242–260. doi: 10.1111/bij.12061

Corning, P. A., and Szathmáry, E. (2015). "Synergistic selection": a darwinian frame for the evolution of complexity. *J. Theor. Biol.* 371, 45–58. doi: 10.1016/j.jtbi.2015.02.002

Darwin, C., and Bynum, W. F. (1859). *The Origin of Species by Means of Natural Selection: or, the Preservation of Favored Races in the Struggle for Life*. Wikisource, Available online at: http://en.wikisource.org/wiki/On_the_Origin_of_Species_%281859%29

Dhurandhar, N. V., Israel, B. A., Kolesar, J. M., Mayhew, G., Cook, M. E., and Atkinson, R. L. (2001). Transmissibility of adenovirus-induced adiposity in a chicken model. *Int. J. Obes. Relat. Metab. Disord.* 25, 990–996. doi: 10.1038/sj.ijo.0801668

Dhurandhar, N. V., Kulkarni, P., Ajinkya, S. M., and Sherikar, A. (1992). Effect of adenovirus infection on adiposity in chicken. *Vet. Microbiol.* 31, 101–107. doi: 10.1016/0378-1135(92)90068-5

Dhurandhar, N. V., Kulkarni, P. R., Ajinkya, S. M., Sherikar, A. A., and Atkinson, R. L. (1997). Association of adenovirus infection with human obesity. *Obes. Res.* 5, 464–469. doi: 10.1002/j.1550-8528.1997.tb00672.x

Ezenwa, V. O., Gerardo, N. M., Inouye, D. W., Medina, M., and Xavier, J. B. (2012). Animal behavior and the microbiome. *Science* 338, 198–199. doi: 10.1126/science.1227412

Faith, J. J., Guruge, J. L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A. L., et al. (2013). The long-term stability of the human gut microbiota. *Science* 341:1237439. doi: 10.1126/science.1237439

Goodman, A. L., Kallstrom, G., Faith, J. J., Reyes, A., Moore, A., Dantas, G., et al. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6252–6257. doi: 10.1073/pnas.1102938108

Holscher, H. D., Caporaso, J. G., Hooda, S., Brulc, J. M., Fahey, G. C. Jr., and Swanson, K. S. (2015). Fiber supplementation influences phylogenetic structure and functional capacity of the human intestinal microbiome: follow-up of a randomized controlled trial. *Am. J. Clin. Nutr.* 101, 55–64. doi: 10.3945/ajcn.114.092064

Hulston, C. J., Churnside, A. A., and Venables, M. C. (2015). Probiotic supplementation prevents high-fat, overfeeding-induced insulin resistance in human subjects. *Br. J. Nutr.* 113, 596–602. doi: 10.1017/s0007114514004097

Kim, Y. S., Carp, R. I., Callahan, S. M., and Wisniewski, H. M. (1987). Scrapie-induced obesity in mice. *J. Infect. Dis.* 156, 402–405. doi: 10.1093/infdis/156.2.402

Klimentidis, Y. C., Beasley, T. M., Lin, H. Y., Murati, G., Glass, G. E., Guyton, M., et al. (2011). Canaries in the coal mine: a cross-species analysis of the plurality of obesity epidemics. *Proc. Biol. Sci.* 278, 1626–1632. doi: 10.1098/rspb.2010.1890

Kozo-Polyansky, B. M. (1924). *A New Principle of Biology*. Moscow: Essay on the Theory of Symbiogenesis.

LeClerc, J. E., Li, B., Payne, W. L., and Cebula, T. A. (1996). High mutation frequencies among escherichia coli and salmonella pathogens. *Science* 274, 1208–1211. doi: 10.1126/science.274.5290.1208

Lee, J. H., Karamychev, V. N., Kozyavkin, S. A., Mills, D., Pavlov, A. R., Pavlova, N. V., et al. (2008). Comparative genomic analysis of the gut bacterium *Bifidobacterium longum* reveals loci susceptible to deletion during pure culture growth. *BMC Genomics* 9:247. doi: 10.1186/1471-2164-9-247

Lei, X., Piao, X., Ru, Y., Zhang, H., Peron, A., and Zhang, H. (2015). Effect of bacillus amyloliquefaciens-based direct-fed microbial on performance, nutrient utilization, intestinal morphology and cecal microflora in broiler chickens. *Asian-Australas J. Anim. Sci.* 28, 239–246. doi: 10.5713/ajas.14.0330

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489, 220–230. doi: 10.1038/nature11550

Manuelidis, L., Chakrabarty, T., Miyazawa, K., Nduom, N. A., and Emmerling, K. (2009). The kuru infectious agent is a unique geographic isolate distinct from creutzfeldt-jakob disease and scrapie agents. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13529–13534. doi: 10.1073/pnas.0905825106

Marcobal, A., Kashyap, P. C., Nelson, T. A., Aronov, P. A., Donia, M. S., Spormann, A., et al. (2013). A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *ISME J.* 7, 1933–1943. doi: 10.1038/ismej.2013.89

Markou, P., and Apidianakis, Y. (2014). Pathogenesis of intestinal pseudomonas aeruginosa infection in patients with cancer. *Front. Cell. Infect. Microbiol.* 3:115. doi: 10.3389/fcimb.2013.00115

McNulty, N. P., Yatsunenko, T., Hsiao, A., Faith, J. J., Muegge, B. D., Goodman, A. L., et al. (2011). The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci. Transl. Med.* 3, 106ra106. doi: 10.1126/scitranslmed.3002701

Meehan, C. J., and Beiko, R. G. (2014). A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biol. Evol.* 6, 703–713. doi: 10.1093/gbe/evu050

Mereschkovsky, K. C. (1909). "The theory of two plasms as the basis of symbiogenesis, new studies about the origins of organisms," in *Proceedings of the Studies of the Imperial Kazan University* (Kazan).

Moradi, M. H., Nejati-Javaremi, A., Moradi-Shahrbabak, M., Dodds, K. G., and McEwan, J. C. (2012). Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genet.* 13:10. doi: 10.1186/1471-2156-13-10

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., et al. (2009). The NIH human microbiome project. *Genome Res.* 19, 2317–2323. doi: 10.1101/gr.096651.109

Quercia, S., Candela, M., Giuliani, C., Turroni, S., Luiselli, D., Rampelli, S., et al. (2014). From lifetime to evolution: timescales of human gut microbiota adaptation. *Front. Microbiol.* 5:587. doi: 10.3389/fmicb.2014.00587

Redinbo, M. R. (2014). The microbiota, chemical symbiosis, and human disease. *J. Mol. Biol.* 426, 3877–3891. doi: 10.1016/j.jmb.2014.09.011

Reeves, G. M., Mazaheri, S., Snitker, S., Langenberg, P., Giegling, I., Hartmann, A. M., et al. (2013). A positive association between *T. gondii* seropositivity and obesity. *Front. Public Health* 1:73. doi: 10.3389/fpubh.2013.00073

Relman, D. A. (2008). 'Til death do us part': coming to terms with symbiotic relationships. *Nat. Rev. Microbiol.* 6, 721–724. doi: 10.1038/nrmicro1990

Ridaura, V. K., Faith, J. J., Rey, F. E., Cheng, J., Duncan, A. E., Kau, A. L., et al. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* 341:1241214. doi: 10.1126/science.1241214

Rubin, C. J., Zody, M. C., Eriksson, J., Meadows, J. R., Sherwood, E., Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591. doi: 10.1038/nature08832

Sanders, M. E., Guarner, F., Guerrant, R., Holt, P. R., Quigley, E. M., Sartor, R. B., et al. (2013). An update on the use and investigation of probiotics in health and disease. *Gut* 62, 787–796. doi: 10.1136/gutjnl-2012-302504

Savinov, A. (2011). Autocenosis and democenosis as individual-and population-level ecological categories in terms of symbiogenesis and systems approach. *Russ. J. Ecol.* 42, 179–185. doi: 10.1134/S1067413611030131

Shang, Q., Wang, H., Song, Y., Wei, L., Lavebratt, C., Zhang, F., et al. (2014). Serological data analyses show that adenovirus 36 infection is associated with obesity: a meta-analysis involving 5739 subjects. *Obesity (Silver. Spring).* 22, 895–900. doi: 10.1002/oby.20533

Song, S. J., Lauber, C., Costello, E. K., Lozupone, C. A., Humphrey, G., Berg-Lyons, D., et al. (2013). Cohabiting family members share microbiota with one another and with their dogs. *Elife* 2:e00458. doi: 10.7554/elife.00458

Strachan, D. P. (2000). Family size, infection and atopy: the first decade of the 'hygiene hypothesis'. *Thorax* 55:S2. doi: 10.1136/thorax.55.suppl_1.S2

Strom, A., Yutzy, B., Kruip, C., Ooms, M., Schloot, N. C., Roden, M., et al. (2014). Foodborne transmission of bovine spongiform encephalopathy to non-human primates results in preclinical rapid-onset obesity. *PLoS ONE* 9:e104343. doi: 10.1371/journal.pone.0104343

Sun, Z., Zhang, W., Guo, C., Yang, X., Liu, W., Wu, Y., et al. (2015). Comparative genomic analysis of 45 type strains of the genus *Bifidobacterium*: a snapshot of its genetic diversity and evolution. *PLoS ONE* 10:e0117912. doi: 10.1371/journal.pone.0117912

Taghavi, S., Garafola, C., Monchy, S., Newman, L., Hoffman, A., Weyens, N., et al. (2009). Genome survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and development of poplar trees. *Appl. Environ. Microbiol.* 75, 748–757. doi: 10.1128/AEM.02239-08

Thrall, P. H., Oakeshott, J. G., Fitt, G., Southerton, S., Burdon, J. J., Sheppard, A., et al. (2011). Evolution in agriculture: the application of evolutionary approaches to the management of biotic interactions in agro-ecosystems. *Evol. Appl.* 4, 200–215. doi: 10.1111/j.1752-4571.2010.00179.x

Torres-Sosa, C., Huang, S., and Aldana, M. (2012). Criticality is an emergent property of genetic networks that exhibit evolvability. *PLoS Comput. Biol.* 8:e1002669. doi: 10.1371/journal.pcbi.1002669

Udayappan, S. D., Hartstra, A. V., Dallinga-Thie, G. M., and Nieuwdorp, M. (2014). Intestinal microbiota and faecal transplantation as treatment modality for insulin resistance and type 2 diabetes mellitus. *Clin. Exp. Immunol.* 177, 24–29. doi: 10.1111/cei.12293

Vrieze, A., de Groot, P. F., Kootte, R. S., Knaapen, M., van Nood, E., and Nieuwdorp, M. (2013). Fecal transplant: a safe and sustainable clinical therapy for restoring intestinal microbial balance in human disease? *Best Pract. Res. Clin. Gastroenterol.* 27, 127–137. doi: 10.1016/j.bpg.2013.03.003

Webster, N. S. (2014). Cooperation, communication, and co-evolution: grand challenges in microbial symbiosis research. *Front. Microbiol.* 5:164. doi: 10.3389/fmicb.2014.00164

Zilber-Rosenberg, I., and Rosenberg, E. (2008). Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiol. Rev.* 32, 723–735. doi: 10.1111/j.1574-6976.2008.00123.x

Zuidhof, M. J., Schneider, B. L., Carney, V. L., Korver, D. R., and Robinson, F. E. (2014). Growth, efficiency, and yield of commercial broilers from 1957, 1978, and 2005. *Poult. Sci.* 93, 2970–2982. doi: 10.3382/ps.2014-04291

Conflict of Interest Statement: Jameson D. Voss, Juan C. Leon, and Frank T. Robb have nothing to declare. Nikhil V. Dhurandhar declares several patents in viral obesity and Adenovirus 36 including uses for E1A, E4orf1 gene and protein, and AKT1 inhibitor. He also declares ongoing grant support from Vital Health Interventions for determining anti-diabetic properties of E4orf1 protein.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 19 March 2015
doi: 10.3389/fmicb.2015.00219

The composition of the global and feature specific cyanobacterial core-genomes

Stefan Simm[1], *Mario Keller*[1], *Mario Selymesi*[1] *and Enrico Schleiff*[1, 2, 3, 4]*

[1] Department of Biosciences, Molecular Cell Biology of Plants, Goethe University, Frankfurt am Main, Germany, [2] Cluster of Excellence Frankfurt, Goethe University, Frankfurt am Main, Germany, [3] Center of Membrane Proteomics, Goethe University, Frankfurt am Main, Germany, [4] Buchmann Institute of Molecular Life Sciences, Goethe University, Frankfurt am Main, Germany

Cyanobacteria are photosynthetic prokaryotes important for many ecosystems with a high potential for biotechnological usage e.g., in the production of bioactive molecules. Either asks for a deep understanding of the functionality of cyanobacteria and their interaction with the environment. This in part can be inferred from the analysis of their genomes or proteomes. Today, many cyanobacterial genomes have been sequenced and annotated. This information can be used to identify biological pathways present in all cyanobacteria as proteins involved in such processes are encoded by a so called core-genome. However, beside identification of fundamental processes, genes specific for certain cyanobacterial features can be identified by a holistic genome analysis as well. We identified 559 genes that define the core-genome of 58 analyzed cyanobacteria, as well as three genes likely to be signature genes for thermophilic and 57 genes likely to be signature genes for heterocyst-forming cyanobacteria. To get insights into cyanobacterial systems for the interaction with the environment we also inspected the diversity of the outer membrane proteome with focus on β-barrel proteins. We observed that most of the transporting outer membrane β-barrel proteins are not globally conserved in the cyanobacterial phylum. In turn, the occurrence of β-barrel proteins shows high strain specificity. The core set of outer membrane proteins globally conserved in cyanobacteria comprises three proteins only, namely the outer membrane β-barrel assembly protein Omp85, the lipid A transfer protein LptD, and an OprB-type porin. Thus, we conclude that cyanobacteria have developed individual strategies for the interaction with the environment, while other intracellular processes like the regulation of the protein homeostasis are globally conserved.

Keywords: cyanobacteria, *Anabaena* sp. PCC 7120, core-genome, genotypic and phenotypic differences, ortholog search, comparative genomics

OPEN ACCESS

Edited by:
Eamonn P. Culligan, University College Cork, Ireland

Reviewed by:
Loren John Hauser, Oak Ridge National Laboratory, USA
Wolfgang R. Hess, University of Freiburg, Germany

***Correspondence:**
Enrico Schleiff, Department of Biosciences, Molecular Cell Biology of Plants, Goethe University, Max von Laue Str. 9, Frankfurt am Main, 60438, Germany
schleiff@bio.uni-frankfurt.de

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *04 November 2014*
Accepted: *04 March 2015*
Published: *19 March 2015*

Citation:
Simm S, Keller M, Selymesi M and Schleiff E (2015) The composition of the global and feature specific cyanobacterial core-genomes. Front. Microbiol. 6:219. doi: 10.3389/fmicb.2015.00219

Introduction

Cyanobacteria are ancient, multifarious, photosynthetic prokaryotes. They are of biotechnological importance and are used for approaches to produce bioactive molecules, biofuels or other energy sources (Jones and Mayfield, 2012; Neilan et al., 2013; Wijffels et al., 2013; Oliver and Atsumi, 2014). In addition, cyanobacteria are considered as model organisms to study general aspects of bacteria

and cellular processes. In focus are the analysis of the function and evolution of photosynthetic systems (Shih et al., 2013; Croce and van Amerongen, 2014), nitrogen fixation (Bothe et al., 2010; Zehr, 2011), cell to cell communication (Flores and Herrero, 2010; Hahn and Schleiff, 2014), cell differentiation (Muro-Pastor and Hess, 2012), and cell wall function (Nicolaisen et al., 2009; Singh and Montgomery, 2011) to name just a few examples. However, most of the information was established for selected model cyanobacteria and still need to be generalized.

