17th International Conference on Availability, Reliability and Security - **ARES 2022**
This is a self-archived pre-print version of this article.
The final publication is available at ACM via
https://doi.org/10.1145/3538969.3538985.

Distance-based Techniques for Personal Microbiome Identification*

Markus Hittmeir
SBA Research
Vienna, Austria
mhittmeir@sba-research.org

Rudolf Mayer
SBA Research
Vienna, Austria
rmayer@sba-research.org

Andreas Ekelhart
SBA Research
Vienna, Austria
aekelhart@sba-research.org

ABSTRACT

Due to its high potential for analysis in clinical settings, research on the human microbiome has been flourishing for several years. As an increasing amount of data on the microbiome is gathered and stored, analysing the temporal and individual stability of microbiome readings and the ensuing privacy risks has gained importance. In 2015, Franzosa et al. demonstrated the feasibility of microbiome-based identifiability on datasets from the Human Microbiome Project, thus posing privacy implications for microbiome study designs. Their technique is based on the construction of body site-specific metagenomic codes that maintain a certain stability over time.

In this paper, we establish a distance-based technique for personal microbiome identification, which is combined with a solution for avoiding spurious matches. In a direct comparison with the approach from Franzosa et al., our method improves upon the identification results on most of the considered datasets. Our main finding is an increase of the average percentage of true positive identifications of 30% on the widely studied microbiome of the gastrointestinal tract. While we particularly recommend our method for application on the gut microbiome, we also observed substantial identification success on other body sites. Our results demonstrate the potential of privacy threats in microbiome data gathering, storage, sharing, and analysis, and thus underline the need for solutions to protect the microbiome as personal and sensitive medical data.

CCS CONCEPTS

• **Security and privacy → Data anonymization and sanitization**; *Privacy protections.*

KEYWORDS

Data Privacy, Re-Identification, Human Microbiome

ACM Reference Format:
Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. 2022. Distance-based Techniques for Personal Microbiome Identification. In *ARES '22: 17th International Conference on Availability, Reliability and Security, August 23–26, 2022, Vienna, Austria.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3538969.3538985

*This work was partially funded by the Austrian Research Promotion Agency FFG under grant 877173 (GASTRIC). SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.

ARES '22, August 23–26, 2022, Vienna, Austria

ACM ISBN 978-1-4503-9670-7/22/08…$15.00
https://doi.org/10.1145/3538969.3538985

1 INTRODUCTION

The human microbiome and its relations to health, diet, exercise and illness is subject to extensive research. The bacteria, viruses, fungi and protists living on various body sites, such as the gastrointestinal tract, appear to have a great influence on our general well-being. The microbiome is studied for prediction, diagnosis and therapy of diseases, and new results about certain correlations and interactions are published on a regular basis. For example, changes in the gut microbiome may be related to gastrointestinal diseases [3], obesity [8], diabetes [12], and depression [14]. While this research field has flourished for several years, there is a great amount of questions and unexplored problems. The clinical importance of the microbiome leads to increasing amounts of funding and investments, and will motivate research efforts for years to come.

Besides its huge potential for analysis in clinical settings, the human microbiome is personal and sensitive medical data, and thus should be treated with the same care and standards as other types of medical data. In some aspects, microbiome data is even more affected by issues of individual privacy such as re-identification attacks. In 2015, Franzosa et al. revealed that the strain variation in clade-specific marker genes of the microbiome can be used to uniquely characterize hundreds of individuals [5]. Using follow-up samples collected 30-300 days later, about 30% of the individuals could still be matched correctly. The gastrointestinal microbiome appeared to be exceptionally stable and allowed to match up to 80% of individuals. The authors conclude that their work demonstrates the feasibility of microbiome-based identifiability, which poses ethical implications for microbiome study designs. The stability of an individual's gut microbiome over time has also been studied by Wang et al. in 2018 [17]. They refer to Franzosa et al., but introduce a different technique called *GePMI*. Unlike the method from the study [5], *GePMI* is applied directly to the genomic sequence data of the microbiome samples. Nevertheless, their results also demonstrate that the difference between the microbiomes of any two hosts is usually greater than the difference between two samples from the same individual.

In 2013, Sweeney et al. [15] demonstrated the feasibility of identifying participants in the Personal Genome Project[1]. There are anonymization techniques for protecting the privacy of human DNA sequence and other genomic databases ([9], [11]). In 2016, Wagner et al. used secure computation for the analysis of microbiome data ([16]). Besides that, there appears to be little previous

[1] https://www.personalgenomes.org/

work on privacy-preserving techniques for the microbiome. Our contribution in this paper concerns the preceding step of exposing, analyzing and quantifying the threat. To this end, we study distance-based techniques for personal microbiome identification (PMI). One major disadvantage of these approaches is the high number of false positive results. We establish a technique for comparing microbiome samples that deals with this issue and is able to reduce the false positive count, while simultaneously keeping a high number of true positives. In a direct comparison of our method with the technique used by Franzosa et al. [5], we show that, on the same datasets, we are able to improve upon the performance of the method and the results established in [5].

The other mentioned approach for microbiome identification, GePMI, is not directly comparable to our evaluation. This is due to (i) their limited focus on data from just one body site, and (ii) the fact that it is based on different assumptions concerning the input data. GePMI relies on the availability of genomic sequence data and is thus more comparable to approaches related to the task of forensic identification on the microbiome (e.g., [4] and [19]). Based on similarity measures such as nucleotide diversity, forensic identification often uses a much richer form of input data than the already rather condensed, tabular, feature-based representation utilized in [5] and in our approach. While working with richer input data leads to more accurate results, there is also a desire to publish, share and study microbiome datasets as discussed in [5]. In addition, privacy risks established on such datasets can serve as a minimal baseline one has to expect when publishing microbiome profiles. Hence, we consider it as the primary benchmark for our evaluation.

The rest of this paper is structured as follows. After a more detailed discussion of the related work in Section 2, we explain the threat model in Section 3. Our own technique is introduced in Section 4 and evaluated in Section 5, where we present the results of our experiments and the detailed comparison to earlier approaches. Finally, Section 6 summarizes this paper and discusses future work.

2 PRELIMINARIES AND RELATED WORK

We start by discussing microbiome data in greater detail. Besides the already mentioned microbiome of the gastrointestinal tract, samples may be taken from several other body sites, such as saliva, throat, anterior nares (the external portion of the nose), supragingival plaque (at the teeth) or buccal mucosa (at the inside of the cheek). In this paper, we consider up to 18 different body sites in the datasets from [5], which are obtained from raw microbiome data that is publicly available through the Human Microbiome Project's repository[2]. For gigabytes of metagenomic sequences as input, the authors of [5] applied both 16S ribosomal gene sequencing and whole metagenome shotgun sequencing and derived structured tabular datasets with different feature types. For comparison, the GePMI method [17] is not based on the construction of such tabular data, but extracts only a small subset of sequence information from the raw microbiome data.

The scope of our work is not the preprocessing of raw microbiome data, but rather techniques for personal microbiome identification applied to structured datasets with already extracted features.

[2]https://www.hmpdacc.org

We have before mentioned two approaches that will now be discussed in greater detail. The technique of [5] aims to find so-called metagenomic codes that are unique for each individual, and are based on the concept of hitting sets. For a collection of nonempty sets $\{M_1, M_2, ..., M_n\}$, a hitting set S is a set that has at least one element in common with each M_i. S is said to be minimal if removing any element from S would cause it to not hit at least one M_i. Consider a population of individuals $\{1, \ldots, k\}$, each with a set of metagenomic features U_i, $i = 1, \ldots, k$. Moreover, let $A_{i,j} := U_i \setminus U_j$ be the set that contains the features present in individual i, but absent in individual j. The authors then use a greedy algorithm to construct a hitting set S_i for the collection of the $A_{i,j}$, $j = 1, \ldots, k$. The returned hitting set might not be minimal. However, it establishes a metagenomic code *unique* to individual i. If they are stable enough over time, the comparison of these codes may be used to find pairs of samples that belong to the same individual.

