frontiers in Microbiology

TECHNOLOGY AND CODE
published: 23 March 2021
doi: 10.3389/fmicb.2021.644861

Forensic Microbiome Database: A Tool for Forensic Geolocation Meta-Analysis Using Publicly Available 16S rRNA Microbiome Sequencing

Harinder Singh[1†], Thomas Clarke[1†], Lauren Brinkac[2], Chris Greco[3] and Karen E. Nelson[1]*

[1]J. Craig Venter Institute, Rockville, MD, United States, [2]Noblis, Reston, VA, United States, [3]GeneDX, Gaithersburg, MD, United States

OPEN ACCESS

Edited by:
Nikos Kyrpides, Lawrence Berkeley National Laboratory, United States

Reviewed by:
Kostas Konstantinidis, Georgia Institute of Technology, United States
Thorsten Stoeck, University of Kaiserslautern, Germany
Ilias Lagkouvardos, Technical University of Munich, Germany

***Correspondence:**
Harinder Singh
hsingh@jcvi.org

[†]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: *22 December 2020*
Accepted: *03 March 2021*
Published: *23 March 2021*

Citation:
Singh H, Clarke T, Brinkac L, Greco C and Nelson KE (2021) Forensic Microbiome Database: A Tool for Forensic Geolocation Meta-Analysis Using Publicly Available 16S rRNA Microbiome Sequencing. Front. Microbiol. 12:644861. doi: 10.3389/fmicb.2021.644861

The human microbiome has been proposed as a tool to investigate different forensic questions, including for the identification of multiple personal information. However, the fragmented state of the publicly available data has retarded the development of analysis techniques and, therefore, the implementation of microbiomes as a forensic tool. To address this, we introduce the forensic microbiome database (FMD), which is a collection of 16S rRNA data and associated metadata generated from publicly available data. The raw data was further normalized and processed using a pipeline to create a standardized data set for downstream analysis. We present a website allowing for the exploration of geolocation signals in the FMD. The website allows users to investigate the taxonomic differences between microbiomes harvested from different locations and to predict the geolocation of their data based on the FMD sequences. All the results are presented in dynamic graphics to allow for a rapid and intuitive investigation of the taxonomic distributions underpinning the geolocation signals and prediction between locations. Apart from the forensic aspect, the database also allows exploration and comparison of microbiome samples from different geolocation and between different body sites. The goal of the FMD is to provide the scientific and non-scientific communities with data and tools to explore the possibilities of microbiomes to answer forensic questions and serve as a model for any future such databases.[1]

Keywords: microbiome, forensic, geolocation, 16S rRNA, database

INTRODUCTION

Advances in the depth of DNA sequencing over the last couple of decades, labeled as next generation sequencing (NGS), has greatly expanded the knowledge of the diversity of bacteria living on or within humans (microbiomes). Examinations of human microbiomes *via* multiple methods, including directed sequencing of 16S ribosomes (rDNA genes), allow for an estimation of the taxonomic diversity and the distribution of the contributory bacterial species. Experiments have demonstrated that human microbiomes are constantly interfacing with external microbiomes,

[1]Database URL: http://fmd.jcvi.org

both from other people, animals (Song et al., 2013; Misic et al., 2015), and from environments (Flores et al., 2011; Hewitt et al., 2012; Luongo et al., 2017). Studies have also demonstrated that the species makeup of human microbiomes is partially shaped by personal factors, including age (Odamaki et al., 2016), diet (De Filippo et al., 2010; Yatsunenko et al., 2012; David et al., 2014), habits (Moon et al., 2015; Wu et al., 2016), disease state (Peters et al., 2016), and geolocation (Yatsunenko, et al., 2012; Zhang et al., 2015; Lund et al., 2017; Brinkac et al., 2018), with the location on the body the strongest determinant (Human Microbiome Project, 2012). As such, this ability to capture and leverage these differences in the human microbiome presents exciting new possibilities for forensic science (Clarke et al., 2017; Hampton-Marcell et al., 2017), including the possibility of linking specific human subjects to objects and locations in the crime scene (Lax et al., 2015) and determining the country of origin for different samples.

Personal identification using microbial biosignatures is still an emerging field, and additional work is necessary for it to become highly effective in forensic science as would be required to be judicially acceptable as evidence. Single sample studies, while sufficient to identify differences between individuals along with a forensic question, are often too restricted in size and scope, such as only addressing one location or one metadata variable. Likewise, though the number of available microbiome samples are rapidly increasing, the diversity of sampling techniques and a lack of uniformity in reporting the metadata associated with the data retards the attempts to use this data in a meta-analysis.