Aside from being of biotechnological importance, cyanobacteria are part of the phytoplankton (Sommer, 2005), but inhabit a diverse range of environments like rocks, lakes and deserts as well (e.g., Mur et al., 1999). It is estimated that all cyanobacteria on earth reach a total biomass of 10^{15} g (Garcia-Pichel et al., 2003), which marks these bacteria as an important component of ecosystems. Moreover, due to their high acclimation capacity in fluctuating environments, some cyanobacterial species are thought to show a higher adaptability to climate changes compared to other species. It is discussed that this can result in overgrowing other phytoplankton species within the communities (Carey et al., 2012; Elliott, 2012). The latter requires an efficient uptake of nutrients as well as efficient mechanisms to compete for trace elements. The uptake of solutes depends on outer membrane proteins (OMP; Mirus et al., 2010). Most OMPs are β-barrel proteins, which act in the recognition and transport of solutes, metabolites and proteins (e.g., Nicolaisen et al., 2009; Mirus et al., 2010). Such β-barrel proteins are characteristic for the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts (Sommer et al., 2011). While the transporters of the inner membrane were studied in some detail, not much, however, is known about the existence and function of the outer membrane β-barrel proteins of cyanobacteria (Hahn and Schleiff, 2014).

One measure to generalize the findings and to learn more about cyanobacteria is the pan- and core-genome determination. The pan-genome describes the entire gene set composed of all genes of all strains analyzed (Medini et al., 2005; Collingro et al., 2011). Therefore, it can be determined for an entire phylum like the cyanobacterial phylum (spelled in capitals below to emphasize that the entire phylum is analyzed: PAN-GENOME), or for a reduced set of organisms within the cyanobacterial phylum (spelled in small letters below to indicate that only a part of the PAN-GENOME is assigned: pan-genome). A pan-genome includes a core-genome, a dispensable-genome as well as unique genes (Reno et al., 2009). The dispensable-genome is the set of genes, which occurs in an intersection of at least two, but not all analyzed genomes. Unique genes are found in a single genome only. The core-genome includes those sets of genes that exist in each of the strains analyzed (Kettler et al., 2007). Again, we use capital letters (CORE-GENOME) in case the whole phylum is analyzed and small letters (core-genome) for the analysis of selected cyanobacteria only.

The selection of a subset of strains (clade) for core- and pan-genome analysis can be based on their phylogenetic positioning according to 16S rRNA sequence analysis (e.g., Valério et al., 2009) or traditional morphological features (e.g., Komárek and Anagnostidis, 1986, 1989; Anagnostidis and Komárek, 1987, 1990). In addition, classification of cyanobacteria with respect to their growth habitat offers the opportunity to determine feature-specific sets of genes. The prerequisite for this classification is the definition of morphological, biochemical and physiological features as well as of the typical growth habitat for each strain. Most of this information is deposited in the Integrated Microbial Genomes database (Markowitz et al., 2012). Based on this information, and refined by an exhaustive literature search, we classified the cyanobacterial strains according to 13 distinct features (**Table 1**, Additional File 1 in Supplementary Material).

Previous studies of gene sets have focused on the identification of intra-species gene sets needed to fully describe a species (Medini et al., 2005). The pan-genome analysis was developed as a consequence of the expanding number of sequenced genomes (Medini et al., 2005; Tettelin et al., 2008). Subsequently, this analysis was applied to study single genera like *Prochlorococcus* (Kettler et al., 2007), *Legionella* (D'Auria et al., 2010), or *Streptococcus* (Donati et al., 2010). Today, pan-genome analysis is used to define core-genomes for model organisms like human (Li et al., 2010) or yeast (Dunn et al., 2012). Similarly, core-genome definition of inter-species comparisons in a single phylum was used to gain information on sequence similarity (Tettelin et al., 2005), phylogenetic relations (Kettler et al., 2007) or evolutionary relations, as for example in *Chlamydiae* (Collingro et al., 2011) or cyanobacteria (Beck et al., 2012). Based on core-genome determination for a specific clade of species, the term "signature genes" has been introduced to denote genes with a limited phylogenetic distribution (Dutilh et al., 2008). Core-genome and signature gene definition was used to define a set of genes specific for cyanobacteria against eucaryotes containing chloroplasts (Martin

TABLE 1 | Phenotypical, ecological and physiological features analyzed.

	Feature	Sub-categories	CWI
1	Habitat	Sea/Ground/Fresh water/Salt meadow/Host/Water surface/Coast/Mud/Hot spring	56
2	Occurence	Lab/Nature	42
3	Nitrogen fixation	Yes or No	29
4	Toxin production and export	Yes or No	14
5	Trichome	Yes or No	52
6	Cell composition	Unicellular/Filament/Chain/Pairs	56
7	Cell shape	Spherical/Filamentous/Helical/Coccoid/Rod shaped/Oval	52
8	Heterocyst	Yes or No	54
9	Hormogonia	Yes or No	6
10	Akinete	Yes or No	7
11	Temperature range	Mesophilic/Thermophilic	56
12	Oxygen demand	Aerobic/Anaerobic/Facultative aerobic	47
13	Motility	Mobile/Immobile	51

Given is the number (column 1) and name of the feature analyzed (column 2), the categories of the feature (column 3), and the number of cyanobacteria with known information on the specific feature (CWI, column 4). Detailed information are given in Additional File 1 in Supplementary Material.

et al., 2003) or specific for the various clades of cyanobacteria (Gupta and Mathews, 2010). This approach has contributed to our knowledge on the origin of photosynthesis (Mulkidjanian et al., 2006) and diversity of metabolism (Beck et al., 2012).

Interestingly, pan- and core-genome analysis was not used to identify feature-specific gene sets yet. Therefore, we investigated gene sets for specific features based on 58 cyanobacterial genomes. We confirmed that the selected genomes are sufficient to define the cyanobacterial CORE-GENOME. In addition, for each genome we determined the genes part of the dispensable-genome and unique genes. Subsequently, cyanobacteria were clustered according to their sequence or feature similarities and we defined the pan- and core-genomes of different clades. This analysis yielded the identification of some genes specific for thermophilic cyanobacteria and for heterocyst forming cyanobacteria. To study the conservation and diversity of the outer membrane proteome, we developed a method for identification of genes coding for β-barrel proteins. The majority of OMPs identified in the PAN-GENOME is not present in the CORE-GENOME. The core-set of β-barrel OMPs in all 58 cyanobacteria is composed of only three proteins, while the majority of the β-barrel OMPs is strain-specific or shared by a small fraction of up to 15 cyanobacteria only. We conclude that the outer membrane proteome is largely adapted to the individual live style and environment of each cyanobacterial strain.

Materials and Methods

Ortholog Search and pan-Genome Construction

Literature and databases were searched for completely sequenced cyanobacterial genomes or assembled drafts. The respective literature is cited in the Section Introduction. Cyanobacterial nucleotide and protein sequences and other relevant information was taken from Cyanobase (Nakao et al., 2010) and the Integrated Microbial Genomes database of the Joint Genome Institute (Markowitz et al., 2012). The ORFs for each strain were categorized in known and hypothetical based on the deposited description. For the construction of the PAN- and CORE-GENOME, the dispensable-genome and the unique genes we used the complete proteomes of all 58 cyanobacteria. We used OrthoMCL (Chen et al., 2006) for prediction of CLiques of Orthologous Genes (CLOGs). OrthoMCL excluded poor-quality sequences with a length below 10 amino acids or a stop codon frequency higher than 20%. By this approach, all CLOGs containing at least two sequences were detected. Sequences not assigned to a cluster by OrthoMCL were subsequently determined as single-sequence clusters (CLOGs of unique genes).

CLOGs defined by OrthoMCL were evaluated by the Pan-Genome Analysis Pipeline (PGAP) to construct CLOGs of different orders containing more than one strain in their respective orthologous groups (Zhao et al., 2012). The PGAP implemented algorithm used (–method MP) is based on the combination of InParanoid and MultiParanoid (Ostlund et al., 2010). The input files of PGAP had to fulfill the following criteria: (i) a 3:1 relation between the CoDing Sequence (CDS) and protein sequence length had to exist to avoid wrongly annotated protein sequences; (ii) the same amount of CDS to protein sequences for each annotated gene was expected; (iii) the identifier had to be unique. In the end, pan-genomes for Nostocales, Prochlorales, Chroococcales, and Oscillatoriales were created using the parameters for clustering and pan-genome construction (–cluster; –pan-genome). For the PAN-GENOME assignment we used the results of OrthoMCL.

For confirmation of feature specific cyanobacterial signature genes we used all available genomes for Viridiplantae and bacteria (except cyanobacteria) available at NCBI non-redundant (nr) database. We used the sequences of the proteins found in *Thermosynechococcus elongatus* BP-1 (thermophile habitat) or *Anabaena* sp. PCC 7120 (soil living, heterocysts) to blast for similar sequences with at least 80% coverage of the bait sequence and an *e*-value of 1.0 e^{-10} or smaller.

To determine the putative function of each CLOG we assigned a functional classification to each sequence of the cyanobacteria (Tatusov et al., 1997) by the Bacterial Annotation System (BASys; van Domselaar et al., 2005) and the information from the WEB-server for Meta-Genome Analysis (WebMGA; Wu et al., 2011).

Construction of the Tanimoto-Like Index and Clustering

The Tanimoto-like index (e.g., Cooper et al., 1993) was used to transform the different features of the cyanobacteria (Additional File 1 in Supplementary Material) in a binary code (bit strings) and calculate the similarity and distance (the latter equals 1-similarity) between two cyanobacteria (Additional File 2 in Supplementary Material). The Tanimoto-like index consists of the sum of bit strings per feature. Each feature may contain more than one subcategory (e.g., habitat: sea, soil, freshwater, host, mud, hot spring, salt marsh) and the amount of subcategories determines the length of each feature bit string. Each subcategory was classified as present (1) or absent (0) based on literature (Additional File 1 in Supplementary Material). Features with no available information were classified as unknown (u). By comparison of two strains we determined whether the feature is (i) unknown in both strains, (ii) known in one strain or (iii) known in both strains. The first case was excluded from further calculations, whereas in the second case the denominator value was increased by 0.5. For the third case we added the sum of ones in the intersection to the numerator and the sum of ones in the union to the denominator (Additional File 2 in Supplementary Material).

Tree Construction

The Tanimoto-like index was used to calculate pair wise distances between strains based on 13 different features (Additional File 3 in Supplementary Material). The distance matrix was used to create the neighbor-joining feature tree (Additional File 4 in Supplementary Material). The CLOG distance neighbor-joining tree (Additional File 4 in Supplementary Material) was based on the CLOG distances (equals 1-similarity) between two strains. The CLOG similarity between two strains was calculated by dividing the number of all shared CLOGs by the number of CLOGs which contained at least one sequence of the two strains. Furthermore, 16S rRNA and average amino acid identity (AAI)

neighbor-joining trees were calculated (Additional File 5 in Supplementary Material). The 16S rRNA neighbor-joining tree was based on a multiple alignment via Multiple Alignment using Faster Fourier Transform (MAFFT; Katoh and Standley, 2013). The AAI neighbor-joining tree was built using the 420 CLOGs of the CORE-GENOME that contained one orthologous sequence per strain only. Pairwise global alignments between strains were calculated for each CLOG and the AAI over all CLOGs per pair of strains determined. Neighbor-joining trees were built with the molecular evolutionary genetics analysis package 6 (MEGA6; Tamura et al., 2013). The tree morphology was compared by calculating the patristic distance correlation (between 1 correlation and -1 anti-correlation) using the Mesquite software (Maddison and Maddison, 2011; http://mesquiteproject.org).

β-Barrel Protein Prediction and Clustering

The first step of Trans-Membrane Beta-barrel Prediction (TMBp) was based on the Beta-barrel Outer Membrane protein Predictor (BOMP; Berven et al., 2004), the K-Nearest Neighbor method based predictor (KNN; Hu and Yan, 2008) and the Trans-Membrane Beta-barrel Discriminator (TMBetaDisc; Ou et al., 2008) that are based on physicochemical features and the primary amino acid sequence. The TMBp approach was supported by a program established in our group (Mirus and Schleiff, 2005) in combination with TMHMM (Moller et al., 2001). Sequences detected as β-barrel proteins by more than one predictor were called *probable* β-barrel proteins.

The second step of β-barrel prediction was based on a Profile Hidden Markov Model (pHMM)-approach using the program HMMer (Eddy, 2011). We used the Protein Family (Pfam) database (Finn et al., 2014), OPM (Lomize et al., 2012), OMPdb (Tsirigos et al., 2011), which provide information on domain architecture and structures of β-barrel OMPs to build HMM profiles for each known β-barrel OMP family. These profiles were used to search for β-barrel OMPs in all cyanobacterial proteomes. Protein sequences with at least one detected β-barrel domain were considered as *probable* β-barrel OMP.

In the third step we defined two minor criteria. First, other domains than β-barrel OMP characterizing domains were identified by searching against the complete Pfam database (Finn et al., 2014). A protein was assigned to have the potential to be β-barrel OMP if an amino acid stretch longer than 79 amino acids was not characterized by such a Pfam domain. Secondly, we analyzed the CLOGs containing sequences representing β-barrel OMPs. If more than 50% of all sequences of a CLOG have been assigned as β-barrel OMP by TMBp and pHMM, the assigned proteins were considered as *detected*.

All proteins were subsequently classified (**Table 2**), namely in proteins detected by all four criteria [category (a)], proteins which fulfill the two main criteria and at least one minor criterion [category (b)], proteins which fulfill the two main criteria only [category (c)] and all other proteins [category (d)]. For all sequences of category (c) we performed *in silico* 3D structure analyzes with Phyre2 (Kelley and Sternberg, 2009). The results were manually inspected resulting in 37 putative β-barrel proteins [category (c); **Table 2**].

Results and Discussion

The General Composition of Cyanobacterial Genomes

Sequenced and annotated genomes of 58 cyanobacterial strains representing 45 species from six cyanobacterial orders were used to build the PAN-GENOME (**Table 3**). We used the amino acid sequences of the proteins encoded by all annotated genes present in the according genome and determined the CLiques of Orthologous Genes (CLOGs). CLOGs with sequences of only one cyanobacterial genome and genes not assigned to any CLOG were classified as "CLOGs of unique genes" for unification of the nomenclature. CLOGs with sequences from a certain set of strains (range from two to 57 strains) were annotated as "CLOGs of the dispensable-genome," and CLOGs with at least one sequence from each of the 58 strains as "CLOGs of the CORE-GENOME." We identified 44,831 CLOGs in total. 28,520 of all CLOGs are "CLOGs of unique genes" (**Figure 1A**). However, it needs to be mentioned that uncertain annotations of hypothetical ORFs can cause a high number of unique genes. Indeed, in Cya7, Cya6, Cya5, Cya4, Cya3, Cya2, Cya1, ProC, Tri1, Mic1, Cya8, Nod1, Glo1 genomes more than 50% of all genes are annotated as "hypothetical". The outcome of this is that 23,781 "CLOGs of unique genes" are "hypothetical" based on the protein sequence description. Moreover, 1725 of the "CLOGs of unique genes" contain two or more sequences from one strain representing putative paralogs. 15,752 are CLOGs of the dispensable-genome, but most of these CLOGs contain only sequences from

TABLE 2 | β-Barrel probability categorization.

Category	Major criteria		Minor criteria		*Anabaena* sp. *PCC 7120*	*All cyanobacteria*
	TMBp	pHMM	Pfam	CLOGs		
(a)	Probable	Probable	Potential	Detected	39	703
(b)	Probable	Probable	One of the two criteria		7	179
(c)	Probable	Probable	–	–	4[a]/0[b]	78[a]/37[b]
(d)	Others				6089	228,326

Shown is the category of the β-barrel prediction (column 1), the major criteria based on TMBp (column 2) and pHMM (column 3) analysis, the minor criteria based on Pfam search for non-β-barrel domains (column 4) or analysis of the CLOG composition (column 4); the number of identified genes in Anabaena sp. PCC 7120 (column 6) or in all cyanobacteria (column 7). [a]before and [b]after structural prediction by Phyre2 and manual inspection.

TABLE 3 | Classification and genome size of the analyzed 58 cyanobacterial strains.