In the GePMI method [17], the pairwise Jaccard similarity index is estimated for human gastrointestinal microbiome samples. The method is based on reference-free, long k-mer features, i.e. on features extracted from subsequences of length k contained within the original sequence. The Jaccard index of two sets A and B is the ratio of the number of elements in their intersection to the number of elements in their union. As the Jaccard index cannot be computed directly due to the mentioned dimensions of the data, a technique called MinHash [1] is used to construct k-mer signatures and estimate the index. In order to evaluate whether two samples are taken from the same individual, a beta distribution is fitted to determine the significance of the similarity. This approach is based on the assumption that samples taken from the same individual at different times will show larger similarities than those taken from two different individuals. The results showed that most of the human gut microbiomes can be identified (auROC=0.9470, auPRC=0.8702), and, even after antibiotic treatment, maintain a certain specificity.

The two described approaches differ both on the features used as input to the technique, as well as on the method used in the process of identifying microbiome samples. The input to the GePMI algorithm[3] is a matrix containing the Jaccard similarity indices for all samples in the database. The estimates of these indices are based on the extracted k-mer features as discussed above. In Franzosa et al. [5], the input to the algorithm[4] is tabular data, where the samples are columns and the features are rows. Four different feature types are considered, namely operational taxonomic unit abundance ('OTU'), bacterial and archaeal species abundance ('Species'), species-specific marker genes ('Markers') and tiled kilobase windows ('KBWindows'). Depending on the feature type, the units for each sample in the columns are either measured in *relative abundance*, or in *reads per kilobase per million sample reads* (RPKM). Relative abundance means that the sum of all components in each sample vector equals 1, while RPKM is normalized by the sequence size. Note that in the approach of [5], a feature is understood as either present or absent, depending on a detection and a nondetection threshold that correspond to the general unit size. Therefore, the sample vectors are effectively transformed into binary 0, 1 vectors, and the original level of abundance is used to prioritize features

[3]Python implementation available at: https://github.com/princello/GePMI
[4]Data and Python implementation available at: http://huttenhower.sph.harvard.edu/idability

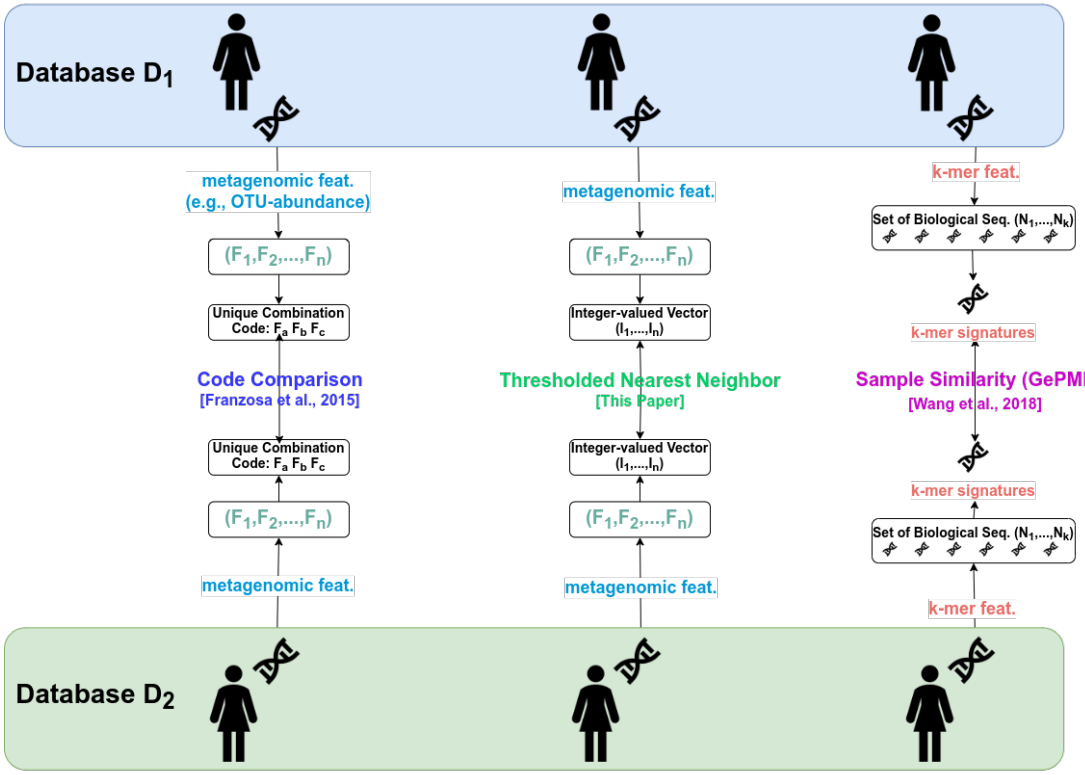


Figure 1: Overview of different techniques for personal microbiome identification: metagenomic code matching by [5] (left), our distance-based method via thresholded nearest neighbor search (centre), and the k-mer based GePMI method [17] (right). The method in [5] as well as our distance-based method take metagenomic feature abundance vectors as input. In [5], a subset of the features present in a sample is used to construct a unique code, which is then used for comparison. In our method, we transform the complete feature vectors to integer-valued vectors and compute distances to find nearest neighbors (see Section 4). Finally, the GePMI approach is based on k-mer features extracted directly from the original genetic sequences in the sample. The similarity between samples is estimated via the construction of k-mer signatures.

in the already mentioned greedy construction of the metagenomic codes. More detailed information on the datasets can be found in Table 1, where we extended a respective table from [5].

The task of PMI is similar to record linkage ([6]), as we try to identify records in datasets that belong to the same individual. Usually, a certain subset of (quasi-)identifying attributes is used for the comparison in record linkage. A well-known example is the identification of records in the high-dimensional and sparse Netflix Prize dataset ([13]) by using limited background knowledge from the Internet Movie Database as a source. *Deterministic* record linkage only links on unique, exactly matching keys (i.e. all attributes have equal values) of the records, and a link is either established, or not. *Probabilistic* record linkage matches two records using possibly not-exactly matching keys, and computes probabilities of how likely they are matches, based on the key values.

In the case of PMI on microbiome data, we do not have specific (quasi)-identifying attributes, but we are instead able to utilize the information from all attributes. As the microbiome readings change over time, the matching can be considered to be equivalent to the probabilistic record linkage task. Thus, we have to solve specific challenges related to the preprocessing of the data and the definition of what we accept as matching records. Figure 1 provides an overview on the differences between the discussed PMI techniques. Our method, which will be discussed in greater detail in Section 4, may be applied to input data that satisfies the assumptions of Franzosa et al. In order to analyze the privacy risks on such kind of data, we will thus focus on a comparison of our technique with their approach established in [5]. The first important difference is that the matching of samples in our method is not based on the construction of metagenomic codes via hitting sets, but on computing pairwise distances with a certain metric. In [5, p.E2936], distance-based microbiome identification has already been considered and compared to the metagenomic-codes technique. They used two metrics, namely the Bray-Curtis dissimilarity and the Canberra distance. The first quantifies the dissimilarity between two sample vectors a and b via $\sum_i |a_i - b_i| / \sum_i |a_i + b_i|$, where a_i and b_i are the components of a and b. The second is a weighted version of the L_1 (Manhattan) distance and given by $\sum_i |a_i - b_i| / (|a_i| + |b_i|)$. Subsequently, the authors of [5] took the nearest neighbor in the respective metric as the candidate for the pairings of the samples, and these pairings are either true positives (i.e., both samples are from the same individual) or false positives. While the experiments suggested that the two metrics lead to reasonably strong true positive rates, the high number of false positives in the form of spurious

Table 1: Feature types of the datasets, their units and detection thresholds (as in Table 1 in [5])

Feature description	Short name	Sequencing basis	Units	Confident detection threshold	Relaxed detection threshold	Confident nondetection threshold	Body sites	Paired samples per body site	Number of features
Operational taxonomic units	OTUs	16S rRNA gene	Relative abundance	$> 1e^{-3}$	$1e^{-4}$	$< 1e^{-5}$	18	25-105	968-2,663
Microbial species	Species	Whole metagenome shotgun	Relative abundance	$> 1e^{-3}$	$1e^{-4}$	$< 1e^{-5}$	6	14-50	196-317
Species-specific marker genes	Markers	Whole metagenome shotgun	RPKM	> 5	> 0.5	< 0.05	6	14-50	154,328-349,779
Kilobase windows from microbial reference genomes	kbwindows	Whole metagenome shotgun	RPKM	> 5	> 0.5	< 0.05	6	9-45	23,878-263,847

matches was understood as a major disadvantage of distance-based PMI techniques. A first solution to this problem was presented in [18], where usage of the Canberra distance together with a constant threshold for accepting or rejecting nearest-neighbor candidates was proposed. In this paper, we elaborate on the approaches of [5] and [18] in regards to distance-based PMI by (i) considering a different representation of the sample vectors as integer vectors to perform the matching on, and (ii) providing a dynamic threshold for containing the high number of false positives observed by [5].

