Forensic Microbiome: Novel applications of the human microbiome in forensic science

Celia Díez López

ISBN: 978-94-6416-753-5

Author: Celia Díez López

Cover design: Celia Díez López, created with BioRender

Layout: Celia Díez López

Printed by: Ridderprint

Forensic Microbiome: Novel applications of the human microbiome in forensic science

Forensisch Microbioom: Nieuwe toepassingen van het menselijke microbioom voor forensische wetenschappen

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. F.A. van der Duijn Schouten

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on
Tuesday 28 September at 13:00 hrs
by

Celia Díez López
born in Burgos, Spain.

Erasmus University Rotterdam

Doctoral Committee:

Promotor:	Prof.dr. M. Kayser
Other members:	Prof.dr. A.G. Uitterlinden Prof.dr. T. Sijen Assoc.-Prof.dr. J.P. Hays
Copromotor:	Assist.-Prof.dr. A. Vidaki

CONTENTS

Chapter 1

Introduction and aims of the thesis

My herewith presented PhD thesis deals with analysing the human microbiome for addressing forensically relevant questions commonly encountered in forensic casework practise. These questions are: How did the crime happen? When did the crime happen? Who is the perpetrator of the crime? To answer these important forensic questions by use of the human microbiome, I applied state of the art lab methods to *de novo* generate human microbiome data such as massively parallel sequencing, collected and used publicly available human microbiome data such as from the Human Microbiome Project, and developed and applied new methods for microbiome data analysis and interpretation such as deep learning neural networks. In the following, I will provide introductory background information on the main topics of this thesis.

The human microbiome

The microorganisms that live in and on the human body, including bacteria, archaea, lower and higher eukaryotes and viruses, are referred as the human microbiota (1, 2). The term microbiome refers to these microorganisms, their genomes and the surrounding environmental conditions (2). However, the term microbiome is also employed to refer to the collection of genes and genomes of the microbiota (2). This thesis focuses on the bacterial component of the microbiota, though in principle many of the discussed aspects apply more generally.

Microorganisms are found in almost every part of the body, from which the gut hosts the most complex and abundant microbial community and has been the focus of the majority of the human microbiome studies so far (3). Some researchers are also currently focusing on traditionally considered sterile body sites in medicine, like blood (4, 5) and placenta (6, 7), though much rigorous analysis pipelines are still needed to confirm the existence of indigenous microbial communities in such sites. Regarding the magnitude of human microbiota, initial estimates suggested that microbial cells surpass human cells by a factor of ten, although a recent revised estimate suggests a ratio closer to 1:1 (8). The genetic diversity of the human microbiome is estimated to surpass the human one by several orders of magnitude (1). For instance, just the gut microbiome is estimated to contain at least 100 times as many genes as the human genome (9), even though more dedicated studies are still needed to accurately quantify this. The human microbiota has co-evolved with humans over thousands of years to form a complex mutualistic relationship (10-12), providing essential functions that humans did not need to evolve in their own (9). The general interest in studying the human microbiome is greatly due to its key role in essential processes such as nutrient absorption (13, 14), immune system development (15-17) and protection against pathogens (18-20), among others. The ubiquity and main functions of the microbiome open up new possibilities in understanding health and disease, searching for potential preventive interventions and treatments, as well as new non-clinical applications such as in forensics.

Advances in high-throughput sequencing and bioinformatics analysis tools allowed for population-level surveys of the human microbiome. The Human Microbiome Project (HMP) Consortium was an initiative launched in 2007 funded by the United States National Institutes of Health (NIH) (21, 22). The first phase (HMP1) aimed to comprehensively characterize the human microbiome by means of 16S rRNA gene amplicon sequencing and whole genome shotgun (WGS) sequencing. For that, 300 healthy 18- to 40-years-old adults were sampled at

five major body sites (gastrointestinal tract, oral cavity, nasal cavity, skin and urogenital tract, primarily vaginal) with a total of 15 or 18 body sites depending on individual's sex. Individuals were sampled at one to three times, for a total of over 11,000 samples (21, 22). The second phase, the integrative Human Microbiome Project (iHMP), aimed to characterize the microbiome in three cohorts of microbiome-associated conditions: pregnancy and preterm birth, inflammatory bowel disease (IBD) and type 2 diabetes mellitus (T2DM) (23). This second phase was a multi-omics effort that integrated 16S rRNA gene, WGS, metatranscriptomics and metabolomics data. Over the course of the HMP, more than 3,000 microbial genomes from the human body were isolated and sequenced to also serve as reference in future studies (24, 25). Moreover, the Data Analysis and Coordinating center (DAAC) (26) was established for the development of new experimental and computational tools as well as resource repositories. Also in 2007, the European Union launched the project on Metagenomics of the Human Intestinal Tract (MetaHIT) (27). This project specifically focused on the gut microbiome of 124 subjects through WGS sequencing, exploring host-microbiome associations as well as potential treatments in various cohorts of different health statuses, including IBD and obesity (27, 28). These big projects did not only boost the knowledge on the human microbiome, but they also developed standard sampling and sequencing protocols as well as standards and quality-control guidelines to guide other future studies to produce accurate, comparable and reliable data (29).

Human microbiome variation

Across body sites

The human microbiome is body site-specific with striking differences, for instance in the communities inhabiting the gut, skin, vagina and oral sites (30, 31) (Figure 1). All body sites have the dominance of one to several genera, together with many less abundant taxa (32). Within a single body site, there are also notorious community differences based on sub-site's characteristics (30, 33). This can be expected in sites like skin, with large dimensions (around 25 square meters) of epithelial surface interacting with microbes (34) and differentiated physiological characteristics dominated by different taxa (30). For example, sebaceous areas on the face are dominated by *Propionibacterium* and *Staphylococcus* species, whereas moist sites like the armpit is dominated by *Corynebacterium* species (35). On the contrary, dry sites like the forearm is dominated by a diverse community of *β-Proteobacteria* and *Flavobacteriales* (35). Notably, there are even differences between the right and left skin sites of the same individual (36, 37). Costello *et al.* (36) reported that the right and left hands of the same individual share on average only 17% operational taxonomic units (OTUs) at species level, whereas Fierer *et al.* described an average of 13.5% for the right and left forearms of the same individual (37). Other sites with a much smaller microbial-interacting surface like the oral cavity also presents unique microbial assemblies on different sites (38, 39). For instance, buccal mucosa, keratinized gingiva and hard palate are dominated by *Streptococcus* species (40). Other oral niches such as saliva, tongue, tonsils, throat and gingival plaque are not dominated by a single genus but present a mixed microbial population that is evenly distributed (40). These body site differences are closely related with the microbial functions

that promote the host's well-being. For example, the vaginal microbiome is distinctly composed of lactic acid-producing bacteria, which prevent infectious diseases by lowering the pH of the vaginal environment, by producing antimicrobial compounds and through competitive exclusion (41, 42).

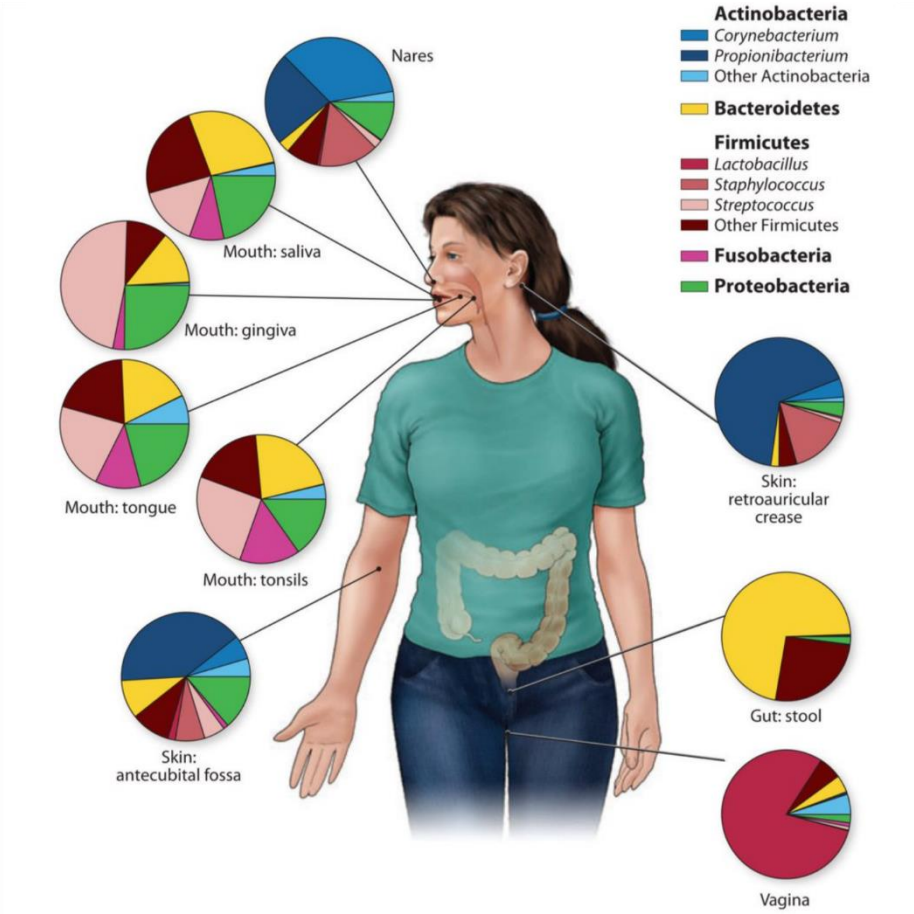


Figure 1. Human microbiome variation across body sites. Body site differences are related to specific microbial functions that promote the host's well-being. There are community differences within a single body site (e.g. oral cavity, skin) based on each sub-site's physiological characteristics. (Picture courtesy of Darryl Leja, National Human Genome Research Institute, NHGRI).

Among individuals

The high specificity of the human microbiome to different body sites leads to the definition of a 'core' microbiome which refers to those microbes shared among unrelated individuals (39). These core taxa have traditionally been described as highly prevalent (common in the population) and abundant (43). This is the case of *Fusobacterium*, *Streptococcus*, *Pasteurellaceae* and *Veillonella* in the oral cavity; *Propionibacterium* and *Staphylococcus* in

skin sites; and *Lachnospiraceae*, *Faecalibacterium*, *Oscillibacter* and *Bacteroides* in the gut (44). However, a minor core exists, although prevalent but less abundant, and should not be confused with transient microbes (45). The percentage of the entire microbial community that constitutes the core differs among body sites, with the highest number in the oral cavity, followed by the gut, skin and vagina (30, 43). The magnitude of the core also depends greatly on the analysed taxonomic level, where the taxa shared among individuals diminish down to the genus, species and strain levels (46). Moreover, individuals can be categorized into clusters based on the abundance of key 'core' bacteria for a particular body site (47). For example, stool samples can be categorized into four community types based on the relative abundance of five main genera: *Bacteroides*, *Prevotella*, *Ruminococcus*, *Alistipes* and *Faecalibacterium* (14, 48). Interestingly, Ding *et al.* (48) showed that the community type in one body site can be predictive of the community type at other body site. In particular, subjects with high *Prevotella* abundances in saliva were also likely to present the stool community type with the highest levels of that genus.

Nevertheless, there are also taxa present at low prevalence and low abundance, which also makes the human microbiome, at least to some extent, individual-specific (30, 49, 50); even between monozygotic twins (51). Inter-individual variation is higher in skin and vaginal sites compared to oral and gut sites (32). For instance, vaginal *Lactobacillus* can be the only detected genus in some individuals, while being absent in others (0-100% relative abundance range) that present a community dominated by other genera such as *Atopobium*, *Prevotella*, *Bifidobacterium* and unclassified members of the *Firmicutes* phylum (48). In the case of oral microbes the relative abundance range narrows down, e.g. *Prevotella* that presents a 2-47% relative abundance range (32). To note, despite individual-specific, co-habiting individuals have more similar microbiomes based on direct or indirect microbial transfer as well as shared environments and habits that shape their microbiota (52, 53).

Over time

The development and foundation of the human microbiome is a dynamic process parallel to host aging. Sequential succession of taxa occurs through infancy, childhood, adulthood and elderly with differences in terms of diversity and variation between the developmental stages (54). The first human colonizers depend on birth mode originating from the mother's vagina in case of naturally (vaginally) delivered babies or from the mother's skin in case of caesarean deliveries (55). From then on, the microbiome evolves until it converges on the adult one, though this process is body site-specific. For example, the phylogenetic diversity in the infant's gut microbiota is enriched linearly, resembling the adult's gut microbiome by the age of 2-3 years (56). However, the process is slowed down in the oral microbiome, where differences with the adult microbiome are still notorious by the age of 18 (57).

One of the pioneering and most comprehensive studies of the natural time series variation of the human microbiome at adulthood was carried out by Caporaso *et al.* (31). The authors took daily samples of the gut, skin and oral sites of the study participants for months, which were then analysed at a high-resolution species level. The authors reported that only a small fraction of the species found within a body site were present across all time points for an individual. This 'temporal core' was bigger in the analysed oral sites, followed by the gut

and skin sites. Other studies have reported similar results where the oral cavity presented the highest time-wise stability, followed by the gut and then less stable sites such as the skin, nares and vagina (30, 32, 49, 58-60). Remarkably, even though taxa preserve through time, their relative abundances change considerably (30, 59, 61). The least stable taxa are normally of low prevalence (rare in the population), which suggests that they are transient microbes (32, 58). This is particularly well-observed in the skin sites as these transient taxa could result from environmental exposure, such as transfer from other skin sites from the individual itself (skin-to-skin contact), between individuals (52, 62) or the built environment (63, 64). Apart from site-dependent, temporal stability is also individual-specific, especially in less stable sites like skin and vagina (32, 65). Nonetheless, intra-individual differences at different time points are smaller than inter-individual differences as observed in the oral cavity (31, 61), gut (31, 59) and skin sites (36, 37, 50).

In addition to the time stability of the human microbiome *in vivo*, some studies have also investigated its stability *ex vivo*, since sample storage and transportation can influence to some degree the microbial communities after leaving the human body, ultimately affecting the accuracy of the observations. When targeting specific species, oral *Streptococci* have been detected using quantitative PCR (qPCR) up to 92 days after deposition on cardboard or cotton fabric (66), and up to 62 days on other fabric types (67). At the community level, a couple of studies reported substantial changes in the community composition of faecal samples stored at room temperature for 72 h, particularly in the *Actinobacteria* and *Firmicutes* phyla abundances (68, 69); though these observations were not reproduced by a different study (70). For skin communities, time stability have been reported up to two weeks on swabs stored at room temperature (71). In 'touched' objects temporal stability is highly affected by environmental factors, such as season (52). However, a study analysing the microbial communities from surfaces in public spaces (e.g. floor, soap dispenser, toilet seat), mainly comprised of skin and gut associated taxa, reported community stability up to eight weeks (72). Lastly, the communities associated with pubic hair have been reported to be stable at room temperature for up to six weeks (73).

Associated with host's and environmental factors

Host's genetics play a role in shaping the microbiome (51, 74). Using the WGS sequencing data produced in the HMP, Blekhman *et al.* (75) reported host's genetics-microbiome interactions in ten out of the 15 analysed body sites. These were mainly induced by genetic variation in immunity-related pathways and occurred in genomic regions different among human populations, which indicate potential adaptation to environment-specific microbes. Also, host's sex (36, 76, 77) and ancestry (78-80) are factors associated with observed differences in microbial communities. One study analysing the vaginal microbiome found that women of European ancestry are more likely to present a *Lactobacillus*-dominated community, whereas women of African ancestry are more likely to have a much diverse community dominated by *Gardnerella vaginalis* and the uncultivated bacterial vaginosis-associated bacterium-1 (BVAB1) (81). A different study analysing the oral microbiome in individuals from Congo found that a high percentage of the community consisted of *Enterobacter,* while this genus was absent in individuals from China, Germany, Poland, Turkey and California (82).

An individual's lifestyle choices can also induce qualitative and quantitative changes in the microbiome with positive or negative effects. Diet in particular is one of the most powerful influences on the gut microbial communities (83); for example, the Western-like high-fat diet has been associated with an increased *Firmicutes:Bacteroidetes* ratio associated with weight gain and obesity-related pathologies (84, 85). However, exercise alone have the opposite effect in the gut microbiome (86, 87), not only affecting the *Firmicutes:Bacteroidetes* ratio, but also by the presence of other taxa directly associated with a low body mass index and a healthy metabolic profile, such as the genus *Akkermansia* (88).

The consumption of certain substances has also been associated with changes in the microbial communities. For instance, cigarette smoking is well-associated with alterations in the oral microbiota. More precisely, lower diversity and enrichment of taxa that lead to shifts in functional pathways have implications in smoking-related diseases (89-91). Also, alcohol can have a beneficial or detrimental effect in the microbiome depending on the alcoholic beverage and quantity. On the one hand, moderate and chronic alcohol consumption has been associated with altered gut microbial communities (92, 93). On the other hand, moderate intake of specific red wine polyphenols, even for only four weeks can have a prebiotic effect in the gut microbiome (94).

Shifts in the microbial composition has also been associated with multiple diseases such as inflammatory bowel disease (IBD) (95), obesity (96), type 2 diabetes mellitus (T2DM) (97), atherosclerosis (98), asthma (99) and autism (100), among others. However, it still remains unclear whether the differences in the microbiota observed in many diseases are a symptom of the disease or a contributing factor (101). Nevertheless, what is certain is that the treatment drugs have an impact on the microbial communities. Probably, the most obvious example is the use of antibiotics, which are broad-range drugs that do not only get rid of the pathogen, but also of an important percentage of the healthy bacterial community (102). Interestingly, the effect of antibiotics also depends on the body site, with the throat and saliva communities recovering faster than the gut (103, 104). Other broadly used drugs have also been associated with microbial shifts, such as non-steroid anti-inflammatory drugs (NSAIDs) (e.g. celecoxib, ibuprofen). These have effects on the gut microbiome, with a higher effect observed with the type of NSAID than with the number of drug administrations (105). There are even discriminatory taxa for individuals using NSAIDs alone or NSAIDs in combination with antidepressants and laxatives (105).

Generating human microbiome data

Traditional microbiology studies individual species as isolated units. However, only a very small percentage of the vast microbial diversity has been successfully isolated as viable specimens for analysis. This is because microbes need very specific environmental conditions to growth which are extremely difficult, or cannot be, experimentally reproduced. Advances in high-throughput sequencing in the mid 2000's was a key contributor to the development of the new field of metagenomics, allowing for the analysis of the whole microbial communities and moving beyond culture-based methods.

In DNA-based human microbiome studies, sample collection is performed via a low-invasive procedure whenever possible (e.g. swabbing), which ensures sufficient amounts of recovered microorganisms for downstream analysis (106). For DNA isolation, there is a great variety of protocols for different sample types (106). Regarding the sequencing platform, different options for insert size, read length, depth, sequence accuracy and cost exits (107). Since the Roche 454 pyrosequencing platform (used in the HMP) become outdated in 2013, Illumina platforms are the preferred ones in microbiome studies based on their high throughput and typical low error rates, even though there are other alternatives, such as the Ion Torrent S5 and S5 XL instruments (108). Also, long-read sequencing technologies, including Oxford Nanopore MinION and Pacific Biosciences Sequel platforms, have recently gained popularity based on the huge amounts of throughput data they can produce, though further lowering their current error rates is an ongoing task (108). DNA-based microbiome studies normally fall in one of the two following categories: (a) targeted amplicon sequencing studies targeting one or a few marker genes to characterize the microorganisms present in a sample and (b) whole genome shotgun (WGS) sequencing studies which also reveal the genetic content in a sample. Both approaches have been proven to be very useful and the choice of one or the other depends on the aims and applications of a study. (Figure 2).

Targeted amplicon sequencing

Targeted amplicon sequencing technique is based on the amplification of small fragments of one or a few marker genes. The sequences from these fragments are obtained via high-throughput sequencing and compared against a reference for taxonomy assignation. By this, researchers can characterize the microorganisms present in a sample. The prokaryotic ribosomal small subunit - 16S ribosomal RNA (rRNA) - gene is the most-widely used phylogenetic marker for the characterization of bacteria and archaea in a sample. Other taxa such as fungi can also be targeted by using other markers, such as the eukaryotic 18S rRNA gene (109) or the internal transcribed spacer (ITS) (110). These ribosomal genes are essential and hence, present in all living organisms. Furthermore, they contain conserved (slowly evolving) regions that can be used to design broad-spectrum PCR amplification primers and hypervariable (fast evolving) regions that can be used to classify organisms (106). However, the amplification PCR primer set of choice can be a source of technical bias (111). For example, the V2 and V4 hypervariable regions of the 16S rRNA gene have been proven to be much more taxonomically informative than the V6 hypervariable region (112). Also, commonly-used primer pairs lack sensitivity for some taxa, such as the F27-R338 for *Bifidobacterium* (113), which is an important community member of the gut microbiota. Regarding classification there are a number of large databases of data produced for 16S rRNA gene and other marker genes that facilitate the process, such as Greengenes (114), SILVA (115), the Ribosomal Database Project (RDP) (116) as well as more specialised ones like the Human Oral Microbiome Database (HOMD) (117). With a targeted approach, classification resolution normally goes down to the genus level (118). Other limitations of this approach include copy number variation (119) and horizontal transfer of the entire marker gene region (120).

After PCR amplification, short indices (or tags) are attached to the PCR amplicons specific to a given sample. This strategy allows for the parallel sequencing of a certain number of samples and the subsequent assignment of the generated sequences to each sample (121). Following sequencing, the traditional approach is to minimize sequencing errors and PCR-based chimeric sequences by clustering similar sequences together based on a percent dissimilarity threshold, normally 3% for species and 5% for genus (122); generating the so-called Operational Taxonomic Units (OTUs). These OTUs can be defined by one of the following three clustering methods: (i) the *closed reference* method in which sequences are clustered against a reference database (114-117), where sequences that do not match the database are discarded; (ii) the *de novo* method in which sequences are grouped into OTUs as a function of their pairwise similarities independently of any external database; and (iii) the *open reference* method which combines the closed reference method first followed by the de novo method, so sequences that do not match the reference database are not discarded (123, 124). However, in the last few years, the microbiome community is moving towards single-nucleotide resolution methods to deal with errors, particularly Amplicon Sequence Variants (ASVs) (125), which distinguish sequences that differ in as little as one nucleotide. This method builds an error model in which real biological sequences are discriminated from error sequences based on frequency probabilities. ASV methods allow for higher resolution than OTU methods and have been proven to be better at discriminating ecological patterns (125, 126). In studies focusing on either OTUs or ASVs, the subsequent step is generally taxonomy assignation, by which each OTU or ASV is labelled with a taxonomy label (i.e. kingdom, phylum, class, order, family, genus – sometimes also species) (Figure 2).

Whole genome shotgun sequencing

The information provided by WGS sequencing is much broader than with targeted approaches since observations are not based on a single or a few marker genes, but in an overview of the gene composition of a microbial community (127). However, in WGS additional mechanical or chemical steps should be taken during sample processing to ensure that minimal host (human) DNA is obtained, which could 'overwhelm' the microbial fraction of interest during the sequencing process (127, 128). For those samples that yield very small amounts of DNA (e.g. biopsies), the amplification of the starting material might be required to comply with the library production DNA amount requirements from the majority of the sequencing technologies (127). For that, one method is multiple displacement amplification (MDA) that employs random hexamers and phage phi29 polymerase (129). As with any amplification method, sequence bias in the amplification and chimera formation can introduce undesired variation on subsequent analysis.

Particularly, WGS sequencing consists on the untargeted ('shotgun') sequencing of all the genomes present in a sample (108). In this method, the DNA present in the sample is randomly breaking up in smaller fragments prior to sequencing. The resulting sequences can be used directly as short fragments or they can be assembled to obtain longer genomic regions (contigs). Similarly to what is done in targeted amplicon sequencing, assembly can be *reference-based* or *de novo* (127). Assembly allows for longer sequences; hence, more information provided. It also results in a higher quality of information since multiple reads cover

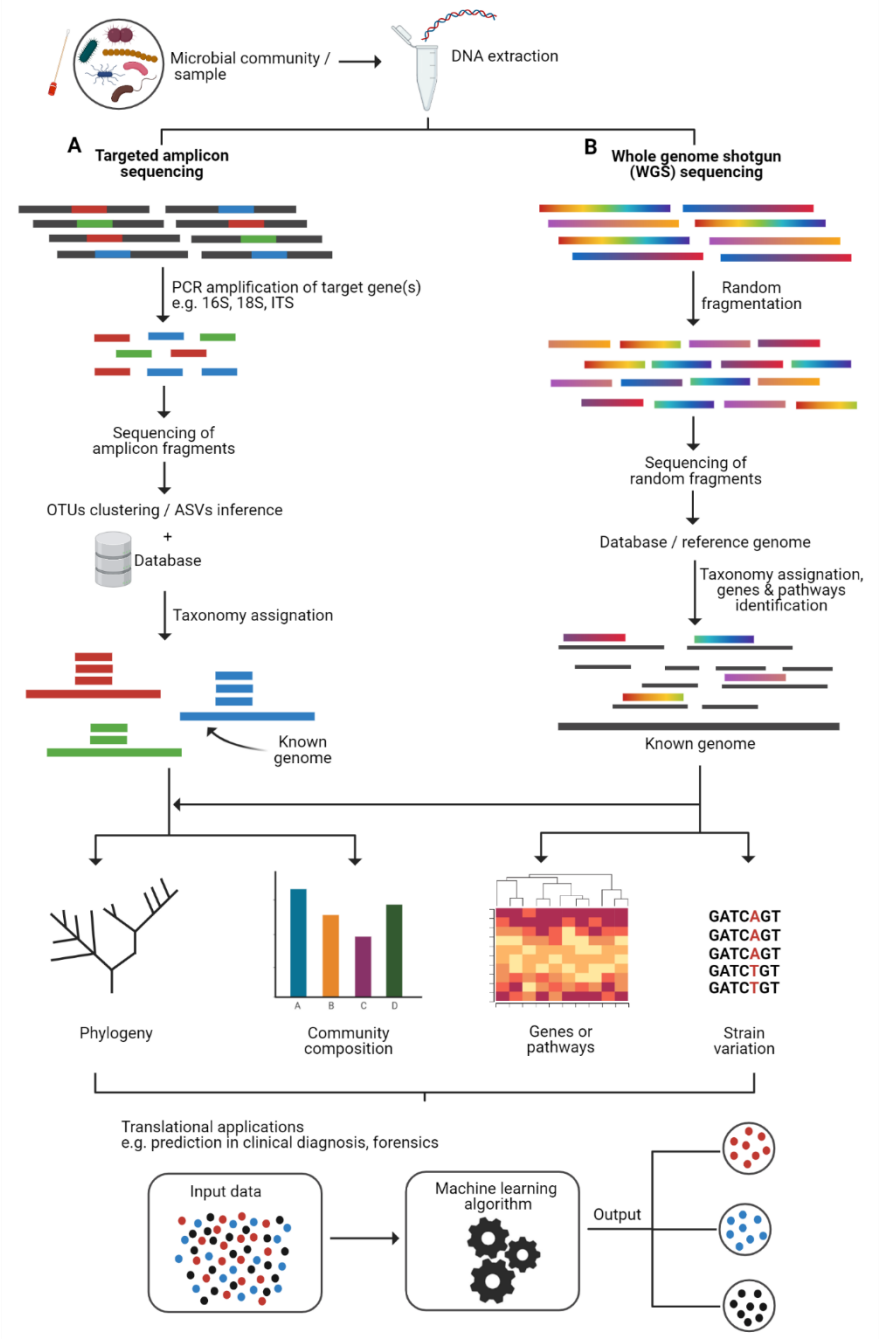


Figure 2. Overview of experimental and analytical tools for studying the human microbiome from obtaining microbial DNA of a sample to its translational applications. **A**. Targeted amplicon sequencing. **B**. Whole genome shotgun (WGS) sequencing. (Picture created with BioRender).

the same segment of genetic information in those cases where sequencing depth allows for it (127). Moreover, the sequences can be sort into groups that might represent an individual genome or genomes from closely related organisms in a process known as binning. The sequences are annotated to identify which species, or even strains, are present in a sample (106, 108) by comparison to reference databases of whole genomes (e.g. Kraken (130)) or selected marker genes (e.g. MetaPhlAn (131), mOTU (132)). The sequences can also be used to perform other analyses, such as metabolic function profiling and antibiotic resistance gene profiling, which are indispensable to uncover the immensity of microbial diversity (106, 108) (Figure 2).

Analysing and interpreting human microbiome data

The data obtained from a microbiome study can have many forms, such as matrices of taxa abundances, diversity measures within (i.e. alpha-diversity) and between (i.e. beta-diversity) samples and PCs of principal coordinates analysis (PCoA) (123). All these can reveal microbial ecology patterns that might be associated with certain health and physiological states (123, 133). It is important though to consider the particular characteristics of high-throughput sequencing microbiome data, particularly when analysing OTUs/ASVs or taxa abundances. The microbiome data typically presents a zero-inflated distribution, since many taxa are not shared among the majority of the analysed samples. Moreover, the highly dimensional microbiome data is of compositional nature. This means that its component taxa compete to make up the constant sum constraint (134), which is determined by the sequencing depth (135). In other words, the sequencing platforms have a fixed number of 'slots' that limits how many sequencing reads they can deliver (136). The absolute number of sequencing reads obtained for each taxa is irrelevant since the count cannot be related back to the actual number of microbial cells in the original sample. The actual information is in the proportions between the counts where the change in the relative abundance of one taxon inevitably affects the relative abundance of others based on the constant sum constraint (135, 136).

Nowadays, the microbiome field is moving from characterization and association towards translational applications. These include clinical diagnosis (137-140), outcome prediction in personalised interventions (141, 142) as well as more specialized applications such as forensics (73, 143-145). This current trend stems in the advances in high-throughput sequencing technologies, the development of analysis tools and the increasing availability of public microbiome data. For such applications, machine learning (ML) methods are commonly employed, not only to identify microbial community differences between groups of samples (e.g. control *vs.* disease), but also to learn from the data and predict which group a new sample belongs to (123, 146) (Figure 2). In supervised ML a known dataset (training set) is used to develop the predictive model. Each training data point consists of a set of input features (e.g. relative taxa abundances) and a dependent variable giving the correct classification (i.e. qualitative variable) or regression (i.e. quantitative variable) of that data point (133). The prediction model is developed on the basis of learning the relationships between the input

features including their discriminatory power for the dependent variable. The prediction performance of the developed model is then assessed in new (unknown) data (test set).

The microbiome field tends to use complex models, such as random forest (RF) (73, 138, 139, 142, 145, 147) and neural networks (NN) (140, 141), because of their ability in capturing nonlinear relationships in the data. However, these models are considered as 'black boxes', offering very limited interpretability. RF is a tree-based ensemble method that combines several models to improve the outcome predictions. It generates a large number of decision trees on different subsamples and combines their outputs using averages at the end of the learning process. For example, Rohit *et al.* (147) used RF based on 37 bacterial species for the accurate diagnose (AUC=0.94) of advanced fibrosis in non-alcoholic fatty liver disease. NN consist of interconnected groups of nodes that mimic the neurons in a brain in a simplified manner in order to recognize patterns in the data. Ashwin *et al.* (141) employed a NN algorithm that combined microbiome and clinical data to predict patient responses to inflammatory bowel disease (IBD) treatment with high accuracy (AUC=0.87). Other simpler models like logistic regression (LR) (143, 148) and support vector machine (SVM) (139, 149) are not so popular, even though they have previously shown high performance in different prediction problems and are easier to interpret. LR is a parametric method that assumes a linear dependency between the input features and the categorical outcome. The output of the LR linear function is a probability (x) between 0 and 1, where if $x<0.5$ the categorical outcome is one class, otherwise the other class. Beck *et al.* (148) used LR for the classification of bacterial vaginosis with a 95% of accuracy. SVM methods can support linear and non-linear solutions. These methods take the data points and find a separating hyperplane between them. For example, Lan *et al.* (149) employed SVM for host's age prediction from gut metagenomes with an accuracy of 65%.

Previous applications of the human microbiome in forensic science

The human microbiome presents features relevant for forensics, such as ubiquity (8), response to changes in the environment (83, 89) and capacity of being shed, deposited and exchanged (62, 150). Various studies have analysed the microbiome in different forensically relevant human samples like vaginal secretions (151), saliva (152), skin (52) or hair (73), in order to answer relevant forensic questions, such as individual identification (52, 63) and post-mortem interval estimation (153, 154), among others (Figure 3). However, the use of the human microbiome in forensic science is in its infancy. When this thesis work started, the use of complete microbiome information, instead of selected candidate species, was limited in some applications, such as in tissue and body fluid identification. With further research and appropriate forensic validation, forensic microbiome can be a very useful investigative tool, especially in cases where traditional non-microbial (human) biomarkers present limitations.

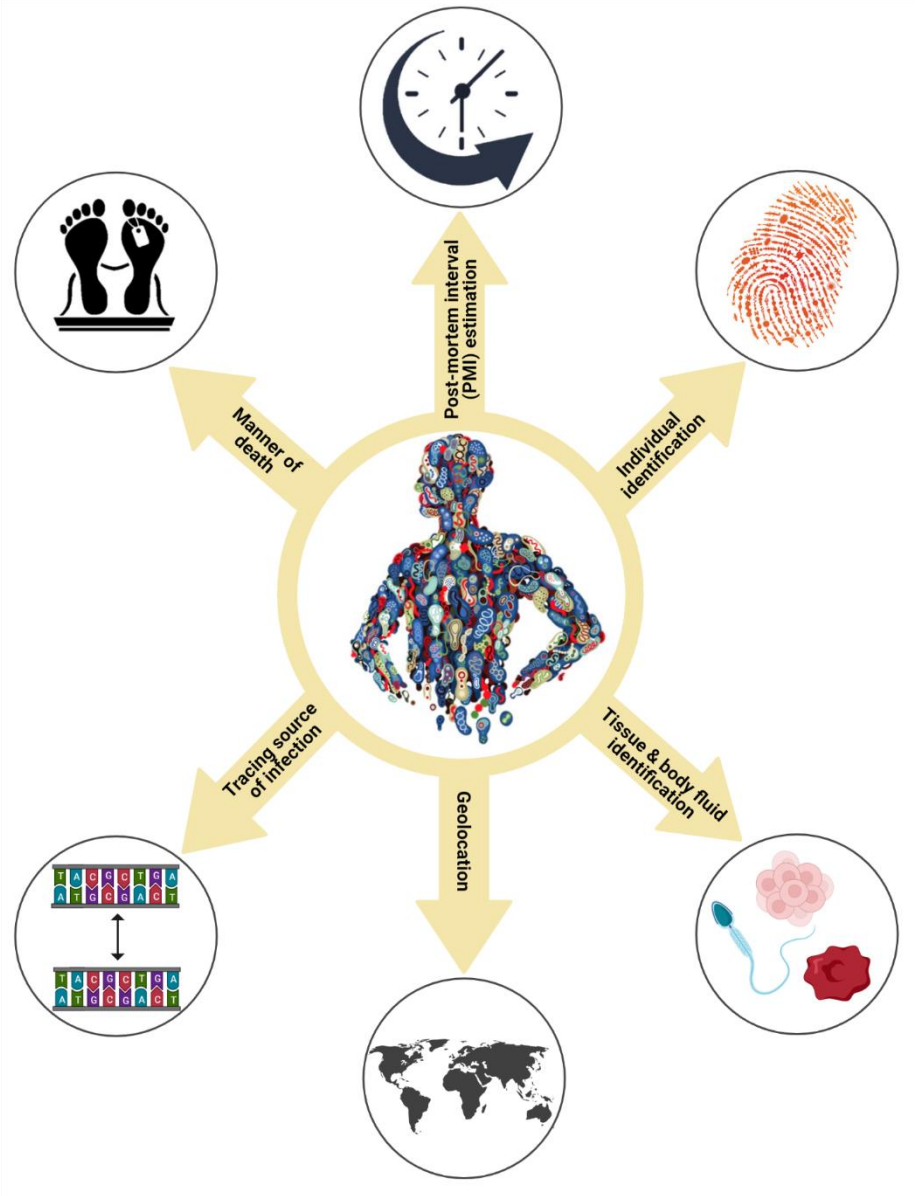


Figure 3. Applications of the human microbiome in forensic science when this thesis work started. (Picture created with BioRender).

Individual identification

The incredible variation in the microbial community among individuals (30, 50) and particularly the one that is also relatively stable over time (30, 32), open up the possibility of exploring the human microbiome as a 'fingerprint' to distinguish individuals among populations. Notably, humans shed around 30 million bacterial cells per hour in their surroundings (150), which can

be transmitted by direct contact to objects (155, 156) and other individuals (62), or on indoor surfaces in the form of aerosols (157). This can be especially relevant in investigations where the recovered human DNA is not of sufficient quantity and/or quality to obtain a full individualized DNA profile based on short tandem repeats (STRs); for example, when dealing with 'touched' samples (158).

Franzosa *et al*. (58) analysed gut metagenomes at the strain-level resolution and developed a model of matching probabilities that resulted in the identification of 80% of the studied individuals (N=242) after a period of a year. To note, the authors of that study raised privacy concerns for individuals enrolled in microbiome research projects. Moreover, using 16S rRNA gene amplicon sequencing various studies have linked objects (e.g. mobile phones, computer equipment) with their owners by comparing the similarities in the microbial composition of the 'touched' sample with the one obtained on the skin of the study individuals (52, 63, 71, 156). For instance, Wilkins *et al.* (52) reported a 67% of accuracy when linking individuals with household surfaces. However, accuracy diminished as the time span between sample deposition and sample collection increased, due to the loss of individualised taxa on the surfaces. Leake *et al.* (159) demonstrated that the saliva microbiome is also capable of differentiating individuals by using a combined amplicon sequencing approach of the 16S rRNA, rpoB genes and hierarchical clustering. Streptococcal DNA from teeth has also being used to identify individuals from bite marks using logistic regression with the highest possible probability of correct identification being reported for the rpoB gene amplicon sequencing (143). Additionally, the microbiome of pubic hair has also been investigated in this context based on its relevance in alleged sexual assaults. Williams *et al.* (73) built a RF model using 16S rRNA gene data from pubic hairs and predicted both the individuals and their gender. A different study even linked sexual partners when the collection of pubic hair samples happened up to 18 h after intercourse and both individuals had showered in the interim time (53). Overall, individual identification has been investigated so far on the basis of comparing individuals in the same study; however, for real forensic value microbiome 'fingerprints' should be able to individualise any random person from the population, which would require appropriate statistical frameworks.

Post-mortem interval estimation

The succession of the human microbiome after death has been proposed as a suitable biomarker for the post-mortem interval (PMI) estimation of cadavers. Using 16S rRNA gene amplicon sequencing various studies have characterized the microbiome composition at different body sites during the decay process, including spleen, liver, brain, heart, blood, bones, gut, skin and oral cavity (144, 145, 153, 154, 160-162). Some of these studies have reported confounding effects of the individual (162), sex (160) and abiotic factors (e.g. ambient temperature, solar irradiance) (161, 162) on the microbiome succession. Johnson *et al.* (144) sampled four decomposing human cadavers at the nasal and ear canals every two-three days until the tissues were too decomposed for being accessed (500 accumulated degree days (ADD), a time measurement that accounts for temperature effects on decomposition rate). The authors built a k-nearest neighbor (KNN) regression model using the skin microbiome data and reported an estimation error of ±55 ADD. In a later study, Metcalf *et al.* (145) analysed the microbial succession on skin and cadaver-associated soil in two human corpses

left outdoors in winter and other two corpses in spring. The microbial succession seemed to be robust across soil types and seasons. In the winter experiment, daily sampling provided the most accurate PMI estimation - root-mean-square error (RMSE) of 4 days.

Geolocation

Geolocation can be performed from body samples as well as from human microbiome traces left on objects and surfaces ('touched' samples). Regarding the former, higher ratios of *Firmicutes*:*Bacteroidetes* phyla in the gut have been associated with higher latitudes, an observation robust to both sex and age (163, 164). Also, community composition differences between developed and underdeveloped countries, as well as urban and rural populations, have been described for the gut, skin and oral microbiomes (165-168), although cultural and societal components are cofounded with the observed geographical differences (169). Remarkably, differences between people in developed countries and uncontacted indigenous people are dramatic, with the second presenting communities of the highest diversity and genetic functions reported so far (167). In a forensic case context, the geographic emergence of the strains of the stomach commensal *Helicobacter pylori* (170, 171) was proved useful to identify the geographical origins of unidentified cadavers (172). Regarding 'touched' samples, community composition of objects from owners living in the same city were more similar between them compared to objects from owners living in different cities (63, 64, 173).

Tissue and body fluid identification

Identifying the tissue/body fluid that a biological crime stain originated from can be very helpful in reconstructing the crime scene and associated sequence of events. The human microbiome can be an outstanding tissue/body fluid biomarker based on the differences of microbial communities across body sites (30, 33) and because the intra-individual differences among body sites are larger than the inter-individual differences for the same body site (30, 31). Most studies so far have focused on the identification of vaginal secretions by targeting vaginal commensals via qPCR, capillary electrophoresis or microarray analysis, mainly *Lactobacillus* species such as *L. crispatus* and *L. gasseri* (151, 174-177). Fleming *et al.* (151) detected those two *Lactobacillus* species in 14 women with very different hormonal conditions (pregnant, pre-menstrual, post-menstrual and hysterectomy). However, other studies failed to detect the same two species in all participants (174-176), as they were replaced by other bacteria performing the same lactic-acid producing function (78, 175). Also, the targeted species were detected in other body fluids, such as female urine (probably as a result of contamination from vaginal secretions), semen, gastric mucosa and anorectal mucosa (174, 176), which can ultimately lead to false positive results. From those studies, only the one carried out by Benschop *et al.* (175) had a relatively big sample size (N=240).

Other studies have attempted to identify other body fluids using PCR and qPCR approaches. For instance, streptococcal DNA has been used to identify saliva (152) as well as expirated blood (forced by airflow out of the nose or mouth) (66, 67). However, as with the previous vaginal examples, not all the targeted *Streptococcus* species were detected in all the

20 participants (152). A different study targeting *Bacteroides* species to identify faeces also reported inter-individual variation in the 20 analysed subjects (178).

Other forensic applications

Other microbiome applications in forensic medicine that have been reported in the literature include identification of the manner of death. For example, the presence of candidate oral commensals or water faecal pollutants in heart blood can serve as indicators of drowning (179, 180). On a different point, genotyping microbes responsible for sexually transmitted diseases can be useful to trace the source of infection, when no samples are taken at the immediate time of an alleged sexual assault or in cases where DNA evidence is not incriminating (181). By this, the suspect and victim can be linked, which has proven useful in child abuse investigations (182, 183).

Aims of this thesis and summary per chapter

In the last years, the acquired knowledge regarding the human microbiome has resulted from advances in high-throughput sequencing, analysis tools and increasing publicly available data. The microbiome field is very active, regarding not only biological-based research, but also experimental and analytical tools. The microbiome insights have brought new perspectives and possibilities to many fields, including forensic science. This thesis aims to investigate novel uses of the human microbiome in forensic science, both by shedding light in novel unsolved forensic questions and by assessing new approaches for common forensic questions where other non-microbiome methods show limitations.

This thesis is divided into three main parts. ***Part I*** describes the use of the human microbiome to infer the body site of origin of single-source human biological traces found at crime scenes. This was achieved based on a taxonomy-independent microbiome approach and deep neural networks for classification purposes, together with a laboratory tool suitable for forensics. ***Chapter 2*** focuses on human biological traces of epithelial origin; namely, saliva, skin and vaginal secretions. ***Chapter 3*** focuses on blood traces of different sites of origin; namely, nasal blood, menstrual blood, fingerprick blood and venous blood. ***Part II*** describes the use of the human microbiome for estimating the time since deposition of human biological traces. ***Chapter 4*** describes the time since deposition estimation of dried saliva stains. This was achieved by analysing time-wise informative bacterial species using qPCR and the development of individualised multiple linear regression models for the time prediction. ***Part III*** describes the use of the human microbiome to infer an individual's lifestyle habits from human biological traces. Particularly, ***Chapter 5*** describes the prediction of an individual's current cigarette smoking habit from saliva microbiome data. This was achieved by following a generalizable strategy to deal with the common problem of imbalanced data in supervised binary classification. Finally, ***Chapter 6*** provides a general discussion on the research presented in *Chapters 2-5* as well as future perspectives and various considerations in the research and application of the human microbiome in forensic science.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The Human Microbiome Project. Nature 449:804-810.
2. Marchesi JR, Ravel J. 2015. The vocabulary of microbiome research: a proposal. Microbiome 3:31.
3. Dave M, Higgins PD, Middha S, Rioux KP. 2012. The human gut microbiome: current knowledge, challenges, and future directions. Transl Res 160:246-257.
4. Paisse S, Valle C, Servant F, Courtney M, Burcelin R, Amar J, Lelouvier B. 2016. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. Transfusion 56:1138-1147.
5. Damgaard C, Magnussen K, Enevold C, Nilsson M, Tolker-Nielsen T, Holmstrup P, Nielsen CH. 2015. Viable bacteria associated with red blood cells and plasma in freshly drawn blood donations. Plos One 10:e0120826.
6. Cao B, Stout MJ, Lee I, Mysorekar IU. 2014. Placental microbiome and its role in preterm birth. Neoreviews 15:e537-e545.
7. Antony KM, Ma J, Mitchell KB, Racusin DA, Versalovic J, Aagaard K. 2015. The preterm placental microbiome varies in association with excess maternal gestational weight gain. Am J Obstet Gynecol 212:653.e1-16.
8. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 14:e1002533.
9. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of the human distal gut microbiome. Science 312:1355-1359.
10. Moeller AH, Li Y, Mpoudi Ngole E, Ahuka-Mundeke S, Lonsdorf EV, Pusey AE, Peeters M, Hahn BH, Ochman H. 2014. Rapid changes in the gut microbiome during human evolution. Proc Natl Acad Sci U S A 111:16431-16435.
11. Zaneveld J, Turnbaugh PJ, Lozupone C, Ley RE, Hamady M, Gordon JI, Knight R. 2008. Host-bacterial coevolution and the search for new drug targets. Curr Opin Chem Biol 12:109-114.
12. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R. 2008. Evolution of mammals and their gut microbes. Science 320:1647-1651.
13. Jumpertz R, Le DS, Turnbaugh PJ, Trinidad C, Bogardus C, Gordon JI, Krakoff J. 2011. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. Am J Clin Nutr 94:58-65.
14. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto J-M. 2011. Enterotypes of the human gut microbiome. Nature 473:174-180.
15. Gaboriau-Routhiau V, Rakotobe S, Lécuyer E, Mulder I, Lan A, Bridonneau C, Rochet V, Pisi A, De Paepe M, Brandi G. 2009. The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. Immunity 31:677-689.
16. Olszak T, An D, Zeissig S, Vera MP, Richter J, Franke A, Glickman JN, Siebert R, Baron RM, Kasper DL. 2012. Microbial exposure during early life has persistent effects on natural killer T cell function. Science 336:489-493.
17. Ivanov II, de Llanos Frutos R, Manel N, Yoshinaga K, Rifkin DB, Sartor RB, Finlay BB, Littman DR. 2008. Specific microbiota direct the differentiation of IL-17-producing T-helper cells in the mucosa of the small intestine. Cell Host Microbe 4:337-349.
18. Aagaard K, Riehle K, Ma J, Segata N, Mistretta TA, Coarfa C, Raza S, Rosenbaum S, Van den Veyver I, Milosavljevic A, Gevers D, Huttenhower C, Petrosino J, Versalovic J. 2012. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. PLoS One 7:e36466.
19. Frank DN, Feazel LM, Bessesen MT, Price CS, Janoff EN, Pace NR. 2010. The human nasal microbiota and Staphylococcus aureus carriage. PLoS One 5:e10598.
20. Gao Z, Tseng C, Strober BE, Pei Z, Blaser MJ. 2008. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. PLoS One 3:e2719.
21. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng QD, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Brooks JP, Buck GA, Buhay CJ, Busam DA, Campbell JL, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207-214.
22. Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature 486:215-21.

23. Integrative HMP (iHMP) Research Network Consortium. 2014. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe 16:276-89.
24. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C, Cole JR, Ding Y, Dugan S, Field D, Garrity GM, Gibbs R, Graves T, Han CS, Harrison SH, Highlander S, Hugenholtz P, Khouri HM, Kodira CD, Kolker E, Kyrpides NC, Lang D, Lapidus A, Malfatti SA, Markowitz V, Metha T, Nelson KE, Parkhill J, Pitluck S, Qin X, Read TD, Schmutz J, Sozhamannan S, Sterk P, Strausberg RL, Sutton G, Thomson NR, Tiedje JM, Weinstock G, Wollam A, Genomic Standards Consortium Human Microbiome Project Jumpstart C, Detter JC. 2009. Genomics. Genome project standards in a new era of sequencing. Science 326:236-7.
25. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, et al. 2010. A catalog of reference genomes from the human microbiome. Science 328:994-9.
26. Human Microbiome Project Data Analysis and Coordination Center. 2017. https://www.hmpdacc.org/hmp.
27. Ehrlich SD, The MetaHIT Consortium. 2011. MetaHIT: The European Union Project on metagenomics of the human intestinal tract, p 307-316, Metagenomics of the human body. Springer.
28. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59-65.
29. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M, Grp NHW. 2009. The NIH Human Microbiome Project. Genome Res 19:2317-2323.
30. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694-1697.
31. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R. 2011. Moving pictures of the human microbiome. Genome Biol 12:R50.
32. Zhou Y, Gao H, Mihindukulasuriya KA, La Rosa PS, Wylie KM, Vishnivetskaya T, Podar M, Warner B, Tarr PI, Nelson DE. 2013. Biogeography of the ecosystems of the healthy human body. Genome Biol 14:1-18.
33. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73.
34. Gallo RL. 2017. Human skin is the largest epithelial surface for interaction with microbes. J Invest Dermatol 137:1213-1214.
35. Schommer NN, Gallo RL. 2013. Structure and function of the human skin microbiome. Trends Microbiol 21:660-668.
36. Fierer N, Hamady M, Lauber CL, Knight R. 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci U S A 105:17994-17999.
37. Gao Z, Tseng CH, Pei ZH, Blaser MJ. 2007. Molecular analysis of human forearm superficial skin bacterial biota. Proc Natl Acad Sci U S A 104:2927-2932.
38. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43:5721-5732.
39. Zaura E, Keijser BJF, Huse SM, Crielaard W. 2009. Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9:259.
40. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C, Izard J. 2012. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13:R42.
41. Witkin SS, Linhares IM, Giraldo P. 2007. Bacterial flora of the female genital tract: function and immune regulation. Best Pract Res Clin Obstet Gynaecol 21:347-354.
42. Graver MA, Wade JJ. 2011. The role of acidification in the inhibition of Neisseria gonorrhoeae by vaginal lactobacilli during anaerobic growth. Ann Clin Microbiol Antimicrob 10:1-5.
43. Li K, Bihan M, Methé BA. 2013. Analyses of the stability and core taxonomic memberships of the human microbiome. PLoS One 8:e63139.
44. Huse SM, Ye YZ, Zhou YJ, Fodor AA. 2012. A core human microbiome as viewed through 16S rRNA sequence clusters. Plos One 7:e34242.

45. Reeder J, Knight R. 2009. The 'rare biosphere': a reality check. Nature Methods 6:636-637.
46. Blaser MJ. 2010. Harnessing the power of the human microbiome. Proc Natl Acad Sci U S A 107:6125-6126.
47. Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R. 2013. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol 9.
48. Ding T, Schloss PD. 2014. Dynamics and associations of microbial community types across the human body. Nature 509:357-360.
49. Rajilić-Stojanović M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, Smidt H, De Vos WM. 2009. Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol 11:1736-1751.
50. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA, Progra NCS. 2009. Topographical and temporal diversity of the human skin microbiome. Science 324:1190-1192.
51. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human genetics shape the gut microbiome. Cell 159:789-799.
52. Wilkins D, Leung MHY, Lee PKH. 2017. Microbiota fingerprints lose individually identifying features over time. Microbiome 5:1.
53. Tridico SR, Murray DC, Addison J, Kirkbride KP, Bunce M. 2014. Metagenomic analyses of bacteria on human hairs: a qualitative assessment for applications in forensic science. Investig Genet 5:16.
54. Dominguez-Bello MG, Blaser MJ, Ley RE, Knight R. 2011. Development of the human gastrointestinal microbiota and insights from high-throughput sequencing. Gastroenterology 140:1713-1719.
55. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, Knight R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A 107:11971-11975.
56. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108:4578-4585.
57. Crielaard W, Zaura E, Schuller AA, Huse SM, Montijn RC, Keijser BJF. 2011. Exploring the oral microbiota of children at various developmental stages of their dentition in the relation to their oral health. BMC Med Genomics 4:1-13.
58. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930-E2938.
59. Rajilić-Stojanović M, Heilig HGHJ, Tims S, Zoetendal EG, de Vos WM. 2013. Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 15:1146-1159.
60. Martínez I, Muller CE, Walter J. 2013. Long-term temporal analysis of the human fecal microbiota revealed a stable core of dominant bacterial species. PLoS One 8:e69621.
61. Lazarevic V, Whiteson K, Hernandez D, François P, Schrenzel J. 2010. Study of inter-and intra-individual variations in the salivary microbiota. BMC Genomics 11:523.
62. Meadow JF, Bateman AC, Herkert KM, O'Connor TK, Green JL. 2013. Significant changes in the skin microbiome mediated by the sport of roller derby. PeerJ 1:e53.
63. Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen JA, Gilbert JA. 2015. Forensic analysis of the microbiome of phones and shoes. Microbiome 3:21.
64. Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, Siegel J, Caporaso JG. 2016. Geography and location are the primary drivers of office microbiome composition. mSystems 1:e00022-16.
65. Oh J, Byrd AL, Park M, Kong HH, Segre JA, Sequencing NC. 2016. Temporal stability of the human skin microbiome. Cell 165:854-866.
66. Donaldson AE, Taylor MC, Cordiner SJ, Lamont IL. 2010. Using oral microbial DNA analysis to identify expirated bloodspatter. Int J Legal Med 124:569-576.
67. Power DA, Cordiner SJ, Kieser JA, Tompkins GR, Horswell J. 2010. PCR-based detection of salivary bacteria as a marker of expirated blood. Sci Justice 50:59-63.
68. Choo JM, Leong LEX, Rogers GB. 2015. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5:1-10.
69. Roesch LFW, Casella G, Simell O, Krischer J, Wasserfall CH, Schatz D, Atkinson MA, Neu J, Triplett EW. 2009. Influence of fecal sample storage on bacterial community diversity. Open Microbiol J 3:40-46.
70. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett 307:80-6.

71. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477-6481.
72. Gibbons SM, Schwartz T, Fouquier J, Mitchell M, Sangwan N, Gilbert JA, Kelley ST. 2015. Ecological succession and viability of human-associated microbiota on restroom surfaces. Appl Environ Microbiol 81:765-773.
73. Diana W. Williams GG. 2017. Individualization of pubic hair bacterial communities and the effects of storage time and temperature. Forensic Sci Int Genet 26:9.
74. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, Deelen P, Vatanen T, Schirmer M, Smeekens SP, Zhernakova DV, Jankipersadsing SA, Jaeger M, Oosting M, Cenit MC, Masclee AAM, Swertz MA, Li Y, Kumar V, Joosten L, Harmsen H, Weersma RK, Franke L, Hofker MH, Xavier RJ, Jonkers D, Netea MG, Wijmenga C, Fu J, Zhernakova A. 2016. The effect of host genetics on the gut microbiome. Nat Genet 48:1407-1412.
75. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG. 2015. Host genetic variation impacts microbiome composition across human body sites. Genome Biol 16:191.
76. De Cárcer DA, Cuív PÓ, Wang T, Kang S, Worthley D, Whitehall V, Gordon I, McSweeney C, Leggett B, Morrison M. 2011. Numerical ecology validates a biogeographical distribution and gender-based effect on mucosa-associated bacteria along the human colon. ISME J 5:801-809.
77. Dominianni C, Sinha R, Goedert JJ, Pei ZH, Yang LY, Hayes RB, Ahn JY. 2015. Sex, body mass index, and dietary fiber intake influence the human gut microbiome. PLoS One 10:e0124599.
78. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108 Suppl 1:4680-4687.
79. Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, Foster JA, Forney LJ. 2007. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME J 1:121-133.
80. Lee S, Sung J, Lee J, Ko G. 2011. Comparison of the gut microbiotas of healthy adult twins living in South Korea and the United States. Appl Environ Microbiol 77:7433-7437.
81. Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, Strauss JF, Jefferson KK, Buck GA, Consortium VM. 2014. Differences in vaginal microbiome in African American women versus women of European ancestry. Microbiology 160:2272-2282.
82. Nasidze I, Li J, Quinque D, Tang K, Stoneking M. 2009. Global diversity in the human salivary microbiome. Genome Res 19:636-643.
83. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI. 2008. Evolution of mammals and their gut microbes. Science 320:1647-1651.
84. Fava F, Gitau R, Griffin BA, Gibson GR, Tuohy KM, Lovegrove JA. 2013. The type and quantity of dietary fat and carbohydrate alter faecal microbiome and short-chain fatty acid excretion in a metabolic syndrome 'at-risk'population. Int J Obes 37:216-223.
85. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li HZ, Bushman FD, Lewis JD. 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science 334:105-108.
86. Monda V, Villano I, Messina A, Valenzano A, Esposito T, Moscatelli F, Viggiano A, Cibelli G, Chieffi S, Monda M, Messina G. 2017. Exercise modifies the gut microbiota with positive health effects. Oxid Med Cell Longev 2017:3831972.
87. O'Sullivan O, Cronin O, Clarke SF, Murphy EF, Molloy MG, Shanahan F, Cotter PD. 2015. Exercise and the microbiota. Gut Microbes 6:131-136.
88. Clarke SF, Murphy EF, O'Sullivan O, Lucey AJ, Humphreys M, Hogan A, Hayes P, O'Reilly M, Jeffery IB, Wood-Martin R. 2014. Exercise and associated dietary extremes impact on gut microbial diversity. Gut 63:1913-1920.
89. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. ISME J 10:2435-2446.
90. Tsigarida AA, Dabdoub SM, Nagaraja HN, Kumar PS. 2015. The influence of smoking on the peri-implant microbiome. J Dent Res 94:1202-1217.
91. Kato I, Vasquez AA, Moyerbrailean G, Land S, Sun J, Lin H-S, Ram JL. 2016. Oral microbiome and history of smoking and colorectal cancer. J Epidemiol Res 2:92-101.
92. Ece AM, Patrick MG, Huzefa R, Masoumeh S, Ammar N, Phillip AE, Mary K, Cynthia KL, Ali K. 2012. Colonic microbiome is altered in alcoholism. Am J Physiol Gastrointest Liver Physiol 302:G966-G978.

93. Gabbard SL, Lacy BE, Levine GM, Crowell MD. 2014. The impact of alcohol consumption and cholecystectomy on small intestinal bacterial overgrowth. Dig Dis Sci 59:638-644.
94. Queipo-Ortuño MI, Boto-Ordóñez M, Murri M, Gomez-Zumaquero JM, Clemente-Postigo M, Estruch R, Cardona Diaz F, Andrés-Lacueva C, Tinahones FJ. 2012. Influence of red wine polyphenols and ethanol on the gut microbiota ecology and biochemical biomarkers. Am J Clin Nutr 95:1323-1334.
95. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13:R79.
96. Turnbaugh PJ, Gordon JI. 2009. The core gut microbiome, energy balance and obesity. J Physiol 587:4153-4158.
97. Hartstra AV, Bouter KEC, Bäckhed F, Nieuwdorp M. 2015. Insights into the role of the microbiome in obesity and type 2 diabetes. Diabetes Care 38:159-165.
98. Slocum C, Kramer C, Genco CA. 2016. Immune dysregulation mediated by the oral microbiome: potential link to chronic inflammation and atherosclerosis. J Intern Med 280:114-128.
99. Noval Rivas M, Crother TR, Arditi M. 2016. The microbiome in asthma. Curr Opin Pediatr 28:764-771.
100. Vuong HE, Hsiao EY. 2017. Emerging roles for the gut microbiome in autism spectrum disorder. Biol Psychiatry 81:411-423.
101. Ursell LK, Clemente JC, Rideout JR, Gevers D, Caporaso JG, Knight R. 2012. The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. J Allergy Clin Immunol 129:1204-1208.
102. Blaser MJ. 2016. Antibiotic use and its consequences for the normal microbiome. Science 352:544-545.
103. Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L. 2010. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. PLoS One 5:e9836.
104. Zaura E, Brandt BW, de Mattos MJT, Buijs MJ, Caspers MPM, Rashid M-U, Weintraub A, Nord CE, Savell A, Hu Y. 2015. Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. mBio 6:e01693-15.
105. Rogers MAM, Aronoff DM. 2016. The influence of non-steroidal anti-inflammatory drugs on the gut microbiome. Clin Microbiol Infect 22:178.e1-178.e9.
106. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, Knight R. 2012. Experimental and analytical tools for studying the human microbiome. Nat Rev Genet 13:47-58.
107. Ferretti P, Farina S, Cristofolini M, Girolomoni G, Tett A, Segata N. 2017. Experimental metagenomics and ribosomal profiling of the human skin microbiome. Exp Dermatol 26:211-219.
108. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833-844.
109. Embong Z, Hitam WHW, Yean CY, Rashid NHA, Kamarudin B, Abidin SKZ, Osman S, Zainuddin ZF, Ravichandran M. 2008. Specific detection of fungal pathogens by 18S rRNA gene PCR in microbial keratitis. BMC Ophthalmol 8:1-8.
110. Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H. 2010. ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. BMC Microbiol 10:1-9.
111. Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M. 2007. Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. FEMS Microbiol Ecol 60:341-350.
112. Liu Z, DeSantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res 36:e120.
113. Hayashi H, Sakamoto M, Benno Y. 2004. Evaluation of three different forward primers by terminal restriction fragment length polymorphism analysis for determination of fecal Bifidobacterium spp. in healthy subjects. Microbiol Immunol 48:1-6.
114. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72:5069-5072.
115. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-D596.
116. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141-D145.

117. Chen T, Yu WH, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. 2010. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database (Oxford) 2010:baq013.
118. Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45:2761-2764.
119. Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328-1333.
120. Schouls LM, Schot CS, Jacobs JA. 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. J Bacteriol 185:7241-7246.
121. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112-5120.
122. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, Zhou H-W, Rognes T, Caporaso JG, Knight R. 2016. Open-source sequence clustering methods improve the state of the art. mSystems 1:e00003-15.
123. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE. 2014. Conducting a microbiome study. Cell 158:250-262.
124. Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 3:e1487.
125. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581-583.
126. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639-2643.
127. Thomas T, Gilbert J, Meyer F. 2012. Metagenomics-a guide from sampling to data analysis. Microb Inform Exp 2:1-12.
128. Feehery GR, Yigit E, Oyola SO, Langhorst BW, Schmidt VT, Stewart FJ, Dimalanta ET, Amaral-Zettler LA, Davis T, Quail MA. 2013. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. PLoS One 8:e76096.
129. Lasken RS. 2009. Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochem Soc Trans 37:450-453.
130. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.
131. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9:811-4.
132. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 10:1196-1199.
133. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343-359.
134. Aitchison J. 1982. The statistical analysis of compositional data. J R Stat Soc 44:139-160.
135. Tsilimigras MCB, Fodor AA. 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of Epidemiology 26:330-335.
136. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224.
137. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol 10:766.
138. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Med 8:1-10.
139. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977.
140. Geman O, Chiuchisan I, Covasa M, Doloc C, Milici M-R, Milici L-D. 2016. Deep learning tools for human microbiome big data. Soft Computing Applications. Advances in Intelligent Systems and Computing, vol 633. Springer, Cham.
141. Ananthakrishnan AN, Luo C, Yajnik V, Khalili H, Garber JJ, Stevens BW, Cleland T, Xavier RJ. 2017. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. Cell Host Microbe 21:603-610.e3.