Order	Species	Strain	Abbr.	Size (Mb)	ORFs	Put. ORFs (%)
Chroococcales	Acaryochloris marina	*Acaryochloris marina* MBIC11017	Aca1	8.36	8383	52.75
	Crocosphaera watsonii	*Crocosphaera watsonii* WH 8501	Cro1	6.24	5958	44.56
	Cyanothece sp. ATCC 51142		Cya1	5.46	5304	56.73
	Cyanothece sp. PCC 7424		Cya2	6.55	5710	36.18
	Cyanothece sp. PCC 7425		Cya3	5.79	5327	33.40
	Cyanothece sp. PCC 8801		Cya4	4.79	4367	29.86
	Cyanothece sp. ATCC 51472		Cya5	5.43	5109	31.45
	Cyanothece sp. CCY 0110		Cya6	5.88	6475	61.64
	Cyanothece sp. PCC 7822		Cya7	7.84	6981	46.48
	Cyanothece sp. PCC 8802		Cya8	4.80	4648	34.47
	Cyanobium sp. PCC 7001		Cyn1	2.83	2771	32.52
	Microcystis aeruginosa	*Microcystis aeruginosa* NIES-843	Mic1	5.84	6311	53.40
	Synechococcus elongatus	*Synechococcus elongates* PCC 6301	Syn2	2.70	2525	44.40
		Synechococcus elongates PCC 7942	Syn7	2.74	2662	38.92
	Synechocystis sp. PCC 6803		Syc1	3.95	3672	50.03
	Synechococcus sp. WH 8102		Syn1	2.43	2526	46.00
	Synechococcus sp. CC9311		Syn3	2.61	2892	38.00
	Synechococcus sp. PCC 7002		Syn4	3.41	3186	31.17
	Synechococcus sp. CC9902		Syn5	2.23	2304	39.67
	Synechococcus sp. CC9605		Syn6	2.51	2638	45.94
	Synechococcus sp. JA-2-3B	*Synechococcus* sp. JA-2-3B'a(2–13)	Syn8	3.05	2862	32.29
	Synechococcus sp. JA-3-3Ab		Syn9	2.93	2760	31.88
	Synechococcus sp. RCC307		SynA	2.22	2535	36.25
	Synechococcus WH7803		SynB	2.37	2533	33.48
	Synechococcus sp. BL107		SynC	2.29	2507	44.28
	Synechococcus sp. CB0205		SynD	2.43	2719	42.18
	Synechococcus sp. PCC 7335		SynF	5.97	5586	45.95
	Synechococcus sp. WH 7805		SynG	2.63	2883	49.60
	Synechococcus sp. WH 8016		SynH	2.69	2990	35.55
	Synechococcus sp. WH 8109		SynI	2.12	2577	39.74
	Thermosynechococcus elongatus	*Thermosynechococcus elongatus* BP-1	The1	2.59	2476	42.37
Gloeobacterales	*Gloeobacter violaceus*	*Gloeobacter violaceus* PCC 7421	Glo1	4.66	4431	57.98
Nostocales	*Anabaena* sp. PCC 7120		Ana1	7.21	6135	56.95
	Anabaena variabilis	*Anabaena variabilis* ATCC 29413	Ana2	7.11	5661	34.98
	Nodularia spumigena	*Nodularia spumigena* CCY9414	Nod1	5.32	4860	50.41
	Trichormus azollae	*Nostoc azollae* 0708	Nos2	5.49	5321	60.42
	Nostoc punctiforme	*Nostoc punctiforme* PCC 73102	Nos3	9.06	6690	39.07
Oscillatoriales	*Lyngbya* sp. CCY 8106		Lyn1	7.04	6142	53.61
	Coleofasciculus chthonoplastes	*Microcoleus chthonoplastes* PCC 7420	Mil1	8.68	8294	57.14
	Arthrospira platensis	*Arthrospira platensis* NIES-39	Art1	6.79	6630	61.70
	Arthrospira maxima	*Arthrospira maxima* CS-328	Art2	6.00	5690	36.50
	Arthrospira sp. PCC 8005		Art3	6.17	5675	46.70
	Oscillatoria sp. PCC 6506		Osc1	6.68	5822	53.98
	Trichodesmium erythraeum	*Trichodesmium erythraeum* IMS101	Tri1	7.75	4451	39.00
Prochlorales	*Prochlorococcus marinus*	*Prochlorococcus marinus* SS120	Pro1	1.75	1882	27.52
		Prochlorococcus marinus MED4	Pro2	1.66	1713	29.83
		Prochlorococcus marinus MIT 9313	Pro3	2.41	2267	32.91
		Prochlorococcus marinus str. NATL2A	Pro4	1.84	2163	40.41

(Continued)

TABLE 3 | Continued

Order	Species	Strain	Abbr.	Size (Mb)	ORFs	Put. ORFs (%)
		Prochlorococcus marinus str. MIT 9312	Pro5	1.71	1962	35.68
		Prochlorococcus marinus str. AS9601	Pro6	1.67	1921	35.97
		Prochlorococcus marinus str. MIT 9515	Pro7	1.70	1906	36.41
		Prochlorococcus marinus str. MIT 9303	Pro8	2.68	2997	50.75
		Prochlorococcus marinus str. NATL1A	Pro9	1.86	2193	46.69
		Prochlorococcus marinus str. MIT 9301	ProA	1.64	1907	35.19
		Prochlorococcus marinus str. MIT 9215	ProB	1.74	1983	37.17
		Prochlorococcus marinus str. MIT 9211	ProC	1.69	1855	37.20
		Prochlorococcus marinus str. MIT 9202	ProF	1.69	1890	33.17
Stigonematales	*Fischerella* sp. JSC-11		Fis1	5.38	4627	27.34

Given is the order (column 1), the species according to NCBI and PATRIC taxonomy (column 2; Wattam et al., 2014) and the strain if not identical with the species (column 3) for each cyanobacteria included in this study. Column 4 gives the abbreviation used in here, column 5 gives the genome size of both, chromosomes and plasmids in megabases (Mb) and column 6 gives the number of protein coding open reading frames (ORFs) on the chromosomes and plasmids. Column 7 gives the percentage of the ORFs only annotated as putative/hypothetical.

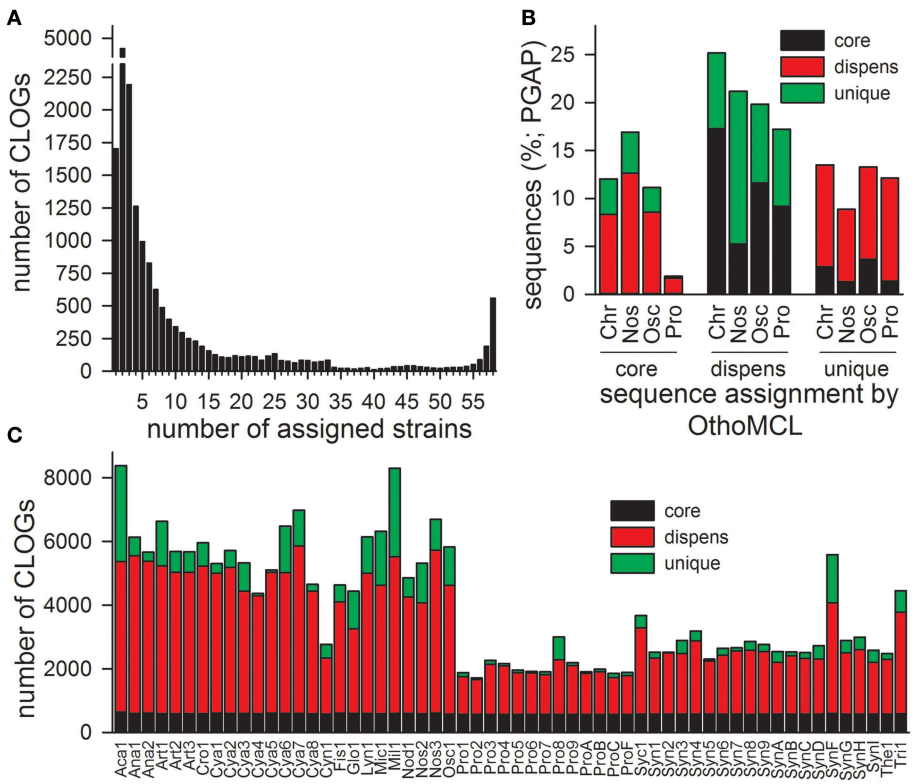

FIGURE 1 | CLOG distribution of the 58 cyanobacteria. (A) The numbers cyanobacterial strains from which sequences are included in an individual CLOG was determined. The number of CLOGs containing genes from a given number of strains is shown. **(B)** CLOGs representing unique, dispensable (dispens) or CORE-genes (core) were determined by OrthoMCL for all genomes or by PGAP for the genome of the cyanobacterial order Chroococcales, Nostocales, Oscillatoriales, and Prochlorales. Shown is the frequency of assignment of genes of a certain CLOG category detected by OrthoMCL to another CLOG category by PGAP (CORE-gene: black; dispensable gene: red; unique gene: green). **(C)** Shown is the number of sequences of the individual cyanobacteria represented by a CLOG of the CORE-GENOME (black), by a CLOG of the dispensable-genome (red; dispens), and by a CLOG of unique genes (green).

up to 10 strains (**Figure 1A**). Finally, 559 CLOGs of the CORE-GENOME (Additional File 6 in Supplementary Material) were identified as they contain sequences of all 58 cyanobacterial strains (**Figure 1A**). This is consistent with the earlier postulation that the CORE-GENOME of cyanobacteria has a size of 500–600 genes (Beck et al., 2012).

The distribution of the sequences in the different CLOG categories is by large comparable to the results of the PGAP

analysis, which created individual pan-genomes of different cyanobacterial orders (**Figure 1B**, Zhao et al., 2012). The discrepancy of about 10% observed by the two approaches is expected, because for CLOG definition by OrthoMCL all genomes were analyzed, while due to computational limitations for the PGAP analysis only the genomes of strains of one order could be used.

With respect to the strains we realized that the majority of the genes of each individual strain was assigned to CLOGs of the dispensable-genome (**Figure 1C**; red). The total number of genes identified in CLOGs of unique genes varies between the different strains (**Figure 1C**; green) and is primarily related to the genome size (**Table 1**). This is expected, because smaller genomes generally code for a lower number of proteins (**Table 3**) and thus, the portion of the genes found in CLOGs of the CORE-GENOME and of the dispensable-genome is larger. However, this rule does not apply to *Prochlorococcus marinus* strain MIT 9303 (Pro8). Nevertheless, the strain MIT 9303 has the largest genome with most annotated ORFs of all *P. marinus* strains, which might explain the larger portion of unique genes. The "additional" genes in *P. marinus* str. MIT 9303 by large encode proteins with putative functions in membrane synthesis and transport (Kettler et al., 2007), which might hint to specific features of this strain when compared to other strains of *P. marinus*.

Further, exceptions from the rule are *Cyanothece* sp. PCC 8801 (Cya4), *Cyanothece* sp. ATCC 51472 (Cya5) and *Cyanothece* sp. PCC 8802 (Cya8), which have the smallest genome as well as assigned proteome of all *Cyanothece* species (**Table 3**). These three species show a large content of genes assigned either to the CORE-GENOME or the dispensable-genome, but a small content of unique genes when compared to other *Cyanothece* species. Thus, the genome of these three strains might be composed of genes for the basic functions of *Cyanothece* only.

The Size of the Cyanobacterial Core- and PAN-Genome

Based on the analysis of the 58 cyanobacterial strains a CORE-GENOME size of 559 genes was observed. To judge whether the 45 species represented by the 58 strains are sufficient to define the CORE-GENOME of cyanobacteria, we determined the CORE-GENOME size dependence on the number of genomes analyzed. We determined the size of the core-genome for a given number of randomly selected genomes from the 58 organisms. The random selection was 1000 times repeated and the average calculated (**Figure 2A**). The number of sequences found in the core-genome changed only little when more than 40 cyanobacterial strains were considered. The result was not dependent on number of repetitions, as for only 100 or even 10,000 random selections the same result was observed (Additional File 7 in Supplementary Material).

The robustness of our result prompted us to compare the CORE-GENOME determined in here with the CORE-GENOMES defined earlier analyzing eight (Martin et al., 2003; 179 CORE-GENES asssigned), 15 (Mulkidjanian et al., 2006; 1044 CORE-GENES asssigned) or 16 cyanobacterial genomes (Beck et al., 2012; 704 CORE-GENES asssigned). The overlap between previously assigned CORE-GENOMES and the one defined in here consists of 520 and 526 sequences for the two larger studies, respectively. On the one hand, this shows that almost all genes of the CORE-GENOME identified in here are present in the previous CORE-GENOME sets, on the other hand it documents that the low number was not sufficient, which is consistent with our simulation (**Figure 2A**). Both conclusions support the notion that the CORE-GENOME of cyanobacteria most likely covers about 500 genes.

We determined the functional categories based on the sequences of *Anabaena* sp. PCC 7120 for the CORE-GENOME. Here we used the functional annotation previously established for clusters of orthologous groups (COG) for seven complete genomes from five major phylogenetic lineages (Tatusov et al., 1997). In part, the result was manually compared to the KEGG annotations (Kanehisa and Goto, 2000). We realized that proteins encoded by 231 sequences of the CORE-GENOME (representing ~40%) are involved in metabolic processes in *Anabaena* sp. PCC 7120 (**Table 4**). Thereof, 59 proteins are assigned to be

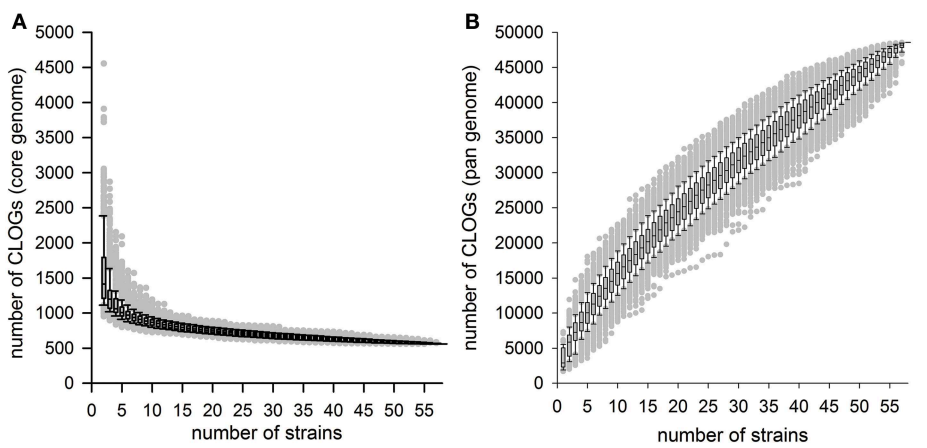

FIGURE 2 | The cyanobacterial core- and pan-genome. (A,B) The number of CLOGs of the cyanobacterial **(A)** core-genome or **(B)** pan-genome for a given number of organisms is shown. The box plots were created for the results of 1000 different random selections of different cyanobacterial strains. Further simulations are shown in Additional File 6 in Supplementary Material.

TABLE 4 | Functional categories and processes according to COG.

Functional category	Functional process	Abbr.	CORE CLOGs
Information storage and processing	Translation, ribosomal structure and biogenesis	J	90
	Transcription	K	11 (3)*
	Replication, recombination and repair	L	37 (3)
	TOTAL		141
Cellular processes and signaling	Cell cycle control, cell division, chromosome partitioning	D	11
	Defense mechanisms	V	1
	Signal transduction mechanisms	T	8
	Cell wall/membrane/envelope biogenesis	M	27
	Cell motility	N	–
	Intracellular trafficking, secretion, and vesicular transport	U	10 (1)
	Posttranslational modification, protein turnover, chaperons	O	40 (1)
	TOTAL		103
Metabolism	Energy production and conversion	C	45 (2)
	Carbohydrate transport and metabolism	G	22 (1)
	Amino acid transport and metabolism	E	49 (10)
	Nucleotide transport and metabolism	F	23 (4)
	Coenzyme transport and metabolism	H	46 (6)
	Lipid transport and metabolism	I	15 (3)
	Inorganic ion transport and metabolism	P	13 (2)
	Secondary metabolites biosynthesis, transport and catabolism	Q	3 (2)
	TOTAL		213
Poorly characterized	General function prediction only	R	35
	Function unknown	S	77
	mixed process**	X	17
	TOTAL		129

Given is the global functional category (column 1), the functional process (column 2), the one letter code for the functional process (column 3) and number of proteins per functional assignment of all proteins encoded by the CORE-GENOME of Anabaena sp. PCC 7120. The CLOG annotation is exemplarily for "Energy production and conversion" to the KEGG annotation (Additional File 7 in Supplementary Material).