3 THREAT MODEL

The well-known identifying power of human DNA, together with the possibility to predict subject traits on its basis, has led to increased concerns for privacy in genomics research [10]. As human DNA appears as contaminant in microbiome samples, this awareness established the routine of removing human DNA sequences from the data before publication [2]. However, it has been concluded in [5], and it will be strengthened by the present paper, that the possibility of accurately tracing the data back to their original sources may be based on microbiome data alone, and thus remains even after DNA cleansing. As the human microbiome can be associated with a variety of individual traits, including the health status, it has to be considered as sensitive medical data and re-identification as a severe threat.

As studies and research on the human microbiome gain importance, an increasing amount of data will be processed and stored in tabular formats based on the metagenomic feature types described in the previous section. We will strengthen the conclusion of [5] that the sample vectors in such datasets and their components, i.e. the feature counts, maintain a temporal stability that allows to identify individuals in the database for extended periods of time. The following threat model shows why microbiome sample matching on metagenomic features is a privacy concern:

Victim. Individuals who provided their microbiome samples in the course of, e.g., medical studies, diagnosis and therapy of diseases, and personal health and fitness advise. Their microbiome data, and potentially analysis results and additional metadata, are electronically available.

Adversary. An adversary is any party in possession of microbiome samples with the intention to link them to other microbiome samples in order to accumulate information about the underlying individual and/or identify the individual (or data subject in the terminology of the General Data Protection Regulation, GDPR). The adversary has multiple options to obtain microbiome samples, including, e.g., public microbiome databases, cyberattacks against healthcare facilities and research organizations, data exfiltration via insiders, and potentially, directly from the human victim (e.g., saliva).

Threat. We assume an adversary already possesses a sample of a certain individual and wants to match it with samples from another database for multiple reasons:

(i) To find out if an individual participated in a study, e.g., in the context of specific diseases. Even if the microbiome samples connected to diseases include no identifying metadata, a match with a known sample will identify the individual.
(ii) The attacker could obtain (previously unknown) metadata linked to the identified sample from the new database (e.g., medical and personal data provided in the course of a study).
(iii) Furthermore, by identifying matching samples, even in the absence of metadata, the attacker gets hold of new microbiome samples from the same individual and could thereby learn about changes over time in the individual's human microbiome. Such changes could, e.g., point to diseases, depression, and changes in diet.
(iv) Finally, ongoing research efforts increasingly associate microbiome samples with several other individual traits, such as the age or geographical background [20], which an attacker could learn. Thereby, collecting and linking multiple samples from the same individual could also aid an attacker in identifying the person behind a sample.

4 PMI VIA THRESHOLDED NEAREST-NEIGHBOR SEARCH

Our main contribution is to introduce a new approach for distance-based microbiome identification, which we combine with a remedy for the problem of high false positive counts discussed at the end of Section 2. By establishing a threshold for accepting and rejecting closest neighbors as possible candidates, we will show that we are able to reduce the false positives and simultaneously keep a high true positive rate. Our results in Section 5 will demonstrate that this technique is suitable for several feature types and applications in the area of PMI. Given the resulting extension of the toolkit for achieving re-identification on microbiome datasets, it may be

concluded that the related privacy risks are even higher than what was demonstrated up until now.

In order to explain our method, let us start by discussing the input data and the task from a formal point of view. Assume that we have two datasets D_1 and D_2, where D_1 contains the microbiome samples of individuals at some initial time point, and D_2 contains the microbiome samples of (some of) these individuals at another point in time. The task is to match all the samples from the individuals in D_2 with the samples in D_1 from the same individual. In the following, we assume that the datasets are comprised of the feature types discussed in Table 1. Also note that, in this scenario, D_1 is considered to be the dataset under attack, while D_2 contains the samples that an attacker might wish to compare with those in D_1. For this reason, our method does not assume any dependence between the samples in D_2, and also works in cases where D_2 contains only a single record.

4.1 Preprocessing via feature abundance limits

The first phase of our approach is to prepare our input data for the application of the distance metric. This step is somewhat similar to Franzosa et al.'s encoding of the sample vectors via certain thresholds (see Section 2). However, instead of just using a detection and a nondetection threshold, we will apply five feature abundance limits, which allows us to split up the complete range of values of the respective unit size. The goal is to transform all the values in the cells in D_1 and D_2 to either 0, 1, 2, 3, 4 or 5. For example, let us consider taxon-type features such as Species and OTUs. The values in the cells of the tables are measured in relative abundance, which means that they range from 0 to 1 and that each sample vector sums up to 1. Franzosa et al. used 0.001 as detection threshold and

Table 2: Feature abundance limits

Features	t_0	t_1	t_2	t_3	t_4
Taxon-Level	0.00005	0.0005	0.005	0.05	0.5
Gene-Level	0.005	0.05	0.5	5	50

0.00001 as nondetection threshold. We will instead use the five feature abundance limits t_0, t_1, t_2, t_3 and t_4 as given in Table 2. As it has been discussed in [5, p.E2936], the incorporation of low-abundance features can influence the performance. In particular, taxon-level features with low abundance (e.g., 10^{-6}) are very unstable, and we generally recommend to ignore all values below our choice for t_0, namely 0.00005. Franzosa et al. also mentioned that this problem is less pronounced on the gene-level, which agrees with our own observations. Therefore, we recommend the value 0.005 for t_0 in these cases. The higher abundance limits in Table 2 split up the complete range of values in an (geometrically) even way.

Let x be any value in the original cells of D_1 and D_2. The vectors in the input datasets are transformed by the following rule.

$$x \leftarrow \begin{cases} 0 & \text{if } x < t_0, \\ i & \text{if } t_{i-1} \le x < t_i \text{ for some } 1 \le i \le 4, \\ 5 & \text{if } x \ge t_4. \end{cases} \tag{4.1}$$

In order to fully understand the purpose of this rule, we again take a look at Figure 1 and remind us of the Franzosa et al.'s approach, which is based on metagenomic codes. Said codes focus on the presence and absence of features. They are built by considering well constructed subsets of the present features that are unique for the respective individual. With this approach in mind, our data preprocessing step presented above may be interpreted as the construction of much more complex metagenomic codes. We do not only encode the presence or absence of features, but also their level of abundance. Our subsequent comparison of individual samples is not based on subsets of present features, but on the complete encoded feature vectors. By taking much more of the original data into account, this transformation of the sample vectors to integer-valued vectors therefore greatly benefits the performance of our algorithm. In addition, it reduces the computations from float-valued to integer-valued vectors.

Finally, we want to select features that help us to distinguish between the individuals in D_1. We hence find all feature rows in D_1 which have been reduced to the 0-vector by applying (4.1). Subsequently, we delete those features from both D_1 and D_2. The number of remaining features only depends on the limit t_0. Due to the general sparsity of the datasets, this feature elimination step also reduces the required runtime of the nearest-neighbor search.[5]

4.2 Computing nearest neighbors

In the second phase of the procedure, we want to compute the distance between all the pairs (a, b), where a is a column (i.e. individual) in D_1 and b is a column (i.e. individual) in D_2. We may use any measure, metric or similarity coefficient for this task. Examples could be the $L1$ (Manhattan) and $L2$ (Euclidean) distances or the already mentioned Bray-Curtis and Canberra measures. For the remainder of this paper, the measure used for the nearest-neighbor search will be denoted by Δ.

For every b in D_2, we want to find the column a in D_1 for which $\Delta(a, b)$ is minimized. To improve the efficiency of this method, we use the following procedure.

(1) Construct a ball tree[6] on D_1 based on Δ.
(2) For each column b in D_2, use the ball tree to find the nearest neighbor of b in D_1. Save the resulting pair.

