frontiers in Genetics

PERSPECTIVE
published: 09 August 2019
doi: 10.3389/fgene.2019.00721

On the Role of Bioinformatics and Data Science in Industrial Microbiome Applications

*Bartholomeus van den Bogert[1], Jos Boekhorst[2], Walter Pirovano[3] and Ali May[1]**

[1] Research and Development Dept., BaseClear, Leiden, Netherlands, [2] NIZO Food Research, Ede, Netherlands, [3] Bioinformatics Dept., BaseClear, Leiden, Netherlands

Advances in sequencing and computational biology have drastically increased our capability to explore the taxonomic and functional compositions of microbial communities that play crucial roles in industrial processes. Correspondingly, commercial interest has risen for applications where microbial communities make important contributions. These include food production, probiotics, cosmetics, and enzyme discovery. Other commercial applications include software that takes the user's gut microbiome data as one of its inputs and outputs evidence-based, automated, and personalized diet recommendations for balanced blood sugar levels. These applications pose several bioinformatic and data science challenges that range from requiring strain-level resolution in community profiles to the integration of large datasets for predictive machine learning purposes. In this perspective, we provide our insights on such challenges by touching upon several industrial areas, and briefly discuss advances and future directions of bioinformatics and data science in microbiome research.

Keywords: DNA sequencing, microbiome, industrial biotechnology, probiotics, 16S rRNA gene profiling, metagenomics, bioinformatics, data science

OPEN ACCESS

Edited by:
Jens Stoye,
Bielefeld University, Germany

Reviewed by:
Gavin Douglas,
Dalhousie University, Canada
Christopher Fields,
University of Illinois at Urbana-Champaign, United States

***Correspondence:**
Ali May
ali.may@baseclear.nl

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *15 April 2019*
Accepted: *09 July 2019*
Published: *09 August 2019*

Citation:
van den Bogert B, Boekhorst J, Pirovano W and May A (2019) On the Role of Bioinformatics and Data Science in Industrial Microbiome Applications. Front. Genet. 10:721. doi: 10.3389/fgene.2019.00721

INTRODUCTION

Microbial communities play important roles in industrial processes such as the production of food, beverages, probiotics, paper, and cleaning products (for a review, see Singh et al., 2016). It has become an industrial standard to study the taxonomic composition and functional capabilities of these microorganisms using marker gene (e.g. 16S rRNA) and shotgun metagenome sequencing for product development, optimization, and quality control (Costessi et al., 2018). In addition, data from other omics sources such as metatranscriptomics and metabolomics can be used in integrative studies to generate leads, for instance in enzyme discovery. Some of the questions asked in these microbiome studies are related to determining the efficacy of probiotics and require strain-level characterization of the community composition (McFarland et al., 2018). Other studies focus on assessing the capability of microbial communities to produce certain compounds and necessitate recovering bacterial genomes from complex (e.g. soil) microbiomes (Howe et al., 2014). Extending microbiome applications to the public for actionable results, for example, to control blood sugar levels, requires a combination of advanced computational methods from bioinformatics, data mining, and machine learning (Zeevi et al., 2015).

In this perspective, we give an overview of several industrial microbiome applications with their bioinformatic and data science challenges. In addition, we highlight some of the advances that have the potential to provide valuable insights into the challenges facing these applications.

We conclude with sharing our view on the future directions and requirements of industrial microbiome applications in terms of their computational components.

CURRENT APPLICATIONS AND PRODUCTS

Dairy Starter Cultures

Microbial populations (e.g. of lactic acid bacteria) are used in a variety of food and beverage production processes including the manufacture of cheese, yoghurt, meat, and wine. Specifically, their role in taste and structure formation is essential, for instance during cheese ripening. These processes are governed by the presence or absence of strain-specific enzymes (Escobar-Zepeda et al., 2016). Studying such enzymes through strain isolation is often costly and time-consuming since culturing strain representatives is difficult due to laborious or unknown growth conditions (Lagier et al., 2016). Alternatively, these enzymes can be studied by metagenome sequencing, assembly, and annotation, for instance, in product optimization (De Filippis et al., 2017). In addition, metagenome assembly plays an important role in analyzing bacteriophage populations in cultures in terms of their abundance, diversity, and development (Muhammed et al., 2017), which is important not only in the prevention of phage infections that cause fermentation failures, but also for unlocking the potential of phages against food-borne pathogens (Fernández et al., 2017).