**The number of proteins in the bracket is the count of proteins assigned to two process (e.g., translation, ribosomal structure and biogenesis and transcription), and the protein is counted for each of the processes.*

***The number proteins assigned to more than two process.*

involved in amino acid transport and metabolism (category E), 52 as coenzyme transport and metabolism (category H) and 47 in energy production and conversion (category C). The observation that not all components of the photosystems are encoded by the CORE-GENOME was confirmed by the analysis of the distribution of the proteins involved in oxidative phosphorylation, photosynthesis and antenna proteins annotated by KEGG (Additional File 8 in Supplementary Material). In addition, 90 proteins coded by the CORE-GENOME genes in *Anabaena* sp. PCC 7120 are assigned to be involved in translation, ribosomal structure and biogenesis (category J), while 41 encoded proteins function in posttranslational modification, protein turnover and chaperones and 40 in replication, recombination and repair (**Table 4**).

Next, we investigated the PAN-GENOME formed by the 44,831 CLOGs observed for the 58 strains defined. Again, we randomly selected the genes of a given number of strains for the determination of the pan-genome and this random selection was repeated 100, 1000, and 10,000 times (**Figure 2B**; Additional File 7 in Supplementary Material). As for the core-genome analysis, the result was not dependent on the number of random selections used in here. Previously it was postulated that increase of the PAN-GENOME follows the power law with respect to number of genomes included (Tettelin et al., 2008; **Figure 2B**). For *P. marinus* it was reported that addition of new strains into the analysis would always yield an increase of the pan-genome size (a so called "open pangenome"), however with a low rate (the according factor is $\alpha = 0.80$ suggesting a low increase of the PAN-GENOME size by addition of the genomic information of an additional strain; Tettelin et al., 2008). For all cyanobacteria we obtained an α of 0.35 ± 0.07. This suggests that the PAN-GENOME of all cyanobacteria is

(i) a so called open PAN-GENOME and increases with addition of new cyanobacterial strains, because only for $\alpha > 1$ a limit exists, and (ii) the PAN-GENOME of all cyanobacteria increases more rapidly by addition of new genomes as the pan-genome for a single species of cyanobacteria like *P. marinus*.

Habitat Specific Cyanobacterial Proteins

We gathered information about ecological, morphological and physiological features for all analyzed strains from the Integrated Microbial Genomes database of the Joint Genome Institute (Markowitz et al., 2012) and from selected publications (Additional File 1 in Supplementary Material; Huber, 1985; Stal and Krumbein, 1985; Jones, 1992; Cohen et al., 1994; Rouhiainen et al., 1995; Kaneko and Tabata, 1997; Gruber and Bryant, 1998; Nakamura et al., 2002; Zhou and Wolk, 2002; El-Shehawy et al., 2003; Lesser, 2003; Urmeneta et al., 2003; Tuit et al., 2004; Araoz et al., 2005; Allewalt et al., 2006; Dworkin et al., 2006; Su et al., 2006; Takaichi et al., 2006; Gao et al., 2007; Kaneko et al., 2007; Kettler et al., 2007; Kim et al., 2007; Campbell et al., 2008; Stockel et al., 2008; Swingley et al., 2008; Bolhuis et al., 2010; Fujisawa et al., 2010; Mejean et al., 2010; Ran et al., 2010; Scott et al., 2010; Carrieri et al., 2011; Larsson et al., 2011; Ploug et al., 2011; Markowitz et al., 2012; Nguyen et al., 2012; Stewart et al., 2012) and extracted 13 different features (**Table 1**, Additional file 1 in Supplementary Material). In some cases information was logically assumed. For example, unicellular organisms should not contain features characterizing multicellular cyanobacteria like heterocysts, akinetes or hormogonia.

Next, we determined genes specific for a subset of cyanobacterial strains with either thermophilic character, with common growth environment or the capability to differentiate heterocysts, because for the remaining features the assignment for the cyanobacteria is laregely incomplete (Additional file 1 in Supplementary Material). For the identification of such genes we searched for CLOGs containing exclusively sequences of cyanobacterial strains with a certain feature. Subsequently, only the CLOGs of the latter pool with sequences of all cyanobacterial strains with this feature were selected. In our set of organisms we had three thermophilic cyanobacteria, namely *T. elongatus* BP-1 (The1), *Synechococcus* sp. JA-3-3Ab (Syn9), and *Synechococcus* sp. JA-2-3B'a(2–13) (Syn8). We obtained four CLOGs with genes of these three strains only. In *T. elongatus* BP-1 these genes are tlr0324, tlr0548, tlr0547, and tsr0549 (Nakamura et al., 2002, 2003). The protein tlr0324 putatively contains a DNAJ-domain and is predicted to be a Heat shock protein (HSP), while the proteins encoded by the second gene cluster, tlr0548, tlr0547, and tsr0549, are of unknown function. Next we analyzed whether the identified genes are specific to cyanobacteria by searching for similar sequences in plants and bacteria (see Materials and Methods). Sequences with similarity to tlr0548 have been identified in bacterial strains with extreme habitats of the genus Acidithiobacillus (5) and the species *Haliangium ochraceum* (1), *Halothiobacillus neapolitanus* (1), *Sorangium cellulosum* (2), or *Thiothrix nivea* (1), but not in plants. In turn, we did not identify sequences with similarity to tlr0324, tlr0547, and tsr0549 in the bacterial or plant genomes by the approach applied (see Materials and Methods). Thus, these three genes likely represent "signature genes" for thermophilic cyanobacteria.

With respect to the growth habitat we obtained 34 cyanobacterial strains assigned to live in salt water, 15 in fresh water, three in fresh water as well as on soil, three require a host organism, one is exclusively soil-living and one occurrs in both salt and fresh water (Additional File 1 in Supplementary Material). However, we did not find a CLOG including sequences of all cyanobacteria growing in salt or fresh water. The same holds true for the three host-living cyanobacteria. Thus, either a habitat-specific core-genome does not exist with respect to salt/fresh water and host-living strains, or for some of the strains the assignment found in literature is incomplete.

Five CLOGs for the cyanobacterial strains assigned as capable of soil-living (*Anabaena* sp. PCC 7120, *Anabaena variabilis* ATCC 29413, *Gloeobacter violaceus* PCC 7421, *Nostoc punctiforme* PCC 73102) were identified. We again aimed for confirmation of the specificity of the identified genes for cyanobacteria. However, similar sequences to the identified oxidoreductase (encoded by all0827 in *Anabaena* sp. PCC 7120) was found in many other plant and bacterial genomes. Similarily, sequences with similarity to the protein with similarity to acetyltransferases (encoded by alr3061), the membrane-spanning subunit DevC of the heterocyst-specific ABC transporter (encoded by alr4974) and the six-bladed β-propeller TolB-like domain containing protein (encoded by all0351) were identified in many bacterial genomes. Only for the protein of unknown function encoded by alr7204 sequences with similarity could not be identified in the analyzed eucaryotic or prokaryotic genomes. Summing up, we propose the existence of at least three signature genes for thermophilic and one signature gene for soil-living cyanobacteria, while we could not identify signature genes for salt or fresh water living cyanobacteria.

Heterocyst Specific Cyanobacterial Proteins

We aimed for the detection of CLOGs unifying heterocyst-forming cyanobacteria. In our set six cyanobacteria are assigned as heterocyst-forming (Additional File 1 in Supplementary Material; *Anabaena* sp. PCC 7120, *Anabaena variabilis* ATCC 29413; *Fischerella* sp. JSC-11; *Nodularia spumigena* CCY9414; *Nostoc azollae* 0708; *Nostoc punctiforme* PCC 73102), while for four cyanobacteria information was not available (*Oscillatoria* sp. PCC 6506, *Synechococcus* sp. WH 8016; *Synechococcus elongates* PCC 6301; *Synechococcus* sp. CB0205). To judge whether we have to include the latter four as heterocyst forming, we inspected CLOGs containing genes known to be essential for heterocyst differentiation, but not related to global families like the ABC transporters. We selected 17 of such genes (**Figure 3A**). Sequences of all confirmed heterocyst-forming cyanobacteria (Additional File 1 in Supplementary Material; Ana1, Ana2, Fis1, Nod1, Nos2, Nos3) are in 14 of the 17 CLOGs formed by the selected heterocyst marker genes (**Figure 3B**, red bars). Only PatS (asl2301, *Anabaena* sp. PCC 7120), HetN (alr5358, *Anabaena* sp. PCC 7120), and PbpB (alr5101, *Anabaena* sp. PCC 7120) could not be detected in all strains by the method applied.

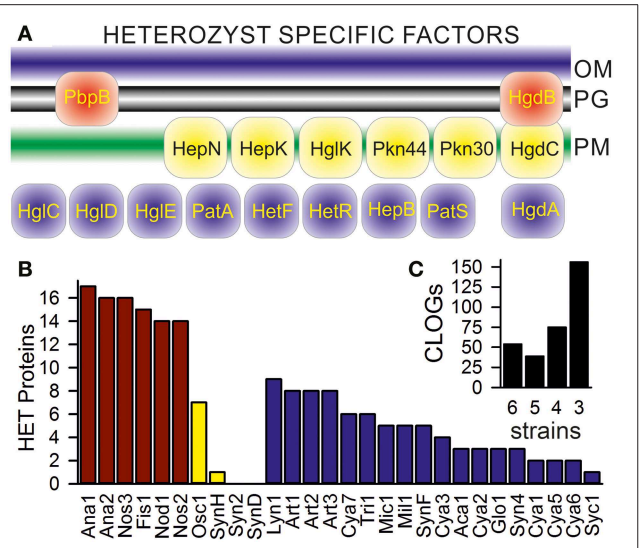

FIGURE 3 | CLOGs of genes involved in heterocyst formation. (A) A scheme of the localization of the 17 selected heterocyst specific proteins is shown: the penicillin-binding protein 2 (alr5101 in *Anabaena* PCC 7120, PbpB; Lazaro et al., 2001), the pentapeptide-repeat protein HglK (all0813; Black et al., 1995), the oxidoreductase HgdA (all5345; Nicolaisen et al., 2009), the glycolipid deposition proteins HgdB and HgdC (all5347 and all5346; Nicolaisen et al., 2009), the HstK family proteins with two-component sensor domain Pkn44 and Pkn30 (all1625 and all3691; Shi et al., 2007), the sensory protein-histidine kinase of a two-component regulatory system (all4496; HepK; Golden and Yoon, 2003), the ketoreductase HetN (alr5358; Corrales-Guerrero et al., 2014), the heterocyst differentiation control proteins HetR (alr2339; Du et al., 2012) and HetF (alr3546; Ionescu et al., 2010), the poly-peptides controlling heterocyst pattern formation PatA (all0521; Zhang et al., 2007) and PatS (asl2301; Nicolaisen et al., 2009), the heterocyst envelope polysaccharide synthesis factor HepB (alr3698; Wang et al., 2007) and the heterocyst glycolipid synthases HglC, HglD, and HglE (alr5355, alr5354, and alr5351; Fan et al., 2005). **(B)** CLOGs including *Anabaena* sp. PCC 7120 sequences mentioned in the text have been analyzed concerning the cyanobacteria the sequences originated from. The number of detected proteins known to be involved in heterocyst formation/function in cyanobacteria known to form heterocysts (red), not to form heterocysts (yellow) or for which information about heterocysts formation is not available (blue) is shown. **(C)** The inset on the right shows the number of CLOGs with the sequences of the six heterocyst-forming cyanobacteria only.

Nine CLOGs of genes known to be essential to heterocyst differentiation contain sequences of the filamentous *Lyngbya* sp. CCY 8106; and eight CLOGs contain sequences of each of the *Arthrospira* strains, though for these cyanobacteria heterocyst formation is not reported (**Figure 3B**, blue bars, Additional File 1 in Supplementary Material). These eight CLOGs represent PbpB, HglK, HgdA, HetR, HetF, Pkn44, Pkn30, and HepK. The meaning of this observation needs to be explored in future.

Of the four strains with unknown assignment to heterocyst formation, sequences of the filamentous *Oscillatoria* sp. PCC 6506 are present in seven of the 17 CLOGs of the selected heterocyst specific genes (**Figure 3B**, yellow bar). As expected sequences of the three most likely unicelluar strains (Syn2, SynD, SynH) are detectable in at most one of the 17 CLOGs (**Figure 3B**, yellow bar). Consequently, from our inspection of the distribution of genes specific for heterocysts we conclude that only the six confirmed heterocyst forming cyanobacteria should be included in the analysis of the core-genome of genes specific for heterocyst forming cyanobacteria.

At first we identified all CLOGs with sequences from the six heterocyst-forming strains only. We observed 54 CLOGs with sequences from all six strains, 39 with sequences from five, 75 from four and 156 from three heterocyst-forming cyanobacteria (**Figure 3C**). The number of CLOGs with sequences of only five strains prompted us to consider the 93 genes of the CLOGS containing sequences of at least five of the six strains as core-genome of heterocyst-forming cyanobacteria (**Tables 5**, **6**). Fourteen of these 93 genes have been experimentally characterized and for 28 a function can be predicted (**Table 5**), while for 51 genes a function is not assigned (**Table 6**). Eight of the 93 genes were shown to be exclusively expressed upon nitrogen starvation in *Anabaena* PCC 7120, while another 48 genes are at least two-fold higher expressed either 12 or 21 h after nitrogen stepdown (**Tables 5**, **6**, Flaherty et al., 2011). In turn, only one gene is not expressed in *Anabaena* PCC 7120 after nitrogen starvation (asl1933) and one is significantly downregulated (asr1289; **Table 5**, Flaherty et al., 2011).

We inspected whether the genes identified are heterocyst specific signature genes of cyanobacteria. We realized that six of the experimentally characterized genes and eight genes with putative function are indeed specific for cyanobacteria based on our criteria (see Materials and Methods), because sequences with similarity could not be identified in the analyzed plants and bacteria (**Table 5**). In addition, for four proteins encoded by the genes identified in the CLOGs formed by heterocyst forming cynobacteria only one other bacterial strain containing a similar sequence was detected (**Table 5**). In addition, for 44 of the not yet characterized factors similar sequences could not be detected in any of the analyzed genomes, while for additional four only one or two sequences with similarity could be detected (**Table 6**). We therefore propose that eight of the identified genes are highly specific for heterocyst forming cyanobacteria, while 58 represent heterocyst specific cyanobacterial signature genes. It is worth mentioning, the majority thereof have not yet been characterized.

The Composition of the Core-Genomes of the Different Clades of Cyanobacteria

We calculated a Tanimoto-like index for each pair of cyanobacteria (see Materials and Methods, Additional File 2 in Supplementary Material), which at first transfers the obtained information on cyanobacterial features into a binary code and subsequently determines the similarity of two strains. These indices were used to group the strains (Additional File 3 in Supplementary Material) and to calculate a neighbor-joining tree (**Figure 4B**, Additional File 4 in Supplementary Material). In parallel, we used the determined CLOGs to calculate the difference between two cyanobacterial strains and used this "pairwise CLOG distance" for calculation of a second neighbor-joining tree (**Figure 4A**, Additional File 4 in Supplementary Material).

By large, the two trees show a comparable branching (patristic distance correlation coefficient: 0.51). This suggests a correlation between the proteome setup and the analyzed cyanobacterial features. For further verification we compared the CLOG

TABLE 5 | Genes with known or putative function in heterocyst-specific CLOGs.

