Forensic Science International: Genetics Supplement Series 8 (2022) 181–184

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

The Microbiome Forensics Database [UZH]

Natasha Arora [a,*], João F. Matias Rodrigues [b], Meghna Swayambhu [a], Pim Witlox [c]

[a] Institute of Forensic Medicine, University of Zurich, Switzerland
[b] Department of Life Sciences, University of Zurich, Switzerland
[c] S3IT, University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Microbiome forensics
Body fluid/tissue identification
Training dataset
Machine learning
Database

ABSTRACT

Microbial communities in biological stains can provide valuable information to assist forensic scientists identify the body fluid/tissue present in these. As these microbial communities are characteristic of body habitats, DNA sequencing of microbes can be used to predict bodily origin. Promising predictive results have been obtained with supervised machine learning algorithms trained on bacterial abundance data from human body sites. Importantly, prediction accuracy is dependent on the training dataset, yet compiling a large and comprehensive training reference is a non-trivial issue requiring substantial efforts. Here we present a new online database and associated data-mining platform which is, to our knowledge, the first one customised for forensic scientists investigating body fluids/tissues. Our database features samples originating from ten human body sites, with selection options through an online platform. Users can download bacterial abundance as well as taxonomic data, which can then be used to train predictive models and test their accuracy. Future stages of the development of the platform will include curation of the samples to decrease potential errors in sample labelling, as well as access to an online tool to conduct exploratory analyses.

1. Introduction

Microbiome sequencing studies conducted in recent years have highlighted the potential of microbial markers for forensic investigations, particularly for the identification of body fluids/tissues (BFI) of crime scene stains [1–4]. Several characteristics of microbes render them useful for BFI: they are abundant, and they form complex communities that are tissue specific [4,5]. Thus, investigation of community composition through next-generation DNA sequencing provides information on bodily origin. A widely adopted sequencing approach is the targeted amplification of regions of the prokaryotic 16S rRNA gene, although shotgun and other sequencing approaches have also been tested. To draw associations between the microbial reads and the body sites, numerous supervised machine learning models have been employed, yielding high prediction accuracies for both 16S rRNA gene region as well as shotgun sequence data [2,3,6–8]. Supervised models are trained on labelled datasets, that is, samples for which the bodily origin is known. Thus, their performance is dependent on the size of the training datasets, as illustrated by Tackmann et al. [9]. Additionally, the breadth and quality of the dataset, that is, the number of relevant body sites included in the training, and the accurate labelling of these, also affect predictive accuracy. In forensic studies of microbiome-based BFI, the training datasets used so far have generally been limited in size, given the challenges in compiling large and comprehensive datasets from publicly available data [2,3,10]. Here, we have generated a customised database to provide forensic scientists access to a large selection of samples, so that they may produce a suitable training dataset for BFI supervised machine learning models.

2. Methods

The Microbiome Forensics Database [UZH] was constructed leveraging data from the 2018 version of the Microbe Atlas Project Database (MAPdb; www.microbeatlas.org). MAPdb contains publicly available microbiome read data downloaded from the NCBI SRA and processed with MAPseq [11] to generate OTU abundance as well as taxonomic classification files. From the ca. 1 million samples in MAPdb in 2018, we selected samples from the "human" sub-environment category, and those which had the following body sites in their metadata keywords: saliva, skin, semen, feces, urine, peripheral blood, vaginal fluid, nostril, menstrual blood, amniotic fluid and breast milk. As part of a preliminary curation step, we removed samples for which bodily origin and human

* Correspondence to: Winterthurerstrasse 190, 8057 Zurich, Switzerland.
E-mail address: natasha.arora@irm.uzh.ch (N. Arora).

https://doi.org/10.1016/j.fsigss.2022.10.028
Received 19 September 2022; Accepted 17 October 2022
Available online 18 October 2022