We have addressed these limitations through the creation of a new database with an associated website that collects and collates publicly available microbiome datasets. The database is populated with ~20,000 human 16S rRNA NGS samples from multiple body sites from various public repositories, which have been subsequently processed using a single pipeline. Apart from sequences, we also capture the metadata associated with the samples including geolocation, healthy or non-healthy status, and other variables. The associated website allows users to compare microbiomes from different geographic locations and body sites, as well as to upload data that can be compared to microbiomes in the database and for which the geolocation of the sample can be predicted (**Figure 1**). The results from these analyses are provided in dynamic visualizations, which show the taxonomic distribution underpinning the analyses and how the individual samples compare to each other.

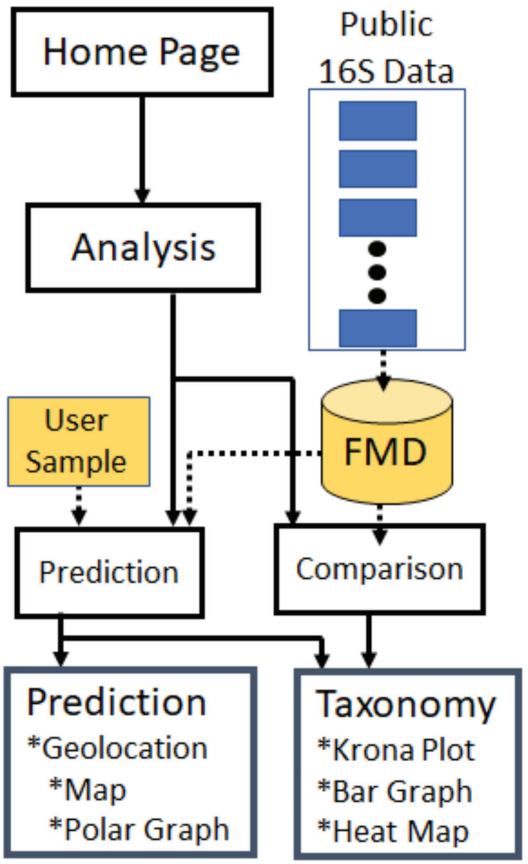


FIGURE 1 | Forensic microbiome database (FMD) and Website Flowchart. Publicly available 16S rRNA sequences were collected from multiple sites, processed, and deposited in the FMD database (detailed in **Figure 2**). This data can be explored through the FMD website, where it is used to predict the geolocation of user-provided samples (see also **Figure 4**) and for comparisons of the microbial taxonomic distribution in different geographic and body sites (see also **Figure 3**).

DATABASE CONSTRUCTION

Data Collection

We surveyed the literature and public repositories for microbiome studies based on 16S rRNA sequencing. Only projects with 16S rRNA sequences sampled from humans with the sampled body site, geographic location, and publishing work all documented were included. Samples of raw sequencing data for each project were downloaded along with available metadata from publicly available databases, including NCBI SRA, EBI, and MG-RAST. The databases include buccal mucosa and stool samples recently collected as part of the forensic microbiome database (FMD) project (PRJNA545251) from adult females (18–26), born and currently living in Barbados (n = 32), Santiago (n = 32), Pretoria (n = 37), and Bangkok (n = 60), and described more fully in Clarke et al., *submitted*. Additional metadata values, including age, gender, and healthy/non-healthy status, were used when available either in the public database or in the citing manuscript. The data was processed with the JCVI pipeline based on UPARSE and SILVA database. Diseases such as IBS and Crohn's disease can have a significant effect on the microbiome (Carroll et al., 2012; Morgan et al., 2012; Zhou et al., 2018). Since disease states can markedly change the microbiome comparison, only samples not explicitly labeled with a disease state are included.

The 16S rRNA Pipeline

Each project was processed separately, and operational taxonomic units (OTUs) were generated *de novo* from raw 454 or Illumina