Acc. Number	Name	Function	FC		CA	V/B	References
			12 h	21 h			
all0521	PatA	Heterocyst formation regulating two-component response regulator	1,6	1,4		0/0	Liang et al., 1992
all1866	TrxA2	Thioredoxin A2	2,8	3,7	Fis1	391/499	Ehira and Ohmori, 2012
all2356	PhnE	Phosphonate ABC transport permease	5,9	6,1	Nos2	0/490	Pernil et al., 2010
alr2392	FraC/SepJ	Filament integrity protein	−1,7	1,9		0/0	Bauer et al., 1995
alr2834	HepC	Glycosyl transferase	47,3	19,2		0/0	Zhu et al., 1998
alr2837		Glycosyl transferase of group 2	Up	up		0/27	Huang et al., 2005
alr3234		Similar to heterocyst formation protein HetP	−1,2	−1,3	Fis1	0/0	Higa and Callahan, 2010
alr3287	NrtB	Nitrate transport protein	1,1	1,9	Nod1	0/479	Herrero et al., 2001
alr3732	PknE	Protein serine-threonine kinase	3,8	1,2		0/0	Zhang et al., 1998
alr4368	PknD	Serine/threonine kinase	3,0	1,4		0/0	Zhang and Libs, 1998
all5341	HglT	Glycosyl transferase of group 1	up	up		48/485	Awai and Wolk, 2007
all5344		Unknown	not	up		0/141	Fan et al., 2005
all5346	HgdC	Membrane spanning subunit of heterocyst specific ABC-transporter	not	34,6		0/85	Fan et al., 2005
all5347	HgdB	Membrane fusion protein of heterocyst specific ABC-transporter	2,3	115,8		0/62	Fan et al., 2005
all0059		Lipopolysaccharide biosynthesis protein	53,6	19,2		0/71	None
all1345		Probable glycosyl transferase	−1,2	−1,3		0/185	None
all1862		Putative peptidase	22,2	9,6	Fis1	0/0	None
all2008		Serine proteinase	1,2	1,2		6/198	None
all2068		Alpha/beta hydrolase fold protein	1,3	1,0		59/482	None
all2357		Phosphonate ABC transport ATP-binding component	4,9	3,3	Nos2	485/497	None
all2358		Periplasmic phosphonate binding protein	6,3	2,9		0/148	None
alr2463		Aminoglycoside phosphotransferase	9,8	3,6		0/1[a]	None
alr3125		Heme oxygenase	−2,5	2,4	Nod1	0/385	None
alr3235	TrpC	Indole-3-glycerol phosphate synthase	up	up	Fis1	89/498	None
alr3246		Pyridoxamine 5′ phosphate oxidase Related protein	up	up	Fis1	0/429	None
all3306		Pentapeptide repeat containing protein	up	up	Fis1	0/21	None
all3559		Putative peptidase	−1,7	1,5	Nod1	0/0	None
alr3774		Rhomboid like protein	3,5	2,4		0/419	None
alr3931		Rhomboid family protein	1,1	−1,0	Nos2	0/485	None
alr3948	CbiQ	Cobalt transport protein	6,8	4,2		0/1[b]	None
all3984		Predicted ATP-dependent protease	2,1	1,0		0/0	None
all4051		Prc barrel domain containing protein	2,3	2,7		0/30	None
all4538		Mannose-6-phosphate isomerase	1,5	−1,2		0/107	None
all4729		Putative metalloprotein	−1,0	100,8		0/1[c]	None
asl4754	PetM	Cytochrome b6f complex subunit	−2,5	−1,8		0/0	None
all4768		ErfK/YbiS/YcfS/YnhG family protein	2,7	7,5	Nod1	0/11	None
alr4812	PatN	Heterocyst differentiation related protein	1,3	1,4	Fis1	0/0	None
alr4877		WD40-repeat protein	2,5	2,7	Nod1	0/0	None
alr4898		Transcriptional regulator	2,1	1,6	Fis1	3/90	None
alr4984		Peptidoglycan binding domain 1 containing protein	25,4	5,7		0/1[d]	None
asr5289		Similar to subunit X of photosystem I	1,2	1,0		0/0	None
all5304		Secretion protein HlyD family protein	6,0	3,2		0/491	None
ava0606		Transmembrane protein	not	not	Ana1	0/0	None

Shown is the accession number of Anabaena sp. PCC 7120 or Anabaena variables ATCC 29413; column 1, the name and function of the gene if assigned (column 2, 3), the fold change (FC) of expression after 12 and 21 h of nitrogen starvation compared to 0 h (Flaherty et al., 2011; column 4, 5; up, infinite; not, not expressed), the cyanobacteria for which no sequence is identified in the according CLOG (CA, column 6), the number of sequences found in the genomes of Viridplantae or bacteria (V/B, column 7) and a references for functional relevance for heterocyst function or development (column 8).

[a] *Candidatus Solibacter usitatus.*

[b] *Thalassospira profundimaris.*

[c] *Rhodopseudomonas palustris.*

[d] *Paenibacillus mucilaginosus.*

TABLE 6 | Genes of unknown function in heterocyst-specific CLOGs.

Acc. number	Fold change		Cyanob. absent	V/B
	12 h	21 h		
asl0176	1,9	4,8		0/0
alr0255	8,5	4,9		0/0
all0307	5,8	3,0	Fis1	0/0
asr0460	1,6	not	Nos2	0/0
asr0461	−1,0	−1,9	Nos2	0/0
all0463	7,7	10,6	Nos2	0/0
asr0680	−1,6	−1,9	Fis1	0/19
alr0805	1,4	1,2		2/0[a]
asl0842	−1,5	−1,6	Fis1	0/0
all0997	−4,9	−1,8	Fis1	0/0
alr1137	−1,6	−2,7		0/0
alr1146	9,1	5,3		0/1[b]
alr1147	2,5	1,8	Nos2	0/2[c]
alr1148	8,7	7,7		0/0
asr1289	−2,7	−2,7	Fis1	0/0
all1395	up	up	Nos2	0/0
asl1412	3,6	3,4		0/0
asr1775	1,9	2,2	Nos2	0/0
all1814	15,5	5,9		0/0
asl1933	not	not	Fis1	0/0
all2003	4,0	1,8		0/1[d]
all2089	1,8	1,3	Nos2	0/0
all2344	1,5	−1,1		0/0
alr2366	−1,1	−1,1	Nos2	0/0
alr2374	3,3	2,3		0/0
alr2522	up	up		0/0
asr3134	−1,7	−2,6	Nod1	0/0
asr3279	4,8	7,7	Nos2	0/0
all3520	2,5	2,4	Fis1	0/0
alr3562	1,3	−1,7		0/0
all3568	1,1	−1,0		0/445
all3696	13,2	6,1		0/243
alr3720	9,3	3,6		0/0
all3745	−1,6	−1,6	Fis1	0/0
alr3910	−2,1	−1,5	Nos2	0/0
all4073	4,6	8,5		0/0
asl4098	1,6	1,3		0/0
all4117	2,4	1,8		0/0
all4381	5,5	5,1		0/0
alr4534	1,5	1,2		0/0
all4555	2,2	1,5	Nod1	0/0
asl4565	1,1	2,5		0/0
alr4684	1,6	4,2	Nos2	0/0
alr4714	2,6	1,8		0/0
asl4743	4,1	1,2		0/0
alr4788	1,7	1,7		0/0
asl4860	−1,2	1,7		0/0
all4962	9,7	5,6	Nos2	0/0
alr5005	1,3	1,1		0/0
asr5071	−1,4	1,1		0/0

Shown is the accession number of Anabaena sp. PCC 7120 (column 1), the fold change of expression after 12 and 21 h of nitrogen starvation compared to 0 h (Flaherty et al., 2011; column 2, 3; up, infinite; not, not expressed), the cyanobacteria for which no sequence is identified in the CLOG (column 5) and he number of sequences found in the genomes of Viridplantae or bacteria (V/B, column 7).

[a] Glycine max, Solanum lycopersicum.

[b] Streptomyces aurantiacus.

[c] Frankia sp. EUN1f, Streptomyces aurantiacus.

[d] Nitrosococcus halophilus.

and feature tree with a tree based on the 16S rRNA and the average amino acid identity (AAI) (Additional File 5 in Supplementary Material). As expected, the correlation between CLOG and IAA tree is the highest with a coefficient of 0.83, while the correlation between the feature tree and the two trees was lower but still detectable (correlation of 0.65 and 0.55, respectively). However, some alterations were observed (**Figure 4**). The CLOG assignment relates the filamentous *Nodularia spumigena* CCY9414 (Nod1) to Nostocales, whereas the feature assignment introduces a shift to Oscillatoriales (Osc1 and Lyn1), because they show similarity in growth habitat, trichome formation and toxin production (**Figure 4**, Additional File 1 in Supplementary Material). As expected the filamentous *Arthrospira* (Art1–Art3) clustered with Oscillatoriales in the CLOG tree, but not in the feature tree. This shift most likely reflects the assignment of *Arthrospira* as not nitrogen fixing, facultative aerobic, cells with helical cell shape and fresh water living, which is distinct from other Oscillatoriales (Additional File 1 in Supplementary Material). Finally, two Prochlorales strains (*P. marinus* MIT 9313, Pro3; *P. marinus* str. MIT 9303, Pro8) are not assigned to Prochlorales, but to the Chroococcales in the CLOG tree (Additional File 4 in Supplementary Material). For *P. marinus* MIT 9313 which has the second largest genome of all analyzed *P. marinus* strains, we speculate that observed clustering in the CLOG tree results from the large number of genes in "CLOGs of dispensable genes" that contain many genes from other species than *P. marinus* (**Figure 1**).

We used the two defined trees (**Figure 4**) to analyze the branch-specific core-genomes with focus on branches including the model system *Anabaena* sp. PCC 7120 (Ana1). At first we compared the size of the core-genomes of the different branches to the expected random average size of core-genomes with the same number of strains (**Figure 2A**). We realized that the core genome for the strains in clade I (**Figure 4A**), A and B (**Figure 4B**) is two-fold larger than expected from our analysis. This could be due the large cyanobacterial genomes in this clade (>5 Mb) when compared to the small genomes from

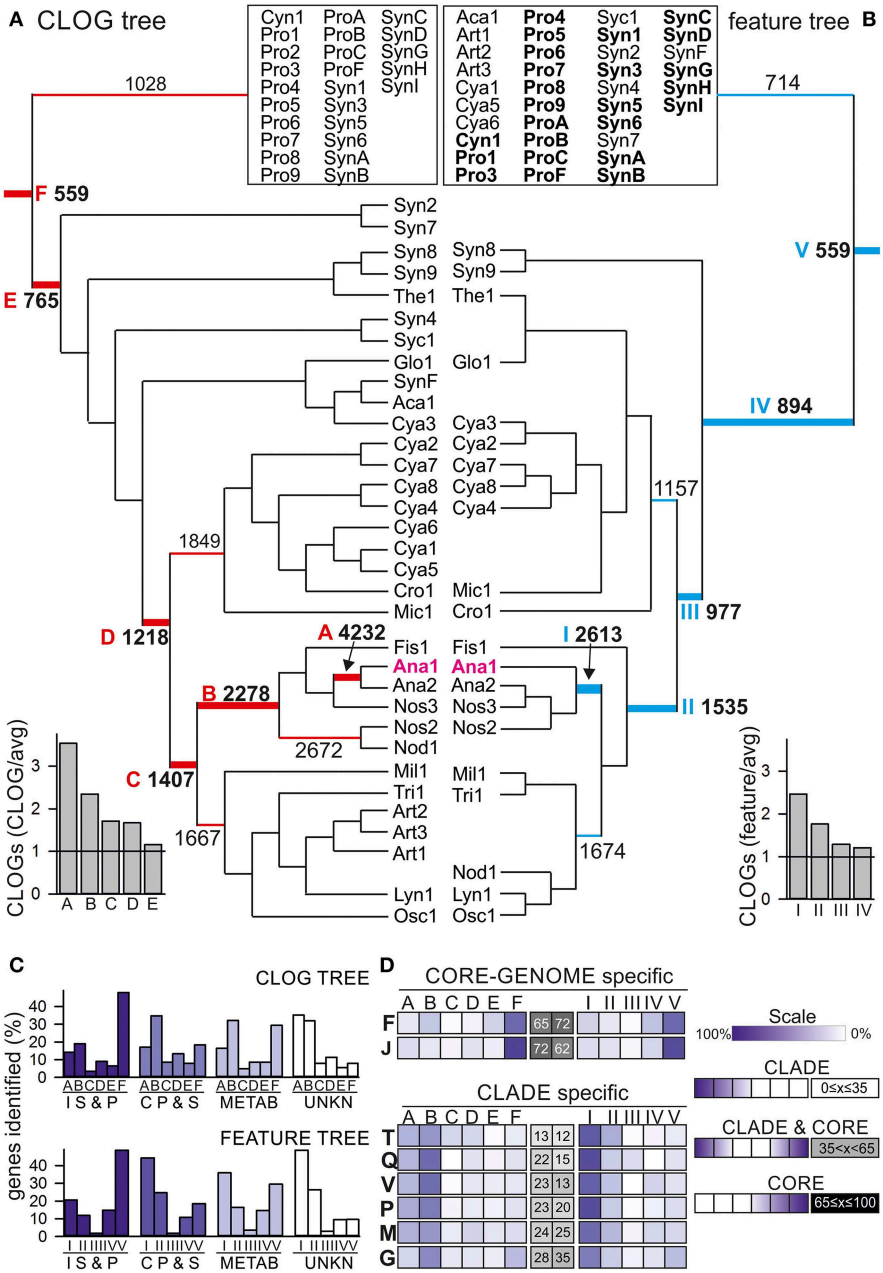

FIGURE 4 | Feature and shared CLOG correlation tree. (A, B) The neighbor-joining tree of the 58 cyanobacteria based **(A)** on pair-wise shared CLOGs as distances or **(B)** on the similarities in the 13 selected features as distances was calculated. The root for the different branches from the deepest root (CORE-GENOME) to *Anabaena* sp. PCC 7120 are marked by letter in **A** (F–A) or roman numerals in **B** (I–VI), and the number of CLOGs defining the core-genome for the branch with this root is given. The ratio of the core-genomes of the branches with different roots to the average size of the core-genome expected for this number (**Figure 2**) is indicated on the bottom left. For simplicity, only branches discussed are shown, while all strains of the remaining part of the tree are clustered in the box on top. The full tree is shown in Additional File 4 in Supplementary Material.**(C)** Each core-genome with the root indicated in **(A,B)** was determined and the number of proteins of a specific category/process (**Table 4**) additionally found to the core-genome of the deeper roots was counted and is deposited in Additional Files 8, 9 in Supplementary Material. Shown is the occurrence of unique proteins (in percent of all identified proteins) assigned to the four categories "Information storage and processing" (I, S, and P), "Cellular processes and signaling" (C, P, and S), "Metabolism" (METAB) and unknown (UNKN) in the different clade specific core genomes defined for the CLOG tree (top) and feature tree (bottom). **(D)** Shown is the occurrence of unique proteins assigned to the individual processes (indicated by one letter code shown in **Table 4**). The distribution for proteins for each process is shown as color code indicated on the right (Scale). For each distribution the profile was analyzed by an inversed gaussian distribution and the position of the minimum was used to assign the process as CLADE specific defined, CLADE and CORE-GENOME defined or CORE genome defined (scale is shown on the right, position of the minimum is given in percent: 0% = exclusive detection in core genome of CLADE A or I, 100% = exclusive detection in CORE-GENOME. The results for equally distributed (CORE and CLADE) genes are shown in Additional File 10 in Supplementary Material.

Chroococcales included in the CORE-GENOME calculation. However, this is in agreement with the close relation of the cyanobacteria in these clades. Next, we determined the functional categories based on the sequences of *Anabaena* sp. PCC 7120 for the core-genomes of different branches defined by the indicated roots (**Figure 4**) of the CLOG (Additional Files 4, 8 in Supplementary Material) and feature-based tree (Additional Files 4, 9 in Supplementary Material) by the strategy described for the CORE-GENOME classification.