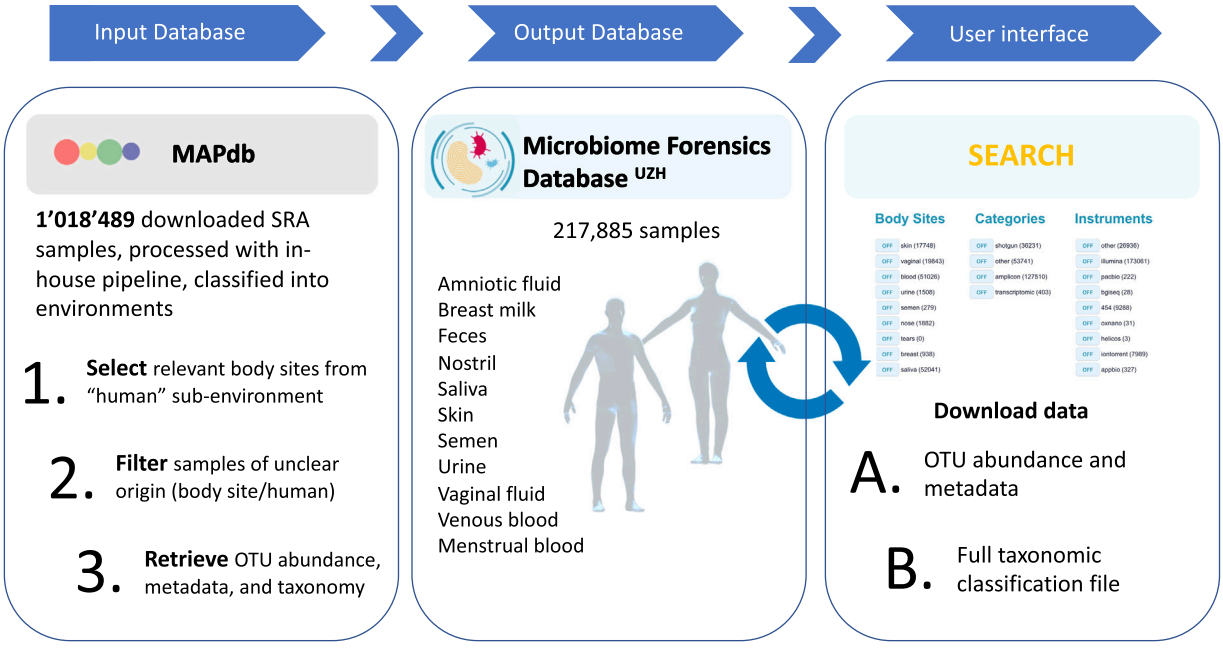

Fig. 1. Database construction and access through the website. A summary of the steps involved in the database construction is shown in the figure. The database can be accessed through various sample selection options in the "Search" tab of the website.