Probiotics

Probiotics are microbes that are intended to benefit the host health when consumed in adequate amounts. Identification of novel probiotics is a laborious process that starts with constructing a strain library using a culturomics approach (Lagier et al., 2016). This is followed by *in vitro* and computational research on the obtained strains to functionally characterize them, for instance for their bile resistance and potential to survive the passage of the stomach. Each of these steps reduces the list of high-potential candidates that as a final step must pass regulatory offices such as the European Food Safety Authority (EFSA, FEEDAP et al., 2018). We believe that the findings from comparative studies of the gut microbiome that highlight associations between phenotypic traits such as inflammation (Andoh et al., 2012) and obesity (Kasai et al., 2015) and specific bacterial populations, when integrated with other sources like metabolomic, demographic, dietary, and lifestyle datasets, may allow automated (e.g. machine learning-based) identification of candidate probiotic strains and reduce the time and financial cost of probiotics screening.

Small differences in the gene content of otherwise genetically identical bacterial strains can lead to different phenotypes (Zeevi et al., 2019), which in return may result in different outcomes *in vivo*. Therefore, well-conducted clinical trials are necessary to prove that the probiotic candidate itself confers the health effect. To make sure that the observed effects are not elicited by other (closely related) organisms and can be ascribed solely to the consumed probiotic, metagenomic, and bioinformatic methods that enable strain-level identification and tracking of the studied probiotic strain are required. For instance, in the genus *Bifidobacterium*, genetic differences between different strains of the same species underlie differences in carbohydrate utilization profiles (Arboleya et al., 2018). As these phenotypic traits are important in the development of probiotics for infant nutrition, applying shotgun metagenomics instead of amplicon sequencing for strain-level characterization may have substantial advantages.

Quality Control

Products like probiotics and dairy starter cultures contain live organisms that are either sold directly to consumers or used to manufacture consumer products. Next to the checks performed for raw materials, quality control of the end product is necessary to ensure the presence of correct strains and the absence of pathogens (Fenster et al., 2019). As mentioned above, microbial strains of the same species can have vastly different phenotypes, making strain-level identification in the quality control process crucial for recognizing possible contaminants (Huys et al., 2013). Traditional typing approaches such Random Amplification of Polymorphic DNA-PCR (RAPD-PCR) can be used for identifying single-strain probiotics contaminants, but require cultivation (Mohkam et al., 2016), making them unsuitable for high-throughput screening of products with complex communities (e.g. probiotics and dairy products). Whole-metagenome sequencing and analysis has the potential not only to circumvent these lengthy processes in providing strain-level information, but also to enable screening of undesired traits such as (spore) heat-resistance based on the presence of associated genes (Berendsen et al., 2016).

Cosmetics

The cosmetics industry has a growing interest in studies that aim to explore the skin microbiome as a potential therapeutic target for disorders including acne, eczema, and *Malassezia* folliculitis (Wallen-Russell, 2019). Unfortunately, these studies are typically hampered by the low biomass of skin samples, where small contaminations (e.g. from adjacent skin or reagents) can easily lead to incorrect outcomes (Kong et al., 2017). Furthermore, the human skin microbiome is strongly subject-specific (Zeeuwen et al., 2012), making it difficult to determine the effect of skin products on the general population. While this opens a potential market for personalized skin products, it also raises the need for personal longitudinal studies, where statistical methods such as redundancy analysis and principle response curve (Van den Brink and Braak, 1999) help assess correlations between taxonomic or functional composition and sample characteristics (environmental variables). Furthermore, the data can be corrected for one of the variables, such as 'subject' so that the variance from that covariate is removed before the actual analysis is performed, which facilitates determining the effect of the treatment.

Enzyme Discovery

A wide range of industrial enzymes, such as those used in the production of cleaning agents, laundry detergents, paper, and textile, have the continuous demand to become cheaper, greener, and more efficient. Among others, marine, soil, and lake microbiomes,

with their extremely high and mainly uncharacterized biodiversity, constitute exciting functional mines not only in the search for new enzymes with such desired properties, but also for the discovery of novel enzymes that can catalyze challenging reactions (Popovic et al., 2015). A notable example of the latter is the recent discovery of two enzymes that enable the production of a renewable alternative to toluene, a petrochemical with a market of 29 million tons per year, from complex microbial communities that live in sewages and lakes (Beller et al., 2018).

Two main bioinformatic challenges in metagenomic enzyme discovery arise from the same fact that makes the chosen environment (e.g. soil) attractive in the first place: its high and uncharacterized biodiversity. The large number of different genomes in the environment and their highly skewed abundance distribution make it difficult to obtain contiguous and complete assemblies (Ayling et al., 2019), an outcome that negatively impacts gene prediction. The next challenge lies in functionally annotating the predicted genes, where commonly a high percentage of sequences are labeled as "hypothetical" or with unknown function.