4.3 Thresholding for finding true matches

As mentioned at the end of Section 2, the third and last phase of our technique solves the task of deciding which of the obtained pairs we should keep and which we should reject. Setting a constant threshold for the distance between these closest neighbors would be a natural solution. However, there are several problems. Since there is a lot of variation between the average distances on different datasets, the threshold would need to be defined specifically for each D_1 and D_2, thus becoming a hyper-parameter of the algorithm. Furthermore, we also want to allow a scenario where D_2 might contain only one sample that is checked against the database D_1, so the definition of the threshold should not assume a minimum size of D_2.

In order to solve this challenge, we consider the similarity of each neighboring pair (a, b) in relation to the whole dataset D_1.

[5] The runtime complexity of our algorithm is analysed in Section 7.1 in the appendix.
[6] A ball tree is a binary tree that partitions the data points into multidimensional balls. This structure allows for an efficient nearest-neighbor search. We used Python's sklearn-package (https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.html).

The threshold for accepting or rejecting neighboring samples is defined and applied by the following procedure.

(1) Let (a, b) be a neighboring pair. For all columns $r \neq a$ in D_1, compute $\Delta(a, r)$. Denote the list of these distances by $\mathcal{L}_a$.
(2) If $\Delta(a, b) < \min\{d : d \in \mathcal{L}_a\}$, accept (a, b). Otherwise, reject.

This way, we obtain a dynamic nearest-neighbor threshold that is different for each of the pairs (a, b), and compares the similarity between a and b to the similarity between a and all the samples in D_1. Only those pairs are accepted for which a is more similar to b than to any sample in D_1. This implies that, in addition to a being the nearest neighbor for b, b is also the nearest-neighbor for a considering all samples in the union of D_1 and $\{b\}$. The defined threshold is independent of possible other samples in D_2, as we desired. Moreover, we want to stress that there are various possibilities to relax or tighten the proposed condition of taking the minimum of $\mathcal{L}_a$. For the sake of clarity, however, our experiments in the next section are based on the procedure as outlined above. In addition, we will also discuss the true positive rate in the case of disabling the threshold and accepting all candidate pairs.

The only thing left to discuss is the choice of the distance function Δ. First of all, we want to mention the possibility of using one distance function Δ_N for the nearest-neighbor search and another distance function Δ_T for the thresholding. However, we will not consider this option in our experiments in the subsequent section. There, we will use a distance function based on the Pearson correlation coefficient[7], namely

$$\Delta_{cor}(a, b) = 1 - \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2}\sqrt{\sum_i (b_i - \bar{b})^2}}, \tag{4.2}$$

where a_i and b_i are the components and $\bar{a}$ and $\bar{b}$ are the means of the vectors a and b, respectively. Since the Pearson coefficient, which is given by $1 - \Delta_{cor}(a, b)$, measures linear correlation and returns a value in $[-1, 1]$ based on whether the correlation is positive or negative, the distance (4.2) returns a value in $[0, 2]$. We will see that this measure performs well in both finding true positive matches and in thresholding for reducing false positives.

5 EVALUATION

In our evaluation, we use the datasets from Franzosa et al. [5]. As already discussed in Section 2, there are four different feature types, namely Species, OTUs, Markers and KBWindows. For each of these feature types, we have a number of different body sites. For each of the body sites, we have two datasets D_1 and D_2 of the same size. D_1 contains a first microbiome sample from a set of individuals, and D_2 contains follow-up samples from the same individuals. Further information on the number of body sites and the size of the datasets can be found in Table 1.

We now explain all possible outcomes of both methods for PMI and the corresponding notation used in our description. For every sample in D_2, the PMI methods construct a set of matches of samples from D_1. For instance, these sets contain the nearest-neighbor in distance-based approaches, or samples that match the metagenomic code in the approach from [5]. In general, we count the number of sets of matches containing only the correct individual (true positives, TP), sets of matches containing only wrong individuals (false positives, FP), sets of matches containing the correct and also wrong individuals (TP+FP), the number of individuals that incorrectly have not been matched (FN), and the number of individuals for which the technique of [5] was not able to construct a unique metagenomic code (NA). Note that NA leads to an incorrect non-match comparable to the occurrences counted by FN. While the sets of matches in the approach of [5] may contain more than one individual, the sets of matches in our correlation-distance approach contain at most one individual, hence the only valid outcomes of this technique are TP, FP and FN. We also want to mention that, for both methods, there cannot be any true negatives, since the compared datasets D_1 and D_2 contain the same individuals, which means that there would always be a correct match for any sample in D_2. The capability of detecting true negatives is nonetheless investigated in Section 7.3 in the appendix, where we modified the datasets to simulate a suitable experimental setup. However, in the main experiments of this paper that are presented below, we use the datasets in an unmodified form to ensure comparability of our results with the ones from [5].

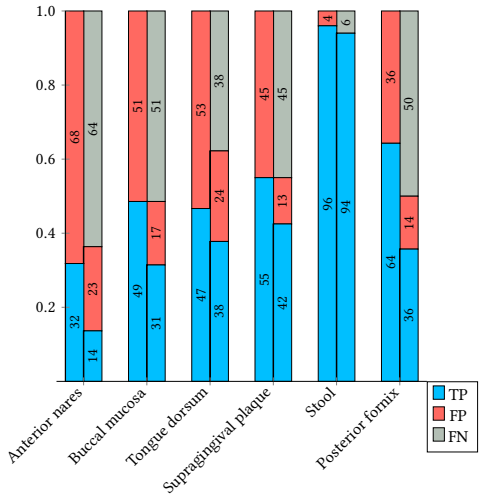
In order to get a first impression of the advantages of our method, we start by considering the results of the datasets for the taxon-level feature types Species and OTUs. On these datasets, the approach from [5] achieved rather low true positive rates compared to the methods based on the Bray-Curtis and the Canberra measure, which have been described in Section 2. The results of these distance-based approaches can be found in Figure S7 in the appendix of [5].

5.1 Taxon-level features

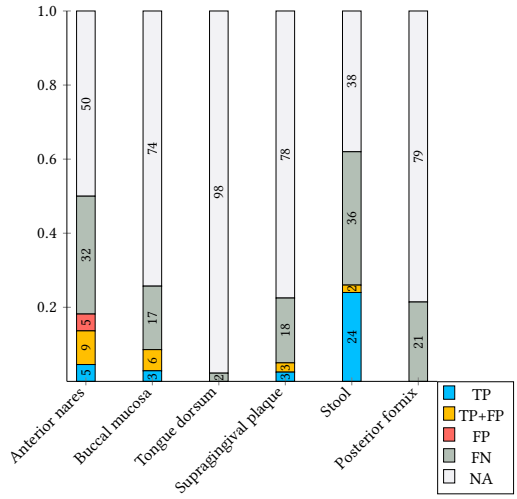
Figures 2 and 3 show the results on the taxon-level features, and are based on the counts in Tables 5 to 8 in Section 7.4 in the appendix. Figure 2 shows the results for the datasets with Species features. Figure 2a concerns the results of our method described in Section 4, while Figure 2b shows the results from [5, Fig.3], which we were able to reproduce exactly by applying the implementation with standard settings. We can see that the true positive (TP) rate of our correlation-distance method exceeds the true positive rate of the metagenomic-codes approach on every body site, and by a substantial margin. Considering the difference between the TP count with enabled nearest-neighbor threshold (right bars) to the TP count without the threshold (left bars) shown in Figure 2a, we can see that our approach in the third phase of the algorithm performs well in distinguishing between true and false positives. We may conclude that our thresholded correlation-distance method is able to preserve most of the true positives that are not found by the metagenomics-code approach, while it resolves the disadvantages of distance-based methods by reducing the false positives and improving the precision, i.e. the ratio of true positives among all instances marked as positives, TP/(TP+FP). In addition, we want to mention that the TP counts without nearest-neighbor threshold (left bars) appear[8] to be also improving upon the results obtained

[7] Here, we used Python's scipy package (https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.correlation.html).
An overview on distance measures provided by scipy is available at https://docs.scipy.org/doc/scipy/reference/spatial.distance.html

[8] The appendix of [5] does not provide exact values, but only a plot of the results, thus we estimated the result values based on this plot.



(a) Results in % for the correlation-distance approach. The left bar shows the result for the method applied without a nearest-neighbor threshold (i.e., without the third phase of the technique), and the right bar shows the result when the threshold is enabled.



(b) Results in % for the metagenomic-codes approach. The TP+FP count indicates the number of occasions where the set of matches contained the correct as well as incorrect individuals. The NA count shows the number of individuals with non-unique metagenomic code.