142. Thaiss CA, Itav S, Rothschild D, Meijer MT, Levy M, Moresi C, Dohnalová L, Braverman S, Rozin S, Malitsky S. 2016. Persistent microbiome alterations modulate the rate of post-dieting weight regain. Nature 540:544-551.
143. Kennedy DM, Stanton J-AL, García JA, Mason C, Rand CJ, Kieser JA, Tompkins GR. 2012. Microbial analysis of bite marks by sequence comparison of streptococcal DNA. PLoS One 7:e51757.
144. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, Lents NH. 2016. A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. PLoS One 11:e0167370.
145. Metcalf JL, Xu ZZ, Weiss S, Lax S, Van Treuren W, Hyde ER, Song SJ, Amir A, Larsen P, Sangwan N. 2016. Microbial community assembly and metabolic function during mammalian corpse decomposition. Science 351:158-162.
146. Metcalf JL, Xu ZZ, Bouslimani A, Dorrestein P, Carter DO, Knight R. 2017. Microbiome tools for forensic science. Trends in Biotechnol 35:814-823.
147. Loomba R, Seguritan V, Li W, Long T, Klitgord N, Bhatt A, Dulai PS, Caussy C, Bettencourt R, Highlander SK, Jones MB, Sirlin CB, Schnabl B, Brinkac L, Schork N, Chen C-H, Brenner DA, Biggs W, Yooseph S, Venter JC, Nelson KE. 2017. Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human nonalcoholic fatty liver disease. Cell Metab 25:1054-1062.e5.
148. Beck D, Foster JA. 2014. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS One 9:e87830.
149. Lan Y, Kriete A, Rosen GL. 2013. Selecting age-related functional characteristics in the human gut microbiome. Microbiome 1:2.
150. Qian J, Hospodsky D, Yamamoto N, Nazaroff WW, Peccia J. 2012. Size-resolved emission rates of airborne bacteria and fungi in an occupied classroom. Indoor Air 22:339-51.
151. Fleming RI, Harbison S. 2010. The use of bacteria for the identification of vaginal secretions. Forensic Sci Int Genet 4:311-315.
152. Nakanishi H, Kido A, Ohmori T, Takada A, Hara M, Adachi N, Saito K. 2009. A novel method for the identification of saliva by detecting oral streptococci using PCR. Forensic Sci Int Genet 183:20-23.
153. Damann FE, Williams DE, Layton AC. 2015. Potential use of bacterial community succession in decaying human bone for estimating postmortem interval. J Forensic Sci 60:844-850.
154. Hauther KA, Cobaugh KL, Jantz LM, Sparer TE, DeBruyn JM. 2015. Estimating time since death from postmortem human gut microbial communities. J Forensic Sci 60:1234-1240.
155. Meadow JF, Altrichter AE, Kembel SW, Moriyama M, O'Connor TK, Womack AM, Brown GZ, Green JL, Bohannan BJM. 2014. Bacterial communities on classroom surfaces vary with human contact. Microbiome 2:1-7.
156. Meadow JF, Altrichter AE, Green JL. 2014. Mobile phones carry the personal microbiome of their owners. PeerJ 2:e447.
157. Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown GZ, Green JL, Bohannan BJM. 2015. Humans differ in their personal microbial cloud. PeerJ 3:e1258.
158. Aditya S, Sharma AK, Bhattacharyya CN, Chaudhuri K. 2011. Generating STR profile from "Touch DNA". J Forensic Leg Med 18:295-298.
159. Leake SL, Pagni M, Falquet L, Taroni F, Greub G. 2016. The salivary microbiome for differentiating individuals: proof of principle. Microbes Infect 18:399-405.
160. Javan GT, Finley SJ, Can I, Wilkinson JE, Hanson JD, Tarone AM. 2016. Human thanatomicrobiome succession and time since death. Sci Rep 6:29598.
161. Can I, Javan GT, Pozhitkov AE, Noble PA. 2014. Distinctive thanatomicrobiome signatures found in the blood and internal organs of humans. J Microbiol Methods 106:1-7.
162. Hyde ER, Haarmann DP, Petrosino JF, Lynne AM, Bucheli SR. 2015. Initial insights into bacterial succession during human decomposition. Int J Legal Med 129:661-671.
163. Suzuki TA, Worobey M. 2014. Geographical variation of human gut microbial composition. Biol Lett 10:20131037.
164. Escobar JS, Klotz B, Valdes BE, Agudelo GM. 2014. The gut microbiota of Colombians differs from that of Americans, Europeans and Asians. BMC Microbiol 14:311.
165. Ying S, Zeng D-N, Chi L, Tan Y, Galzote C, Cardona C, Lax S, Gilbert J, Quan Z-X. 2015. The influence of age and gender on skin-associated microbial communities in urban and rural human populations. PLoS One 10:e0141842.
166. Tyakht AV, Kostryukova ES, Popenko AS, Belenikin MS, Pavlenko AV, Larin AK, Karpova IY, Selezneva OV, Semashko TA, Ospanova EA. 2013. Human gut microbiota community structures in urban and rural populations in Russia. Nat Commun 4:1-9.

167. Blaser MJ, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Estrada I, Gao Z, Clemente JC, Costello EK, Knight R. 2013. Distinct cutaneous bacterial assemblages in a sampling of South American Amerindians and US residents. ISME J 7:85-95.
168. Li J, Quinque D, Horz HP, Li MK, Rzhetskaya M, Raff JA, Hayes MG, Stoneking M. 2014. Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. BMC Microbiol 14:316.
169. Miller GE, Engen PA, Gillevet PM, Shaikh M, Sikaroodi M, Forsyth CB, Mutlu E, Keshavarzian A. 2016. Lower neighborhood socioeconomic status associated with reduced diversity of the colonic microbiota in healthy adults. PLoS One 11:e0148952.
170. McNulty SL, Mole BM, Dailidiene D, Segal I, Ally R, Mistry R, Secka O, Adegbola RA, Thomas JE, Lenarcic EM, Peek RM, Jr., Berg DE, Forsyth MH. 2004. Novel 180- and 480-base-pair insertions in African and African-American strains of Helicobacter pylori. J Clin Microbiol 42:5658-5663.
171. Kersulyte D, Kalia A, Gilman RH, Mendez M, Herrera P, Cabrera L, Velapatiño B, Balqui J, Paredes Puente de la Vega F, Rodriguez Ulloa CA, Cok J, Hooper CC, Dailide G, Tamma S, Berg DE. 2010. Helicobacter pylori from Peruvian Amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an amerind strain. PLoS One 5:e15076.
172. Nagasawa S, Motani-Saitoh H, Inoue H, Iwase H. 2013. Geographic diversity of Helicobacter pylori in cadavers: Forensic estimation of geographical origin. Forensic Sci Int Genet 229:7-12.
173. Hewitt KM, Gerba CP, Maxwell SL, Kelley ST. 2012. Office space bacterial abundance and diversity in three metropolitan areas. PLoS One 7:e37849.
174. Akutsu T, Motani H, Watanabe K, Iwase H, Sakurada K. 2012. Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. Leg Med 14:160-162.
175. Benschop CCG, Quaak FCA, Boon ME, Sijen T, Kuiper I. 2012. Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? Int J Legal Med 126:303-310.
176. Giampaoli S, Berti A, Valeriani F, Gianfranceschi G, Piccolella A, Buggiotti L, Rapone C, Valentini A, Ripani L, Romano Spica V. 2012. Molecular identification of vaginal fluid by microbial signature. Forensic Sci Int Genet 6:559-564.
177. Giampaoli S, Alessandrini F, Berti A, Ripani L, Choi A, Crab R, De Vittori E, Egyed B, Haas C, Lee HY, Korabecná M, Noel F, Podini D, Tagliabracci A, Valentini A, Romano Spica V. 2014. Forensic interlaboratory evaluation of the ForFLUID kit for vaginal fluids identification. J Forensic Leg Med 21:60-63.
178. Nakanishi H, Shojo H, Ohmori T, Hara M, Takada A, Adachi N, Saito K. 2013. Identification of feces by detection of Bacteroides genes. Forensic Sci Int Genet 7:176-179.
179. Lucci A, Campobasso CP, Cirnelli A, Lorenzini G. 2008. A promising microbiological test for the diagnosis of drowning. Forensic Sci Int Genet 182:20-26.
180. Suto M, Kato N, Abe S, Nakamura M, Tsuchiya R, Hiraiwa K. 2009. PCR detection of bacterial genes provides evidence of death by drowning. Leg Med (Tokyo) 11 Suppl 1:S354-6.
181. Gunn A, Pitt SJ. 2012. Microbes as forensic indicators. Trop Biomed 29:311-330.
182. Black CM, Driebe EM, Howard LA, Fajman NN, Sawyer MK, Girardet RG, Sautter RL, Greenwald E, Beck-Sague CM, Unger ER, Igietseme JU, Hammerschlag MR. 2009. Multicenter study of nucleic acid amplification tests for detection of Chlamydia trachomatis and Neisseria gonorrhoeae in children being evaluated for sexual abuse. Pediatr Infect Dis J 28:608-13.
183. Martin IMC, Foreman E, Hall V, Nesbitt A, Forster G, Ison CA. 2007. Non-cultural detection and molecular genotyping of Neisseria gonorrhoeae from a piece of clothing. J Med Microbiol 56:487-490.

PART I

Human microbiome for assigning body site of origin of forensic traces

Chapter 2

Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials

Celia Díez López, Athina Vidaki, Arwin Ralf, Diego Montiel González, Djawad Radjabzadeh, Robert Kraaij, André G Uitterlinden, Cordula Haas, Oscar Lao, Manfred Kayser

Forensic Sci Int Genet. 2019 Jul; 41:72-82.

ABSTRACT

Correct identification of different human epithelial materials such as from skin, saliva and vaginal origin is relevant in forensic casework as it provides crucial information for crime reconstruction. However, the overlap in human cell type composition between these three epithelial materials provides challenges for their differentiation and identification when using previously proposed human cell biomarkers, while their microbiota composition largely differs. By using validated 16S rRNA gene massively parallel sequencing data from the Human Microbiome Project of 1,636 skin, oral and vaginal samples, 50 taxonomy-independent deep learning networks were trained to classify these three tissues. Validation testing was performed in *de novo* generated high-throughput 16S rRNA gene sequencing data using the Ion Torrent™ Personal Genome Machine from 110 test samples: 56 hand skin, 31 saliva and 23 vaginal secretion specimens. Body-site classification accuracy of these test samples was very high as indicated by AUC values of 0.99 for skin, 0.99 for oral, and 1 for vaginal secretion. Misclassifications were limited to 3 (5%) skin samples. Additional forensic validation testing was performed in mock casework samples by *de novo* high-throughput sequencing of 19 freshly-prepared samples and 22 samples aged for 1 up to 7.6 years. All of the 19 fresh and 20 (91%) of the 22 aged mock casework samples were correctly tissue-type classified. Moreover, comparing the microbiome results with outcomes from previous human mRNA-based tissue identification testing in the same 16 aged mock casework samples reveals that our microbiome approach performs better in 12 (75%), similarly in 2 (12.5%), and less good in 2 (12.5%) of the samples. Our results demonstrate that this new microbiome approach allows for accurate tissue-type classification of three human epithelial materials of skin, oral and vaginal origin, which is highly relevant for future forensic investigations.

INTRODUCTION

The gold standard in routine forensic investigations is DNA profiling based on short tandem repeats (STRs) to determine if a known suspect is the sample donor, or not (1). However, for evaluating the crime relevance of the biological trace used for DNA-based individual identification, as well as to reconstruct the events taken place at the crime scene, it is necessary to establish the cell or tissue type a biological trace belongs to. Currently applied (mainly biochemical) test systems used for forensic tissue identification only provide presumptive answers, since substances other than the indicated body fluid can also provide positive test outcomes (with the exception of microscopic sperm cell identification) (2). Recently developed test systems based on human mRNA, miRNA, and DNA methylation markers are generally more specific (2, 3), but the overlap of cell types between different forensically-relevant tissues provides challenges. Tissue misclassification has been particularly noted for those forensically relevant tissues that contain epithelial cells from different body sites, such as skin, saliva, and vaginal secretion samples, which are often collected from crime scenes (2, 3). However, these different body sites harbor specialized microbial communities (4), and microbial cells are found in equal or larger quantities than human cells at the respective body sites (5). Therefore, the human microbiome is expected to provide a suitable resource for correctly identifying the body site source of such challenging epithelial samples.

Previous attempts to identify forensically relevant tissue types via microbial DNA analysis focused on a limited number of microbial species (2, 6, 7), sometimes in combination with human RNA (8) or DNA methylation (9) markers. The main limitation of specifically targeting a limited number of bacterial species for forensic tissue identification is their large variation in frequency composition within (such as over time) and between individuals, typically not sufficiently quantified in small-sized studies (10, 11). Moreover, some target species occupy a specific niche within a body site (i.e. *Streptococcus mutans* inhabits dental plaques within the oral cavity) that may not always be represented in the sample collected from a crime scene (i.e. saliva sample) (6). Another limitation is that for some of those microbial target species, non-specific signals have been reported in non-target tissues (12-15), in food (7), and in other animal species (16), which can lead to false positive results. In addition, bacterial-specific nucleic acids isolation methods are not usually employed in forensics, which can complicate the detection of certain species such as gram-positive bacteria (17). Furthermore, sample size in previous forensically-oriented microbial DNA studies was typically small, in principle not allowing for obtaining reliable outcomes.

The limitations of target species for differentiating among epithelial materials from different body sites, together with the availability of multi-tissue data from the Human Microbiome Project (HMP) (18), prompted us to investigate the use of microbiome profiling for the identification of skin, saliva and vaginal secretions. In contrast to previous studies, we aimed to take advantage of the full spectrum of the microbial community present in a sample (not only a limited number of target species) by applying a taxonomy-independent microbiome approach, and used a large number of reference data from the HMP, which both - to our knowledge - has not been previously reported for this purpose. To achieve tissue classification from microbiome composition, we first developed a new taxonomy-independent deep learning (DL) approach for tissue classification based on 16S ribosomal RNA (rRNA) gene sequencing

data, and trained the DL networks with HMP data from a total of 1,636 skin, oral and vaginal samples. Next, we validated this novel approach in 110 newly-collected test samples from skin, saliva, and vaginal secretion, for which we *de novo* generated high-throughput 16S rRNA gene sequencing data using the Ion Torrent™ Personal Genome Machine (PGM) and finally explained the rarely observed classification outliers by subsequently applying a taxonomy-based approach (QIIME). Moreover, we assessed the robustness and forensic suitability of our new microbiome approach in mock casework samples by using freshly-prepared mock samples as well as mock samples aged for 1 up to 7.6 years. In the study design and execution, we paid particular attention towards harmonizing newly and previously generated microbiome data in terms of DNA extraction and 16S rRNA gene sequencing to allow their combined use.

MATERIALS AND METHODS

Reference data preparation

16S rRNA gene sequencing data from the Human Microbiome Project 16S production phase I (16S-PP1) were obtained from the HMP data analysis and coordination center (DACC) website (19). Sequencing data from skin, oral and female urogenital sites were obtained in FASTA file format. Files were validated to contain both V1-3 and V3-5 16S rRNA gene hypervariable regions target sequences. Validated data comprised a total of 1,636 samples (Supplementary File S1) that were assigned to three main body sites categories as described in Table 1.

Table 1. HMP reference samples with 16S rRNA gene sequencing data used for DL training purposes.

Body site category	Body site	Number of validated samples
Skin	Left antecubital fossa	172
	Right antecubital fossa	176
	Left retroauricular crease	88
	Right retroauricular crease	98
	Total	**534**
Oral	Buccal mucosa	98
	Hard palate	96
	Keratinized gingiva	98
	Palatine tonsils	101
	Saliva	83
	Subgingival plaque	103
	Supragingival plaque	105
	Throat	163
	Tongue dorsum	104
	Total	**951**
Vagina	Mid vagina	52
	Posterior fornix	51
	Vaginal introitus	48
	Total	**151**

The sequencing reads were aligned to the *Escherichia coli* str. K 12 substr. MG1655 (nc_000913.3) 16S rRNA gene using BWA-MEM (20) aligner software with parameters: -B1, -O1, -E1 and –L1 to allow mapping of sequencing reads that have many variations compared to the reference. SAMtools (21) was used to convert the SAM files to sorted BAM files. SAMtools mpileup command was used to pile up all aligned positions. From each sample and

position of the 16S rRNA gene, the number of reads with adenine (A), guanine (G), thymine (T), cytosine (C), insertion (+) and deletion (-) were retrieved. Subsequently, a matrix of individuals by sites was generated by merging all positions detected at least in one sample, representing 1,166 positions. In order to select the most informative sequence positions, we applied Correspondence Analysis (CA) technique by displaying the available 1,166 sequence positions in a contingency table for all the 1,636 reference samples from the HMP dataset. Subsequently, sequence positions showing low inertia (<0.6) were filtered out using the R package ca (22). The excluded sequence positions were highly similar among the three body site categories (skin, oral and vagina), being no informative for the purpose of differentiating the three epithelial materials. The final input matrix for DL consisted of 1,636 samples and 240 sequence positions (inertia>0.6), each position representing six cells (A, G, T, C, +, -).

Test samples and mock casework samples for validation

Biological samples including hand skin, saliva and vaginal secretion were collected from a total of 32 donors, 24 females and 8 males. An overview of the collected 110 test samples is provided in Table 2. Participants were asked to follow specific instructions before sample donation in order to minimize potential inter-individual variation. Similarly, to diminish both inter-individual and study-to-study systematic protocol biases, sample collection procedures were adapted to be as similar as possible to those specified by the Manual of Procedures for HMP (23) and are detailed in Supplementary Material S2. In the case of hand skin specimens, the index fingers from both hands were sampled, with the difference that the skin of one hand was 'protected' from the surrounding environment by a sterile glove worn for a specific time (1 hour) before removal and sample collection. This was to test whether body site classification was influenced by the presence (hand without glove) or absence (hand with glove) of potential microbiota picked up from the environment in hand skin samples.

Table 2. Type and number of test samples for validation purposes.

	Hand skin (without glove)	**Hand skin (with glove)**	**Saliva**	**Vaginal secretion**
Female	20	20	23	23
Male	8	8	8	-
Total	**28**	**28**	**31**	**23**

In addition, a total of 19 mock casework samples, mimicking realistic and often encountered forensic scenarios, were freshly prepared: 4 skin, 10 saliva and 5 vaginal secretion samples. Biological material was deposited on different substrates (fabric, eating and office utensils) and exposed to various environmental conditions, including temperature and relative humidity. DNA isolation was performed in fresh and 48 hours old samples. More detailed information on mock casework sample preparation and collection procedures are available in Supplementary Material S3.

Additionally, a total of 22 aged mock casework samples that were used in previous mRNA-based tissue identification studies (24-27) were tested: 7 skin, 9 oral and 6 vaginal secretion samples. The tissue materials from which these samples were obtained for the present study had previously been included in collaborative exercises with a validation focus on human mRNA-based tissue identification organized on behalf of the European DNA Profiling Group (EDNAP) (24-26) and also by the European Forensic Genetics Network of Excellence (EUROFORGEN_NoE) (27). Tissue materials were initially stored at room

conditions in the dark for a time period of up to a maximum of 6 months before RNA extraction on parts of the material were carried out for use in the previous studies. The excess material remained stored at room conditions in the dark for an additional 6 months to 7.1 years depending on the sample before DNA extraction was carried out for the purpose of this study, using the same protocol as applied for the test samples and the freshly-prepared mock samples. Hence, the total storage time for the aged mock samples used in the present study ranged from 1 to 7.6 years. All these samples were analyzed without prior knowledge of their body site origin ('blind' testing). Detailed information on sample features, collection procedures and EDNAP/EUROFORGEN_NoE exercises is available in Supplementary Table S4.

All samples for *de novo* microbiome analysis were obtained with informed consent. The Kantonale Ethikkommission Zürich (KEK) approved the sample collection for this study. Declaration of no objection No.24-2015.

DNA isolation and quantification

DNA isolation was performed using the DNeasy PowerSoil kit (Qiagen, Germany), following the manufacturer's instructions, but only modified at two points for maximizing DNA yield. Firstly, as proposed by Castelino *et al.* (28), the 2 mL PowerBead tubes containing 750 µl solution and the saliva or swab head (skin and vaginal secretion samples) were incubated for 15 minutes at 70°C at the cell lysis step after the addition of PowerSoil® Solution C1. In addition, as recommended by the Manual of Procedures for HMP (23), centrifugation time in step 12 was increased from 1 to 2 minutes. A blank swab was included with each extraction. Isolated bacterial DNA was quantified with the Femto™ Bacterial DNA Quantification kit (Zymo Research, CA, USA) following the manufacturer's instructions on the C1000 Touch™ Thermal Cycler (Bio-Rad).

Library preparation from test samples and mock casework samples

This study was based on a target amplicon approach of hypervariable regions of the 16S rRNA gene. Since the freely available dataset from the HMP was used as a reference, the same primers suggested by the 16S 454 Sequencing Protocol from the HMP Consortium (29) were used in this study: V1-3 region of the 16S rRNA gene was amplified using the F27/R534 combination (5'-AGAGTTTGATCCTGGCTCAG-3', 5'- ATTACCGCGGCTGCTGG-3') and the V3-5 region using the F357/R926 combination (5'-CCTACGGGAGGCAGCAG-3', 5'-CCGTCAATTCMTTTRAGT-3'). Using the same primers reduces amplification biases observed with different sets of primers and increases the compatibility of the results obtained in different studies.

Library preparation was performed following the *Prepare Amplicon Libraries without Fragmentation Using the Ion Plus Fragment Library Kit* protocol for 400 bp target read length from the Ion Community webpage using the Ion Plus Fragment Library Kit (Thermo Fisher Scientific, USA). The total amount of bacterial DNA template used for the amplicon generation in the test samples was between 0.07-2.64 ng for hand skin without glove, 0.01-0.60 ng for hand skin with glove, 0.06-2 ng for saliva and 4-32 ng for vaginal secretion samples. For the mock casework samples, bacterial DNA template was between 0.05-36.71 ng for freshly-prepared samples and 0.01-29.03 ng for the aged samples. An annealing temperature of 58°C was settled for V1-3 region and 50°C for V3-5 region. Amplicons were quantified with Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific) following the manufacturer's

instructions and equimolar amplicon pools were prepared (30 ng of the amplicon pool in a total volume of 50 µL). Libraries were quantified using the Ion Library Taqman® Quantitation Kit (Thermo Fisher Scientific) following manufacturer's instructions and diluted accordingly down to 50 pM. The diluted libraries harboring different barcodes were pooled together in an equal concentration.

Template preparation and massively parallel sequencing

Template preparation was performed using the Ion PGM Hi-Q View OT2 kit and loaded onto the Ion One Touch™ 2 (OT2) instrument for emulsion PCR following the manufacturer's protocol (Thermo Fisher Scientific). Subsequently, samples were enriched with template-positive ion sphere particles (ISPs) on the Ion Torrent™ Enrichment System (ES) following the manufacturer's protocol (Thermo Fisher Scientific). Sequencing was performed using the Ion PGM Hi-Q View Sequencing kit and Ion 318™ v2 chip on the Ion Torrent™ PGM (Thermo Fisher Scientific) according to the supplier's instructions. We used 850 flows in order to get sufficiently long sequencing reads. Sequencing data was obtained from the Torrent server as FASTQ file format using the FastqCreator (v3.4.56313) plugin. On average, test samples sequencing runs generated ~3.65 million of total reads after removal of sequencing adapters and barcodes. On average, mock casework samples sequencing runs generated ~4.78 million of total reads. Sequencing reads were trimmed by quality (Q<20) and length (<100 bp) using Trimmomatic (30).

Deep learning implementation

The deep learning (DL) artificial neural network (ANN) for human epithelial material identification was implemented in Tensorflow (31) and Keras (32). Input data comprised 1,440 features (six nucleotide categories – A, G, T, C, (+), (-) – by 240 retained sequence positions). Output was the classification into one of the three body site categories skin, oral, or vagina. The DL consisted on a four layer ANN, with two hidden layers of ten neurons each one using the hyperbolic tangent function (TANH) (33) and a softmax activation function as output. The TANH activation function is a rescaling version of the Sigmoid function showing a more efficient convergence during the training.

The training algorithm of the neural network was performed using Adam optimizer (34) and drop out (35) at a rate of 0.1 to regularize the model. The DL was trained for 10,000 generations or until the categorical accuracy error was smaller than 0.001.

Given that the number of features was relatively large compared to the number of training samples, a model averaging/bagging was implemented (35) by independently training 50 DL networks with the above proposed neural architecture. Test and mock samples were given assigned probabilities to each site category (skin, oral, vagina) by means of averaging the output from the different 50 DL networks.

Area under the receiver operating characteristic curve (AUC) values from the classification performed by the 50 DL networks in the test samples were calculated using the scikit-learn (36) package. Two-dimensional plots from principal component analysis (PCA), using the activation values of the hidden neurons of each network ((10+10)*50 neurons per sample), were generated using ggplot2 R package (37).

Taxonomic profiling of classification outliers

16S rRNA gene sequencing data for each analyzed sample was obtained from the Torrent server as FASTQ files. Chimera reads, reads shorter than 100 bp and reads with total expected error higher than 1% were removed from the files using UCHIME version 8.1 (38). Remaining reads were then mapped to the SILVA rRNA database version 128 (39) using RDP Naïve Bayesian Classifier (40) in QIIME environment using the closed-reference method (41). The final Operational Taxonomic Unit (OTU) table was divided into 5 sub-tables at different taxonomic levels (in QIIME environment): phylum, class, order, family and genus.

RESULTS

Training of taxonomy-independent deep learning approach

Overall, 16s rRNA gene V1-3 and V3-5 hypervariable sequences of 1,636 samples from the HMP 16S-PP1 with skin, oral and vaginal sites origin were validated. This sample set was used to train 50 DL networks independently from each other (see Materials and Methods section for network architecture). The output consisted of the probability of assignment of each sample to each of the three body site categories skin, oral and vagina. In order to provide a visual summary of the DL performance, principal component analysis (PCA) using the activation values of the hidden neurons of each network ((10+10)*50 neurons per sample) was performed on the HMP training dataset. The first two principal components explained 92% of the variation present in these data. As evident from Figure 1, the 1,636 training HMP samples appeared to cluster in three main groups, corresponding to the three body sites from which these samples were initially obtained i.e. skin, oral cavity and vagina.

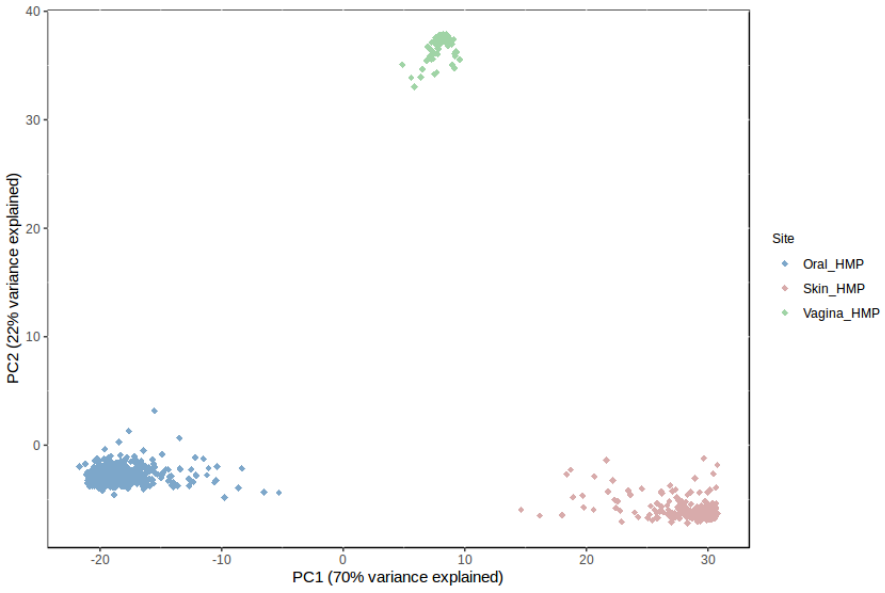


Figure 1. Principal component analysis (PCA) of the deep learning (DL) training samples from HMP. Projection of the activation values of the hidden neurons of each DL using the HMP training samples (Oral_HMP, Skin_HMP, Vagina_HMP) in the first two dimensions of a PC space together explaining 92% of the total variation. Three distinct data groups correspond to the three body site categories from where the samples were collected.

Validation of taxonomy-independent deep learning approach using test samples

The 50 DL networks that were trained based on HMP reference data were subsequently applied to the 16S rRNA gene sequencing data *de novo* generated with the Ion Torrent™ PGM from the 110 newly-collected test samples: 56 hand skin - with and without glove, 31 saliva and 23 vaginal secretion specimens. Each sample was given a probability of assignment to each of the three body site categories by means of averaging the output from the 50 DL networks. A probability of assignment threshold of 0.7 was used for considering conclusive (≥0.7) or inconclusive (<0.7) classifications. The resulting AUC values were very high at 0.99 for oral, 0.99 for skin and 1 for vagina, demonstrating that the trained DL networks had very high classification power. Overall, 95% (53/56) of the skin samples were correctly classified to the skin category, 100% (23/23) of the vaginal specimens were correctly classified to the vagina category and 100% (31/31) of the oral samples were accurately assigned to the oral category (Table 3).

Table 3. True (rows) *versus* estimated (columns) body-site classification of 110 test samples obtained with the novel taxonomy-independent deep learning microbiome approach.

	Skin	Oral	Vagina
Skin (n=56)	53 (95%)	3 (5%)	0
Oral (n=31)	0	31 (100%)	0
Vagina (n=23)	0	0	23 (100%)

Body site misclassification i.e. samples with a probability of assignment ≥0.7 in a category other than the true body site of origin, was limited to three skin samples only (5% of total skin samples), which were all misclassified as oral (explaining the AUC of 0.99 for skin and oral). Without considering the three misclassifications, skin samples were correctly classified with an average probability of assignment at 0.99885 (SD±8.25E-03). No substantial differences in the probabilities of assignment were noted for the correctly-classified skin samples with and without the usage of gloves for the same donors (Supplementary Figure S5). Oral samples presented an average probability of assignment at 0.99602 (SD±1.31E-02) and vaginal samples at 0.99939 (SD±2.77E-03). All probabilities for each of the test samples are shown in Supplementary Table S6.

Next, the *de novo* analyzed test samples were superimposed on the principal component analysis defined by the activation values of each of the 50 DL networks of the training HMP dataset (Figure 2). When considering the first two PCs (explain 92% of variation), the test samples largely overlap with the HMP training samples for all of the three groups, respectively. Exceptions were the three misclassified skin samples: d18_Swo (A), d19_Swo (B) and d23_Sw (C) (Figure 2), which did not cluster in the skin body site group, but in the oral group instead.

Forensic validation of taxonomy-independent deep learning approach using mock casework samples

As additional forensic validation step, we applied our new microbiome approach to freshly-prepared and aged mock casework samples mimicking realistic and often encountered forensic scenarios. From the 41 mock casework samples used in total across all three tissue types, our microbiome approach classified 39 (95%) correctly and only one was classified incorrectly (2.5%) while one (2.5%) was inconclusive (Table 4).

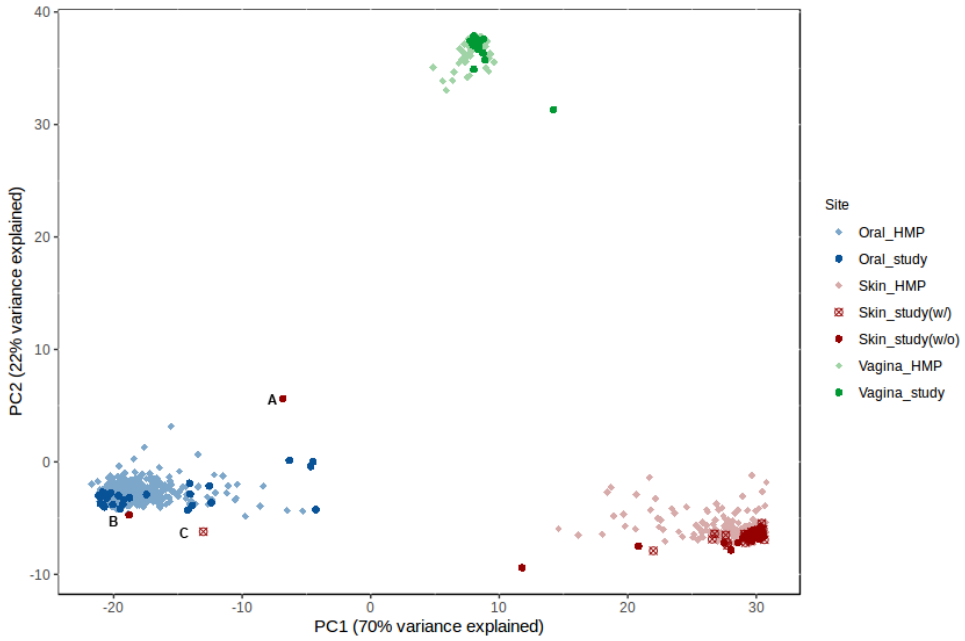


Figure 2. Principal component analysis (PCA) of the deep learning training samples from HMP with superimposed *de novo* analyzed test samples. All *de novo* analyzed test samples cluster with the HMP training samples in agreement with the respective body-site sample category except for the three misclassified skin samples d18_Swo (A), d19_Swo (B) and d23_Sw (C) that have assigned probabilities of 0.98324, 0.99999 and 0.99992 to the oral category, respectively.

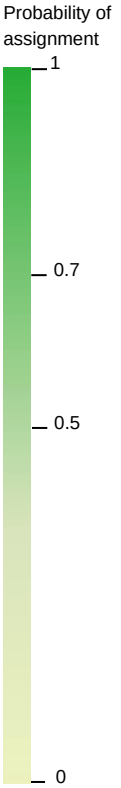
Regarding the freshly-prepared mock casework samples, all 19 were correctly classified to the respective body site they were collected from with probabilities of assigned to the correct body-site category >0.9999. In particular, freshly-prepared mock casework skin samples were correctly classified with an average probability of assignment at 0.99999 (SD±6.84E-0.7), oral samples at 0.99999 (SD±2.99E-08), and vaginal samples at 0.99997 (SD±0.2.88E-06) (Table 4). Notably, accurate body-site classification was obtained independently of the substrate of the sample (fabric, eating and office utensils), storage time until DNA isolation (fresh or 48 hours) and storage conditions (room temperature, 30°C, 4°C or 50% relative humidity), as considered in the design.

Regarding the 22 aged mock casework samples, all the 7 skin mock samples were correctly classified to the skin category with an average probability of assignment at 0.99998 (SD±3.67E-06). Out of the 9 aged oral mock samples, 8 were classified correctly, while one was misclassified as skin, with an assigned skin probability of 0.99977. The correctly-classified aged oral mock samples (8/9) presented an average probability of assignment at 0.99999 (SD±9.93E-06). Out of the 6 aged vaginal mock samples, 5 (83%) were correctly classified with an average probability of assignment at 0.94966 (SD±1.06E-01) using the 0.7 threshold. Of those, 4 (67%) had probabilities of vagina assigned ≥ 0.99339 while one had a probability of vagina assignment of 0.76002 (together with a probability of skin assignment of 0.16395 and oral assignment of 0.07602). The remaining one aged vaginal mock sample

yielded inconclusive results following the 0.7 probability threshold used, with a probability of vagina assignment of 0.55730 and a probability of skin assignment of 0.44223 (Table 4).

Table 4. Body site assignment probabilities of the freshly-prepared and aged mock casework samples obtained with the novel taxonomy-independent deep learning microbiome approach.

True site	Sample ID	Trace storage time (years)	Predicted site probabilities		
			Skin	Oral	Vagina
Skin	F3A	0	0.99999	7.44E-06	6.29E-06
	F3B	0	0.99999	7.44E-06	6.26E-06
	F3C	0	0.99999	7.53E-06	6.32E-06
	F3D	0	0.99998	8.65E-06	6.46E-06
	ZHF12	1	0.99999	7.76E-06	6.45E-06
	ZHF8	2	0.99999	7.38E-06	6.29E-06
	ZHF4	4.9	0.99999	7.66E-06	6.44E-06
	ZHF21	5	0.99999	7.36E-06	6.38E-06
	ZHF22	5	0.99998	1.66E-05	7.11E-06
	ZHF23	5	0.99998	8.06E-06	7.07E-06
	ZHF25	5	0.99999	7.47E-06	6.27E-06
Oral	F2A	0	3.64E-06	0.99999	2.92E-06
	F2B	0	3.59E-06	0.99999	2.96E-06
	F2C	0	3.69E-06	0.99999	2.82E-06
	F2D	0	3.62E-06	0.99999	2.91E-06
	F2E	0	3.58E-06	0.99999	2.98E-06
	F2F	0	3.61E-06	0.99999	2.94E-06
	F2G	0	3.58E-06	0.99999	2.98E-06
	F2H	0	3.62E-06	0.99999	2.91E-06
	F2I	0	3.70E-06	0.99999	2.83E-06
	F2J	0	3.60E-06	0.99999	2.94E-06
	ZHF10	1	3.64E-06	0.99999	2.87E-06
	ZHF5	2	3.87E-06	0.99999	2.71E-06
	ZHF6	2	3.72E-06	0.99999	2.76E-06
	ZHF24	5	3.61E-06	0.99999	2.96E-06
	ZHF2	7	3.66E-06	0.99999	2.96E-06
	ZHF17	7	3.18E-05	0.99997	2.96E-06
	ZHF18	7	4.11E-06	0.99999	2.69E-06
	ZHF14	7.6	3.61E-06	0.99999	2.93E-06
	ZHF15	7.6	0.99977	2.21E-04	1.19E-05
Vagina	F1A	0	1.92E-05	1.11E-05	0.99997
	F1B	0	1.87E-05	1.12E-05	0.99997
	F1C	0	1.67E-05	1.24E-05	0.99997
	F1D	0	1.79E-05	1.15E-05	0.99997
	F1E	0	2.57E-05	1.03E-05	0.99996
	ZHF11	1	1.87E-05	1.12E-05	0.99997
	ZHF7	2	1.64E-01	7.60E-02	0.76002
	ZHF3	5.9	0.44223	4.73E-04	0.55730
	ZHF20	6	2.00E-05	1.05E-04	0.99987
	ZHF19	6.6	4.95E-03	1.51E-05	0.99504
	ZHF16	7	6.44E-03	1.77E-04	0.99339



P≥0.7 threshold for conclusive classification. Misclassified and inconclusively classified samples are highlighted in bold.

Understanding classification outliers by subsequent taxonomic profiling

We aimed at finding an explanation for the misclassified samples. For this, we performed taxonomy identification from the microbiome data of these samples, by mapping the sequencing reads to the SILVA rRNA database in QIIME (41) environment using the closed-reference method. Operational Taxonomic Unit (OTU) tables at the phylum and genus levels were used to compare the retrieved microbiome profiles (Supplementary File S7).

From the three misclassified test samples, two misclassified skin test samples (d19_Swo and d23_Sw) that had a probability of assignment to oral at 0.99999 and 0.99992, respectively, revealed at the phylum level a reduced relative abundance of *Actinobacteria* and *Proteobacteria,* and an increased relative abundance of *Firmicutes,* when compared to the correctly classified skin samples (Figure 3a). At the genus level, these samples showed a reduced relative abundance of *Propionibacterium* and the uncultured genus of the *Neisseriaceae* family, and an increased relative abundance of *Streptococcus* (Figure 3b). The DL approach considering three sites resulted in the classification of these skin samples to the closest oral cluster that also showed low relative abundances of *Propionibacterium* and the uncultured genus of the *Neisseriaceae* family. For the third misclassified skin test sample (d18_Swo) that had a probability of assignment to oral at 0.98324, we noted an insufficiently low quality of the sample reads to perform taxonomic analyses.

Same procedure was followed for the one misclassified and one inconclusively classified mock casework samples. The misclassified oral sample (ZHF15) that had a probability of assignment to skin at 0.99977 revealed at the phylum level a reduced relative abundance of *Firmicutes* and an increased relative abundance of *Actinobacteria* and *Proteobacteria* when compared to the correctly classified oral samples (Figure 4a). At the genus level, this sample revealed a reduced relative abundance of *Streptococcus*, *Veillonella*, *Haemophilus*, *Fusobacterium*, *Neisseria*, *Blautia* and *Lactobacillus* compared to the correctly-classified oral samples. Interestingly, the most abundant genera present in this sample were *Ralstonia*, *Microbacterium, Phyllobacteriae* family and *Corynebacterium* (Figure 4b). These genera were almost absent in the correctly-classified samples from all the three categories (skin, oral and vagina). *Corynebacterium* was also present in the correctly-classified skin samples in a similar abundance which, together with the decreased abundance of oral-associated genera, could explained the assignation of this sample to the skin category.

The inconclusively classified mock vaginal sample (ZHF3) and the correctly-classified vaginal sample (ZHF7) that had a reduced (but still above the 0.7 threshold) vagina assignment probability of 0.76, both showed a slightly reduced relative abundance of *Firmicutes* at the phylum level. The former one also showed an increased relative abundance of *Actinobacteria* (Figure 4a). At the genus level, both samples showed decreased relative abundance of *Lactobacillus* while other genera appeared to be more abundant than in the correctly-classified vaginal samples. The inconclusively-classified vaginal sample (ZHF3) showed an increased relative abundance of *Staphylococcus* and *Corynebacterium*, while *Streptococccus* was more abundant in the correctly-classified one (ZHF7) (Figure 4b). These particular observations resembled to the relative abundances of these genera in the correctly-classified skin samples and could explain the probability of assignment of the two samples to the skin category.

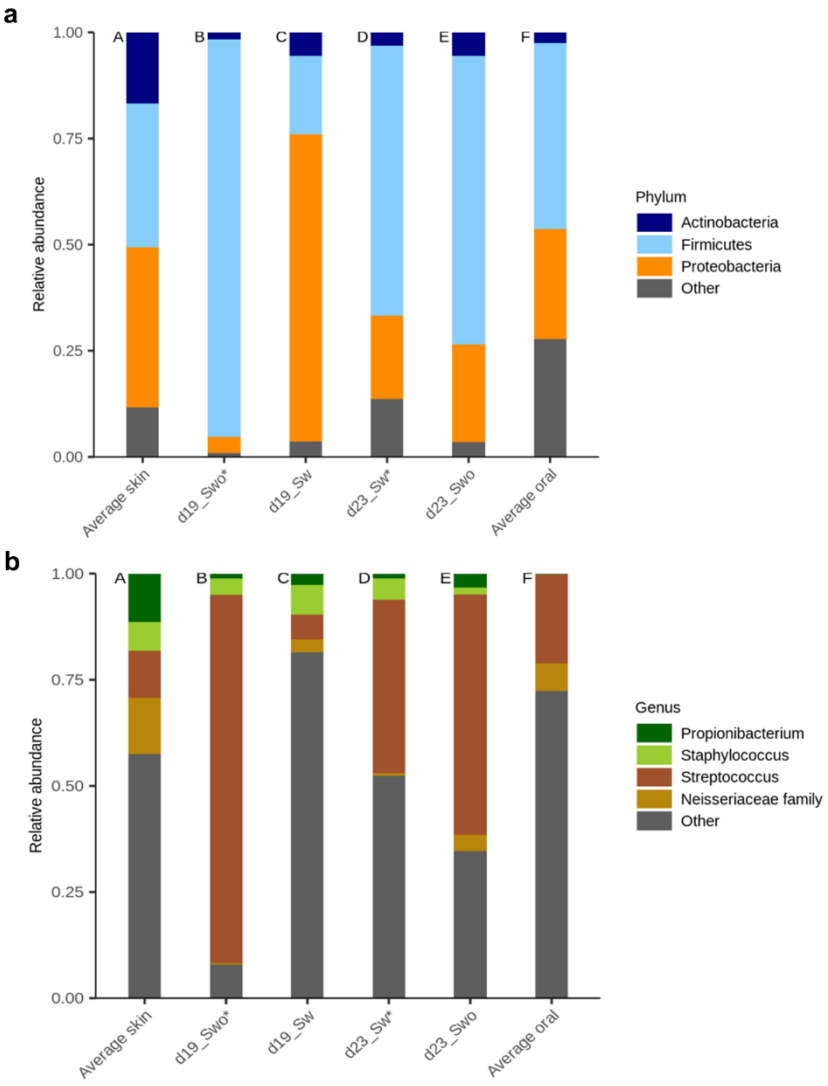


Figure 3. Taxonomic profile of misclassified skin samples at the phylum **(a)** and genus **(b)** level. (A) Overview of the relative abundance of the bacteria present in the correctly classified newly-analyzed test and mock skin samples. (B) Relative abundance of the bacteria in the misclassified hand skin sample without glove (d19_Swo) with a probability of assignment of 0.99999 to oral site. (C) Relative abundance of the bacteria in the well-classified skin sample with glove (d19_Sw). (D) Relative abundance of the bacteria in the misclassified hand skin sample with glove (d23_Sw) with a probability of assignment of 0.99992 to oral site. (E) Relative abundance of the bacteria in the well-classified skin sample without glove (d23_Swo). (F) Overview of the relative abundance of the bacteria present in the correctly classified newly-analyzed test and mock oral samples. Asterisks (*) refer to misclassified samples.

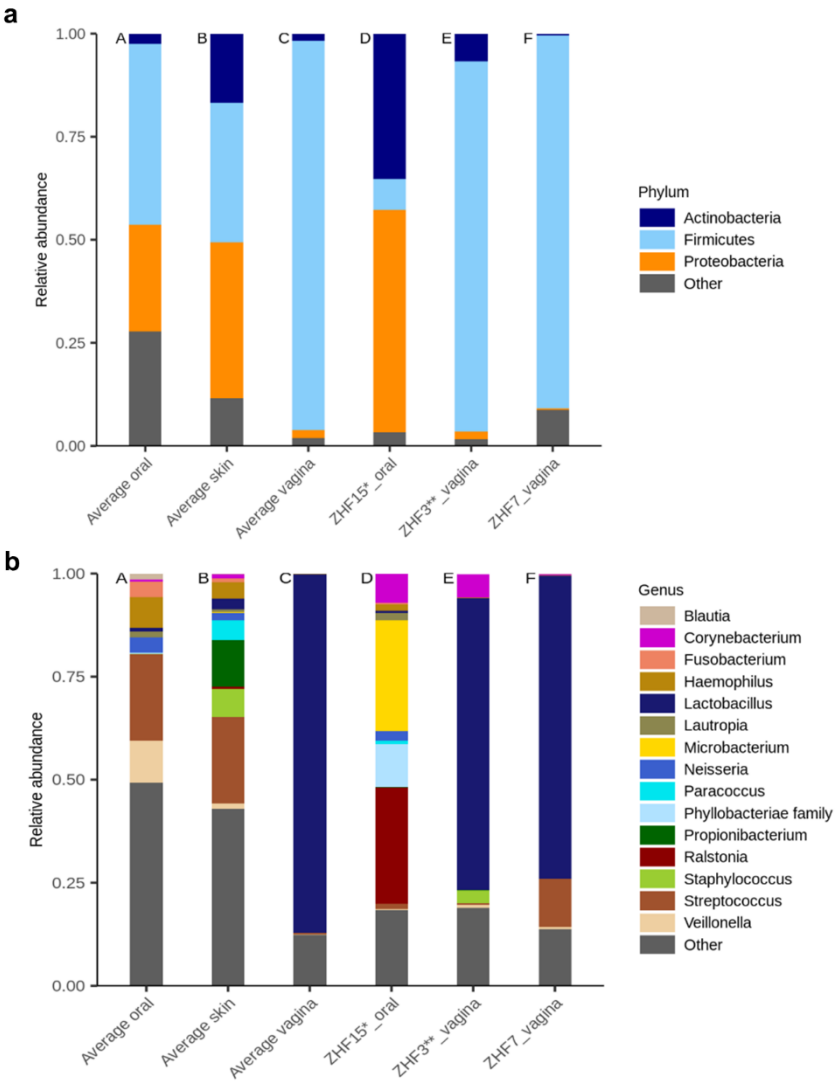


Figure 4. Taxonomic profile of mock samples outliers at the phylum **(a)** and genus **(b)** level. (A), (B) and (C) Overview of the relative abundance of the bacteria present in the correctly classified newly-analyzed test and mock oral, skin and vaginal samples, respectively. (D) Relative abundance of the bacteria in the misclassified oral sample (ZHF15) with a probability of assignment of 0.99977 to skin site. (E) Relative abundance of the bacteria in the inconclusively classified vaginal sample (ZHF3) with a probability of assignment of 0.55730 and 0.44223 to vagina and skin site, respectively. (F) Relative abundance of the bacteria in the correctly-classified vaginal sample (ZHF7), though with a low-probability of 0.16395 of assignment to the skin site. Asterisk (*) refer to misclassified sample, double-asterisk (**) refer to inconclusively classified sample.

Comparison of new microbiome with previous human mRNA tissue classification

The 22 aged mock casework samples had previously been included in collaborative inter-laboratory exercises on human mRNA-based tissue identification organized by the EDNAP group (24-26) and the EUROFORGEN_NoE Consortium (27) (Supplementary Table S4). For three out of the 22 samples mRNA-based tissue identification data were not available for comparison. Additionally, three samples were human RNA-based analyzed to test cross-reactivity with other tissue types rather than for their respective true tissue type (ZHF16, ZHF19, ZHF24). For the remaining 16 aged mock samples, the comparison of tissue classification based on the new microbiome approach and the previously applied human mRNA approach revealed clearly improved classification for 12 (75%) of the 16 samples as achieved with our microbiome approach. One aged oral mock sample (ZHF2, 5µL saliva on paper) was found challenging with the mRNA approach since saliva human mRNA markers were detected in 4 or less of the 13 participating laboratories (24). However, with our microbiome approach accurate classification to the oral category was achieved by obtaining a high probability of assignment at 0.99999. On another aged oral mock sample (ZHF6, buccal swab), mRNA outcomes concluded the presence of saliva, as well as a possible minor contribution of vaginal secretion, the latter being incorrect based on the sampling set-up (27). With our microbiome approach, we achieved correct oral assignment with a high probability of 0.99999. For three oral samples, better outcomes were obtained with our microbiome-based approach than with mRNA-based analyses: for ZHF14 (1/4 chewing gum), ZHF17 (5 µl saliva on glass slide) and ZHF18 (licked plastic spoon) saliva mRNA markers were detected in 8 or less of the 18 participating laboratories for the former and in 9 or less of the 13 participating laboratories for the other two samples (24). We achieved correct oral assignment with a high probability of ≥0.99997 for the three of them. Additionally, 87% of the target mRNA MPS reads in ZHF5 (50 µl saliva) concluded saliva (27); however, our microbiome approach assigned the sample to the oral category with a probability of 0.99999. In one of the aged oral mock samples our microbiome approach revealed problems not seen with the previous human mRNA approach. In the aged oral mock sample (ZHF15, 5µL saliva on pad) that was misclassified as skin with our microbiome approach with a probability of 0.99977, the previous mRNA data concluded saliva. However, in the previous mRNA studies the sample was only tested for saliva mRNA markers, but not for mRNA markers for skin and vaginal secretion (24); hence the comparison is not completely fair.

Overall, for the aged skin mock samples (ZHF4, ZHF21, ZHF22, ZHF23 and ZHF25, small swab from palm, hand print on glossy paper, key from computer keyboard, finger print on glass slide and scraped skin from back of hand, respectively), the mRNA-based approach performed poorly (26, 27) whereas our microbiome approach achieved correct skin assignment with high probabilities at ≥0.99998 for all of them. In the remaining aged skin mock sample (ZHF8, skin swab (27)) both approaches showed correct results. The improved tissue identification achieved with the microbiome approach in these samples may be explained by higher tissue specificity of our microbiome approach compared to the previously used mRNA approach, a higher abundance of microbial over human cells in the analyzed samples, sample degradation impacting on human mRNA but not microbial DNA, or a combination of these possible factors.

One aged vaginal mock sample (ZHF20) was classified with our microbiome approach at the vagina category with a probability of 0.99987, improving the mRNA-based

results (25). Another aged vaginal mock sample (ZHF7) was correctly assigned to the vagina category with our microbiome approach with a probability of 0.76002, with similar performance to previous RNA approaches. The human RNA massively parallel sequencing (MPS)-based results for this sample showed true vaginal classification (78% of target reads), though with minor contributions from other tissues than vaginal secretion such as blood (9%), semen (3%), saliva (7%) and skin (2%) (27). For the inconclusively classified aged mock vaginal sample with the microbiome approach (ZHF3; 0.55730 probability of assignment to the vagina category and 0.44223 to the skin category), human vaginal mRNA markers were detected in 6 or more of the 22 participating laboratories (25).

DISCUSSION

Tissue identification in a forensic investigation allows for the evaluation of the crime relevance of the biological trace found at a crime scene and used for DNA-based individual identification and may provide crucial information about the activity level of the DNA-identified trace donor. Previous attempts on forensic tissue identification using human biomarkers highlighted difficulties when dealing with human epithelial materials derived from skin, oral and vaginal sites, which is explained by the overlap in human cell composition between these different epithelial materials, leading to cross-identification, reduced specificity and thus limited forensic use (2). The human microbiome, however, provides a suitable resource to overcome this limitation, since many body sites including oral cavity, skin and vagina harbor distinctive microbial communities (4, 42, 43).

Here, we introduce a novel taxonomy-independent deep learning microbiome profiling approach for identifying skin, oral and vaginal samples that uses sequence information from all the bacteria represented in a DNA sample. Moreover, our new approach that is based on massively parallel sequencing uses a large number of reference samples for training purposes, which allows us to achieve reliable and accurate tissue identification in different sets of validation samples, for which we *de novo* generated microbiome data in a way that allows combined use together with the reference data established by the HMP.

Our taxonomy-independent approach represents a single-test approach applicable for all three different tissue types simultaneously, which is decisive in forensic investigations where the biological material is often limited. A general advantage of microbiome-based tissue identification relative to human mRNA-based methods currently in forensic use in some countries (3, 44) is that STR profiling for individual identification (1) and tissue identification can be performed on the very same DNA extract. By using bacterial-specific nucleic acids isolation methods, both human and microbial DNA are obtained for downstream analyses. Although the combination of STR-based individual identification and mRNA-based tissue identification from the same human biological sample was demonstrated recently via DNA/RNA co-isolation and simultaneous DNA/RNA targeted massively parallel sequencing, this approach required independent treatment of the RNA and DNA fractions before sequencing (45).

A limitation of previous microbial genetic attempts on forensic tissue identification is the small study sample size leading to unreliable outcomes due to under-powered detection of individual variation. Here, we aimed to overcome this limitation by using a large reference dataset from the HMP for training the deep learning neural networks. By harmonizing DNA

extraction and 16S rRNA gene sequencing protocols we made sure that the previously established HMP dataset used here for training purposes and the *de novo* generated datasets used for validation purposes could indeed be combined in the data analysis. To represent skin in the reference dataset, we used samples from four skin HMP categories i.e. right and left antecubital fossa, and right and left retroauricular crease. We are aware of the fact that these specific skin areas represented in the training sample set do not overlap with those represented in the test sample set (i.e. hand skin, more specifically the down side of the distal phalange of the index fingers). However, as implied by the high skin probabilities we achieved in the skin test samples, these skin HMP categories are good proxies for hand skin.

To test for the influence of touching other body parts or other objects, we used hand skin samples with gloves worn for one hour before sampling, and without gloves. We saw, however, that the probabilities of assignment to skin from samples with and without glove use were nearly the same for all the donors, excluding misclassified and inconclusively classified skin samples (Supplementary Figure S5). This indicates that with the microbiome approach we introduce here, the possible pick-up of non-authentic microflora (which would be reduced by wearing gloves) is not a serious problem for most of the cases. This agrees with a previous study in which participants were asked to wash their hands one hour before sampling, resulting in a minimal but insignificant community difference compared to non-washed hand samples (46).

Regarding oral HMP samples used for DL training purposes, all nine oral categories present in the HMP dataset (Table 1) were used. Though dissimilarities in the microbial communities between saliva and dental sites have been reported (47, 48), in our analysis all oral HMP samples used cluster together and separately from the samples of the other two body sites (Figure 1). This suggests that microbial differences between the different oral categories represented by the HMP samples are much smaller than those between oral and the other two body sites. Similar findings were obtained for the samples from the three vaginal sample HMP categories used (Table 1, Figure 1).

With our novel approach, we achieved highly reliable and accurate tissue identification from epithelial materials collected at the three body sites: skin, oral and vagina. Misclassifications achieved in the test samples were very low (3 skin samples). In the oral-classified skin sample (d19_Swo) *Streptococcus* was the dominant genus, while the correctly-classified skin sample with glove from the same donor (d19_Sw) presented very low levels of this genus (Figure 3b). Thus, it could be argued that the origin of *Streptococcus* in the oral-classified skin sample without glove came from an environmental source, such as commonly-consumed fermented products (49), leading to the *Streptococcus* overrepresentation in this sample. In contrast, in the case of the inconclusively classified skin sample with glove (d23_Sw), both skin samples from donor 23, with and without glove usage, showed high levels of *Streptococcus* (Figure 3b), which suggests that the high skin load of this bacteria genus was of different than external origin (or the external *Streptococcus* sources was introduced to the skin prior to glove usage). Given the distinct general taxonomic profiles, this individual could suffer from any common skin disorder, in agreement to that reported by other studies (50). This particular case points out the necessity of establishing a more diverse DL training dataset that includes different skin microbiota states (i.e. common skin disorders) that can ultimately allow for correct assignation of such samples to the skin category in the future.

The highly accurate tissue-type classification our novel microbiome approach achieved in the test samples was also seen in the forensic mock casework samples (39 of 41 samples correctly classified). Here we used freshly-prepared mock samples as well as aged mock samples, the latter to additionally investigate the impact of sample storage time on the tissue-classification outcome. In the freshly-prepared mock samples, we varied sample substrate (fabric, eating and office utensils), storage condition (temperature or relative humidity) or minor storage time until DNA isolation (fresh or 48 hours). Given the completely accurate tissue-classification our microbiome approach revealed for all of the 19 freshly-prepared mock casework samples from all three body sites, we conclude that these external parameters have no significant impact on the performance of our approach. This observation agreed with previous reports in which bacterial communities were largely unaffected by differences in short-term storage conditions (51). Moreover, the average probabilities of assignment at each body-site category obtained for these freshly-prepared mock samples were very similar to those obtained in the 110 test samples, which provides further confidence in our approach.

The good performance of our proposed approach as revealed in the test samples and the freshly prepared mock casework samples was further supported by the results obtained in the aged mock casework samples, where 20 (91%) of the 22 samples tested were tissue-type classified correctly and only one was incorrectly classified, while one was inconclusive. Notably, in contrast to the test samples and the freshly-prepared mock casework samples, the aged mock samples were analysed in blind way, since these samples were prepared in a different laboratory (Zurich) than were the DNA testing was performed (Rotterdam). The proportion of correctly tissue-type assignment was similar in samples stored for 1 year and those stored for ≥7 years, underlining that storage time is not a major factor impacting on the performance of our microbiome approach.

The one aged mock sample misclassified with our microbiome approach was a sample prepared by putting 5 µl saliva on a pad (ZHF15), which our microbiome approach classified as skin sample with a high probability of 0.99977. This misclassification might be caused by the long storage time (7.6 years). However, that storage time alone cannot explain misclassification or inconclusive classification is indicated by our findings for yet other four aged oral mock samples stored similarly long (≥7 years) were correctly classified high probabilities for oral assignment of 0.99999. The taxonomic analysis revealed increased presence of *Ralstonia, Microbacterium, Corynebacterium* and *Phyllobacteriae* family at the genus level (Figure 4b), as well as decrease relative abundance of oral-associated genera. In any case, one shall keep in mind that it is not typical that a human biological crime scene trace is recovered for forensic analysis many years after sample deposition at the crime scene.

For the vaginal samples, we noted that the probabilities for correct vaginal assignment in two of the aged mock samples (ZHF7 and ZHF3, 5x5 mm piece of worn underpants and ¼ vaginal swab, respectively) were lower than for the other ones; where for one sample (ZHF7) the vagina probability was still above the 0.7 threshold used, resulting in correct assignment, while for the other (ZHF3) it was well below, leading to an inconclusive classification outcome. As the taxonomic analysis revealed, some genera in these two samples resembled to those in the correctly-classified skin samples (*Streptococcus* in ZHF7 and *Corynebacterium* and *Staphylococcus* in ZHF3) (Figure 4b). This could be explained by the swabbing collection procedures applied for these vaginal samples, being difficult to avoid any contact with the

closest skin areas, hence with the skin microbiome, resulting in some low probability of assignment to the skin category.

Overall, a successful microbiome application depends on both technical and biological issues. Especially when microbiome profiles from different data sources are aimed to be combined, as done here for training and validation samples, care needs to be applied on method homogenization. Previous studies demonstrated that DNA extraction methods impact on outcomes of microbiome profiling (17). In this study, sample collection, DNA isolation and 16S rRNA gene amplification primers used to generate the validation data were adapted to be as similar as those previously used to generate the HMP data we applied here for training purposes to reduce technical biases. Because of these method homogenizations steps, an important quality of our study is that *de novo* generated data from our validation samples, in contrast with other reports (52), clustered with the previous generated data from the reference samples used here for training purposes (Figure 2). Moreover, data from the HMP training and the validation samples were largely comparable with each other, even though different MPS technologies were used to generate these two datasets. We therefore demonstrate that although these different MPS technologies generate sequences of different length and depth, no significant impact was noted as long as other methodological points are homogenized.

Regarding future work, from the technical side, further forensic validation testing of our new microbiome approach shall be performed in more complex mock casework samples, such as mixed multi-tissue samples, to establish data evidence and guidelines for interpreting data for mixed tissue samples. Additionally, from a biological side, potential host's factors affecting specific body sites microbial communities shall be further investigated to better characterize the whole picture of microbial inter- and intra-individual variation, including microbiome consequences of disease conditions. Depending on the outcomes of such studies, the training dataset of the DL approach may need to be revised to consider such outcomes in future tissue identification using our microbiome approach.

CONCLUSION

We introduce a novel microbiome approach that is based on massively parallel sequencing of the 16S rRNA gene and taxonomic-independent deep learning networks trained with large reference data, for body source identification of skin, oral and vaginal samples that are difficult to differentially identify with human cell biomarkers. The results of the *de novo* analysed validation test samples demonstrate the accuracy and robustness of our microbiome approach. The forensic suitability of this new approach is highlighted by the outcomes of the *de novo* analysed forensic mock casework samples. We envision this novel microbiome approach introduced here to be applied for identifying human epithelial materials in future forensic casework and any other applications where tissue-type knowledge is unavailable but required. The 16S rRNA gene sequencing data generated in this study as well as the codes and executables of the taxonomy-independent deep learning approach, and a user manual are all publicly available through the website of the Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam (https://www.erasmusmc.nl/genetic_identification/resources/).

ACKNOWLEDGEMENTS

The authors are grateful to all the volunteers for their kindly donation of the biological samples, and to the HMP for making their data publicly available.

AUTHOR'S CONTRIBUTIONS

AR and MK conceived this study, with the collaboration of CDL and AV. CH contributed to sample collection. CDL performed the experiments. CDL, DMG and OL prepared the data. OL designed and implemented the deep learning and contributed to the corresponding section in the manuscript. DR performed the taxonomic profiling with QIIME supported by RK and AGU. DMG provided bioinformatics support. CDL and MK wrote the manuscript with contributions from AR, AV and OL. All authors read and approved the final manuscript.

FUNDING

No particular funding was received for this study. The work of CDL, AR, AV, DR, RK, AGU and MK is supported by Erasmus MC, University Medical Center Rotterdam. OL was supported by a Ramón y Cajal grant from the Spanish Ministerio de Economía y Competitividad (MINECO) with reference RYC-2013-14797, and with the support of Secretaria d'Universitats i Recerca del Department d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 937).

DATA AVAILABILITY STATEMENT

Raw 16S rRNA gene sequencing data generated in this study using the Ion Torrent™ PGM (Thermo Fisher Scientific), codes and executables of the taxonomy-independent deep learning approach, and a manual of the user are freely available through the website of the Department of Genetic Identification, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands (https://www.erasmusmc.nl/genetic_identification/resources/).