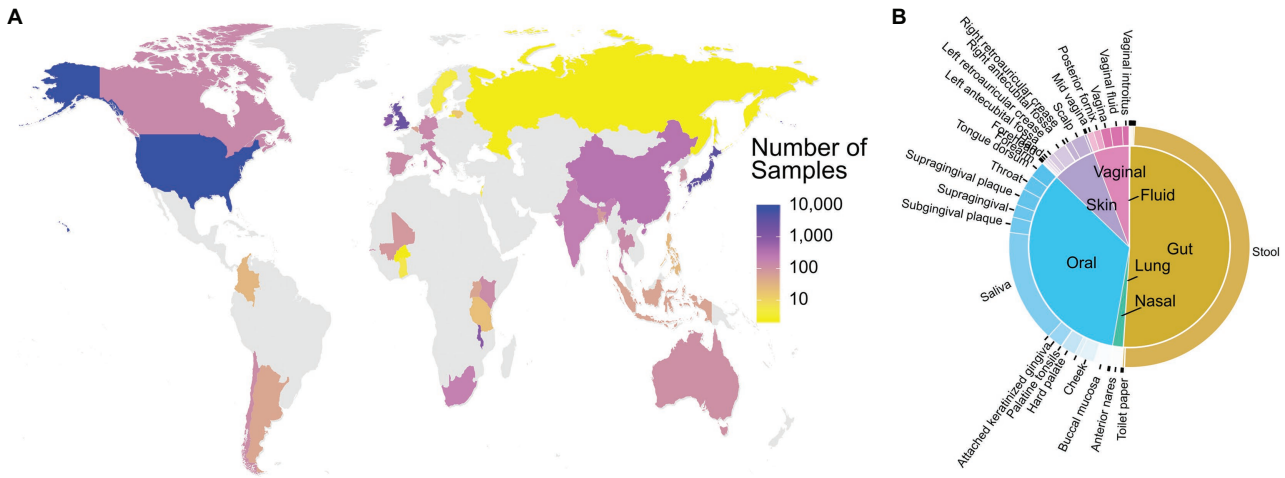


FIGURE 2 | Data Availability in the FMD. The number of samples in the FMD from different countries **(A)** and body sites **(B)**. Only samples from healthy individuals are shown.

sequence reads using the UPARSE pipeline (Edgar, 2013). Paired-end reads were trimmed from the adapter sequences, barcodes, and primers prior to assembly. Sequences of low quality and singletons were discarded. Sequences were further subjected to de-replication and chimera filtering during clustering. Mothur (Schloss et al., 2009) was used to report full taxonomies with 100 iterations for the wang classifier (iters = 100) wand, only including sequences where 80 or more of the 100 iterations are reporting similar assignment (cutoff = 80). The RDP classifier in mothur and version 123 of the SILVA 16S ribosomal RNA database (Quast et al., 2013) were used for the taxonomy assignment of OTUs. Rare OTUs or taxa are strongly affected by sequencing errors, and statistical conclusions relying on them are typically unstable (He et al., 2015). The OTUs with less than 10 total reads in each project dataset were considered rare OTUs using the phyloseq (McMurdie and Holmes, 2013) package in R and were removed along with OTUs that were either unknown or unclassed at the genera level. Quality control was also performed on all samples, and the OTUs with samples containing more than 20% of their reads in unknown or unclassified genera or less than 2,000 reads were removed (Amir et al., 2017; Singh et al., 2017). We removed these samples because OTUs with no genera classification will introduce biases in the composition plots, average calculation and impact the prediction module. The trimmed samples were then normalized to their proportion of reads in each OTU and combined into a master OTU table using the phyloseq merge function. All of the microbiome data present in the FMD database are at the genus level. The phyloseq tax_glom function to merge the same genera into one single genera in each separate project was used.

Database Architecture

Forensic microbiome database is built on Apache HTTP server 2.2 with MySQL server 5.1.47 as the back end and PHP 5.2.9, HTML, and JavaScript as the front end.

Database Summary

The current version of FMD has 20,820 samples from 95 projects with 79 PubMed references. These 96 projects contain 16S rRNA data obtained from 54 different body sites of individuals from 35 different countries, 91 states, provinces or equivalent, and 138 cities. The samples in the database are highly concentrated in developed countries, with the United States (9,492 samples) the most significant contributor in the FMD, followed by Japan (4,054 samples) and the United Kingdom (2,722 samples; **Figure 2A**). The majority of 16S rRNA data (~50%) was obtained from stool samples, followed by saliva and other oral locations (**Figure 2B**). Detailed descriptions of the included data are available at http://fmd.jcvi.org/stat.php.

Web Interface

The FMD website contains two separates but connected modules. The first allows the user to explore the loaded 16S rRNA data and compare various geolocation and body sites using the processed and loaded data described above. To explore the FMD data, a user can compare the taxonomic abundance profile of individuals' microbiomes from multiple geolocations and body sites. The first option is a bar plot of the top twenty abundant genera (ranked based on the first selected geolocation in the query) of the selected geolocations (**Figure 3A**). The second option is the Krona charts of all the selected geolocations (Ondov et al., 2011), with individual Krona charts available for full-screen visualization (**Figure 3B**). The final option is a heatmap showing the relative abundances of the top ten most abundant genera, ranked similarly to the bar plots, of all the selected geolocations.