Figure 2: Results of Personal Microbiome Identification on datasets with the *Species*-type features

from the Bray-Curtis and Canberra measures as shown in Figure S7 in the appendix of [5]. This observation supports our general approach of encoding the datasets by using feature abundance limits and subsequently using the correlation-distance, instead of the binary encoding used in [5].

Similar conclusions may be drawn for datasets with OTUs features, the results for which are shown in Figure 3. Again, our method doubles the TP count over the metagenomic-codes approach on almost every body site. Sometimes, the improvement is even more substantial than that. Interestingly, the metagenomic-codes approach shows a significant number of TP+FP matches on several body sites. However, it has to be noted that TP+FP matches only offer a 1/2 chance (or worse) of choosing the correct individual. While our thresholding technique works fine on most occasions, it still leaves a precision of 50% or less on some of the sites, such as antecubital fossa and posterior fornix. This might indicate that, when applied to these body sites, the choice of stricter thresholds or the selection of other distance measures for the thresholding in our method might be advisable whenever a high number of true positives is less important than precision. The interested reader may find a more detailed discussion of possibilities for adapting our method in Section 7.2 in the appendix.

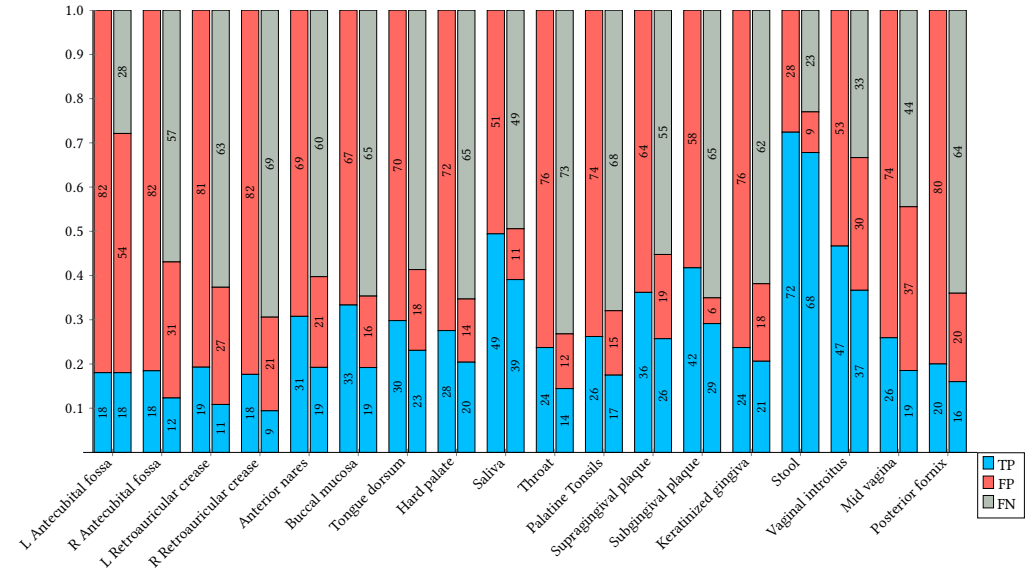
5.2 Gene-level features

Figures 4 and 5 are based on the counts in Tables 9 and 10 in Section 7.4 in the appendix. Figure 4 shows the results for the datasets with feature type Markers. Again, our method improves the results of the metagenomic-codes approach on 4 out of 6 body sites, namely buccal mucosa, tongue dorsum, supragingival plaque and stool. A particularly noteworthy improvement may be observed for buccal mucosa, where the percentage of true positives is more than twice as high, while the number of false positives is the same as for the metagenomic-codes approach. A worse performance can be observed on the body site anterior nares, where our method has a rather low TP count. On posterior fornix, we have a high number of true positives (79%) in the unthresholded version of our method. However, this is reduced to 50% after thresholding.

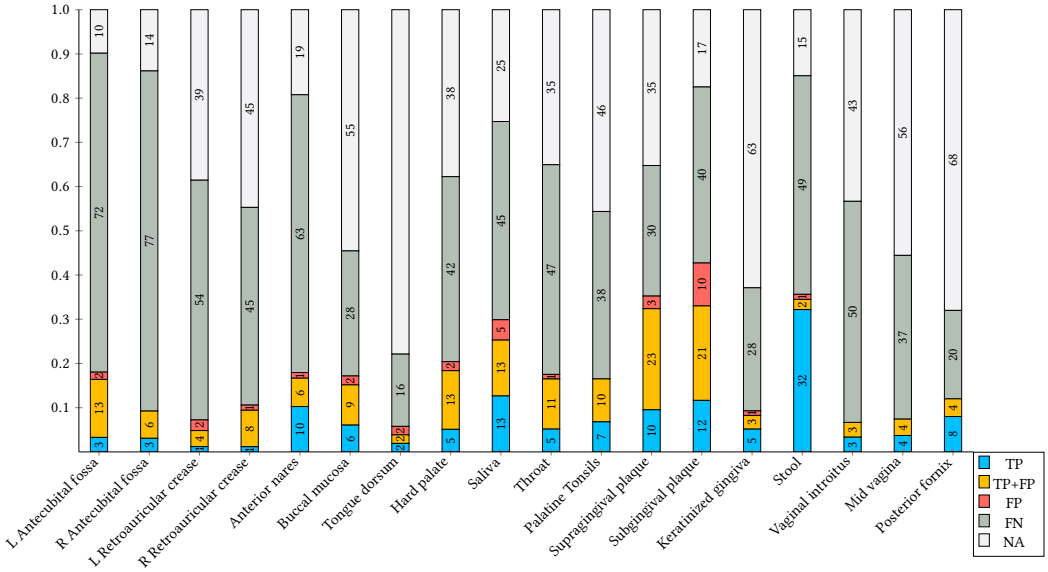
Regarding the results on the datasets with feature type KBWindows as shown in Figure 5, we can note that our method appears to bring no benefit compared to the metagenomic-codes technique, except on the body site stool, where our method outperforms the approach on metagenomic codes by identifying 80% of all true positives. For buccal mucosa, tongue dorsum and supragingival plaque, our method produces the same or fewer true positives, while having a larger false positive rate. On anterior nares and posterior fornix, we achieve the same number of confirmed true positives, but still a higher number of false positive. Also, on posterior fornix, [5] provides a relatively large number of result sets that are either true or false positives, depending on which item of that set is chosen.

5.3 Average performance comparison

We have seen that the differences between the performances of the methods depend on the body sites and feature types. In general, our approach works particularly well on taxon-level features, while the differences to the metagenomic-codes technique are less pronounced on the gene-level. One explanation for this behaviour is the so-called "curse of dimensionality". Our distance-based technique works particularly well on shorter sample vectors. On Species-type features, where we only have 300 features or less (see Table 1), our improvement is most substantial. For microbiome samples of the gastrointestinal tract, however, our approach appears to improve upon the method from [5] regardless of the considered feature type. On the other hand, a converse statement seems to be true for anterior nares.



(a) Results in % for the correlation-distance approach

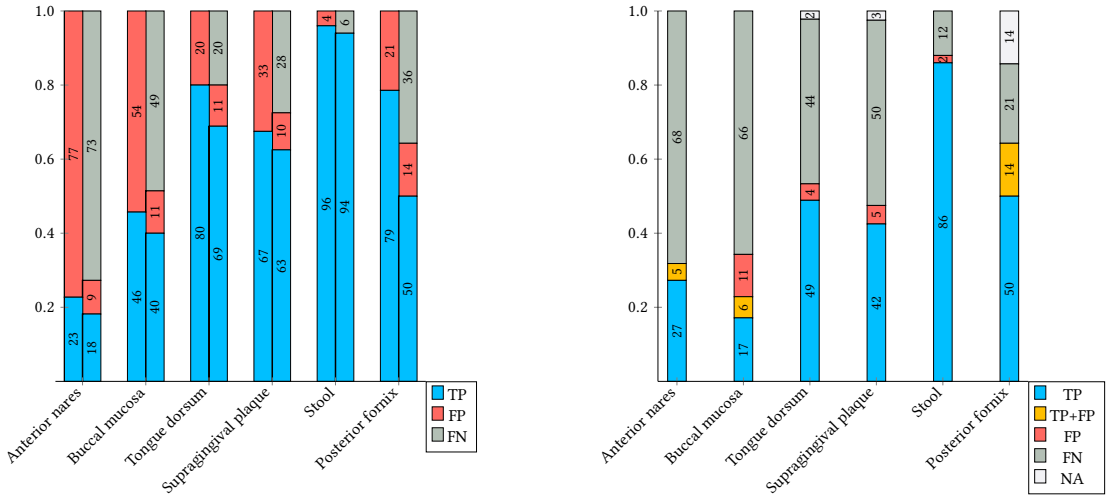


(b) Results in % for the metagenomic-codes approach

Figure 3: Results of Personal Microbiome Identification on datasets with *OTU*-type features