REFERENCES

1. Butler JM. 2004. Short tandem repeat analysis for human identity testing. Curr Protoc Hum Genet Chapter 14:Unit 14.8.
2. Harbison S, Fleming R. 2016. Forensic body fluid identification: state of the art. Res Rep Forensic Med Sci (Dove Med Press) 6:13.
3. Sijen T. 2015. Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers. Forensic Sci Int Genet 18:21-32.
4. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694-1697.
5. Sender R, Fuchs S, Milo R. 2016. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell 164:337-340.
6. Nakanishi H, Kido A, Ohmori T, Takada A, Hara M, Adachi N, Saito K. 2009. A novel method for the identification of saliva by detecting oral streptococci using PCR. Forensic Sci Int Genet 183:20-23.
7. Giampaoli S, Berti A, Valeriani F, Gianfranceschi G, Piccolella A, Buggiotti L, Rapone C, Valentini A, Ripani L, Romano Spica V. 2012. Molecular identification of vaginal fluid by microbial signature. Forensic Sci Int Genet 6:559-564.
8. Fleming RI, Harbison S. 2010. The use of bacteria for the identification of vaginal secretions. Forensic Sci Int Genet 4:311-315.
9. Choi A, Shin KJ, Yang WI, Lee HY. 2014. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med 128:33-41.
10. Witkin SS, Linhares IM, Giraldo P. 2007. Bacterial flora of the female genital tract: function and immune regulation. Best Pract Res Clin Obstet Gynaecol 21:347-354.
11. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108:4680-4687.
12. Zou K-N, Ren L-J, Ping Y, Ma K, Li H, Cao Y, Zhou H-G, Wei Y-L. 2016. Identification of vaginal fluid, saliva, and feces using microbial signatures in a Han Chinese population. J Forensic Leg Med 43:126-131.
13. Benschop CCG, Quaak FCA, Boon ME, Sijen T, Kuiper I. 2012. Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? Int J Legal Med 126:303-310.
14. Doi M, Gamo S, Okiura T, Nishimukai H, Asano M. 2014. A simple identification method for vaginal secretions using relative quantification of Lactobacillus DNA. Forensic Sci Int Genet 12:93-99.
15. Akutsu T, Motani H, Watanabe K, Iwase H, Sakurada K. 2012. Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. Leg Med 14:160-162.
16. Manuel A, Rao JV, John K, Aranjani JM. 2014. Biofilm Production and Antibiotic Susceptibility of Planktonic and Biofilm Bacteria of Canine Dental Tartar Isolates. Acta Sci Vet 42.
17. Rantakokko-Jalava K, Jalava J. 2002. Optimal DNA isolation method for detection of bacteria in clinical specimens by broad-range PCR. J Clin Microbiol 40:4211-4217.
18. Human Microbiome Project C. 2012. A framework for human microbiome research. Nature 486:215-21.
19. Human Microbiome Project Data Analysis and Coordination Center. 2019. Accessed 24, May 2018. https://www.hmpdacc.org/hmp/(2017).
20. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.
22. Nenadic O, Greenacre M. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. J Stat Softw 20:13.
23. Manual of Procedures for Human Microbiome Project. 2010. Accessed 24, May 2018. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd003190.2.
24. Haas C, Hanson E, Anjos MJ, Banemann R, Berti A, Borges E, Carracedo A, Carvalho M, Courts C, De Cock G, Dotsch M, Flynn S, Gomes I, Hollard C, Hjort B, Hoff-Olsen P, Hribikova K, Lindenbergh A, Ludes B, Maronas O, McCallum N, Moore D, Morling N, Niederstatter H, Noel F, Parson W, Popielarz C, Rapone C, Roeder AD, Ruiz Y, Sauer E, Schneider PM, Sijen T, Court DS, Sviezena B, Turanska M, Vidaki A, Zatkalikova L, Ballantyne J. 2013. RNA/DNA co-analysis from human saliva and semen stains--results of a third collaborative EDNAP exercise. Forensic Sci Int Genet 7:230-9.
25. Haas C, Hanson E, Anjos MJ, Ballantyne KN, Banemann R, Bhoelai B, Borges E, Carvalho M, Courts C, De Cock G, Drobnic K, Dotsch M, Fleming R, Franchi C, Gomes I, Hadzic G, Harbison SA, Harteveld J, Hjort B,

Hollard C, Hoff-Olsen P, Huls C, Keyser C, Maronas O, McCallum N, Moore D, Morling N, Niederstatter H, Noel F, Parson W, Phillips C, Popielarz C, Roeder AD, Salvaderi L, Sauer E, Schneider PM, Shanthan G, Court DS, Turanska M, van Oorschot RA, Vennemann M, Vidaki A, Zatkalikova L, Ballantyne J. 2014. RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: results of a fourth and fifth collaborative EDNAP exercise. Forensic Sci Int Genet 8:203-12.

26. Haas C, Hanson E, Banemann R, Bento AM, Berti A, Carracedo A, Courts C, Cock G, Drobnic K, Fleming R, Franchi C, Gomes I, Hadzic G, Harbison SA, Hjort B, Hollard C, Hoff-Olsen P, Keyser C, Kondili A, Maronas O, McCallum N, Miniati P, Morling N, Niederstatter H, Noel F, Parson W, Porto MJ, Roeder AD, Sauer E, Schneider PM, Shanthan G, Sijen T, Syndercombe Court D, Turanska M, van den Berge M, Vennemann M, Vidaki A, Zatkalikova L, Ballantyne J. 2015. RNA/DNA co-analysis from human skin and contact traces--results of a sixth collaborative EDNAP exercise. Forensic Sci Int Genet 16:139-147.
27. Ingold S, Dorum G, Hanson E, Berti A, Branicki W, Brito P, Elsmore P, Gettings KB, Giangasparo F, Gross TE, Hansen S, Hanssen EN, Kampmann ML, Kayser M, Laurent FX, Morling N, Mosquera-Miguel A, Parson W, Phillips C, Porto MJ, Pospiech E, Roeder AD, Schneider PM, Schulze Johann K, Steffen CR, Syndercombe-Court D, Trautmann M, van den Berge M, van der Gaag KJ, Vannier J, Verdoliva V, Vidaki A, Xavier C, Ballantyne J, Haas C. 2018. Body fluid identification using a targeted mRNA massively parallel sequencing approach - results of a EUROFORGEN/EDNAP collaborative exercise. Forensic Sci Int Genet 34:105-115.
28. Castelino M, Eyre S, Moat J, Fox G, Martin P, Ho P, Upton M, Barton A. 2017. Optimisation of methods for bacterial skin microbiome investigation: primer selection and comparison of the 454 versus MiSeq platform. BMC Microbiol 17.
29. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. 2010.16S 454 Sequencing Protocol HMP Consortium.
30. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
31. Chen H-Y. 2016. TensorFlow–A system for large-scale machine learning. Proc USENIX Symp Oper Syst Des Implement. OSDI 265-283S.
32. Chollet F. 2015. Keras. GitHub. https://github.com/fchollet/keras.
33. Karlik B, Olgac AV. 2011. Performance analysis of various activation functions in generalized MLP architectures of neural networks. Int J Artif Intell and Expert Syst 1:111-122.
34. Kingma DP, Ba JL. 2014. Adam: A method for stochastic optimization. Proc 3rd Int Conf Learn Representations.
35. Goodfellow I, Bengio Y, Courville A. 2016. Deep learning. MIT Press.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825-2830.
37. Hadley W. 2009. ggplot2: Elegant graphics for data analysis. Springer Publishing Company, Incorporated.
38. Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996-8.
39. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-D596.
40. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261-7.
41. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335-6.
42. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng QD, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Brooks JP, Buck GA, Buhay CJ, Busam DA, Campbell JL, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207-214.
43. Huse SM, Ye YZ, Zhou YJ, Fodor AA. 2012. A core human microbiome as viewed through 16S rRNA sequence clusters. Plos One 7:e34242.

44. van den Berge M, Carracedo A, Gomes I, Graham EAM, Haas C, Hjort B, Hoff-Olsen P, Maroñas O, Mevåg B, Morling N, Niederstätter H, Parson W, Schneider PM, Court DS, Vidaki A, Sijen T. 2014. A collaborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results. Forensic Sci Int Genet 10:40-48.
45. Zubakov D, Kokmeijer I, Ralf A, Rajagopalan N, Calandro L, Wootton S, Langit R, Chang C, Lagace R, Kayser M. 2015. Towards simultaneous individual and tissue identification: A proof-of-principle study on parallel sequencing of STRs, amelogenin, and mRNAs with the Ion Torrent PGM. Forensic Sci Int Genet 17:122-128.
46. Meadow JF, Altrichter AE, Green JL. 2014. Mobile phones carry the personal microbiome of their owners. PeerJ 2:e447.
47. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C, Izard J. 2012. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13:R42.
48. Zaura E, Keijser BJF, Huse SM, Crielaard W. 2009. Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9:259.
49. Beresford TP, Fitzsimons NA, Brennan NL, Cogan TM. 2001. Recent advances in cheese microbiology. Int Dairy J 11:259-274.
50. Gao Z, Tseng C, Strober BE, Pei Z, Blaser MJ. 2008. Substantial alterations of the cutaneous bacterial biota in psoriatic lesions. PLoS One 3:e2719.
51. Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N. 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett 307:80-6.
52. Hanssen EN AE, Rudi K, Gil P, Snipen L. 2017. Body fluid identification from microbial patterns for forensic application. Forensic Sci Int Genet 30:10-17.

SUPPLEMENTARY MATERIALS

Supplementary File S1. SRA numbers of the 1,636 skin, oral and vaginal samples retrieved from HMP 16S-PP1 and validated for V1-3 and V3-5 target sequences. These samples comprise the training set of the 50 DL neural networks. This file can be found in the online version of the manuscript (doi: 10.1016/j.fsigen.2019.03.015).

Supplementary Material S2. Sample collection procedures and instructions for participants before sample donation.

Skin samples: One hour before sample collection donors were asked to wash both hands with water, no hygienic products and immediately afterwards to wear a nitrile powder-free examination glove in one of the two hands (the non-dominant hand). After one hour and removal of the glove, two skin samples were collected from each donor, down side of the distal phalange of the index finger from both right and left hands. This was to test whether body site classification was influenced by the presence (hand without glove) or absence (hand with glove) of potential microbiota picked up from the environment in hand skin samples. To avoid sample cross-contamination, the person in charge of sample collection used a new pair of gloves with each collected specimen. Skin surface specimens were collected using PurFlock Ultra 6" sterile standard flock swabs (Puritan, USA) moistened with sterile SCF-1 solution (50mM Tris buffer (pH7.6), 1mM EDTA (pH8.0), and 0.5% Tween-20). Swab shaft was hold parallel to the skin surface and swab head was rubbed back and forth approximately 50 times along the sample site, applying firm pressure.

Vaginal secretion samples: Female donors were asked not to have sexual vaginal activity, use genital wipes, feminine sprays or spermicides 48 hours before vaginal fluid collection. Participants were provided with the previously mentioned swabs to collect the specimen by themselves in private. They were asked to place the swab at the vaginal introitus and rotate it along the lumen with a circular motion 5 times.

Saliva samples: One and a half hours prior to saliva collection donors were requested not to brush their teeth, use mouthwash, eat any food, use chewing gums and drink any other drinks than non-sparkling water. Participants were asked to collect saliva in the mouth for at least one minute and spit into a sterile tube and repeat the process multiple times up to 5mL of saliva.

Immediately after sample collection, the swab head (skin and vaginal secretion samples) was broken off into a 2mL PowerBead tube containing 750µl solution (Qiagen, Germany) in those cases in which DNA isolation was carried out straightaway. For those samples not directly processed, swab head was broken off into a 1.5mL Eppendorf tube and stored at -80°C until DNA isolation. In the case of saliva samples, tubes were centrifuged at 2600g for 15 minutes at room temperature to separate solid from liquid material. Saliva supernatant (500µl) was transferred to a 2mL PowerBead tube containing 750µl solution (Qiagen, Germany). In those cases in which the saliva specimens were not directly processed, collection tubes were stored at -80°C until DNA isolation.

Supplementary Material S3. Freshly-prepared mock casework samples collection procedures.

Skin samples: A total of 4 mock casework skin samples were prepared. A keyboard, a light switch, a door latch and a key were firstly treated with DNA Away™ Surface Decontaminant (Thermo Fisher Scientific). After 48 hours being normally touched by people, sample sites were swabbed. Before sample collection, the key was kept into a plastic box for 48 hours under controlled relative humidity conditions (50% RH) as described by Forney *et al*. (1992).

Saliva samples: A total of 10 mock casework saliva samples were prepared, 5 using aluminium spoons and 5 using glass cups as sample substrates. Eating and drinking actions using the spoons and cups were mimicked for approximately 2 minutes in order to obtain realistic saliva samples. For each substrate type, one sample was immediately collected after preparation and processed for DNA isolation (fresh). The four remaining spoons/cups were stored at different conditions (storage at room temperature, into an incubator at 30°C, into a fridge at 4°C, and into a plastic box at 50% RH (Forney *et al*. (1992))) and were swabbed at the sample area and processed for DNA isolation 48 hours afterwards.

All mock casework skin and saliva samples were collected using PurFlock Ultra 6'' sterile standard flock swabs (Puritan, USA) moistened with sterile SCF-1 solution (50mM Tris buffer (pH7.6), 1mM EDTA (pH8.0), and 0.5% Tween-20). In the case of the skin samples, sample area comprised the entire surface of the target objects. For the saliva samples, both sides of the head of the spoons as well as the rim surface in the cups were swabbed. The swab shaft was hold parallel

to the sample area and swab head was rubbed applying firm pressure while rotating to ensure the entire swab surface made contact with the sample site. The swab was not rotated more than once to avoid redepositing biological material. Immediately after collection, the swab head was broken off into a 2mL PowerBead tube containing 750µl solution (Qiagen, Germany) and processed or stored at -20°C until DNA isolation.

Vaginal secretion samples: A total of 5 mock casework vaginal secretion samples were collected in private by the donor herself by using PurFlock Ultra 6'' sterile standard flock swabs (Puritan, USA). The swab head was placed at the vaginal introitus and rotated along the lumen with a circular motion five times. Samples were collected at intervals of 45min during the same day to allow for enough biological material to be loaded in all samples. Immediately after sample collection, fresh vaginal secretion was deposited on a fabric by holding the swab shaft parallel to it and rubbing the swab head with a circular motion applying firm pressure along the fabric sample site (square of 4 cm^2). Sterile scissors were used to cut a square of approximately 1 cm^2 of the sample area on the fabric. Afterwards, one sample was immediately processed for DNA isolation (fresh). The other four mock casework samples were stored at different conditions (storage at room temperature, 30°C, 4°C, and 50% RH, as previously mentioned) immediately after preparation and were processed for DNA isolation 48 hours afterwards.

Forney CF, Brandl DG. 1992. Control of humidity in small controlled environment chambers using glycerol-water solutions. HortTechnology 2:52-54.

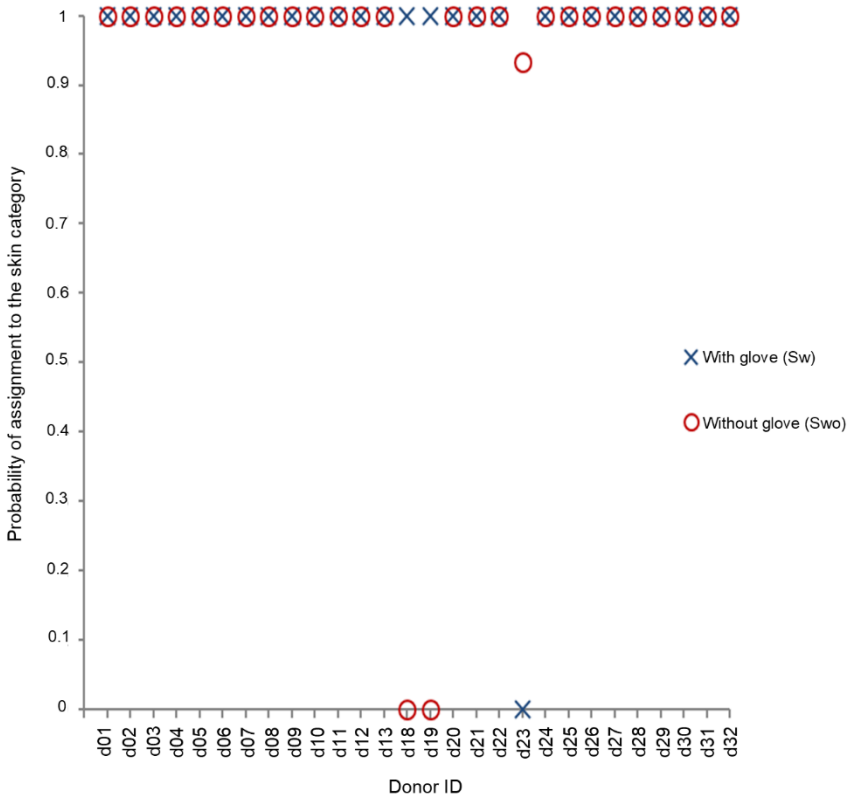
Supplementary Table S4. Aged mock casework samples description, collection procedures and EDNAP/ EUROFORGEN_NoE exercises references (in ascendant age order).

Sample ID	Body site category	Description	Age (years)	Sample collection specifications	Sex of donor	EDNAP/ EUROFOR-GEN exercise (reference)	EDNAP/ EUROFOR-GEN sample ID
ZHF10	Oral	50 µL saliva	1	Saliva collected in sterile microtube.	F	*	-
ZHF11	Vagina	¼ vaginal swab	1	Vaginal secretion taken using sterile cotton swab.	F	*	-
ZHF12	Skin	Skin swab	1	Hand rubbed using a moistened (RNase-free H_2O), sterile cotton swab.	F	*	-
ZHF5	Oral	50µL saliva	2	Saliva collected in sterile microtube.	F	MPS1 (27)	15
ZHF6	Oral	Buccal swab	2	Buccal sample taken using sterile cotton swab.	F	MPS1 (27)	16
ZHF7	Vagina	¼ vaginal swab	2	Vaginal secretion taken using sterile cotton swab.	F	MPS1 (27)	19
ZHF8	Skin	Skin swab	2	Forearm rubbed using a moistened (80% ethanol), sterile cotton swab.	F	MPS1 (27)	23
ZHF4	Skin	Swab from palm	4.9	Palm rubbed using a moistened (RNase-free H_2O), sterile cotton swab.	M	6 (26)	1
ZHF21	Skin	Hand print on glossy paper	5	Moistened swab used over the entire surface of the glossy paper.	M	6 (26)	2
ZHF22	Skin	Key from computer keyboard	5	Moistened swab used over the upward side of the key.	Mix	6 (26)	3
ZHF23	Skin	Fingerprint on glass slide	5	Moistened swab used over the upward side of glass slide.	M	6 (26)	4
ZHF24	Oral	Saliva swab	5	Saliva collected in sterile microtube.	F	6 (26)	7
ZHF25	Skin	Scraped skin from back of hand	5	Hand rubbed using a moistened (RNase-free H H_2O), sterile cotton swab.	M	6 (26)	8
ZHF3	Vagina	5x5 mm piece of worn underpants	5.9	Underpants already in use for some time and washed as usual. After being worn for 1 day, the closest part in contact with the vaginal introitus was cut into pieces (5x5 mm) using a sterile blade.	F	5 (25)	10

ZHF20	Vagina	5x5 mm from sanitary towel	6	After wearing a sanitary towel for 1 day, the closest part in contact with the vaginal introitus was cut into pieces (5x5 mm) using a sterile blade.	F	5 (25)	16
ZHF19	Vagina	¼ vaginal swab	6.6	Vaginal secretion taken using sterile cotton swab.	F	4 (25)	5
ZHF2	Oral	5µL saliva on paper	7	Saliva collected in sterile microtube and 5 µL spotted on paper.	F	3 (24)	11
ZHF16	Vagina	¼ vaginal swab	7	Vaginal secretion taken using sterile cotton swab.	F	3 (24)	13
ZHF17	Oral	5µL saliva on glass slide	7	Moistened swab used over the glass surface.	F	3 (24)	14
ZHF18	Oral	Licked plastic spoon	7	Moistened swab used over both sides of the head of the spoon.	M	3 (24)	19
ZHF14	Oral	¼ chewing gum	7.6	Chewed chewing gum.	F	3 (24)	6
ZHF15	Oral	5 µL saliva on pad	7.6	Saliva collected in sterile microtube and 5 µL spotted on pad.	M	3 (24)	10

Asterisks (*) refer to samples part of the MPS2 exercise which have not been analyzed for mRNA-based tissue identification yet.

Where required, samples were collected using PurFlock Ultra 6'' sterile standard flock swabs (Puritan, USA) moistened with sterile SCF-1 solution (50mM Tris buffer (pH 7.6), 1mM EDTA (pH 8.0), and 0.5% Tween-20). Immediately after collection, the swab head or the available sample material was deposited into a 2mL PowerBead tube containing 750µl solution (Qiagen, Germany) and processed or stored at -20°C until DNA isolation.



Supplementary Figure S5. Classification of the newly-analyzed test hand skin samples. No substantial differences in the probabilities of assignment are noted for the majority of the skin samples with and without glove for the same donor. Exceptions are misclassified skin samples (d18_Swo, d19_Swo and d23_Sw) that are assigned to the oral category.

Supplementary Table S6. Probabilities of assignment of the 110 test samples to each body site category (skin, oral, vagina) with the proposed taxonomy-independent DL approach. Misclassified samples are highlighted in gray.

		Predicted site		
sample ID	**True site**	**Skin**	**Oral**	**Vagina**
d01_S	oral	1.33E-05	0.999984	3.13E-06
d02_S	oral	3.67E-06	0.999994	2.82E-06
d03_S	oral	3.97E-06	0.999993	2.75E-06
d04_S	oral	4.45E-06	0.999993	2.78E-06
d05_S	oral	3.87E-06	0.999993	2.79E-06
d06_S	oral	7.26E-06	0.999989	3.72E-06
d07_S	oral	4.03E-06	0.999993	2.78E-06
d08_S	oral	3.85E-06	0.999993	2.76E-06
d09_S	oral	3.28E-02	0.967154	5.04E-05
d10_S	oral	6.43E-05	0.999932	3.96E-06
d11_S	oral	3.13E-03	0.996765	1.06E-04
d12_S	oral	3.65E-06	0.999994	2.86E-06
d13_S	oral	3.68E-06	0.999994	2.80E-06
d15_S	oral	4.15E-06	0.999993	2.83E-06
d16_S	oral	6.28E-02	0.936968	2.02E-04
d17_S	oral	9.68E-06	0.999987	3.74E-06
d18_S	oral	2.40E-02	0.975876	1.37E-04
d19_S	oral	3.69E-06	0.999994	2.77E-06
d20_S	oral	3.63E-06	0.999994	2.86E-06
d21_S	oral	1.34E-05	0.999982	4.94E-06
d22_S	oral	4.09E-06	0.999993	2.78E-06
d23_S	oral	3.81E-06	0.999994	2.76E-06
d24_S	oral	1.12E-05	0.999986	2.99E-06
d25_S	oral	3.69E-06	0.999994	2.79E-06
d26_S	oral	3.72E-06	0.999994	2.76E-06
d27_S	oral	3.69E-06	0.999993	2.82E-06
d28_S	oral	3.73E-06	0.999994	2.76E-06
d29_S	oral	3.69E-06	0.999994	2.85E-06
d30_S	oral	3.72E-06	0.999994	2.79E-06
d31_S	oral	4.06E-06	0.999993	2.69E-06
d32_S	oral	3.62E-06	0.999994	2.91E-06
d01_Swo	skin	0.99999	7.32E-06	6.36E-06
d01_Sw	skin	0.99999	7.52E-06	6.35E-06
d02_Swo	skin	0.99999	7.36E-06	6.34E-06
d02_Sw	skin	0.99999	7.23E-06	6.47E-06
d03_Swo	skin	0.99999	8.13E-06	6.22E-06
d03_Sw	skin	0.99999	7.68E-06	6.20E-06
d04_Swo	skin	0.99999	8.36E-06	6.27E-06
d04_Sw	skin	0.99999	8.17E-06	6.24E-06
d05_Swo	skin	0.99999	7.22E-06	6.51E-06
d05_Sw	skin	0.99999	7.33E-06	6.35E-06
d06_Swo	skin	0.99999	7.28E-06	6.41E-06
d06_Sw	skin	0.99999	7.25E-06	6.44E-06
d07_Swo	skin	0.99999	7.25E-06	6.45E-06
d07_Sw	skin	0.99999	7.27E-06	6.41E-06
d08_Swo	skin	0.99999	7.28E-06	6.39E-06
d08_Sw	skin	0.99999	7.34E-06	6.35E-06
d10_Swo	skin	0.99999	7.28E-06	6.40E-06
d10_Sw	skin	0.99999	8.06E-06	6.41E-06
d11_Swo	skin	0.99999	7.24E-06	6.48E-06
d11_Sw	skin	0.99999	7.22E-06	6.48E-06
d12_Swo	skin	0.99999	7.28E-06	6.40E-06
d12_Sw	skin	0.99999	7.29E-06	6.39E-06
d13_Swo	skin	0.99999	7.26E-06	6.42E-06
d13_Sw	skin	0.99999	7.38E-06	6.32E-06
d15_Swo	skin	0.99999	7.49E-06	6.29E-06
d15_Sw	skin	0.99999	7.24E-06	6.44E-06
d18_Swo	skin	3.51E-03	0.98324	1.32E-02
d18_Sw	skin	0.99999	7.31E-06	6.38E-06
d19_Swo	skin	4.53E-06	0.99999	2.60E-06

d19_Sw	skin	0.99999	8.77E-06	6.29E-06
d20_Swo	skin	0.99999	7.67E-06	6.27E-06
d20_Sw	skin	0.99999	8.33E-06	6.21E-06
d21_Swo	skin	0.99999	7.69E-06	6.24E-06
d21_Sw	skin	0.99999	7.52E-06	6.47E-06
d22_Swo	skin	0.99999	7.27E-06	6.41E-06
d22_Sw	skin	0.99999	7.47E-06	6.26E-06
d23_Swo	skin	0.93994	6.01E-02	1.28E-05
d23_Sw	skin	7.76E-05	0.99992	2.92E-06
d24_Swo	skin	0.99999	7.24E-06	6.44E-06
d24_Sw	skin	0.99999	7.24E-06	6.45E-06
d25_Swo	skin	0.99999	7.25E-06	6.44E-06
d25_Sw	skin	0.99999	7.54E-06	6.29E-06
d26_Swo	skin	0.99999	7.29E-06	6.38E-06
d26_Sw	skin	0.99999	7.49E-06	6.31E-06
d27_Swo	skin	0.99999	7.23E-06	6.48E-06
d27_Sw	skin	0.99999	7.53E-06	6.32E-06
d28_Swo	skin	0.99992	7.45E-05	6.62E-06
d28_Sw	skin	0.99996	3.42E-05	5.95E-06
d29_Swo	skin	0.99999	7.25E-06	6.43E-06
d29_Sw	skin	0.99999	7.27E-06	6.41E-06
d30_Swo	skin	0.99999	7.23E-06	6.52E-06
d30_Sw	skin	0.99999	7.10E-06	6.90E-06
d31_Swo	skin	0.99999	7.25E-06	6.45E-06
d31_Sw	skin	0.99998	9.35E-06	6.22E-06
d32_Swo	skin	0.99999	7.25E-06	6.45E-06
d32_Sw	skin	0.99999	7.25E-06	6.45E-06
d01_V	vagina	1.76E-05	1.15E-05	0.99997
d02_V	vagina	1.76E-05	1.16E-05	0.99997
d04_V	vagina	1.70E-05	1.17E-05	0.99997
d05_V	vagina	1.70E-05	1.17E-05	0.99997
d06_V	vagina	1.78E-05	1.15E-05	0.99997
d07_V	vagina	1.72E-05	1.16E-05	0.99997
d08_V	vagina	1.73E-05	1.16E-05	0.99997
d09_V	vagina	1.79E-05	1.14E-05	0.99997
d10_V	vagina	1.33E-02	8.30E-06	0.98667
d11_V	vagina	1.70E-05	1.17E-05	0.99997
d13_V	vagina	1.73E-05	1.17E-05	0.99997
d14_V	vagina	1.71E-05	1.17E-05	0.99997
d16_V	vagina	1.78E-05	1.15E-05	0.99997
d17_V	vagina	1.72E-05	1.16E-05	0.99997
d18_V	vagina	1.84E-05	1.15E-05	0.99997
d19_V	vagina	1.75E-05	1.16E-05	0.99997
d21_V	vagina	1.74E-05	1.17E-05	0.99997
d23_V	vagina	1.75E-05	1.15E-05	0.99997
d28_V	vagina	1.77E-05	1.20E-05	0.99997
d29_V	vagina	1.93E-05	1.21E-05	0.99997
d30_V	vagina	1.89E-05	1.33E-05	0.99997
d31_V	vagina	1.83E-05	1.13E-05	0.99997
d32_V	vagina	1.76E-05	1.14E-05	0.99997

Supplementary File S7. Operational taxonomic unit (OTU) tables of the 109 test samples (skin sample d18_Swo no analyzed due to insufficiently low quality reads) and 41 mock casework samples (skin, oral, vagina). This file can be found in the online version of the manuscript (doi: 10.1016/j.fsigen.2019.03.015).

Chapter 3

Microbiome-based body site of origin classification of forensically relevant blood traces

Celia Díez López, Diego Montiel González, Cordula Haas, Athina Vidaki, Manfred Kayser

Forensic Sci Int Genet. 2020 Jul; 47:102280.

ABSTRACT

Human blood traces are amongst the most commonly encountered biological stains collected at crime scenes. Identifying the body site of origin of a forensic blood trace can provide crucial information in many cases, such as in sexual and violent assaults. However, means for reliably and accurately identifying from which body site a forensic blood trace originated are missing, but would be highly valuable in crime scene investigations. With this study, we introduce a taxonomy-independent deep neural network approach based on massively parallel microbiome sequencing, which delivers accurate body site of origin classification of forensically-relevant blood samples, such as menstrual, nasal, fingerprick, and venous blood. A total of 50 deep neural networks were trained using a large 16S rRNA gene sequencing dataset from 773 reference samples, including 220 female urogenital tract, 190 nasal cavity, 213 skin, and 150 venous blood samples. Validation was performed with *de novo* generated 16S rRNA gene massively parallel sequencing (MPS) data from 94 blood test samples of four different body sites, and achieved high classification accuracy with AUC values at 0.992 for menstrual blood (N=23), 0.978 for nasal blood (N=16), 0.978 for fingerprick blood (N=30), and 0.990 for venous blood (N=25). The obtained highly accurate classification of menstrual blood was independent of the day of the menses, as established in additional 86 menstrual blood test samples. Accurate body site of origin classification was also revealed for 45 fresh and aged mock casework blood samples from all four body sites. Our novel microbiome approach works based on the assumption that a sample is from blood, as can be obtained in forensic practise from prior presumptive blood testing, and provides accurate information on the specific body source of blood, with high potentials for future forensic applications.

INTRODUCTION

Forensic cell type or tissue identification is a key topic of interest in crime scene investigation, as it allows for establishing the crime relevance of human biological evidence found at crime scenes and for crime scene reconstruction. Blood is one of the most commonly encountered body fluids at crime scenes, and represents an important type of human biological evidence because it can serve as indicator of violent crime. Several mainly biochemical tests are available for the presumptive detection of blood in general (1). However, many forensic cases, where the presence of blood has been concluded from such presumptive testing, would benefit from more detailed knowledge of the particular body site the recovered bloodstain originated to provide additional crime-relevant information. Obvious scenarios are the differentiation between menstrual blood and venous blood in alleged sexual assault cases, and between nasal blood and venous blood in alleged violent assault cases.

For the identification of menstrual *vs.* venous blood, human RNA-based methods had been introduced using RNA markers with gene expression differences between blood from these two different sources, including messenger RNA (mRNA) (2-5), microRNA (miRNA) (6-8), circular RNA (circRNA) (9, 10), and piwi-interacting RNA (piRNA) (11) markers. The major challenge with such mostly qualitative mRNA-testing for forensic tissue identification is the complex composition of menstrual blood, containing different cell types i.e. blood cells, vaginal epithelium cells, and skin epithelium cells, which can show different expression patterns at the applied mRNA markers. Moreover, certain previously suggested mRNA markers demonstrated instable expression patterns towards the end of menstruation (12), which can lead to misclassifications. Previously suggested miRNA markers for menstrual and venous blood still require validation testing regarding their expression stability during menses and further development of current laboratory and data analysis workflows (13). Knowledge about more recently introduced circRNA (9, 10) and piRNA markers (11) for menstrual and venous blood is even more limited thus far, requiring more research to investigate their suitability for reliably identifying these two blood sources separately. For differentiating nasal *vs.* venous blood, a combination of nasal mucosa and venous blood mRNA markers have been previously proposed (14, 15). However, their reported time-wise stability is limited, which provides challenges for forensic tissue identification in aged blood samples (14). In a different study, a nasal mucosa mRNA marker was proposed (15), but was not detected in all the nasal blood samples analyzed. Moreover, the majority of the analyzed nasal blood samples in this study (15) were wrongly classified as saliva based on the mRNA multiplex employed, with some sporadic signals of vaginal secretion and menstrual blood.

DNA methylation markers have also been proposed for the differentiation of menstrual *vs.* venous blood (16-18). However, their methylation levels between these two tissues overlap, which provides challenges for correct classification (18). Moreover, there are discrepancies in the performance of certain menstrual blood DNA methylation markers between different studies (16, 17, 19, 20), which may be explained by intra- and inter-individual variation, given the aforementioned complex nature of menstrual blood, and thus, requires further research before practical forensic applications.

On the other hand, the human microbiome is expected to serve as a promising forensic resource for classifying human biological samples, including blood samples, according to their body site of origin, because different human body sites constitute niches for specialized

microbial communities (21). Previous microbial studies dealing with blood are of limited use, because of the targeted microbial species approach they applied. For instance, it had been reported that the use of DNA markers for particular bacterial species found in the vaginal environment i.e. *Lactobacillus crispatus*, *Lactobacillus gasseri*, *Lactobacillus jensenii* and/or *Lactobacillus iners*, in combination with human mRNA (15) and DNA methylation (22) markers, can identify menstrual blood by using multiplex PCR systems. However, for some analyzed menstrual blood samples, no DNA of the target bacterial species was detected with this method, while in other samples only some of the bacterial target species were detected (15, 22), which might be caused by intra- and inter-individual variation of the bacterial target species considered. Moreover, some *Lactobacillus* species, including *L. crispatus* and *L. gasseri*, have been detected on the penile environments independent of prior vaginal contact (23), which can lead to wrong conclusions.

The disadvantage of applying a species-targeted microbial approach is that typically only a small number of bacterial species is targeted, which can be overcome by applying a microbiome-based approach covering the majority of the microbial genetic information a sample contains. However, microbiome-based studies of forensically relevant blood samples are scarce and are yet unavailable regarding blood source classification. A previous microbiome study assessed the stability of the microbial community in menstrual and peripheral blood samples exposed to indoor conditions for 30 days based on a 16S rRNA gene target sequencing approach (24). In this study, all but one peripheral blood sample yielded no bacterial DNA, making their approach not suitable at least for venous blood.

Since previous studies demonstrated that the human microbiome of the vagina, nasal mucosa and skin are largely different (21), it can be expected that blood samples originated from such body sites (menstrual, nasal and fingerprick blood) can be differentiated and identified by use of a suitable microbiome approach. To the best of our knowledge, there is no published study demonstrating the feasibility and performance of such microbiome approach for identifying different forensically relevant blood samples. To overcome previous limitations, we developed and tested a taxonomy-independent deep neural network (DNN) microbiome method for classifying forensically relevant blood samples according to their body site of origin, by using the principle of our recently published microbiome approach for identifying forensically relevant epithelial materials (25), which showed high performance for our purpose. This required i) a new selection of the bacterial 16S rRNA gene sequence positions for the DNN building, ii) a newly ascertained large reference dataset set from samples of all relevant body sites for DNN training that we partly established *de novo* and partly took from publicly available sources of the Human Microbiome Project (HMP), and iii) a large test dataset from blood samples of all relevant body sites for DNN validation that we established *de novo*, as well as the testing of forensic mock casework blood samples of different sources. Finally, using taxonomic profiling, we explained the misclassified and inconclusively classified test and mock casework blood samples that our new microbiome DNN method rarely delivered.

MATERIALS AND METHODS

Reference samples and data

16S rRNA gene sequencing data from female urogenital, nasal cavity, and skin sites from the HMP 16S production phase I (16S-PP1) were obtained from the HMP data analysis and coordination center (DACC) website (26) in FASTA file format. Data comprised a total of 623 samples containing both V1-3 and V3-5 16S rRNA gene hypervariable regions. Additionally, 16S rRNA gene sequencing reference data from 150 venous blood samples were produced in-house, following the same procedure as with the test and mock casework blood samples. We used blood samples donated by healthy blood donors, who gave consent for their blood to be used for biomedical research purposes. Samples were used completely anonymously. The study was carried out in accordance with the Code of Conduct of the Federation of Medical Scientific Societies in the Netherlands (http://www.fmwv.nl). The total set of 773 reference samples used for model training were assigned to four main body site categories: vaginal epithelium (VE), nasal mucosa (NM), skin epithelium (SE), and venous blood (VB), as described in Table 1.

Table 1. Samples of which 16S rRNA gene sequencing data were used as reference data for the DNN training.

Body site category	Original body site	Number of samples
Skin epithelium (SE)	Left antecubital fossa	45
	Right antecubital fossa	44
	Left retroauricular crease	54
	Right retroauricular crease	70
Nasal mucosa (NM)	Anterior nares	190
Vaginal epithelium (VE)	Mid vagina	73
	Posterior fornix	71
	Vaginal introitus	76
Venous blood (VB)	Venous blood	150
	Total	773

Test and mock casework samples

The testing set comprised a total of 180 blood samples, including 109 menstrual blood (MB), 16 nasal blood (NB), 30 fingerprick blood (FB), and 25 venous blood (VB) samples collected from volunteers with informed consent and ethical approval. More specifically regarding the MB samples, 23 female volunteers were asked to collect a sample swab every day during their menses (ranging between 2-10 days depending on the donor), resulting in 109 MB samples in total of which 23 (i.e. one randomly drawn sample per female individual) were used for model testing and the remaining 86 for investigating the impact of menses day on body site classification accuracy.

Additional freshly-prepared mock casework blood samples were employed to mimic realistic and often encountered forensic scenarios. These consisted of 24 specimens: 6 MB, 6 NB, 6 FB and 6 VB samples. The biological material was deposited on different substrates (nylon swabs, fabric, tissue paper) and exposed to various temperatures (4°C, room temperature, 37°C) for 48 hours before DNA isolation was performed. Further, a total of 21 aged mock casework samples previously used in human mRNA-based tissue identification studies (4, 12, 27, 28) were analyzed with our approach: 6 MB, 2 NB and 13 VB samples (with the second not being included in the previously published reports). These mock casework blood samples had been previously included in collaborative exercises with a focus on

validating human mRNA-based tissue identification organized on behalf of the European DNA profiling Group (EDNAP) (12, 27, 28) and also by the European Forensic Genetics Network of Excellence (EUROFORGEN_NoE) (4). Prior to DNA extraction for the purpose of the present study, bloodstains were stored in the dark at room conditions for a time period from 1.1 years to 21 years depending on the sample. More detailed information on all mock casework blood samples is available in Supplementary Table S1.

Ethical approval for the study was obtained from the Kantonale Ethikkommission Zürich (KEK) (declaration of no objection No. 24-2015) and the Medische Ethische Toetsings Commissie Erasmus MC (MEC-2018-1731).

DNA isolation and quantification

DNA was isolated using the DNeasy PowerSoil kit (Qiagen, Germany), according to the manufacturer's instructions but with two adjustments in order to maximize DNA yield. Firstly, as proposed by Castelino *et al.* (29), the 2 mL PowerBead tubes containing both 750 µl PowerSoil® Solution C1 and the sample were incubated at 70°C for 15 minutes during cell lysis. Additionally, as recommended by the Manual of Procedures of the HMP Consortium (30), the centrifugation time in step 12 was increased from 1 to 2 minutes. Isolated bacterial DNA was quantified with the Femto™ Bacterial DNA Quantification kit (Zymo Research, CA, USA) following the manufacturer's instructions on a C1000 Touch™ Thermal Cycler (Bio-Rad).

Library preparation

Since data from the HMP was used as reference data, and in order to both reduce amplification biases reported by different primer sets and increase concordance in the results between studies, the same primers employed by the 16S 454 Sequencing Protocol of the HMP Consortium (31) were used for *de novo* MPS in this study. The V1-3 region of the 16S rRNA gene was amplified using the F27/R534 combination (5'-AGAGTTTGATCCTGGCTCAG-3', 5'-ATTACCGCGGCTGCTGG-3') and the V3-5 region using the F357/R926 combination (5'-CCTACGGGAGGCAGCAG-3', 5'-CCGTCAATTCMTTTRAGT-3').

Library preparation was performed following the *Prepare Amplicon Libraries without Fragmentation Using the Ion Plus Fragment Library Kit* protocol for 400 bp target read length (Thermo Fisher Scientific, USA). An annealing temperature of 58°C was settled for the V1-3 region and 50°C for the V3-5 region, according to the 16S 454 Sequencing Protocol of the HMP Consortium (31). Amplicons were quantified with the Quant-iT™ PicoGreen™ dsDNA Assay Kit (Thermo Fisher Scientific) following the manufacturer's instructions, based on which equimolar amplicon pools were prepared. Libraries were quantified using the Ion Library Taqman® Quantitation Kit (Thermo Fisher Scientific) following manufacturer's instructions. Libraries harboring different barcodes were diluted and pooled together in equal concentrations of 100 pM.

Template preparation and massively parallel sequencing

Emulsion PCR, bead enrichment, and chip loading were automatically performed on an Ion Chef™ instrument (Thermo Fisher Scientific) using Ion 520™ & Ion 530™ ExT Kits (Thermo Fisher Scientific). Ion 530™ chips were used (Thermo Fisher Scientific). Sequencing runs for each chip were planned within the Ion Torrent Suite™ Software v.5.10.1 (Thermo Fisher Scientific). From the pooled libraries, 50 µL were transferred to the library sample tubes of the

Ion Chef™ reagent cartridges (Thermo Fisher Scientific). The Ion Chef™ instrument was loaded as recommended by the manufacturer.

Templates were sequenced on an Ion S5™ sequencer (Thermo Fisher Scientific) using the Ion S5 ExT Sequencing Kit (Thermo Fisher Scientific) with a number of flows set at 1,350. Sequencing data were obtained from the S5 Torrent Server (Thermo Fisher Scientific) in FASTQ file format. Sequencing reads were trimmed by quality (Q<20) and length (<100bp) using Trimmomatic (32).

Ascertaining gene positions for deep neural network building

The sequences of all 773 reference samples were aligned to *Escherichia coli* str. K 12 substr. MG1655 (nc_000913.3) 16S rRNA gene using BWA-MEM (33) with parameters: -B1, -O1, -E1 and -L1. SAM files were converted into BAM files using SAMtools (34) and *mpileup* command was used to pile up all aligned positions. From each sample and position of the 16S rRNA gene, the number of reads with adenine (A), guanine (G), thymine (T), cytosine (C), insertion (+) and deletion (-) were retrieved. We decided to maximize the informative positions employed in the DNN training given that the microbial communities found on the NM and SE have many commensals in common, making it more challenging to distinguish between each other. To this, it has to be added that the reference samples obtained from the HMP and our validation samples (test and mock casework blood samples) originated from the same body site but did not represent exactly the same tissue, meaning they did not necessarily contain the exact same human or microbial cells (vaginal epithelium *vs.* menstrual blood, nasal mucosa *vs.* nasal blood, and skin epithelium *vs.* fingerprick blood). Firstly, we selected the most representative positions among all samples and subsequently, a matrix of individuals by sites was generated by merging all positions detected in at least one sample, representing a total of 920 positions. The final input matrix consisted of 773 samples and 920 positions, each position representing six cells – A, G, T, C, (+), (-). The input matrix was preprocessed with a min-max normalization (35) using the scikit-learn package (36), in order to improve the gradient algorithm during optimization and converge more efficiently, by scaling the features within a range of [0-1] for each sample.

Developing deep neural network classifier

The deep neural network (DNN) classifier for the body site of origin of human blood traces being vaginal epithelium, nasal mucosa, skin epithelium and venous blood were implemented in Tensorflow (37) and Keras (38). Input data comprised 5,520 features (six nucleotides categories - A, G, T, C, (+), (-) - multiplied by 920 retained sequence positions). The DNN architecture was based on four layers: an input, two 10-neuron hidden layers (HLs) and an output. The input-HL1 and HL1-HL2 connections consisted of a hyperbolic tangent activation function (TANH) (39), while the HL2-output was a softmax activation function. The DNN algorithm was trained using the Adam optimizer (40) at a learning rate of 0.001. Given that the number of features was much larger than the number of reference samples, we took measures to promote regularization; hence, preventing overfitting. Namely, a 5-fold stratified cross-validation was implemented to validate the loss and accuracy of our model. Moreover, to reduce the natural variance of the DNN model, we independently trained each fold 10 times (averaging/bagging) (41). Finally, we used a hidden layer dropout (41) at a rate of 0.2 and early stopping (42) with a patience parameter of 50. As a result, a total of 50 independent

DNN models were obtained. The DNN output consisted of the four assigned probabilities to each of the four body site categories vaginal epithelium (VE), nasal mucosa (NM), skin epithelium (SE), and venous blood (VB).

Statistical testing of classification accuracy

The performance of the DNN approach was evaluated based on the body site classification results of 94 blood test samples using the scikit-learn package (36). The area under the receiving operating characteristic (AUC) values shows the sensitivity of the classifier, considering both true and false positives. The precision, or positive predictive value (PPV), scores the probability that a positive prediction made by the classifier is actually true. The recall refers to the fraction of false negatives in the prediction rate (used instead of sensitivity in multi-class classifications). The F1 score considers both precision and recall. The negative predictive value (NPV) scores the probability that a negative prediction made by the classifier is actually negative. All these classification metrics were scored [0-1] in each of the four categories (MB, NB, FB, and VB), with 0 being the worst score and 1 being the best score (43). Lastly, the Mathews correlation coefficient (MCC) that takes into account all four possible categories (MB, NB, FB, and VB) considering true positives, false positives, true negatives, and false negatives, was estimated where -1 being the worst score and 1 being the best score (36). Regarding menstrual blood test samples used for classification accuracy testing, the *sample( )* function in R was used to randomly select a single sample per each donor (total of 23 samples) from the several samples we had available per donor (the remaining menstrual blood samples were used to investigate the effect of menstruation day on menstrual blood classification, see below). Two-dimensional plots from the principal component analysis (PCA) based on the first and the second principal components (PCs) using the activation values of the hidden neurons of each network ((10+10)*50 neurons per sample) were generated with the ggplot2 R-package (44).

Taxonomic profiling

For misclassified and inconclusive samples, initial filtering and de-noising steps were carried out from the 16S rRNA gene sequencing data obtained from the S5 Torrent Server (Thermo Scientific) as FASTQ files using the DADA2 R-package (v.1.6.0) (45) with parameters maxN=0 and maxEE=2, preventing unambiguous nucleotides and setting the maximum number of ‘expected errors’ allowed in a read to 2. After sample inference of true sequence variants, an amplicon sequence variants (ASVs) table was constructed and chimeric sequences were removed. The IdTaxa taxonomic classification method (46) was used for assigning taxonomy via the DECIPHER Bioconductor package (v.2.6.0) (47) at taxonomic ranks of interest (domain, phylum, class, order, family and genus) using the SILVA SSU r132 database as reference (48). In the venous blood reference samples used for DNN building, those sequencing reads assigned as ‘NA’ (not assigned) at the domain level were further investigated using standard nucleotide blast and the nucleotide collection database from the National Center for Biotechnology Information (NCBI).

RESULTS

Training taxonomy-independent deep neural networks

16S rRNA gene sequencing data from a total of 773 reference samples i.e. 220 VE, 190 NM, 213 SE and 150 VB samples were employed for training 50 DNNs in classifying blood samples according these four body sites of origin (see Materials and methods for network architecture). The performance of the DNNs in the reference dataset was visualized by generating two-dimensional PCA plots in which the first two principal components explained a large proportion of 60% of the variation present in the dataset. As evident from Figure 1a, these 773 reference samples clustered in four well-defined and largely separated data groups that correspond to the four body sites of origin of these samples i.e. VE, NM, SE, and VB.

Validating taxonomy-independent deep neural networks

The trained DNN classifier was subsequently applied to *de novo* generated 16S rRNA gene sequencing data from 94 human blood test samples from four different body sites used for validation purposes i.e. 23 MB, 16 NB, 30 FB and 25 VB samples. Based on the DNN outcomes, each of these blood samples was given a probability of assignment to each of the four body site categories VE, NM, SE and VB. The resulting classification performance metrics for each of the four blood categories are provided in Table 2. AUC values were very high at 0.992 for MB, 0.978 for NB, 0.978 for FB and 0.990 for VB (Figure 2a) . When applying the principle of our recently published approach (25) and thus excluding 11 inconclusively classified samples with probabilities of assignment below the 0.7 conclusive classification threshold at any of the four categories (Figure 2b), the resulting AUC values remained almost the same (Supplementary Table S2, Supplementary Figure S3).

Table 2. DNN classification performance metrics in 94 blood test samples.

	Menstrual blood (MB)	Nasal blood (NB)	Fingerprick blood (FB)	Venous blood (VB)
AUC	0.992	0.978	0.978	0.990
Precision	0.955	0.750	0.897	0.957
Recall	0.913	0.938	0.867	0.880
F1 score	0.933	0.833	0.881	0.917
NPV	0.972	0.986	0.938	0.958
Support	23	16	30	25
MCC	0.858			

AUC=area under the receiver operating characteristic curve; NPV=negative predictive value; MCC=Mathews correlation coefficient.

To test if the day of the MB sample donor's menses impacts on the microbiome-based VE classification accuracy, the DNN classifier was applied to *de novo* generated 16S rRNA gene sequencing data of 86 additional MB test samples collected from 23 donors at different menses days. Their obtained high probabilities of assignment to the VE category, and their consequent correct identification as MB for the majority of these samples, suggests that the

day of the menses was not a determinant factor of the performance of the DNN classifier regarding MB assignment to the correct VE category (Supplementary Figure S4).

When combining all test samples, from the total number of 180 blood samples included in this test dataset, the output of the DNN approach comprised eight (4%) misclassified samples i.e. samples with a probability of assignment ≥ 0.7 at a category other than the true body site of origin (Figure 2c). In particular, five MB samples and one FB sample were wrongly assigned to the NM category, and one NB as well as one VB sample were wrongly assigned to the SE category (see below for their taxonomy-based explanations). When the test dataset was superimposed on the two-dimensional PCA plot of the reference dataset (61% variation explained by PC1 and PC2 together for the combined dataset), the test samples largely overlapped well with their body site corresponding reference samples (Figure 1b). Rarely observed exceptions were 17 (9%) test samples placed in-between reference sample clusters, which correspond to the inconclusively classified ones, and eight (4%) test samples were grouped in the wrong cluster, which correspond to the misclassified samples. Body site assignment probabilities of all 180 human blood test samples obtained with the DNN approach are available in Supplementary Table S5.

Forensic validation using mock casework blood samples

The performance of the DNN classifier was further validated from a forensic standpoint using 24 freshly prepared and 21 aged mock casework blood samples (ranging 1.1 - 21 years) produced from blood collected from all four body sites (see Materials and methods for sample details). Applying the probability assignment threshold of 0.7 led to 6 (13%) inconclusively classified mock casework blood samples (Figure 2d). From the 24 freshly prepared mock casework blood samples, 4 (17%) samples were misclassified i.e. one FB sample and three VB samples were wrongly assigned to the VE category. Moreover, from the 21 aged mock casework sample set, 13 (62%) samples were misclassified i.e. one MB, two NB and nine VB samples were wrongly assigned to the SE category, while one MB sample was wrongly assigned to the NM category. Body site assignment probabilities of all 45 mock casework samples resulted from the DNN approach are available in Supplementary Table S6.

Explaining classification outliers by taxonomic profiling

Next, we aimed at finding an explanation for the misclassified and inconclusively classified test and mock casework blood samples and therefore performed taxonomic profiling in these samples. For this, an ASVs table was constructed from the *de novo* generated 16S rRNA sequencing data using the DADA2 R-package (45) and the IdTaxa taxonomic classification method (46) was used at different taxonomic levels.

From the seven test and mock casework MB misclassified samples, six were wrongly assigned to the NM category, which might be explained by a commonly observed decrease in the relative abundance of the *Firmicutes* phylum (Figure 3a), which corresponds to the VE-specific *Lactobacillus* at the genus level (Figure 4a). These samples also showed an increase in the relative abundances of common NM commensals, such as *Staphylococcus*, *Corynebacterium 1*, and *Anaerococcus* at the genus level (Figure 4a). The three misclassified NB samples were wrongly assigned to the SE category. Although not very clear at the phylum level (Figure 3b), these samples showed a decrease or absence in the relative abundances of common NM commensals at the genus level, such as *Corynebacterium 1* and *Moraxella*,

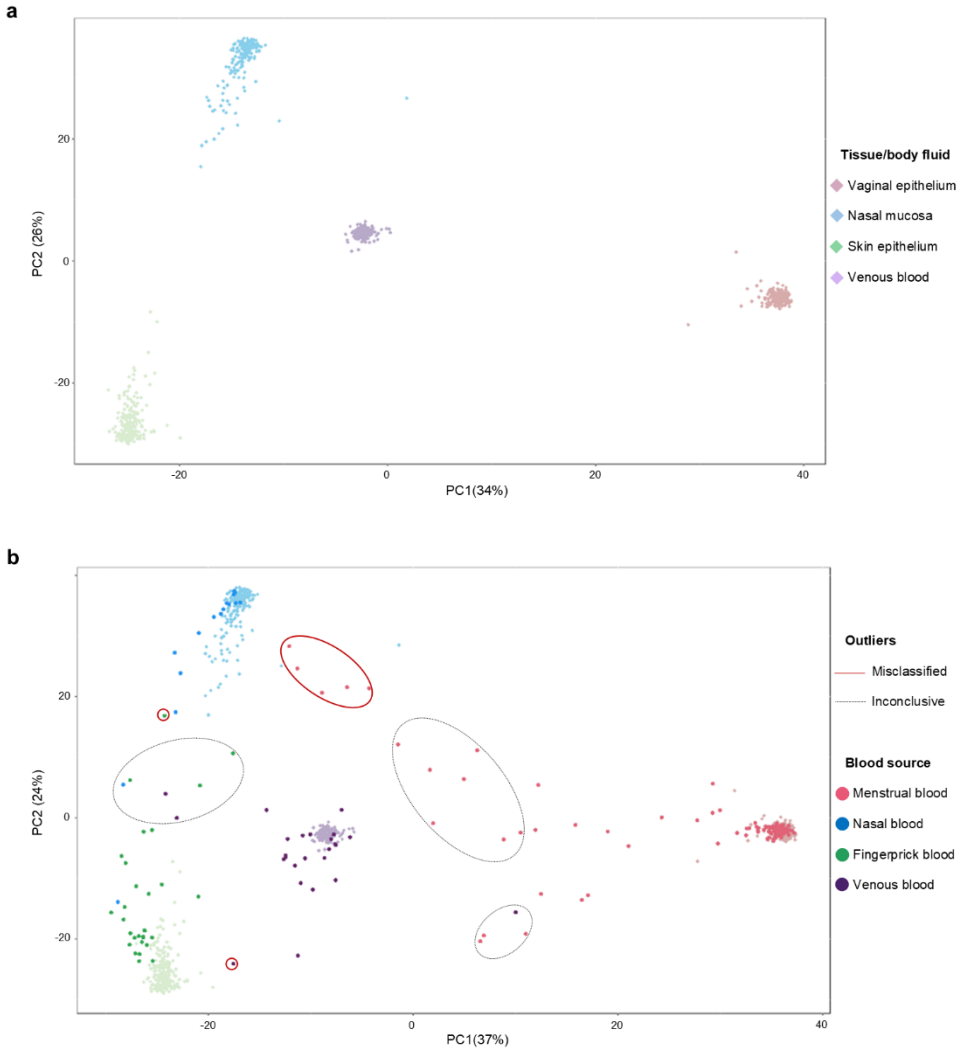


Figure 1. (a) Two-dimensional plot from principal component analysis (PCA) of the deep neural network (DNN) reference samples from four body sites. Projection of the activation values of the hidden neurons of each DL using the reference samples in the first two PCs together explain 60% of the total variation. Four distinct and largely separated clusters corresponding to the four body site categories from where the reference samples originated i.e. vaginal epithelium, nasal mucosa, skin epithelium and venous blood highlighted in light color shades. **(b)** Two-dimensional plot from principal component analysis (PCA) of the DNN reference samples (squared symbols in light color shades) with superimposed *de novo* analyzed blood test samples from four body sites (round symbols in dark color shades). The first two PCs explain 61% of the total variation. The vast majority of *de novo* analyzed blood samples cluster closely together with their body site corresponding reference samples. Exceptions are misclassified blood test samples (red continuous line) and inconclusively classified (black discontinued line) blood test samples based on the probability of assignment threshold of 0.7 for conclusive classification.

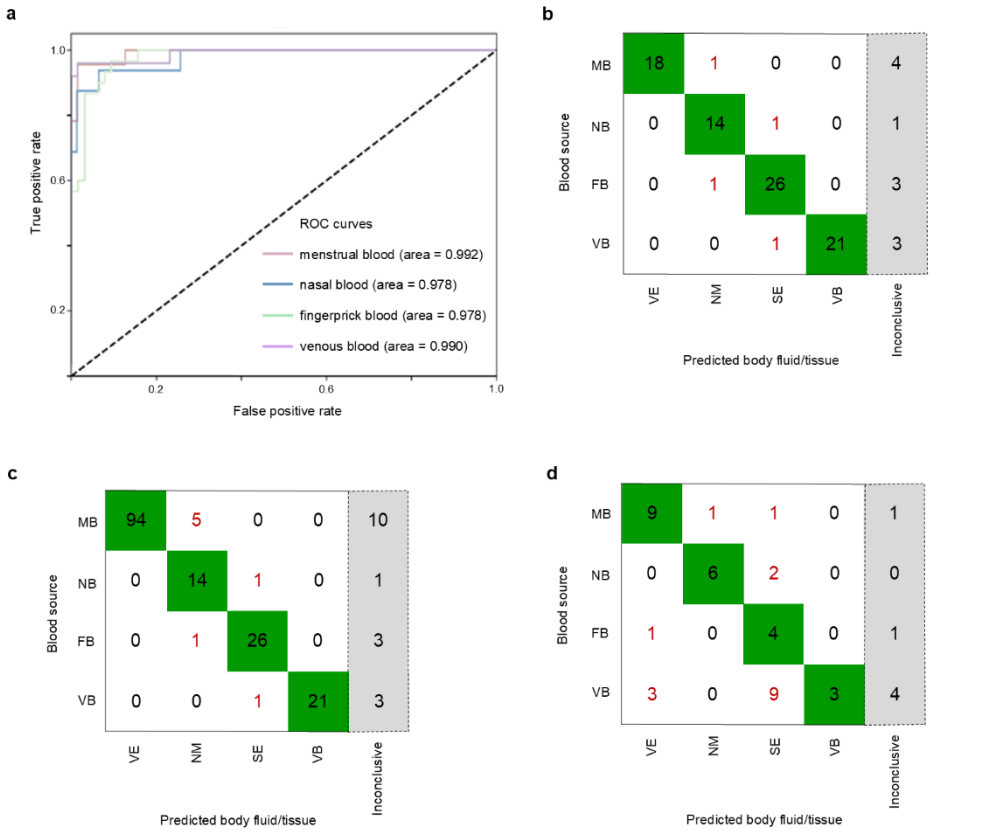


Figure 2. (a) Area under the receiver operating characteristic curve (AUC) of the 94 blood test samples analyzed with the deep neural network classifier. Confusion matrix of **(b)** the 94 blood test samples, **(c)** the 180 blood test samples including the 94 blood test samples together with 86 additional menstrual blood samples, and **(d)** 45 mock casework blood samples, all analyzed with the deep neural network classifier. "Inconclusive" represents those samples with probabilities of assignment below the 0.7 conclusive classification threshold at any of the four categories.

which increased their resemblance to the SE category (Figure 4b). Of the two misclassified FB samples, the NM-misclassified test sample (FB_D11) presented an increase in the relative abundances of *Actinobacteria* and *Firmicutes* at the phylum level (Figure 3c), corresponding to *Corynebacterium 1* and *Staphylococcus* at the genus level, respectively (Figure 4c), which resulted in its increased similarity to the composition of the NM category. The VE-misclassified mock casework FB sample (FB_F13) showed an increase in the relative abundance of *Firmicutes* at the phylum level (Figure 3c), corresponding to *Lactobacillus* at the genus level (Figure 4c), which resulted in its increased similarity to the composition of the VE category. A very distinctive profile was observed in the ten SE-misclassified VB samples, in which an increase in the relative abundances of *Actinobacteria*, *Bacteriodetes* and *Firmicutes* among other minor phyla was observed (Figure 3d). At the genus level an increase in the relative abundances of skin commensals, such as *Cuticubacterium*, *Corynebacterium 1* and *Staphylococcus*, among others was seen (Figure 4d). The three remaining VE-misclassified VB samples showed a slight increase of the VE-specific *Lactobacillus* genus (Figure 4d).

Overall, the eleven inconclusively classified MB test and mock casework samples showed a clear decrease in the relative abundance of *Firmicutes* and an increase of *Bacteriodetes* at the phylum level, corresponding to *Lactobacillus* and *Prevotella* genera, respectively. The inconclusively classified NB test sample showed a taxonomic profile in-between the samples correctly assigned to the NM and SE categories. The four inconclusively classified FB test and mock casework samples presented a taxonomic profile in-between the SE and NM categories. Finally, the taxonomic profiles of the seven inconclusively classified VB test and mock casework samples revealed the presence of typical SE commensals in the majority of them. Taxonomic profiles of all inconclusively classified test and mock casework blood samples at the phylum and genus level are available at Supplementary Figures S7 and S8, respectively.

Comparison with previous human mRNA classification

Out of the 21 aged mock casework samples, 19 had been previously analyzed for human mRNA markers for the purpose of forensic tissue identification as part of collaborative inter-laboratory exercises organized by EDNAP group (12, 27, 28) and the EUROFORGEN_NoE Consortium (4) (Supplementary Table S1), whereas two aged mock casework nasal blood samples were not included in the previously published reports. The overlapping 19 aged mock casework samples ranged from 1.1 years to 21 years old when analyzed with our DNN approach, while they had been less than 6 months old at the time of the previous human mRNA analysis, except for VB_B10, VB_B11 and MB_B14 that were 3, 11 and 5 years old, respectively. Despite the large sample storage time increase between the time of mRNA analysis and the time of our microbiome analysis, three out of six MB samples (MB_B12, MB_B14 and MB_B20) were classified to the VE category. The other three MB samples where either misclassified or inconclusively classified with our microbial approach. Based on previous human mRNA analysis, all these samples were classified as MB (12). Additionally, the analysis of MB_B20 using a MPS approach also showed some signals of venous blood. Regarding the 13 aged mock casework VB samples one (VB_B9) was correctly assigned to the VB category with our microbiome approach, whereas nine were misclassified to the SE category. The remaining three VB samples were inconclusively classified with our approach, with high probabilities assigned to the VB and SE categories (but still below the probability of conclusive assignment threshold of 0.7). In the previous human mRNA study using four of the aged mock casework VB samples, 8 to 15 laboratories of the 16 participators detected venous blood human mRNA markers (27). In a second previous mRNA study using six of the aged mock casework VB samples, 0 to 18 laboratories of the 18 participators detected venous blood human mRNA markers (28). In the previous MPS study using three of the aged mock casework VB samples, venous blood mRNA reads ranging from 92% to 99.5% of total reads were reported (4). (See Supplementary Table S1 for further sample's details). However, a direct comparison between the classification outcomes of the new microbiome approach and the human mRNA approach previously applied to the overlapping mock casework samples is not possible, because of the largely increased storage time of the samples when used for the two approaches. It may be expected that a similarly long sample storage time as applied here for the microbiome analysis may have caused issues for the mRNA approach due to RNA degradation.

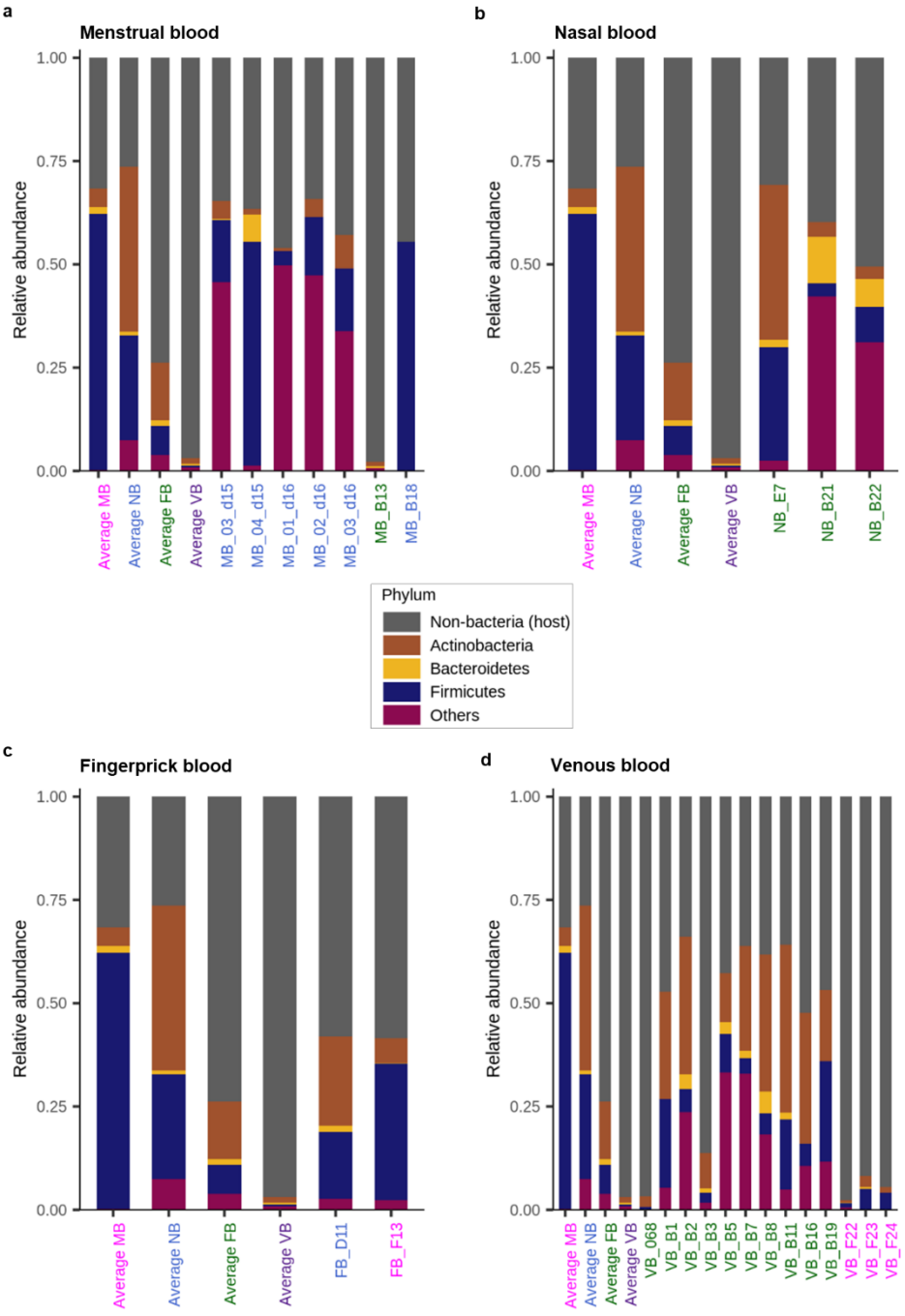


Figure 3. Taxonomic profiles at the phylum level of the DNN-misclassified blood test samples and mock casework blood samples. **(a)** menstrual blood, **(b)** nasal blood, **(c)** fingerprick blood, **(d)** venous blood samples. First four columns represent the average taxonomic profiles of the correctly classified samples. Sample IDs are colored according to the category they are assigned to: pink for vaginal epithelium, blue for nasal mucosa, green for skin epithelium, and purple for venous blood.