Microbiome-Based Health and Personalized Nutrition

Companies and citizen science projects such as MyMicroZoo[1], Biovis[2], and American Gut[3] offer affordable microbiome analysis services to general consumers. While operationally their analyses are the same as those used for research, they must pay far more attention to the clarity of the results to ensure correct interpretations by the end-users even if the results are stated not to be interpreted as diagnosis. In practice, this means that the end-user should be guided through the (actionable) results with the help of trained healthcare professionals [e.g. dieticians and general practitioners (GPs)], who should take the limitations of a given analysis into account to prevent overinterpretation.

While basing health-related advice on published research findings is a good practice, the fact that most studies focus on a defined cohort and report "averaged" population trends makes it questionable whether results can be translated back to individuals. Such translations to the individual may be less complicated with function-based approaches through metagenomics as the 'personalized' effects are less pronounced in these datasets (Lloyd-Price et al., 2017). Nonetheless, the predictive value of a person's gut microbiome for health was demonstrated by an inspirational study by Zeevi and colleagues (2015), which integrated blood parameters, dietary habits, anthropometrics, physical activity, and the gut microbiome data into a machine learning algorithm that predicted the post meal glycemic responses of the subjects. Ultimately, 72 taxonomic or functional features of the microbiome were included in the predictive model. This approach, validated further with another independent cohort, is now offered to the public by DayTwo[4], which is a good example of how extensive datasets from scientific studies and data science can be combined in an industrial setting for providing customers with evidence-based health-related recommendations.

[1]mymicrozoo.com
[2]biovis-diagnostik.eu
[3]humanfoodproject.com/americangut/
[4]daytwo.com

CURRENT ADVANCES

Metagenome Assembly, Binning, and Annotation

Metagenome assembly enables gene prediction, annotation, and abundance profiling, and therefore is an important computational step when studying the functional composition and capacity of microbiomes. Many (de Bruijn graph-based) metagenome assembly methods that differ in terms of their ease of use, scalability, running time, and memory requirement exist, making it important to carefully choose the one that serves the research question at hand the best (Van der Walt et al., 2017). For instance, in comparative studies with large cohorts where the impact of probiotics on the abundances of gene groups and pathways is analyzed, tools that are computationally less intensive, such as MEGAHIT (Li et al., 2015), are preferred. In contrast, studies with a low number of samples, such as those in enzyme discovery applications, can make use of assembly tools like metaSPAdes (Nurk et al., 2017) that include optimizations such as error correction but with a subsequent runtime trade-off. When higher read depth for assembling low abundance members or recovering full genomes is needed, data from (not too) different samples (e.g. dairy starter cultures) can be combined using co-assembly methods like crass (Dutilh et al., 2012) which also facilitates metagenomic comparison between samples. Finally, binning methods such as MetaBAT (Kang et al., 2015), MaxBin (Wu et al., 2014), and COCACOLA (Lu et al., 2017) facilitate extracting individual (draft) genomes from metagenome assemblies, which helps look at a specific organism in more detail e.g. in enzyme discovery applications where identifying the genome that encodes the target enzyme is important.

In a recent study of cow rumen microbiome, a valuable environment for biomass-degrading enzyme discovery, Stewart et al. (2018) showed that 90% of the proteins predicted to be involved in the studied mechanism (carbohydrate metabolism) did not have a good match in public databases. Such findings highlight the relatively large room for improvement in microbiome annotation.

Hypothesis-Driven Functional Analyses

Exhaustively analyzing all functional aspects and querying all potential longitudinal and cross-sectional aspects of a microbiome dataset is generally considered a hopeless task. Even when computationally feasible, multiple testing issues lead to a severe reduction of the analysis power. Although approaches like the removal of collinear variables and validation of potential correlations in independent datasets can in part address these issues (Falony et al., 2016), delineating the relevant functional aspects is a big step in overcoming these limitations. Using a specific database to answer a particular hypothesis, such as in the case with certain enzyme classes or a set of enzymatic pathways, is such an approach. Examples of such databases

and tools are Resfams (Gibson et al., 2015), dbCAN (Yin et al., 2012), and antiSMASH (Blin et al., 2017), focusing on antibiotic resistance, carbohydrate utilization, and secondary metabolite synthesis, respectively. Methods developed for the elucidation of gene function, such as the guilt by association approaches implemented in STRING (Szklarczyk et al., 2014), can be used to identify genes that are not directly flagged by comparison to specific functional datasets such as the ones described above, but have distribution patterns similar enough to genes that are represented in the reference set. A drawback of functional analyses that require protein sequences is the need for assembly and gene prediction, which can be computationally intensive as described above. Tools like HUMAnN2 (Franzosa et al., 2018) work directly with short-read data without requiring an assembly for profiling protein family abundance.