We are now interested in analyzing and comparing the average performance of the two considered methods for PMI. In order to summarize the results for all datasets, we consider the 6 body sites that have been investigated for every feature type, namely anterior nares, buccal mucosa, tongue dorsum, supragingival plaque, stool and posterior fornix. In Table 3, 'MetaC' stands for the metagenomic-codes approach, while 'CorrD' shows the results of our correlation-based method. We consider precision, recall and the F1-score. While precision has already been defined, recall is usually understood as the fraction of the relevant instances (in our case, matches) that were actually found. Since in our datasets every instance is relevant in the sense that there is a match for each sample, and there are thus no true negatives, recall equals the fraction TP/n, where n is the number of individuals in the dataset. As a result, we may consider recall as the TP count in % that is shown in the figures above. Finally, the F1-score is defined as the harmonic mean of precision and recall, i.e.: $2 \cdot \text{precision} \cdot \text{recall}/(\text{precision} + \text{recall})$

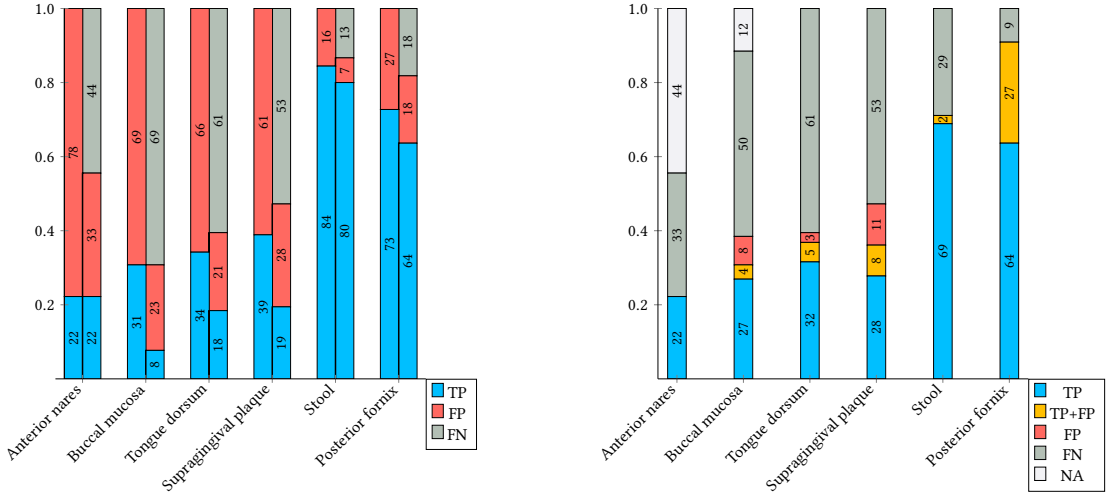
The cells in the table show the mean of the respective measure over the four datasets for the different feature types. In particular, we stress that the mean F1-score is not directly obtained from the values of the mean precision and recall in the same column in terms of the equation above. This is particularly true for tongue



(a) Results in % for the correlation-distance approach

(b) Results in % for the metagenomic-codes approach

Figure 4: Results of Personal Microbiome Identification on datasets with *Marker*-type features



(a) Results in % for the correlation-distance approach

(b) Results in % for the metagenomic-codes approach

Figure 5: Results of Personal Microbiome Identification on datasets with *KBWindows*-type features

Table 3: Mean precision, recall and F1-score on six body sites over the four considered feature types (Highlighted: Better score)

Body Site	Ant. Nares		Buccal Mucosa		Tongue Dorsum		Suprag. Plaque		Stool		Post. Fornix	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
Mean Precision	**0.760**	0.481	**0.619**	0.554	**0.747**	0.623	**0.692**	0.655	0.946	**0.951**	**0.779**	0.679
Mean Recall	**0.214**	0.183	0.225	**0.246**	0.224	**0.370**	0.290	**0.375**	0.544	**0.840**	**0.418**	0.413
Mean F1-Score	**0.332**	0.262	0.287	**0.338**	0.305	**0.455**	0.386	**0.473**	0.660	**0.890**	0.440	**0.505**

dorsum and posterior fornix, as for these the metagenomic-codes approach did not find any TP or FP matches for the species datasets. While the recall and F1-scores are set to 0, precision is undefined in these cases, hence only the other three feature types have been considered in computing the respective precision means, which provides an optimistic estimation of the average. Before discussing the results, we also note that we have accounted for the TP+FP matches of the metagenomic-codes approach by adding them both to the TP as well as to the FP count before computing the values of the measures. Compared to the second possible option, namely

ignoring TP+FP outcomes in the computation, this choice improves recall more than it reduces the precision, and thus increases the mean F1-scores of this technique on each of the six body sites. As a result, this choice of dealing with TP+FP matches favours the metagenomic-codes method's F1-score the most, and can thus be seen as an optimistic bound to compare our method to. Nevertheless, we can see that the correlation-based technique outperforms the F1-scores of the metagenomic-codes approach on each of the body sites, except for anterior nares. In particular, we observe an increase of the mean recall percentage of almost 30% for CorrD on the gut microbiome (measured by stool), while also improving upon the precision. As the harmonic mean of precision and recall, the F1-score may be understood as a measure for the general performance of the methods. While precision is better in five of six columns of MetaC, the main advantage of CorrD is the improvement of the TP count and, hence, the recall, at a generally significant lower increase of false positives, i.e. a decrease in precision. Considering the remaining 12 body sites that have only been investigated for OTUs features (Figure 3), we can see that this tendency seems to extend to most of them. For instance, on saliva, subgingival plaque or hard palate, we can see that the increase of the TP count is much more significant than the decrease of precision. However, precision is notably worse in our method for body sites such as antecubital fossa and retroauricular crease.

6 CONCLUSION

In this paper, we studied approaches for personal microbiome identification. While the temporal stability and the possibility of associating individual microbiome samples may come with benefits for clinical analysis, they also pose a privacy threat to the affected individuals. We presented a distance-based technique for PMI and showed that the approach improves upon previous results in this research area. A substantial performance gain can be observed for data containing certain feature types, and for microbiome samples obtained from certain body sites. In particular, our method showed an increase of the average percentage of true positive identifications of 30% for gut microbiome samples. Since the gut microbiome is widely studied, this is our most important finding.

Based on our results, we conclude that the privacy threats related to microbiome identifiability are even higher than what was already acknowledged. In general, we observed a strong dependence of the performance of PMI techniques on the body site. While the temporal stability of samples from the gastrointestinal tract allows for true positive identifications in a high number of cases, the performance on body sites such as anterior nares is drastically worse. In addition, we also observed a pronounced influence of the feature type on the performance. Regardless of the applied PMI technique, a comparison between Figure 3 and Figure 4 shows that, on average, the performance on OTU-type features appears to be worse than on Marker-type features. A similar example arises from the generally weaker performance we observed on KBWindows features compared to Markers, which are both on the gene level. Overall, such findings may be useful in efforts to mitigate the privacy risks related to microbiome research.

Due to the current lack of public data availability, further experiments on larger datasets remain a prospect for future research (see Section 7.1). In addition, we will focus on mitigating the threats exposed in this paper. One approach that appears to work well against PMI is data *synthetization* ([7]). Another promising approach is data *anonymization* via micro-aggregation on microbiome samples.