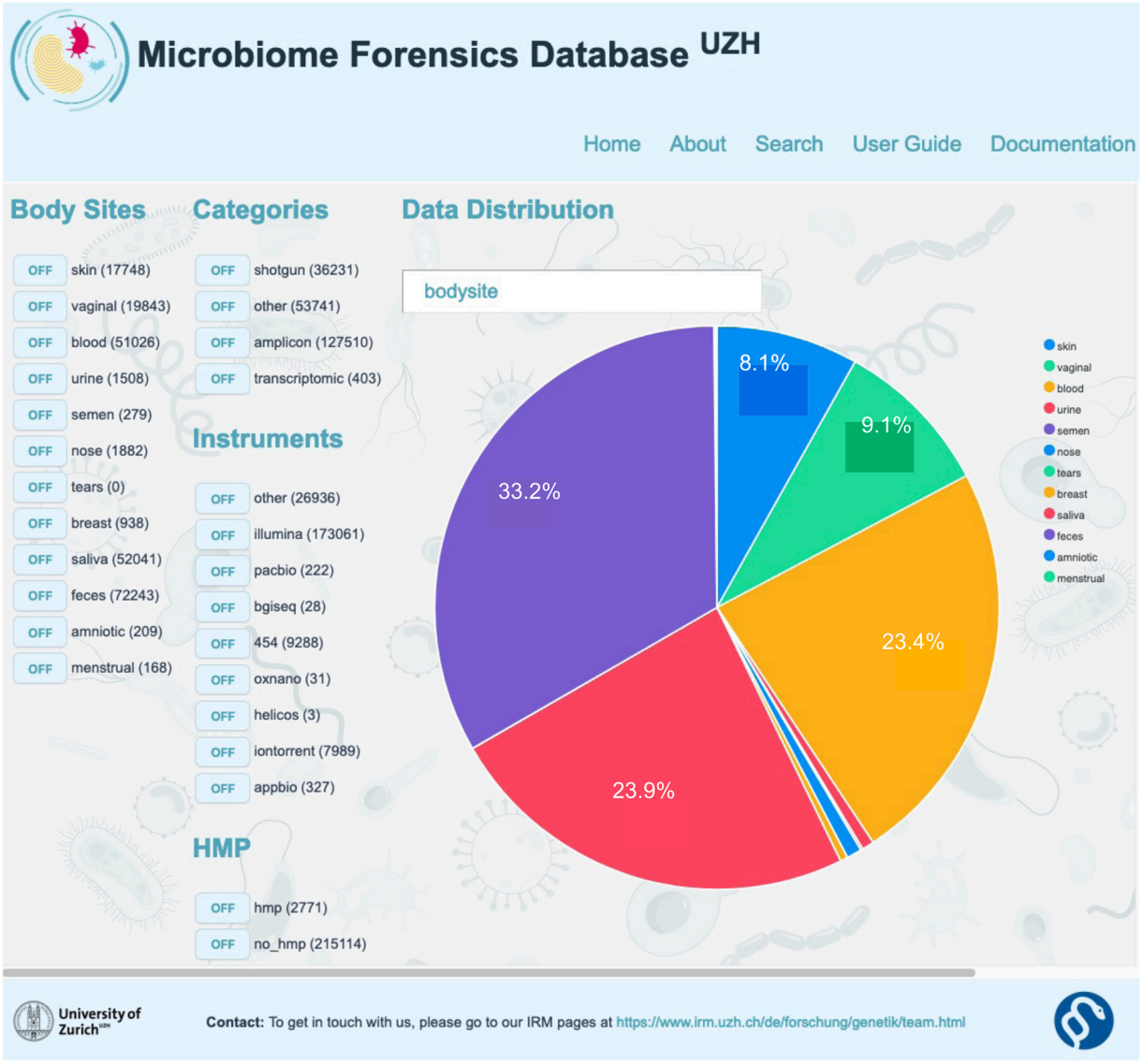

Fig. 2. Screenshot of the Search website. The four selection options are shown together with the pie chart for visualisation of selected data.

origin were ambiguous (those containing more than one body site as well as non-human animals in their keywords). The selected samples were used to populate a database with the associated data files, and a website was developed to provide access to the data (Fig. 1).

3. Results

Our database comprises a total of 217,885 samples from eleven human body sites. The associated OTU abundance and taxonomic classification files for these samples can be accessed through our online data mining platform https://mfdb.irm.uzh.ch. The "Search" tab (Fig. 2) of the platform enables users to select samples based on four categories: body site, sequencing, instrument, type of sequencing data and whether the reads are from the Human Microbiome Project (HMP). For the selected samples, users can download a combined file containing the sample metadata and the read counts for OTUs clustered at the 97 %, 98 % and 99 % OTU identity thresholds. In addition, users can download a taxonomic classification file for all samples in the database. In the "User Guide" tab we provide information and an R script to assist processing these files to generate suitable tables for downstream statistical analyses.

4. Conclusion

The Microbiome Forensics Database [UZH] and the online data mining platform provide access to microbiome data from a growing number of sequencing studies. By making this data easily accessible, and by providing possibilities to generate a comprehensive training dataset, our platform is expected to help improve machine learning models in the field of microbiome forensics. In further stages of development of our platform, we aim to provide additional features and functionalities including curated training datasets and online tools for exploratory analyses of user data.

Conflict of interest statement

The authors declare no conflicts of interest.

Acknowledgments

We are grateful to Margot Riggi for the illustrations and careful review and feedback, and to Peter Resutik for his assistance. We are thankful for the funding received from the Emma Louis Kessler Foundation for this project.

References

[1] A. Dobay, et al., Microbiome-based body fluid identification of samples exposed to indoor conditions, Forensic Sci. Int. Genet. 40 (2019) 105–113.

[2] C.D. López, et al., Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials, Forensic Sci. Int.: Genet. 41 (2019) 72–82.

[3] E.N. Hanssen, E. Avershina, K. Rudi, P. Gill, L. Snipen, Body fluid prediction from microbial patterns for forensic application, Forensic Sci. Int. Genet. 30 (2017) 10–17.

[4] A. Fernández-Rodríguez, F. González-Candelas, N. Arora, Omics for forensic and post-mortem microbiology, in: J. Moran-Gilad, Y. Yagel (Eds.), Application and Integration of Omics-powered Diagnostics in Clinical and Public Health Microbiology, Springer International Publishing, 2021, pp. 219–240.

[5] E.K. Costello, et al., Bacterial community variation in human body habitats across space and time, Science 326 (2009) 1694–1697.

[6] A. Statnikov, et al., A comprehensive evaluation of multicategory classification methods for microbiomic data, Microbiome 1 (2013) 11.

[7] D. Knights, E.K. Costello, R. Knight, Supervised classification of human microbiota, FEMS Microbiol. Rev. 35 (2011) 343–359.

[8] A.L. Tan-Torres Jr, J.P. Brooks, B. Singh, S. Seashols-Williams, Machine learning clustering and classification of human microbiome source body sites, Forensic Sci. Int. 328 (2021), 111008.

[9] J. Tackmann, N. Arora, T.S.B. Schmidt, J.F.M. Rodrigues, C. von Mering, Ecologically informed microbial biomarkers and accurate classification of mixed and unmixed samples in an extensive cross-study of human body sites, Microbiome 6 (2018) 192.

[10] C. Díez López, D. Montiel González, C. Haas, A. Vidaki, M. Kayser, Microbiome-based body site of origin classification of forensically relevant blood traces, Forensic Sci. Int. Genet. 47 (2020), 102280.

[11] J.F. Matias Rodrigues, T.S.B. Schmidt, J. Tackmann, C. von Mering, MAPseq: improved speed, accuracy and consistency in ribosomal RNA sequence analysis, bioRxiv 126953 (2017), https://doi.org/10.1101/126953.