In the second module, users can upload their processed microbiome data to predict the potential geolocation of the user-provided data and compare it with any existing geolocation data in FMD. To geolocate the user sample, it is compared

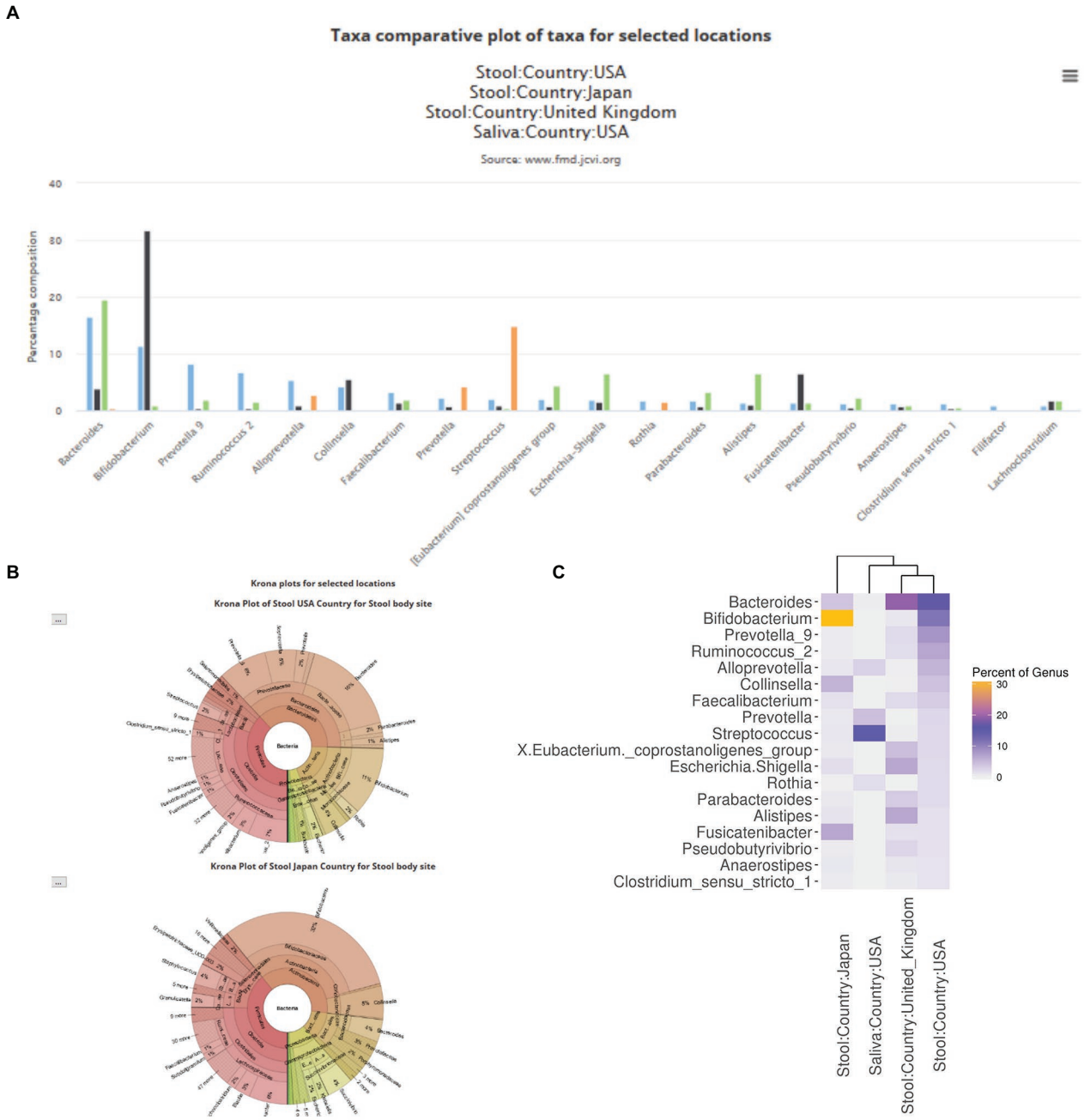
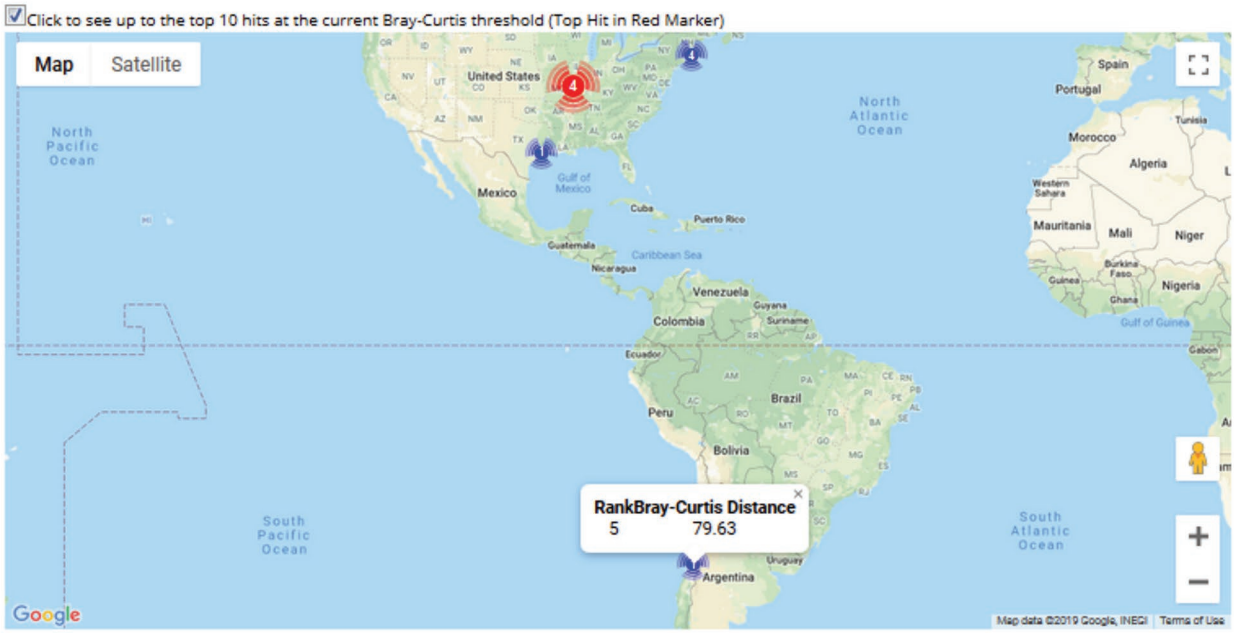