REFERENCES

[1] A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*. IEEE Computer Society, Salerno, Italy, 21–29. https://doi.org/10.1109/SEQUEN.1997.666900

[2] Human Microbiome Project Consortium. 2012. A framework for human microbiome research. *Nature* 486, 7402 (2012), 215–221. https://doi.org/10.1038/nature11209

[3] Eleonora Distrutti, Lorenzo Monaldi, Patrizia Ricci, and Stefano Fiorucci. 2016. Gut microbiota role in irritable bowel syndrome: New therapeutic strategies. *World Journal of Gastroenterology* 22, 7 (2016), 2219–2241. https://doi.org/10.3748/wjg.v22.i7.2219

[4] Noah Fierer, Christian L. Lauber, Nick Zhou, Daniel McDonald, Elizabeth K. Costello, and Rob Knight. 2010. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences* 107, 14 (April 2010), 6477–6481. https://doi.org/10.1073/pnas.1000162107

[5] Eric A. Franzosa, Katherine Huang, James F. Meadow, Dirk Gevers, Katherine P. Lemon, Brendan J. M. Bohannan, and Curtis Huttenhower. 2015. Identifying personal microbiomes using metagenomic codes. *PNAS* 112, 22 (2015), E2930–E2938. https://doi.org/10.1073/pnas.1423854112

[6] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.* 42, 4, Article 14 (2010), 53 pages. https://doi.org/10.1145/1749603.1749605

[7] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. 2022. Utility and Privacy Assessment of Synthetic Microbiome Data. In *Data and Applications Security and Privacy XXXVI*. Springer International Publishing, To appear.

[8] Ruth E. Ley, Peter J. Turnbaugh, Samuel Klein, and Jeffrey I. Gordon. 2006. Human gut microbes associated with obesity. *Nature* 444, 7122 (2006), 1022–1023. https://doi.org/10.1038/4441022a

[9] Guang Li, Yadong Wang, and Xiaohong Su. 2012. Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. *Computer Methods and Programs in Biomedicine* 108, 1 (2012), 1–9. https://doi.org/10.1016/j.cmpb.2011.02.013

[10] William W. Lowrance and Francis S. Collins. 2007. Identifiability in Genomic Research. *Science* 317, 5838 (2007), 600–602. https://doi.org/10.1126/science.1147699

[11] Bradley Malin. 2005. Protecting genomic sequence anonymity with generalization lattices. *Methods Inf Med.* 44, 5 (2005), 687–692.

[12] Giovanni Musso, Roberto Gambino, and Maurizio Cassader. 2010. Obesity, Diabetes, and gut microbiota: the hygiene hypothesis expanded? *Diabetes Care* 33, 10 (2010), 2277–2284. https://doi.org/10.2337/dc10-0556

[13] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, Oakland, CA, USA, 111–125. https://doi.org/10.1109/SP.2008.33

[14] G.B. Rogers, D.J. Keating, R.L. Young, M.L. Wong, J. Licinio, and S. Wesselingh. 2016. From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. *Mol Psychiatry* 21, 6 (2016), 738–748. https://doi.org/10.1038/mp.2016.50

[15] Latanya Sweeney, Akua Abu, and Julia Winn. 2013. Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment). (2013).

[16] Justin Wagner, Joseph N. Paulson, Xiao Wang, Bobby Bhattacharjee, and Héctor Corrada Bravo. 2016. Privacy-preserving microbiome analysis using secure computation. *Bioinformatics* 32, 12 (02 2016), 1873–1879. https://doi.org/10.1093/bioinformatics/btw073

[17] Zicheng Wang, Huazhe Lou, Ying Wang, Ron Shamir, Rui Jiang, and Ting Chen. 2018. GePMI: A statistical model for personal intestinal microbiome identification. *npj Biofilms and Microbiomes* 4, 20 (2018). https://doi.org/10.1038/s41522-018-0065-2

[18] Hikaru Watanabe, Issei Nakamura, Sayaka Mizutani, Yumiko Kurokawa, Hiroshi Mori, Ken Kurokawa, and Takuji Yamada. 2018. Minor taxa in human skin microbiome contribute to the personal identification. *PLOS ONE* 13, 7 (July 2018), e0199947. https://doi.org/10.1371/journal.pone.0199947

[19] August E. Woerner, Nicole M.M. Novroski, Frank R. Wendt, Angie Ambers, Rachel Wiley, Sarah E. Schmedes, and Bruce Budowle. 2019. Forensic human identification with targeted microbiome markers using nearest neighbor classification. *Forensic Science International: Genetics* 38 (Jan. 2019), 130–139. https://doi.org/10.1016/j.fsigen.2018.10.003

[20] Tanya Yatsunenko, Federico E. Rey, Mark J. Manary, et al. 2012. Human gut microbiome viewed across age and geography. *Nature* 486, 7402 (2012), 222–227. https://doi.org/10.1038/nature11053

7 APPENDIX

In the following, we provide a runtime analysis, a discussion of possible variants of the technique, an additional experiment on the ability to detect true negatives, and additional auxiliary material.

7.1 Runtime analysis and scalability

We consider the overall complexity of our method and estimate the cost of handling data with a higher number of individuals and/or features.

For the sake of clarity, we first consider the cost of checking one single sample b against a dataset D_1. Let d be the number of features and n be the number of individuals in D_1. Our algorithm consists of three steps, namely the data preprocessing, the nearest-neighbor search and the thresholding. In the first step, we have to transform each cell of D_1, thus have to perform $O(dn)$ comparisons of floating point numbers. The cost for the feature selection step is negligible. The complexity of the nearest-neighbor search[9] in the second step is dominated by the cost for constructing a ball tree on D_1, which may be bounded by $O(dn \log n)$. The query time to find the nearest neighbor a in D_1 of our sample b then depends on the size of d and is often close to $O(d \log n)$. However, it can also reach $O(dn)$ in the worst case. Finally, for the thresholding in the third step of the algorithm, we first need to find the distance of a to its second-nearest neighbor in D_1 and then check if the distance $\Delta(a, b)$ is still smaller than that. Of course, we may again use the ball tree constructed in the second step, which implies that the cost of the thresholding is similar to that of a query. To summarize, we obtain the following runtime bounds.

(1) The cost for building the infrastructure, i.e. the preprocessing of D_1 and the construction of the ball tree, is bounded by $O(dn \log n)$.
(2) Subsequently, the cost for queries and the thresholding for every sample we want to check against D_1 is often reduced from $O(dn)$ to $O(d \log n)$.

We want to stress that these are worst-case upper bounds for the complexities. For instance, the number d of features is usually reduced in the feature selection in the first step of the algorithm, which improves the runtime of the following steps. We also want to add that the construction of a ball tree is not recommended if only a few samples are checked against D_1. In this case, the brute-force computation of distances might actually be faster.

As can be seen in Table 1, the dimensionality d of the feature vectors varies from small to large[10], namely from about 10^2 to 10^5. This allows us to estimate the influence of larger values for the number of individuals n on the runtime complexity. In any case, we expect to run our algorithm on datasets with thousands of individuals in a matter of seconds to minutes. While it would certainly be interesting to study the accuracy of our method on significantly larger datasets, we are not aware of publicly available microbiome data that contains more than a few hundred individuals and fits our experimental setup by containing two samples for each individual within a certain time frame. At the moment, such an analysis and the application of our method (or variants thereof) to more diverse datasets remain prospects for future research.

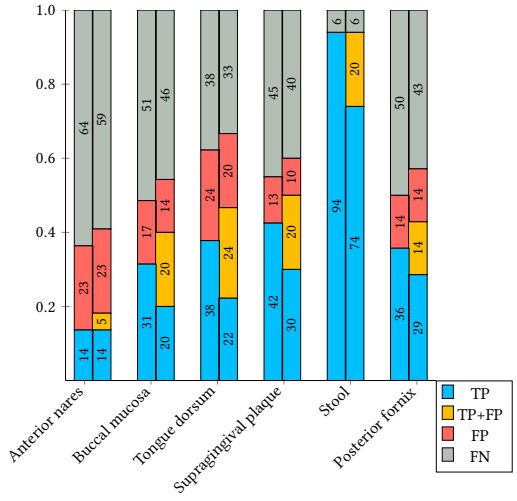


Figure 6: Results on Species in %. The left bar shows CorrD with $k = 1$ (see Figure 2a), the right bar shows CorrD with $k = 3$.