Assembly-Independent Strain-Level Characterization

Probiotic members such as *Bifidobacterium longum* subsp. *longum* and *Bifidobacterium longum* subsp. *infantis*, which have two distinct phenotypes with relevant functional implications in infant nutrition (Underwood et al., 2015), differ only slightly in their16S rRNA gene sequences (Lawley et al., 2017). Such differences are lost in classical operational taxonomic unit (OTU) clustering-based taxonomic analyses. Novel methods like UNOISE2 (Edgar, 2016) and DADA2 (Callahan et al., 2016) circumvent clustering and apply sequence filtering steps, enabling distinguishing between sequences on a single-nucleotide level by grouping reads in amplicon sequence variants (ASVs). This has a great potential to improve the phylogenetic depth at which microbiome studies can be interpreted. Notable applications of these new algorithms provided new, sub-species level insights into oral (Mukherjee et al., 2018) and vaginal microbiomes (Callahan et al., 2017).

In cases where multiple strains of a species of interest have identical 16S rRNA sequences, algorithms such as StrainPhlAn (Truong et al., 2017) and PanPhlAn (Scholz et al., 2016) enable strain-level analyses from shotgun metagenome datasets without the need for metagenome assembly (**Figure 1**). These methods open the possibility for routine compositional analyses to verify the presence of desired strains or identify potential pathogens in end products.

Long-Read Sequencing and Other Advances

Although their use in microbiome studies is currently not common, long-read sequencing platforms Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) offer exciting opportunities for several industrial applications mentioned above. For instance, circular consensus sequencing application by PacBio, which allows multiple reads generated from a circularized amplicon molecule to be bioinformatically combined into a high-quality, full-length (16S) sequence (Callahan et al., 2018), provides the necessary phylogenetic resolution for applications such as fermentation studies, which is unfeasible with short-read amplicon sequencing. The on-demand sequencing nature of ONT, on the other hand, seems suitable for quality control applications for detecting distinct pathogens, although the high error rate is limiting for accurate, strain-level detection.

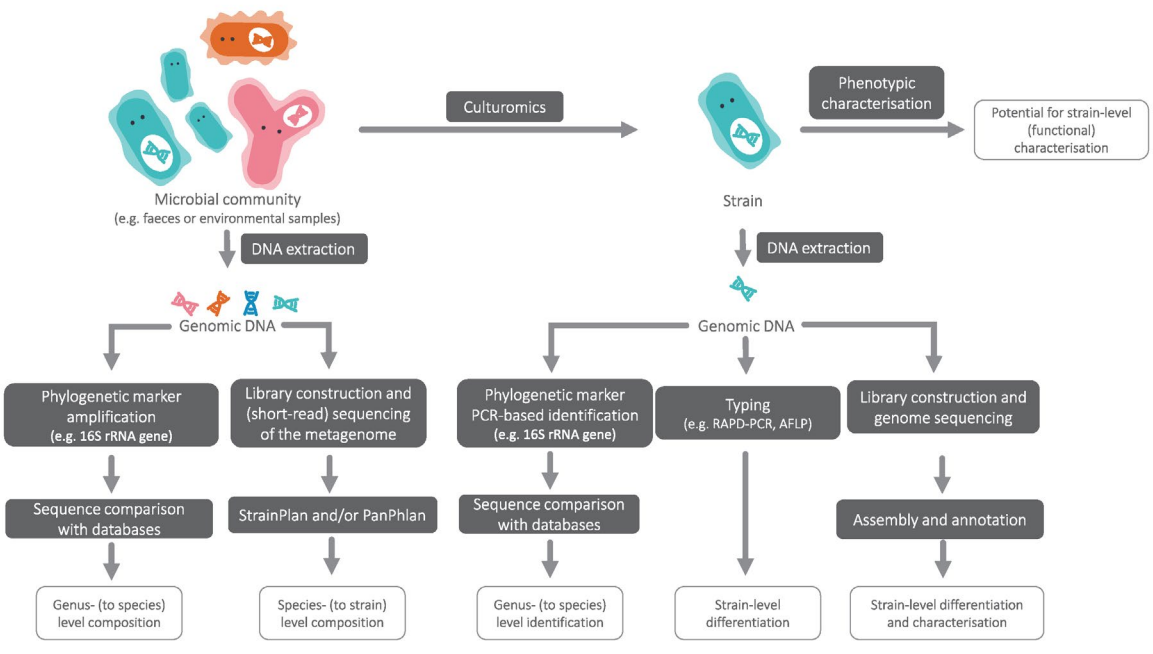

FIGURE 1 | An overview of approaches to achieve taxonomic resolution at different levels.