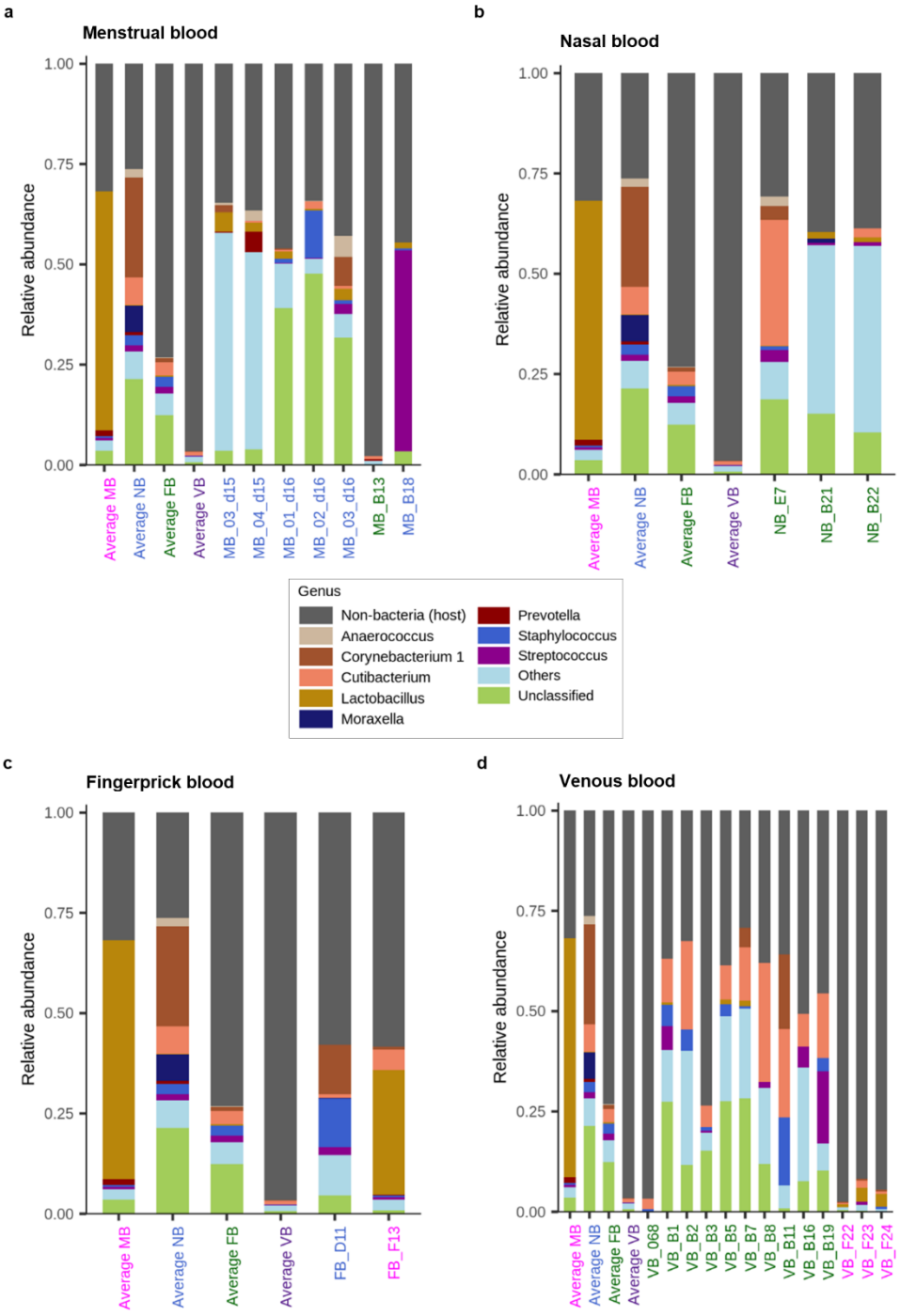


Figure 4. Taxonomic profiles at the genus level of the DNN-misclassified blood test samples and mock casework blood samples. **(a)** menstrual blood, **(b)** nasal blood, **(c)** fingerprick blood, **(d)** venous blood samples. First four columns

represent the average taxonomic profiles of the correctly classified samples. Sample IDs are colored according to the category they are assigned to: pink for vaginal epithelium, blue for nasal mucosa, green for skin epithelium, and purple for venous blood.

DISCUSSION

Confirmatory forensic tissue identification of human biological evidence is relevant for crime scene reconstruction and particularly important for differentiating and identifying forensic blood traces originating from different body sites. Here, we introduced and tested a taxonomy-independent deep neural network microbiome method to identify the body site of origin of different forensically relevant human blood traces, i.e. menstrual blood, nasal blood, fingerprick blood and venous blood. Given its underlying principle and training data used, our MPS-based DNN method relies on *a priori* knowledge that the sample analyzed contains any type of human blood. In forensic practice, knowledge of blood presence can be delivered by presumptive blood testing using commercial kits together with visual aids (dark/red color).

For building and training our DNN method, we employed a large 16S rRNA gene sequencing reference dataset publicly available from the HMP, which was comprised of microbiome data representative of three body sites from which forensic blood stains could originate i.e. vaginal epithelium (for menstrual blood), nasal mucosa (for nasal blood) and skin epithelium (for bloodstains originated with major skin contact). Due to the absence of suitable publicly available microbiome sequencing data from venous blood samples as required for DNN building and training purposes, we *de novo* generated 16S rRNA gene sequencing data from 150 venous blood samples. In line with previous studies (49, 50), we revealed by dedicated data analysis that the majority of the sequencing data we obtained from venous blood samples corresponded to non-specific, co-amplified human host DNA products. These findings were in agreement with the low bacterial DNA quantification values we obtained from the venous blood samples. Nevertheless, the number of sequencing reads obtained from microbiome sequencing of venous blood samples was similar to the expected average. At first glance, this appeared puzzling, but could finally be explained by assigning most of the venous blood sequencing data to non-specific, co-amplified human host sequences. The observed low abundance of bacterial taxa in venous blood was in agreement with previous reports on a healthy blood-associated microbiome (50-52). Sequencing data from MB, NB, and FB also contained human host DNA originating from their human blood component. However, in these non-venous blood samples, the human host DNA component was of much lower abundance, whereas the microbial DNA component was of much higher abundance. Given the high and highly specific bacterial load in MB, NB and FB samples we identified, our use of large HMP reference data from non-blood samples of the three relevant body sites VE, NM and SE, and our observation that VB contains very limited microbial DNA but instead large amounts of human sequence data used in the DNN, all together, allowed us to accurately classify MB, NB, FB (as proxy for blood with major skin contact) and VB (as proxy for blood without major skin contact) samples.

Because of our findings and under the assumption that the tested sample is from blood, we regard it as justified to conclude that a blood sample showing a higher than threshold probability for the VE category it most likely represents MB, for the NM category it most likely represents NB, for the SE category it most likely represents FB or other blood mixed with skin

cells e.g. caused by a bleeding skin injury, and for the VB category it most likely represents VB or any blood with none or minor skin contact.

Overall, test MB samples presented a taxonomic profile dominated by the *Lactobacillus* genus (Figure 4a), in agreement with previous vaginal microbiome (53, 54) and menstrual blood (24) analysis. We should highlight that accurate body site classification was observed independently of the day of the menses (ranging from one to ten days, depending on the donor) (Supplementary Figure S4). The stability of vaginal microbiome during menses has been previously reported (55), and it is especially noticeable in microbial communities dominated by the *Lactobacillus* genus (56), corresponding to the overall MB communities in the present study (Figure 4a). This is a clear advantage of our DNN microbiome approach over previously reported approaches, such as human mRNA-based ones, in which certain menstrual blood markers are not stable during menses (12). It must also be noted that misclassified and inconclusively classified MB test samples corresponded in their majority to samples of different menstruation days from the same donors. These samples presented a common decrease in the relative abundance of the *Lactobacillus* genus, replaced by other genera, such as *Staphylococcus* or *Prevotella* (Figure 4a and Supplementary Figure S8a), which could be an indication of healthy women with microbial communities other than the mostly represented in this study (53, 54). Given these taxonomic profiles, these women could also suffer from bacterial vaginosis (BV) (57), a condition with varying prevalence by ethnic group within countries (ranging from 0.4 to 60.8%, although BV tends to be of higher prevalence in sub-Saharan Africa and lower in Asia/Australia/western Europe, there are populations with high and low BV prevalence in all of these regions) (58), and in which half of the BV-positive individuals are asymptomatic (59).

The average taxonomic profiles observed in the correctly classified test NB samples (Figure 3b and 4b) agreed with previous nasal mucosa microbiome studies (60, 61). It has been reported that the relative proportions of NM commensals can vary substantially among individuals (60), which might be the underlying explanation for the misclassified and inconclusively classified NB test samples in this study. The time between nose-blowing and sample collection could also explain the similarity of these NB samples to the SE category. Nose-blowing pushes out the mucus dragging out some nasal bacteria, while leaving, for a time, other commensals in closer contact with the nasal epithelium more exposed for sample collection. Unknown nose-blowing activities prior to nasal blood donation can provide a source of skin-misclassification of nasal blood samples collected at crime scenes.

The taxonomic profiles observed in the correctly classified test FB samples (Figure 3c and 4c) agreed with previous reports (62, 63). The misclassification of one test FB sample (FB_D11) to the NM category is not surprising given the similar bacterial communities in the SE and NM. Regarding the VE-misclassified sample (FB_F13), the presence of the *Lactobacillus* genus has previously been reported in skin samples (62, 64), though not in such a high relative proportion. Its high abundance in the FB sample could be explained by an external source of contamination occurring before sample collection.

Regarding the mock casework blood samples, in the three VE-misclassified fresh VB mock samples (VB_F22-24) the *Lactobacillus* genus was observed in their taxonomic profiles (Figure 4d), though in low relative abundance, which might explain their assignation to the VE category. This genus has not been reported in previous blood microbiome studies (50-52, 65), and its presence due to external contamination during sample processing cannot be excluded.

The sample collection itself could also be the main factor explaining the majority of misclassified and inconclusively classified aged mock casework NB and VB samples. For example, skin commensals present in the skin areas surrounding the nasal cavities could be dragged into the NB samples by using a tissue paper to collect them. Additionally, 'contamination' of VB samples with skin can also occur via venipuncture for sample collection.

To our knowledge, this is the first study using a microbiome-based approach for the identification of the body site of origin of human blood samples, which based on the performance outcome presented here has high potentials for future casework applications provided further validation work. From the technical side, more samples should be used for training the DNN classifier, such as VE samples from donors suffering from BV or with microbial communities less frequent in the general population. This can ultimately account for intra- and inter- individual variation, increasing the accuracy of our approach when these type of samples are analyzed. Additionally, formal forensic developmental validation should be performed on the laboratory tool including sensitivity testing, concordance testing and including more mock casework samples mimicking realistic and often encountered forensic scenarios. It shall be emphasized again that the microbiome-based approach we introduced here for body site of origin identification of blood samples fully relies on the assumption that an analyzed sample is from or at least contains blood, an information that in practical forensic applications can be obtained from presumptive blood testing. In the future, we envision to combine our recently reported DNN microbiome approach for saliva, vaginal secretion and skin identification with our newly introduced DNN microbiome approach for menstrual blood, nasal blood, fingerprick blood and venous blood identification into a single taxonomy-independent DNN-based MPS microbiome approach for the identification of the body site or origin of forensically relevant human biomaterials.

ACKNOWLEDGEMENTS

The authors are grateful to all volunteers for their kind donation of biological samples, and to the HMP for making their data publicly available. The authors express their gratitude to Antoinette Hollestelle for sharing blood biobank samples and to Oscar Lao for valuable guidance in the deep learning implementation.

AUTHOR'S CONTRIBUTIONS

AV conceptualized. CDL, AV and MK designed this study. CDL performed the experiments and generated the sequencing data. CDL and DM prepared the sequencing data. DM designed and implemented the deep neural network approach and contributed to the corresponding section in the manuscript. CDL performed the taxonomic profiling. CH contributed to sample collection. CDL, AV and MK wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

The work of CDL, DMG, AV and MK is supported by the Erasmus MC, University Medical Center Rotterdam. AV is also supported by an EUR fellowship by the Erasmus University Rotterdam.

REFERENCES

1. Vennemann M, Scott G, Curran L, Bittner F, Tobe SS. 2014. Sensitivity and specificity of presumptive tests for blood, saliva and semen. Forensic Sci Med Pathol 10:69-75.
2. Lindenbergh A, Maaskant P, Sijen T. 2013. Implementation of RNA profiling in forensic casework. Forensic Sci Int Genet 7:159-66.
3. Hanson EK, Ballantyne J. 2013. Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis. F1000Res 2:281.
4. Ingold S, Dorum G, Hanson E, Berti A, Branicki W, Brito P, Elsmore P, Gettings KB, Giangasparo F, Gross TE, Hansen S, Hanssen EN, Kampmann ML, Kayser M, Laurent FX, Morling N, Mosquera-Miguel A, Parson W, Phillips C, Porto MJ, Pospiech E, Roeder AD, Schneider PM, Schulze Johann K, Steffen CR, Syndercombe-Court D, Trautmann M, van den Berge M, van der Gaag KJ, Vannier J, Verdoliva V, Vidaki A, Xavier C, Ballantyne J, Haas C. 2018. Body fluid identification using a targeted mRNA massively parallel sequencing approach - results of a EUROFORGEN/EDNAP collaborative exercise. Forensic Sci Int Genet 34:105-115.
5. Xu Y, Xie J, Cao Y, Zhou H, Ping Y, Chen L, Gu L, Hu W, Bi G, Ge J, Chen X, Zhao Z. 2014. Development of Highly Sensitive and Specific mRNA Multiplex System (XCYR1) for Forensic Human Body Fluids and Tissues Identification. PLoS One 9:e100123.
6. Sauer E, Reinke AK, Courts C. 2016. Differentiation of five body fluids from forensic samples by expression analysis of four microRNAs using quantitative PCR. Forensic Sci Int Genet 22:89-99.
7. Li ZL, Bai P, Peng D, Long B, Zhang L, Liang WB. 2015. Screening and confirmation of microRNA markers for distinguishing between menstrual and peripheral blood. Forensic Sci Int Genet 5:E353-E355.
8. Sirker M, Fimmers R, Schneider PM, Gomes I. 2017. Evaluating the forensic application of 19 target microRNAs as biomarkers in body fluid and tissue identification. Forensic Sci Int Genet 27:41-49.
9. Song F, Luo H, Xie M, Zhu H, Hou Y. 2017. Microarray expression profile of circular RNAs in human body fluids. Forensic Sci Int Genet 6:e55-e56.
10. Zhang Y, Liu B, Shao C, Xu H, Xue A, Zhao Z, Shen Y, Tang Q, Xie J. 2017. Evaluation of the inclusion of circular RNAs in mRNA profiling in forensic body fluid identification. Int J Legal Med:43-52.
11. Wang S, Wang Z, Tao R, He G, Liu J, Li C, Hou Y. 2019. The potential use of Piwi-interacting RNA biomarkers in forensic body fluid identification: A proof-of-principle study. Forensic Sci Int Genet 39:129-135.
12. Haas C, Hanson E, Anjos MJ, Ballantyne KN, Banemann R, Bhoelai B, Borges E, Carvalho M, Courts C, De Cock G, Drobnic K, Dotsch M, Fleming R, Franchi C, Gomes I, Hadzic G, Harbison SA, Harteveld J, Hjort B, Hollard C, Hoff-Olsen P, Huls C, Keyser C, Maronas O, McCallum N, Moore D, Morling N, Niederstatter H, Noel F, Parson W, Phillips C, Popielarz C, Roeder AD, Salvaderi L, Sauer E, Schneider PM, Shanthan G, Court DS, Turanska M, van Oorschot RA, Vennemann M, Vidaki A, Zatkalikova L, Ballantyne J. 2014. RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: results of a fourth and fifth collaborative EDNAP exercise. Forensic Sci Int Genet 8:203-12.
13. Li Z, Peng D, Tian H, Wang J, Xia Y, Zhang L, Bai P. 2017. Expression changes of microRNAs in menstrual blood samples of different menstrual cycle collection days. Forensic Sci Int Genet 6:e159-e161.
14. Sakurada K, Akutsu T, Watanabe K, Yoshino M. 2012. Identification of nasal blood by real-time RT-PCR. Leg Med 14:201-204.
15. van den Berge M, Bhoelai B, Harteveld J, Matai A, Sijen T. 2016. Advancing forensic RNA typing: On non-target secretions, a nasal mucosa marker, a differential co-extraction protocol and the sensitivity of DNA and RNA profiling. Forensic Sci Int Genet 20:119-129.
16. Holtkotter H, Beyer V, Schwender K, Glaub A, Johann KS, Schrenkamp M, Sibbing U, Banken S, Wiegand P, Pfeiffer H, Dennany L, Vennemann M, Vennemann M, Consortium E-N. 2017. Independent validation of body fluid-specific CpG markers and construction of a robust multiplex assay. Forensic Sci Int Genet 29:261-268.
17. Lee HY, Park MJ, Choi A, An JH, Yang WI, Shin KJ. 2012. Potential forensic application of DNA methylation profiling to body fluid identification. Int J Legal Med 126:55-62.
18. Holtkötter H, Schwender K, Wiegand P, Pfeiffer H, Vennemann M. 2018. Marker evaluation for differentiation of blood and menstrual fluid by methylation-sensitive SNaPshot analysis. Int J Legal Med 132:387-395.
19. An JH, Choi A, Shin KJ, Yang WI, Lee HY. 2013. DNA methylation-specific multiplex assays for body fluid identification. Int J Legal Med 127:35-43.
20. Lee HY, An JH, Jung S-E, Oh YN, Lee EY, Choi A, Yang WI, Shin K-J. 2015. Genome-wide methylation profiling and a multiplex construction for the identification of body fluids using epigenetic markers. Forensic Sci Int Genet 17:17-24.

21. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694-1697.
22. Choi A, Shin KJ, Yang WI, Lee HY. 2014. Body fluid identification by integrated analysis of DNA methylation and body fluid-specific microbial DNA. Int J Legal Med 128:33-41.
23. Benschop CCG, Quaak FCA, Boon ME, Sijen T, Kuiper I. 2012. Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? Int J Legal Med 126:303-310.
24. Dobay A, Haas C, Fucile G, Downey N, Morrison HG, Kratzer A, Arora N. 2019. Microbiome-based body fluid identification of samples exposed to indoor conditions. Forensic Sci Int Genet 40:105-113.
25. Díez López C, Vidaki A, Ralf A, Montiel González D, Radjabzadeh D, Kraaij R, Uitterlinden AG, Haas C, Lao O, Kayser M. 2019. Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. Forensic Sci Int Genet 41:72-82.
26. Human Microbiome Project Data Analysis and Coordination Center. 2019. Accessed 4, February 2019. https://www.hmpdacc.org/hmp/(2017).
27. Haas C, Hanson E, Bar W, Banemann R, Bento AM, Berti A, Borges E, Bouakaze C, Carracedo A, Carvalho M, Choma A, Dotsch M, Duriancikova M, Hoff-Olsen P, Hohoff C, Johansen P, Lindenbergh PA, Loddenkotter B, Ludes B, Maronas O, Morling N, Niederstatter H, Parson W, Patel G, Popielarz C, Salata E, Schneider PM, Sijen T, Sviezena B, Zatkalikova L, Ballantyne J. 2011. mRNA profiling for the identification of blood--results of a collaborative EDNAP exercise. Forensic Sci Int Genet 5:21-6.
28. Haas C, Hanson E, Anjos MJ, Bar W, Banemann R, Berti A, Borges E, Bouakaze C, Carracedo A, Carvalho M, Castella V, Choma A, De Cock G, Dotsch M, Hoff-Olsen P, Johansen P, Kohlmeier F, Lindenbergh PA, Ludes B, Maronas O, Moore D, Morerod ML, Morling N, Niederstatter H, Noel F, Parson W, Patel G, Popielarz C, Salata E, Schneider PM, Sijen T, Sviezena B, Turanska M, Zatkalikova L, Ballantyne J. 2012. RNA/DNA co-analysis from blood stains--results of a second collaborative EDNAP exercise. Forensic Sci Int Genet 6:70-80.
29. Castelino M, Eyre S, Moat J, Fox G, Martin P, Ho P, Upton M, Barton A. 2017. Optimisation of methods for bacterial skin microbiome investigation: primer selection and comparison of the 454 versus MiSeq platform. BMC Microbiol 17:23.
30. Manual of Procedures for Human Microbiome Project. 2010. Accessed 4, February 2019. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd003190.2.
31. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. 16S 454 Sequencing Protocol HMP Consortium. 2010.
32. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
33. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997.
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.
35. Santhakumaran TJaA. 2011. Statistical Normalization and Back Propagation for Classification. Int J Comput Theory Eng 3:1793-8201.
36. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825-2830.
37. Chen H-Y. 2016. TensorFlow–A system for large-scale machine learning. Proc. USENIX Symp. Oper. Syst. Des. Implement. OSDI (OSDI): 265-283
38. Chollet F. 2015. Keras. GitHub. Acessed 4, February 2019.https://github.com/fchollet/keras.
39. Karlik B, Olgac AV. 2011. Performance analysis of various activation functions in generalized MLP architectures of neural networks. Int J Artif Intell and Expert Syst 1:111-122.
40. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In ICLR.
41. Goodfellow I, Bengio Y, Courville A. 2016. Deep learning. MIT Press.
42. Rich C, Steve L, Lee G. 2000. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. Adv Neural Inf Proc Syst:402-408.
43. Sokolova M, Lapalme G. 2009. A systematic analysis of performance measures for classification tasks. Inform Process Manage 45:427-437.
44. Hadley W. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer Publishing Company.
45. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581-583.
46. Murali A, Bhargava A, Wright ES. 2018. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. Microbiome 6:140.

47. Wright ES. 2016. Using DECIPHER v2. 0 to analyze big biological sequence data in R. R J 8.
48. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-D596.
49. Kommedal O, Simmon K, Karaca D, Langeland N, Wiker HG. 2012. Dual priming oligonucleotides for broad-range amplification of the bacterial 16S rRNA gene directly from human clinical specimens. J Clin Microbiol 50:1289-1294.
50. Olde Loohuis LM, Mangul S, Ori APS, Jospin G, Koslicki D, Yang HT, Wu T, Boks MP, Lomen-Hoerth C, Wiedau-Pazos M, Cantor RM, de Vos WM, Kahn RS, Eskin E, Ophoff RA. 2018. Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. Transl Psychiatry 8:96.
51. Paisse S, Valle C, Servant F, Courtney M, Burcelin R, Amar J, Lelouvier B. 2016. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. Transfusion 56:1138-1147.
52. Whittle E, Leonard MO, Harrison RD, Gant TW, Tonge DP. 2018. Multi-method characterisation of the human circulating microbiome. Front Microbiol 9:3266.
53. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SSK, McCulle SL. 2011. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108:4680-4687.
54. Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, Foster JA, Forney LJ. 2007. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME J 1:121-133.
55. Chaban B, Links MG, Jayaprakash TP, Wagner EC, Bourque DK, Lohn Z, Albert AYK, van Schalkwyk J, Reid G, Hemmingsen SM, Hill JE, Money DM. 2014. Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. Microbiome 2:23.
56. Gajer P, Brotman RM, Bai GY, Sakamoto J, Schutte UME, Zhong X, Koenig SSK, Fu L, Ma ZS, Zhou X, Abdo Z, Forney LJ, Ravel J. 2012. Temporal dynamics of the human vaginal microbiota. Sci Transl Med 4:132ra52.
57. Chen H-M, Chang T-H, Lin F-M, Liang C, Chiu C-M, Yang T-L, Yang T, Huang C-Y, Cheng Y-N, Chang Y-A, Chang P-Y, Weng S-L. 2018. Vaginal microbiome variances in sample groups categorized by clinical criteria of bacterial vaginosis. BMC Genomics 19:876.
58. Kenyon C, Colebunders R, Crucitti T. 2013. The global epidemiology of bacterial vaginosis: a systematic review. Am J Obstet Gynecol 209:505-23.
59. Srinivasan S, Fredricks DN. 2008. The human vaginal bacterial biota and bacterial vaginosis. Interdiscip Perspect Infect Dis 2008:750479.
60. Biswas K, Hoggard M, Jain R, Taylor MW, Douglas RG. 2015. The nasal microbiota in health and disease: variation within and between subjects. Front Microbiol 6:134.
61. Camarinha-Silva A, Jáuregui R, Chaves-Moreno D, Oxley APA, Schaumburg F, Becker K, Wos-Oxley ML, Pieper DH. 2014. Comparing the anterior nare bacterial community of two discrete human populations using Illumina amplicon sequencing. Environ Microbiol 16:2939-2952.
62. Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, Berg-Lyon D, Ackermann G, Moeller Christensen GJ, Nakatsuji T, Zhang L, Borkowski AW, Meehan MJ, Dorrestein K, Gallo RL, Bandeira N, Knight R, Alexandrov T, Dorrestein PC. 2015. Molecular cartography of the human skin surface in 3D. Proc Natl Acad Sci U S A 112:E2120-9.
63. Sanford JA, Gallo RL. 2013. Functions of the skin microbiota in health and disease. Semin Immunol 25:370-7.
64. Staudinger T, Pipal A, Redl B. 2011. Molecular analysis of the prevalent microbiota of human male and female forehead skin compared to forearm skin and the influence of make-up. J Appl Microbiol 110:1381-9.
65. Castillo DJ, Rifkin RF, Cowan DA, Potgieter M. 2019. The healthy human blood microbiome: fact or fiction? Front Cell Infect Microbiol 9:9.

SUPPLEMENTARY MATERIALS

Supplementary Table S1. Detailed information of mock casework samples: sample ID, description, age and EDANP/ EUROFORGEN exercise.

Freshly-prepared mock casework blood samples			
Sample ID	**Description**	**Age**	**EDNAP/EUROFORGEN exercise (reference)**
MB_F1	Menstrual blood swab at room temperature	48 h	NA
MB_F2	Menstrual blood swab at 4 °C	48 h	
MB_F3	Menstrual blood swab at 37 °C	48 h	
MB_F4	Menstrual blood swab at room temperature	48 h	
MB_F5	Menstrual blood swab at 4 °C	48 h	
MB_F6	Menstrual blood swab at 37 °C	48 h	
VB_F7	50 µl of venous blood on fabric at room temperature	48 h	
VB_F8	50 µl of venous blood on fabric at 4 °C	48 h	
FB_V9	50 µl of venous blood on fabric at 37 °C	48 h	
FB_F10	Fingerprick blood swab at room temperature	48 h	
FB_F11	Fingerprick blood swab at 4 °C	48 h	
FB_F12	Fingerprick blood swab at 37 °C	48 h	
FB_F13	Fingerprick blood on glass at room temperature	48 h	
FB_F14	Fingerprick blood on glass at 4 °C	48 h	
FB_F15	Fingerprick blood on glass at 37 °C	48 h	
NB_F16	Nasal blood on tissue at room temperature	48 h	
NB_F17	Nasal blood on tissue at 4 °C	48 h	
NB_F18	Nasal blood on tissue at 37 °C	48 h	
NB_F19	Nasal blood on tissue at room temperature	48 h	
NB_F20	Nasal blood on tissue at 4 °C	48 h	
NB_F21	Nasal blood on tissue at 37 °C	48 h	
VB_F22	100 µl of venous blood on fabric at room temperature	48 h	
VB_F23	100 µl of venous blood on fabric at 4 °C	48 h	
VB_F24	100 µl of venous blood on fabric at 37 °C	48 h	

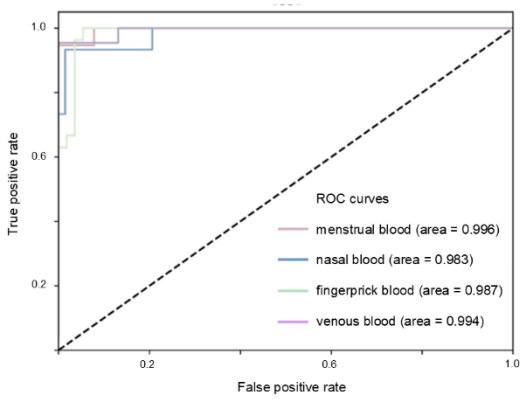
Aged mock casework blood samples			
Sample ID	**Description**	**Age**	**EDNAP/EUROFORGEN exercise (reference)**
NB_B21	Nasal blood in tissue (1x1 cm)	1.1 yr	NA
VB_B19	50 µl of venous blood in swab	1.8 yr	NA
NB_B22	Nasal blood in tissue (1x1 cm)	1.9 yr	NA
MB_B20	¼ menstrual blood swab	2 yr	(4)
VB_B16	50 µl of venous blood in swab	2.9 yr	(4)
VB_B17	50 µl of venous blood in cellulose swab	2.9 yr	(4)
MB_B18	¼ menstrual blood swab	2.9 yr	(4)
MB_B15	¼ menstrual blood swab	7.2 yr	(12)
MB_B13	¼ menstrual blood swab	7.3 yr	(12)
MB_B12	Menstrual blood on sanitary towel	8.4 yr	(12)
VB_B9	Non-EDTA venous blood on t-shirt (2 mm circle)	10 yr	(28)
VB_B8	10 µl EDTA-venous blood on swab	10 yr	(28)
VB_B7	10 µl EDTA-venous blood on a bandage	10 yr	(28)
VB_B6	Non-EDTA venous blood on t-shirt (5 mm circle)	10 yr	(28)
VB_B5	10 µl EDTA-venous blood on swab	10 yr	(27)
VB_B3	10 µl EDTA-venous blood on swab	11 yr	(27)
VB_B2	10 µl EDTA-venous blood on swab	11 yr	(27)
VB_B1	10 µl EDTA-venous blood on swab	11 yr	(27)
MB_B14	Menstrual blood on sanitary towel	12 yr	(12)
VB_B10	10 µl non-treated venous blood on swab	13 yr	(28)
VB_B11	EDTA-blood on cotton cloth	21 yr	(28)

NA (not applicable).

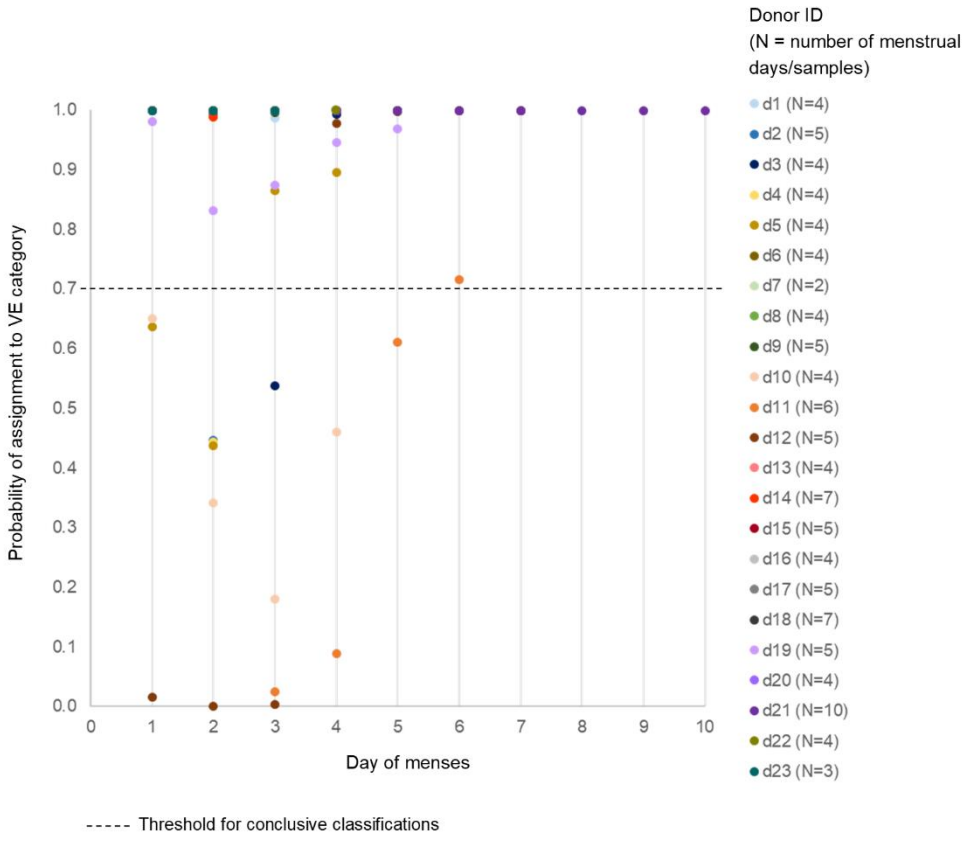
Supplementary Table S2. DNN classification performance metrics of the 83 blood test samples – excluding the inconclusively classified samples with probabilities of assignment below the 0.7 conclusive classification threshold at any of the four categories.

	Menstrual blood (MB)	Nasal blood (NB)	Fingerprick blood (FB)	Venous blood (VB)
AUC	0.996	0.983	0.987	0.994
Precision	1.000	0.875	0.929	1.000
Recall	0.947	0.933	0.963	0.955
F1 score	0.973	0.903	0.946	0.977
NPV	0.985	0.985	0.982	0.984
Support	19	15	27	22
MCC	0.935			

AUC=area under the operating characteristic curve; NVP=negative predictive value; MCC=Mathews correlation coefficient.



Supplementary Figure S3. Area under the receiver operating characteristic curve (AUC) of the 83 blood test samples – excluding the inconclusively classified samples with probabilities of assignment below the 0.7 conclusive classification threshold at any of the four categories.



Supplementary Figure S4. Overview of probabilities of assignment to vaginal epithelium (VE) category of the 109 menstrual blood test samples from different menses days from the participating female donors.

Supplementary Table S5. Body site of origin assignment probabilities, i.e. vaginal epithelium, nasal mucosa, skin epithelium and venous blood, of the 180 human blood test samples resulted from the DNN approach. Green indicates correctly assigned samples (threshold ≥0.7); yellow indicates inconclusively classified samples (<0.7 in any category); red indicates misclassified samples (≥0.7 in any category other than the true body site of origin). Asterisk (*) indicates menstrual blood samples randomly selected for classification performance metrics calculations.

Sample ID	Blood type	Venous blood	Vaginal epithelium	Nasal mucosa	Skin epithelium
FB_D01	fingerprick	0.0008	0.0002	0.0007	0.9983
FB_D02	fingerprick	0.0029	0.0003	0.0015	0.9953
FB_D05	fingerprick	0.0014	0.0002	0.0009	0.9974
FB_D07	fingerprick	0.0010	0.0002	0.0010	0.9979
FB_D08	fingerprick	0.0008	0.0002	0.0008	0.9982
FB_D09	fingerprick	0.0042	0.0002	0.0060	0.9896
FB_D10	fingerprick	0.0016	0.0002	0.0014	0.9967
FB_D11	fingerprick	0.0264	0.0006	0.9207	0.0523
FB_D12	fingerprick	0.0443	0.0009	0.2328	0.7220
FB_D13	fingerprick	0.0010	0.0002	0.0008	0.9980
FB_D14	fingerprick	0.0511	0.0009	0.5451	0.4029
FB_D15	fingerprick	0.0006	0.0002	0.0004	0.9988
FB_D16	fingerprick	0.0014	0.0002	0.0013	0.9970
FB_D17	fingerprick	0.0218	0.0004	0.0774	0.9004
FB_D18	fingerprick	0.0011	0.0002	0.0012	0.9976
FB_D19	fingerprick	0.0771	0.0054	0.5453	0.3721
FB_E5	fingerprick	0.0088	0.0003	0.0075	0.9834
FB_E6	fingerprick	0.0786	0.0160	0.6595	0.2460
FB_E7	fingerprick	0.0549	0.0011	0.1886	0.7555
FB_ZH01	fingerprick	0.1640	0.0020	0.0206	0.8133
FB_ZH02	fingerprick	0.0005	0.0002	0.0004	0.9989
FB_ZH04	fingerprick	0.1427	0.0010	0.0300	0.8263
FB_ZH05	fingerprick	0.0009	0.0002	0.0011	0.9978
FB_ZH07	fingerprick	0.0006	0.0002	0.0005	0.9987
FB_ZH08	fingerprick	0.0010	0.0002	0.0010	0.9978
FB_ZH10	fingerprick	0.0006	0.0002	0.0004	0.9988
FB_ZH16	fingerprick	0.0727	0.0006	0.0276	0.8991
FB_ZH19	fingerprick	0.0905	0.0008	0.0836	0.8251
FB_ZH21	fingerprick	0.0325	0.0007	0.0173	0.9494
FB_ZH25	fingerprick	0.0023	0.0002	0.0035	0.9940
MB_01_D01	menstrual	0.0003	0.9992	0.0003	0.0002
MB_01_D02	menstrual	0.0003	0.9992	0.0003	0.0002
MB_01_D03	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_D04	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_D05	menstrual	0.1929	0.6369	0.0007	0.1696
MB_01_ZH01	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_ZH02	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_ZH04	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_ZH05	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_ZH07	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_ZH10	menstrual	0.0082	0.9800	0.0017	0.0101
MB_01_ZH16	menstrual	0.0003	0.9992	0.0003	0.0002
MB_01_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_01_ZH21	menstrual	0.0003	0.9992	0.0003	0.0002
*MB_01_ZH25	menstrual	0.0002	0.9993	0.0003	0.0002
*MB_02_D01	menstrual	0.0005	0.9988	0.0004	0.0003
*MB_02_D02	menstrual	0.1529	0.4461	0.0013	0.3997
*MB_02_D03	menstrual	0.0020	0.9961	0.0010	0.0008
*MB_02_D04	menstrual	0.1265	0.4431	0.2142	0.2162
*MB_02_D05	menstrual	0.2100	0.4372	0.0008	0.3520
*MB_02_ZH01	menstrual	0.0002	0.9993	0.0003	0.0002
MB_02_ZH02	menstrual	0.0029	0.9887	0.0081	0.0003
*MB_02_ZH04	menstrual	0.0003	0.9991	0.0003	0.0002

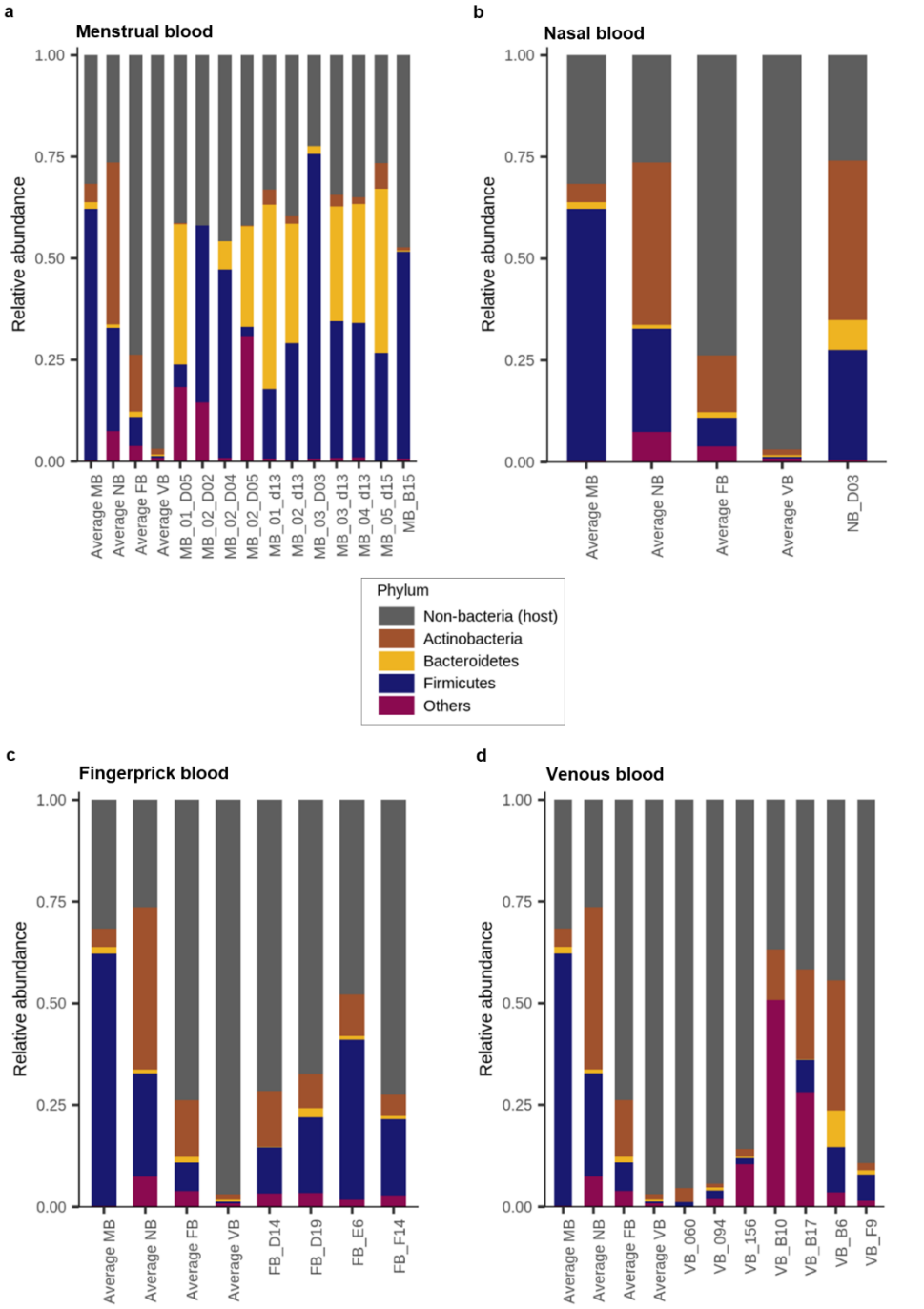
*MB_02_ZH05	menstrual	0.0003	0.9992	0.0003	0.0002
*MB_02_ZH07	menstrual	0.0003	0.9992	0.0003	0.0002
MB_02_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
*MB_02_ZH10	menstrual	0.0293	0.8311	0.1296	0.0100
*MB_02_ZH16	menstrual	0.0003	0.9992	0.0003	0.0002
MB_02_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
*MB_02_ZH21	menstrual	0.0003	0.9991	0.0003	0.0002
MB_02_ZH25	menstrual	0.0002	0.9993	0.0003	0.0002
MB_03_D01	menstrual	0.0101	0.9868	0.0010	0.0020
MB_03_D02	menstrual	0.0004	0.9990	0.0003	0.0003
MB_03_D03	menstrual	0.0722	0.5371	0.3767	0.0139
MB_03_D04	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_D05	menstrual	0.0620	0.8646	0.0007	0.0727
MB_03_ZH01	menstrual	0.0002	0.9993	0.0003	0.0002
*MB_03_ZH02	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_ZH04	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_ZH05	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_ZH07	menstrual	0.0003	0.9992	0.0003	0.0002
*MB_03_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
MB_03_ZH10	menstrual	0.0622	0.8739	0.0345	0.0295
MB_03_ZH16	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_ZH21	menstrual	0.0003	0.9992	0.0003	0.0002
MB_03_ZH25	menstrual	0.0002	0.9992	0.0003	0.0002
MB_04_D01	menstrual	0.0006	0.9987	0.0003	0.0004
MB_04_D02	menstrual	0.0006	0.9987	0.0004	0.0003
MB_04_D03	menstrual	0.0046	0.9925	0.0009	0.0020
MB_04_D04	menstrual	0.0003	0.9992	0.0003	0.0002
MB_04_D05	menstrual	0.0500	0.8956	0.0007	0.0537
MB_04_ZH01	menstrual	0.0002	0.9993	0.0003	0.0002
MB_04_ZH02	menstrual	0.0003	0.9992	0.0003	0.0002
MB_04_ZH04	menstrual	0.0003	0.9992	0.0003	0.0002
MB_04_ZH05	menstrual	0.0003	0.9992	0.0003	0.0002
MB_04_ZH07	menstrual	0.0002	0.9992	0.0003	0.0002
MB_04_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
MB_04_ZH10	menstrual	0.0240	0.9448	0.0193	0.0119
MB_04_ZH16	menstrual	0.0003	0.9992	0.0003	0.0002
*MB_04_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_04_ZH21	menstrual	0.0003	0.9992	0.0003	0.0002
MB_05_D02	menstrual	0.0007	0.9986	0.0003	0.0004
MB_05_ZH02	menstrual	0.0002	0.9993	0.0003	0.0002
MB_05_ZH04	menstrual	0.0003	0.9991	0.0003	0.0002
MB_05_ZH07	menstrual	0.0003	0.9992	0.0003	0.0002
MB_05_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
MB_05_ZH10	menstrual	0.0151	0.9688	0.0069	0.0092
MB_05_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_06_ZH02	menstrual	0.0002	0.9993	0.0003	0.0002
MB_06_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
MB_06_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_07_ZH02	menstrual	0.0003	0.9992	0.0003	0.0002
MB_07_ZH08	menstrual	0.0002	0.9993	0.0003	0.0002
MB_07_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_08_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_09_d5	menstrual	0.0004	0.9990	0.0003	0.0003
MB_09_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_10_ZH19	menstrual	0.0003	0.9992	0.0003	0.0002
MB_01_d13	menstrual	0.2946	0.6504	0.0258	0.0293
MB_01_d15	menstrual	0.0002	0.9993	0.0003	0.0002
MB_01_d16	menstrual	0.0334	0.0154	0.9417	0.0094
*MB_02_d13	menstrual	0.3590	0.3412	0.2653	0.0345
*MB_02_d15	menstrual	0.0002	0.9993	0.0003	0.0002
*MB_02_d16	menstrual	0.0146	0.0012	0.9815	0.0027

MB_03_d13	menstrual	0.2671	0.1806	0.5206	0.0317
MB_03_d15	menstrual	0.0511	0.0250	0.9221	0.0018
MB_03_d16	menstrual	0.0218	0.0038	0.9712	0.0032
MB_04_d13	menstrual	0.2996	0.4606	0.2093	0.0305
MB_04_d15	menstrual	0.0587	0.0894	0.8483	0.0035
MB_04_d16	menstrual	0.0170	0.9773	0.0022	0.0034
MB_05_d15	menstrual	0.3112	0.6101	0.0239	0.0549
MB_05_d16	menstrual	0.0013	0.9975	0.0004	0.0008
MB_06_d1	menstrual	0.0003	0.9992	0.0003	0.0002
MB_06_d15	menstrual	0.1788	0.7151	0.0018	0.1043
*MB_06_d2	menstrual	0.0075	0.9890	0.0004	0.0031
MB_06_d3	menstrual	0.0002	0.9993	0.0003	0.0002
MB_06_d4	menstrual	0.0002	0.9993	0.0003	0.0002
*MB_07_d1	menstrual	0.0003	0.9992	0.0003	0.0002
MB_07_d2	menstrual	0.0003	0.9992	0.0003	0.0002
MB_08_d1	menstrual	0.0002	0.9993	0.0003	0.0002
*MB_08_d2	menstrual	0.0002	0.9993	0.0003	0.0002
MB_08_d3	menstrual	0.0002	0.9993	0.0003	0.0002
MB_08_d4	menstrual	0.0002	0.9993	0.0003	0.0002
MB_09_d1	menstrual	0.0010	0.9981	0.0003	0.0006
*MB_09_d2	menstrual	0.0002	0.9993	0.0003	0.0002
MB_09_d3	menstrual	0.0027	0.9956	0.0003	0.0014
MB_09_d4	menstrual	0.0003	0.9992	0.0003	0.0002
NB_D01	nasal	0.0108	0.0003	0.9783	0.0106
NB_D03	nasal	0.0416	0.0011	0.5262	0.4311
NB_D07	nasal	0.0002	0.0003	0.9993	0.0002
NB_D08	nasal	0.0032	0.0003	0.9933	0.0031
NB_D10	nasal	0.0002	0.0003	0.9993	0.0002
NB_D12	nasal	0.0061	0.0003	0.9881	0.0055
NB_D13	nasal	0.0173	0.0006	0.9351	0.0471
NB_D14	nasal	0.0004	0.0003	0.9987	0.0006
NB_D15	nasal	0.0003	0.0003	0.9991	0.0003
NB_D16	nasal	0.0002	0.0003	0.9992	0.0003
NB_D19	nasal	0.0012	0.0003	0.9971	0.0014
NB_D21	nasal	0.0071	0.0003	0.9869	0.0057
NB_D23	nasal	0.0003	0.0003	0.9990	0.0004
NB_E5	nasal	0.0142	0.0004	0.9706	0.0148
NB_E6	nasal	0.0002	0.0003	0.9993	0.0002
NB_E7	nasal	0.0044	0.0002	0.0072	0.9881
VB_047	venous	0.9982	0.0006	0.0003	0.0010
VB_054	venous	0.7548	0.0010	0.0978	0.1464
VB_060	venous	0.3009	0.0154	0.0009	0.6828
VB_068	venous	0.0290	0.0040	0.0006	0.9664
VB_075	venous	0.9769	0.0032	0.0005	0.0194
VB_082	venous	0.9783	0.0031	0.0004	0.0181
VB_094	venous	0.3839	0.4810	0.0009	0.1342
VB_107	venous	0.9982	0.0005	0.0004	0.0009
VB_113	venous	0.9952	0.0006	0.0012	0.0031
VB_125	venous	0.9739	0.0006	0.0206	0.0048
VB_135	venous	0.9990	0.0004	0.0003	0.0003
VB_142	venous	0.9989	0.0004	0.0003	0.0004
VB_150	venous	0.9988	0.0004	0.0003	0.0005
VB_151	venous	0.9707	0.0021	0.0005	0.0266
VB_153	venous	0.9987	0.0003	0.0004	0.0006
VB_154	venous	0.9988	0.0004	0.0004	0.0004
VB_155	venous	0.9950	0.0006	0.0006	0.0038
VB_156	venous	0.4696	0.0010	0.2897	0.2398
VB_157	venous	0.9970	0.0006	0.0004	0.0020
VB_158	venous	0.9990	0.0004	0.0003	0.0003
VB_159	venous	0.9983	0.0007	0.0006	0.0003
VB_160	venous	0.9977	0.0005	0.0004	0.0014
VB_161	venous	0.9983	0.0004	0.0003	0.0009

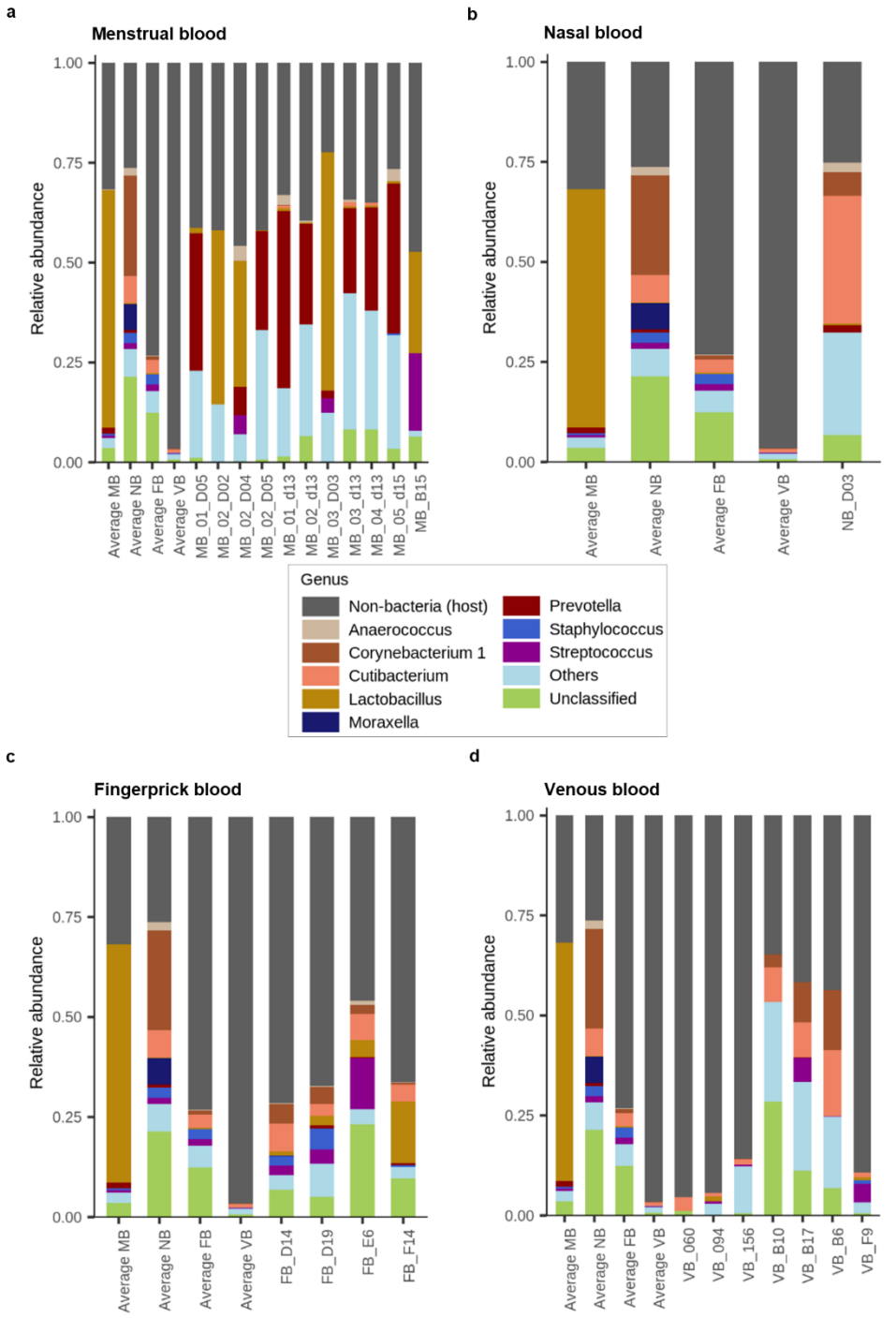
VB_162	venous	0.9991	0.0003	0.0003	0.0003
VB_163	venous	0.9974	0.0004	0.0005	0.0017

Supplementary Table S6. Body site of origin assignment probabilities, i.e. vaginal epithelium, nasal mucosa, skin epithelium and venous blood, of the 45 human blood mock casework samples resulted from the DNN approach. Green indicates correctly assigned samples (threshold ≥0.7); yellow indicates inconclusively classified samples (<0.7 in any category); red indicates misclassified samples (≥0.7 in any category other than the true body site of origin).

Sample ID	Mock type	Blood type	Venous blood	Vaginal epithelium	Nasal mucosa	Skin epithelium
VB_B1	aged	venous	0.0679	0.0024	0.0059	0.9238
VB_B10	aged	venous	0.5802	0.0024	0.0075	0.4099
VB_B11	aged	venous	0.0017	0.0002	0.0016	0.9964
VB_B16	aged	venous	0.2195	0.0038	0.0088	0.7680
VB_B17	aged	venous	0.2219	0.0009	0.0934	0.6838
VB_B19	aged	venous	0.0181	0.0004	0.0141	0.9674
VB_B2	aged	venous	0.0010	0.0003	0.0005	0.9983
VB_B3	aged	venous	0.0184	0.0016	0.0009	0.9790
VB_B5	aged	venous	0.0402	0.0010	0.0212	0.9376
VB_B6	aged	venous	0.1643	0.0005	0.1461	0.6891
VB_B7	aged	venous	0.0725	0.0030	0.0019	0.9227
VB_B8	aged	venous	0.0098	0.0005	0.0042	0.9856
VB_B9	aged	venous	0.7192	0.0016	0.0526	0.2265
VB_F22	fresh	venous	0.0426	0.9497	0.0018	0.0060
VB_F23	fresh	venous	0.0599	0.9300	0.0027	0.0073
VB_F24	fresh	venous	0.2194	0.7658	0.0097	0.0051
VB_F7	fresh	venous	0.9682	0.0019	0.0011	0.0288
VB_F8	fresh	venous	0.8321	0.0017	0.0089	0.1574
VB_F9	fresh	venous	0.3141	0.0045	0.5856	0.0958
MB_B12	aged	menstrual	0.0002	0.9993	0.0003	0.0002
MB_B13	aged	menstrual	0.0526	0.0006	0.0193	0.9275
MB_B14	aged	menstrual	0.0015	0.9970	0.0006	0.0009
MB_B15	aged	menstrual	0.0342	0.4165	0.5421	0.0072
MB_B18	aged	menstrual	0.0144	0.0142	0.9665	0.0048
MB_B20	aged	menstrual	0.0003	0.9991	0.0004	0.0002
MB_F1	fresh	menstrual	0.0002	0.9993	0.0003	0.0002
MB_F2	fresh	menstrual	0.0002	0.9993	0.0003	0.0002
MB_F3	fresh	menstrual	0.0002	0.9993	0.0003	0.0002
MB_F4	fresh	menstrual	0.0003	0.9992	0.0003	0.0002
MB_F5	fresh	menstrual	0.0003	0.9992	0.0003	0.0002
MB_F6	fresh	menstrual	0.0003	0.9992	0.0003	0.0002
NB_B21	aged	nasal	0.0429	0.0071	0.0006	0.9495
NB_B22	aged	nasal	0.0235	0.0024	0.0025	0.9716
NB_F16	fresh	nasal	0.0120	0.0004	0.9815	0.0061
NB_F17	fresh	nasal	0.0082	0.0003	0.9863	0.0052
NB_F18	fresh	nasal	0.0119	0.0004	0.9802	0.0074
NB_F19	fresh	nasal	0.0002	0.0003	0.9992	0.0003
NB_F20	fresh	nasal	0.0041	0.0003	0.9921	0.0035
NB_F21	fresh	nasal	0.0002	0.0003	0.9993	0.0002
FB_F10	fresh	fingerprick	0.0036	0.0002	0.0049	0.9913
FB_F11	fresh	fingerprick	0.0012	0.0002	0.0013	0.9973
FB_F12	fresh	fingerprick	0.0007	0.0002	0.0007	0.9985
FB_F13	fresh	fingerprick	0.0083	0.9872	0.0004	0.0040
FB_F14	fresh	fingerprick	0.1749	0.5916	0.0010	0.2325
FB_F15	fresh	fingerprick	0.0079	0.0117	0.0021	0.9782



Supplementary Figure S7. Taxonomic profiles at the phylum level of the inconclusively classified test and mock casework blood samples; **(a)** menstrual blood, **(b)** nasal blood, **(c)** fingerprick blood, **(d)** venous blood.



Supplementary Figure S8. Taxonomic profiles at the genus level of the inconclusively classified test and mock casework blood samples; **(a)** menstrual blood, **(b)** nasal blood, **(c)** fingerprick blood, **(d)** venous blood.

PART II

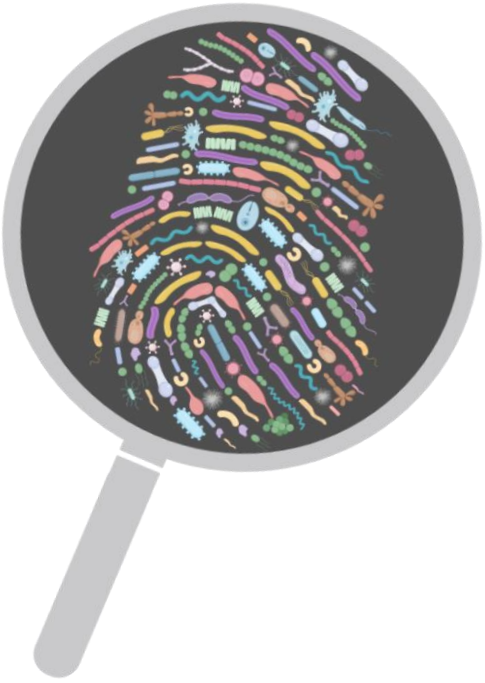
Human microbiome for estimating the time since deposition of forensic traces

Chapter 4

Estimating the time since deposition of saliva stains with a targeted bacterial DNA approach: a proof-of-principle study

Celia Díez López, Manfred Kayser, Athina Vidaki

Front Microbiol 12: 647933.



ABSTRACT

Information on the time when a human biological stain was deposited at a crime scene can be valuable in forensic investigations. It can help link a DNA-identified stain donor with a crime or provide with post-mortem interval estimation in cases with human cadavers. Available methods for estimating stain deposition time so far have limitations of different types and magnitudes. In this proof-of-principle study we investigated the use of microbial DNA for this purpose in human saliva stains. First, we identified the most abundant and frequent bacterial species in saliva using publicly available 16S rRNA gene next generation sequencing (NGS) data from 1,848 samples. Next, we assessed time-dependent changes in 15 identified species using *de novo* 16S rRNA gene NGS in saliva stains of two individuals exposed to indoor conditions up to 1 year. We selected four bacterial species i.e. *Fusobacterium periodonticum*, *Haemophilus parainfluenzae*, *Veillonella dispar*, and *Veillonella parvula*, showing statistically significant time-dependent changes and developed a 4-plex qPCR assay for their targeted analysis. Then, we analysed saliva stains of 15 individuals exposed to indoor conditions up to 1 month. We observed that bacterial counts generally increased with time, which explained 54.9% of the variation (p=<2.2E-16). Time since deposition explained ≥86.5% and ≥88.9% of the variation in each individual and species, respectively (p=<2.2E-16). Finally, based on sample duplicates we built and tested multiple linear regression models for predicting the time since deposition at an individual level, resulting in an average mean absolute error (MAE) of 5 days (ranging 3.3-7.8 days). Overall, the deposition time of 181 (81.5%) stains was correctly predicted within 1 week. Prediction models were also assessed in stains exposed to the same conditions seven months later, resulting in an average MAE of 8.8 days (ranging 3.9-16.9 days). Our proof-of-principle study suggests the potential of DNA profiling of human commensal bacteria for estimating time since deposition of saliva stains in the forensic scenario, which may be expanded to other forensically relevant tissues. Before our novel approach is considered for practical applications, forensic developmental validation and implementation criteria need to be met via future more dedicated studies.

INTRODUCTION

In routine forensic investigations, DNA profiling based on short tandem repeats (STRs) is the gold standard for identifying individuals who left a biological sample at the crime scene (1). However, the presence of a person's DNA at a crime scene does not necessarily allow us to conclude that the sample donor is the perpetrator, which is typically done in court using additional (non-genetic) information. One important additional information that can be crucial to solve a case is knowledge on (the time frame) when the DNA-identified person left the biological stain behind at the scene. Knowing the time since deposition of a crime scene stain can help the police assessing alibis given by known suspects or provide investigative information to search for the right suspect. Moreover, when multiple biological traces belonging to different donors are found at a scene, information on their time since deposition may help investigators select the ones with the highest investigative value for further analysis, in cases where the time of the crime is known. Furthermore, in some missing person cases, such knowledge on the time of stains deposition on relevant items (such as clothing) might also be linked with the time gone missing if this is unknown. Lastly, in crime scenes involving (parts of) a corpse, estimation of the time since deposition of stains found around/on the body can serve as an additional method to determine the time since death i.e. post-mortem interval (PMI).

For estimating the time since deposition of human biological stains, the most studied molecular approach so far has been the differential time-dependent degradation of human RNA; mainly mRNA markers (2-7) but also miRNA markers (4, 8). In principle, RNA decay continues *ex vivo* after a stain has been deposited, even if the biological material is dehydrated (9). However, not all markers investigated for their potential time-dependent degradation turned out informative in published studies due to their reported time-stability for months (4, 8, 10) or even a year (8). Nevertheless, the majority of studies to date have reported time-dependent decay of selected RNA markers and a few have also attempted to use this for estimating the time since deposition of body fluid stains.