We inspected the distribution of the genes of the four functional categories (**Figure 4C**). For proteins involved in the metabolism (METAB) we found a comparable number in the CORE-GENOME (root F, V; entire tree) as in the clade specific core-genome (root A, B, I, II), while most of the proteins assigned as "Information storage and processing (IS and P)" are found already in the CORE-GENOME (root F, V; **Figure 4C**). Proteins of unknown function (UNKN) and of "Cellular processes and signaling (CP and S)" are largely found in the clade specific core genomes (root A, B, I, II, **Figure 4C**). On the one hand this suggests that many strain specific processes have not yet been characterized, on the other hand it can be postulated that cyanobacterial signaling strategies are largely strain specific.

To substantiate the latter notion, we analyzed the distribution of the proteins assigned to the various biological processes (**Table 1**) in the different clade specific core-genomes. We realized that proteins of most categories are found in the CORE-GENOME of all cyanobacteria as well as in clade specific core genomes (Additional Files 9–11 in Supplementary Material). Only proteins of category N (cell motility) are not represented by the CORE-GENOME, but the detected proteins are equally found in all clade specific core-genomes (Additional Files 11 in Supplementary Material). However, we observed two processes for which most of the proteins are encoded by the CORE-GENOME, namely translation, ribosomal structure and biogenesis (category J), as well as in nucleotide metabolism and transport (category F; **Figure 4D**). This finding is not unexpected as the process of protein synthesis and nucleotide metabolism were previously identified to be very ancient even existing in the last universal common ancestor (e.g., Poole et al., 1999; Armenta-Medina et al., 2014). In contrast, many proteins classified to be involved in signal transduction and defense mechanisms show a clade specific occurrence (categories V and T, **Figure 4D**). This supports the above formulated notion that cyanobacterial signaling strategies are largely strain specific.

In addition, proteins involved in inorganic ion, secondary metabolite and carbohydrate metabolism and transport (categories G, P, and Q) as well as in cell wall and cell envelope biogenesis (category M; **Figure 4D**) are largely CLADE specific. This finding suggests that not only signaling strategies, but also the mechanisms to interact with the environment are specific for small clades of cyanobacteria and even for individual strains.

The β-Barrel Proteins in Cyanobacteria

To confirm the notion that the proteome for the interaction with the environment, particularly for the uptake and secretion of molecules is highly clade specific, we aimed for the identification of putative OMPs as they are involved in such processes. We focused on proteins characterized by a membrane-embedded β-barrel domain as representative subset of the outer membrane proteome. We developed a consensus approach for the prediction of β-barrel OMPs in the cyanobacterial proteomes (see Materials and Methods). This approach yielded 703 putative β-barrel proteins detected by all criteria [category (a); **Table 2**], 179 which fulfill the two main criteria and at least one minor criterion [category (b); **Table 2**] and 37 which fulfill the two main criteria only, but are confirmed by tertiary structure prediction [category (c); **Table 2**]. All other proteins were not considered as putative β-barrel proteins [category (d); **Table 2**]. We clustered the sequence stretches representing the putative β-barrel domains of all selected proteins to assign functional properties as previously established (Mirus et al., 2009). We detected 21 clusters of β-barrel proteins with more than four sequences, which represent 12 functional groups based on domains defined by Pfam (**Table 7**, Additional File 12 in Supplementary Material).

Sequences of three β-barrel protein families are found in almost all strains analyzed, namely the OMP of 85 kDa (Omp85; Pfam: Bac_surface_Ag; Moslavac et al., 2005), the lipopolysaccharide transport protein D (LptD; Pfam: DUF3769; Haarmann et al., 2010), and the carbohydrate-selective porin (Pfam: OprB–OMP from *Pseudomonas aeruginosa;* **Table 7**). Omp85 and LptD are the two central proteins of outer membrane biogenesis of Gram-negative bacteria and belong to the most ancient outer membrane proteins (e.g., Bredemeier et al., 2007; Hahn and Schleiff, 2014), while a porin like OprB is generally required for solute transport. However, only Omp85 is a true component of the CORE-GENOME of cyanobacteria (**Figure 5**), because orthologs to LptD could not be identified *Acaryochloris marina* and *Synechococcus* sp. *CB0205,* although proteins with low similarity exist. For OprB we realized that the identified sequences cluster in different CLOGs, which is consistent with the detection of the protein family in all strains but the absence in the CORE-GENOME.

In addition, sequences with the broad signature for outer membrane localized β-barrel proteins (OmpA_Pfam/OmpA_OMPdb/OMP-β-brl; cluster 11, 13–16, and 18, **Table 7**, Additional File 11 in Supplementary Material) are found in the genome of 33 strains of all six cyanobacterial orders, which suggests that most of the cyanobacterial strains have additional outer envelope transporters to OrpB. However, they appear to be strain specific as they are not encoded by any clade specific core genome (**Figure 5**). The same holds true for the TonB dependent transporter involved in metal transport (Mirus et al., 2009), which was identified in all cyanobacterial orders, but only in 22 strains (**Table 7**).

All other identified β-barrel protein families are restricted to a lower number of strains and cyanobacterial orders. For example, proteins with a domain characteristic of autotransporters are specific for *Synechococcus* strains (**Table 7**). Moreover, β-barrel proteins with the INTIMIN/INVASIN domain are only found in five strains of the Prochlorales, in nine *Synechococcus* strains, in *Acaryochloris marina* MBIC11017 and in *Microcoleus chthonoplastes* PCC 7420. Such domains are usually found in virulence

TABLE 7 | Clusters of β-barrel representing sequences.

Pfam nomenclature for β-barrel domain	Cluster	Strain		Orders*		Sequences	
(Glucose selective) OprB	9	7	58	2	6	8	295
	20	9		2		15	
	21	58		6		274	
Omp85	10	58		6		155	
LptD (DUF3769)	7	56		6		56	
TonB_dep_Rec/TBDT	6	22		6		124	
OmpA_Pfam/OMPdb	11	14	20	5	5	15	27
	14	3		3		5	
	16	7		2		7	
Omp_β-brl	13	7	17	4	5	10	27
	15	9		1		11	
	18	5		4		6	
DUF3442 Intimin/Invasin	3	5	16	2	3	6	32
	4	3		3		6	
	5	10		2		20	
Fasciclin	17	5		3		5	
DUF481[a]	1	4	15	1	2	6	17
	8	11		2		11	
OmpW	19	5		2		5	
Cellulose synthesis complex barrel/BcsC	2	5		2		5	
Autotransporter	12	4		1		5	

Shown are the names of the Pfam domains characteristic for the β-barrel families (column 1), the number of the cluster according to Additional File 11 in Supplementary Material (column 2), number of strains of which a sequence is present in the cluster (column 3) or in all clusters of the same family (column 4), the number of orders of which sequences are in the cluster (column 5) or in all clusters of the same family (column 6), and the number of different sequences in the cluster (column 7) or in all clusters of the same family (column 8).

[*] Orders: Chroococcales, Gloeobacterales, Nostocales, Oscillatoriales, Prochlorales, Stigonematales.

[a] DUF, domain of unknown function.

factors of enteropathogenic bacteria, mediating invasion into and adherence to host cells (Bodelon et al., 2013). All strains with such proteins are unicellular (except *M. chthonoplastes* PCC 7420) and live in the sea, which might require proteins with such domain for the association of cells to other organisms of the community.

Furthermore, OMPs with a domain characteristic for the cellulose synthase subunit with β-barrel (BcsC) or a FASCLINE domain are found in only eight strains, namely the heterocyst-forming *Anabaena* sp. PCC 7120 (both proteins), *Anabaena variabilis* ATCC 29413 (BcsC), *Nostoc punctiforme* PCC 73102 (both), *Fischerella* sp. JSC-11 (FASCLINE), *Nodularia spumigena* CCY9414 (FASCLINE) as well as in *Acaryochloris marina* MBIC11017 (BscC), *Synechococcus* sp. PCC 7002 (BscC) and *Oscillatoria* sp. PCC 6506 (FASCLINE). BcsC is involved in poly-β-1,6-N-acetyl-D-glucosamine or cellulose export (Keiski et al., 2010). Thus, such a protein might be involved in the formation of the heterocyst specific glycolipid layer and the heterocyst polysaccharide envelope (e.g., Nicolaisen et al., 2009). The FASCICLIN domain is an ancient cell adhesion domain (Borner et al., 2002) that might link the heterocyst specific layer to the outer membrane. In line, the gene of *Anabaena* sp. PCC 7120 (alr3754) with the BscC domain is highly induced (∼10-fold) by nitrogen starvation (Flaherty et al., 2011) and the protein with FASCLINE domain was found in heterocyst membrane proteome (Moslavac et al., 2007). Thus, we propose that the function of two OMP families with BcsC or FASCICLIN domains identified in cyanobacteria is most likely related to heterocyst formation, although the experimental evidence is still missing.

From the inspection of the β-barrel proteome we conclude that the basic set for fundamental processes of outer membrane biogenesis represented by Omp85 and LptD and the basic principle of solute exchange represented by OprB are indeed globally conserved, while the majority of the β-barrel OMPs has

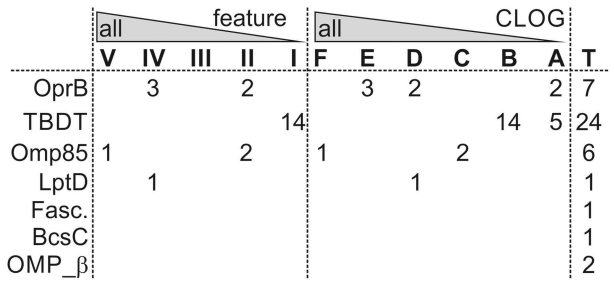

FIGURE 5 | β-barrel proteins in various core-genomes. Given are the numbers of OMPs characterized by the indicated domains (**Table 7**) found in *Anabaena* sp. PCC 7120, which are present in the indicated core-genome of the feature or CLOG tree (**Figure 3**). T indicates the total number of identified sequences.

evolved clade or strain specific to adapt to environmental situations. The large number of proteins with a membrane anchoring domain with general β-barrel signature in various analyzed strains (**Table 7**: OmpA, Omp_β, DUF481, and OmpW), but with distinct properties leading to a distinct CLOG assignment (**Figure 5**) supports the above formulated notion that mechanisms to interact with the environment are specific for small clades of cyanobacteria and even for individual strains.

Conclusion

The analysis of the protein sequences of 58 cyanobacterial strains of six different orders (**Table 3**) revealed a PAN-GENOME of about 44,831 genes (**Figure 2**). The cyanobacterial PAN-GENOME is considered to be open, which means that it will increase with each additional genome. In contrast, the CORE-GENOME of the 58 organisms is composed of 559 genes, and it is expected to level off at about 500 sequences (**Figure 2**). Roughly 20% of the CORE-GENOME is composed of genes involved in protein homeostasis, whereas most of the other genes perform housekeeping functions (**Table 4**). The individual genomes of cyanobacteria are largely composed of genes of the so-called dispensable-genome genomes, while unique genes are the minority (**Figure 1**). Based on the comparability of the trees calculated on the base of the genetic information or on features of the cyanobacteria (**Figure 3**, **Table 1**) we confirm that features dominate the genomic content. On the one hand, this is supported by the observation that for some features like "heterocyst formation" specific genes can be assigned (**Tables 5, 6**). On the other hand, analysis of clade specific core-genomes shows the ancient occurrence of processes like translation, ribosomal biogenesis and nucleotide metabolism, while processes involved in reactions to the environment like signal transduction and cell wall biogenesis are highly clade specific (**Figure 4**). The latter is also supported by the analysis of a specific protein family, namely the β-barrel shaped OMPs. Proteins involved in fundamental processes like outer membrane biogenesis (Omp85, LptD, **Figure 5**, **Table 7**) are globally conserved, while the majority of the β-barrel proteins are rather specific for clades of common features or even strain specific (**Figure 5**). Thus, while the CORE-GENOME describes the housekeeping and protein homeostasis functions, the proteins involved in environment response mechanisms are largely individualized for the various cyanobacteria.

Author Contributions

ES conceptualized, designed and headed the project. SS and MK performed the literature survey, the computational pan-genome and core-genome analysis. SS and MS implemented the β-barrel prediction approach. All authors were involved in analyzing the *in silico* results. ES, MK, and SS were involved in writing the manuscript.

Acknowledgments

We thank our colleagues for careful reading of the Manuscript, particularly B. Weis. The work was supported by grants from the Deutsche Forschungsgemeinschaft DFG SCHL 585-3 and 585-7 to ES. We thank Nadine Flinner, Oliver Mirus, Sotirios Fragkostefanakis, and Mara Stevanovic for critical discussion of the manuscript.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00219/abstract

Supplementary File 1 (Table)—Features of the 58 Cyanobacterial Strains

For all 58 cyanobacteria information on 13 selected features is presented. For better readability, the information is split in three sub-tables. In Tables S1A, S1B the abbreviation assigned to each strain (**Table 1**) is given (column 1) and the following columns list the information on growth habitat, growth temperature, cultivation in the Lab or collected from nature, cell shape, cell order, mobility and toxin production (Table S1A), as well as on the ability to form Heterocysts, Akinetes, Hormogonia, or Trichome, on the ability to fix nitrogen as well as on their oxygen demand (Table S1B). The source of the information is represented in brackets and the reference is given in Table S1C.

Supplementary File 2 (Figure)—Calculation of the Tanimoto-Like Index

The feature similarity between two strains was calculated by a Tanimoto-like index (see Materials and Methods). Each feature is divided in categories (rectangles) (like feature habitat is divided in the subcategories: mud; fresh water; sea etc.). For each subcategory in each feature a value for present (1), not present (0), or unknown (u) is added. For each feature three different cases could occur: (I) unknown feature in one of two strains (1. feature) counts 0.5 in the denominator, (II) unknown feature of both strains (2. feature) is excluded from counting, and

(III) known feature in both strains (3. feature) counts as the quotient of the intersection in the numerator and union in the denominator.

Supplementary File 3 (Figure)—Heat Map of Feature-Based Distances of Cyanobacteria

Shown is the heat map of the distance of the 58 cyanobacterial strains analyzed in here based on the Tanimoto-like index for the 13 different features. The pair-wise distance is represented in a color code based on percentage calculated by the Tanimoto-like index. Black, 0% distance—related to each other with respect to the features analyzed; white, 100% distance—not related to each other with respect to the features analyzed.

Supplementary File 4 (Figure)—Neighbor-Joining Trees of Figure 3

(**A, B**)—The neighbor-joining tree of the 58 cyanobacterial organisms is based on their pairwise shared CLOGs (**A**) or the feature distance (**B**). In A the number of shared CLOGs including two organisms is used for distance calculation. In B, the feature distance was calculated by a pairwise Tanimoto-like index based on the intersection of 13 features. The patristic distance correlation had a value of 0.51.

Supplementary File 5 (Figure)—Neighbor-Joining Trees Of 16s rRNA and AAI

(A, B)—The neighbor-joining tree of the 58 cyanobacterial strains is based on their alignment of 16S rRNA sequences (**A**) or average amino acid identity (AAI) (**B**). In A the 16S rRNA sequences were multiple aligned by MAFFT. In B, 420 CLOGs of the CORE-GENOME with a single ortholog per strain were pairwise globally aligned the average over the CLOGs calculated to define a distance for each pair of strains. The patristic distance correlation between both trees is 0.76 meaning a strong correlation.

Supplementary File 6 (Table)—Clogs of the Core-Genome

Shown are the groups of the OrthoMCL ortholog search representing the CLOGs of the CORE-GENOME (column 1) and for each cyanobacterial strain the gene accessions (column 2–59).