Even with high dataset coverage and advanced methods, assemblies from short-read datasets commonly remain very fragmented, especially in samples from complex communities like soil. Soon, we expect the integration of long-read sequencing to be more common in assembly-oriented studies for obtaining full, chromosome-level microbial genomes. Correspondingly, we see potential in adapting hybrid assembly methods such as hybridSPAdes (Antipov et al., 2015) to enable their use with long- and short-read metagenome datasets. Other promising developments revolve around using barcoded short reads that have long-range information, such as those provided by 10x Genomics (http://10xgenomics.com), in microbiome research. We see the emergence of tailored bioinformatic methods such as the Athena assembler (Bishara et al., 2018), which uses barcode information in short-reads and improves the contiguity of metagenome assemblies.

Machine Learning and Data Science

With decreasing sequencing costs, the size of datasets in microbiome studies and the depth of sequencing per sample have increased. This led to studies with higher statistical power, and consequently to the transition of OTU tables and functional profiles from end-goal deliverables into starting material for downstream analyses such as machine learning (ML) applications (Pasolli et al., 2016). Methods like random forests (RF) have been successfully used by many within a disease context, for instance, for accurately predicting irritable bowel syndrome (Saulnier et al., 2011) and bacterial vaginosis (Beck and Foster, 2014) based on taxonomic profiles (for a review, see LaPierre et al., 2019 and Qu et al., 2019). On the other hand, Sze and Schloss (2016) used 10 previously published obesity datasets and showed that RF ML models trained on one of the datasets and tested on the remaining nine had a median accuracy of only 56.68%, suggesting that i) the method may not be applicable for some diseases, or ii) the disease signal may be more apparent at the level of differentially expressed functions (gene transcripts) of the microbiome.

Industrial microbiome applications of ML include building classification models based on soil microbiome data for detecting oil sites (Miranda et al., 2019) and above-mentioned personalized health-related lifestyle (diet) recommendation services that are partly based on gut microbiome data. As mentioned in *Probiotics*, we expect dataset integration and ML to have an impact also on areas such as screening of novel probiotics. To meet the overall demand for user-friendly ML in microbiome research, software suites like QIIME 2 (Bokulich et al., 2018), MicrobiomeAnalyst (Dhariwal et al., 2017), and USEARCH (Edgar, 2010) started incorporating ML methods that can be used by researchers who aren't necessarily trained as bioinformaticians.

CONCLUSIONS AND OUTLOOK

The vast number of experimental and computational methods available for microbiome research have led to a broad collection of choices. While creation of guidelines and standardization for increased comparability and reproducibility is essential, achieving a global consensus in methods used remains a challenge. What constrains researchers to their current practices is mainly the laborious nature of adopting other (new) protocols, which may have an ironically detrimental effect on comparability between different studies, or even within studies that run over prolonged periods. Like Knight et al. (2018), we think that a primary objective of microbiome studies should be to standardize the documentation of used methods, tools, data formats, and data processing parameters, and to publish these "logs" next to the final results and interpretations. While complete disclosure is scientifically ideal, it raises commercial concerns for microbiome analysis providers like BaseClear[5], NIZO food research[6], Clinical Microbiomics[7], Vedanta Biosciences[8], and COSMOSID[9], as it would mean releasing a substantial part of their, sometimes unique, intellectual property.

With reducing costs, we soon expect long-read sequencing technologies to be commonly used in microbiome studies, which will benefit from enhanced taxonomic resolution with full-length marker gene sequencing, as well as improved functional analyses thanks to more contiguous metagenome assemblies. Here, the focus in developments is likely to be on the translation of bioinformatic protocols already established for short reads to long-read versions, for instance in denoising and read classification approaches.

Other challenges relate to shotgun metagenome analyses in large studies, where expensive calculations used in *de novo* assembly and annotation may result in capacity issues. For companies that cannot afford large on-premise compute infrastructures, the cloud provides a flexible alternative, where know-how of cloud-computing becomes essential.