A qPCR-based study on degradation profiles of two human hypoxia sensitive mRNA markers up to a month obtained mean absolute error (MAE) values of 2.7, 3.5 and 6.4 days in blood, saliva and semen stains, respectively (2). However, sample size was small (n=5 for each body fluid) and stain exposure did not mimic realistic forensic scenarios. In the cases of saliva and semen, fluids were left in tubes until swabbed rather than left to dry as stains. Bauer e*t al.* (7) analysed two mRNA markers in dried bloodstains using qPCR under the hypothesis that the 5'-end degrades at a faster rate than the 3'-end in mRNA and that relative degradation could be used to estimate the time since deposition of the stains. Significant levels in mRNA degradation were only reported in stains with large deposition time differences of at least 4-5 years, resulting in very large estimation intervals of several months or even a few years. Another qPCR-based study analysed four mRNA markers in dried bloodstains under the same previous hypothesis and reported a time estimation error of 2-4 weeks for stains exposed less than six months and 4-6 weeks for stains exposed between six and twelve months (6). Furthermore, Alshehhi *et al.* (8) employed qPCR targeting two and four mRNA/miRNA markers in dried saliva and semen stains, respectively, up to a year. On the one hand, the mRNA markers showed large fluctuation, no degradation or were not detected at all due to the assay's sensitivity past 90 days. On the other hand, the miRNA markers remained stable

across all analysed time points, making them not suitable to investigate time-dependent changes. Lastly, another study employed RNA next-generation sequencing (NGS) to analyse potential time-dependency of transcripts present in dried blood, saliva, semen and vaginal fluid stains up to a year (3). Particularly for the saliva transcripts, abundance values decreased rapidly and erratically; hence, no comprehensive analysis could be performed. For other stains the time-dependency of transcripts was useful for a limited time period less than a year. Overall, an RNA-based approach for estimating the time since deposition of stains could be promising but suffers from significant drawbacks including inter-individual degradation variation.

The human microbiome has been recently proposed as a promising tool in forensic science, especially for forensically relevant topics for which other approaches present challenges and limitations. For example, the human microbiome has proven to be a promising forensic tool for post-mortem interval (PMI) estimation based on predictable succession and colonization of microorganisms over time at different body sites (11-14). However, caution must be taken due to environmental (11, 15) and individual-specific (11) factors affecting time-dependent changes. The human microbiome, particularly the skin microbiome, can also serve as a kind of 'fingerprint' that is transferred to touched objects, promising for individual identification in cases where recovered human DNA is not sufficient for obtaining an STR profile ('touched' samples) (16-18). Additionally, we also showed that the human microbiome is suitable for the identification of the body site of origin of human body fluid stains, which can be of great value in crime scene reconstruction. For instance, when it comes to crime scene stains that contain epithelial cells from different body sites of origin including skin, saliva and vaginal fluids (19), and bloodstains from different body sites of origin including venous/arterial blood, menstrual blood, nasal blood or blood from skin injuries (20), where previous molecular approaches such as RNA-based ones have limitations (21-23). On top of these previously investigated forensic microbiome applications, we envisioned it a promising tool for estimating the time since deposition of human biological stains at a crime scene, which has not been studied yet.

Particularly, the oral human microbiome has been extensively characterized (24) and microorganisms living in the oral cavity comprise the second largest and most diverse microbial community of the human body (25) after the gut. Notably, one millilitre (mL) of saliva in healthy adults is estimated to contain approximately 100 million bacterial cells (26). Considering the normal salivary flow rate to be around 750 mL/day, $8x10^{10}$ bacterial cells are shed daily from the oral surfaces (26). As a result, human saliva samples are likely to contain a high number of bacterial cells, including dried saliva stains found at crime scenes that often are small and based on just a few microliters (µL) of liquid saliva. Additionally, it has been shown that the 'core' oral microbiome, which can be defined as the taxa shared among unrelated individuals (27), is quite large; actually larger than in other body sites such as the gut or skin (28). Finally, the oral microbiome has shown a high degree of *in vivo* time-wise stability within an individual, with no significant changes over months (28-30) and even a few years (31). Time-stable biological information of any biomarker used in forensics is crucial in investigations, especially in approaches where old crime scene samples may be used retrospectively, such as for estimating the time since stain deposition. So far, only a few studies investigated time-dependent microbiome changes in dried saliva samples exposed to

indoor conditions (32, 33) and due to their small sample size no meaningful conclusions could be made on how suitable the approach is for forensic purposes.

In this proof-of-principle study, we investigated for the first time the potential of genetic profiling of human saliva commensal bacteria for estimating the time since deposition of saliva stains, with promises in a future forensic scenario. To achieve this, we first identified the most abundant and most frequent bacterial species in saliva from a large publicly available 16S rRNA gene next-generation sequencing (NGS) dataset. Next, we assessed time-dependent changes in the relative abundance of the top identified bacterial target species in 16S rRNA gene NGS data we produced from dried saliva stains exposed to indoor long-term conditions up to 1 year. Based on the observed significant time-dependent changes, we further selected four bacterial species for which we developed a multiplex qPCR assay. Finally, we used this assay to analyse dried saliva stains exposed to indoor short-term conditions with various sample storage times up to 1 month.

MATERIALS AND METHODS

Saliva microbiome datasets

Publicly available human saliva 16S rRNA gene next generation sequencing (NGS) data from two previously published studies were obtained from the European Bioinformatics Institute (EMBL-EBI). These studies included data from three cohorts: the American Cancer Society Cancer Prevention Study II (ACS CPS-II) (N=543) (34) and the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial (N=661) (34) from which the produced microbiome data was published as part of the same study; and the American Gut Project (AGP) (N=1,089) (35). Accession numbers were PRJNA434300, PRJNA434312 and PRJEB11419, respectively. The studies' metadata were accessed via the National Center for Biotechnology Information (NCBI) and matched to the corresponding sample identifiers using custom Python scripts to create flat metadata tables. In the first study (ACS CPS-II/PLCO), quality control sample replicates were removed to avoid data redundancy. We also discarded samples with missing metadata information for age, sex and/or ethnicity. We also removed samples obtained from donors less than 15 years old given that differences between adult and youth saliva microbiomes are expected (36).

Most abundant and frequent bacterial species in saliva

We analyzed the above-mentioned human saliva 16S rRNA gene NGS data to identify the most abundant and frequent bacterial species across individuals included in the two selected studies (Figure 1A). Primer sequences were obtained from the original studies and were removed from the raw sequencing reads using cutadapt (v2.6) (37) by setting the minimum-length to 100 to discard processed reads shorter than 100 bp. The resulting FASTQ files were quality-filtered and de-noised using DADA2 (v1.12.1) (38). Parameter maxN was set to 0 in the ACS CPS-II/PLCO study to prevent unambiguous nucleotides in the sequencing reads, whereas maxN was set to 1 in the AGP study to avoid too few reads passing the filtering. Parameter maxEE for the maximum number of 'expected errors' in the reads was set to 2 in the two studies. Parameter truncLen was set based on the read quality profiles ensuring to

maintain an overlap between forward and reverse reads to be merged later. Following sample inference of true sequence variants, an amplicon sequence variants (ASV) table was constructed and chimeric sequences were removed. At this point, only high-coverage samples (>1,000 reads) were chosen for downstream analysis resulting in 525 (ACS CPS-II), 452 (PLCO) and 871 (AGP) samples. To assign taxonomy the ASV table was processed via the *assignTaxonomy* function in DADA2 at taxonomic ranks of interest (phylum, class, order, family, genus, species). The expanded Human Oral Microbiome Database (eHOMD) (v15.2) (24) was transformed to a DADA2-compatible training format and used as the reference database. Since the eHOMD database is bacteria-exclusive, the ASV table was further compared against the SILVA SSU r132 database (39) to check for sequences assigned to Eukarya, mitochondria or chloroplasts, which were subsequently filtered out. We additionally filtered out taxa present in less than 0.005% of relative abundance. Finally, in each study, we ordered the bacterial species according to their relative abundance (number of bacterial species sequencing reads divided by the total number of sequencing reads) and frequency (number of samples in which the bacterial species is reported divided by the total number of samples). As abundant and frequent bacterial species in saliva we selected the top 15 common ones across the studies. The processed NGS data in the form of relative abundance tables of the identified taxa can be found in Supplementary Data Sheet 1.

Saliva collection

Sample collection, handling and subsequent analysis adhered to the Medische Ethische Toestsings Commissie Erasmus MC (MEC-2018-1731). All individuals who donated saliva provided signed informed consent for the purpose of this study. One individual included in the long-term experiment (individual No 1) was also included in the short-term one (individual No 2) with a time span of two years between saliva collection for each of the experiments. In summary, we included 11 females and 5 males with an average age of 26.3 years and from various ethnic backgrounds, but mainly white Europeans (13/16). Information on our donors' sex, age, ethnicity and sample donation can be found on Supplementary Table S1. In brief, as part of donating saliva, individuals were asked to avoid all of the following at least for one hour before saliva donation: brushing their teeth, use mouthwash, eat food, use chewing gum and drink nothing but non-sparkling water. Individuals were independently asked to collect saliva in their mouth for a minute and spit into a sterile tube, repeating the process several times until reaching ~5 mL of saliva. Subsequently, for each individual, we prepared all stains per time point each consisting of 150 µL of saliva deposited on a sterile swab (PurFlock Ultra 6" sterile standard flock swabs, Puritan, Guilford, ME, USA). Therefore, for each individual all saliva stains were individual samples, though collected at the same time point. As substrate, we chose sterile swabs that are routinely used in forensics, for example to collect suspected saliva stains from objects found at the crime scene for molecular analysis. Also, their sterile nature allows us to conclude the absence of microbial contamination introduced before the saliva was deposited.

Dried saliva stains

With the exception of fresh (t0) saliva swabs that were processed straightaway, the prepared swabs were dried and directly exposed to our laboratory's environment apart from each other for a specific time prior to bacterial DNA isolation. The swabs were stored at standard room

temperature (20-25 °C), with relative humidity (30-50 %) and daily ambient light (8-11 hours). Ambient light included both artificial and natural light sources as the swabs were placed four meters away from a window (though not directly hit by the sun). First, for the long-term time-dependent bacterial composition analysis, saliva was collected from two individuals (Figure 1B). To sufficiently cover selected time points over a one-year period, seven saliva swabs were prepared per individual (single sample replicates per time point) and processed at day 1 (t1), 7 (t2), 30 (t3), 120 (t4), 180 (t5) and 365 (t6) after deposition. Second, for the short-term time-dependent bacterial marker analysis, saliva was collected from 15 individuals (Figure 1C). To sufficiently cover selected time points over a one-month period, 32 saliva swabs were prepared per individual (double replicates per time point); with the exception of one individual (No 1) for which there was insufficient volume of saliva to prepare the last time point. In this case, dried saliva swabs were processed at day 2 (t1), 4 (t2), 6 (t3), 8 (t4), 10 (t5), 12 (t6), 14 (t7), 16 (t8), 18 (t9), 20 (t10), 22 (t11), 24 (t12), 26 (t13), 28 (t14) and 30 (t15) after deposition. Additionally, six of these fifteen individuals also donated saliva seven months after the first collection date (Figure 1D). For this, eight saliva swabs were prepared per individual (single replicates per time point) and processed at day 2 (t1), 6 (t2), 10 (t3), 14 (t4), 18 (t5), 22 (t6), 26 (t7), 28 (t8 for individual No 1) and 30 (t8 for the rest of individuals) after deposition, corresponding to one-month time frame as in the first collection date. Additionally, swabs with no biological material were prepared as background blanks in both the long-term and short-term experiments and were exposed and processed in parallel at the same time points as the dried saliva stains.

Bacterial DNA isolation and quantification

Bacterial DNA isolation was performed with the QIAamp DNA Mini Kit (Qiagen, Germany) following the buccal swab spin protocol to simplify the isolation of DNA from human saliva samples deposited on a swab. We chose a kit that can co-isolate both bacterial and human DNA present in the sample to simultaneously allow for short tandem repeats (STR) profiling, necessary to identify the sample's donor; therefore, increasing the forensic applicability of the proposed approach. After appropriate optimization, we slightly modified the manufacturer's instructions for maximizing DNA yield. More specifically, the incubation time in step four was increased from 10 to 30 minutes, the elution was performed with nuclease-free water using a reduced 50 µL elution volume, spin columns were incubated for 5 minutes at room temperature following the addition of nuclease-free water and before centrifugation, the centrifugation time and speed were increased to 2 minutes and 12,000 rpm and finally, a second elution step using the eluate was added. Isolated bacterial DNA was quantified with the Femto™ Bacterial DNA Quantification kit (Zymo Research, Irvine, CA, USA) following the manufacturer's instructions on a CFX38 Touch™ Real-Time PCR System (Bio-Rad, Hercules, CA, USA).

Library preparation and sequencing

For the long-term time-dependent bacterial composition analysis we sequenced the obtained bacterial DNA from the dried saliva stains of the two individuals (N=14) as well as the background controls (N=7). Additionally, to assess the performance of the workflow we sequenced one negative control sample, one smart control (SC) sample for monitoring the library construction process and potential introduced contamination and one positive control -

a commercial microbial community DNA standard sample (ZymoBIOMICS™ Microbial Community DNA Standard, ZymoResearch). Library preparation was performed using the QIAseq™ 16S/ITS Panel Kit (Qiagen) for sequencing the V4-V5 regions of the 16S rRNA bacterial gene. Library quality control was performed with the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA) using a high sensitivity DNA chip following the manufacturer's instructions. Library quantification was performed using the KAPA Library Quantification Kit (Kapa Biosystems, Inc. Wilmington, MA, USA) following the manufacturer's instructions on a CFX384 Touch™ Real-Time PCR System (Bio-Rad). Libraries where diluted down to 2 nM, or to the highest possible concentration in case the library concentration was <2 nM and pooled together for 2x276 bp paired-end sequencing on a MiSeq platform using the MiSeq v3 Reagent Kit (Illumina, San Diego, CA, USA).

Long-term time-dependent differential bacterial abundance analysis

We performed differential abundance analysis to identify changes in the relative abundance of bacterial species over time in the 16S rRNA gene NGS data derived from the long-term dried saliva stains. Phased primer sequences were removed from the raw sequencing reads using a custom Python script. Subsequent filtering, de-noising, ASV table construction and taxonomy annotation was carried out as previously described for the publicly available saliva microbiome datasets. We chose gneiss (40) for the differential abundance analysis since it acknowledges the compositional nature of microbiome data. Based on this compositional nature it is only possible to infer relative, but not absolute, abundance changes with time since abundance change of one species influences the abundance changes in the other species. Gneiss was run using the q2-gneiss plugin in QIIME2 (v.2019.10) (41). Input data comprised of a microbial profile sub-selection of the 15 most abundant and most frequent bacterial species in saliva as previously identified. First, a bifurcating tree is built relating bacterial species to each other based on how they co-occur by using Ward's hierarchical clustering via the *correlation-clustering* command. Each balance (internal nodes in the tree) is calculated by taking the log ratio of geometric means of subtrees via the *ilr-transform* command. Each balance is indicated as '*y*' followed by an ordinal number, being *y0* the first balance in the root of the constructed tree. The taxa on one side of the balance are termed as numerators and on the other side as denominators. Each log ratio's numerical value depends on the balance between the numerator's and denominator's taxa, and can be either positive, negative or null. Differences in the log ratio balances can be compared between sample groups to infer relative changes in the microbial composition. These log-transformed balances were used to construct a multivariate response linear model using the time since deposition and the individual ID as covariates using the *ols-regression* command where 10-fold cross validation of 10 partitions showed no overfitting. The regression summary showed the contributions of the covariates to the abundances of the selected bacterial species. Balances significantly affected by the covariates were determined with a *p* value cutoff at 0.05 after Bonferroni correction. These *p* values were based on relative, rather than absolute, values resulting from inter-dependent taxa. Significant balances for time since deposition but not for individual ID were selected as the most informative for the purpose of this study and used to analyze the informative bacterial species via qPCR in the short-term dried saliva stains.

4-plex qPCR assay design and optimization

Based on the differential abundance analysis results in total four bacterial species were selected for qPCR analysis in the short-term dried saliva stains. The selected species were *Fusobacterium periodonticum, Haemophilus parainfluenzae, Veillonella dispar and Veillonella parvula.* We aimed to design a suitable 4-plex qPCR assay based on TaqMan probe technology that would allow for the simultaneous analysis of all four selected bacterial species using species-specific primers that target single-copy genes. For *F. periodonticum*, *V. dispar* and *V. parvula* we chose the beta subunit of RNA polymerase gene (rpoB) as the target gene; for *H. parainfluenzae* we chose the translation initiation factor IF-2 gene (infB). A literature search was conducted to find previously designed suitable primers resulting in the reverse primers for *V. dispar* and *V. parvula* (42). The rest of primer sequences as well as the probe sequences were manually designed using the PrimerQuest Tool (Integrated DNA Technologies, IDT, Coralville, IA, US). The fluorescent dyes labeled to the 5'-end of the probe sequences were: 6-carboxyfluorescein (6-FAM) for *F. periodonticum*, cyanine 5 (Cy5) for *H. parainfluenzae*, Texas red-615 (TEX-615) for *V. dispar*, and hexachloro-fluorescein (HEX) for *V. parvula.* To test primer pair specificity each pair was compared against the nucleotide collection database from the National Center for Biotechnology Information (NCBI) using Primer BLAST. The Autodimer software (43) was also used to assess the potential formation of primer dimers and hairpins under our experimental conditions. Final primer and probe sequences are summarized in Supplementary Table S2.

The 4-plex qPCR assay was developed based on the CFX384 Touch™ Real-Time PCR System (Bio-Rad). The assay was optimized according to various parameters including annealing temperature and primer/probe concentrations. The optimal oligo concentrations varied for each bacterial target and were determined as follows (primers/probe): *F. periodonticum* (0.7/0.5 µM)*, H. parainfluenzae* (0.6/0.5 µM)*, V. dispar* (0.2/0.05 µM) *and V. parvula* (0.9/0.5 µM). Synthetic double stranded DNA fragments (gBlocks, IDT) for each of the bacterial target gene fragments were used as standard samples (positive controls) (Supplementary Table S2). Concentrations were converted to copy numbers by using the formula:

$$(C) * (M) * (1*10^{-15}\ mol/fmol) * (Avogadro's\ number) = copy\ number/\mu L$$

where C is the concentration of the gBlock gene fragment in ng/µL and M is the molecular weight in fmol/ng. gBlocks were mixed in known concentrations ranging from 125,000 down to 61 copies per each bacterial target gene fragment. The assay was performed in a 20 µL reaction in triplicate, including 10 µL of iQ Multiplex Powermix (Bio-Rad), 4 µL of each primer (forward and reverse) and probe mix (5X), 1 µL of 25 µM of $MgCl_2$ (Thermo Fisher Scientific, Waltham, MA, USA), 0.5 µL of 20 mg/mL of bovine serum albumin (BSA) (New England Biolabs, Ipswich, MA, USA), 1 µL of bacterial DNA (corresponding to 2 ng) and 3.5 µL of nuclease-free water. The thermocycling program included an initial denaturation and polymerase activation step at 95 °C for 3 min, followed by 35 PCR cycles of 95 °C for 10 s and an extension step of 60 °C for 45 s.

qPCR data analysis

Using our developed and optimized 4-plex qPCR assay we analyzed the short-term dried saliva stains of 15 individuals. The standard samples with known concentrations per bacterial target gene fragment (gBlocks) were used to create the best-fitted linearity curve. The

efficiency of each qPCR assay was calculated from the slope of the serially diluted standard samples according to the equation (44):

$$E = 10^{-(1/slope)}.$$

For each reaction we obtained the quantification cycle (*Cq*) value, the point at which fluorescence above the threshold level is detectable. To standardize, the threshold was set to 100 relative fluorescence units (RFU) for all reactions and fragments. The copy number (*cn*) for each bacterial target gene fragment was calculated according to the equation:

$$cn = e^{-Cq}.$$

Since we target single-copy genes, reported copy numbers can be translated to bacterial cell counts. Between-plate variation was removed using the Factor-qPCR tool (45) in the stains produced at the first donation time point and seven months later for six individuals. We set the qPCR plate ID as the variable causing the variation to be removed, while the bacterial marker and time since deposition as the variables for which preserve their effects. Resulting between-plate correction can result in negative count values in some cases.

Short-term time-dependent bacterial analysis

We aimed to investigate the statistical relationship between the time since deposition and the four selected bacteria cell counts in the short-term dried saliva stains. Since the focus was on dried saliva stains, fresh (t0) samples were excluded from this analysis. Sample duplicates collected at each time point for each donor were analysed independently from each other to assess the magnitude of sample variation. For each analysed stain qPCR triplicates were considered as separate samples to account for potential reaction variation. Various linear regression models were built using the *lm()* function in the lme4package (v.1.1.20) (46) in R (v.3.6.1 [2019-07-05]). The linear models were based on the functions below, where C refers to the bacterial cell count, I to the individual, S to the bacterial species and finally, T to the time since deposition (in days). Interactions between variables are indicated with an asterisk (*). The statistical relationship between each bacterial species cell count for each individual and the time since deposition was calculated based on the function:

$$lm\ (C \sim T).$$

The statistical relationship between the four species cell count and the time since deposition, species and their interaction for each individual was calculated based on the function:

$$lm\ (C \sim T*S).$$

The statistical relationship between each species cell count and the time since deposition, individual and their interaction was calculated based on the function:

$$lm\ (C \sim T*I).$$

Finally, the statistical relationship between the four species cell count and the time since deposition, species, individual and their interaction was calculated based on the function:

$$lm\ (C \sim T*S + T*I).$$

Adjusted R^2 and *p* values were evaluated for all the linear models. Sample variation was assessed by testing for equality between the coefficients in the linear regression models of each of the sample duplicate sets using the Chow test implemented in the gap R-package (v.1.2.2) (47). Significant *p* values were determined with a value cutoff at 0.05 following

Benjamini-Hochberg (BH) correction. All plots were generated with the ggplot2 R-package (v.3.3.2) (48).

Time since deposition prediction modelling

We further investigated the possibility to predict the time since deposition in the short-term dried saliva stains. We once again excluded fresh (t0) samples based on the notion that it is not feasible to collect a purely fresh saliva sample at a real crime scene. We first attempted a generalized time since deposition prediction model based on random forest (RF) regression using the randomForest R-package (v.4.6.14) (49). To evaluate the generalizability of this approach and to avoid prediction biases, we built a model based on the average detected microbial DNA cell counts of the four targeted species per time point. We then used data from all time points from 14 of the analysed individuals as the training set, while keeping all the time points of the remaining individual as the testing set. By this, the tested individual was not present in the training set to mimic real-life applications. We repeated this process 15 times given the 15 individuals in our dataset. The 15 RF models were based on a 5-fold cross-validation repeated for three times and 500 trees with the four variables (targeted bacterial species) sampled at each split. NA values were replaced with column medians using the *na.roughfix* command. The average performance of the generalized RF models was assessed using the mean absolute error (MAE) that measures the discrepancies between predicted and real values according to the formula below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |yi - y|$$

where n is the total number of data points, y_i is the real value and y is the predicted value. Pearson's correlation was used to calculate the correlation (r) between real and predicted values. MAE and r were calculated with the Metrics R-package (v.0.1.4) (50).

For the individualized modelling approach, sample duplicates collected at each time point for each donor were considered separately; namely, one duplicate was used as the training sample, while the other was used as the testing sample, mimicking a potential future forensic scenario of having both references and crime scene samples. As predictors we chose the bacterial cell counts of the four selected species at each time point. Multiple linear regression (MLR) models were built using the *lm()* function in the lme4 R-package (46) based on the function below:

$$lm\ (T \sim C_1 + C_2 + C_3 + C_4)$$

where C refers to each bacterial species cell counts, and T to the time since deposition in days. Additionally, the follow-up dried saliva stains of the selected six individuals were also analysed as testing samples. In this case, as time since deposition predictors we did not only consider the bacterial cell counts for the four selected species but all the possible combinations of also one, two and three predictor species in order to select the model with the lowest error for each individual. The donor-specific prediction models were evaluated based on the adjusted R^2 and *p* values, where significant *p* values were determined with a value cutoff at 0.05 following Benjamini-Hochberg (BH) correction. The average model performances were assessed using the mean absolute error (MAE) and Pearson's correlation was used to calculate the correlation (r) between real and predicted values. All plots were generated with the ggplot2 R-package (v.3.3.2) (48). The processed qPCR data used to build and test the prediction models can be found in Supplementary Data Sheet 1.

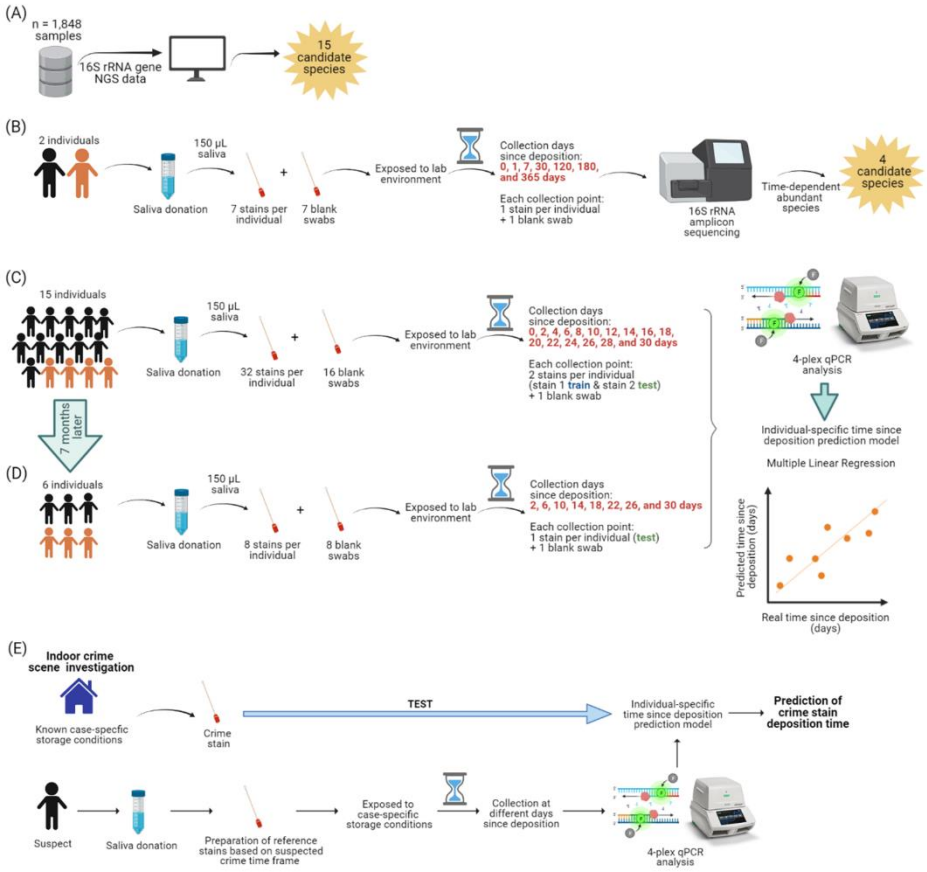


Figure 1. Overview of the study. **(A)** Identification of the most abundant and frequent bacterial species in saliva from publicly available 16S rRNA gene next-generation sequencing (NGS) data. **(B)** Long-term (up to 1 year) time-dependent differentially abundance analysis in saliva stains from two individuals. The individuals are color-coded; black is female and orange is male. **(C)** Short-term (up to 1 month) targeted analysis of species informative for time since deposition of saliva stains from 15 individuals and **(D)** from six individuals re-sampled seven months later. **(E)** Hypothesized individualized solution to saliva trace timing as viewed in our dataset.

RESULTS

Selection of most abundant and most frequent bacterial species in human saliva from large 16S rRNA gene sequencing data

Publicly available 16S rRNA gene NGS data from 1,848 human saliva samples were analysed to identify the most abundant and most frequent bacterial species across the studies they were retrieved from (34, 35). A total of 10,326,403 sequencing reads were retrieved from the ACS CPS-II/PLCO study and 31,046,365 sequencing reads from the AGP study. In the ACS CPS-II/PLCO study, 218 bacterial species from 35 families were identified, while in the AGP study 471 bacterial species from 88 families were found. We then selected the top 15 most abundant and most frequent bacterial species from a total of 10 families that were common

across these studies, namely *Actinomyces sp. HMT 180*, *Fusobacterium periodonticum*, *Granulicatella adiacens, Haemophilus parainfluenzae, Leptotrichia sp. HMT 417, Porphyromonas pasteri, Prevotella melaninogenica, Prevotella salivae, Prevotella veroralis, Rothia mucilaginosa, Streptococcus oralis subs. dentisani clade 058, Streptococcus parasanguinis clade 411, Streptococcus salivarus*, *Veillonella dispar* and *Veillonella parvula* (Figure 2A). These 15 identified common species accounted for 66.0% (6,817,142) of the sequencing reads in the ACS CPS-II/PLCO study and 55.1% (17,095,402) of the sequencing reads in the AGP study. In the ACS CPS-II/PLCO study, *S. oralis subs. dentisani clade 058* was the most abundant species accounting for 24.9% (2,567,719) of the reads, whereas *L. sp. HMT 417* was the least abundant accounting for 0.59% (61,475) of the reads. In the AGP study, *R. mucilaginosa* was the most abundant species accounting for 10.7% (3,305,923) of the reads, whereas *P. salivae* was the least abundant accounting for 0.47% (146,083) of the reads. Overall, these 15 common species were similar in abundance across the analysed studies, with the exception of *S. oralis subs. dentisani clade 058* which was markedly more abundant in the ACS CPS-II/PLCO than in the AGP study (24.9% *vs.* 9.1% of total reads) and *R. mucilaginosa* which was more abundant in the AGP than in the ACS CPS-II/PLCO study (10.6% *vs.* 5.4% of total reads). The most frequent species was *S. oralis subs. dentisani clade 058* present in 97.0% (948) of the ACS CPS-II/PLCO study's individuals and in 95.2% (829) of the AGP study's individuals. The less frequent species in the ACS CPS-II/PLCO study was *L. sp. HMT 417* present in 46.2% (451) of the individuals, whereas in the AGP study it was *P. veroralis* present in 33.9% (295) of the individuals.

Selection of bacterial species with time-dependent relative abundance in long-term dried saliva stains using *de novo* 16S rRNA gene sequencing

Dried saliva stains produced from two individuals and exposed to indoor conditions for different time periods up to 1 year were analysed. The obtained 16S rRNA microbial profiles were clearly distinct from the background blanks (empty swabs) for each time point, indicating a low level of bacterial contamination (Supplementary Figure S1). We then extracted the data of the 15 most abundant and most frequent bacterial species identified in our previous *in silico* analysis (Figure 2A) to identify for which species their relative abundance significantly changed over time. The fit of the overall multivariate response linear model was R^2=0.21, with the time since deposition accounting for 7% and the individual person accounting for 15% of the bacterial variation.

A total of 13 log ratio balances (from y0 to y12) were generated as internal nodes in the built tree. Log ratio balances y4 (p=0.008), y5 (p=0.004) and y7 (p=0.022) were significantly different for the time since deposition (Table 1). Balance y4 was composed of *A. sp. HMT 180, S. oralis subsp. dentisani clade 058, S. parasanguinis clade 411, P. pasteri, P. melaninogenica* and *P. veroralis* as numerator's taxa; and *F. periodonticum, P. melaninogenica, V. dispar* and *V. parvula* as denominator's taxa. Balance y5 was composed of *F. periodonticum, H. parainfluenzae* and *V. dispar* as numerator's taxa; and *V. parvula* as denominator's taxa. Balance y7 was composed of *F. periodonticum* and *H. parainfluenzae* as numerator's taxa; and *V. dispar* as denominator's taxa. It has to be noted that balance y7 is a subdivision of balance y5 numerator (Table 1). An overview of the generated log ratio balances, intercept and *p* values for the time since deposition and individual person can be found in Supplementary Table S3.

Table 1. Significant log ratio balances for time since deposition in the differential abundance analysis. The top 15 most abundant and frequent bacterial species in human saliva were sub-selected from the *de novo* generated 16S rRNA gene NGS data obtained from the long-term dried saliva stains. Each balance is composed of the numerator's bacterial taxa and the denominator's bacterial taxa.

Balance	**Bacterial species**	***p* value; Time since deposition**	***p* value; Individual person**
$y4_{numerator}$	*A. sp. HMT 180* *S. oralis subsp. dentisani clade 058* *S. parasangunis clade 411* *P. pasteri* *P. melaningenica* *P. veroralis*	0.008	0.390
$y4_{denominator}$	*F. periodonticum* *P. melaningenica* *V. dispar* *V. parvula*		
$y5_{numerator}$	*F. periodonticum* *H. parainfluenzae* *V. dispar*	0.004	0.911
$y5_{denominator}$	*V. parvula*		
$y7_{numerator}$	*F. periodonticum* *H. parainfluenzae*	0.022	0.022
$y7_{denominator}$	*V. dispar*		

For the purpose of this study we sub-selected the log ratio balance y5 as our reference because of its strongest significant time dependency (p=0.004) in both individuals. We preferred y5 over its subdivision y7 since balances toward the root of the tree capture more information as they contain more tree tips. Furthermore, three of the four bacterial species in y5 were also present in y4, which also showed strong significant time dependency (p=0.008), albeit less strong than y5 (Table 1). For both individuals there was a similar pattern in the log ratio evolution of balance y5 through time since saliva stain deposition, though the rate of change was individual-specific (Figure 2B). From 7 to 365 days, the general trend was the increase of the log ratios' values for both individuals. Looking at the relative abundances of the four species from balance y5 at day 7 and day 365 since deposition we observed that for *H. parainfluenzae* relative abundance increased in both individuals, for *V. dispar* and *V. parvula* relative abundances decreased in both individuals and for *F. periodonticum* relative abundance increased in individual 1, whereas it slightly decreased in individual 2 (Figure 2C). Based on these results, the four species composing balance y5 were selected for developing a 4-plex qPCR assay for their targeted analysis in the short-term dried saliva stains. Parallel to the NGS analysis, the relative abundance of the four selected species in the background (blank) swabs was very low (≤1%) on average (mean±standard deviation), as follows: *F. periodonticum* (0.008±0.010), *H. parainfluenzae* (0.010±0.009), *V. dispar* (0.005±0.006) and *V. parvula* (0.006±0.009).

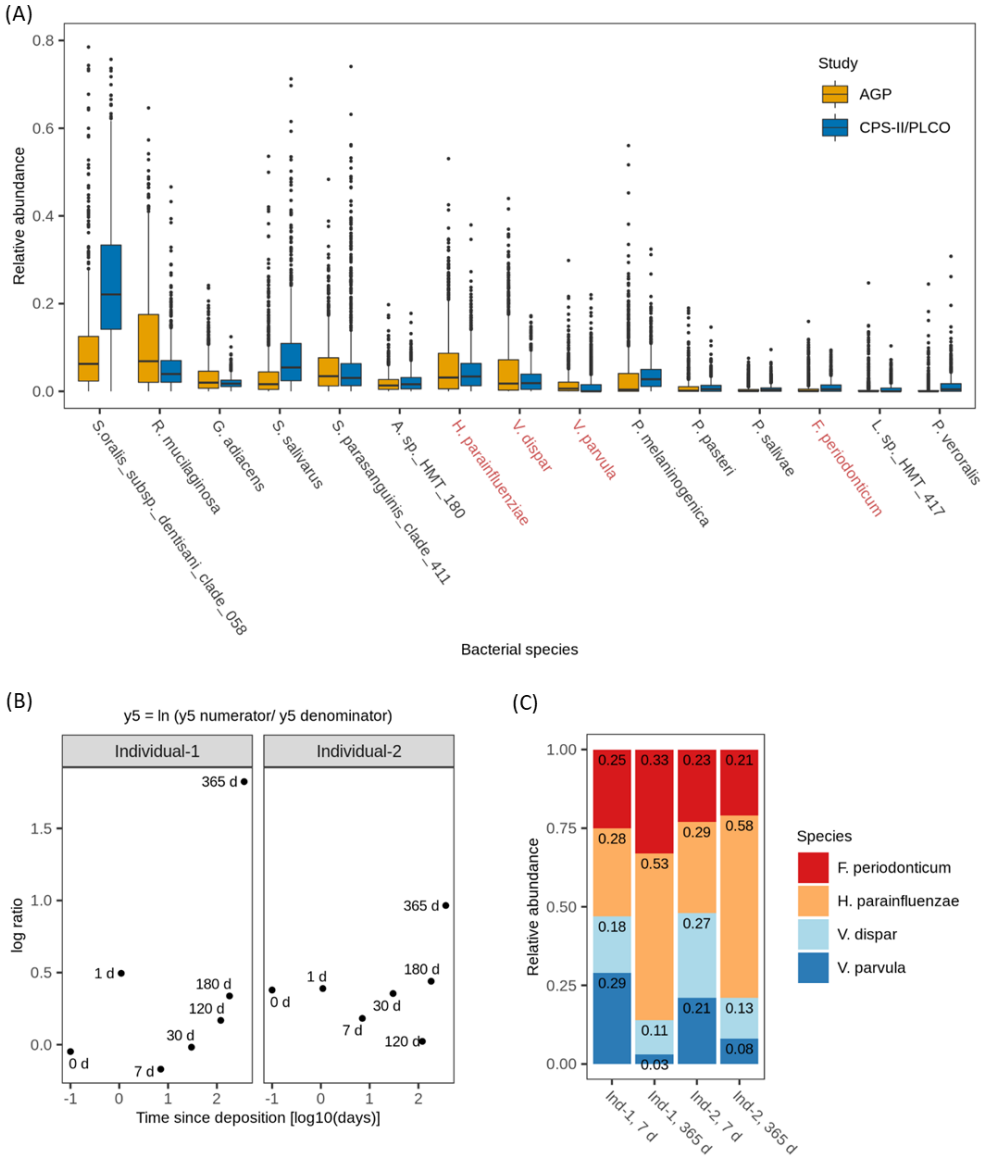


Figure 2. (A) Relative abundances of the 15 most abundant and frequent bacterial species in human saliva from adults across the analysed publicly available 16S rRNA gene NGS datasets (N=1,848). Highlighted in red are the four bacterial species subsequently included in the 4-plex qPCR assay. **(B)** Log ratio balance y5 significantly different for the time since deposition in the *de novo* generated 16S rRNA gene NGS data from the long-term (up to 1 year) dried saliva stains. Time since deposition days (x-axis) were log-transformed to facilitate the visualization of the earliest time points. Each dot represents a long-term dried saliva stain with its corresponding time since deposition. **(C)** Relative abundances of the four bacterial species from balance y5 (*Fusobacterium periodonticum, Haemophilus parainfluenzae, Veillonella dispar* and *Veillonella parvula)* in the long-term (up to 1 year) dried saliva stains at day 7 and day 365 since stain deposition for both analysed individuals.

Relationship between bacterial abundance and time since deposition in short-term dried saliva stains using multiplex qPCR

Dried saliva stains from 15 individuals up to 1 month since deposition were analysed using our developed 4-plex qPCR assay. Parallel to the qPCR analysis, no signal above the set threshold was reported in the background (blank) swabs for any of the four bacterial markers and time point. The qPCR results obtained from the fresh (t0) samples confirmed that the four selected bacterial species were abundant and frequent in the saliva of all 15 individuals, although we observed high inter-individual variation within and between species. For each of the four species, the average and standard deviation (mean±SD) as well as the minimum and maximum value (range) of qPCR-derived cell counts in 1 µL of isolated bacterial DNA solution (equivalent to 2 ng of total bacterial DNA) were as follows: *F. periodonticum* (23,386±24,598; range 2,698-105,567)*, H. parainfluenzae* (83,854 ± 80,412; range 13,167-331,667)*, V. dispar* (20,071±27,406; range 399-91,937) and *V. parvula* (9,825±21,354; range 309-92,347). A figure of the bacterial cell count distribution in fresh saliva samples for each of the four species can be found in Supplementary Figure S2.

We first investigated the time dependency of each of the four species in each individual in the dried saliva stains ranging from 2 days (t1) up to 30 days (t15) since deposition. The Chow test for equality showed no significant differences in the great majority of the compared time point swab duplicates' regressions. Exceptions were the univariate linear regressions for *F. periodonticum* in individual 2 (p=0.010), individual 7 (p = 0.005) and individual 15 (p = 0.020) and for *H. parainfluenzae* in individual 6 (p=0.020). We observed high inter-individual differences in terms of the amount of variation explained by time for each species' cell count (Figures 3, 4, 5, 6). For example, in individual 5 the variation explained for *F. periodonticum* cell count was high in both duplicates (R^2=0.663, p=1.06E-10 in duplicate 1; and R^2=0.522, p=6.92E-08 in duplicate 2) (Figure 3). However, the variation explained for *V. parvula* was much lower, even close to zero (R^2=0.001, p=0.374 in duplicate 1; and R^2=0.062, p=0.092 in duplicate 2) (Figure 6). The univariate regression results including R^2 values, BH-corrected p values and significance testing can be found in Supplementary Table S4.

For *F. periodonticum*, the qPCR-derived cell count increased over time for most individuals except individual 11, although at different rates in the different individuals. The average and standard deviation fold-change between day 2 (t1) and day 30 (t15) was 2.0±0.6, range 1.1-3.3. The highest time-dependent bacterial increase was reported for individual 2 (R^2=0.825, p=2.59E-14, sample duplicate 1) (Figure 3). For *H. parainfluenzae,* the time-dependent behavior varied in an individual-specific manner meaning either increasing (individuals 1, 4, 5, 6, 7, 10, 13, 14), decreasing (individuals 9, 11) or barely changing (individuals 2, 3, 8, 12, 15) (Figure 4). For *V. dispar*, the cell count increased with time in the majority of the individuals at different rates except individual 11. The average and standard deviation fold-change between day 2 (t1) and day 30 (t15) was 1.7±1.1, range 1.1-7.0. The highest time-dependent bacterial increase was reported for individual 14 (R^2=0.623, p=8.70E-10, sample duplicate 1) (Figure 5). Lastly, for *V. parvula,* cell count increased with time for some individuals (individuals 1, 2, 4, 7, 10, 13, 14), whereas for others it decreased (individuals 11, 12, 15) or barely changed (individuals 3, 5, 6, 9) (Figure 6).

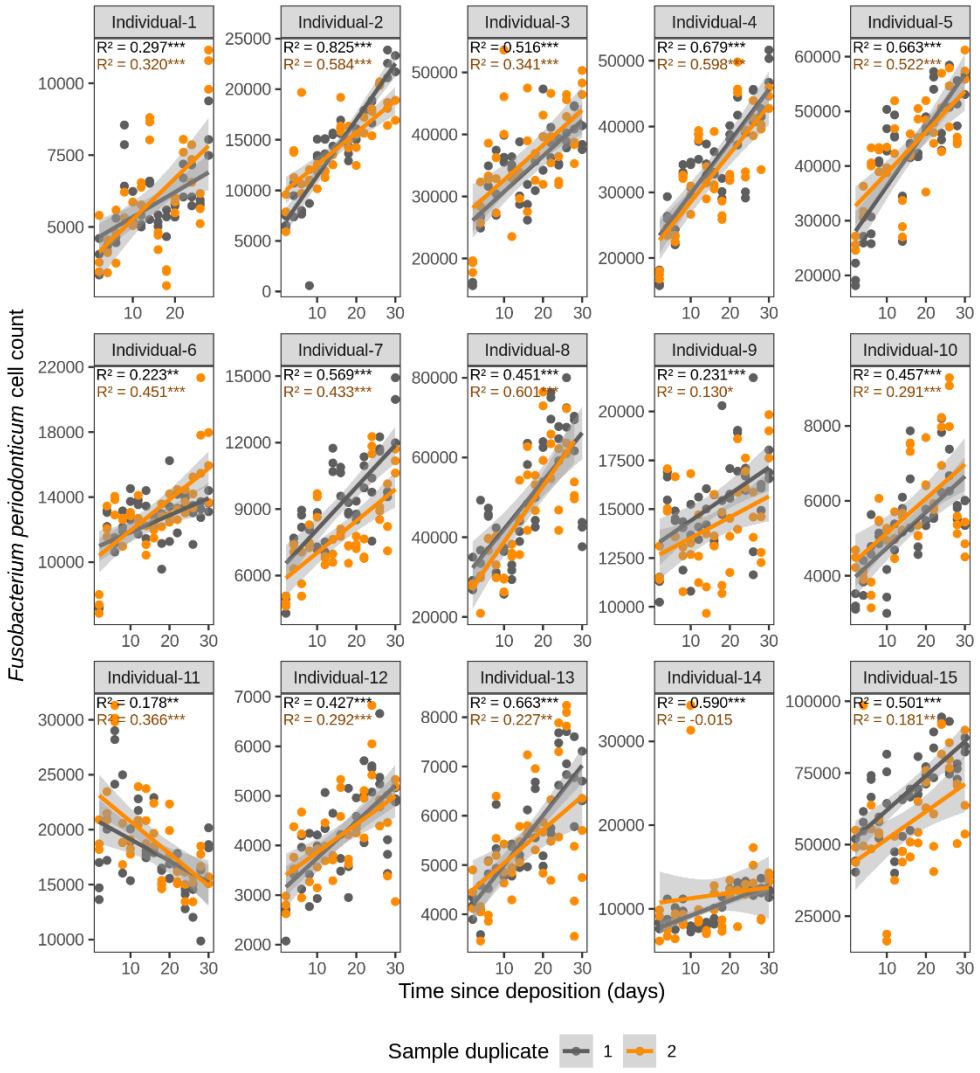


Figure 3. Time-dependency of qPCR-based *Fusobacterium periodonticum* cell count for each individual in the short-term (up to 1 month) dried saliva stains from 15 individuals. Sample duplicate set 1 is indicated in gray and sample duplicate set 2 in orange. Each dot represents a qPCR-run sample triplicate. R^2 values indicate the variation explained by the time since deposition in the bacterial cell count. Asterisks indicate the significance level of the Benjamini-Hochberg corrected *p* values as follows: 0.001 ***, 0.01 ** and 0.05 *.

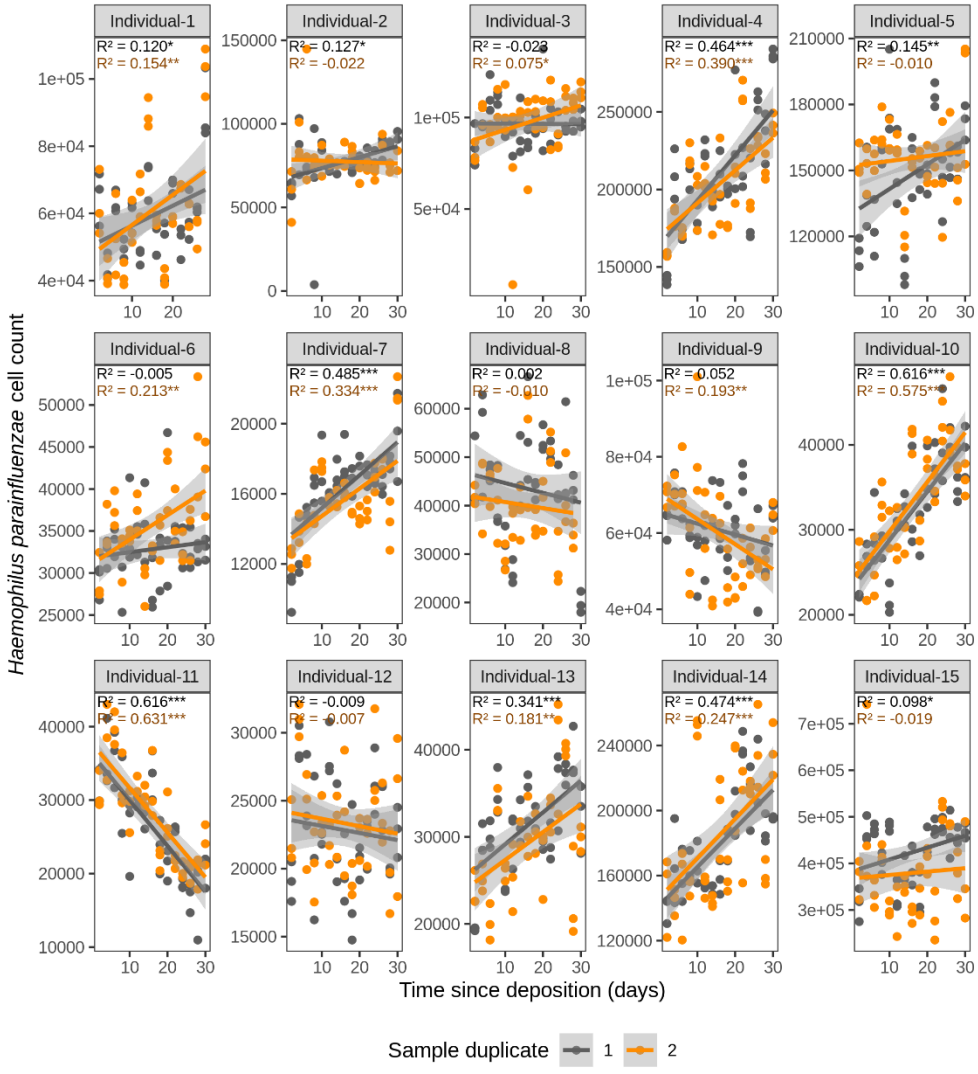


Figure 4. Time-dependency of qPCR-based *Haemophilus parainfluenzae* cell count for each individual in the short-term (up to 1 month) dried saliva stains from 15 individuals. Sample duplicate set 1 is indicated in gray and sample duplicate set 2 in orange. Each dot represents a qPCR-run sample triplicate. R^2 values indicate the variation explained by the time since deposition in the bacterial cell count. Asterisks indicate the significance level of the Benjamini-Hochberg corrected *p* values as follows: 0.001 ***, 0.01 ** and 0.05 *.

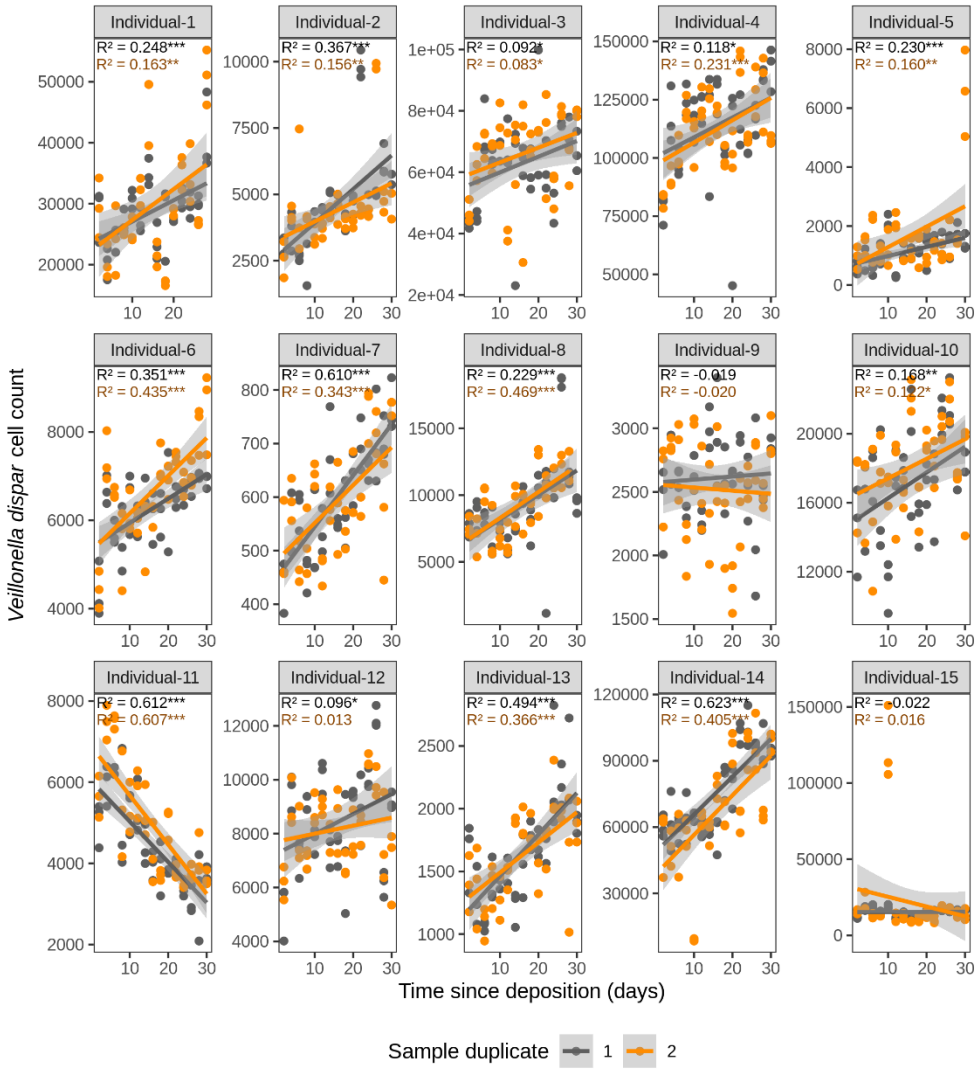


Figure 5. Time-dependency of qPCR-derived *Veillonella dispar* cell count for each individual in the short-term (up to 1 month) dried saliva stains from 15 individuals. Sample duplicate set 1 is indicated in gray and sample duplicate set 2 in orange. Each dot represents a qPCR-run sample triplicate. R^2 values indicate the variation explained by the time since deposition in the bacterial cell count. Asterisks indicate the significance level of the Benjamini-Hochberg corrected *p* values as follows: 0.001 ***, 0.01 ** and 0.05 *.

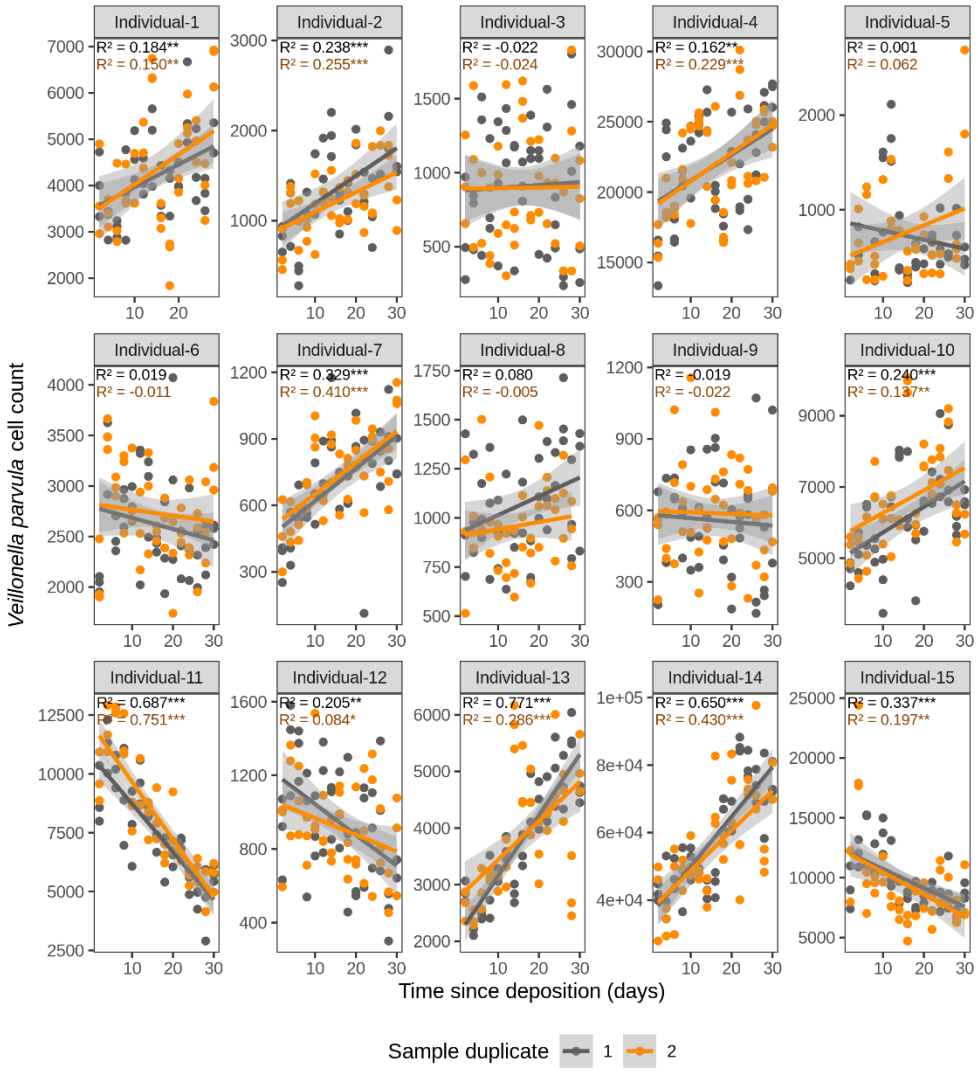


Figure 6. Time-dependency of qPCR-derived *Veillonella parvula* cell count for each individual in the short-term (up to 1 month) dried saliva stains from 15 individuals. Sample duplicate set 1 is indicated in gray and sample duplicate set 2 in orange. Each dot represents a qPCR-run sample triplicate. R^2 values indicate the variation explained by the time since deposition in the bacterial cell count. Asterisks indicate the significance level of the Benjamini-Hochberg corrected *p* values as follows: 0.001 ***, 0.01 ** and 0.05 *.

We next investigated the time dependency of the four bacterial species altogether in each individual (Table 2). The variation explained by time in the four species cell count varied among individuals but it was very high in the majority of them with R^2 values ranging between 0.865 and 0.979 (*p*=<2.20E-16). The individual with the highest variation explained on average was individual 5 (R^2=0.964 in sample duplicate 1, R^2=0.979 in sample duplicate 2, *p*=<2.20E-16). The individual with the lowest variation explained on average was individual 8 (R^2=0.874 in duplicate 1, R^2=0.908 in duplicate 2, *p*=<2.20E-16). We also investigated the time dependency of each bacterial species considering the 15 individuals altogether (Table 2). The variation explained by time for each of the four bacterial species cell count in all individuals together was very high and significant. The strongest time dependency variation on average (R^2 in duplicate 1, R^2 in duplicate 2, *p* value) was observed for *V. parvula* (R^2=0.969, R^2=0.948, *p*=<2.20E-16), followed by *H. parainfluenzae* (R^2=0.964, R^2=0.920, p=<2.20E-16), *V. dispar* (R^2=0.959, R^2=0.906, p=<2.20E-16) and finally, *F. periodonticum* (R^2=0.941, R^2=0.889, p=<2.20E-16). Finally, we investigated the time dependency of the four bacterial species together in the 15 individuals altogether (Table 2) resulting in a significant variation explained in the dataset of R^2=0.544 in duplicate 1 and R^2=0.548 in duplicate 2, p=<2.20E-16.

Table 2. Time-dependency of the bacterial marker cell counts in the dried saliva stains exposed to indoor conditions up to 1 month for sample duplicate sets 1 and 2. lm(C~T*S): time-dependency of the four selected bacterial species altogether in each individual, lm(C~T*I): time-dependency of each selected bacterial species considering the 15 individuals altogether and lm(C~T*I + T*S): time-dependency of the four selected bacterial species and 15 individuals altogether. R^2 values indicate the variation explained by the time since stain deposition in the qPCR-derived bacterial cell counts. Asterisks indicate the significance level of the Benjamini-Hochberg corrected *p* values as follows: 0.001 ***, 0.01 ** and 0.05 *.

		Sample duplicate 1		Sample duplicate 2	
Linear regression analysis		**R^2**	***p* value**	**R^2**	***p* value**
lm(C~T*S)	Individual 1	0.922	<2.2E-16***	0.865	<2.2E-16***
	Individual 2	0.952		0.948	
	Individual 3	0.940		0.907	
	Individual 4	0.955		0.966	
	Individual 5	0.964		0.979	
	Individual 6	0.968		0.962	
	Individual 7	0.973		0.966	
	Individual 8	0.874		0.908	
	Individual 9	0.965		0.946	
	Individual 10	0.953		0.950	
	Individual 11	0.912		0.927	
	Individual 12	0.936		0.946	
	Individual 13	0.968		0.937	
	Individual 14	0.959		0.905	
	Individual 15	0.968		0.902	
lm(C~T*I)	*F. periodonticum*	0.941		0.889	
	H. parainfluenzae	0.964		0.920	
	V. dispar	0.959		0.906	
	V. parvula	0.969		0.948	
lm(C~T*I + T*S)	Overall	0.546		0.551	

Estimating the time since deposition of dried human saliva stains based on bacterial DNA

We finally investigated the possibility to estimate the time since deposition of the dried saliva stains exposed to indoor short-term conditions up to 1 month using a generalized RF regression model. The correlation between real and predicted time since deposition values was very low (r=0.11) (Supplementary Figure S3) and the average MAE was 8 days. The real and predicted times since deposition for each individual are summarized in Supplementary Table S5. From these predicted values it is clear that the generalized approach is unable to discriminate between early and late times since deposition in the analysed interval of 1 month. This is clear, for instance, in individual 4 for whom all times since deposition are predicted as either 20 or 21 days and individual 7 for whom all times since deposition were predicted as 16, 17 or 18 days (Supplementary Table S5). Hence, the time-dependent variation in the four targeted bacterial species was surpassed by the high inter-individual variation in our dataset. The high inter-individual variation can be clearly observed using principal component analysis (PCA) (Figure 7) where the saliva stains cluster based on the individual. Hence, because of the high inter-individual variation we observed, also as described in section 3.3, which limited the implementation of a generalized model in our dataset, we built individual-specific models.

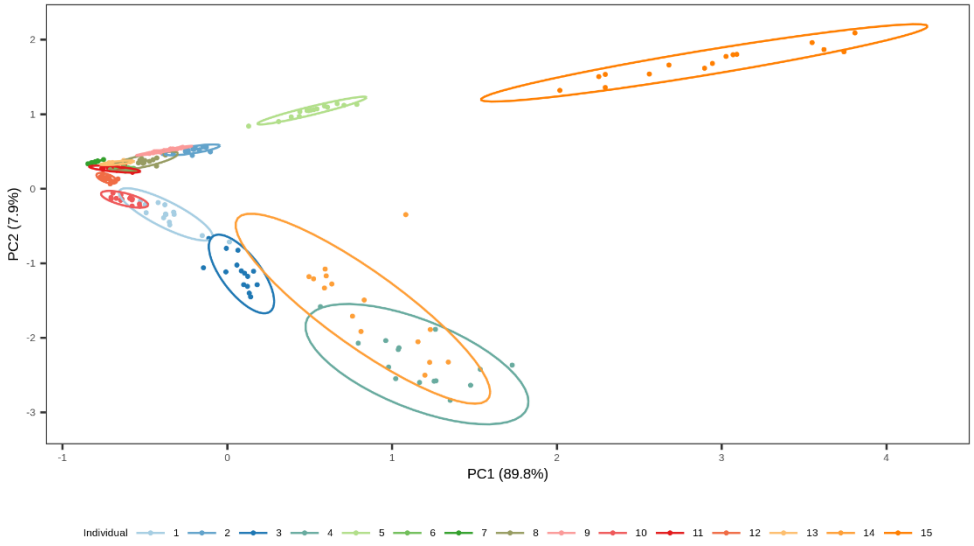


Figure 7. Two-dimensional plot from principal component analysis (PCA) of the short-term dried saliva stains from 15 individuals. Projection of the counts of our four targeted species in the first two PCs together explain 97.7 % of the total variation. The stains cluster based on the individual which highlights the high inter-individual variation in the dataset.

With this individual-specific modelling approach, we presented an individualized solution to saliva trace timing in our dataset where the model training and testing data are obtained from the individual's reference saliva stored under the same environmental conditions for a specific time period (Figure 1E). We hypothesize that for certain indoor crimes, where environmental conditions are rather stable, various parameters (temperature, humidity, etc.) could be measured at the crime scene when the stain is collected and applied to the reference saliva stains used to generate the model underlying data with flexibility on the time

window. Consequently, in our experiments we applied the same environmental parameters to saliva samples stored up to 1 month that we used for model training and for model testing. The R^2 and BH-corrected *p* values of each individual-specific model can be found in Supplementary Table S6. Overall, the average model fit was R^2=0.752. The best model fit was obtained for individual 11 (R^2=0.921, *p*=4.80E-05), whereas for individual 1 the model barely fit (R^2=0.178, *p*=0.233).

Considering the testing stains of all 15 individuals, the average correlation between the true and predicted time since deposition was r=0.742, while the average MAE was 5 days (16.7% of the analysed time frame of 1 month). The model for individual 8 presented the lowest MAE of 3.3 days with a correlation between real and predicted values of r=0.905 (Figure 8). The model for individual 14 presented the highest MAE of 7.8 days with a correlation between real and predicted values of r=0.235 (Figure 8). The real and predicted times since deposition for each individual are summarized in Supplementary Table S5. We further investigated errors in the time since deposition prediction in the individual-specific approach (Supplementary Figure S4). There was no clear pattern relating certain individuals or time points with lower errors. The time since deposition of eleven stains (5.0%) was correctly predicted with zero days error. In the rest of the predictions, there was a similar distribution in the stains in which time since deposition was underestimated (113 stains, 50.9%) or overestimated (98 stains, 44.1%). In 81.5% of the cases (181 stains) the error of the predicted time since deposition fell within 1 week (up to +/- 7 days error). From those, more than half of the samples fell within three days (+/- 3 days error). More precisely, 16.7% (37 stains) were predicted with +/- 1 day, 17% (38 stains) with +/- 2 days and 9% (20 stains) with +/- 3 days error.