FIGURE 3 | Examples of comparative taxa abundance module in the FMD. The FMD allows for a visual comparison of the taxa abundance of different body sites in different geolocations, which is shown *via* bar charts **(A)**, Krona plots **(B)**, and heatmaps **(C)**. Stool samples from the United States, Japan, and the United Kingdom, as well as saliva samples from the United States, are shown.

against all the samples present in the FMD using the Bray-Curtis distance matrix score, and the results are ranked. The results are visually explorable by both locations and by sample. First, the page displays a world map showing the location and counts of the high-ranked matches, which the site with the top match in a different color (**Figure 4A**). The distance between the user sample and the high matching samples in the site can be displayed by mousing over the respective site. A second tab contains a sample-level visualization of the Bray-Curtis distances between the user and the FMD samples with distances less than the cutoff points on a polar graph (**Figure 4B**). The cutoff distance for displayed values can be altered to examine as many sites as desired. The FMD-sample points are colored by metadata values, beginning with geolocation but changeable

A



B

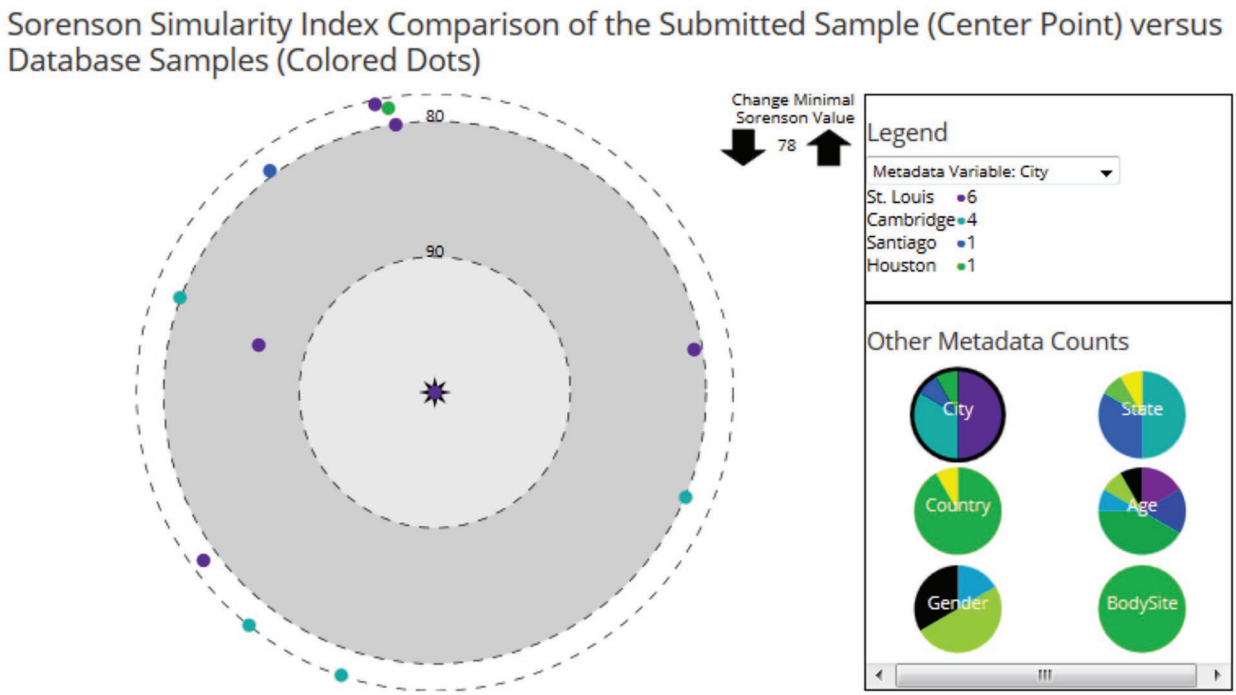


FIGURE 4 | Examples of user data geo-location prediction in the FMD. Geolocation of user-uploaded data is predicted by finding the taxonomically closest FMD samples, which are shown both on a world map, with the number of hits per city shown, **(A)** and as individual samples on a polar graph with their similarity to the sample indicated by the distance to the center **(B)**. 16S rRNA microbiomes obtained from stool samples of an individual residing in St. Louis (Sample ID: SRS015854) is shown here.

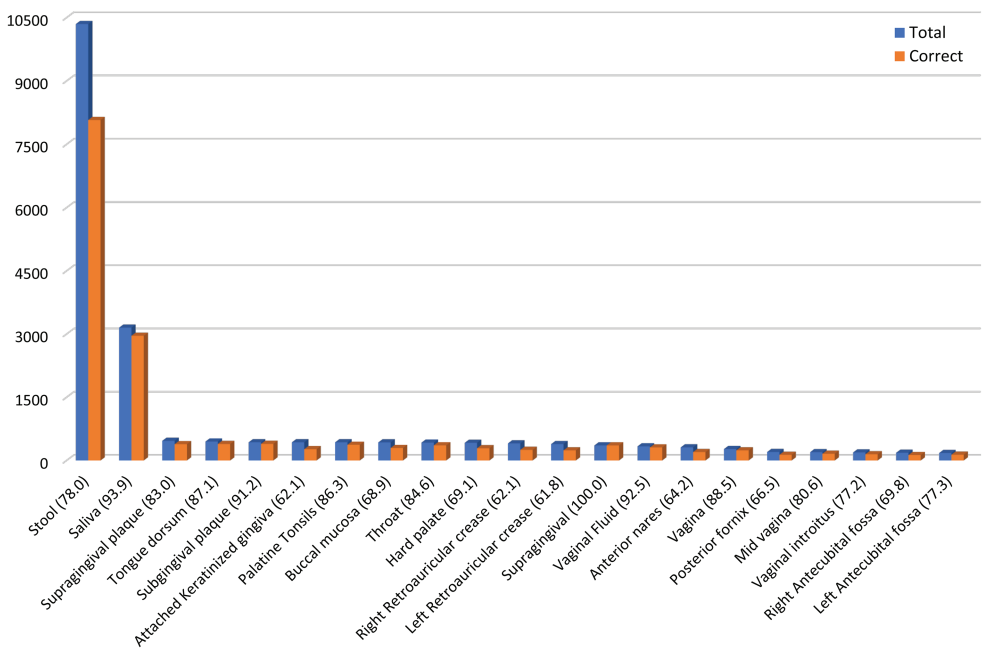


FIGURE 5 | Similarity-based prediction performance of different body sites. A leave-one-out cross-validation of the data was performed using a similarity-based prediction approach. The results for body sites with more than 150 samples are shown. The bar graph shows the total number of samples for a body site and the number of samples with correct geolocation prediction.