Finally, the rapid translation of microbiome research into important industrial applications in healthcare, energy, and food production will continue to stimulate collaborations between academic and industrial communities. We expect the role of bioinformatics and data science to become only more significant in this relationship.

AUTHOR CONTRIBUTIONS

All authors were involved in the writing and final preparation of the article.

ACKNOWLEDGMENTS

The authors thank Thomas Battaglia for his help in the preparation of the final manuscript.

[5]baseclear.com
[6]nizo.com
[7]clinical-microbiomics.com
[8]vedantabio.com
[9]cosmosid.com

REFERENCES

(FEEDAP), E., Rychen, G., Aquilina, G., Azimonti, G., Bampidis, V., Bastos, M., et al. (2018). Guidance on the characterisation of microorganisms used as feed additives or as production organisms. *EFSA J.* 16 (3), e05206. doi: 10.2903/j.efsa.2018.5206

Andoh, H., Kizuoka, H., Tsujikawa, T, Nakamura, S, Hirai, F., Suzuki, Y., et al. (2012). Multicenter analysis of fecal microbiota profiles in Japanese patients with Crohn's disease. *J. Gastroenterol.* 47 (12), 1298–1307. doi: 10.1007/s00535-012-0605-0

Antipov, D., Korobeynikov, A., McLean, J., and Pevzner, P. (2015). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32 (7), 1009–1015. doi: 10.1093/bioinformatics/btv688

Arboleya, S., Bottacini, F., O'Connell-Motherway, M., Ryan, C., Ross, R., Van Sinderen, D., et al. (2018). Gene-trait matching across the Bifidobacterium longum pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. *BMC Genomics* 19 (1), 33. doi: 10.1186/s12864-017-4388-9

Ayling, M., Clark, M., and Leggett, R. (2019). New approaches for metagenome assembly with short reads. *Brief Bioinform.* 1–11. doi: 10.1093/bib/bbz020

Beck, D., and Foster, J. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PloS One* 9 (2), e87830. doi: 10.1371/journal.pone.0087830

Beller, H., Rodrigues, A., Zargar, K., Wu, Y.-W., Saini, A., Saville, R., et al. (2018). Discovery of enzymes for toluene synthesis from anoxic microbial communities. *Nat. Chem. Biol.* 14 (5), 451. doi: 10.1038/s41589-018-0017-4

Berendsen, E., Boekhorst, J., Kuipers, O., and Wells-Bennik, M. (2016). A mobile genetic element profoundly increases heat resistance of bacterial spores. *ISME J.* 10 (11), 2633. doi: 10.1038/ismej.2016.59

Bishara, A., Moss, E., Kolmogorov, M., Parada, A., Weng, Z., Sidow, A., et al. (2018). High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* 36, 1067–1075. doi: 10.1038/nbt.4266

Blin, K., Wolf, T., Chevrette, M., Lu, X., Schwalen, C., Kautsar, S., et al. (2017). antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* 45 (W1), W36–W41. doi: 10.1093/nar/gkx319

Bokulich, N., Dillon, M., Bolyen, E., Kaehler, B., Huttley, G., and Caporaso, J. (2018). q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Source Softw.* 3 (30), 934. doi: 10.21105/joss.00934

Callahan, B., DiGiulio, D., Goltsman, D., Sun, C., Costello, E., Jeganathan, P., et al. (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc. Natl. Acad. Sci* 114 (37), 9966–9971. doi: 10.1073/pnas.1705899114

Callahan, B., McMurdie, P., Rosen, M., Han, A., Johnson, A., and Holmes, S. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13 (7), 581. doi: 10.1038/nmeth.3869

Callahan, B., Wong, J., Heiner, C., Oh, S., Theriot, C., Gulati, A., et al. (2018). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* gkz569. https://doi.org/10.1093/nar/gkz569

Costessi, A., van den Bogert, B., May, A., Ver Loren van Themaat, E., Roubos, J., Kolkman, M., et al. (2018). Novel sequencing technologies to support industrial biotechnology. *FEMS Microbiol. Letters* 365 (16), fny103. doi: 10.1093/femsle/fny103

De Filippis, F., Parente, E., and Ercolini, D. (2017). Metagenomics insights into food fermentations. *Microb. Biotechnol.* 10 (1), 91–102. doi: 10.1111/1751-7915.12421

Dhariwal, A., Chong, J., Habib, S., King, I., Agellon, L., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45 (W1), W180–W188. doi: 10.1093/nar/gkx295