Supplementary File 7 (Figure)—Core- and PAN-Genome Size Dependence on the Number of Analyzed Strains

Shown are the numbers of total CLOGs in the core-genome (**A**) or the pan-genome (**B**) derived from the analysis of the given number of organisms (x-axis), which have been randomly selected 100 (left) or 10,000 times (right). The results are plotted as box-plots. Values for 1000 iterations are shown in **Figure 2**.

Supplementary File 8 (Table)—Distribution of Anabaena sp. PCC 7120 Proteins Involved in Oxidative Phosphorylation and Photosynthesis According to KEGG Assignment in the Core-Genomes of Different Clades of the Feature Based Tree

The table gives: the root of the clade of the feature based tree for which the core genome was defined (column 1), the KEGG number of the protein (column 2), the name of the protein (column 3), the accession number of the according gene in *Anabaena* sp. PCC 7120 (column 4) and the functional category according to KEGG (column 6) and the functional category according to COG (column 7: Energy prod, energy production and conversion; non, no functional assignment in COG, other, a functional assignment distinct from energy production and conversion).

Supplementary File 9 (Table)—Functional Categories of the Core-Genomes Based on the Clog-Based Tree Exemplified for Anabaena sp. PCC 7120

Given is the functional category (column 1), the abbreviation of the COG of the functional process (column 2) and the number of sequences of *Anabaena* sp. PCC 7120 assigned to the different core-genomes (columns 3–8) based on the CLOG tree (**Figure 4A**).

Supplementary File 10 (Table)—Functional Categories of the Core-Genomes Based on the Feature Tree Exemplified for Anabaena sp. PCC 7120

Given is the functional category (column 1), the abbreviation of the COG of the functional process (column 2) and the number of sequences of *Anabaena* sp. PCC 7120 assigned to the different core-genomes (columns 3–8) based on the feature tree (**Figure 4B**).

Supplementary File 11 (Figure)—Proteins Found in Core and Clade Genes

Shown is the occurrence of unique proteins assigned to the individual processes (indicated by one letter code shown in **Table 2**). The distribution for proteins for each process is shown as color code indicated in **Figure 4D**. For each distribution the profile was analyzed by an inversed Gaussian distribution and the position of the minimum was used to assign the process as CLADE and CORE-GENOME defined.

Supplementary File 12 (Figure)—Clustering of Predicted β-Barrel Proteins

Shown are clusters of amino acid sequences sections of putative cyanobacterial β-barrel proteins of category (a), (b), and (c) (**Table 2**) via CLANS. The clusters were numbered and colored according to their predicted function (**Table 7**). Distances below $1.0 \times e^{-20}$ are shown and contain the same functional or domain annotation.

References

Allewalt, J. P., Bateson, M. M., Revsbech, N. P., Slack, K., and Ward, D. M. (2006). Effect of temperature and light on growth of and photosynthesis by *Synechococcus* isolates typical of those predominating in the octopus spring microbial mat community of Yellowstone National Park. *Appl. Environ. Microbiol.* 72, 544–550. doi: 10.1128/AEM.72.1.544-550.2006

Anagnostidis, K., and Komárek, J. (1987). Modern approach to the classification system of Cyanophytes. 3. Oscillatoriales. *Algol. Stud.* 50–53, 327–472.

Anagnostidis, K., and Komárek, J. (1990). Modern approach to the classification system of Cyanophytes. 5. Stigonematales. *Algol. Stud.* 59, 1–73

Araoz, R., Nghiem, H. O., Rippka, R., Palibroda, N., De Marsac, N. T., and Herdman, M. (2005). Neurotoxins in axenic oscillatorian cyanobacteria: coexistence of anatoxin-a and homoanatoxin-a determined by ligand-binding assay and GC/MS. *Microbiology* 151, 1263–1273. doi: 10.1099/mic.0.27660-0

Armenta-Medina, D., Segovia, L., and Perez-Rueda, E. (2014). Comparative genomics of nucleotide metabolism: a tour to the past of the three cellular domains of life. *BMC Genomics* 15:800. doi: 10.1186/1471-2164-15-800

Awai, K., and Wolk, C. P. (2007). Identification of the glycosyl transferase required for synthesis of the principal glycolipid characteristic of heterocysts of *Anabaena* sp. strain PCC 7120. *FEMS Microbiol. Lett.* 266, 98–102. doi: 10.1111/j.1574-6968.2006.00512.x

Bauer, C. C., Buikema, W. J., Black, K., and Haselkorn, R. (1995). A short-filament mutant of *Anabaena* sp. strain PCC 7120 that fragments in nitrogen-deficient medium. *J. Bacteriol.* 177, 1520–1526.

Beck, C., Knoop, H., Axmann, I. M., and Steuer, R. (2012). The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics* 13:56. doi: 10.1186/1471-2164-13-56

Berven, F. S., Flikka, K., Jensen, H. B., and Eidhammer, I. (2004). BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.* 32, W394–W399. doi: 10.1093/nar/gkh351

Black, K., Buikema, W. J., and Haselkorn, R. (1995). The hglK gene is required for localization of heterocyst-specific glycolipids in the cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 177, 6440–6448.

Bodelon, G., Palomino, C., and Fernandez, L. A. (2013). Immunoglobulin domains in *Escherichia coli* and other enterobacteria: from pathogenesis to applications in antibody technologies. *FEMS Microbiol. Rev.* 37, 204–250. doi: 10.1111/j.1574-6976.2012.00347.x

Bolhuis, H., Severin, I., Confurius-Guns, V., Wollenzien, U. I., and Stal, L. J. (2010). Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium *Microcoleus chthonoplastes*. *ISME J.* 4, 121–130. doi: 10.1038/ismej.2009.99

Borner, G. H., Sherrier, D. J., Stevens, T. J., Arkin, I. T., and Dupree, P. (2002). Prediction of glycosylphosphatidylinositol-anchored proteins in *Arabidopsis*. *A genomic analysis. Plant Physiol* 129, 486–499. doi: 10.1104/pp.010884

Bothe, H., Schmitz, O., Yates, M. G., and Newton, W. E. (2010). Nitrogen fixation and hydrogen metabolism in cyanobacteria. *Microbiol. Mol. Biol. Rev.* 74:529–551. doi: 10.1128/MMBR.00033-10

Bredemeier, R., Schlegel, T., Ertel, F., Vojta, A., Borissenko, L., Bohnsack, M. T., et al. (2007). Functional and phylogenetic properties of the pore-forming β-barrel transporters of the Omp85 family. *J. Biol. Chem.* 282, 1882–1890. doi: 10.1074/jbc.M609598200

Campbell, E. L., Christman, H., and Meeks, J. C. (2008). DNA microarray comparisons of plant factor- and nitrogen deprivation-induced Hormogonia reveal decision-making transcriptional regulation patterns in *Nostoc punctiforme*. *J. Bacteriol.* 190, 7382–7391. doi: 10.1128/JB.00990-08

Carey, C. C., Ibelings, B. W., Hoffmann, E. P., Hamilton, D. P., and Brookes, J. D. (2012). Eco-physiological adaptations that favour freshwater cyanobacteria in a changing climate. *Water Res.* 46, 1394–1407. doi: 10.1016/j.watres.2011.12.016

Carrieri, D., Ananyev, G., Lenz, O., Bryant, D. A., and Dismukes, G. C. (2011). Contribution of a sodium ion gradient to energy conservation during fermentation in the cyanobacterium *Arthrospira (Spirulina) maxima* CS-328. *Appl. Environ. Microbiol.* 77, 7185–7194. doi: 10.1128/AEM.00612-11

Chen, F., Mackey, A. J., Stoeckert, C. J. Jr., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34, D363–D368. doi: 10.1093/nar/gkj123

Cohen, M. F., Wallis, J. G., Campbell, E. L., and Meeks, J. C. (1994). Transposon mutagenesis of *Nostoc* sp. strain ATCC 29133, a filamentous cyanobacterium with multiple cellular differentiation alternatives. *Microbiology* 140, 3233–3240. doi: 10.1099/13500872-140-12-3233

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R. C., et al. (2011). Unity in variety—the pan-genome of the Chlamydiae. *Mol. Biol. Evol.* 28, 3253–3270. doi: 10.1093/molbev/msr161

Cooper, D. L., Mort, K. A., Allan, N. L., Kinchington, D., and McGuigan, C. (1993). Molecular similarity of anti-HIV phospholipids. *J. Am. Chem. Soc.* 115, 12615–12616. doi: 10.1021/ja00079a063

Corrales-Guerrero, L., Mariscal, V., Nurnberg, D. J., Elhai, J., Mullineaux, C. W., Flores, E., et al. (2014). Subcellular localization and clues for the function of the HetN factor influencing heterocyst distribution in *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 196, 3452–3460. doi: 10.1128/JB.01922-14

Croce, R., and van Amerongen, H. (2014). Natural strategies for photosynthetic light harvesting. *Nat. Chem. Biol.* 10, 492–501. doi: 10.1038/nchembio.1555

D'Auria, G., Jimenez-Hernandez, N., Peris-Bondia, F., Moya, A., and Latorre, A. (2010). *Legionella pneumophila* pangenome reveals strain-specific virulence factors. *BMC Genomics* 11:181. doi: 10.1186/1471-2164-11-181

Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N. J., Angiuoli, S. V., et al. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol.* 11:R107. doi: 10.1186/gb-2010-11-10-r107

Du, Y., Cai, Y., Hou, S., and Xu, X. (2012). Identification of the HetR recognition sequence upstream of hetZ in *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 194, 2297–2306. doi: 10.1128/JB.00119-12

Dunn, B., Richter, C., Kvitek, D. J., Pugh, T., and Sherlock, G. (2012). Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res.* 22, 908–924. doi: 10.1101/gr.130310.111

Dutilh, B. E., Snel, B., Ettema, T. J., and Huynen, M. A. (2008). Signature genes as a phylogenomic tool. *Mol. Biol. Evol.* 25, 1659–1667. doi: 10.1093/molbev/msn115

Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H., and Stackebrandt, E. (eds.). (2006). *The Prokaryotes: Vol. 3: Archaea. Bacteria: Firmicutes, Actinomycetes.* New York, NY: Springer Science and Business Media.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Ehira, S., and Ohmori, M. (2012). The redox-sensing transcriptional regulator RexT controls expression of thioredoxin A2 in the cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Biol. Chem.* 287, 40433–40440. doi: 10.1074/jbc.M112.384206

Elliott, J. A. (2012). Is the future blue-green? A review of the current model predictions of how climate change could affect pelagic freshwater cyanobacteria. *Water Res.* 46:1364, 1371. doi: 10.1016/j.watres.2011.12.018

El-Shehawy, R., Lugomela, C., Ernst, A., and Bergman, B. (2003). Diurnal expression of hetR and diazocyte development in the filamentous non-heterocystous cyanobacterium *Trichodesmium erythraeum*. *Microbiology* 149, 1139–1146. doi: 10.1099/mic.0.26170-0

Fan, Q., Huang, G., Lechno-Yossef, S., Wolk, C. P., Kaneko, T., and Tabata, S. (2005). Clustered genes required for synthesis and deposition of envelope glycolipids in *Anabaena* sp. strain PCC 7120. *Mol. Microbiol.* 58, 227–243. doi: 10.1111/j.1365-2958.2005.04818.x

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Flaherty, B. L., van Nieuwerburgh, F., Head, S. R., and Golden, J. W. (2011). Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. *BMC Genomics* 12:332. doi: 10.1186/1471-2164-12-332

Flores, E., and Herrero, A. (2010). Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nat. Rev. Microbiol.* 8, 39–50. doi: 10.1038/nrmicro2242

Fujisawa, T., Narikawa, R., Okamoto, S., Ehira, S., Yoshimura, H., Suzuki, I., et al. (2010). Genomic structure of an economically important cyanobacterium, *Arthrospira (Spirulina) platensis* NIES-39. *DNA Res.* 17, 85–103. doi: 10.1093/dnares/dsq004

Gao, K., Yu, H., and Brown, M. T. (2007). Solar PAR and UV radiation affects the physiology and morphology of the cyanobacterium *Anabaena* sp. PCC 7120. *J. Photochem. Photobiol. B.* 89, 117–124. doi: 10.1016/j.jphotobiol.2007.09.006

Garcia-Pichel, F., Johnson, S. L., Youngkin, D., and Belnap, J. (2003). Small-scale vertical distribution of bacterial biomass and diversity in biological soil crusts from arid lands in the Colorado plateau. *Microb. Ecol.* 46, 312–321. doi: 10.1007/s00248-003-1004-0

Golden, J. W., and Yoon, H. S. (2003). Heterocyst development in *Anabaena*. *Curr. Opin. Microbiol.* 6, 557–563. doi: 10.1016/j.mib.2003.10.004

Gruber, T. M., and Bryant, D. A. (1998). Characterization of the alternative sigma-factors SigD and SigE in *Synechococcus* sp. strain PCC 7002. SigE is implicated in transcription of post-exponential-phase-specific genes. *Arch. Microbiol.* 169, 211–219. doi: 10.1007/s002030050563

Gupta, R. S., and Mathews, D. W. (2010). Signature proteins for the major clades of Cyanobacteria. *BMC Evol. Biol.* 10:24. doi: 10.1186/1471-2148-10-24

Haarmann, R., Ibrahim, M., Stevanovic, M., Bredemeier, R., and Schleiff, E. (2010). The properties of the outer membrane localized Lipid A transporter LptD. *J. Phys. Condens. Matter* 22, 454124. doi: 10.1088/0953-8984/22/45/454124

Hahn, A., and Schleiff, E. (2014). "The cell envelope," in *The Cell Biology of Cyanobacteria*, eds E. Flores and A. Herrero (Norfolk: Caister Academic Press), 29–87.

Herrero, A., Muro-Pastor, A. M., and Flores, E. (2001). Nitrogen control in cyanobacteria. *J. Bacteriol.* 183, 411–425. doi: 10.1128/JB.183.2.411-425.2001

Higa, K. C., and Callahan, S. M. (2010). Ectopic expression of hetP can partially bypass the need for hetR in heterocyst differentiation by *Anabaena* sp. strain PCC 7120. *Mol. Microbiol.* 77, 562–574. doi: 10.1111/j.1365-2958.2010.07257.x

Hu, J., and Yan, C. (2008). A method for discovering transmembrane beta-barrel proteins in Gram-negative bacterial proteomes. *Comput. Biol. Chem.* 32, 298–301. doi: 10.1016/j.compbiolchem.2008.03.010

Huang, G., Fan, Q., Lechno-Yossef, S., Wojciuch, E., Wolk, C. P., Kaneko, T., et al. (2005). Clustered genes required for the synthesis of heterocyst envelope polysaccharide in *Anabaena* sp. strain PCC 7120. *J Bacteriol* 187, 1114–1123. doi: 10.1128/JB.187.3.1114-1123.2005

Huber, A. L. (1985). Factors affecting the germination of akinetes of *Nodularia spumigena* (Cyanobacteriaceae). *Appl. Environ. Microbiol.* 49, 73–78.

Ionescu, D., Voss, B., Oren, A., Hess, W. R., and Muro-Pastor, A. M. (2010). Heterocyst-specific transcription of NsiR1, a non-coding RNA encoded in a tandem array of direct repeats in cyanobacteria. *J. Mol. Biol.* 398, 177–188. doi: 10.1016/j.jmb.2010.03.010

Jones, C. S., and Mayfield, S. P. (2012). Algae biofuels: versatility for the future of bioenergy. *Curr. Opin. Biotechnol.* 23, 346–351. doi: 10.1016/j.copbio.2011.10.013

Jones, K. (1992). Diurnal nitrogen fixation in tropical marine cyanobacteria: a comparison between adjacent communities of non-heterocystous *Lyngbya* sp. and heterocystous Calothrix sp. *Br. Phycol. J.* 27, 107–118. doi: 10.1080/00071619200650121

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kaneko, T., Nakajima, N., Okamoto, S., Suzuki, I., Tanabe, Y., Tamaoki, M., et al. (2007). Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res.* 14, 247–256. doi: 10.1093/dnares/dsm026

Kaneko, T., and Tabata, S. (1997). Complete genome structure of the unicellular cyanobacterium *Synechocystis* sp. PCC6803. *Plant Cell Physiol.* 38, 1171–1176. doi: 10.1093/oxfordjournals.pcp.a029103

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Keiski, C. L., Harwich, M., Jain, S., Neculai, A. M., Yip, P., Robinson, H., et al. (2010). AlgK is a TPR-containing protein and the periplasmic component of a novel exopolysaccharide secretin. *Structure* 18, 265–273. doi: 10.1016/j.str.2009.11.015

Kelley, L. A., and Sternberg, M. J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* 4, 363–371. doi: 10.1038/nprot.2009.2

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., et al. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet.* 3:e231. doi: 10.1371/journal.pgen.0030231

Kim, C. J., Jung, Y. H., and Oh, H. M. (2007). Factors indicating culture status during cultivation of *Spirulina (Arthrospira) platensis*. *J. Microbiol.* 45, 122–127.