to age, gender, and body site. The percentage of samples with distances above the cutoff with different values for each metadata variable is also shown. The taxonomic distribution user-submitted samples can also be visualized similarly to the FMD database samples, either with Krona charts or compared with any geographic site data present in the FMD, represented as the average of all the samples of that particular site, as a bar chart and heatmap. A detailed description of the website's usage can be obtained using the user manual available on the website.[2]

To better understand the similarity-based approach's prediction capability, we performed the leave-one-out cross-validation to estimate the prediction module's performance. Body sites with more than 150 samples in the database were considered, which constitute 96% of the data. As observed in **Figure 5**, the overall accuracy is 80.5% for cities, 81.5% for state/region, and 92.1% for countries. The accuracy ranges from 61% for retroauricular crease to 93% for saliva samples (**Figure 5**). We were able to achieve 78% prediction accuracy for stool samples, which constitute half of the samples collected from all around the world. Further, we explored the impact of similar body sites on the prediction module performance in **Supplementary Figure S1**. We observed that similar body sites are cross-predicted, i.e., the supragingival samples can be predicted as Subgingival plaque samples and vice versa. There is negligible cross prediction between the oral cavity, skin, vagina, and stool samples which validates the unique microbiome composition of different body sites. Next, we analyzed the remaining incorrect 20.5% samples to understand the impact of distance on incorrect predictions. In the case of incorrectly predicted vagina samples which constitute 13% of all vagina samples, the average distance is ~7,000 km. On average, the incorrect prediction has ~1,000 km distance (**Supplementary Figure S2**). We examine the incorrect vagina samples that were predicted as stool samples are dominated by the same genus, which suggests either cross-contamination or biological/technical contamination, which explains the considerable variation in incorrect samples' distance. When we remove the samples where a single genus is more than 60% of microbiome composition, only eight vagina samples were predicted as a stool instead of 35 wrong predictions (**Supplementary Figure S3**).