Dutilh, B., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R., et al. (2012). Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* 28 (24), 3225–3231. doi: 10.1093/bioinformatics/bts613

Edgar, R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26 (19), 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgar, R. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv* 081257. doi: 10.1101/081257

Escobar-Zepeda, A., Sanchez-Flores, A., and Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol.* 57, 116–127. doi: 10.1016/j.fm.2016.02.004

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science* 352 (6285), 560–564. doi: 10.1126/science.aad3503

Fenster, K., Freeburg, B., Hollard, C., Wong, C., Rønhave Laursen, R., and Ouwehand, A. (2019). The production and delivery of probiotics: a review of a practical Approach. *Microorganisms* 7 (3), 83. doi: 10.3390/microorganisms7030083

Fernández, L., Escobedo, S., Gutiérrez, D., Portilla, S., Martínez, B., García, P., et al. (2017). Bacteriophages in the dairy environment: from enemies to allies. *Antibiotics* 6 (4), 27. doi: 10.3390/antibiotics6040027

Franzosa, E., McIver, L., Rahnavard, G., Thompson, L., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15 (11), 962. doi: 10.1038/s41592-018-0176-y

Gibson, M., Forsberg, K., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* 9 (1), 207. doi: 10.1038/ismej.2014.106

Howe, A., Jansson, J., Malfatti, S., Tringe, S., Tiedje, J., and Brown, C. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci.* 111 (13), 4904–4909. doi: 10.1073/pnas.1402564111

Huys, G., Botteldoorn, N., Delvigne, F., De Vuyst, L., Heyndrickx, M., Pot, B., et al. (2013). Microbial characterization of probiotics–Advisory report of the Working Group "8651 Probiotics" of the Belgian Superior Health Council (SHC). *Mol. Nutr. Food Res.* 57 (8), 1479–1504. doi: 10.1002/mnfr.201300065

Kang, D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3, e1165. doi: 10.7717/peerj.1165

Kasai, C., Sugimoto, K., Moritani, I., Tanaka, J., Oya, Y., Inoue, H., et al. (2015). Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC Gastroenterol.* 15 (1), 100. doi: 10.1186/s12876-015-0330-2

Knight, R., Vrbanac, A., Taylor, B., Aksenov, A., Callewaert, C., Debelius, J., et al. (2018). Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 4 (16), 410–422. doi: 10.1038/s41579-018-0029-9

Kong, H., Andersson, B., Clavel, T., Common, J., Jackson, S., Olson, N., et al. (2017). Performing skin microbiome research: a method to the madness. *J. Investig. Dermatol.* 137 (3), 561–568. doi: 10.1016/j.jid.2016.10.033

Lagier, J.-C., Khelaifia, S., Alou, M., Ndongo, S., Dione, N., Hugon, P., et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat. Microbiol.* 1 (12), 16203. doi: 10.1038/nmicrobiol.2016.203

LaPierre, N., Ju, C.-T., Zhou, G., and Wang, W. (2019). MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction. *Methods.* doi: 10.1016/j.ymeth.2019.03.003

Lawley, B., Munro, K., Hughes, A., Hodgkinson, A., Prosser, C., Lowry, D., et al. (2017). Differentiation of Bifidobacterium longum subspecies longum and infantis by quantitative PCR using functional gene targets. *PeerJ.* 5, e3375. doi: 10.7717/peerj.3375

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 (10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550 (7674), 61. doi: 10.1038/nature23889

Lu, Y., Chen, T., Fuhrman, J., and Sun, F. (2017). COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics* 33 (6), 791–798. doi: 10.1093/bioinformatics/btw290

McFarland, L., Evans, C., and Goldstein, E. (2018). Strain-specificity and disease-specificity of probiotic efficacy: a systematic review and meta-analysis. *Front. Med.* 5, 124. doi: 10.3389/fmed.2018.00124

Miranda, J., Seoane, J., Esteban, A., and Espí, E. (2019). *Microbial Exploration Techniques: An Offshore Case Study.* Eds. J. Miranda, J. Seoane, A. and Esteban, E. Espí. Boca Raton, Florida: CRC Press

Mohkam, M., Nezafat, N., Berenjian, A., Mobasher, M., and Ghasemi, Y. (2016). Identification of Bacillus probiotics isolated from soil rhizosphere using 16S rRNA, recA, rpoB gene sequencing and RAPD-PCR. *Probiotics Antimicrob. Proteins* 8 (1), 8-18. doi: 10.1007/s12602-016-9208-z