Additionally, we estimated the time since deposition of short-term dried saliva stains from six individuals that we collected and exposed to indoor conditions seven months after the first collection time point. Considering the tested stains of all six individuals, the average MAE was 8.8 days (29.3% of the analysed time frame of 1 month). The one-variable predictor model for individual 1 presented the lowest MAE of 3.9 days. The four-variable predictor model for individual 2 presented the highest MAE of 16.9 days. The MAE values of all the individual-specific models of one, two, three and four species as predictors for time since deposition can be found in Supplementary Table S7.

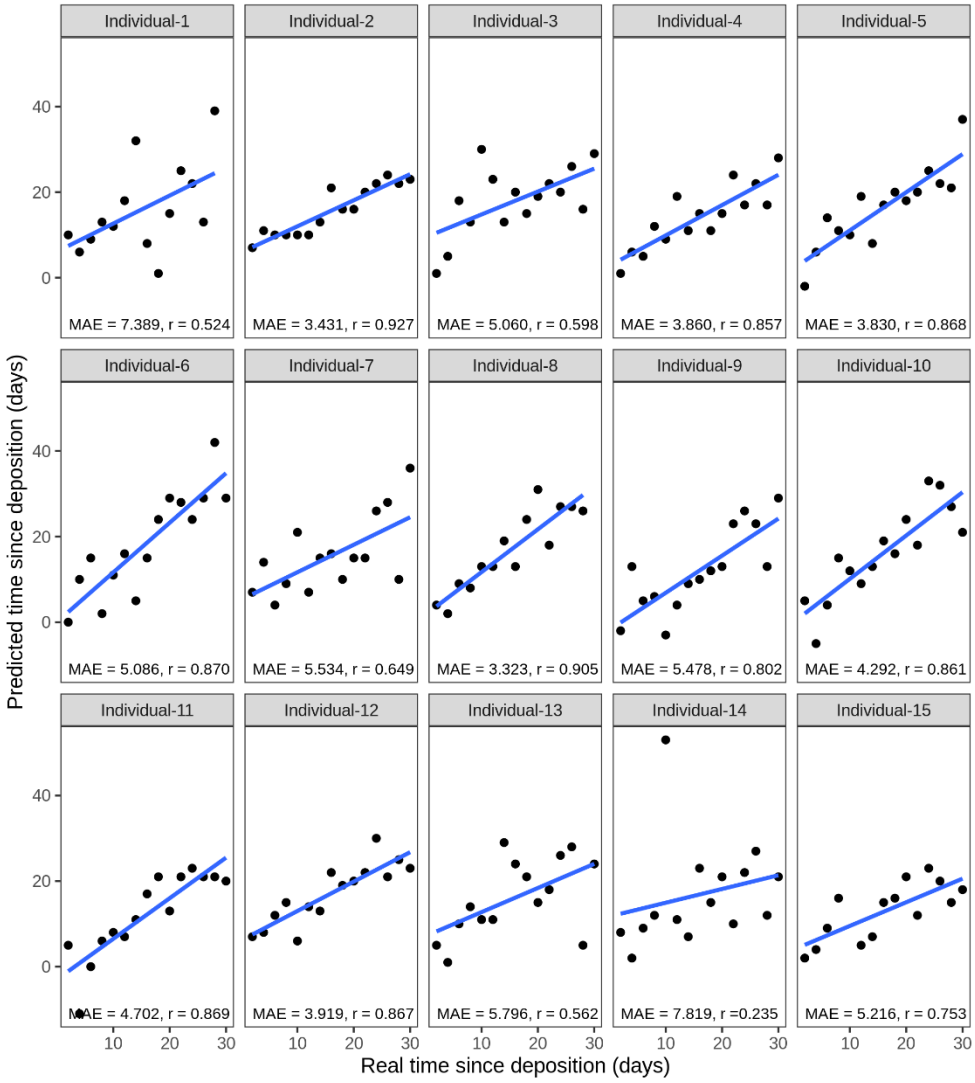


Figure 8. Individual-specific model performance for the prediction of the time since deposition of saliva stains using data of the short-term (up to 1 month) stored saliva stains based on four bacterial species from 15 individuals. Data from sample duplicates 2 were used for model testing while data from sample duplicates 1 were used for model building. The mean absolute error (MAE) measures the discrepancies between the real and the predicted time since deposition values. The correlation between real and predicted values is indicated with r.

DISCUSSION

Knowledge on the time when a human biological stain was left at a crime scene – also known as time since deposition – can be of great forensic value in assessing alibis of known suspects, searching for suspects, selecting the stains with the highest informative value for further analysis, helping in missing person cases when the time gone missing is unknown and in estimating PMI in scenes involving a corpse or parts thereof. In this study, we evaluated for the first time a microbial DNA-based approach for estimating the time since deposition of dried saliva stains exposed to indoor conditions.

First, we identified the most abundant and most frequent bacterial species in human saliva samples from two publicly available 16S rRNA gene sequencing datasets (34, 35). On the one hand, we looked for abundant species to ensure their detection in forensic-type saliva samples, which are often as small as a few microliters (µL) in volume. On the other hand, we also looked for frequent species, meaning that they are more likely to be present in the general population; hence, in the great majority of saliva stains found at crime scenes. For our final choice, we focused on species that are both abundant and frequent at the same time since taxa 'exclusive' to an individual often account for a significant percentage of an individual's microbiome profile (27-29, 51-55). The top 15 bacterial species identified as both most abundant and most frequent in our samples belonged to genera previously reported to be predominant taxa in saliva and part of the so-called 'core' oral microbiome (27-29, 51-55). Our observations agreed with these previous studies, reporting that the saliva microbiome is dominated by just a few taxa, while most of the taxa detected per individual are rare.

Secondly, we analysed microbiome profiles of dried saliva stains exposed to our laboratory environment long-term (up to 1 year) and focused our time-dependent analysis on the previously selected 15 most abundant and frequent bacterial species. Based on differential abundance analysis, we identified four species the abundance of which significantly changed over time since deposition: *Fusobacterium periodonticum, Haemophilus parainfluenzae, Veillonella dispar* and *Veillonella parvula.* It is noted that three of the species are obligate anaerobes (*F. periodonticum, V. dispar, V. parvula*), while the fourth one is a facultative anaerobe (*H. parainfluenzae*). Theoretically, obligate anaerobes might be depleted upon exposure to an oxygen-rich environment such as our laboratory, as a previously study also reported this for the case of *Veillonella* genus (32). We hypothesize that the selected bacterial species co-aggregate *ex vivo* with other saliva microbes forming biofilms and having access to nutrients and molecules for survival and protection. The removal of oxygen by aerobic and facultative anaerobes could create 'pockets' of anoxia that support the growth of obligate anaerobes, similarly to what happens in the human dental plaque both *in vivo* (56) and *ex vivo* (57).

Also, obligate anaerobic organisms can metabolize oxygen and produce protective enzymes in response to oxidative stress (58, 59). Additionally, the bacterial *ex vivo* co-aggregates could be advantageous for a more efficient utilization of nutrients and molecules found in saliva, as previously reported (60-62). The method we employed reported only relative abundance changes of the four species over time, though different scenarios could explain the direction of these changes. For example, the increase in the selected reference log ratio balance from 7 to 365 days since deposition could be explained by one of the following five scenarios of absolute abundance changes: i) the numerator's taxa increased on

average; ii) the denominator's taxa decreased on average; iii) a combination of the previous two happened; iv) both the numerator's and denominator's taxa increased, but the numerator's taxa increased more compared to the denominator's taxa; v) the numerator's and denominator's taxa both decreased, but the denominator's taxa decreased more compared to the numerator's taxa.

In line with our findings, other published studies have also indicated time-dependent microbiome changes in dried saliva samples exposed to indoor conditions. Though not the main aim of their study, Salzmann *et al.* (32) analysed the microbial communities in both fresh and dried saliva samples exposed to their laboratory environment for both five and nine months. Via differential abundance analysis using DESeq2 (63) they showed that four facultative and obligate anaerobic bacteria were significantly depleted upon exposure to indoor conditions: *Actinomyces, Staphylococcus, Veillonella* and an unclassified genus from the *Leptotrichiaceae* family. However, no definitive conclusions could be drawn due to the small sample size in the study (n=4, two fresh and two dried saliva stains). Moreover, the analysis employed for the differential abundance testing (DESeq2 (63)) was originally developed for RNA-Seq data and requires further development for general use on microbiome data (64). While the authors did not report their results at the species level, looking at the genus level two of the four bacterial species we selected as time-dependent markers belong to the genus *Veillonella* and one of the 15 species identified as most abundant and most frequent belong to the *Actinomyces* genus and another to the *Leptotrichiaceae* family.

We are aware of the limitation that only two individuals were studied in the differential abundance analysis of the long-term dried saliva stains. This was mainly due to technical and financial restrictions of our NGS analysis, but we consider it sufficient for a proof of principle study. In the future, analysis of more individuals for targeted time periods will add on our results and perhaps reveal additional promising biomarkers. Also, in our study we only analysed the four most promising bacterial species via a targeted analysis. However, based on the promising reported follow-up results, future work could focus on the analysis of all the top 15 most abundant and frequent species as identified from the publicly available adult human saliva 16S gene NGS datasets. It could be expected that the other eleven bacterial species similarly participate in an *ex vivo* microbial consortium (65). Hence, potential time-dependent changes in their abundance might additionally serve as powerful estimators of the time since deposition of saliva stains.

Based on the four differentially abundant and frequent bacterial species, we developed a 4-plex qPCR assay to test the forensic applicability in dried saliva stains exposed to indoor short-term conditions up to 1 month. A prerequisite for applying such qPCR assay in a stain would be to confirm its body fluid source being saliva. For this, it is possible to apply another microbiome-based approach for the conclusive identification of saliva stains, as we recently showed in a previous study (19). The four species targeted with the 4-plex qPCR assay were detected in the fresh saliva (t0) of all 15 analysed individuals, confirming that they are abundant and frequent enough for forensic use. Interestingly, the qPCR-reported cell counts of our four targeted bacterial species increased with time since deposition for the majority of the analysed individuals. However, high inter-individual differences were observed in the variation explained by time in the species abundance. Nevertheless, there was no clear relationship between a higher explained variation and initial species abundance (t0). We are not very surprised by this, as this variation between individuals could be explained by bacterial

interactions. For example, other bacteria taxa present in the sample might have interacted with our targeted species in different ways (i.e. mutualism, syntrophism, commensalism, proto-cooperation, antagonism, competition, parasitism, predation). Also, the presence and abundance of certain nutrients and molecules in the saliva at the time since deposition could favour or impair some of these interactions. Finally, while qPCR is a well-established and suitable method for the purpose of this study, it will be possible to transfer the protocol to newer, more sensitive methods, such as digital droplet PCR (ddPCR), in the future.

For estimating the time since deposition of our saliva stains, we first built a generalized prediction model. By this, we attempted to estimate the time when an 'unknown' test stain was deposited based on a previously established model. However, the high inter-individual variation in our dataset limited the possibility of implementing such a model. Though time-dependent changes in the four targeted bacteria occurred in the short-term dried saliva stains (up to 1 month) from all the 15 analysed individuals, the magnitude and evolution through time of those changes were very specific to each individual. As a result, the estimation of the time since deposition of an individual's stains based on a model trained with stains from other individuals was not feasible with our dataset. In order to further explore the possibility of a generalized model, future studies might employ a much bigger sample size; not only regarding the number of individuals but also regarding the tested environmental conditions. This could allow to better understand whether a broad range of inter-individual and different environmental effects can be captured during model building. These effects, together with the bacterial-based time-dependent information, might result in a generalized model being applicable to unknown stains originated from any random individual in the population and exposed to different environmental conditions.

Based on the limitation of applying a generalized model in our dataset, we decided to build individual-specific models to predict the time since deposition of these short-term dried saliva stains. For each individual, we employed the first sample duplicate set for model training and the second for model testing. With this, we aimed to mimic forensic investigations with our dataset, where the estimation of the time since deposition of one or various stains is possible based on a model built from a reference set of dried saliva stains from the same individual exposed to adequate storage conditions and time frames (e.g. indoor storage conditions for a particular period of time) (Figure 1E). We acknowledge that different factors might affect the model building and accuracy; particularly, the environment the stain is exposed to; i.e. temperature, relative humidity, ambient light, availability of nutrients and molecules, the template bacterial community present in the stain at the time of deposition (which seems to be affected by individual characteristics) and the time since deposition itself. All these factors are expected to influence the time-dependent changes in the bacterial biomarkers, which could be accounted for by an individualized solution as we do in our dataset in more dedicated future studies. The reported mean absolute error (MAE) values in these individual-specific models further highlighted the observed inter-individual differences, ranging from 3.3 to 7.8 days (average of 5 days). In contrast to RNA-based studies, we did not observe increased prediction errors with increased storage times. There is only one published RNA-based study so far on time since deposition estimation of dried saliva stains we can compare our results with (2). That study reported a slightly lower MAE value than ours (3.5 *vs.* 5 days); however, sample size was much larger in our study (222 samples from 15 different individuals *vs.* 5 samples). A different study analysing dried blood stains reported a time estimation error

of 2-4 weeks for stains exposed less than six months, which we improved in our study for dried saliva stains (6).

We observed an increase in the MAE values in the short-term dried saliva stains collected seven months later, ranging from 3.9 to 16.9 days (average of 8.8 days) compared to a range of 3.3 to 7.8 days (average of 5 days) from the first sample collection. In the same way that different environmental factors can affect PMI of human cadavers (11, 66), the time since deposition estimation of dried body fluid stains could be affected by variations in the exposure conditions. We hypothesize that one factor affecting our predictions might be the season, since the first round of stains were collected and exposed during spring and the second round during autumn with the consequent differences in average temperatures (slightly lower in autumn) and daylight duration (8 *vs.*11 hours), even in indoor conditions (since stains were placed four meters away from a window). It could also happen that the *in vivo* abundance of some or the four selected bacterial species varied between the two saliva collection time points, seven months apart from each other, which could happen due to changes in an individual's health status or lifestyle habits affecting the subsequent time-dependent bacterial abundance changes in the prepared stains. However, given the more extensive previous datasets demonstrating time-wise microbiome stability in saliva *in vivo* (28-31) and the absence of available information on potential changes in our volunteers' health status and lifestyle habits, our preliminary data need to be considered with care and larger data evidence needs to be established in the future.

More dedicated future research might focus on increasing the reliability of the individual-specific prediction models for example, by increasing the model training sample size. For instance, instead of preparing and collecting dried stains every two days, shorter time frames could be analysed (i.e. daily or every a few hours), which may better reflect the rapid division rates of bacterial cells when the conditions are favourable. A bigger training set would also mean the possibility of investigating more complex prediction models that can capture other time-dependent changes than linear. Additionally, DNA-based analysis can be reliable in time-dependent bacterial 'growing' patterns but might present limitations in 'decaying' patterns since living and dead cells cannot be distinguished. An alternative could be bacterial RNA-based analysis in which only the live bacterial fraction is analysed or a combined approach of bacterial DNA/RNA analysis. Before our proposed approach is considered for future forensic applications various forensic developmental and implementation criteria will need to be met. For instance, future research should deal with the suitability of the approach under different scenarios, such as sample volume and sample substrate (e.g. cigarette butts, chewing gums, food utensils, fabrics) as well as environmental factors (e.g. average temperature, percentage of air humidity, ambient daylight hours).

CONCLUSION

To summarize, this research shows from a forensic standpoint and to the best of our knowledge, the first time using commensal human bacteria absolute abundance changes to estimate the time since deposition of dried saliva stains. We focused on abundant and frequent commensal bacterial species of the saliva of human adults, aiming for applications in forensic-type saliva stains from the general population. We observed that, though high inter-

individual variation was found, the four selected bacterial species present a high and significant correlation between their abundance in saliva stains and the time since deposition of saliva stains. We present an individual-specific solution on estimating stain time since deposition in our dataset. We hypothesize that this might be forensically feasible when a saliva reference sample is used to produce the prediction model underlying data. This would be based on samples stored for different time intervals under specific environmental conditions that resemble those to which the crime scene stain was exposed, such as in cases of indoor crimes, though more dedicated future research is needed to confirm our hypothesis. While we consider 1 month as a forensically realistic time frame between stain deposition at a crime scene and reference sample collection from potential suspects for the majority of forensic cases, shorter or longer time spans could be studied in more detail to analyse the extended potential forensic utility of our approach. Our proof-of-principle study suggests that like in other forensic applications, the human microbiota offers high promises on the future forensic application of our findings for estimating the time since deposition of a saliva stain at a crime scene. In the future this novel approach may be expanded to other forensically relevant human stains containing microbial DNA. Before such microbiome-based stain timing can be further considered for practical forensic applications, further microbiome research is needed to better understand and model all the factors contributing to the bacterial time-dependent changes. Additionally, various forensic developmental and forensic implementation criteria will need to be met via future more dedicated studies.

ACKNOWLEDGEMETNS

We are grateful to all volunteers for their kind donation of biological samples and to the authors of ACS CPS-II, PLCO and AGP for making their data publicly available. We thank Benjamin Planterose and Diego Montiel González (Erasmus MC) for their valuable guidance in statistical analysis, as well as Eric M. J. Bindels (Erasmus MC) for providing us with access and help to use the Illumina® MiSeq platform. Finally, we also thank Qiagen for supplying the QIAseq™ 16S/ITS Panel Kit as a prize to CDL for having been selected amongst the top 10 finalists in the 2018 Qiagen Microbiome Award (PhD student category).

AUTHOR'S CONTRIBUTIONS

AV conceptualized this work. CDL and AV designed the study with contributions by MK. CDL performed all experiments and data analyses. CDL and AV interpreted the data with contributions by MK. MK provided resources. All authors wrote the manuscript and approved its final version.

FUNDING

The work of all authors is supported by Erasmus MC, University Medical Center Rotterdam. AV was additionally supported with an EUR fellowship by Erasmus University Rotterdam.

DATA AVAILABILITY STATEMENT

The raw saliva NGS data used in this study are publicly available via EMBL-EBI (https://www.ebi.ac.uk/) while metadata can also be accessed via NCBI (https://www.ncbi.nlm.nih.gov/) with accession numbers PRJNA434300, PRJNA434312 and PRJEB11419. The processed NGS and qPCR data produced and used in this study to perform our analysis and derive our conclusions are included in Supplementary Data Sheet 1. This file can be found in the online version of the manuscript (doi:10.3389/fmicb.2021.647933).

REFERENCES

1. Butler JM. 2004. Short tandem repeat analysis for human identity testing. Curr Protoc Hum Genet Chapter 14:Unit 14.8.
2. Asaghiar F, Williams GA. 2020. Evaluating the use of hypoxia sensitive markers for body fluid stain age prediction. Sci Justice 60:547-554.
3. Weinbrecht KD FJ, Payton ME, Allen RW. 2017. Time-dependent loss of mRNA transcripts from forensic stains. Dovepress 7:1-12.
4. Amany TM KS, Ali AH, Awad A. 2018. Validation of mRNA and microRNA profiling as tools in qPCR for estimation of the age of bloodstains. Life Sci 15.
5. Anderson SE, Hobbs GR, Bishop CP. 2011. Multivariate analysis for estimating the age of a bloodstain. J Forensic Sci 56:186-93.
6. Fu J, Allen RW. 2019. A method to estimate the age of bloodstains using quantitative PCR. Forensic Sci Int Genet 39:103-108.
7. Bauer M, Polzin S, Patzelt D. 2003. Quantification of RNA degradation by semi-quantitative duplex and competitive RT-PCR: a possible indicator of the age of bloodstains? Forensic Sci Int 138:94-103.
8. Alshehhi S, Haddrill PR. 2019. Estimating time since deposition using quantification of RNA degradation in body fluid-specific markers. Forensic Sci Int 298:58-63.
9. Bauer M. 2007. RNA in forensic science. Forensic Sci Int Genet 1:69-74.
10. Watanabe K, Akutsu T, Takamura A, Sakurada K. 2017. Practical evaluation of an RNA-based saliva identification method. Sci Justice 57:404-408.
11. Dash HR, Das S. 2020. Thanatomicrobiome and epinecrotic community signatures for estimation of post-mortem time interval in human cadaver. Appl Microbiol Biotechnol 104:9497-9512.
12. Pechal JL, Schmidt CJ, Jordan HR, Benbow ME. 2018. A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. Sci Rep 8:5724.
13. Adserias-Garriga J, Quijada NM, Hernandez M, Rodríguez Lázaro D, Steadman D, Garcia-Gil LJ. 2017. Dynamics of the oral microbiota as a tool to estimate time since death. Mol Oral Microbiol 32:511-516.
14. Hyde ER, Haarmann DP, Lynne AM, Bucheli SR, Petrosino JF. 2013. The living dead: bacterial community structure of a cadaver at the onset and end of the bloat stage of decomposition. PLoS One 8:e77733.
15. Pechal JL, Schmidt CJ, Jordan HR, Benbow ME. 2017. Frozen: Thawing and its effect on the postmortem microbiome in two pediatric cases. J Forensic Sci 62:1399-1405.
16. Yang J, Tsukimi T, Yoshikawa M, Suzuki K, Takeda T, Tomita M, Fukuda S. 2019. Cutibacterium acnes (Propionibacterium acnes) 16S rRNA genotyping of microbial samples from possessions contributes to owner identification. mSystems 4:e00594-19.
17. Schmedes SE, Woerner AE, Novroski NMM, Wendt FR, King JL, Stephens KM, Budowle B. Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. Forensic Sci Int Genet 32:50-61.
18. Schmedes SE, Woerner AE, Budowle B. 2017. Forensic human identification using skin microbiomes. Appl Environ Microbiol 83:e01672-17.
19. Díez López C, Vidaki A, Ralf A, Montiel González D, Radjabzadeh D, Kraaij R, Uitterlinden AG, Haas C, Lao O, Kayser M. 2019. Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. Forensic Sci Int Genet 41:72-82.
20. Díez López C, Montiel González D, Haas C, Vidaki A, Kayser M. 2020. Microbiome-based body site of origin classification of forensically relevant blood traces. Forensic Sci Int Genet:102280.
21. Haas C, Hanson E, Anjos MJ, Ballantyne KN, Banemann R, Bhoelai B, Borges E, Carvalho M, Courts C, De Cock G, Drobnic K, Dotsch M, Fleming R, Franchi C, Gomes I, Hadzic G, Harbison SA, Harteveld J, Hjort B, Hollard C, Hoff-Olsen P, Huls C, Keyser C, Maronas O, McCallum N, Moore D, Morling N, Niederstatter H, Noel F, Parson W, Phillips C, Popielarz C, Roeder AD, Salvaderi L, Sauer E, Schneider PM, Shanthan G, Court DS, Turanska M, van Oorschot RA, Vennemann M, Vidaki A, Zatkalikova L, Ballantyne J. 2014. RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: results of a fourth and fifth collaborative EDNAP exercise. Forensic Sci Int Genet 8:203-12.
22. Holtkotter H, Beyer V, Schwender K, Glaub A, Johann KS, Schrenkamp M, Sibbing U, Banken S, Wiegand P, Pfeiffer H, Dennany L, Vennemann M, Vennemann M, Consortium E-N. 2017. Independent validation of body fluid-specific CpG markers and construction of a robust multiplex assay. Forensic Sci Int Genet 29:261-268.
23. Ingold S, Dorum G, Hanson E, Berti A, Branicki W, Brito P, Elsmore P, Gettings KB, Giangasparo F, Gross TE, Hansen S, Hanssen EN, Kampmann ML, Kayser M, Laurent FX, Morling N, Mosquera-Miguel A, Parson W, Phillips C, Porto MJ, Pospiech E, Roeder AD, Schneider PM, Schulze Johann K, Steffen CR,

Syndercombe-Court D, Trautmann M, van den Berge M, van der Gaag KJ, Vannier J, Verdoliva V, Vidaki A, Xavier C, Ballantyne J, Haas C. 2018. Body fluid identification using a targeted mRNA massively parallel sequencing approach - results of a EUROFORGEN/EDNAP collaborative exercise. Forensic Sci Int Genet 34:105-115.

24. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. 2018. New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract. mSystems 3:e00187-18.
25. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng QD, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Brooks JP, Buck GA, Buhay CJ, Busam DA, Campbell JL, et al. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207-214.
26. Curtis MA, Zenobia C, Darveau RP. 2011. The relationship of the oral microbiotia to periodontal health and disease. Cell Host Microbe 10:302-306.
27. Zaura E, Keijser BJF, Huse SM, Crielaard W. 2009. Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9:259.
28. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694-1697.
29. Lazarevic V, Whiteson K, Hernandez D, François P, Schrenzel J. 2010. Study of inter-and intra-individual variations in the salivary microbiota. BMC Genomics 11:523.
30. Zhu X, Wang S, Gu Y, Li X, Yan H, Yan H, Miyoshi S-i, Shi L. 2012. Possible variation of the human oral bacterial community after wearing removable partial dentures by DGGE. World J Microbiol Biotechnol 28:2229-2236.
31. Stahringer SS, Clemente JC, Corley RP, Hewitt J, Knights D, Walters WA, Knight R, Krauter KS. 2012. Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Res 22:2146-2152.
32. Salzmann AP, Russo G, Aluri S, Haas C. 2019. Transcription and microbial profiling of body fluids using a massively parallel sequencing approach. Forensic Sci Int Genet:102149.
33. Dobay A, Haas C, Fucile G, Downey N, Morrison HG, Kratzer A, Arora N. 2019. Microbiome-based body fluid identification of samples exposed to indoor conditions. Forensic Sci Int Genet 40:105-113.
34. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. ISME J 10:2435-2446.
35. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, et al. 2018. American Gut: an open platform for citizen science microbiome research. mSystems 3:e00031-18.
36. Burcham ZM, Garneau NL, Comstock SS, Tucker RM, Knight R, Metcalf JL, Genetics of Taste Lab Citizen S. 2020. Patterns of oral microbiota diversity in adults and children: A crowdsourced population study. Sci Rep 10:2133.
37. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 17:3.
38. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581-583.
39. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590-D596.
40. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vázquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. 2017. Balance trees reveal microbial niche differentiation. mSystems 2:e00162-16.
41. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet

C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnology 37:852-857.

42. Mashima I, Theodorea CF, Thaweboon B, Thaweboon S, Nakazawa F. 2016. Identification of Veillonella species in the tongue biofilm by using a novel one-step polymerase chain reaction method. PLoS One 11:e0157516.
43. Vallone PM, Butler JM. 2004. AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques 37:226-31.
44. Kubista M, Andrade JM, Bengtsson M, Forootan A, Jonák J, Lind K, Sindelka R, Sjöback R, Sjögreen B, Strömbom L. 2006. The real-time polymerase chain reaction. Mol Aspects Med 27:95-125.
45. Ruijter JM, Villalba AR, Hellemans J, Untergasser A, van den Hoff MJB. 2015. Removal of between-run variation in a multi-plate qPCR experiment. Biomol Detect Quantif 5:10-14.
46. Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. J Stat Softw 67.
47. Zhao JH. 2007. gap: Genetic analysis package. J Stat Softw 23:1-18.
48. Hadley W. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer Publishing Company.
49. Liaw A, Wiener M. 2002. Classification and regression by randomForest. R news 2:18-22.
50. Hamner B, Frasco M, LeDell E. 2012. Metrics: Evaluation metrics for machine learning. R package version 01 1.
51. Nasidze I, Li J, Quinque D, Tang K, Stoneking M. 2009. Global diversity in the human salivary microbiome. Genome Res 19:636-643.
52. Li J, Quinque D, Horz HP, Li MK, Rzhetskaya M, Raff JA, Hayes MG, Stoneking M. 2014. Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. BMC Microbiol 14:316.
53. Huse SM, Ye YZ, Zhou YJ, Fodor AA. 2012. A core human microbiome as viewed through 16S rRNA sequence clusters. PLoS One 7:e34242.
54. Li K, Bihan M, Methé BA. 2013. Analyses of the stability and core taxonomic memberships of the human microbiome. PLoS One 8:e63139.
55. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, Neufeld JD, Beiko RG, Senadheera DB. 2017. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. NPJ Biofilms Microbiomes 3:1-7.
56. Schaechter M. 2009. Encyclopedia of microbiology. Academic Press.
57. Diaz PI, Zilm PS, Rogers AH. 2002. Fusobacterium nucleatum supports the growth of Porphyromonas gingivalis in oxygenated and carbon-dioxide-depleted environments. Microbiology 148:467-472.
58. Marquis RE. 1995. Oxygen metabolism, oxidative stress and acid-base physiology of dental plaque biofilms. J Ind Microbiol 15:198-207.
59. Jabłońska J, Tawfik DS. 2019. The number and type of oxygen-utilizing enzymes indicates aerobic vs. anaerobic phenotype. Free Radic Biol Med 140:84-92.
60. Periasamy S, Kolenbrander PE. 2009. Aggregatibacter actinomycetemcomitans builds mutualistic biofilm communities with Fusobacterium nucleatum and Veillonella species in saliva. Infect Immun 77:3542-3551.
61. Kuramitsu HK, Ellen RP. 2000. Oral bacterial ecology: the molecular basis. Horizon Scientific.
62. Bradshaw DJ, Homer KA, Marsh PD, Beighton D. 1994. Metabolic cooperation in oral microbial communities during growth on mucin. Microbiology 140:3407-3412.
63. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15.
64. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5:27.
65. Kolenbrander PE. 2011. Multispecies communities: interspecies interactions influence growth on saliva as sole nutritional source. Int J Oral Sci 3:49-54.
66. Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, Knight R, Metcalf JL. 2018. Microbiome data accurately predicts the postmortem interval using random forest regression models. Genes (Basel) 9:104.

SUPPLEMENTARY MATERIALS

Supplementary Table S1. Information on sex, age, ethnicity and sample collection of saliva donors. Asterisks (*) indicate that the individual is the same with a time span of two years between the two saliva donation time points.

Saliva samples	Individual ID	Sex	Age (years)	Ethnicity	Second saliva donation 7 months after first donation time point
Long-term	1*	F	27	White	
	2	M	21	White	
Short-term	1	M	32	Admixed (Black-White)	X
	2*	F	29	White	X
	3	F	25	White	X
	4	M	31	Hispanic	X
	5	M	25	Asian	
	6	F	28	White	
	7	F	27	White	
	8	M	25	White	X
	9	F	33	White	X
	10	F	23	White	
	11	F	26	White	
	12	F	25	White	
	13	F	24	White	
	14	F	23	White	
	15	F	25	White	

Supplementary Table S2. Primers, probes and gBlocks used in the 4-plex qPCR approach to analyse the short-term dried saliva stains. In the gBlocks, forward and reverse primer sequences are highlighted in red and probe sequences are highlighted in blue. Asterisks (*) at the 5'-end of the gBlocks indicate that additional nucleotides had to be added to the sequence for successful production as required by Integrated DNA Technologies (IDT, Coralville, IA, US).

Assay	Target gene	Amplicon length (bp)	Forward primer sequence (5'>3')	Reverse primer sequence (5'>3')	Probe primer sequence (5'>3')	Probe dye	gBlock sequence (5'>3')
Fusobacterium periodonticum	rpoB	87	TGCTGAAAAACG TAAGATAACTG	TCTGCAGGTAAT ACTCTTGAAACA	TGTCAGGAAGAC ATGGAAACAAA GGGGT	6-Carboxyfluo-rescein (6-FAM)	*CGTAGAGTGTAATGTATTAG **TGCTGAAAAACGTAAGATA** **ACTG**TAGGGGATAAGA**TGTC** **AGGAAGACATGGAAACAAA** **GGGGTTGTTTCAGAGTATT** **ACCTGCAGA**GCTATGCTAG CTTAGTCAGC
Haemophilus parainfluenzae	infB	115	TAGAACAAGAAT TACTTCAACACG	ATGGCATCTAATA AGTCATCAACAC	CGGTGGTGATGT GCAATTCGTTCC	Cyanine 5 (Cy5)	*CGTAGA**TAGAACAAGAATT** **ACTTCAACACG**AAGTGATTT CTGAGAAATT**CGGTGGTGAT** **GTGCAATTCGTTCC**TGTTTC TGCGAAGAAAGGTACC**GGT** **GTTGATGACTTATTAGATGC** **CAT**GCTATG
Veillonella dispar	rpoB	86	TCAAAGCTACAG ACCTTCACATT	GCGAATAGCGTC AATTTGTC	AAAGAAGCTCGT ACTGATGCATCT TATTGGG	Texas Red-615 (TEX-615)	*CGTAGAGTGTAATGTATTAG **TCAAAGCTACAGACCTTCA** **CATT**CAAAATGTATT**GAAAG** **AAGCTCGTACTGATGCATCT** **TATTGGGGACAAATTGACG** **CTATTCGC**GCTATGCTAGCT TAGTCAGC
Veillonella parvula	rpoB	138	CAAAGCTACAGA CCTTCACATTC	CGTAACATCTTC CGAAACTTTC	TGCAATCCGCGA CCTTTATGCT	Hexachloro-fluo-rescein (HEX)	**CAAAGCTACAGACCTTCAC** **ATTC**AAAACGTGTTGAAAGA AGCCCGTACTGATGCATCTT ACTGGGAACAAATCGA**TGCA** **ATCCGCGACCTTTATGCTCA** AATTGAAGCATTGGAGC**GAA** **AGTTTCGGAAGATGTTACG**

Supplementary Table S3. Log ratio balances, intercept and coefficient FDR *p* values for time since deposition and individual person in the multivariate linear model using gneiss analysis in the long-term dried saliva stains exposed to indoor conditions for up to 1 year. Balances significantly affected by the covariates time since deposition and individual person were determined with a *p* value cutoff at 0.05. Log ratio balances significant for time since deposition are highlighted in gray.

Log ratio balances	**Intercept**	***p* value; Time since deposition**	***p* value; Individual person [T.individual-2]**
y0	0.005	0.558	0.387
y1	0.138	0.475	0.018
y2	0.248	0.414	0.294
y3	0.462	0.462	0.462
y4	0.690	0.008	0.390
y5	0.911	0.004	0.911
y6	0.645	0.645	0.008
y7	0.001	0.022	0.022
y8	0.000	0.397	0.078
y9	0.836	0.958	0.836
y10	0.040	0.316	0.891
y11	0.000	0.416	0.718
y12	0.628	0.766	0.142

Supplementary Table S4. Univariate linear regression results for each of the four species cell count and the time since deposition (lm(C~T)) for each individual and sample duplicate set. Results include adjusted R^2 values, Benjamini-Hochberg (BH) corrected *p* values and significance testing. Significance codes: 0.001 ***, 0.01 ** and 0.05 *.

	Individual	1		2		3		4		5	
	Sample duplicate	1	2	1	2	1	2	1	2	1	2
Fusobacterium periodonticum	**R^2 adj.**	0.297	0.320	0.825	0.584	0.516	0.341	0.679	0.598	0.663	0.522
	***P* value (BH FDR)**	2.461E-04	1.315E-04	2.596E-14	8.541E-09	1.252E-07	3.950E-05	5.239E-11	2.095E-09	1.062E-10	6.919E-08
	Significance	***	***	***	***	***	***	***	***	***	***
Haemophilus parainfluenzae	**R^2 adj.**	0.120	0.154	0.127	-0.022	-0.023	0.075	0.464	0.390	0.145	-0.010
	***p* value (BH FDR)**	0.019	0.009	0.014	0.783	0.944	0.048	6.701E-07	8.290E-06	0.008	0.509
	Significance	*	**	*			*	***	***	**	
Veillonella dispar	**R^2 adj.**	0.248	0.163	0.367	0.156	0.092	0.083	0.118	0.231	0.230	0.160
	***p* value (BH FDR)**	9.562E-04	0.007	2.189E-05	0.008	0.031	0.039	0.016	9.735E-04	9.827E-04	0.006
	Significance	***	**	***	**	*	*	*	***	***	**
Veillonella parvula	**R^2 adj.**	0.184	0.150	0.238	0.255	-0.022	-0.024	0.162	0.229	0.001	0.062
	***p* value (BH FDR)**	0.004	0.009	9.836E-04	6.850E-04	0.818	0.944	0.006	9.836E-04	0.374	0.092
	Significance	**	**	***	***			**	***		

	Individual	6		7		8		9		10		11	
	Sample duplicate	1	2	1	2	1	2	1	2	1	2	1	2
Fusobacterium periodonticum	**R^2 adj.**	0.223	0.451	0.569	0.433	0.451	0.601	0.231	0.130	0.457	0.291	0.178	0.366
	***p* value (BH FDR)**	0.001	1.016E-06	7.670E-09	1.905E-06	1.016E-06	6.397E-09	9.735E-04	0.012	8.626E-07	1.790E-04	0.004	1.873E-05
	Significance	**	***	***	***	***	***	***	*	***	***	**	***
Haemophilus parainfluenzae	**R^2 adj.**	-0.005	0.213	0.485	0.334	0.002	-0.010	0.052	0.193	0.616	0.575	0.616	0.631
	***p* value (BH FDR)**	0.439	0.001	2.982E-07	4.745E-05	0.360	0.504	0.089	0.003	1.082E-09	6.346E-09	1.082E-09	5.952E-10
	Significance		**	***	***				**	***	***	***	***
Veillonella dispar	**R^2 adj.**	0.351	0.435	0.610	0.343	0.229	0.469	-0.019	-0.020	0.168	0.122	0.612	0.607
	***p* value (BH FDR)**	2.991E-05	1.770E-06	1.239E-09	3.781E-05	9.836E-04	1.261E-06	0.719	0.754	0.005	0.015	1.180E-09	1.377E-09
	Significance	***	***	***	***	***	***			**	*	***	***
Veillonella parvula	**R^2 adj.**	0.019	-0.011	0.329	0.410	NA	NA	-0.019	-0.022	0.240	0.137	0.687	0.751
	***p* value (BH FDR)**	0.222	0.527	5.56E-05	4.210E-06	NA	NA	0.729	0.873	7.763E-04	0.010	3.570E-11	3.347E-13
	Significance			***	***	NA	NA				**	***	***

Individual		12		13		14		15	
	Sample duplicate	1	2	1	2	1	2	1	2
Fusobacterium periodonticum	R^2 adj.	0.427	0.292	0.663	0.227	0.590	-0.015	0.501	0.181
	p value (BH FDR)	2.216E-06	1.735E-04	1.062E-10	0.001	3.047E-09	0.618	1.613E-07	0.004
	Significance	***	***	***	**	***		***	**
Haemophilus parainfluenzae	R^2 adj.	-0.009	-0.007	0.341	0.181	0.474	0.247	0.098	-0.019
	p value (BH FDR)	0.504	0.467	3.950E-05	0.004	4.633E-07	6.644E-04	0.027	0.714
	Significance			***	**	***	***	*	
Veillonella dispar	R^2 adj.	0.096	0.013	0.494	0.366	0.623	0.405	-0.022	0.016
	p value (BH FDR)	0.029	0.255	2.058E-07	1.876E-05	8.697E-10	4.937E-06	0.864	0.243
	Significance	*		***	***	***	***		
Veillonella parvula	R^2 adj.	0.205	0.084	0.771	0.286	0.650	0.430	0.337	0.197
	p value (BH FDR)	0.002	0.038	8.201E-14	2.071E-04	2.050E-10	2.073E-06	4.444E-05	0.003
	Significance	**	*	***	***	***	***	***	**

Supplementary Table S5. Real and predicted time since deposition values (in days) in the short-term dried saliva stains for each individual based on the generalized random forest (RF) approach and the individualized multiple linear regression (MLR) approach.

		Predicted time since deposition (days)	
Individual	**Real time since deposition (days)**	**Generalized approach (RF)**	**Individualized approach (MLR)**
1	2	9	10
1	4	11	6
1	6	15	9
1	8	22	13
1	10	21	12
1	12	21	18
1	14	19	32
1	16	18	8
1	18	11	1
1	20	20	15
1	22	20	25
1	24	20	22
1	26	19	13
1	28	18	39
2	2	13	7
2	4	10	11
2	6	9	10
2	8	11	10
2	10	11	10
2	12	11	10
2	14	11	13
2	16	11	21
2	18	12	16
2	20	13	16
2	22	11	20
2	24	11	22
2	26	11	24
2	28	11	22
2	30	11	23
3	2	19	1
3	4	11	5
3	6	13	18
3	8	13	13
3	10	14	30
3	12	12	23
3	14	13	13
3	16	17	20
3	18	13	15
3	20	14	19
3	22	14	22
3	24	13	20
3	26	15	26
3	28	19	16
3	30	16	29
4	2	21	1
4	4	20	6
4	6	20	5
4	8	21	12
4	10	20	9
4	12	20	19
4	14	20	11
4	16	20	15
4	18	20	11
4	20	20	15
4	22	20	24
4	24	20	17

4	26	20	22
4	28	20	17
4	30	20	28
5	2	16	-2
5	4	12	6
5	6	14	14
5	8	17	11
5	10	16	10
5	12	19	19
5	14	13	8
5	16	17	17
5	18	17	20
5	20	17	18
5	22	17	20
5	24	17	25
5	26	16	22
5	28	17	21
5	30	20	37
6	2	17	0
6	4	18	10
6	6	18	15
6	8	18	2
6	10	18	11
6	12	18	16
6	14	18	5
6	16	18	15
6	18	17	24
6	20	18	29
6	22	18	28
6	24	18	24
6	26	18	29
6	28	20	42
6	30	19	29
7	2	18	7
7	4	17	14
7	6	17	4
7	8	17	9
7	10	17	21
7	12	17	7
7	14	17	15
7	16	17	16
7	18	17	10
7	20	17	15
7	22	17	15
7	24	17	26
7	26	17	28
7	28	17	10
7	30	16	36
8	2	21	4
8	4	22	2
8	6	21	10
8	8	22	9
8	10	20	12
8	12	21	12
8	14	22	16
8	16	24	12
8	18	24	22
8	20	24	30
8	22	24	18
8	24	24	26
8	26	26	27
8	28	24	23
9	2	6	-2
9	4	9	13
9	6	8	5

9	8	7	6
9	10	8	-3
9	12	8	4
9	14	8	9
9	16	10	10
9	18	8	12
9	20	8	13
9	22	8	23
9	24	8	26
9	26	8	23
9	28	8	13
9	30	8	29
10	2	12	5
10	4	12	-5
10	6	14	4
10	8	18	15
10	10	14	12
10	12	18	9
10	14	19	13
10	16	17	19
10	18	18	16
10	20	19	24
10	22	19	18
10	24	18	33
10	26	16	32
10	28	18	27
10	30	18	21
11	2	19	5
11	4	17	-11
11	6	14	0
11	8	17	6
11	10	20	8
11	12	15	7
11	14	19	11
11	16	22	17
11	18	21	21
11	20	21	13
11	22	21	21
11	24	21	23
11	26	21	21
11	28	22	21
11	30	21	20
12	2	4	7
12	4	5	8
12	6	4	12
12	8	3	15
12	10	4	6
12	12	7	14
12	14	3	13
12	16	5	22
12	18	4	19
12	20	5	20
12	22	6	22
12	24	11	30
12	26	12	21
12	28	4	25
12	30	6	23
13	2	14	5
13	4	6	1
13	6	11	10
13	8	8	14
13	10	13	11
13	12	10	11
13	14	9	29
13	16	9	24
13	18	9	21

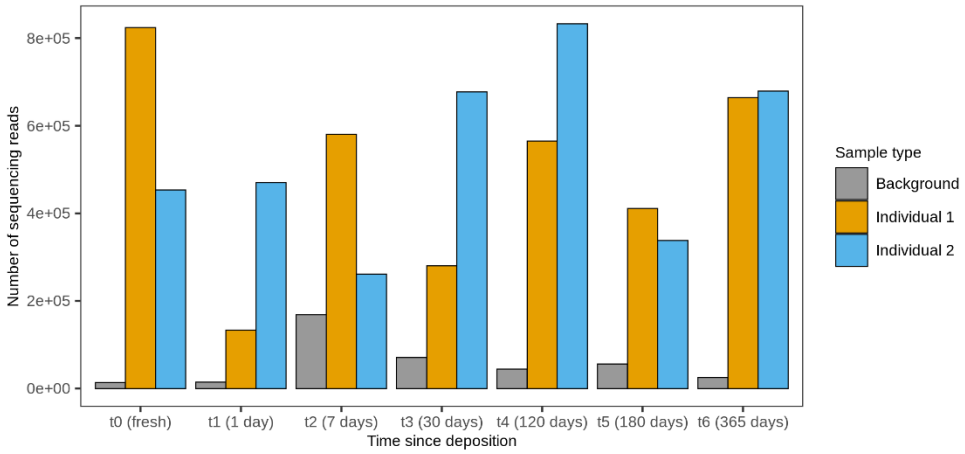
13	20	8	15
13	22	9	18
13	24	11	26
13	26	11	28
13	28	8	5
13	30	9	24
14	2	13	8
14	4	13	2
14	6	13	9
14	8	13	12
14	10	11	53
14	12	13	11
14	14	13	7
14	16	12	23
14	18	13	15
14	20	12	21
14	22	13	10
14	24	14	22
14	26	13	27
14	28	12	12
14	30	13	21
15	2	26	2
15	4	26	4
15	6	26	9
15	8	26	16
15	12	24	5
15	14	26	7
15	16	26	15
15	18	26	16
15	20	26	21
15	22	26	12
15	24	26	23
15	26	26	20
15	28	26	15
15	30	26	18

Supplementary Table S6. Individual-specific multilinear regression model results (lm(T~C1 + C2 + C3 + C4)). Results include adjusted R^2 values, Benjamini-Hochberg (BH) corrected *p* values and significance. Significance codes: 0.001 ***, 0.01 ** and 0.05 *.

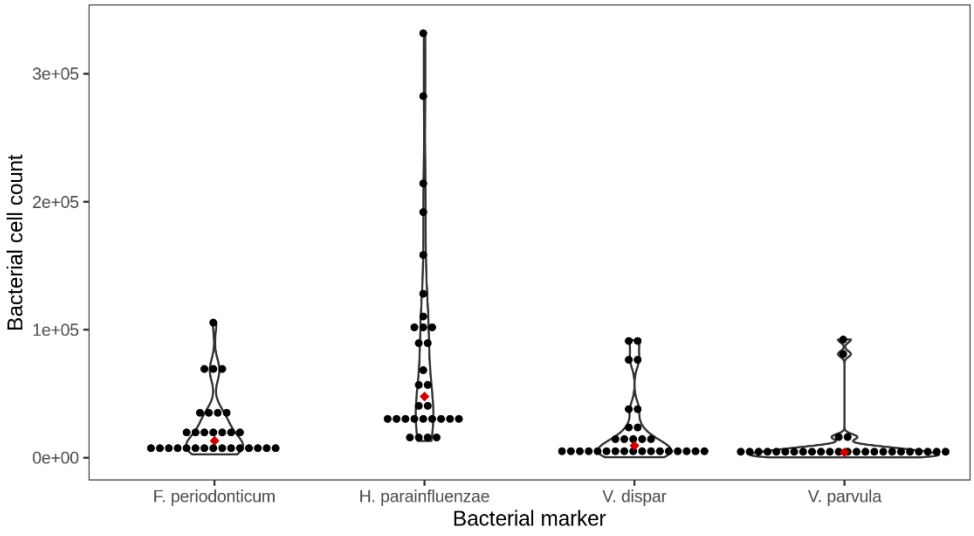
Individual	**R^2 adj.**	***p* value (BH FDR)**	**Significance**
1	0.178	0.233	
2	0.887	1.42E-04	***
3	0.663	0.004	**
4	0.722	0.002	**
5	0.850	4.04E-04	***
6	0.786	7.47E-04	***
7	0.775	8.35E-04	***
8	0.820	1.59E-04	***
9	0.636	0.006	**
10	0.849	2.01E-04	***
11	0.921	4.82E-05	***
12	0.862	1.59E-04	***
13	0.866	1.59E-04	***
14	0.678	0.004	**
15	0.785	7.47E-04	***

Supplementary Table S7. Mean aboslute error (MAE) values of the individual-specific time since deposition prediction models of one, two, three and four species as predictors. The dried saliva stains tested were collected from six individuals (Ind) and exposed short-term (up to 1 month) to indoor conditions seven months after the first collection time point. The model with the lowest MAE value is highlighted in gray for each individual.

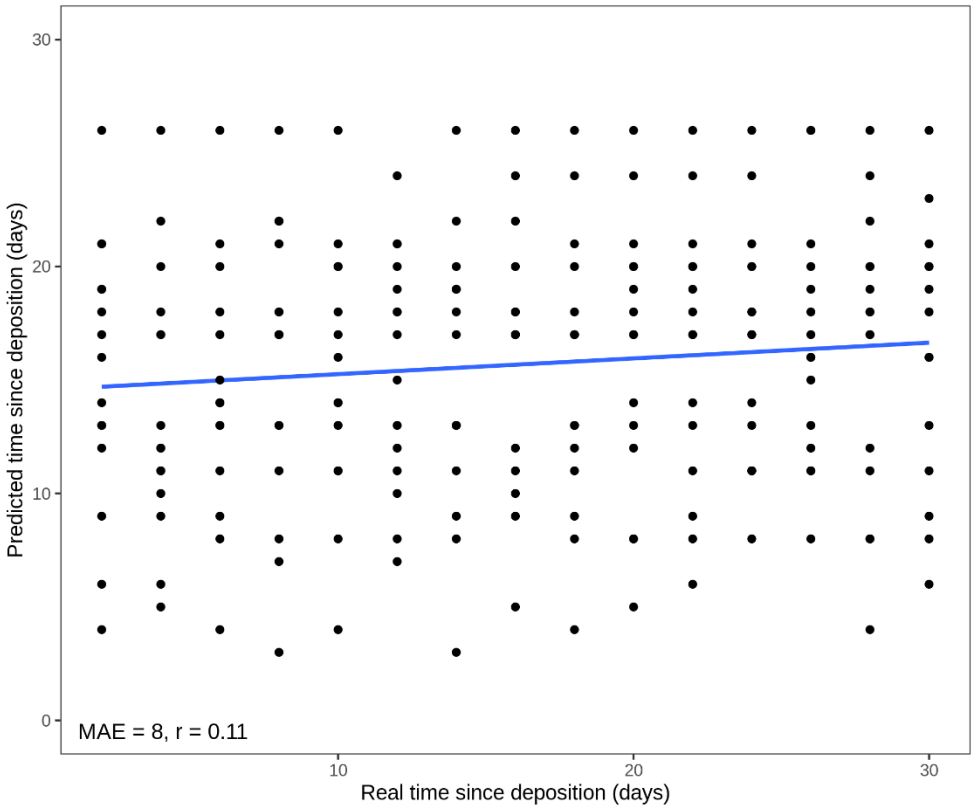
		MAE					
Number of predictors	**Predictor bacterial species**	**Ind 1**	**Ind 2**	**Ind 3**	**Ind 4**	**Ind 8**	**Ind 9**
1 predictor	*F. periodonticum*	23.7	38.9	8.9	23.8	7.2	8.1
	H. parainfluenzae	3.9	47.9	8.2	15.3	8.7	14.9
	V. dispar	12.1	37.0	13.5	13.8	33.8	33.3
	V. parvula	6.8	82.7	8.0	17.9	111.8	66.1
2 predictors	*F. periodonticum* + *H. parainfluenzae*	21.7	30.3	5.0	50.4	7.4	31
	F. periodonticum + *V. dispar*	9.8	38.6	3.7	54.7	7.1	29.3
	F. periodonticum + *V. parvula*	14.2	27.7	10.5	29.3	23.4	57.8
	H. parainfluenzae + *V. dispar*	16.3	56.5	7.4	14.5	26.1	14.2
	H. parainfluenzae + *V. parvula*	4.3	85.2	8.2	19.4	86.9	43.4
	V. dispar + *V. parvula*	10.6	79.3	8.0	18.3	104.6	65.3
3 predictors	*F. periodonticum* + *H. parainfluenzae* + *V. dispar*	8.7	29.6	5.7	65.6	7.0	18.4
	F. periodonticum + *H. parainfluenzae* + *V. parvula*	13.9	17.7	21.7	57.4	22.0	25.7
	F. periodonticum + *V . dispar* + *V. parvula*	9.6	27.3	15.3	52.2	28.4	57.9
	H. parainfluenzae + *V. dispar* + *V. parvula*	12.8	82.0	9.4	19.5	92.0	23.3
4 predictors	*F. periodonticum* + *H. parainfluenzae* + *V. dispar* + *V. parvula*	10.0	16.9	22.4	63.5	27.5	17.2



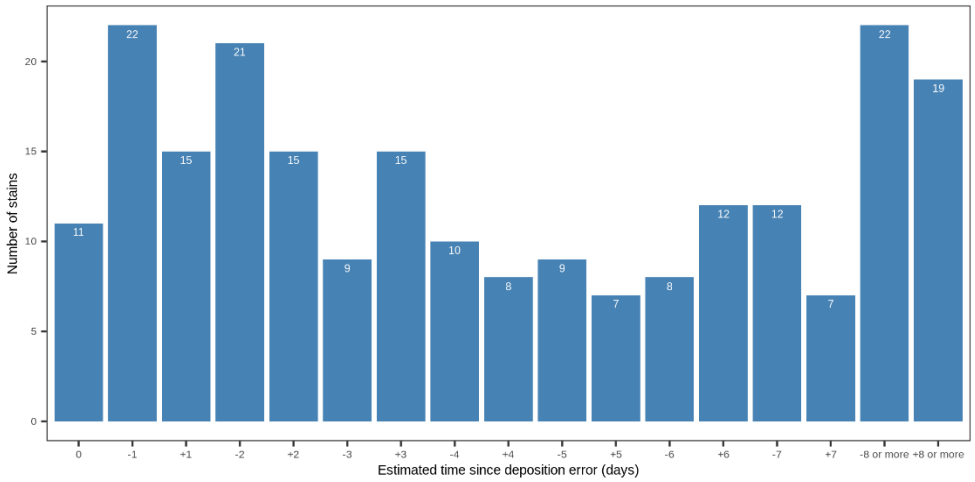
Supplementary Figure S1. Number of 16S rRNA gene (V4-V5 hypervariable regions) next-generation sequencing reads obtained from the long-term dried saliva stains which were exposed to indoor conditions up to 1 year. The bar plot includes the background blanks (empty swabs) and the saliva stains produced from the two individuals analysed.



Supplementary Figure S2. Violin plot of the bacterial cell counts for each of the four species in the fresh (t0) short-term dried saliva stains using our developed 4-plex qPCR approach. Each black dot corresponds to a sample duplicate for each of the 15 individuals analysed. Red dots indicate the average cell count for each bacterial marker considering the 15 individuals altogether. The violin plot reveals high inter-individual variation within and between the bacterial markers.



Supplementary Figure S3. Generalized random forest model performance for the prediction of the time since deposition of saliva stains using data of the short-term (up to 1 month) experiment based on four bacterial species from 15 individuals. The mean absolute error (MAE) measures the discrepancies between the real and predicted values of time since deposition. The correlation between real and predicted values is indicated with r.



Supplementary Figure S4. Error (in days) from individual-specific stain estimation time since deposition using data of the short-term (up to 1 month) stored saliva stains based on four bacterial species from 15 individuals. Data from sample duplicates 1 were used for model training, while data from sample duplicates 2 were used for model testing. Each bar represents the error in days from 0 (exact time predicted) up to +/- 8 or more days of error in the predicted time. The height of each bar corresponds to the number of stains that fall within the specified error.

PART III

Human microbiome for inferring host lifestyle from forensic traces

Chapter 5

Addressing class imbalance in microbiome-based classification using data augmentation and machine learning with smoking habit prediction from saliva as example

Celia Díez López, Diego Montiel González, Athina Vidaki, Manfred Kayser

mSystems - Submitted

ABSTRACT

Human microbiome research is currently moving from characterization and association studies to translational applications, including human trait predictions in medical research, clinical diagnostics, and other practical applications. Machine learning (ML) methods are often employed for such predictions, which face practical challenges. Class imbalance in the available microbiome data is one of them, leading to spurious prediction accuracies and limiting the classifier's generalization. Here, we introduce a novel analytical approach based on data augmentation and ML to address the class imbalance problem in microbiome-based classification, using smoking habit prediction as prominent example. We collected publicly available saliva 16S rRNA gene sequencing data and smoking habit metadata with an unavoidable class imbalance problem i.e. 175 current *vs.*1,070 non-current smokers. Three data augmentation techniques: synthetic minority oversampling technique (SMOTE), adaptive synthetic (ADASYN), tree-based associative data augmentation (TADA) were applied together with seven ML methods: logistic regression, k nearest neighbors, support vector machine with linear and radial kernels, decision trees, random forest, extreme gradient boosting. K-fold nested cross-validation was used with the different augmented data types and baseline non-augmented data to validate the prediction outcome. We found that combining data augmentation with ML generally outperformed baseline methods. The final prediction model combines prior feature selection with TADA and support vector machine with linear kernel. Our approach successfully addresses the class imbalance issue in microbiome data for reliable classification and is generalizable to other data. Moreover, we present the first solution for microbiome-based prediction of smoking habits from saliva despite the unavoidable class imbalance issue.

INTRODUCTION

In the recent years, human microbiome research has elucidated the importance of microbes in the host's well-being and their interplay with different phenotypes (1, 2). Human microbiome research is currently moving from characterization and association studies towards translational applications. These include diagnosis of metabolic diseases such as type 2 diabetes (3-5), chronic inflammation disorders (4, 6, 7) and cancer (4, 6, 8, 9), among others; as well as the prediction of the likely outcomes in personalised interventions such as therapeutic response (7, 10) and nutrition (11, 12). In more specialized applications such as forensics, novel uses of the human microbiome have been reported to help in crime scene reconstruction (13, 14), estimating the post mortem interval (PMI) (15) or identifying potential perpetrator(s) (16, 17). This current trend is possible due to advances in high-throughput sequencing technologies and bioinformatics analysis methods, together with the large amount of microbiome data that become available from public repositories. Often, machine learning (ML) methods are preferred for data analysis, with the random forest standing out as often used method so far. ML methods have shown to be sufficiently robust to the statistical challenges inherent to microbiome data, such as their compositional nature by which they carry relative information (18) and their over-dispersed and zero-inflated distribution, since many taxa are not shared among samples.

Despite the great promises of ML methods in microbiome research in general, and its application in trait prediction in particular, including microbiome-based ones (5, 6, 8), they also face practical challenges. Heterogeneity in methods, such as nucleic acids isolation or target region of the marker gene, are often encountered in cumulative microbiome datasets and are an obstacle for cross-study applications due to introduced study-specific technical variation (19). Avoiding pooling data from different studies can bypass the study-specific effect issue, though greatly reducing the statistical power with negative effects on the reliability of the outcome. Additionally, microbiome data commonly suffer from imbalanced sample distribution (20-23); particularly in (binary) classification applications, it is commonly the case that one class is overrepresented (majority class) while the other is underrepresented (minority class). This class imbalance leads to spurious high classification accuracy favouring the majority class, while research often focuses on the minority class, and also limits the classifier's generalization (20-23). Therefore, class imbalance should be considered in the data analysis approach. However, in many occasions collecting data from more samples is unfortunately not viable and therefore, many public datasets come with serious class imbalance problems. Thus, researchers have to explore novel methods for solving the class imbalance at the data and/or algorithm level (21-23).

At the data level, synthetic sampling methods have been suggested for microbiome research (24), though studies applying them are scarce. With these methods, to balance the classes, new samples are synthesized *in silico* based on existing minority class samples and added to the training dataset, an approach referred to as data augmentation. For example, the synthetic minority over-sampling technique (SMOTE) (25) is one of the most widely used methods to deal with the class imbalance problem in real-life applications, and has been employed in some microbiome studies (26-30). An alternative is the adaptive synthetic sampling approach for imbalanced learning (ADASYN) (31). More recently, the tree-based associative data augmentation (TADA) method (32) has been proposed as a microbiome-

specific data augmentation method, since it takes into account the phylogenetic relationship between the microbial taxa, but has not been widely applied by the microbiome community as of yet.

One potential practical classification application of microbiome data, where typically class imbalance is unavoidable, is the microbiome-based prediction of smoking habits, which is relevant in medical, diagnostic and even forensic scenarios. Previous association studies have stablished the relationship of saliva microbiota with the host's tobacco smoking habit (33-38). These studies found that some bacteria such as those from the *Proteobacteria* phylum are decreased in abundancy in the saliva of smokers, while others such as from the *Actinobacteria* phylum are increased. However, at the lower taxonomic levels there are some discrepancies between studies as well as study-specific associations. Notably, the largest available studies suffer from an imbalance ratio of about 1:5 between the minority class of current smokers and the majority class of non-smokers (33, 37). Moreover, since tobacco smoking has been linked to several systemic diseases (39), applying the association results for predicting individuals' smoking habits based on microbiome data is expected to be useful in medical applications. For example, microbiome-based smoking prediction from easily accessible, non-invasive saliva samples may be applicable in epidemiology and public health research when these data are unavailable or not reliable. Typically, smoking data are collected for research purposes via self-reported questionnaires that are known to be unreliable, which could be overcome by microbiome-based prediction of smoking habits. Additionally, such knowledge can add valuable information on the long-term lifestyle habits of a patient in a diagnostic setting or could even help finding unknown perpetrators of crime by providing an investigative lead, as an unknown person cannot be identified with forensic DNA profiling (40, 41). However, ML methods have not been applied for microbiome-based smoking habit prediction from saliva as of yet.

In this study, we aimed to develop an analytical approach that allows to deal with class-imbalanced data and apply it to the prediction of human traits from unavoidable class-imbalanced human microbiome data using smoking habit prediction as relevant example. Our strategy consists of i) optimization and validation of different data augmentation and ML methods using nested cross-validation based on class-imbalanced microbiome data with the option of prior feature selection, and ii) identifying the best-performing approach for realistic prediction of an individual's smoking habit from class-imbalanced microbiome data. The novel analytical approach we introduce here is generalizable, allowing its use on different class-imbalanced microbiome datasets for the prediction of other host's traits as well as for other research purposes and practical applications.

METHODS

Datasets

Publicly available 16S rRNA gene amplicon sequencing data and associated metadata from two different studies were obtained from the European Bioinformatics Institute (EMBL-EBI). The first study (33) (dataset S1 as referred in this paper) included two cohorts: the American Cancer Society (ACS) Cancer Prevention Study II (CPS-II) Nutrition cohort (N=543) (42) and the National Cancer Institute (NCI) Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer

Screening Trial cohort (N=661) (43). The second study (34) (dataset S2 as referred in this paper) included a single cohort from the New York City Health and Nutrition Examination Survey (NYC HANES) (N=297) (44). We discarded samples based on the following criteria: (i) samples lacking metadata information for age, sex and/or ethnicity; (ii) samples from donors less than 15 years old based on microbial community differences between youth and adults (45); (iii) duplicate samples from the dataset S1 to avoid data redundancy; and (iv) samples from non-smokers with second-hand exposure and 'alternative' smokers from the dataset S2. Selected characteristics of the two studies are described in Table 1. The setup of the experimental studies is described in further detail in Supplementary Table S1.

Table 1. Characteristics of the two microbiome datasets used in this study.

	Dataset S1* (N=1,088)	Dataset S2** (N=157)
Smoking status, N (%)		
Never smoker	473 (43.5)	39 (24.8)
Former smoker	519 (47.7)	39 (24.8)
Current smoker	96 (8.8)	79 (50.4)
Sex, N (%)		
Female	429 (39.4)	88 (56.1)
Male	659 (60.6)	69 (43.9)
Age group, N (%)		
20-29	-	20 (12.7)
30-39	-	31 (19.8)
40-49	-	40 (25.5)
50-59	147 (13.5)	29 (18.5)
60-69	505 (46.4)	21 (13.4)
70-79	377 (34.7)	9 (5.7)
80-89	59 (5.4)	6 (3.8)
≥90	-	1 (0.6)
Ethnicity, N (%)		
European	1,028 (94.5)	59 (37.6)
Non-European	60 (5.5)	98 (62.4)

* Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. ISME J 10:2435-2446.

** Beghini F, Renson A, Zolnik C, Geistlinger L, Usyk M, Moody TU, Thorpe L, Dowd JB, Burk R, Segata N, Jones HE, Waldron LD. 2019. Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. Ann Epidemiol 34:18-25.