SUMMARY

Numerous studies have identified microbiomes' potential to be a valuable forensic investigatory tool, but the translation of these findings into legally actionable information remains incomplete. The development of these tools is hindered by multiple limitations of analyzing the data, such as the diversity of formats in which the information is available, the absence of sufficient metadata, and the large amount of data required to generate any tools. The database introduced here begins to

[2]http://fmd.jcvi.org/help.php

address these limitations and can form the backbone for future explorations and generation of a novel technique to tease apart the signals within microbiomes to detect forensic information. The database and the website will facilitate exploration of the taxonomic underpinnings of geolocation signals, both through dynamic explorations of the taxonomic distributions of microbiomes from different geographic locations through comparisons of the data samples in combination with user-supplied metadata. The key limitation of the database is the unavailability of data from many African and Middle east countries apart from few countries from each continent. Since we considered only good quality microbiome data that was not explicitly labeled with a disease state; we were limited to data availability. We hope that in the future, additional data from these regions will be available from the public database and will be added to the FMD database.

The FMD is designed for rapid and intuitable exploration of geolocation signals in the microbiomes using well-documented and computationally inexpensive algorithms. Currently, the database only uses 16S rRNA sequences for the geolocation analysis, and while metagenomic whole genomic sequencing of microbiomes are a rapidly expanding field (Schmedes et al., 2017; Almeida et al., 2019), the analytical tools available to distinguish the geo-position of metagenomes are not as developed. Likewise, machine learning and other reduced taxonomic comparisons are emerging tools for dissecting taxonomic distributions and looking at forensic questions (Johnson et al., 2016; Sarkar et al., 2017), but these have yet to be adapted for a global analysis.

As the state of forensic analysis of microbiomes continues to develop, the FMD is well adapted to address some of the remaining outstanding issues. As previously documented, the body site sampled remains the primary determinant of the taxonomic distribution differences of microbiomes, and multiple body sites have been shown to have a geographic-specific signal (Zhang et al., 2015; Sarkar et al., 2017; Brinkac et al., 2018). By collecting samples from multiple body sites, the FMD currently allows for comparison of the geolocation signals. While current analysis suggests that this signal is not additive across body sites, future analytical techniques might have an amplification effect across body sites that the FMD would capture. Additionally, both the raw information and subsequent analyses can be highly complex and not readily digestible by non-specialists. As the analysis tools increase in complexity, we believe that the summary dynamics figures used by the FMD provide a useful example of how to engage a non-specialist in an analytical examination of the results.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: http://fmd.jcvi.org/stat.php and http://fmd.jcvi.org/bioproject/PRJNA545251.

AUTHOR CONTRIBUTIONS

HS and TC generated the data, designed/developed the database, and wrote the manuscript. LB curated the data, participated in manuscript writing, and designed the database and project. CG generated the data. KN participated in design and implementation of the project. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by Award No. 2015-R2-CX-K036 awarded by the Office of Justice Programs; National Institute of Justice, Department of Justice. The opinions, findings, and conclusion or recommendations expressed of the project are those of the author(s) and do not necessarily reflect the views of the Department of Justice or the grant-making component.

ACKNOWLEDGMENTS

We would like to acknowledge Matt LaPointe for his assistance with the website's FMD logo design.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.644861/full#supplementary-material

REFERENCES

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., et al. (2019). A new genomic blueprint of the human gut microbiota. *Nature* 568, 499–504. doi: 10.1038/s41586-019-0965-1

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e00191-16. doi: 10.1128/mSystems.00191-16

Brinkac, L., Clarke, T. H., Singh, H., Greco, C., Gomez, A., Torralba, M. G., et al. (2018). Spatial and environmental variation of the human hair microbiota. *Sci. Rep.* 8:9017. doi: 10.1038/s41598-018-27100-1

Carroll, I. M., Ringel-Kulka, T., Siddle, J. P., and Ringel, Y. (2012). Alterations in composition and diversity of the intestinal microbiota in patients with diarrhea-predominant irritable bowel syndrome. *Neurogastroenterol. Motil.* 24, 521.e248-530.e248. doi: 10.1111/j.1365-2982.2012.01891.x

Clarke, T. H., Gomez, A., Singh, H., Nelson, K. E., and Brinkac, L. M. (2017). Integrating the microbiome as a resource in the forensics toolkit. *Forensic Sci. Int. Genet.* 30, 141–147. doi: 10.1016/j.fsigen.2017.06.008