Muhammed, M., Kot, W., Neve, H., Mahony, J., Castro-Mejía, J., Krych, L., et al. (2017). Metagenomic analysis of dairy bacteriophages: extraction method and pilot study on whey samples derived from using undefined and defined mesophilic starter cultures. *Appl. Environ. Microbiol.* 83 (19), e00888–e00817. doi: 10.1128/AEM.00888-17

Mukherjee, C., Beall, C., Griffen, A., and Leys, E. (2018). High-resolution ISR amplicon sequencing reveals personalized oral microbiome. *Microbiome* 6 (1), 153. doi: 10.1186/s40168-018-0535-z

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27 (5), 824–834. doi: 10.1101/gr.213959.116

Pasolli, E., Truong, D., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12 (7), e1004977. doi: 10.1371/journal.pcbi.1004977

Popovic, A., Tchigvintsev, A., Tran, H., Chernikova, T., Golyshina, O., Yakimov, M., et al. (2015). *Metagenomics as a tool for enzyme discovery: hydrolytic enzymes from marine-related metagenomes*. Eds. A. Popovic, A. Tchigvintsev, H. Tran, T. Chernikova, O. Golyshina, and M. Yakimov. Basel, Switzerland: Springer. doi: 10.1007/978-3-319-23603-2_1

Qu, K., Guo, F., Liu, X., Lin, Y., and Zou, Q. (2019). Application of Machine Learning in Microbiology. *Front. Microbiol.* 10, 827. doi: 10.3389/fmicb.2019.00827

Saulnier, D., Riehle, K., Mistretta, T.–A., Diaz, M.–A., Mandal, D., Raza, S., et al. (2011). Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterol.* 141 (5), 1782–1791. doi: 10.1053/j.gastro.2011.06.072

Scholz, M., Ward, D., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13 (5), 435. doi: 10.1038/nmeth.3802

Singh, R., Kumar, M., Mittal, A., and Mehta, P. (2016). Microbial enzymes: industrial progress in 21st century. *3 Biotech.* 6 (2), 174. doi: 10.1007/s13205-016-0485-8

Stewart, R., Auffret, M., Warr, A., Wiser, A., Press, M., Langford, K., et al. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* 9 (1), 870. doi: 10.1038/s41467-018-03317-6

Sze, M., and Schloss, P. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *MBio.* 7 (4), e01018–e01016. doi: 10.1128/mBio.01018-16

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43 (D1), D447–D452. doi: 10.1093/nar/gku1003

Truong, D., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27 (4), 626–638. doi: 10.1101/gr.216242.116

Underwood, M., German, J., Lebrilla, C., and Mills, D. (2015). Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatr. Res.* 77 (1-2), 229. doi: 10.1038/pr.2014.156

Van den Brink, P., and Braak, C. (1999). Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. *Environ. Toxicol. Chem. Int J.* 18 (2), 138–148. doi: 10.1002/etc.5620180207

Van der Walt, A., Van Goethem, M., Ramond, J.-B., Makhalanyane, T., Reva, O., and Cowan, D. (2017). Assembling metagenomes, one community at a time. *BMC Genomics* 18 (1), 521. doi: 10.1186/s12864-017-3918-9

Wallen-Russell, C. (2019). The Role of Every-Day Cosmetics in Altering the Skin Microbiome: a Study Using Biodiversity. *Cosmetics* 6 (1), 2. doi: 10.3390/cosmetics6010002

Wu, Y.-W., Tang, Y.-H., Tringe, S., Simmons, B., and Singer, S. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome.* 2 (1), 26. doi: 10.1186/2049-2618-2-26

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 40 (W1), W445–W451. doi: 10.1093/nar/gks479

Zeeuwen, P., Boekhorst, J., van den Bogaard, E., de Koning, H., van de Kerkhof, P., Saulnier, D., et al. (2012). Microbiome dynamics of human epidermis following skin barrier disruption. *Genome Biol.* 13 (11), R101. doi: 10.1186/gb-2012-13-11-r101

Zeevi, D., Korem, T., Godneva, A., Bar, N., Kurilshikov, A., Lotan-Pompan, M., et al. (2019). Structural variation in the gut microbiome associates with host health. *Nature* 1, 43–48. doi: 10.1038/s41586-019-1065-y

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., et al. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* 163 (5), 1079–1094. doi: 10.1016/j.cell.2015.11.001

Conflict of Interest Statement: BB, WP, and AM work at BaseClear. JB works at NIZO Food Research.