Komárek, J., and Anagnostidis, K. (1986). Modern approach to the classification system of Cyanophytes. 2. Chroococcales. *Algol. Stud.* 43, 157–226.

Komárek, J., and Anagnostidis, K. (1989). Modern approach to the classification system of Cyanophytes. 4. Nostocales. *Algol. Stud.* 56, 247–345.

Larsson, J., Nylander, J. A., and Bergman, B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evol. Biol.* 11:187. doi: 10.1186/1471-2148-11-187

Lazaro, S., Fernandez-Pinas, F., Fernandez-Valiente, E., Blanco-Rivero, A., and Leganes, F. (2001). pbpB, a gene coding for a putative penicillin-binding protein, is required for aerobic nitrogen fixation in the cyanobacterium *Anabaena* sp. strain PCC7120. *J Bacteriol* 183, 628–636. doi: 10.1128/JB.183.2.628-636.2001

Lesser, M. (2003). Advances in Marine Biology cumulative index volumes 20–44. *Adv. Mar. Biol.* 45, 9–312. doi: 10.1016/S0065-2881(03)45002-5

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., et al. (2010). Building the sequence map of the human pan-genome. *Nat. Biotechnol.* 28, 57–63. doi: 10.1038/nbt.1596

Liang, J., Scappino, L., and Haselkorn, R. (1992). The patA gene product, which contains a region similar to CheY of *Escherichia coli*, controls heterocyst pattern formation in the cyanobacterium *Anabaena* 7120. *Proc. Natl. Acad. Sci. U.S.A.* 89, 5655–5659. doi: 10.1073/pnas.89.12.5655

Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., and Lomize, A. L. (2012). OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.* 40, D370–D376. doi: 10.1093/nar/gkr703

Maddison, W. P., and Maddison, D. R. (2011). *Mesquite: A Modular System for Evolutionary Analysis. Version 2.75*. Available online at: http://mesquiteproject.org (725 modules)

Markowitz, V. M., Chen, I. M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., et al. (2012). IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* 40, D115–D122. doi: 10.1093/nar/gkr1044

Martin, K. A., Siefert, J. L., Yerrapragada, S., Lu, Y., McNeill, T. Z., Moreno, P. A., et al. (2003). Cyanobacterial signature genes. *Photosyn. Res.* 75, 211–221. doi: 10.1023/A:1023990402346

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594. doi: 10.1016/j.gde.2005.09.006

Mejean, A., Mazmouz, R., Mann, S., Calteau, A., Medigue, C., and Ploux, O. (2010). The genome sequence of the cyanobacterium *Oscillatoria* sp. PCC 6506 reveals several gene clusters responsible for the biosynthesis of toxins and secondary metabolites. *J Bacteriol* 192, 5264-5265. doi: 10.1128/JB.00704-10

Mirus, O., Hahn, A., and Schleiff, E. (2010). "Outer membrane proteins," in *Prokaryotic Cell Wall Compounds. Structure and Biochemistry,* eds H. König, H. Claus, and A. Varma (Berlin; Heidelberg: Springer-Verlag), 175–230. doi: 10.1007/978-3-642-05062-6_6

Mirus, O., and Schleiff, E. (2005). Prediction of beta-barrel membrane proteins by searching for restricted domains. *BMC Bioinformatics* 6:254. doi: 10.1186/1471-2105-6-254

Mirus, O., Strauss, S., Nicolaisen, K., Von Haeseler, A., and Schleiff, E. (2009). TonB-dependent transporters and their occurrence in cyanobacteria. *BMC Biol.* 7:68. doi: 10.1186/1741-7007-7-68

Moller, S., Croning, M. D., and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646–653. doi: 10.1093/bioinformatics/17.7.646

Moslavac, S., Mirus, O., Bredemeier, R., Soll, J., Von Haeseler, A., and Schleiff, E. (2005). Conserved pore-forming regions in polypeptide-transporting proteins. *FEBS J.* 272, 1367–1378. doi: 10.1111/j.1742-4658.2005.04569.x

Moslavac, S., Reisinger, V., Berg, M., Mirus, O., Vosyka, O., Ploscher, M., et al. (2007). The proteome of the heterocyst cell wall in *Anabaena* sp. PCC 7120. *Biol. Chem.* 388, 823-829. doi: 10.1515/BC.2007.079

Mulkidjanian, A. Y., Koonin, E. V., Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., et al. (2006). The cyanobacterial genome core and the origin of photosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13126–13131. doi: 10.1073/pnas.0605709103

Mur, L. R., Skulberg, O. M., and Utkilen, H. (1999). "Cyanobacteria in the environment," in *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring, and Management*, Chapter 2, eds I. Chorus and J. Bartram (London; New York: E&FN Spon), 15–40.

Muro-Pastor, A. M., and Hess, W. R. (2012). Heterocyst differentiation: from single mutants to global approaches. *Trends Microbiol.* 20, 548–557. doi: 10.1016/j.tim.2012.07.005

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., et al. (2002). Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1 (supplement). *DNA Res.* 9, 135–148. doi: 10.1093/dnares/9.4.135

Nakamura, Y., Kaneko, T., Sato, S., Mimuro, M., Miyashita, H., Tsuchiya, T., et al. (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids (supplement). *DNA Res.* 10, 181–201. doi: 10.1093/dnares/10.4.181

Nakao, M., Okamoto, S., Kohara, M., Fujishiro, T., Fujisawa, T., Sato, S., et al. (2010). CyanoBase: the cyanobacteria genome database update 2010. *Nucleic Acids Res.* 38, D379–D381. doi: 10.1093/nar/gkp915

Neilan, B. A., Pearson, L. A., Muenchhoff, J., Moffitt, M. C., and Dittmann, E. (2013). Environmental conditions that influence toxin biosynthesis in cyanobacteria. *Environ. Microbiol.* 15, 1239–1253. doi: 10.1111/j.1462-2920.2012.02729.x

Nguyen, T. A., Brescic, J., Vinyard, D. J., Chandrasekar, T., and Dismukes, G. C. (2012). Identification of an oxygenic reaction center psbADC operon in the cyanobacterium *Gloeobacter violaceus* PCC 7421. *Mol. Biol. Evol.* 29, 35–38. doi: 10.1093/molbev/msr224

Nicolaisen, K., Hahn, A., and Schleiff, E. (2009). The cell wall in heterocyst formation by *Anabaena* sp. PCC 7120. *J Basic Microbiol* 49, 5–24. doi: 10.1002/jobm.200800300

Oliver, J. W., and Atsumi, S. (2014). Metabolic design for cyanobacterial chemical synthesis. *Photosyn. Res.* 120, 249–261. doi: 10.1007/s11120-014-9997-4

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Ou, Y. Y., Gromiha, M. M., Chen, S. A., and Suwa, M. (2008). TMBETADISC-RBF: discrimination of beta-barrel membrane proteins using RBF networks and PSSM profiles. *Comput. Biol. Chem.* 32, 227–231. doi: 10.1016/j.compbiolchem.2008.03.002

Pernil, R., Herrero, A., and Flores, E. (2010). Catabolic function of compartmentalized alanine dehydrogenase in the heterocyst-forming cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 192, 5165–5172. doi: 10.1128/JB.00603-10

Ploug, H., Adam, B., Musat, N., Kalvelage, T., Lavik, G., Wolf-Gladrow, D., et al. (2011). Carbon, nitrogen and O(2) fluxes associated with the cyanobacterium *Nodularia spumigena* in the Baltic Sea. *ISME J.* 5, 1549–1558. doi: 10.1038/ismej.2011.20

Poole, A., Jeffares, D., and Penny, D. (1999). Early evolution: prokaryotes, the new kids on the block. *Bioessays.* 21, 880–889.

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J. A., Ininbergs, K., Zheng, W. W., et al. (2010). Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE* 5:e11486. doi: 10.1371/annotation/835c5766-5128-41c4-b636-adfe0c503103

Reno, M. L., Held, N. L., Fields, C. J., Burke, P. V., and Whitaker, R. J. (2009). Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8605–8610. doi: 10.1073/pnas.0808945106

Rouhiainen, L., Sivonen, K., Buikema, W. J., and Haselkorn, R. (1995). Characterization of toxin-producing cyanobacteria by using an oligonucleotide probe containing a tandemly repeated heptamer. *J. Bacteriol.* 177, 6021–6026.

Scott, N. L., Xu, Y., Shen, G., Vuletich, D. A., Falzone, C. J., Li, Z., et al. (2010). Functional and structural characterization of the 2/2 hemoglobin from *Synechococcus* sp. PCC 7002. *Biochemistry* 49, 7000–7011. doi: 10.1021/bi100463d

Shi, L., Li, J. H., Cheng, Y., Wang, L., Chen, W. L., and Zhang, C. C. (2007). Two genes encoding protein kinases of the HstK family are involved in synthesis of the minor heterocyst-specific glycolipid in the cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 189, 5075–5081. doi: 10.1128/JB.00323-07

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., et al. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. USA.* 110, 1053–1058. doi: 10.1073/pnas.1217107110

Singh, S. P., and Montgomery, B. L. (2011). Determining cell shape: adaptive regulation of cyanobacterial cellular differentiation and morphology. *Trends Microbiol.* 19, 278–285. doi: 10.1016/j.tim.2011.03.001

Sommer, M. S., Daum, B., Gross, L. E., Weis, B. L., Mirus, O., Abram, L., et al. (2011). Chloroplast Omp85 proteins change orientation during evolution. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13841–13846. doi: 10.1073/pnas.1108626108

Sommer, U. (2005). *Biologische Meereskunde*. Heidelberg: Springer.

Stal, L. J. K., and Krumbein W. E. (1985). Nitrogenase activity in the non-heterocystous cyanobacterium *Oscillatoria* sp. grown under alternating light-dark cycles. *Arch. Microbiol.* 143, 67–71. doi: 10.1007/BF00414770

Stewart, I., Eaglesham, G. K., McGregor, G. B., Chong, R., Seawright, A. A., Wickramasinghe, W. A., et al. (2012). First report of a toxic *Nodularia spumigena* (Nostocales/ Cyanobacteria) bloom in sub-tropical Australia. II. Bioaccumulation of nodularin in isolated populations of mullet (Mugilidae). *Int. J. Environ. Res. Public Health* 9, 2412–2443. doi: 10.3390/ijerph9072412

Stockel, J., Welsh, E. A., Liberton, M., Kunnvakkam, R., Aurora, R., and Pakrasi, H. B. (2008). Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc. Natl. Acad. Sci. U.S.A.* 105, 6156–6161. doi: 10.1073/pnas.0711068105

Su, Z., Mao, F., Dam, P., Wu, H., Olman, V., Paulsen, I. T., et al. (2006). Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium *Synechococcus* sp. WH 8102. *Nucleic Acids Res.* 34, 1050–1065. doi: 10.1093/nar/gkj496

Swingley, W. D., Chen, M., Cheung, P. C., Conrad, A. L., Dejesa, L. C., Hao, J., et al. (2008). Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium A*caryochloris marina*. *Proc. Natl. Acad. Sci. U.S.A.* 105, 2005–2010. doi: 10.1073/pnas.0709772105

Takaichi, S., Mochimaru, M., and Maoka, T. (2006). Presence of free myxol and 4-hydroxymyxol and absence of myxol glycosides in *Anabaena variabilis* ATCC 29413, and proposal of a biosynthetic pathway of carotenoids. *Plant Cell Physiol.* 47, 211–216. doi: 10.1093/pcp/pci236

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631–637. doi: 10.1126/science.278.5338.631

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* 11, 472–477. doi: 10.1016/j.mib.2008.09.006

Tsirigos, K. D., Bagos, P. G., and Hamodrakas, S. J. (2011). OMPdb: a database of {beta}-barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Res.* 39, D324–D331. doi: 10.1093/nar/gkq863

Tuit, C., Waterbury, J., and Ravizza, G. (2004). Diel variation of moldybdenum and iron in marine diazotrophic cyanobacteria. *Limn. Oceanogr.* 49, 978–990. doi: 10.4319/lo.2004.49.4.0978

Urmeneta, J., Navarrete, A., Huete, J., and Guerrero, R. (2003). Isolation and characterization of cyanobacteria from microbial mats of the Ebro Delta, Spain. *Curr. Microbiol.* 46, 199–204. doi: 10.1007/s00284-002-3856-9

Valério, E., Chambel, L., Paulino, S., Faria, N., Pereira, P., and Tenreiro, R. (2009). Molecular identification, typing and traceability of cyanobacteria from freshwater reservoirs. *Microbiology* 155, 642–656. doi: 10.1099/mic.0.022848-0

van Domselaar, G. H., Stothard, P., Shrivastava, S., Cruz, J. A., Guo, A., Dong, X., et al. (2005). BASys: a web server for automated bacterial genome annotation. *Nucleic Acids Res.* 33, W455–W459. doi: 10.1093/nar/gki593

Wang, Y., Lechno-Yossef, S., Gong, Y., Fan, Q., Wolk, C. P., and Xu, X. (2007). Predicted glycosyl transferase genes located outside the HEP island are required for formation of heterocyst envelope polysaccharide in *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 189, 5372–5378. doi: 10.1128/JB.00343-07

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

Wijffels, R. H., Kruse, O., and Hellingwerf, K. J. (2013). Potential of industrial biotechnology with cyanobacteria and eukaryotic microalgae. *Curr. Opin. Biotechnol.* 24, 405–413. doi: 10.1016/j.copbio.2013.04.004

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12:444. doi: 10.1186/1471-2164-12-444

Zehr, J. P. (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol.* 19, 162–173. doi: 10.1016/j.tim.2010.12.004

Zhang, C. C., Friry, A., and Peng, L. (1998). Molecular and genetic analysis of two closely linked genes that encode, respectively, a protein phosphatase 1/2A/2B homolog and a protein kinase homolog in the cyanobacterium *Anabaena* sp. strain PCC 7120. *J. Bacteriol.* 180, 2616–2622.

Zhang, C. C., and Libs, L. (1998). Cloning and characterisation of the pknD gene encoding an eukaryotic-type protein kinase in the cyanobacterium *Anabaena* sp. PCC 7120. *Mol. Gen. Genet.* 258, 26–33. doi: 10.1007/s004380050703

Zhang, W., Du, Y., Khudyakov, I., Fan, Q., Gao, H., Ning, D., et al. (2007). A gene cluster that regulates both heterocyst differentiation and pattern formation in *Anabaena* sp. strain PCC 7120. *Mol. Microbiol.* 66, 1429–1443. doi: 10.1111/j.1365-2958.2007.05997.x

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., and Yu, J. (2012). PGAP: pan-genomes analysis pipeline. *Bioinformatics* 28, 416–418. doi: 10.1093/bioinformatics/btr655

Zhou, R., and Wolk, C. P. (2002). Identification of an akinete marker gene in *Anabaena* variabilis. *J. Bacteriol.* 184, 2529–2532. doi: 10.1128/JB.184.9.2529-2532.2002

Zhu, J., Kong, R., and Wolk, C. P. (1998). Regulation of hepA of *Anabaena* sp. strain PCC 7120 by elements 5' from the gene and by hepK. *J. Bacteriol.* 180, 4233–4242.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on