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

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

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Flores, G. E., Bates, S. T., Knights, D., Lauber, C. L., Stombaugh, J., Knight, R., et al. (2011). Microbial biogeography of public restroom surfaces. *PLoS One* 6:e28132. doi: 10.1371/journal.pone.0028132

Hampton-Marcell, J. T., Lopez, J. V., and Gilbert, J. A. (2017). The human microbiome: an emerging tool in forensics. *Microb. Biotechnol.* 10, 228–230. doi: 10.1111/1751-7915.12699

He, Y., Caporaso, J. G., Jiang, X. T., Sheng, H. F., Huse, S. M., Rideout, J. R., et al. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 3:20. doi: 10.1186/s40168-015-0081-x

Hewitt, K. M., Gerba, C. P., Maxwell, S. L., and Kelley, S. T. (2012). Office space bacterial abundance and diversity in three metropolitan areas. *PLoS One* 7:e37849. doi: 10.1371/journal.pone.0037849

Human Microbiome Project C (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Johnson, H. R., Trinidad, D. D., Guzman, S., Khan, Z., Parziale, J. V., DeBruyn, J. M., et al. (2016). A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS One* 11:e0167370. doi: 10.1371/journal.pone.0167370

Lax, S., Hampton-Marcell, J. T., Gibbons, S. M., Colares, G. B., Smith, D., Eisen, J. A., et al. (2015). Forensic analysis of the microbiome of phones and shoes. *Microbiome* 3:21. doi: 10.1186/s40168-015-0082-9

Lund, J. B., List, M., and Baumbach, J. (2017). Interactive microbial distribution analysis using BioAtlas. *Nucleic Acids Res.* 45, W509–W513. doi: 10.1093/nar/gkx304

Luongo, J. C., Barberan, A., Hacker-Cary, R., Morgan, E. E., Miller, S. L., and Fierer, N. (2017). Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. *Indoor Air* 27, 338–344. doi: 10.1111/ina.12302

McMurdie, P. J., and Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

Misic, A. M., Davis, M. F., Tyldsley, A. S., Hodkinson, B. P., Tolomeo, P., Hu, B., et al. (2015). The shared microbiota of humans and companion animals as evaluated from Staphylococcus carriage sites. *Microbiome* 3:2. doi: 10.1186/s40168-014-0052-7

Moon, J. H., Lee, J. H., and Lee, J. Y. (2015). Subgingival microbiome in smokers and non-smokers in Korean chronic periodontitis patients. *Mol Oral Microbiol* 30, 227–241. doi: 10.1111/omi.12086

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13:R79. doi: 10.1186/gb-2012-13-9-r79

Odamaki, T., Kato, K., Sugahara, H., Hashikura, N., Takahashi, S., Xiao, J. Z., et al. (2016). Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* 16:90. doi: 10.1186/s12866-016-0708-5

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a web browser. *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385

Peters, B. A., Dominianni, C., Shapiro, J. A., Church, T. R., Wu, J., Miller, G., et al. (2016). The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome* 4:69. doi: 10.1186/s40168-016-0218-6

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

Sarkar, A., Stoneking, M., and Nandineni, M. R. (2017). Unraveling the human salivary microbiome diversity in Indian populations. *PLoS One* 12:e0184515. doi: 10.1371/journal.pone.0184515

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

Schmedes, S. E., Woerner, A. E., and Budowle, B. (2017). Forensic human identification using skin microbiomes. *Appl. Environ. Microbiol.* 83, e01672–e01617. doi: 10.1128/AEM.01672-17

Singh, H., Yu, Y., Suh, M. J., Torralba, M. G., Stenzel, R. D., Tovchigrechko, A., et al. (2017). Type 1 diabetes: urinary proteomics and protein network analysis support perturbation of lysosomal function. *Theranostics* 7, 2704–2717. doi: 10.7150/thno.19679

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

Wu, J., Peters, B. A., Dominianni, C., Zhang, Y., Pei, Z., Yang, L., et al. (2016). Cigarette smoking and the oral microbiome in a large study of American adults. *ISME J.* 10, 2435–2446. doi: 10.1038/ismej.2016.37

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

Zhang, J., Guo, Z., Xue, Z., Sun, Z., Zhang, M., Wang, L., et al. (2015). A phylo-functional core of gut microbiota in healthy young Chinese cohorts across lifestyles, geography and ethnicities. *ISME J.* 9, 1979–1990. doi: 10.1038/ismej.2015.11

Zhou, Y., Xu, Z. Z., He, Y., Yang, Y., Liu, L., Lin, Q., et al. (2018). Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. *mSystems* 3:e00188-17. doi: 10.1128/mSystems.00188-17

Conflict of Interest: CG was employed by company GeneDX.

The remaining 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.