Processing of 16S rRNA gene amplicon sequencing data

Data from the two selected studies were processed separately. Primer sequences were obtained from the original studies and were removed from the raw sequencing reads using

cutadapt (v.1.15) (46) by setting minimum length to >100 bp. The resulting FASTQ files were quality-filtered and de-noised using DADA2 (v.1.12.1) (47). Briefly, in both studies parameters maxNN and maxEE were set to 0 to avoid unambiguous nucleotides and 'expected errors' in the sequencing reads, respectively. Additionally, in dataset S1 (single-end) parameter maxLen was set to 500 and in dataset S2 (paired-end) parameter truncLen was set to 200-150 based on the read quality profiles, making sure to maintain overlap between forward and reverse reads to merge them later. After sample inference of true sequence variants, an amplicon sequence variants (ASVs) table was constructed for each study and chimeric sequences were removed using the command *removeBimeraDenovo()* with default parameters. Subsequently, the naïve Bayesian classifier method was employed for taxonomy assignation using the expanded Human Oral Microbiome Database (eHOMD) (v.15.2) (48) as reference. At this point, only high-coverage samples (>1,000 reads) were kept for downstream analysis and species with mean relative abundance < 1E-04 across samples were discarded. Taxa counts were normalized using total-sum scaling (TSS) for relative abundance (49). Moreover, microbiome datasets are normally sparse and characterized by a zero-inflated distribution, where most taxa are not shared among the majority of the samples. This is magnified in cross-study applications with study-specific taxa which can limit the generalizability of the applications. Based on this, we merged the two ASVs tables from the two analysed studies and filtered out study-specific taxa. Afterward, we assessed potential introduced cross-study technical variation using guided principal component analysis (gPCA) (50).

Statistical analyses

The overall differences in the saliva microbial communities between the smoking classes were calculated in QIIME 2 (v.2019.10) (51): current *vs.* never *vs.* former, and current *vs.* non-current (combined never and former). For this, the weighted UniFrac distance matrix was analysed by analysis of similarities (ANOSIM) and permutation multivariate analysis of variance (PERMANOVA) where *q* values ($q<0.05$ for significance) were obtained with default 999 permutations.

Consensus feature selection

Feature selection was performed in the baseline non-augmented data; particularly in the outer *k*-folds of training datasets (Figure 1D1-D7) prior to nCV procedure. Differences between smoking classes (smokers *vs.* non-smokers) across bacterial species were measured using the two-sided Mann-Whitney U test. In each *k*-fold of training dataset ($k=5$), significant features were identified after Benjamini-Hochberg (BH) correction for significant *p* values ($p<0.05$ for significance). This process was repeated for ten times to control for introduced variation by data partitions, resulting in a total of 50 (5*10) lists of significant features. Significant features in at least 90% of the *k*-folds of training datasets (45 out of 50 folds) were selected as consensus features**.**

Data augmentation techniques

For the prediction of an individual's current smoking habit (smoker *vs.* non-smoker) we aimed to employ a binary machine learning (ML) classifier. For that, data imbalance was a marked issue in our dataset with a ratio of about 1:6 between the minority class (N=175 smokers) and the majority class (N=1,070 non-smokers) (Table 1). The problem stems from the ML

algorithms that assume equal number of samples for each class which would lead to spurious high classification accuracy, favouring the majority class and limiting the classifier's generalization. Therefore, we applied different data augmentation techniques to overcome the data imbalance issue at the data level in our dataset. We used two techniques commonly employed in different fields to handle data imbalance named synthetic minority over-sampling technique (SMOTE) (25) and adaptive synthetic sampling approach (ADASYN) (31), as well as a recently-introduced technique specific for microbiome data named tree-based associative data augmentation (TADA) (32).

The general approaches to deal with data imbalance are over-sampling (increase the minority class), under-sampling (decrease the majority class) or a combination of the two. Particularly, SMOTE and ADASYN techniques differ on the generation of synthetic samples in the minority class (over-sampling). For that, SMOTE over-sampling pinpoints the samples belonging to the minority class in an Euclidean space and a random sample is first chosen for which *k* of its nearest neighbors are found. A line is drawn between the original sample and one randomly chosen neighbor, where a new synthetic sample is generated at a random point along the line (linear combination of samples). The process is repeated generating the same number of synthetic samples for each original minority sample until a specific ratio between the minority and majority classes is reached or to equal the majority class. On the other hand, ADASYN adds random small values to the neighbor samples; hence, they are not linearly correlated to the original sample. By this, ADASYN considers a density distribution between the original sample and its neighborhood, which acts as the criterion to set the number of synthetic samples to be generated from each original sample. On another point, with the under-sampling approach random majority class samples are dropped out until a specific ratio between the classes is reached. Both SMOTE and ADASYN techniques were implemented using the imbalanced-learn Python toolbox (v.0.6.1) (52) with default parameters. We employed a combination of over- and under-sampling, indicated as SMOTE-1 and ADASYN-1 in this study. In order to set the final ratio between the minority and majority classes we used the following equation:

$$t = \approx |C_{min} - C_{max}|, \textit{ over-sampling} = \frac{t - C_{min}}{C_{max}}, \textit{ under-sampling} = \frac{C_{max} - t}{C_{min}}$$

where C_{min} is the number of the minority class samples and C_{max} is the number of the majority class samples.

We also used the over-sampling approach alone, indicated as SMOTE-2 and ADASYN-2 in this study, by which the number of the samples in the minority class was equalled to the majority class.

The microbiome-specific TADA technique generates minority class synthetic samples based on a statistical generative model that takes into account the phylogenetic relationships between microbial taxa. We implemented TADA with default parameters which equals the number of samples in the minority class with the majority class. For the rooted phylogenetic tree required as input we used the merged ASVs table of the two studies to obtain a single consensus sequence for all those sequences assigned to the same taxa at the species level. For that, we used the *ConsensusSequence* function in DECIPHER (v.2.12.0) (53) and subsequently we performed multiple sequence alignment of the consensus sequences using MAFFT with auto parameter (v.7.310) (54). A rooted phylogenetic tree was obtained using FastTree (v.2.1.11) (55) with generalized time-reversible (GTR) model.

Machine learning algorithms

We evaluated seven different ML methods for binary classification (0: smoker, 1: non-smoker) included in the scikit-learn (v.0.23.2) Python package (56): logistic regression (LR), k nearest neighbors (KNN), support vector machine with linear (SVML) and radial (SVMR) kernels, decision trees (DT), random forest (RF) and extreme gradient boosting (XGBoost). LR is a parametric ML model that assumes a linear dependency between the input features (taxa) and the categorical outcome. The output of the logistic regression linear function is a probability x between 0 and 1, where if $x<0.5$ the categorical outcome is 0 (smoker), otherwise 1 (non-smoker). KNN is a non-parametric model and as such supports non-linear solutions. It finds the Euclidean distances between a query sample and a k number of its closest samples (nearest neighbors) in the feature space and identifies their most frequent class label. SVM models take the data points and find a separating hyperplane between the two classes. SVML is a linear method that looks for linear dependencies among the input features to separate classes. SVMR is a non-linear method that adds an extra dimension to the data (kernel) so they become linearly separable and then projects back the decision boundary to the original dimension using the dot product of two vectors in the feature space known as the kernel function. DT is a tree-based ML algorithm that mimics a decision diagram. Each input feature constitutes a node in the tree, where based upon a certain condition or rule splits into sub-nodes and extends until the leaf node that represents the classification decision (0 for smoker or 1 for non-smoker). Finally, RF and XGBoost are tree-based ensemble models that combine several models to improve their outcome predictions. RF generates a large number of decision trees on different subsamples and combines their outputs using averages at the end of the learning process. On the contrary, XGBoost combines the decision trees during the learning process for which it uses a gradient descent algorithm. By this, the mistakes done in a previous model are learnt and improved in the subsequent model until no further improvement can be achieved. Hyperparameter optimization for all the ML models was performed using nested k-fold cross-validation (Figure 1F).

Nested k-fold cross-validation

Nested cross-validation (nCV) is a resampling procedure that enables both model optimization and evaluation (57). The difference between non-nested and nested CV approaches is that the former use the same cross-validation dataset for hyperparameter optimization and model evaluation which biases the model to the dataset and leads to optimistically biased classifier's performance; in other words, over-fitting in model selection. The nCV approach overcomes this by evaluating the ML algorithm and the model hyperparameters separately in multiple randomised partitions of the data (58), though it requires more computational time. In nCV, apart from splitting the original training set into k folds of training and test sets (outer folds) (Figure 1E), each k training fold is at the same time split into n folds of training and validation sets (inner folds) for model hyperparameter tuning (Figure 1F). The optimized model is then validated in the corresponding k test fold (Figure 1G). We employed a 5x2 (kxn) nCV where each of the first splits is named outer-fold (k) and each of the inside splits for hyperparameter tuning inner-fold (n). Hyperparameter optimization for the seven ML models was performed using the *RandomizedSearchCV()* function in scikit-learn.

Validation of data types with machine learning methods

Since most ML algorithms operate trying to maximize the classification accuracy, spurious high classification occurs in imbalanced datasets by correctly classifying all or almost all the samples from the majority class at cost of misclassifying many samples from the minority class. Hence, performance metrics such as accuracy or F1 score alone can lead to misleading results in imbalanced datasets (59). In contrast, the Mathew's correlation coefficient (MCC) offers a balanced metric by considering the four confusion matrix categories: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), according to the following equation:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

For a high MCC score the classifier has to correctly predict a high percentage of the samples in both the majority and the minority classes, independent of their ratios in the overall dataset; hence, independent of data imbalance. MCC gives a score ranging [-1, +1], where 1 means perfect prediction, 0 random prediction and -1 perfect inverse prediction. Based on this, in order to compare all the possible combinations of input data type, including original non-augmented data, (Figure 1, d = 6) and ML method (Figure 1, m = 7) we used the MCC metric. However, we also reported the AUC metric to show potential misleading results in those models with baseline non-augmented data (highly imbalanced). Comparisons among the different data types with a given ML method were performed in R environment (v.3.6.1) using the Kruskal-Wallis and Wilcoxon tests. Significant p values were determined with a value cut-off at 0.05 following Benjamini-Hochberg (BH) correction.

Approach setup

The original dataset (N=175 smokers, N=1,070 non-smokers) was split into training (80%) and test (20%) sets maintaining the sample ratio between the classes (Figure1A, B, C). Data augmentation techniques were applied to the training split: ADASYN-1 (over- and under-sampling), ADASYN-2 (over-sampling alone), SMOTE-1 (over- and under-sampling), SMOTE-2 (over-sampling alone) and TADA. We evaluated a total of six training data types (d=6), including the original non-augmented and the five augmented data types (Figure 1B). We performed the proposed strategy without feature selection and filtering by previously identified consensus significant features (Figure 1D) (for details on feature selection procedure, see the *Consensus feature selection* in the Methods section ; Figure 1D1-D7). Considering each data type separately, we optimized and evaluated seven ML methods (LR, KNN, SVML, SVMR, DT, RF and XGBoost) (m=7) using a nCV approach as explained before (Figure 1E, 1F, 1G). This entire process was repeated ten times (j=10) (Figure 1I) aiming to avoid introduced variation by the original data partitions. The performance metrics (MCC and AUC) resulted from the validation of each optimized model in the five outer test folds (k) over ten times (j=10) (total of 50 (5*10) resulting values for each metric) (Figure 1H). The best performing data type with ML method was based on the highest resulting MCC value (Figure 1H) and the final classifier trained in the whole original training set (Figure 1B, J) was validated in the original test set (Figure 1C, K).

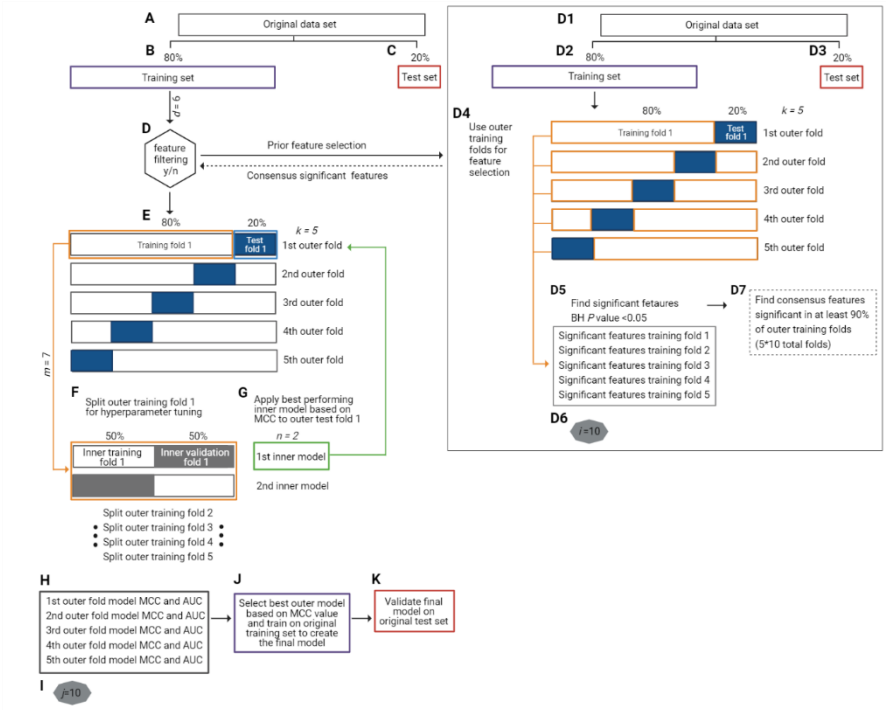


Figure 1. Overview of the study's analytical strategy. **A**, **B**, **C**. The original dataset was split into training (80%) (purple box in B) and test (20%) (red box in C) sets by maintaining the original ratio between classes in the partitions. Data augmentation techniques were applied to the training set, making a total of six different input data types (*d*=6), including baseline non-augmented and differently augmented data types. **D**. The proposed strategy was performed both without feature selection (*f*=124) and with prior feature selection at the species level (*f*=19). **D1**, **D2**, **D3**, **D4**. Prior feature selection was carried out in the outer *k*-folds of the training datasets (orange boxes in D4). **D5**. Significant features were identified in each *k*-fold training dataset. **D6**. Repetition (D1)-(D5) 10 times (*i*=10) to control for introduced variation by data partitions. **D7**. Identified consensus significant features were kept for the downstream analysis steps (E-H) when feature filtering option was chosen (D). **E**. For the nested cross-validation (nCV) approach, the training set was split into 5 outer *k*-folds of training (80%) (orange box in E) and test (20%) (blue box in E) datasets each one. **F**. Each outer *k*-fold was split into 2 inner *n*-folds of training (50%) and validation (50%) datasets (orange box in F) in which seven different machine learning (ML) models (*m*=7) were optimized and validated (inner models). **G**. The best performing *n*-fold inner model (green box in G) was applied to the corresponding *k*-fold test set (green arrow to blue box in E). **H**. For each *k*-fold test set, two performance metrics were obtained: Mathews correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC). Repetition (E)-(H) for all the input data types (*d*=6) with ML method (*m*=7) (total of 42 different approaches). **I**. Repetition (A)-(H) 10 times (*j*=10) to control for introduced variation by data partitions. **J**. Selection of the best performing data type with ML method based on MCC metric and training on full original training dataset (purple box in C) to create the final prediction model. **K**. Validation of final prediction model on original test dataset (red box in D).

RESULTS

Saliva microbiome data

Data comprised saliva 16S rRNA gene amplicon sequencing data and associated metadata from two different studies referred in this research as dataset S1 (33, 42, 43) and dataset S2 (34, 44) (see the *Datasets* in the Methods section for more details). Filtering samples for quality-controlled metadata, de-noising of sequencing reads, and sequencing depth filtering resulted in 1,245 total samples (N=1,088 from dataset S1 and N=157 from dataset S2). In the whole dataset, class imbalance in smoking habits was large regarding current smokers, with 512 (44.1%) never smokers, 558 (44.8%) former smokers, but only 175 (14.1%) current smokers. Female samples accounted for 41.5% of the total and the average age (± standard deviation) was 65.2 (±11.0) years. European ancestry of the saliva sample donors was overrepresented (87.3%), as typically encountered in human microbiome data available thus far. Selected characteristics of the two datasets are described in further detail in Table 1.

Microbial taxonomy assignment using the expanded Human Oral Microbiome Database (eHOMD) (v.15.2) (48) as reference (see *Processing of 16S rRNA gene amplicon sequencing data* in the Methods section for more details) resulted in 591 species from 94 families in dataset S1, and 356 species from 67 families in dataset S2. For downstream analyses, we selected the 124 species from 30 families that were common between the two datasets, to ensure that our proposed strategy was generalizable for the prediction in samples from both datasets. These common species accounted for 86% of the sequencing reads in dataset S1 and 61% in dataset S2.

We assessed potential study-specific introduced technical variation in a total of 12,570,370 sequencing reads from the final 1,245 samples of both datasets i.e. 11,835,011 sequencing reads from 1,088 samples in dataset S1 and 735,359 sequencing reads from 157 samples in dataset S2. Notably, the two studies used different DNA isolation kits and different NGS platforms, and also targeted different amplicons of the 16S rRNA gene while overlap existed in the V4 hypervariable region (Supplementary Table S1). However, no statistically significant study-specific variation was reported using generalized principal component analysis (gPCA) (p=0.24), where the two studies clustered together in the first two principal components (PC) (Supplementary Figure S1).

Classification of smoking habits

The overall saliva microbial communities differed with statistical significance between current and never smokers (ANOSIM R=0.04, q=0.03; PERMANOVA pseudo-F=11.37, q=0.002), current and former smokers (ANOSIM R=0.04, q=0.03; PERMANOVA pseudo-F=11.91, q=0.002), but not between never and former smokers (ANOSIM R=0, q=0.51; PERMANOVA pseudo-F=0.64, q=0.63). Therefore, we grouped the never and former smokers into a single category of non-current smokers, which when compared with the current smokers showed statistically significant differences in the overall microbial communities (ANOSIM R=0.04, q=0.02; PERMANOVA pseudo-F=13.26, q=0.001). Based on these results, we used two classes of non-current and current smokers in all downstream analyses.

Validation of data types and machine learning models for smoking habit prediction based on all microbiome features

A step-by-step overview of our analytical setup can be found in Figure 1. For each input data type (*d*=6), including augmented data and baseline non-augmented data, and each ML model (*m*=7), the resulting classifiers' performance metrics expressed as Mathew's correlation coefficient (MCC) and area under the receiver operating characteristic curve (AUC) are summarized in Figure 2 and Supplementary Table S2. Overall, data augmentation techniques with ML methods outperformed baseline methods, with the exception of the KNN method. Briefly, the MCC values resulting from the baseline non-augmented methods increased on average when applying data augmentation techniques with percentages of increase as follows: XGBoost (99.8%), SVMR (92.7%), DT (48.9%), RF (30.6%), LR (8.8%). The highest increase was observed with SVML where the baseline non-augmented method resulted in random prediction (MCC equal or close to zero), which was highly improved with data augmentation techniques (MCC values 0.31-0.33). Notably, the AUC baseline values did not change so drastically when applying data augmentation techniques (percentage increase or decrease (-)): XGBoost (15.8%), SVML (8%), SVMR (null increase/decrease), RF (-1.0%), KNN (-4.6%), DT (-6.1%), LR (-10.4%).

The SVML method performed the best in predicting smoking habits from microbiome data based on the MCC metric. As the reference metric for comparison purposes, we chose the MCC, since it is independent of data imbalance, which is not the case for the AUC metric. MCC values were significantly higher with each of the five augmented data types compared to non-augmented data (Wilcoxon test, BH-adjusted p=9.93E-20) (Supplementary Table S3). However, there were no statistically significant differences in the MCC metric between the augmented data pair comparisons (Wilcoxon test, BH-adjusted p values between p=0.392 and p=0.882) (Supplementary Table S3). From these results we concluded that SVML with augmented data performed better than with imbalanced non-augmented data.

Validation of data types and machine learning models for smoking habit prediction based on prior selected microbiome features

Feature selection is a commonly used approach to improve the performance of the prediction model by only retaining the features with the highest predictive power and eliminating useless or redundant features (60). It also avoids the computational cost associated with measuring all the features (60). Applying a feature selection analysis in smokers *vs.* non-smokers identified 19 statistically significant differences of the 124 tested features (BH-p<0.05 in at least 90% of the *k*-folds of the training datasets; for further details, see the *Consensus feature selection* in the Methods section) (Supplementary Table S4). These 19 significant features accounted for 14% of the sequencing reads in dataset S1 and 9% of the sequencing reads in dataset S2. The majority of the significant features belonged to the *Firmicutes* phylum (36.8%), followed by *Bacteroidetes* (31.6%), *Proteobacteria* (15.8%) and *Actinobacteria* (15.8%) phyla. The features most associated with smokers were: *Streptococcus parasanguinis clade 411*, *Prevotella veroralis*, *Gemella morbillorum*, *Prevotella nigrescens*, *Alloprevotella tannerae*, *Neisseria cinerea* and *Prevotella intermedia*. The features most associated with non-smokers were: *Actinomyces sp. HMT 180*, *Neisseria NA*, *Corynebacterium durum*, *Lachnoanaerobaculum umeaense*, *Abiotrophia defectiva*, *Capnocytophaga leadbetteri*, *Lachnospiraceae [G-2] bacterium HMT 096*, *Veillonella*

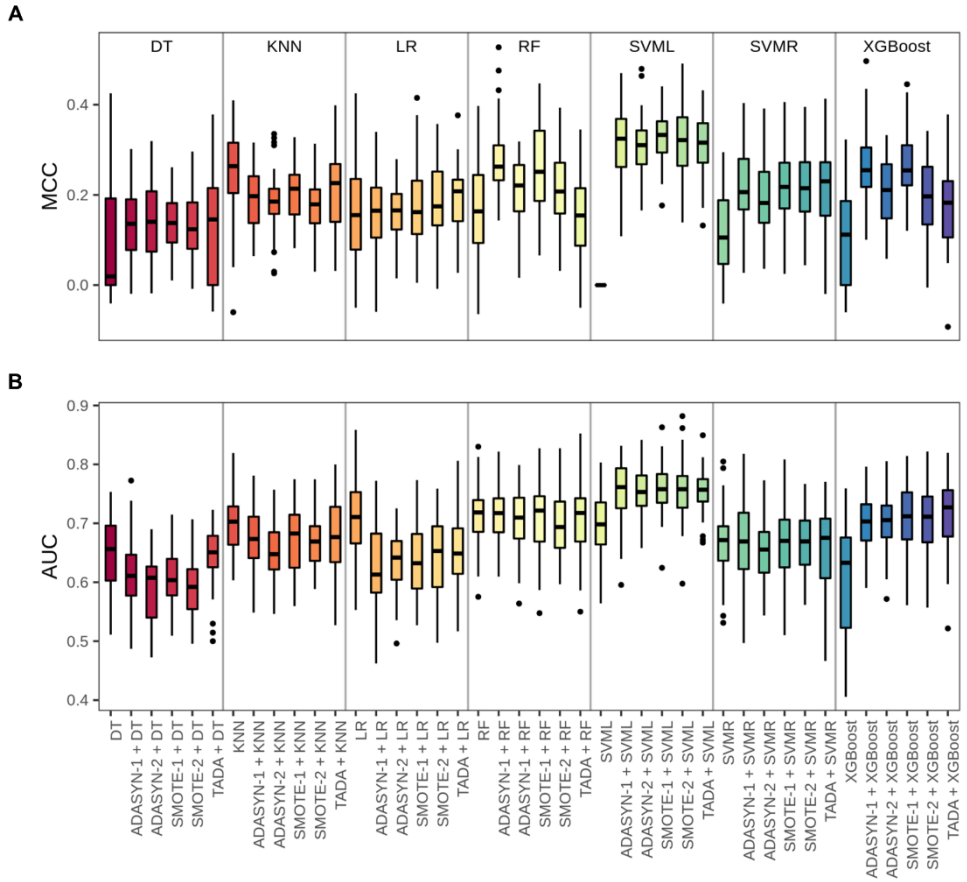


Figure 2. Validation of data types with machine learning (ML) methods for microbiome-based prediction of smoking habits based on the S1 and S2 datasets together. For each ML method, we evaluated six types of input data: baseline non-augmented, and five augmented datasets based on different methods ADASYN-1, ADASYN-2, SMOTE-1, SMOTE-2 and TADA. A. Mathews correlation coefficient (MCC) and B. area under the receiver operating characteristic curve (AUC) values from the five-fold nested crossed-validation repeated for ten times (5*10). For MCC, +1 represents perfect prediction, 0 random prediction and –1 perfect inverse prediction. For AUC, 1 indicates perfectly accurate prediction and 0.5 indicates random prediction. ML method abbreviations: DT, decision trees; KNN, k nearest neighbors; LR, logistic regression; RF, random forest; SVML, support vector machine with linear kernel; SVMR, support vector machine with radial kernel; XGBoost, extreme gradient boosting.

rogosae, *Rothia aeria*, *Capnocytophaga granulosa*, *Neisseria oralis* and *Lactobacillus iners*. An overview of the distribution of the significant features across the smoking habit classes can be found in Figure 3.

The resulting classifiers' performance metrics with prior feature selection – MCC and AUC – are summarized in Figure 4 and Supplementary Table S5. Overall, the general performance trend for each input data type with ML method with this reduced number of 19 features was similar to that observed with the whole set of 124 features where data augmentation techniques with ML methods generally outperformed baseline methods (see the *Validation of data types with machine learning models for smoking habit prediction* in the Results section before). Considering all input data types, MCC values changed based on ML

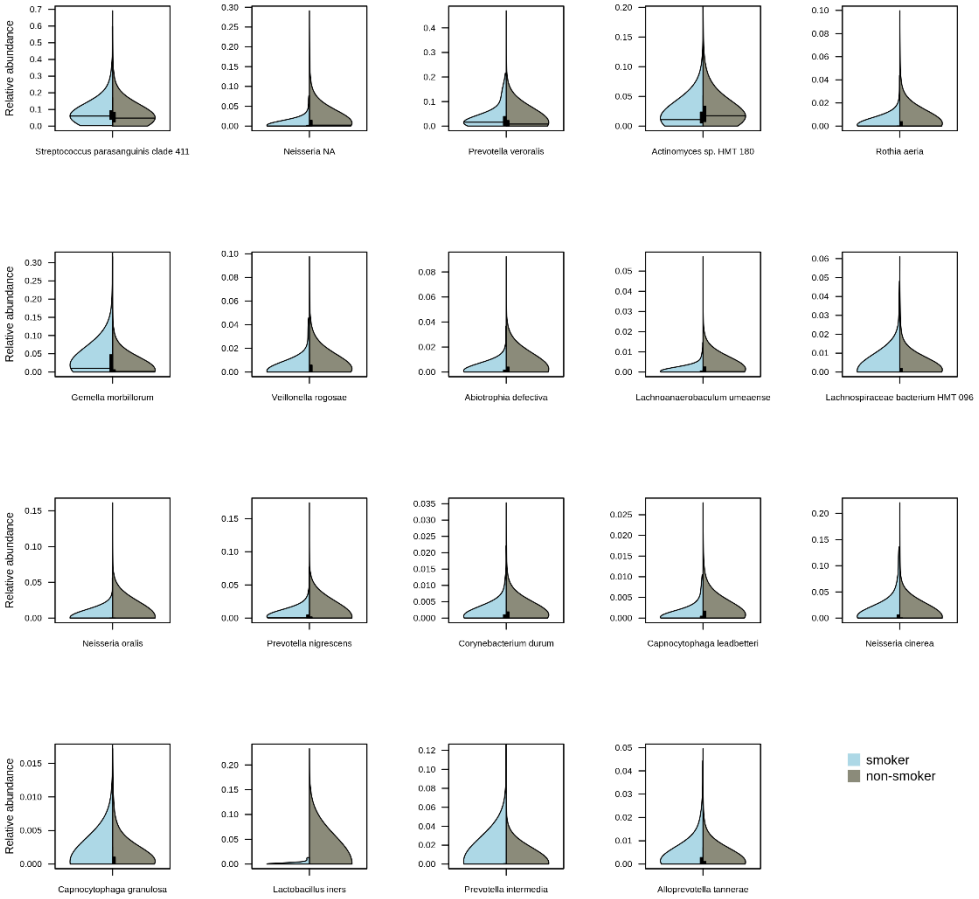


Figure 3. Violin box plots of the relative abundance of the significant 19 differences of the 124 tested features. Smoking class is shown in the x-axis where blue indicates smoker and green indicates non-smoker. Relative abundance is shown in the y-axis.

method when prior feature selection was applied (percentage increase or decrease (-)): LR (58.3%), DT (23.6%), SVMR (14.9%), RF (8.9%), SVML (1.6%), XGBoost (-5.5%) and KNN (-8.8%). AUC values also changed based on the ML method when prior feature selection was applied (percentage increase or decrease (-)): LR (14.4%), SVMR (8.8%), DT (0.4%), RF (-1.4%), KNN (-1.5%), SVML (-1.6%), XGBoost (-2.1%). Overall, data augmentation techniques with ML methods outperformed baseline methods, with the exception of the DT and KNN methods. Briefly, the MCC values resulting from the baseline non-augmented methods increased on average when applying data augmentation techniques, with percentages of increase as follows: XGBoost (130.8%), LR (126.0%) and RF (38.8%). The highest increase was observed with SVM methods, where the baseline non-augmented method resulted in random prediction (MCC equal or close to zero), but was highly improved with data augmentation techniques (MCC: 0.31-0.33 in SVML, MCC: 0.26-0.30 in SVMR). Notably, the

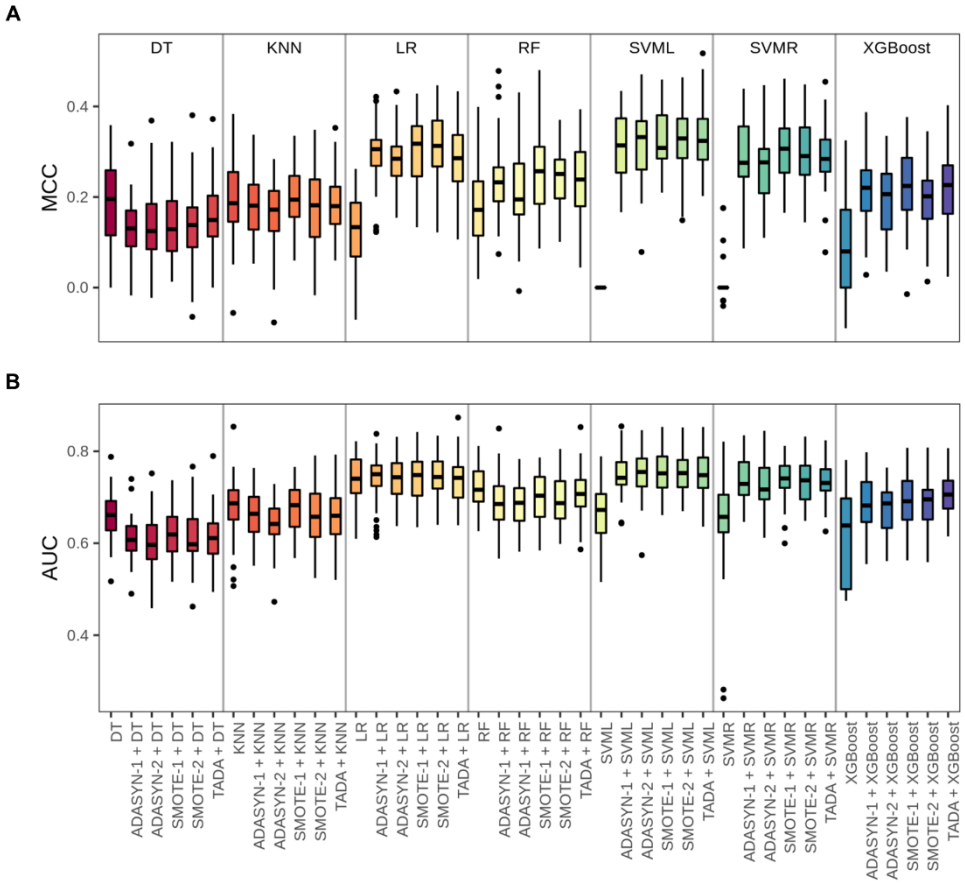


Figure 4. Validation of data types with machine learning (ML) methods for microbiome-based prediction of smoking habits based on the S1 and S2 datasets together after initial feature selection. For each ML method, we evaluated six types of input data: baseline non-augmented, and five augmented dataset based on different methods ADASYN-1, ADASYN-2, SMOTE-1, SMOTE-2 and TADA. A. Mathews correlation coefficient (MCC) and B. area under the receiver operating characteristic curve (AUC) values from the five-fold nested crossed-validation repeated for ten times (5*10). For MCC, +1 represents perfect prediction, 0 random prediction and –1 perfect inverse prediction. For AUC, 1 indicates perfectly accurate prediction and 0.5 indicates random prediction. ML method abbreviations: DT, decision trees; KNN, k nearest neighbors; LR, logistic regression; RF, random forest; SVML, support vector machine with linear kernel; SVMR, support vector machine with radial kernel; XGBoost, extreme gradient boosting.

AUC baseline values did not change so drastically when applying data augmentation techniques (percentage increase or decrease (-)): SVML (13.6%), SVMR (13.2%), XGBoost (12.8%), LR (null increase/decrease), KNN (-2.9%), RF (-3.5%) and DT (-7.6%).

As without prior feature selection, the SVML method performed the best in predicting smoking habits from microbiome data based on the MCC metric. MCC values were significantly higher with each of the five augmented data types compared to baseline non-augmented data (Wilcoxon test, BH-adjusted p=9.93E-20) (Supplementary Table S3). However, there were no statistically significant differences in the MCC metric between the augmented data pair comparisons (Wilcoxon test, BH-adjusted p values between p=0.739 and p=0.983) (Supplementary Table S3). From these results we conclude that SVML with

augmented data performed better than with imbalanced non-augmented data. For the final model training (Figure 1J) and final model validation in the test set (Figure 1K), we selected SVML method with TADA and initial feature selection. We based our decision on the following: i) the SVML method performed the best in predicting smoking habits from microbiome data based on the MCC metric with and without initial feature selection (not statistical difference); ii) we selected TADA as the data augmentation technique since it takes into account the phylogenetic relationship between the microbial taxa; and iii) we selected the approach with initial feature selection as it allows for shorter computational time. The average model performance (standard deviation) metrics were 0.33 (0.07) MCC and 0.75 (0.05) AUC in the training dataset as well as 0.32 (0.06) MCC and 0.75 (0.03) AUC in the test set.

DISCUSSION

With this study, we introduce a generalized strategy to deal with the common issue of data imbalance in human microbiome binary classification with the aim of unlocking the prediction of human host's traits. By applying this strategy, we - for the first time – succeeded in realistically predicting a person's smoking habit from saliva using unavoidable class-imbalanced 16S rRNA gene amplicon microbiome sequencing data, which in the future may become useful in medical and forensic settings.

We selected publicly available saliva microbiome data from two studies that differed in their experimental setup (Supplementary Table S1). We only included studies targeting the saliva microbiome. We did not use data from microbiome-smoking association studies targeting other niches in the oral cavity due to known diverse microbial assemblies on different oral sites (61-63). We only included these two studies based on their big sample sizes, discarding other studies with very small sample sizes that could be source of variation rather than useful information for the prediction. The lack of widespread consensus on microbiome analysis methods, together with the variation introduced at each step of the microbiome pipeline, constitute hurdles for cross-study applications. This lack can sometimes outweigh the factor(s) of interest as well as limit the statistical power and generalization of the application (64-67). Though we could not control for any potential variation introduced during the experimental analysis, we aimed to apply the same or the most similar bioinformatics analysis to the raw sequencing data to avoid study-specific computational variation, from which quality control choices are amongst the largest sources of variation (64, 65). Moreover, we only selected the species common between the two datasets for downstream analyses. On one hand, we are aware that this might have reduced the power of our prediction by discarding informative species in each of the two datasets separately. On the other hand, this procedure ensured that the approach is generalizable for the prediction in samples from both datasets.

Our observations in the overall saliva microbial composition were in agreement with the two original studies (33, 34), where microbiome variation did not significantly differ between never and former smokers (ANOSIM R=0, q=0.51; PERMANOVA pseudo-F=0.64, q=0.63), but significantly differed between never and current (ANOSIM R=0.04, q=0.03; PERMANOVA pseudo-F=11.37, q=0.002), and between former and current smokers

(ANOSIM R=0.04, *q*=0.03; PERMANOVA pseudo-F=11.91, *q*=0.002). One of the two studies (33) also reported significant differences between current smokers and non-current smokers (combined never and former) as we did here (ANOSIM R=0.04, *q*=0.02; PERMANOVA pseudo-F=13.26, *q*=0.001).

The problem of data imbalance, which is commonly encountered in microbiome datasets and in many other real-life applications, is the main focus of our proposed strategy. Our dataset suffered from highly imbalanced data with a ratio of about 1:6 between the minority class (current smokers) and the majority class (non-smokers) (Table 1). Imbalanced data can lead to spurious high accuracy based on the correct classification of the majority of the samples from the majority class at the cost of misclassifying many or even the majority of the samples from the minority class (20-23). This would be translated in the lack of generalizability of our prediction based on the classifier's inability to correctly predict the positive observations for current smoking habits (minority class). A clear example of this was the baseline non-augmented data with the SVML method (Figure 2, Supplementary Table S2), where we obtained a MCC of zero but a relatively high AUC of 0.7. This spurious high AUC occurs in the baseline imbalanced dataset by correctly classifying almost all the samples from the majority non-smoker class at cost of misclassifying most of the samples from the minority smoker class. This highlights the necessity of future studies not relying only on a single accuracy score for model validation when dealing with imbalanced data (59).

The MCC performance metric allowed us for fair comparisons of the validated ML methods for both non-augmented and augmented data, since this performance metric is independent of data imbalance (68, 69). For the great majority of the ML methods, augmented data resulted in higher MCC scores compared to imbalanced non-augmented data; hence, improved classification performance. This demonstrates that microbiome-based classification problems can benefit from data augmentation techniques when needed, in line with previous suggestions (24). In our dataset, the combined over- and under-sampling approach generally performed slightly better (though not statistically significantly) than the over-sampling approach alone (Supplementary Tables S3 and S5). However, we acknowledge that in other datasets under-sampling might be detrimental since, although dropped samples belong to the majority class, they might still be useful for the learning process of the ML method.

The variation in the performance metric values for each input data type with ML method (Figures 2 and 4) highlights the variation introduced in the optimization and validation procedures (Figure 1). This underlines the necessity for a nCV approach for overall model validation and selection that is independent from the different data partitions (57, 58). In our final selected and validated classifier, we confirmed generalization; hence, avoiding overfitting in model selection as demonstrated by the very similar performance metrics between the final training (MCC: 0.33±0.07, AUC: 0.75±0.05) and test (MCC: 0.32±0.06, AUC: 0.75±0.03) datasets, which were very similar to those of the folds in the nCV (MCC: 0.33±0.07, AUC: 0.75±0.05). As it has been suggested before (70), with our strategy, we report the variation in the predictive performance on the different folds of nCV, as well as on both the final training and test datasets, which unfortunately it is not very common practice in microbiome-based trait prediction.

To the best of our knowledge, our study provides the first-of-its-kind solution to predict individuals' smoking habits from saliva 16S rRNA gene microbiome data and despite the class imbalance problem of the data used. Previously, Sato *et al.* (38) predicted smoking habits

from tongue metagenomics data with class imbalance (N=234 never, N=52 current smokers) using a RF approach and conventional non-nested k-fold CV. The authors only reported AUC from the testing dataset (AUC=0.75±0.10). This prediction was improved when using single-nucleotide variants of single species as input data (AUC=0.80-0.93) instead of relative abundances of all species. More recently, Carrieri *et al.* (71) predicted smoking habits from leg skin 16S rRNA gene amplicon sequencing data based on a very small dataset (N=43 never, N=19 current smokers) using the XGboost method and conventional non-nested k-fold CV. In this case, the authors reported the F1 performance metric in the CV folds (F1=0.72±0.12), training dataset (F1=0.98) and test dataset (F1=0.85). The noted differences in the F1 scores might be an indication of introduced variation by the different data partitions and bias towards model selection, which can be overcome using a nCV approach as proposed by us and others. The methods applied in both of these previous studies did not take the class imbalance problem in the used data into account. Therefore, because of the small sample size the prediction accuracies reported in these two studies are not expected to be completely reliable.

Regarding feature selection, we identified bacterial species that were more associated with current smokers, such as species belonging to the *Streptococcus* and *Lactobacillus* genera; as well as other species that were more associated with non-smokers, such as species belonging to the *Abiotrophia*, *Capnocytophaga* and *Corynebacterium* genera, in agreement with the original studies (33, 34). Although both original studies reported that *Neisseria* genera was lower in smokers, we observed that, depending on the species, *Neisseria* genus was more associated with non-smokers (i.e. non-identified species from the *Neisseria* genus and *Neisseria oralis*) or with smokers (i.e. *Neisseria cinerea*) (Figure 3). This further highlights the previously reported study-specific associations with smoking habits that might be partly influenced by the taxonomic level of analysis (33-38).

In our dataset, we acknowledge some metadata-related characteristics that might limit the generalizability of the microbiome-based smoking habits prediction, even when the data imbalance issue is overcome. Precise phenotype descriptions were available in only one of the two studies (Supplementary Table S1), which is a commonly encountered problem in cross-study applications and can limit the interpretation of results (72). Also, the dataset is overrepresented by the age range of 50-79 years old and European ancestry of the sample donors (Table 1), which might result in different prediction performance in other age groups (73, 74) and ethnicities (75, 76). To add, one limitation of the data augmentation techniques is that synthesized metadata associated with the synthetically produced data is not reliable. This limits the possibility of statistically adjusting for covariates (i.e. age, sex, ethnicity) in the ML methods, which can ultimately improve the prediction performance. Hence, the ideal scenario would be to start from a sample that is a good representation of the general population, though this is challenging in real-life applications.

To conclude, our proposed strategy validates different ML methods based on the 'no-free-lunch' theorem that states that there is no best method for classification (77) with the aim for a generalizable strategy that other researchers can apply to different datasets and prediction problems. In parallel, we believe that there is no best data augmentation technique to deal with imbalanced data in our dataset at least; hence, we based our strategy on the validation of different techniques with the aim of unlock the prediction of novel traits. We found that combining data augmentation with ML generally outperformed baseline methods, as other

researchers have also suggested before (24). With the example we chose, we present for the first time a solution to predict a person's smoking habit from saliva 16S rRNA gene microbiome data. Future research might focus on unlocking the microbiome-based prediction of novel traits, even when the issue of imbalanced data is unavoidable, as we did here. A bigger emphasis on the report of the used prediction pipeline might ultimately help other researchers to better assess the best approach for a given application and create awareness about the challenges and limitations encountered with different data types and ML methods. Also, the research community across different fields would benefit from improving some features of the data augmentation techniques, such as reliable synthetized metadata where they are non-available or non-standardized.

ACKNOWLEDGEMENTS

We thank the colleagues for making their microbiome and smoking data publicly available. We also thank Liese Boonstra for contributing to initial data collection and exploration as part of her Bachelor education at Leiden University of Applied Sciences carried out during her internship at Erasmus MC.

AUTHOR'S CONTRIBUTIONS

CDL and AV conceptualized the idea. CDL, DMG and AV designed the study with contributions by MK. DMG implemented and performed the data analysis. CDL contributed to initial data collection and curation, performed some of the statistical analyses and prepared the display items. MK provided resources. All authors wrote the manuscript and approved its final version.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

1. Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260-270.
2. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. 2018. Current understanding of the human microbiome. Nat Med 24:392-400.
3. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Mujagond P. 2018. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 24:1532-1535.
4. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun 8:1784.
5. Reitmeier S, Kiessling S, Clavel T, List M, Almeida EL, Ghosh TS, Neuhaus K, Grallert H, Linseisen J, Skurk T, Brandl B, Breuninger TA, Troll M, Rathmann W, Linkohr B, Hauner H, Laudes M, Franke A, Le Roy CI, Bell JT, Spector T, Baumbach J, O'Toole PW, Peters A, Haller D. 2020. Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. Cell Host Microbe 28:258-272.e6.
6. Su X, Jing G, Sun Z, Liu L, Xu Z, McDonald D, Wang Z, Wang H, Gonzalez A, Zhang Y, Huang S, Huttley G, Knight R, Xu J. 2020. Multiple-disease detection and classification across cohorts via microbiome search. mSystems 5:e00150-20.
7. Zhou Y, Xu ZZ, He Y, Yang Y, Liu L, Lin Q, Nie Y, Li M, Zhi F, Liu S, Amir A, González A, Tripathi A, Chen M, Wu GD, Knight R, Zhou H, Chen Y. 2018. Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. mSystems 3:e00188-17.
8. Zheng Y, Fang Z, Xue Y, Zhang J, Zhu J, Gao R, Yao S, Ye Y, Wang S, Lin C, Chen S, Huang H, Hu L, Jiang GN, Qin H, Zhang P, Chen J, Ji H. 2020. Specific gut microbiome signature predicts the early-stage lung cancer. Gut Microbes 11:1030-1042.
9. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, McKay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579:567-574.
10. Ananthakrishnan AN, Luo C, Yajnik V, Khalili H, Garber JJ, Stevens BW, Cleland T, Xavier RJ. 2017. Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. Cell Host Microbe 21:603-610.e3.
11. Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, Leeming E, Gibson R, Le Roy C, Khatib HA, Francis L, Mazidi M, Mompeo O, Valles-Colomer M, Tett A, Beghini F, Dubois L, Bazzani D, Thomas AM, Mirzayi C, Khleborodova A, Oh S, Hine R, Bonnett C, Capdevila J, Danzanvilliers S, Giordano F, Geistlinger L, Waldron L, Davies R, Hadjigeorgiou G, Wolf J, Ordovás JM, Gardner C, Franks PW, Chan AT, Huttenhower C, Spector TD, Segata N. 2021. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nat Med 27:321-332.
12. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, Suez J, Mahdi JA, Matot E, Malka G, Kosower N, Rein M, Zilberman-Schapira G, Dohnalová L, Pevsner-Fischer M, Bikovsky R, Halpern Z, Elinav E, Segal E. 2015. Personalized nutrition by prediction of glycemic responses. Cell 163:1079-1094.
13. Díez López C, Vidaki A, Ralf A, Montiel González D, Radjabzadeh D, Kraaij R, Uitterlinden AG, Haas C, Lao O, Kayser M. 2019. Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. Forensic Sci Int Genet 41:72-82.
14. Díez López C, Montiel González D, Haas C, Vidaki A, Kayser M. 2020. Microbiome-based body site of origin classification of forensically relevant blood traces. Forensic Sci Int Genet:102280.
15. Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, Knight R, Metcalf JL. 2018. Microbiome data accurately predicts the postmortem interval using random forest regression models. Genes (Basel) 9:104.
16. Yang J, Tsukimi T, Yoshikawa M, Suzuki K, Takeda T, Tomita M, Fukuda S. 2019. Cutibacterium acnes (Propionibacterium acnes) 16S rRNA genotyping of microbial samples from possessions contributes to owner identification. mSystems 4:e00594-19.
17. Woerner AE, Novroski NMM, Wendt FR, Ambers A, Wiley R, Schmedes SE, Budowle B. Forensic human identification with targeted microbiome markers using nearest neighbor classification. Forensic Sci Int Genet 38:130-139.
18. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. Front Microbiol 8:2224.
19. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R. 2016. Tiny microbes, enormous impacts: what matters in gut microbiome studies? Genome biology 17:217-217.

20. Thabtah F, Hammoud S, Kamalov F, Gonsalves A. 2020. Data imbalance in classification: Experimental evaluation. Inf Sci 513:429-441.
21. Japkowicz N, Stephen S. 2002. The class imbalance problem: A systematic study. Intel Data Anal 6:429-449.
22. Abd Elrahman SM, Abraham A. 2013. A review of class imbalance problem. J Netw 1:332-340.
23. Ali A, Shamsuddin SM, Ralescu AL. 2013. Classification with class imbalance problem. Int J Advance Soft Compu Appl 5.
24. Knights D, Costello EK, Knight R. 2011. Supervised classification of human microbiota. FEMS Microbiol Rev 35:343-359.
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321-357.
26. Gomez-Alvarez V, Revetta RP. 2020. Monitoring of nitrification in chloraminated drinking water distribution systems with microbiome bioindicators using supervised machine learning. Front Microbiol 11:571009.
27. Mehta SD, Zhao D, Green SJ, Agingu W, Otieno F, Bhaumik R, Bhaumik D, Bailey RC. 2020. The microbiome composition of a man's penis predicts incident bacterial vaginosis in his female sex partner with high accuracy. Front Cell Infect Microbiol 10:433.
28. Chen L, Li Z, Zeng T, Zhang Y-H, Liu D, Li H, Huang T, Cai Y-D. 2020. Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes. Front Mol Biosci 7:604794.
29. Brooks AW, Priya S, Blekhman R, Bordenstein SR. 2018. Gut microbiota diversity across ethnicities in the United States. PLoS Biol 16:e2006842.
30. Wingfield B, Coleman S, McGinnity TM, Bjourson AJ. 2018. Robust microbial markers for non-invasive inflammatory bowel disease identification. IEEE/ACM Trans Comput Biol Bioinform 16:2078-2088.
31. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning, p 1322-1328. *In* (ed), IEEE,
32. Sayyari E, Kawas B, Mirarab S. 2019. TADA: phylogenetic augmentation of microbiome samples enhances phenotype classification. Bioinformatics 35:i31-i40.
33. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. ISME J 10:2435-2446.
34. Beghini F, Renson A, Zolnik C, Geistlinger L, Usyk M, Moody TU, Thorpe L, Dowd JB, Burk R, Segata N, Jones HE, Waldron LD. 2019. Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. Ann Epidemiol 34:18-25.
35. Rodriguez-Rabassa M, Lopez P, Rodriguez-Santiago RE, Cases A, Felici M, Sanchez R, Yamamura Y, Rivera-Amill V. 2018. Cigarette smoking modulation of saliva microbial composition and cytokine levels. Int J Environ Res Public Health 15:2479.
36. Kato I, Vasquez AA, Moyerbrailean G, Land S, Sun J, Lin H-S, Ram JL. 2016. Oral microbiome and history of smoking and colorectal cancer. J Epidemiol Res 2:92-101.
37. Takeshita T, Kageyama S, Furuta M, Tsuboi H, Takeuchi K, Shibata Y, Shimazaki Y, Akifusa S, Ninomiya T, Kiyohara Y, Yamashita Y. 2016. Bacterial diversity in saliva and oral health-related conditions: the Hisayama Study. Sci Rep 6:22164.
38. Sato N, Kakuta M, Hasegawa T, Yamaguchi R, Uchino E, Kobayashi W, Sawada K, Tamura Y, Tokuda I, Murashita K, Nakaji S, Imoto S, Yanagita M, Okuno Y. 2020. Metagenomic analysis of bacterial species in tongue microbiome of current and never smokers. npj Biofilms and Microbiomes 6:11.
39. Fagerström K. 2002. The epidemiology of smoking. Drugs 62:1-9.
40. Kayser M. 2015. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. Forensic Sci Int Genet 18:33-48.
41. Vidaki A, Kayser M. 2017. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. Genome Biol 18:238.
42. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Data from 'Cigarette smoking and the oral microbiome in a large study of American adults'. European Nucleotide Archive https://www.ebi.ac.uk/ena/browser/view/PRJNA434300.
43. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Data from 'Cigarette smoking and the oral microbiome in a large study of American adults'. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJNA434312.
44. Beghini F, Renson A, Zolnik C, Geistlinger L, Usyk M, Moody TU, Thorpe L, Dowd JB, Burk R, Segata N, Jones HE, Waldron LD. 2019. Data from 'Tobacco exposure associated with oral microbiota oxygen

utilization in the New York City Health and Nutrition Examination Study'. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJNA484874.
45. Burcham ZM, Garneau NL, Comstock SS, Tucker RM, Knight R, Metcalf JL, Genetics of Taste Lab Citizen S. 2020. Patterns of oral microbiota diversity in adults and children: A crowdsourced population study. Sci Rep 10:2133.
46. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet 17:3.
47. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581-583.
48. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. 2018. New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract. mSystems 3:e00187-18.
49. Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10:1200-1202.
50. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, Kocher JP, Eckel-Passow JE. 2013. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics 29:2877-83.
51. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnology 37:852-857.
52. Lemaître G, Nogueira F, Aridas CK. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18:559-563.
53. Wright ES. 2016. Using DECIPHER v2. 0 to analyze big biological sequence data in R. R J 8.
54. Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059-3066.
55. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately maximum-likelihood trees forl large alignments. PLoS One 5:e9490.
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825-2830.
57. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 6:10.
58. Cawley GC, Talbot NLC. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 11:2079-2107.
59. Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:6.
60. Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. Journal of machine learning research 3:1157-1182.
61. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. 2005. Defining the normal bacterial flora of the oral cavity. J Clin Microbiol 43:5721-5732.
62. Zaura E, Keijser BJF, Huse SM, Crielaard W. 2009. Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9:259.
63. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C, Izard J. 2012. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13:R42.
64. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. 2015. The microbiome quality control project: baseline study design and future directions. Genome Biol 16:276.
65. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Abnet CC, Knight R, White O, Huttenhower C, The Microbiome Quality Control Project C. 2017. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol 35:1077-1086.
66. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA, Vaginal Microbiome C. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 15:66.
67. Wang Y, LêCao K-A. 2019. Managing batch effects in microbiome data. Brief Bioinform 21:1954-1970.

68. Boughorbel S, Jarray F, El-Anbari M. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One 12:e0177678.
69. Ballabio D, Grisoni F, Todeschini R. 2018. Multivariate comparison of classification performance measures. Chemom Intell Lab Syst 174:33-44.
70. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. 2020. A framework for effective application of machine learning to microbiome-based classification problems. mBio 11:e00434-20.
71. Carrieri AP, Haiminen N, Maudsley-Barton S, Gardiner L-J, Murphy B, Mayes AE, Paterson S, Grimshaw S, Winn M, Shand C. 2021. Explainable AI reveals changes in skin microbiome composition linked to phenotypic differences. Sci Rep 11:1-18.
72. Huttenhower C, Knight R, Brown CT, Caporaso JG, Clemente JC, Gevers D, Franzosa EA, Kelley ST, Knights D, Ley RE, Mahurkar A, Ravel J, Scientists for Advancement of Microbiome R, White O. 2014. Advancing the microbiome research community. Cell 159:227-230.
73. Lira-Junior R, Åkerman S, Klinge B, Boström EA, Gustafsson A. 2018. Salivary microbial profiles in relation to age, periodontal, and systemic diseases. PLoS One 13:e0189374.
74. Liu S, Wang Y, Zhao L, Sun X, Feng Q. 2020. Microbiome succession with increasing age in three oral sites. Aging (Albany NY) 12:7874.
75. Yang Y, Zheng W, Cai Q, Shrubsole MJ, Pei Z, Brucker R, Steinwandel M, Bordenstein SR, Li Z, Blot WJ, Shu X-O, Long J. 2019. Racial differences in the oral microbiome: Data from low-income populations of African ancestry and European ancestry. mSystems 4:e00639-19.
76. Mason MR, Nagaraja HN, Camerlengo T, Joshi V, Kumar PS. 2013. Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. PLoS One 8:e77287.
77. Gómez D, Rojas A. 2016. An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. Neural Comput 28:216-228.

SUPPLEMENTARY MATERIALS

Supplementary Table S1. Experimental characteristics of the two analysed studies.

Study ID	**S1**	**S2**
EMBL-EBI Accession Number	PRJNA434300 (ACS CPS-II cohort) PRJNA434312 (NCI PLCO cohort)	PRJNA484874
Current smokers definition	NA	>100 lifetime smoked cigarettes, smoked a cigarette in the last 5 days and did not use any alternative tobacco product in the last 5 days
Never smokers definition	NA	<100 lifetime smoked cigarettes, no usage of any tobacco product in the last 5 days and serum cotinine levels less than 0.05 ng/mg
Former smokers definition	NA	>100 lifetime smoked cigarettes, currently not smoking, no usage of any tobacco product in the last 5 days and serum cotinine levels less than 0.05 ng/mL
DNA Isolation Kit	Mo Bio PowerSoil DNA Isolation Kit (Mo Bio Laboratories)	QIAamp DNA Mini Kit (QIAGEN)
16S rRNA gene region(s)	V3-V4	V4
Sequencing Platform	454 Roche FLX Titanium	Illumina MiSeq
Sequencing reads	Single-end	2x300 paired-end

Supplementary Table S2. Validation of data types with machine learning (ML) methods for microbiome-based prediction of smoking habits based on the S1 and S2 datasets together. For each ML method, we evaluated six types of input data as described in the study methods section: baseline non-augmented, and five augmented dataset based on different methods ADASYN-1, ADASYN-2, SMOTE-1, SMOTE-2 and TADA. The Mathews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUC) values are the average (and standard deviation) of the performance metrics resulting from the five-fold nested cross-validation repeated for ten times (5*10). For MCC, +1 indicates perfect prediction, 0 random prediction and –1 perfect inverse prediction. For AUC, 1 indicates perfectly accurate prediction and 0.5 indicates random prediction. ML method abbreviations: DT, decision trees; KNN, k nearest neighbors; LR, logistic regression; RF, random forest; SVML, support vector machine with linear kernel; SVMR, support vector machine with radial kernel; XGBoost, extreme gradient boosting.

ML method	Data type	MCC [average (sd)]	AUC [average (sd)]
DT	Non-augmented	0.09 (0.12)	0.65 (0.06)
	ADASYN-1	0.13 (0.08)	0.61 (0.06)
	ADASYN-2	0.14 (0.09)	0.59 (0.05)
	SMOTE-1	0.14 (0.06)	0.61 (0.05)
	SMOTE-2	0.13 (0.07)	0.59 (0.05)
	TADA	0.13 (0.13)	0.65 (0.05)
KNN	Non-augmented	0.25 (0.10)	0.70 (0.05)
	ADASYN-1	0.19 (0.06)	0.67 (0.05)
	ADASYN-2	0.18 (0.07)	0.65 (0.05)
	SMOTE-1	0.20 (0.06)	0.67 (0.06)
	SMOTE-2	0.18 (0.06)	0.67 (0.05)
	TADA	0.21 (0.09)	0.68 (0.06)
LR	Non-augmented	0.16 (0.11)	0.71 (0.07)
	ADASYN-1	0.16 (0.09)	0.62 (0.07)
	ADASYN-2	0.16 (0.06)	0.63 (0.05)
	SMOTE-1	0.18 (0.09)	0.64 (0.06)
	SMOTE-2	0.18 (0.08)	0.64 (0.06)
	TADA	0.19 (0.07)	0.65 (0.06)
RF	Non-augmented	0.17 (0.10)	0.71 (0.05)
	ADASYN-1	0.28 (0.08)	0.72 (0.04)
	ADASYN-2	0.21 (0.07)	0.70 (0.05)
	SMOTE-1	0.26 (0.09)	0.71 (0.06)
	SMOTE-2	0.21 (0.08)	0.70 (0.05)
	TADA	0.15 (0.09)	0.71 (0.05)
SVML	Non-augmented	0.00 (0.00)	0.70 (0.06)
	ADASYN-1	0.31 (0.07)	0.75 (0.05)
	ADASYN-2	0.31 (0.06)	0.76 (0.04)
	SMOTE-1	0.33 (0.06)	0.76 (0.04)
	SMOTE-2	0.32 (0.08)	0.76 (0.05)
	TADA	0.31 (0.06)	0.75 (0.05)
SVMR	Non-augmented	0.11 (0.09)	0.67 (0.06)
	ADASYN-1	0.22 (0.09)	0.67 (0.07)
	ADASYN-2	0.19 (0.07)	0.65 (0.05)
	SMOTE-1	0.22 (0.07)	0.67 (0.06)
	SMOTE-2	0.22 (0.08)	0.67 (0.05)
	TADA	0.21 (0.09)	0.66 (0.07)
XGBoost	Non-augmented	0.11 (0.10)	0.61 (0.08)
	ADASYN-1	0.26 (0.07)	0.70 (0.05)
	ADASYN-2	0.21 (0.08)	0.70 (0.04)
	SMOTE-1	0.26 (0.08)	0.71 (0.05)
	SMOTE-2	0.20 (0.08)	0.70 (0.06)
	TADA	0.17 (0.08)	0.72 (0.06)

Supplementary Table S3. Wilcoxon test pair comparisons between the different types of input data with the support vector machine with linear kernel (SVML) method without feature selection (*f*=124) and with feature selection (*f*=19). Comparisons are based on the Mathew's correlation coefficient (MCC) values resulted from the five-fold nested crossed-validation repeated for ten times (5*10). Significant *p* values were determined with a value cut-off of 0.05 after Benjamini-Hochberg (BH) correction.

Group 1	**Group 2**	**BH *p* without feature selection (*f*=124)**	**BH *p* with feature selection (*f*=19)**
Non-augmented	ADASYN-1	9.93E-20	9.93E-20
Non-augmented	ADASYN-2	9.93E-20	9.93E-20
Non-augmented	SMOTE-1	9.93E-20	9.93E-20
Non-augmented	SMOTE-2	9.93E-20	9.93E-20
Non-augmented	TADA	9.93E-20	9.93E-20
ADASYN-1	ADASYN-2	0.595	0.983
ADASYN-1	SMOTE-1	0.518	0.799
ADASYN-1	SMOTE-2	0.682	0.739
ADASYN-1	TADA	0.882	0.799
ADASYN-2	SMOTE-1	0.124	0.799
ADASYN-2	SMOTE-2	0.416	0.739
ADASYN-2	TADA	0.682	0.799
SMOTE-1	SMOTE-2	0.741	0.825
SMOTE-1	TADA	0.392	0.960
SMOTE-2	TADA	0.595	0.799

Supplementary Table S4. Feature selection analysis for differences between smokers *vs.* non-smokers across bacterial species using the two-sided Mann-Whitney U test and Benjamini-Hochberg correction for significant *p* values ($p<0.05$) in at least 90% of the *k*-folds of training datasets (45 of 50 total folds).