7.2 Variants of the technique

In this section, we introduce and discuss possible adaptations and variants of the distance-based technique for our PMI method described in Section 4. The first and already discussed possibility would be to consider other choices of the threshold for accepting neighboring sample pairs, or the distance(s) used for finding nearest neighbors and the thresholding. We have performed our experiments with several distances and similarity measures. Besides Canberra, Bray-Curtis and the Pearson coefficient, we investigated L1 (Manhattan), L2 (Euclidean), the Cosine similarity, the Jaccard index and several combinations of all these metrics for Δ_N, the distance used for the nearest-neighbor search, and for Δ_T, the measure used for thresholding. We observed mostly subtle differences in the performances, and the overall tendencies are well represented by the results discussed in Section 5. Still, it may be considered as an advantage of our method that it allows for the selection from a large pool of measures, which may lead to further refinements for certain data types.

The second possibility concerns the number k of nearest neighbors computed in the second phase of the algorithm. Considering not only the nearest neighbor of a sample from D_2, but also the second-nearest and maybe even the third-nearest, we will certainly find matches that are not returned by the method applied with $k = 1$. The adaptation concerns the second phase of the (unmodified) algorithm. For each column b in D_2, we use the ball tree to find and save the k-nearest neighbors of b in D_1, instead of just the nearest neighbor. The third phase would then be applied just like in the original description of the technique. Note that the dynamic definition of the thresholds for the nearest-neighbor pairs may lead to the behavior that the second-nearest neighbor pair is accepted,

[9] More details can be found at https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbor-algorithms
[10] For this reason, we recommend to use a ball tree instead of a K-D tree for personal microbiome identification. The query time of the latter does not scale well with the higher dimensionalities d occurring in the datasets.

while the nearest neighbor pair is rejected. On the other hand, we also expect to obtain a significant number of TP+FP matches in cases where both the nearest and the second-nearest neighbor pairs are accepted and the true match is among them.

In order to get a representative impression of the general effects of this adaptation, we consider $k = 3$ on the datasets for the feature type Species. The left bars in Figure 6 show the results with $k = 1$ discussed in Section 5, and the right bars demonstrate the outcome for the adaptation discussed above, where $k = 3$. We can see that the TP count is always smaller for $k = 3$, except for anterior nares, where it stays the same. However, the FP count is also smaller on three body sites, and equal on the remaining three. Additionally, if we add the TP+FP count to the TP count, the TP count of $k = 1$ is exceeded on every body site except for stool. This corresponds with our expectations described above. While TP+FP matches come with already discussed caveats, we conclude that the introduction of the additional parameter k appears to also allow for additional opportunities to refine our method.

7.3 Investigating true negatives

One capability of our PMI technique that has not been investigated in the main experiment in the previous subsections is to detect true negatives. Due to the already mentioned data availability issues, we modified the datasets from Franzosa et al. to simulate the occurrence of true negatives by randomly deleting samples in D_1. To be more concrete, we performed an experiment in the manner of a 10-fold cross validation. The individual samples in D_1 are shuffled and then divided into ten groups. Hence, each individual is in exactly one test group of negatives. We then start with the first group and delete all of its samples in D_1, which means that the corresponding samples in D_2 do not have a match and should be detected as negatives by our method. This step is repeated for each of the ten groups.

This experiment is also another practical test for our thresholding procedure in the third phase of the algorithm. In general, the true negative detection performance is expected to be high on datasets and bodysites where there are a lot of false negatives and only a few matches in the main experiment. It is thus more interesting to consider datasets where the false negative count is comparably low. We hence conducted the validation on the four datasets for the bodysite stool. While the datasets for this bodysite have a sufficiently large number of samples to allow for the setup described above, they also show a high number of samples that have been matched by our method in the main experiment. We applied our distance-based PMI approach as described in Section 4.

Table 4: True negative detection performance

Stool	Species	OTUs	Markers	KBW	Total
Mean TNR	0.78	0.72	0.74	0.61	0.71
Mean F1-Score	0.96	0.80	0.96	0.87	0.90

The results are shown in Table 4. 'TNR' denotes the true negative rate, i.e. the percentage of samples that have been detected as negative from those samples that are truly negative (those that belong to the group of samples that has been deleted in the respective run). In addition, we also consider the F1-score that has already been discussed in Section 5.3. For each dataset, we computed the mean of these two measures over the ten runs as described above. In the last column, we consider the overall mean of the cells in the same row. We can see that an average of 71% of true negatives is detected as such. Interestingly, our method appears to do better on the feature types Species, OTUs and Markers than on the feature type KBWindows, which also matches the performance pattern observed in the main experiment. While these results demonstrate the general capability of our method to detect true negatives, further investigations on suitable, preferably larger datasets will be necessary to corroborate them.

7.4 Auxiliary tables

In Tables 5 to 10, 'MetaC' shows the results of the approach described in [5, Fig.3], which we reproduced by applying the implementation with standard settings. The column 'CorrD' shows the results of the technique discussed in Section 4. In brackets, we display the TP count for a disabled nearest-neighbor threshold in the third phase of the technique.

Table 5: PMI on Species Features

Body Site	Ant. Nares		Buccal Mucosa		Tongue Dorsum		Suprag. Plaque		Stool		Post. Fornix	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
TP	1	3 (7)	1	11 (17)	0	17 (21)	1	17 (22)	12	47 (48)	0	5 (9)
FP	1	5	0	6	0	11	0	5	0	0	0	2
TP+FP	2	0	2	0	0	0	1	0	1	0	0	0
FN	7	14	6	18	1	17	7	18	18	3	3	7
NA	11	0	26	0	44	0	31	0	19	0	11	0

Table 6: PMI on OTU Features

Body Site	L Ante. Fossa		R Ante. Fossa		L Retro. crease		R Retro. crease		Ant. Nares		Buccal mucosa	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
TP	2	11 (11)	2	8 (12)	1	9 (16)	1	8 (15)	8	15 (24)	6	19 (33)
FP	1	33	0	20	2	22	1	18	1	16	2	16
TP+FP	8	0	4	0	3	0	7	0	6	0	9	0
FN	44	17	50	37	45	52	38	59	49	47	28	64
NA	6	0	9	0	32	0	38	0	15	0	54	0

Table 7: PMI on OTU Features

Body Site	Tongue Dorsum		Hard Palate		Saliva		Throat		Palatine Tonsils		Suprag. Plaque	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
TP	2	24 (31)	5	20 (27)	11	34 (43)	5	14 (23)	7	18 (27)	10	27 (38)
FP	2	19	2	14	4	10	1	12	0	15	3	20
TP+FP	2	0	13	0	11	0	11	0	10	0	24	0
FN	17	61	41	64	39	43	46	71	39	70	31	58
NA	81	0	37	0	22	0	34	0	47	0	37	0

Table 8: PMI on OTU Features

Body Site	Subg. Plaque		Kerat. Gingiva		Stool		Vaginal Introitus		Mid Vagina		Post. Fornix	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
TP	12	30 (43)	5	20 (23)	28	59 (63)	1	11 (14)	1	5 (7)	2	4 (5)
FP	10	6	1	17	1	8	0	9	0	10	0	5
TP+FP	22	0	3	0	2	0	1	0	1	0	1	0
FN	41	67	27	60	43	20	15	10	10	12	5	16
NA	18	0	61	0	13	0	13	0	15	0	17	0

Table 9: PMI on Markers Features

Body Site	Ant. Nares		Buccal Mucosa		Tongue Dorsum		Suprag. Plaque		Stool		Post. Fornix	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
TP	6	4 (5)	6	14 (16)	22	31 (36)	17	25 (27)	43	47 (48)	7	7 (11)
FP	0	2	4	4	2	5	2	4	1	0	0	2
TP+FP	1	0	2	0	0	0	0	0	0	0	2	0
FN	15	16	23	17	20	9	20	11	6	3	3	5
NA	0	0	0	0	1	0	1	0	0	0	2	0

Table 10: PMI on KBWindows Features

Body Site	Ant. Nares		Buccal Mucosa		Tongue Dorsum		Suprag. Plaque		Stool		Post. Fornix	
Method	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD	MetaC	CorrD
TP	2	2 (2)	7	2 (8)	12	7 (13)	10	7 (14)	31	36 (38)	7	7 (8)
FP	0	3	2	6	1	8	4	10	0	3	0	2
TP+FP	0	0	1	0	2	0	3	0	1	0	3	0
FN	3	4	13	18	23	23	19	19	13	6	1	2
NA	4	0	3	0	0	0	0	0	0	0	0	0