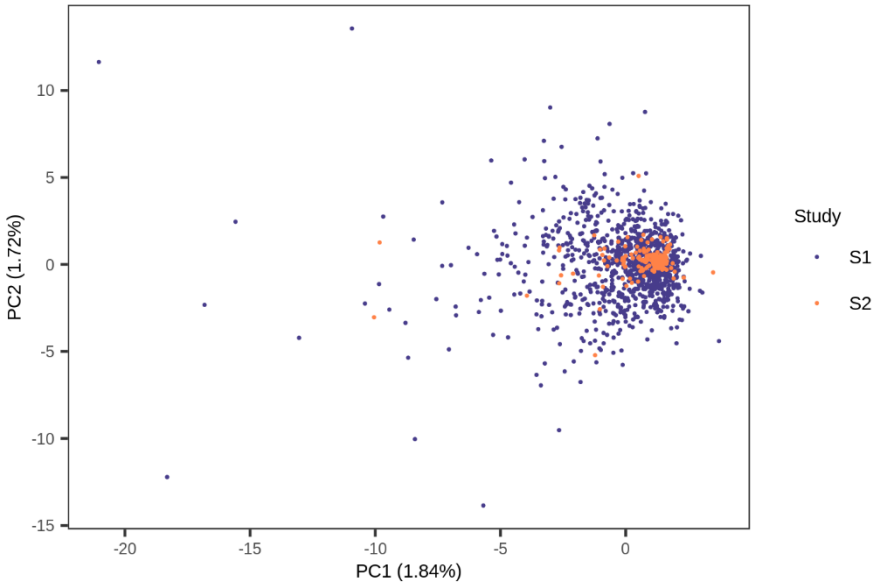
Data partition	Streptococcus parasangunis clade 411	Neisseria NA	Prevotella veroralis	Actinomyces sp. HMT 180	Rothia aeria	Gemella morbillorum	Veillonella rogosae	Abiotrophia defectiva
0	2.02E-02	8.16E-06	5.84E-03	5.18E-03	5.78E-02	5.82E-06	4.35E-04	5.55E-03
1	5.15E-03	3.79E-05	4.02E-04	8.67E-04	2.81E-02	2.82E-05	1.68E-04	1.03E-03
2	1.05E-02	6.06E-08	1.19E-04	4.38E-04	6.55E-03	1.25E-08	3.43E-05	4.40E-03
3	1.14E-02	1.02E-06	2.60E-03	9.21E-03	3.51E-03	7.55E-07	2.41E-04	1.47E-03
4	1.90E-03	1.23E-05	4.72E-03	9.06E-04	6.50E-02	7.73E-06	1.19E-04	1.16E-02
5	1.73E-03	1.56E-07	8.14E-04	4.19E-03	1.51E-03	2.44E-06	6.35E-04	2.89E-04
6	6.67E-03	1.31E-06	9.03E-03	2.15E-03	4.81E-04	7.18E-06	5.42E-04	1.41E-04
7	1.13E-02	4.12E-06	6.46E-04	1.15E-02	9.70E-04	3.34E-05	3.36E-03	1.52E-04
8	2.11E-02	2.00E-07	1.09E-03	1.19E-03	5.50E-04	1.44E-05	6.17E-05	9.57E-05
9	1.03E-03	1.77E-07	1.02E-03	8.17E-03	1.57E-04	1.05E-07	1.19E-03	3.53E-04
10	8.97E-03	1.50E-05	3.36E-02	1.14E-02	9.31E-03	9.25E-07	1.05E-04	4.88E-03
11	3.33E-02	1.20E-05	6.14E-02	3.21E-02	1.09E-02	1.05E-05	5.92E-04	5.69E-03
12	1.07E-02	5.12E-07	2.88E-03	3.02E-02	6.49E-04	4.75E-07	4.00E-04	2.89E-03
13	9.58E-03	3.73E-05	1.12E-02	6.12E-03	1.27E-03	1.14E-07	4.33E-04	6.42E-04
14	2.57E-02	1.55E-07	1.53E-02	5.11E-03	9.34E-04	2.12E-09	2.24E-03	4.40E-03
15	1.32E-02	8.13E-07	1.80E-02	2.26E-03	6.75E-03	5.66E-09	7.36E-04	6.77E-04
16	3.67E-02	4.33E-06	1.11E-02	6.19E-04	5.31E-03	9.94E-08	1.85E-04	2.92E-03
17	3.18E-03	1.57E-06	2.64E-03	4.50E-03	5.68E-04	1.09E-05	2.88E-03	1.69E-03
18	5.69E-02	2.04E-06	2.08E-04	4.60E-03	9.88E-03	2.02E-06	3.78E-05	5.78E-03
19	1.45E-02	2.73E-04	3.16E-03	2.34E-02	6.32E-03	5.25E-04	7.38E-03	6.12E-03
20	6.71E-02	2.44E-07	3.72E-03	1.73E-03	3.47E-03	6.19E-08	6.83E-04	2.02E-02
21	1.95E-02	1.85E-06	4.19E-02	3.40E-03	2.43E-02	1.75E-04	2.30E-04	1.94E-02
22	2.88E-02	7.85E-07	1.37E-04	2.75E-03	4.98E-03	5.92E-07	4.79E-04	1.08E-02
23	1.46E-02	2.27E-06	1.24E-02	3.05E-03	1.17E-02	1.00E-06	5.00E-03	3.05E-03
24	3.13E-02	8.38E-08	4.63E-03	1.78E-04	3.60E-03	5.46E-09	1.86E-04	6.90E-03
25	3.08E-03	3.26E-08	5.04E-02	1.38E-02	7.24E-04	1.54E-05	1.92E-03	5.58E-03
26	6.04E-02	6.02E-08	1.11E-02	1.09E-02	2.78E-03	1.39E-07	1.32E-02	5.45E-04
27	5.75E-02	2.10E-07	1.84E-02	1.90E-02	6.93E-03	4.94E-08	4.32E-03	2.53E-03
28	3.14E-02	1.05E-06	2.69E-02	5.51E-03	1.65E-02	1.35E-08	2.60E-03	1.54E-04
29	7.57E-03	3.48E-08	9.70E-04	7.02E-03	1.88E-03	2.05E-09	9.75E-04	7.48E-03
30	2.48E-04	9.72E-10	8.15E-03	5.37E-05	3.58E-03	8.76E-07	3.02E-06	2.40E-04
31	6.30E-03	1.17E-06	1.19E-02	9.41E-03	8.89E-03	2.93E-05	7.93E-04	9.54E-04
32	1.66E-03	4.81E-06	7.14E-03	5.47E-03	1.71E-03	6.11E-04	1.12E-04	3.34E-04
33	1.13E-03	9.44E-07	8.48E-04	1.45E-02	1.29E-03	7.31E-06	4.06E-04	1.06E-03
34	1.38E-02	7.65E-08	5.26E-03	4.81E-03	6.65E-03	2.30E-08	3.91E-04	1.17E-03
35	5.64E-03	6.00E-06	8.72E-04	2.13E-03	5.50E-03	4.89E-07	9.47E-04	1.79E-03
36	1.24E-02	3.69E-07	1.21E-02	9.58E-04	1.10E-02	1.92E-06	8.78E-06	1.06E-02
37	3.43E-03	3.91E-07	2.24E-03	5.19E-04	5.32E-03	2.05E-07	7.48E-06	1.46E-03
38	6.85E-03	2.49E-06	1.90E-03	9.43E-04	1.24E-02	1.18E-06	3.88E-04	3.12E-03
39	1.63E-03	2.05E-07	4.50E-03	1.24E-04	2.92E-03	6.49E-05	8.50E-05	2.06E-03
40	7.19E-03	2.38E-07	6.10E-04	1.02E-02	5.67E-03	2.72E-07	3.84E-04	4.56E-03
41	8.07E-03	4.62E-06	4.99E-04	2.67E-03	5.69E-03	5.12E-06	2.02E-03	1.71E-03
42	7.23E-03	7.36E-07	1.45E-03	9.09E-03	6.76E-03	1.76E-06	3.89E-04	6.42E-03
43	5.09E-03	4.59E-07	4.25E-03	5.08E-03	5.91E-03	6.75E-06	2.17E-04	3.02E-03
44	2.10E-02	3.05E-06	1.39E-02	4.63E-03	4.51E-02	1.51E-05	6.77E-04	4.21E-03
45	1.72E-03	2.67E-06	3.84E-03	4.96E-02	1.15E-03	4.81E-05	1.14E-03	6.58E-03
46	6.12E-04	7.63E-09	9.51E-05	5.83E-02	1.55E-03	3.78E-05	4.02E-05	5.64E-03
47	3.29E-03	1.06E-07	1.01E-03	4.61E-02	2.32E-03	1.06E-06	3.35E-05	1.76E-03
48	9.52E-03	1.12E-07	1.11E-02	2.35E-02	6.39E-03	4.59E-07	2.90E-03	2.12E-02
49	1.02E-02	1.47E-06	1.81E-04	9.66E-03	3.41E-03	2.27E-06	8.56E-04	3.71E-03

Data partition	Lachnoanaerobaculum umeaense	Lachnospiraceae [G-2] bacterium HMT 096	Neisseria oralis	Prevotella nigrescens	Corynebacterium durum	Neisseria cinerea	Capnocytophaga granulosa	Lactobacillus iners
0	6.03E-06	3.11E-03	5.97E-02	9.50E-03	1.72E-02	5.71E-03	5.61E-03	1.47E-05
1	7.57E-05	5.85E-03	9.23E-02	5.40E-03	1.75E-02	1.19E-02	5.47E-04	3.39E-05
2	2.87E-05	1.09E-02	1.61E-02	3.17E-05	1.03E-02	2.04E-03	2.30E-04	8.11E-07
3	3.30E-05	1.43E-03	2.33E-02	3.84E-04	1.11E-02	1.38E-02	1.11E-02	2.10E-06
4	5.56E-04	6.59E-03	5.84E-02	1.05E-02	7.70E-02	6.58E-04	1.08E-02	4.91E-08
5	1.60E-04	6.28E-04	1.76E-03	1.39E-02	7.77E-03	1.23E-03	5.76E-04	1.25E-03
6	7.99E-06	1.47E-03	1.91E-03	2.29E-02	5.59E-03	6.03E-04	4.47E-04	1.60E-06
7	5.34E-05	1.01E-02	3.04E-03	5.68E-03	5.20E-03	3.09E-04	1.60E-03	6.25E-06
8	1.43E-05	1.17E-03	4.88E-03	4.50E-03	1.15E-03	1.91E-03	1.43E-05	1.57E-07
9	8.31E-05	2.56E-04	1.06E-03	3.72E-03	3.95E-03	3.39E-04	1.16E-04	7.73E-10
10	1.17E-03	1.21E-03	1.69E-02	1.19E-02	5.41E-03	6.13E-04	1.66E-03	1.40E-08
11	5.49E-03	9.93E-03	4.01E-02	3.78E-02	1.51E-02	2.96E-04	2.35E-03	2.22E-04
12	1.57E-03	7.25E-04	6.82E-04	6.74E-04	2.90E-03	8.79E-04	7.17E-04	2.04E-04
13	1.38E-04	4.66E-04	6.78E-03	6.13E-03	3.64E-03	1.19E-05	1.21E-04	4.50E-05
14	9.88E-04	4.10E-04	6.49E-03	4.07E-03	1.16E-02	2.62E-04	1.04E-03	1.43E-04
15	9.77E-04	5.03E-03	7.27E-02	8.66E-04	8.85E-04	4.27E-03	1.15E-02	1.25E-04
16	1.19E-04	3.15E-03	1.53E-02	1.37E-04	1.11E-02	1.32E-03	2.76E-03	5.65E-06
17	1.39E-05	2.33E-03	3.67E-03	7.01E-03	2.50E-03	2.16E-03	2.35E-03	1.04E-05
18	1.09E-04	9.50E-03	3.11E-02	7.11E-03	6.91E-03	5.73E-03	5.91E-03	3.06E-06
19	2.63E-03	2.19E-02	4.72E-02	1.47E-02	3.53E-02	6.00E-03	4.87E-03	2.29E-03
20	6.10E-04	1.50E-03	5.70E-03	4.07E-03	1.06E-02	1.56E-03	1.78E-03	9.76E-08
21	7.36E-04	2.00E-03	8.76E-03	1.60E-02	4.46E-02	8.58E-03	6.00E-03	6.35E-04
22	1.26E-04	8.98E-04	6.77E-03	3.18E-02	1.10E-02	4.15E-04	4.18E-04	9.78E-08
23	3.02E-03	5.04E-04	1.12E-02	1.22E-02	7.42E-03	1.85E-04	3.60E-04	4.14E-09
24	6.30E-05	5.52E-04	3.47E-03	1.84E-03	3.90E-02	1.29E-05	2.30E-03	2.09E-07
25	1.28E-03	7.61E-04	2.95E-03	1.38E-02	1.24E-02	1.01E-02	1.72E-03	7.00E-04
26	3.83E-04	5.52E-03	2.42E-03	2.01E-02	9.41E-03	1.50E-02	3.02E-04	2.32E-05
27	5.52E-04	5.44E-03	9.48E-03	1.85E-02	1.94E-02	2.98E-03	2.88E-03	3.96E-03
28	2.36E-04	2.21E-04	1.99E-02	2.62E-02	2.03E-02	3.77E-04	5.58E-04	1.12E-04
29	1.05E-03	3.67E-03	2.28E-02	1.97E-03	9.84E-03	6.23E-04	9.12E-05	5.24E-06
30	7.50E-07	9.61E-05	7.34E-04	6.22E-02	6.16E-03	6.36E-02	5.42E-04	5.39E-07
31	1.36E-04	2.09E-05	6.85E-03	1.00E-02	1.06E-02	1.50E-03	3.74E-03	1.01E-03
32	1.10E-04	4.93E-04	1.73E-04	4.18E-02	3.27E-02	4.99E-03	1.77E-03	6.51E-09
33	6.25E-05	5.83E-04	9.44E-04	5.11E-02	1.34E-02	1.51E-02	8.54E-03	9.98E-09
34	3.35E-04	1.79E-03	9.31E-03	3.78E-03	1.33E-02	1.34E-02	1.16E-02	8.04E-07
35	8.69E-05	4.68E-03	1.48E-02	9.27E-04	1.64E-03	1.87E-03	8.75E-03	1.32E-10
36	7.11E-05	6.67E-03	6.78E-03	6.02E-03	1.34E-02	7.08E-03	6.98E-03	6.10E-06
37	3.96E-04	1.93E-03	4.46E-03	2.26E-03	4.58E-03	1.50E-03	6.60E-04	1.49E-06
38	2.33E-04	1.75E-03	3.92E-02	1.81E-03	3.09E-03	1.81E-04	2.03E-03	3.46E-06
39	1.15E-05	1.09E-03	3.38E-03	3.34E-03	2.98E-03	3.47E-03	9.60E-04	3.72E-08
40	1.49E-04	5.68E-04	3.37E-02	4.26E-03	9.79E-03	5.69E-04	1.25E-04	3.40E-07
41	4.56E-04	9.50E-04	2.16E-02	5.51E-02	4.81E-03	5.30E-04	2.47E-03	4.89E-07
42	1.29E-04	1.24E-03	7.30E-03	4.62E-02	7.02E-03	1.10E-03	6.37E-03	5.98E-06
43	1.57E-05	2.98E-03	3.20E-02	3.73E-02	2.57E-03	1.07E-02	1.13E-02	1.68E-06
44	4.03E-04	4.23E-03	2.17E-02	2.41E-01	5.62E-03	1.09E-04	3.59E-03	6.91E-06
45	3.47E-04	3.32E-03	1.31E-03	2.60E-03	1.37E-02	1.74E-03	3.55E-03	6.97E-05
46	3.14E-05	5.53E-03	1.86E-04	3.68E-03	8.07E-04	3.72E-03	1.68E-03	2.83E-05
47	1.49E-06	5.04E-03	1.50E-03	5.51E-03	2.47E-02	9.18E-02	1.69E-03	1.27E-06
48	8.79E-05	9.34E-03	4.60E-03	1.71E-03	1.66E-02	1.05E-02	3.29E-03	1.33E-05
49	1.80E-04	8.70E-03	3.44E-03	9.27E-03	5.76E-03	3.56E-03	1.45E-03	7.38E-07

Data partition	Capnocytophaga leadbetteri	Prevotella intermedia	Alloprevotella tannerae
0	1.84E-02	1.59E-02	1.35E-04
1	4.09E-04	1.14E-01	1.81E-03
2	1.13E-02	8.58E-04	1.80E-04
3	3.63E-03	2.22E-02	2.24E-04
4	3.87E-02	4.69E-03	9.03E-03
5	6.83E-04	2.08E-03	6.38E-04
6	4.57E-04	1.75E-02	2.39E-02
7	3.21E-05	9.85E-05	3.96E-03
8	1.37E-05	8.96E-03	5.19E-03
9	3.32E-04	3.39E-04	2.98E-04
10	5.89E-03	3.68E-03	1.30E-03
11	1.86E-02	1.76E-02	2.65E-02
12	2.68E-03	5.34E-03	6.95E-04
13	7.16E-03	1.48E-02	4.55E-03
14	3.11E-02	8.91E-03	2.94E-03
15	4.15E-03	3.54E-03	4.19E-03
16	2.02E-02	5.41E-05	2.17E-03
17	3.55E-03	3.85E-03	3.80E-03
18	2.26E-02	5.97E-03	5.62E-02
19	5.89E-03	6.41E-03	4.39E-03
20	1.03E-02	2.01E-03	2.63E-02
21	1.22E-02	1.16E-02	4.33E-02
22	6.06E-03	2.44E-03	6.53E-04
23	6.04E-03	1.48E-02	2.09E-03
24	7.07E-03	3.65E-03	4.17E-04
25	2.36E-02	1.17E-02	1.56E-02
26	4.27E-04	4.31E-02	1.06E-03
27	6.54E-03	6.85E-03	4.09E-03
28	1.97E-02	3.66E-03	2.02E-02
29	2.14E-02	7.97E-03	2.28E-03
30	9.40E-04	3.08E-02	3.08E-02
31	1.91E-03	1.50E-02	9.24E-02
32	1.87E-03	6.80E-02	3.14E-02
33	5.91E-03	5.44E-02	6.87E-02
34	4.75E-03	4.97E-03	1.37E-02
35	5.22E-03	3.71E-04	4.13E-04
36	8.98E-03	2.84E-03	6.29E-03
37	1.91E-04	4.08E-02	3.61E-03
38	2.25E-02	1.79E-02	2.89E-03
39	3.38E-03	1.15E-02	1.44E-02
40	6.15E-03	1.10E-03	1.10E-02
41	3.29E-02	9.22E-04	1.31E-02
42	1.28E-02	2.36E-03	5.32E-03
43	1.10E-02	2.83E-03	2.03E-02
44	2.36E-02	3.30E-03	3.47E-03
45	1.66E-02	2.69E-03	2.40E-02
46	5.56E-03	8.36E-03	1.07E-02
47	3.98E-03	3.22E-03	5.04E-03
48	2.77E-02	2.16E-02	3.71E-03
49	1.79E-02	3.78E-03	1.72E-03

Supplementary Table S5. Validation of data types with machine learning (ML) methods for microbiome-based prediction of smoking habits based on the S1 and S2 datasets together after initial feature selection (*f*=25). For each ML method, we evaluated six types of input data as described in the study methods section: baseline non-augmented, and five augmented dataset based on different methods ADASYN-1, ADASYN-2, SMOTE-1, SMOTE-2 and TADA. The Mathews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUC) values are the average (and standard deviation) of the performance metrics resulting from the five-fold nested cross-validation repeated for ten times (5*10). For MCC, +1 indicates perfect prediction, 0 random prediction and –1 perfect inverse prediction. For AUC, 1 indicates perfectly accurate prediction and 0.5 indicates random prediction. ML method abbreviations: DT, decision trees; KNN, k nearest neighbors; LR, logistic regression; RF, random forest; SVML, support vector machine with linear kernel; SVMR, support vector machine with radial kernel; XGBoost, extreme gradient boosting.

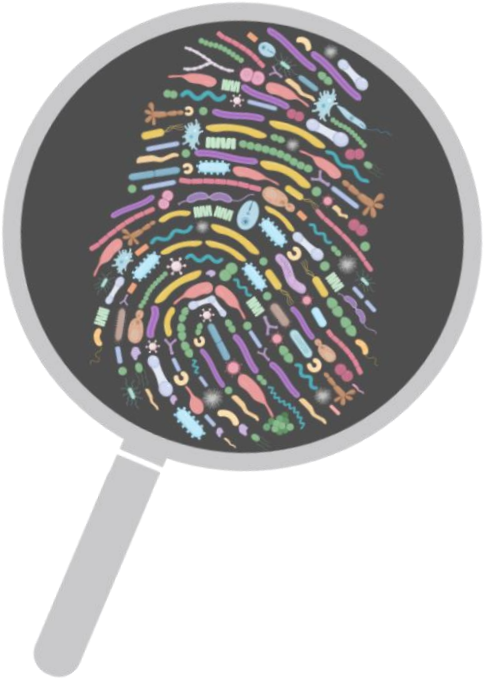
ML method	Data type	MCC [average (sd)]	AUC [average (sd)]
DT	Non-augmented	0.19 (0.09)	0.66 (0.06)
	ADASYN-1	0.13 (0.07)	0.61 (0.04)
	ADASYN-2	0.14 (0.08)	0.60 (0.06)
	SMOTE-1	0.14 (0.08)	0.62 (0.05)
	SMOTE-2	0.14 (0.09)	0.61 (0.06)
	TADA	0.16 (0.07)	0.61 (0.05)
KNN	Non-augmented	0.19 (0.08)	0.68 (0.07)
	ADASYN-1	0.18 (0.06)	0.66 (0.05)
	ADASYN-2	0.16 (0.07)	0.64 (0.05)
	SMOTE-1	0.20 (0.06)	0.68 (0.05)
	SMOTE-2	0.18 (0.08)	0.66 (0.06)
	TADA	0.19 (0.07)	0.66 (0.06)
LR	Non-augmented	0.13 (0.08)	0.74 (0.05)
	ADASYN-1	0.29 (0.06)	0.74 (0.05)
	ADASYN-2	0.28 (0.06)	0.74 (0.05)
	SMOTE-1	0.30 (0.08)	0.74 (0.05)
	SMOTE-2	0.31 (0.08)	0.74 (0.05)
	TADA	0.29 (0.08)	0.74 (0.06)
RF	Non-augmented	0.17 (0.08)	0.72 (0.05)
	ADASYN-1	0.24 (0.08)	0.69 (0.05)
	ADASYN-2	0.21 (0.08)	0.68 (0.05)
	SMOTE-1	0.26 (0.09)	0.70 (0.06)
	SMOTE-2	0.24 (0.07)	0.69 (0.05)
	TADA	0.23 (0.08)	0.71 (0.05)
SVML	Non-augmented	0.00 (0.00)	0.66 (0.06)
	ADASYN-1	0.31 (0.07)	0.75 (0.04)
	ADASYN-2	0.31 (0.08)	0.75 (0.05)
	SMOTE-1	0.33 (0.07)	0.75 (0.05)
	SMOTE-2	0.33 (0.07)	0.75 (0.04)
	TADA	0.33 (0.07)	0.75 (0.05)
SVMR	Non-augmented	0.01 (0.04)	0.65 (0.10)
	ADASYN-1	0.29 (0.08)	0.74 (0.05)
	ADASYN-2	0.26 (0.07)	0.73 (0.05)
	SMOTE-1	0.30 (0.06)	0.74 (0.04)
	SMOTE-2	0.30 (0.07)	0.74 (0.05)
	TADA	0.29 (0.07)	0.74 (0.04)
XGBoost	Non-augmented	0.09 (0.10)	0.61 (0.10)
	ADASYN-1	0.21 (0.08)	0.68 (0.05)
	ADASYN-2	0.19 (0.08)	0.68 (0.05)
	SMOTE-1	0.22 (0.08)	0.69 (0.05)
	SMOTE-2	0.20 (0.07)	0.69 (0.06)
	TADA	0.22 (0.10)	0.70 (0.05)



Supplementary Figure S1. Two-dimensional plot from principal component analysis (PCA) of the microbiome data comprising the 124 species in common between the two analysed studies. In the first two PCs, which together explain 3.56% of the total variation, the samples originating from the two studies cluster together.

Chapter 6

General discussion



This thesis was devised based on the motivation to investigate novel uses of the human microbiome to address forensically relevant questions, commonly encountered in forensic casework practice. These questions include 'What happened at the crime scene?' (*Part I*), 'When did the crime happened?' (*Part II*) and 'Who committed the crime?' (*Part III*). This motivation stemmed from the ample and rapid advances on the characterization of the human microbiome at a wide variety of body sites (1-3) and their consequent association to the host's well-being and various different phenotypes (4-6). At the same time, the continuous development and improvement of the experimental and analysis tools facilitated the potential implementation of microbiome findings from a forensic standpoint. This thesis also benefited from available microbiome data deposited in public repositories, some of which were analysed as part of the presented studies.

Before starting this thesis, the forensic community had already pinpointed specific features of the human microbiome relevant in a forensic context, such as ubiquity (7), ability to response to changes in the environment (8, 9) and capacity of being shed, deposited and exchanged between people and places (10, 11). Various studies had already focused on different applications of the human microbiome in forensics, such as individual identification (12-14), post-mortem interval estimation (15-17), geolocation prediction (18-20) and tissue and body fluid identification (21-23), among others (24-27). However, some of these studies suffer from particular limitations that I, together with my collaborators, identified as points to consider and/or improve in our research. On the one hand, we aimed to assess the investigative feasibility of the human microbiome in forensic questions where other molecular-based approaches showed challenges and limitations (*Part I* and *Part II)*. On the other hand, we aimed to explore the potential of the human microbiome as a novel tool in more recently introduced applications (*Part III)*.

Technical aspects of microbiome research

Based on my own experience and in agreement with the literature, biases can be introduced at every step in the microbiome pipeline. Undesirable variation can originate from cofounding factors of biological, technical and computational nature that are independent of the factor(s) of interest but can sometimes have a comparable effect size (28-30). Such biases not only need to be considered through good study design but also needs to be acknowledged by detailed and fair reports in scientific publications. The Achilles' heel of microbiome research is with no doubt the long-lasting lack of consensus in both experimental and analytical methods. Although there is a steady flow of newly published studies recommending the 'best' practises in microbiome research, these are usually true for a particular study and its aim, and not necessarily transferable to other studies. More recently, multi-collaborative efforts such as the Microbiome Quality Control project (MBCQ) (30) and the International Human Microbiome Standards group (IHMS) (31), focus on the comprehensive evaluation of microbiome methods to promote the best practices through the field and increase the comparability of the results between studies. However, the ease of implementation of the identified practices in laboratories around the globe and the integration of microbiome data produced with other methods remain to be solved still.

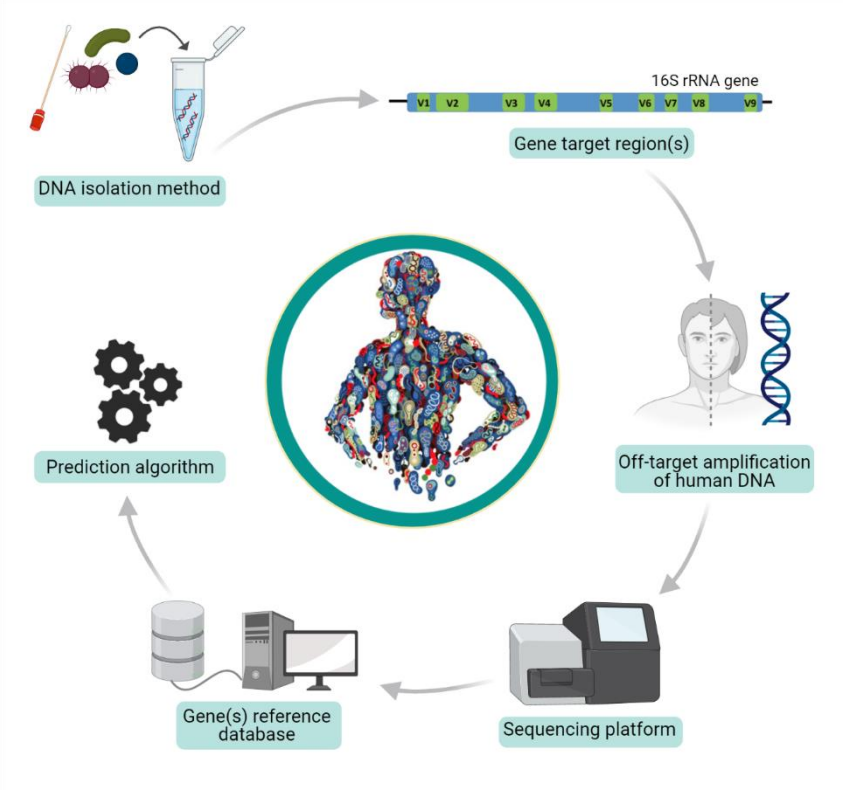


Figure 1. Steps and choices in the microbiome analysis pipeline that can introduce undesirable biases. (Picture created with BioRender).

DNA isolation

The DNA isolation method of choice is the step in the microbiome pipeline that can introduce the greatest variation in the obtained microbiome results (28, 32). This has been demonstrated by numerous studies in which different DNA isolation methodologies led to differences in the abundance of specific bacterial groups (33-36). The reason is that some bacterial cells are harder to lyse than others, such as Gram-positive bacteria compared to Gram-negative bacteria (37). One example is *Actinobacteria* (38), which is found in a variety of body sites from the skin (39) to the gut (40). Moreover, the DNA isolation method of choice also affects the yield and quality of the obtained microbial DNA (35, 41), as well as the presence of inhibitors that can affect downstream PCR efficiency (42), leading to preferential amplification and thus biased detection of taxa. However, this topic is not free of controversy since there are also quite a few studies reporting no significant effects based on the DNA isolation method (43-45). In our research (C*hapters 2, 3* and *4*), we made sure to employ a method that includes a bead-beating step, which has been linked with higher bacterial diversity; a more efficient DNA extraction from Gram-positive bacteria achieving higher DNA yields (38, 46, 47). I also

carried out an in-house optimization of the DNA isolation method to be used prior to analysis, to ensure optimal DNA yield and quality.

Another factor is the contamination introduced by the DNA isolation method itself, known as the 'kitome', which not only varies between kit manufacturers, but also between batches (48). Hence, we always included negative controls (DNA isolation and PCR reaction controls) in the experiments included in this thesis to monitor potential contamination. However, how to best account for the results obtained from these negative controls is still under debate. It is not a good practice to remove the taxa found in the controls since many overlap with those in the study samples (49). As previously recommended (50), in our studies, I combined data from quantitative PCR and sequencing to compare the absolute and relative abundances of potential contaminating taxa.

Finally, the DNA isolation method of choice can also affect the percentage of co-extracted human host DNA that is included in the obtained sequencing reads (35), which can become an important issue in samples with a very big human to bacteria ratio, such as venous blood (further details in *Off-target amplification of human DNA* later).

Gene target region(s)

Following the DNA isolation method of choice, the PCR is the step that can introduce the second greatest variation in the microbiome pipeline (28, 32). A PCR amplification step of a region within the gene of interest, such as the widely used 16S rRNA gene, is needed prior to sequencing since the entire gene is too long to be sequenced using short-read second-generation sequencing platforms. Although there is consensus in the suitability of the 16S rRNA gene, there is none on the 'best' region(s) to be analysed within this marker gene. Moreover, the PCR primer pair of choice affects the phylogenetic resolution, though there is not a real 'universal' set, able to equally amplify all bacterial groups. Yang *et al.* (51) reported that the V4-V6 regions are the most reliable to represent the full-length of the 16S rRNA gene. In our first research (*Chapter 2* and *Chapter 3*), we targeted the V1-V3 and the V3-V5 regions in two different PCR reactions that we subsequently pooled together, based on the publicly available reference microbiome data from the HMP. By this, we aimed to overcome the short-read limitation and expand the informative positions in the 16S rRNA gene that we used as input for the taxonomy-independent deep neural networks approach. In *Chapter 5*, the two original studies from which we obtained the publicly available saliva microbiome data targeted different regions within the 16S rRNA gene, but were overlapping at the V4 region. This, together with other different experimental factors, might partly explain the study-specific taxa we observed in *Chapter 5*.

Furthermore, in *Chapter 4* we looked for alternative essential genes to target using quantitative PCR (qPCR): beta subunit of RNA polymerase gene (rpoB) and translation initiation factor IF-2 gene (infB). Those genes were highly informative allowing us for species-level resolution, while targeting short amplicons. Nevertheless, targeting other genes than the commonly analysed 16S rRNA gene is not such a common practice in the field as it is normally necessary to design, optimize and validate the PCR primer sets.

Off-target amplification of human DNA

Before this thesis, non-specific, co-amplification of human host DNA by 16S rRNA gene PCR primers had been generally overlooked in the literature. However, the study in *Chapter 3* highlighted this artefact when analysing human blood-type samples. At that time, only a couple of studies mentioned this problem; one study that assessed the performance of 16S rRNA gene PCR primers in human clinical specimens (52) and an another that analysed human placenta biopsies (53). It is likely that researchers had not paid special attention to this issue previously, since in 16S rRNA gene data analysis it is a common practice to simply remove any sequencing read falling far from the average sequence length or not classified as of bacteria origin (54).

More recently, some studies have focused on the issue of off-target amplification of human host DNA in microbiome analysis. Sune *et al.* commented on the issue of co-amplification of human mitochondrial DNA that contains variants of the 16S rRNA gene (55). In a different study, Walker *et al.* (54) showed that, compared to the V3-V4 primers, the V1-V2 primers on average diminish 80% of the reads aligning to the human genome. Moreover, Pereira-Marques *et al.* (56) reported that this phenomena only occurred in samples with a very high human to bacteria ratio, particularly in human biopsy samples. However, we observed co-amplified human host DNA sequences in all human blood-type samples included in our study (*Chapter 3*), which could be partly explained by the use of different 16S rRNA gene PCR primers. The presence of sequencing reads aligning to the human genome reference sequence was much more notorious in our analysed venous and fingerprick blood samples than in the menstrual and nasal blood ones. This can be expected because the bacterial load in menstrual and nasal blood samples is higher than in venous and fingerprick blood ones. In other words, the lower the bacterial load, the more problematic these human genome-aligned reads are. It is of high importance that the research community is aware of this problem, especially since we used the 16S rRNA gene PCR primers recommended by the 16S 454 Sequencing Protocol of the Human Microbiome Project (HMP) Consortium (57), that other researchers also use as reference.

Sequencing platform

The sequencing platform is an important choice with specific advantages and issues associated with the insert size, read length, sequence accuracy and cost (58). Fortunately, data is reproducible across sequencing runs within and between different sequencing technologies (59). Even when obtained by different platforms, results have proven to be robust when the datasets are generated following the same protocol (DNA isolation method and 16S rRNA gene PCR primer pair) (60). We observed this in the research described in *Chapter 2* and *Chapter 3,* where the reference data from the HMP was produced with the Roche 454 pyrosequencing platform while ours with Ion Torrent instruments, despite keeping other potential sources of technical bias homogeneous.

The Illumina MiSeq is currently the platform of choice in the majority of 16S rRNA gene microbiome studies. This platform presents a high-throughput option that allows for high sequencing depth (61), it is supposed to produce high-quality data and allows for strict quality control parameters (60). However, even though the Illumina MiSeq errors are reported to be

around 0.01%, Kozich *et al.* (62) described errors as high as 10% and recommended the analysis of reads that provide with complete overlap between the forward and reverse paired-end sequencing to correct for that. We did not carry out specific studies on sequencing platform-specific errors but, from our experience, the Illumina MiSeq (*Chapter 4*) produces higher throughputs than the Ion Torrent instruments (*Chapter 2* and *Chapter 3*). In our research this was not a particular issue, since we followed a taxonomy-independent approach and focused on large microbial community differences across body sites (*Chapter 2* and *Chapter 3*). However, a lower throughput might be relevant in other studies where the diversity of the community is under-represented and especially in case informative rare taxa go unnoticed.

Gene(s) reference databases

The marker gene(s) reference database of choice that is employed for OTU picking and/or taxonomy assignation can sometimes affect the results. For instance, alignments by the Greengenes database are sometimes of poor quality, resulting in artificially-inflated richness and diversity estimates (63). Moreover, mistakes in the taxonomy assignment can occur due to PCR and/or sequencing errors (64) or by incorrect taxa labelling (65). Also, commonly used microbiome databases are biased towards the presence of clinically relevant taxa for humans, with discrepancies observed in environmental bacteria depending on the reference database used (66). This was not a particular issue in the research performed in this thesis, although it might be relevant in future forensic research dealing with samples exposed to other environmental conditions, where more than human-derived bacteria are expected.

Additionally, it has been shown that using niche-specific databases leads to an increase in lower taxonomic assignments, probably due to less competition among sequences than with larger databases (67), as we observed in *Chapter 4* and *Chapter 5* with the expanded Human Oral Microbiome Database (eHOMD) (68). However, care should be taken when using these types of databases, since they might not contain all the bacteria included in commercial mock communities that are commonly used to assess bias during PCR, library preparation, sequencing and initial analysis steps. An alternative would be to compare the mock community with a larger database or use an in-house mock sample that better reflects the bacteria of interest.

Prediction algorithms

Advances in high-throughput sequencing technologies have coincided with improvements in machine learning (ML) and its shift to the biological field for natural, physiological, lifestyle or disease trait prediction (69). Once the research community characterized the human microbiome and learnt from the microbial differences between different health statuses, the next step was to move to prediction. Most studies have focused so far on disease status prediction in the clinic based on microbiome data (70-74), which have proven the promise of ML methods for translational applications. However, one major issue is the lack of generalizable prediction methods applicable to microbiome datasets other than the study-specific one (3, 75, 76). For example, Kurilshikov *et al.* (77) analysed 16S rRNA gene faecal

microbiome data from 24 cohorts (N=18,340) and reported that only 9 out of 410 genera were detected in more than 95% of the samples. The total shared taxa is expected to be even smaller when lower taxonomic levels (e.g. species, strains) are analysed (78). Consequently, the missing data for the majority of the analysed subjects make it very complex to extend and apply one model to other cohorts. The origin of the missing data might be attributable to biological factors (e.g. young *vs.* elderly cohorts) or technical factors that can introduce biases throughout the microbiome pipeline as explained previously. We experienced this issue in the research described in *Chapter 5,* where the input features in the prediction model for smoking habits were only taxa common between the two analysed studies. Altogether, this emphasizes once again the need for standardization in microbiome research.

Moreover, microbiome studies should put more emphasis on the detailed and correct report of the ML pipeline used, as we tried in the research conducted as part of this thesis (*Chapters 2*, *3*, *4* and *5*). For example, the reasons for the ML of choice, the variation in the performance of different folds of cross-validation, as well as between validation and test sets (79). This can ultimately help other researchers better assess the most appropriate ML method for a given application and create awareness about the challenges and limitations encountered with different data types and ML methods, as we aimed in *Chapter 5*.

Microbiome data

The amount of microbiome data deposited in public repositories have increased exponentially in recent years, especially since the advances introduced by big consortia such as the HMP (80) in the mid 2000's. These data originate from big multi-collaborative efforts (e.g. American Gut Project (81)) as well as smaller studies analysing smaller sets of samples. So far, the gut has been the focus of the majority of the microbiome studies. This can be partly explained because the gut hosts the most complex and abundant microbial community in the body (82) and plays a key role in physical (83, 84) and mental health (gut-brain axis) (85). The review by Proctor *et al.* (86) on the microbiome research carried out at the United States National Institutes of Health (NIH) during a 10 year period (2007-2016) is a good proxy of the overall microbiome literature. According to the review, three quarters of the research has focused on just four body sites: gastrointestinal (GI) tract, urogenital tract (primarily vaginal), oral cavity and lung. From those, GI tract corresponds to 40% of the total research, which mainly analysed stool samples. Skin and nares microbiomes each represented 3% of the total research. The remaining included other body sites, tissues and systems, such as ear, eye, liver, blood, cardiovascular system and central nervous system.

This uneven research on different body sites limits the application of the human microbiome in forensics, where other tissues and body fluids rather than the GI tract are more relevant. This pushed us to find alternative solutions in our forensically motivated research. For instance, the HMP data repository we used as reference in *Chapter 3* lacked forensically relevant blood-type fluids. In consequence, as reference we used microbiome data from the body sites from which forensic blood stains could originate i.e. vaginal epithelium for menstrual blood, nasal mucosa for nasal blood and skin epithelium for fingerprick blood. In the case of venous blood we had no other choice but to *de novo* generate the required data due to the

absence of suitable publicly available microbiome sequencing data from venous blood that we could homogenize with our experimental approach in order to avoid undesirable biases. Nevertheless, in the future the forensic field will enormously benefit from public repositories that also include microbiome data from sites not only relevant for the clinic, but also commonly encountered in forensic investigations, such as sexually related stains (e.g. menstrual blood, seminal fluid) that can also be useful in other areas of (medical) research.

Moreover, most microbiome studies so far have focused on studying samples from individuals living in the USA and Europe. Whether their findings can be generalized to other populations is still poorly understood (87). More recently, studies are emerging from a few other countries with similar industrialization profiles and lifestyle including China, Japan and Israel, though there is a clear underrepresentation of populations in Africa, South America and other regions in Asia (e.g. India). These underrepresented regions possess distinct genetic (88, 89), ethnic (90, 91), sociocultural (92, 93) and lifestyle (94, 95) backgrounds that shape their microbiome. Other factors than the population of study are also uneven distributed across the microbiome literature, such as the donors' age. There are efforts from some cohorts to have an evenly distributed age range among participants (96). Nevertheless, this is not the case for the majority of the research, in which elderly people are over-represented based on the interest of age-related disorders or phenotypes with onset at older ages (9, 97, 98). We observed this age pattern in the study sample described in *Chapter 5*. From a forensic standpoint, it is important to have microbiome information from as many diverse samples as possible in order to assess the generalization of any proposed investigative tool. Future research needs to establish whether such human population differences impact on the forensic microbiome applications.

Another problem that needs to be pointed out is the high variability of metadata between studies, as we observed in the research presented in *Chapter 5*, where we used available 16S rRNA gene saliva microbiome data from two different studies. Though recent efforts are trying to set standards to overcome the lack of unified metadata in the field (99), the current situation makes sample reanalysis overly complex, since the available metadata is often misannotated, misleading and non-standardized. Importantly, metadata does not only refer to sample attributes, but also to sample preparation and processing (100). The variability of metadata can limit the conclusions, since the observed patterns might not be attributable to a certain factor(s) (101, 102) that should also not be ignored when selecting study subjects. This metadata issue might also restrict the performance of machine learning algorithms, where covariate adjustment (e.g. age, sex, ethnicity) can improve the predictions based on known influences on the microbiome.

Moreover, metadata that can potentially identify the donor or are of sensitive content (e.g. medical history) is usually kept confidential and only accessible to authorized users. Hence, researchers need to wait some time until getting access to that metadata. This can be a detrimental factor in cross-study applications where one should request metadata to several different sources and face disparate legal and privacy constrains (103, 104). Also, recently, the necessity of validating metadata submitted to public repositories by providing immediate and informative feedback to the submitting researchers has been highlighted (100).

Inter- and intra-individual variation of the human microbiome

In the studies carried out as part of this thesis and in agreement with the literature, we have come across substantial variation in the microbiome between- and within-individuals, which are discussed in this section. As described in the introduction of this thesis (*Chapter 1*), the magnitude of inter- and intra-individual variation depends greatly on the body site where the microbiome is studied (105). Thereby, variation is higher in skin and vaginal sites compared to oral and gut sites (3, 106). This variation can be measurable at a species level, where the same bacteria present different dynamics based on the body site they are detected (107, 108). Moreover, within a single body site, variation is also niche-specific. For instance, the nares or external auditory canal are more stable skin niches compared to the inner forearm (109). Here, I focus on the three main body sites analysed in our research; namely vagina, skin, and oral cavity, specifically saliva.

Regarding the human vagina, the most common microbial resident is *Lactobacilli*, as we also observed in the majority of the vaginal samples we analysed (*Chapter 2* and *Chapter 3*). Ravel *et al.* (110) analysed 396 women of reproductive-age and different ethnic groups (white, black, Hispanic, Asian). Interestingly, the authors found that different *Lactobacillus* species presented distinct prevalence (%) and dominance (%) in the study sample, as follows: *L. iners* (83.5%, 34.0%), *L. crispatus* (64.5%, 26.2%), *L. gasseri* (42.9%, 6.3%) and *L. jensenni* (48.1%, 5.3%). Moreover, 20-30% of healthy subjects were found to lack *Lactobacillus* and were dominated by other array of genera. These genera included *Corynebacterium*, *Atopobium*, *Gardnerella*, *Anaerococcus*, *Prevotella* and *Sneathia*, among others (110-112), as we also observed in a few vaginal samples in our studies (*Chapter 2* and *Chapter 3*). In a forensic context, this ample inter-individual variation underlines the advantage of targeting the entire microbial community, instead of just a few candidate bacteria. Using candidate bacteria, one runs the risk of misclassifying a stain of vaginal origin in case it lacks some or all the targeted bacteria. Moreover, females lacking *Lactobacillus* are more likely to be of African ancestry (113), which is underrepresented in the microbiome literature (see *Microbiome data* section before). This creates an imbalance in the available vaginal microbiome data, where populations are not equally represented. This data imbalance can ultimately limit the generalization of applications and even lead to confusing a healthy, low-prevalent microbial community with a disease state.

Furthermore, the skin is another high-variable body site. To study individuality contribution to skin microbial stability, Oh *et al.* (114) analysed samples over short (1-2 months) and long (1-2 years) time intervals from 12 donors at different skin sites. Samples from the same individual at both the short- and long-time intervals were more similar than those between individuals. At the species level, short-time interval samples were more similar than the long ones. Notably, some species were site- and individual-dependent in multiple individuals, such as the palm *Propionibacterium acnes*. This highlights again the advantage of relying on the entire microbial community profiles rather than a few candidate bacteria that are not prevalent in the entire population and/or might fluctuate over time for the same individual, leading to misleading conclusions.

Finally, the oral microbiome, particularly saliva, presents less variation than the previous two mentioned sites (3, 106). To characterise this, Hall *et al.* (115) analysed daily, weekly and monthly variations of the 16S rRNA gene saliva microbiome of 85 subjects. The

authors reported that the low prevalent microorganisms were highly variable over time within a subject, whereas others were more stable, like the *Fusobacterium* and *Haemophilus* genera. We made use of this and analysed these two genera in the research described in *Chapter 4*, together with *Veillonella* genus, where we sampled the study participants twice within a time interval of seven months between the two sample points. Our results confirmed that *Fusobacterium*, *Haemophilus* and *Veillonella* were stable over time, since they were successfully detected using qPCR seven months after the first sample point. However, their absolute quantification varied between the two sample points in an individual-specific manner. Our observations might also be based on the use of an absolute quantification method (qPCR) rather than a method providing relative information only (16S rRNA gene amplicon sequencing) and on the low taxonomic level we analysed (species level). Moreover, researchers should keep in mind that the saliva microbiome is constantly exposed to external environmental factors; for example, changes in the host's lifestyle habits (e.g. opposite diet, omnivorous *vs.* vegan) can lead to shifts in the microbial community (116) contributing to both inter- and intra-individual variation.

Studying the human microbiome from a forensic perspective

The need to include microbiome analysis in routine forensic investigation has already been pointed out by some authors (117). This is a result of the increasing flux of forensic microbiome studies being published, especially in the last few years, as well as newly established international consortia on the topic (e.g. Metagenomics and Metadesign of Subways and Urban Biomes (MetaSUB) International Consortium, 2015 (118)). This trend in forensic microbiome research resembles to that of the clinical microbiome research, although a few years late, and can be graphically represented in a 'hype' cycle (Figure 2). Briefly, the advances in high throughput sequencing technologies in the mid 2000's and the subsequent production of high amounts of microbiome data in the clinical field sowed interest in the interdisciplinary application of the human microbiome, such as in forensics (innovation trigger). Since then, more and more researchers from both within and outside the forensic community turned their attention to the study of the microbiome to answer forensically relevant questions (peak of inflated expectation). Right now, although of great promise, the human microbiome faces various hurdles in the forensic context that need to be considered and further investigated to ultimately being overcome (slope of enlightenment). In addition to the technical details, data considerations and microbiome features discussed above, particular factors should be considered when analysing the human in forensic applications. Apart from association and application-based studies, the forensic community needs to further carry out specific forensic validation experiments with the final aim of introducing microbiome analysis in the investigation toolkit (Figure 3).

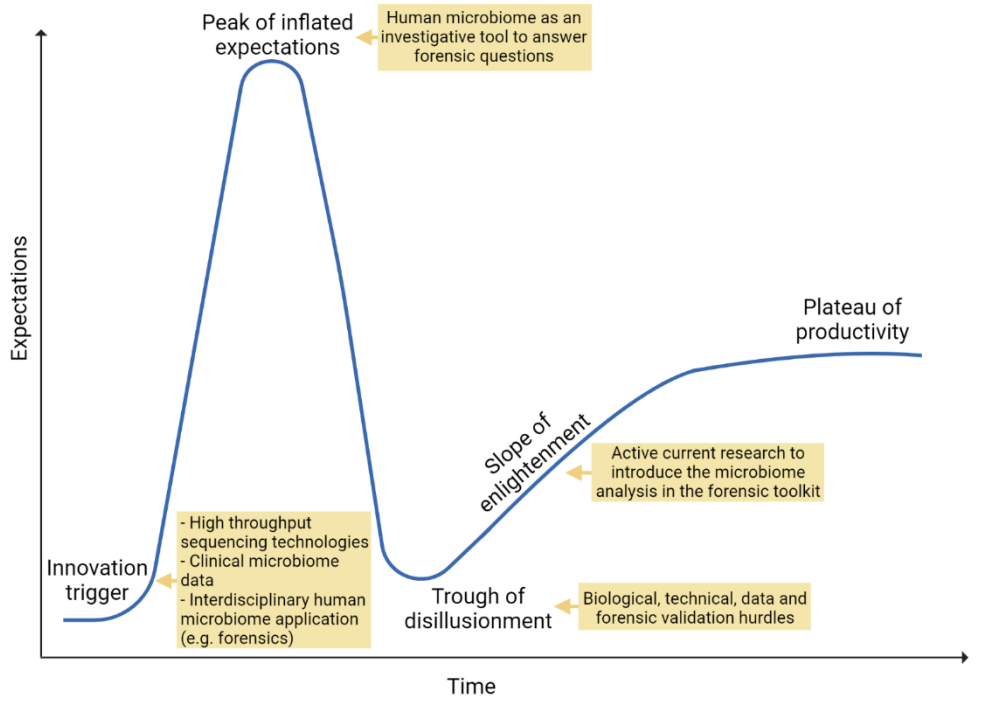


Figure 2. The 'hype' cycle describing the trends in forensic microbiome research in the last few years. (Picture created with BioRender).

Sample deposition and collection

In a real forensic scenario sample deposition cannot be controlled; nevertheless, it can be assessed in studies with controlled conditions. In our research, we observed that the type and level of host's activity before sample deposition may have an effect in the microbiome analysis results. For instance, regarding the skin specimens in *Chapter 2*, both index fingers from participants were sampled, with the difference that one hand was 'isolated' from the surrounding environment by a sterile glove worn for a specific time (1 h). This difference between hands did not affect the body site of origin classification to the skin epithelium with our approach, meaning that potential transient microorganisms picked up from the environment did not negatively contribute to the classification. However, environmental microorganisms might detrimentally affect other applications such as individual identification, where they can be confused with individualised taxa (12). Furthermore, in *Chapter 3* we discussed that the time between nose-blowing and nasal blood deposition might impact how similar a nasal blood stain microbiome is to the reference nasal mucosa or skin epithelium one. In other body sites highly exposed to external factors like the oral cavity (saliva), activities like the time since last teeth brushing might also impact the microbiome results. Moreover, the time during the day when the sample is deposited can additionally introduce variation, as showed by Takayasu *et al.* (119) who reported global circadian rhythm in the saliva

microbiome. Future forensic validation studies should further investigate the magnitude of the impact of sample deposition on the microbiome and ways to deal with it during analysis.

Regarding sample collection, sterile swabs are a good option for sampling, since they do not introduce undesirable contamination and in many cases they can also serve for the collection of both crime and reference samples (e.g. 'touched' object at the crime scene and suspect's skin reference sample). Moreover, swabbing has proven to produce reliable results in low-biomass samples, which are commonly encountered in forensic casework (35).

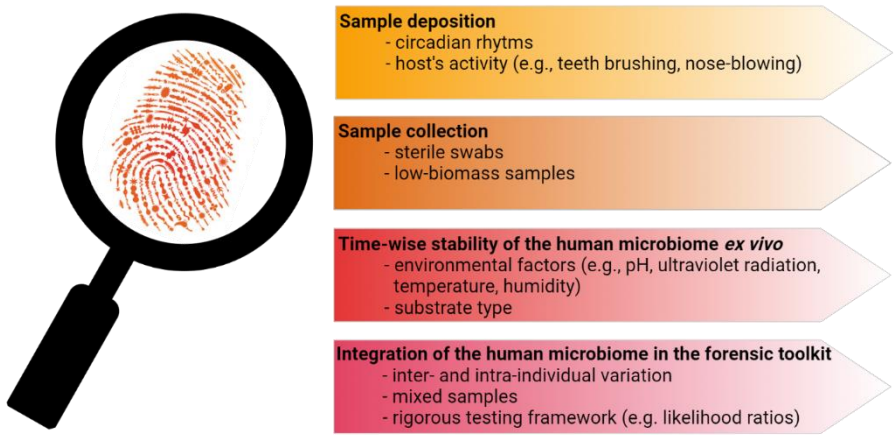


Figure 3. Special considerations in the study of the human microbiome from a forensic standpoint. (Picture created with BioRender).

Time-wise stability of the human microbiome ex vivo

In forensic investigations it is common that biological traces are exposed to the environment for variable periods of time before sample collection and processing. On the one hand, for certain applications it is of interest that the microbial communities remain stable once out of the human body, like in individual and tissue/body fluid identification applications. On the other hand, some traceable level of variation is desired in other applications, such as post-mortem interval and time since stain deposition estimations. The balance between the former and the later greatly falls on the analysed taxonomic level, by which only low-taxonomy analysis (e.g. species level) might reflect *ex vivo* variation, whereas high-taxonomy analysis (e.g. phylum level) might not. A few studies have assessed the time-wise stability of the microbiome *ex vivo* using 16S rRNA gene amplicon sequencing. Fierer *et al.* (14) reported no or little sample storage effect on microbial community composition on skin swabs exposed for two weeks at indoor conditions (temperature of approximate 20 °C and fluorescent light for about 8 h per day). Dobay *et al.* (120) analysed various sample types (saliva, skin, peripheral blood, menstrual blood, vaginal fluid and semen) and showed that samples analysed straightaway after collection and samples exposed to the lab environment for 30 days clustered together

according to the body site of origin. In a different study, Williams *et al.* (121) described random and no significant influence in the taxonomic profiles obtained from pubic hair stored at different temperatures (i.e. 20 °C, refrigerated at 4 °C and frozen at -20°C). Recently, Salzmann *et al*. (122) analysed stains from five body fluids (venous blood, menstrual blood, saliva, semen and vaginal secretions) exposed to indoor conditions (dark and dry environment, room temperature) at different time points, from fresh up to 1.5 years. The authors analysed total RNA using NGS and reported very low variation in the RNA profiles of the samples (domain and phylum level) across all the analysed time points. In *Chapter 2* and *Chapter 3*, we also reported stability, even for several years, in the overall microbial communities of a very diverse set of mock casework samples (e.g. fabrics, food utensils, hygiene products) using 16S rRNA gene amplicon sequencing. However, in *Chapter 4* where we employed qPCR for low-taxonomy level analysis, we confirmed traceable time-dependent variation of the targeted bacterial species.

Moreover, the crime stain and reference sample it is compared to, are likely to be obtained at different environmental conditions (e.g. chemical and physical agents, such as pH, ultraviolet (UV) radiation, temperature and humidity) that can lead to different microbial communities for each sample type. For instance, post-mortem interval estimation predictions are influenced by the season, temperature or location of the cadaver (17, 123, 124). These factors might not only have an influence on the microbial communities present in human cadavers, but also in human-derived traces deposited at crime scenes. This was recently reported in a study that analysed total RNA in different body fluid stains (venous blood, menstrual blood, saliva, semen and vaginal secretions) exposed to outdoor conditions (flat rooftop exposed to the sun and wind but protected from the rain) in which, after 6 months, the prokaryotic RNA fraction decreased while the eukaryotic RNA fraction increased (mainly plant RNA) (122). Nevertheless, this topic still remains to be further explored and might be solved following approaches where reference samples are exposed to specific crime scene conditions (like we did in *Chapter 4*) or investigating alternatives to converge both crime and reference sample types. Moreover, future studies should include forensic validation experiments on the influence of substrate type in the microbiome results. For example, the material a body fluid stain is deposited on might prevent or inhibit the colonization of certain bacterial groups, which are informative for a certain application or, on the contrary, might favour the transfer and growth of others (125-127).

Future integration of the human microbiome in the forensic toolkit

The human microbiome can serve as a very useful investigative tool in those cases where other human (non-microbial) molecular approaches present limitations. This does not mean that those other approaches should be replaced by microbiome analysis but, where possible, to sum forces with the aim of answering challenging forensic questions. For instance, bacterial DNA seems to be more resistant compared to human DNA to harsh environmental conditions (i.e. chemical and physical agents), based on the circular nature of the bacterial DNA molecule and its localization within a cell furnished with a wall of peptidoglycan matrix (128). In our studies on microbiome-based body site of origin classification of human traces (*Chapter 2* and *Chapter 3*), we observed that our proposed microbiome approach performed better than

human RNA-based methods in challenging samples, such as for skin traces deposited on 'touched' objects and for body fluids of a complex nature, like menstrual blood. Moreover, individualising microbiome 'fingerprints' could be useful in cases involving monozygotic twins, since these are identical in their human genome component and indistinguishable using standard short tandem repeats (STR) DNA-based profiling (129). There are even current efforts on creating publicly available forensic microbiome databases like the one by Singh *et al.* (130) for geolocation applications, that include a web interface with a total of 20,850 samples from 95 different projects, including 54 different body sites and individuals of 35 different countries.

However, before the introduction of microbiome analysis in the forensic toolkit, certain hurdles should be overcome. Further research is needed on the influence on the microbiome results of the time span between a trace is deposited at the crime scene and the collection of a reference sample from a known suspect. As discussed before (*Inter-individual and intra-individual variation of the human microbiome*), intra-individual variation in the microbial community between the two time points (e.g. based on change in individual's lifestyle habits, disease onset) could make the comparison between crime and reference samples difficult or not reliable. Additionally, the magnitude of the inter-individual variation on the applications generalizability should be further assessed and tried to be solved at the data or analysis level. Furthermore, most forensic microbiome studies are of small sample size. This can lead to big differences between the analysed sample and the general population that arise simply by chance, making it challenging to interpret the results and generalize the applications. Regarding forensic validation, the analysis and deconvolution of mixed samples, from different persons, is a pending task for the forensic microbiome community. This is already highly complex when analysing mixed human DNA profiles of two or a few individuals and might be much more tedious in microbiome profiles consisting of dozens or hundreds of microorganisms. Lastly, in routine forensic investigations there is a very rigorous testing framework in which the output is the likelihood ratio between the hypothesis being tested (e.g. a known individual is the donor of a crime stain) *vs.* the null hypothesis (e.g. the crime stain donor is any other random individual in the population). Same should be considered before microbiome analysis is introduced in the forensic toolkit, though much further research and validation into this is needed.

Conclusions and outlook

Besides the progress made with the research outcomes summarized in this thesis, and that of others in recent years, the use of the human microbiome in forensic science is still in its infancy. Nevertheless, as demonstrated here, with further research and appropriate validation it holds a great promise in the forensic field, especially where other approaches show limitations and challenges. With the studies performed as part of this thesis, we have proven the potential of the human microbiome as a tool to answer relevant forensic questions. Firstly, to answer the question 'What happened at the crime scene?', we have demonstrated the forensic suitability of the human microbiome to accurately identify the body site of origin of human biological stains by applying a novel taxonomy-independent deep learning approach

(*Chapter 2* and *Chapter 3*). Secondly, to answer the question 'When did the crime happened?', we showed for the first time the future potential of targeting human bacteria for estimating the time since deposition of saliva stains in the forensic scenario (*Chapter 4*). After further research and appropriate forensic validation, this might be expanded to other forensically relevant human stains than saliva ones. Lastly, to answer the question 'Who committed the crime?', we introduced the use of the human microbiome for inferring host's lifestyle habits (e.g. cigarette smoking) in a forensic context. Moreover, we successfully applied a novel analytical approach based on data augmentation and ML to address the class imbalance problem commonly encountered in microbiome datasets, which is not only applicable in forensics but generalizable to other microbiome-based prediction applications in other areas of research (*Chapter 5*). Importantly, in our research we took special care to keep the experimental and analytical analyses as homogenous as possible allowing for the comparison between our studies and reference ones as well as data reanalysis.

Future forensic microbiome studies should carry out longitudinal experiments evaluating the microbiome stability both *in vivo* (intra-individual variation) and *ex vivo* (variation outside the human body after being deposited in the form of stains). With regard to the former, the composition of the human microbiome is influenced by many factors, such as environment (131, 132), health status (6), lifestyle habits (133, 134) or even human relationships (135, 136). It remains to be further explored how these factors can influence the comparison between the crime and reference samples in an investigation. Not only the forensic community can benefit from this research, but it might help in some cases (e.g. presence of disease) to unlock the paradigm whether intra-individual changes are a result of the factor or the driving force of it, as well as how much time on average it takes for a complete community shift under certain circumstances. Additionally, the stability of the microbiome *ex vivo* is a main concern in forensics and may be desired depending on the application. The study of this topic could also benefit other (medical) areas of research, where sample stability (e.g. transportation, storage conditions) can introduce undesirable variation in the microbiome results (137, 138).

Moreover, the forensic microbiome community should test multi-omics approaches for answering relevant forensic questions. Studying 'what organisms are present' in a sample (DNA-based analysis) is not enough to capture the entire complexity of the human microbiome. It is known that community assembly is primarily mediated by functional niches rather than a requirement for specific organisms (139, 140). Therefore, analysing 'what the microorganisms are doing' from metatranscriptomics (141), metaproteomics (142) and metabolomics (143) can help in adding new dimensions to the current knowledge. This can be important, for example, in cases that we know that different bacteria can perform the same function (e.g. lactic acid production in vagina), which based on DNA-based analysis might look different, but not based on other data type, such as metabolomics data.

On another point, the research community should be aware of the ongoing improvements of the long-read sequencing technologies, like Pacific Biosciences Sequel (PacBio) and Oxford Nanopore MiniON platforms. These platforms are said to suffer from high error rates in the range of 5% to 15% (37). However, a study was able to reduce the observed error rate for the V1-V9 region of the 16S rRNA gene from 0.69 to 0.03% for PacBio data, comparable to those of the Illumina systems (144). The long-read third-generation sequencing platforms can offer several advantages over the short-read second-generation sequencing

platforms, as they are able to sequence the full length of the genes of interest increasing taxonomic resolution; hence, opening up new possibilities and applications options.

Finally, after considering the previous factors, the ultimate goal would be to develop and carefully validate an all-in-one microbiome forensic tool. Current nucleic acids isolation methods allow for the parallel extraction of both human and microbial nucleic acids. This is a plus when, if enough material is available, both human- and microbial-based analysis can be performed producing different pieces of information. From the microbial fraction, an all-in-one tool could provide diverse facts based on the investigation questions. For instance, once the microbiome data is produced, they could serve as the input for different prediction algorithms for tissue/body fluid identification, estimation of the time since stain deposition and prediction of the lifestyle habits (e.g. smoking) of the unknown donor of a trace, as we have investigated in this thesis.

References

1. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. 2009. Bacterial community variation in human body habitats across space and time. Science 326:1694-1697.
2. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, Knights D, Gajer P, Ravel J, Fierer N, Gordon JI, Knight R. 2011. Moving pictures of the human microbiome. Genome Biol 12:R50.
3. Zhou Y, Gao H, Mihindukulasuriya KA, La Rosa PS, Wylie KM, Vishnivetskaya T, Podar M, Warner B, Tarr PI, Nelson DE. 2013. Biogeography of the ecosystems of the healthy human body. Genome Biol 14:1-18.
4. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein PC, Knight R. 2016. Microbiome-wide association studies link dynamic microbial consortia to disease. Nature 535:94-103.
5. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG. 2015. Host genetic variation impacts microbiome composition across human body sites. Genome Biol 16:191.
6. Cho I, Blaser MJ. 2012. The human microbiome: at the interface of health and disease. Nat Rev Genet 13:260-270.
7. Sender R, Fuchs S, Milo R. 2016. Revised estimates for the number of human and bacteria cells in the body. PLoS Biol 14:e1002533.
8. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R. 2008. Evolution of mammals and their gut microbes. Science 320:1647-1651.
9. Wu J, Peters BA, Dominianni C, Zhang YL, Pei ZH, Yang LY, Ma YF, Purdue MP, Jacobs EJ, Gapstur SM, Li HL, Alekseyenko AV, Hayes RB, Ahn JY. 2016. Cigarette smoking and the oral microbiome in a large study of American adults. ISME J 10:2435-2446.
10. Meadow JF, Bateman AC, Herkert KM, O'Connor TK, Green JL. 2013. Significant changes in the skin microbiome mediated by the sport of roller derby. PeerJ 1:e53.
11. Qian J, Hospodsky D, Yamamoto N, Nazaroff WW, Peccia J. 2012. Size-resolved emission rates of airborne bacteria and fungi in an occupied classroom. Indoor Air 22:339-51.
12. Wilkins D, Leung MHY, Lee PKH. 2017. Microbiota fingerprints lose individually identifying features over time. Microbiome 5:1.
13. Lax S, Hampton-Marcell JT, Gibbons SM, Colares GB, Smith D, Eisen JA, Gilbert JA. 2015. Forensic analysis of the microbiome of phones and shoes. Microbiome 3:21.
14. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477-6481.
15. Johnson HR, Trinidad DD, Guzman S, Khan Z, Parziale JV, DeBruyn JM, Lents NH. 2016. A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. PLoS One 11:e0167370.
16. Metcalf JL, Xu ZZ, Weiss S, Lax S, Van Treuren W, Hyde ER, Song SJ, Amir A, Larsen P, Sangwan N. 2016. Microbial community assembly and metabolic function during mammalian corpse decomposition. Science 351:158-162.
17. Hyde ER, Haarmann DP, Petrosino JF, Lynne AM, Bucheli SR. 2015. Initial insights into bacterial succession during human decomposition. Int J Legal Med 129:661-671.
18. Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, Siegel J, Caporaso JG. 2016. Geography and location are the primary drivers of office microbiome composition. mSystems 1:e00022-16.
19. Nagasawa S, Motani-Saitoh H, Inoue H, Iwase H. 2013. Geographic diversity of Helicobacter pylori in cadavers: Forensic estimation of geographical origin. Forensic Sci Int Genet 229:7-12.
20. Hewitt KM, Gerba CP, Maxwell SL, Kelley ST. 2012. Office space bacterial abundance and diversity in three metropolitan areas. PLoS One 7:e37849.
21. Fleming RI, Harbison S. 2010. The use of bacteria for the identification of vaginal secretions. Forensic Sci Int Genet 4:311-315.
22. Nakanishi H, Kido A, Ohmori T, Takada A, Hara M, Adachi N, Saito K. 2009. A novel method for the identification of saliva by detecting oral streptococci using PCR. Forensic Sci Int Genet 183:20-23.
23. Akutsu T, Motani H, Watanabe K, Iwase H, Sakurada K. 2012. Detection of bacterial 16S ribosomal RNA genes for forensic identification of vaginal fluid. Leg Med 14:160-162.
24. Lucci A, Campobasso CP, Cirnelli A, Lorenzini G. 2008. A promising microbiological test for the diagnosis of drowning. Forensic Sci Int Genet 182:20-26.
25. Suto M, Kato N, Abe S, Nakamura M, Tsuchiya R, Hiraiwa K. 2009. PCR detection of bacterial genes provides evidence of death by drowning. Leg Med (Tokyo) 11 Suppl 1:S354-6.
26. Black CM, Driebe EM, Howard LA, Fajman NN, Sawyer MK, Girardet RG, Sautter RL, Greenwald E, Beck-Sague CM, Unger ER, Igietseme JU, Hammerschlag MR. 2009. Multicenter study of nucleic acid

amplification tests for detection of Chlamydia trachomatis and Neisseria gonorrhoeae in children being evaluated for sexual abuse. Pediatr Infect Dis J 28:608-13.
27. Martin IMC, Foreman E, Hall V, Nesbitt A, Forster G, Ison CA. 2007. Non-cultural detection and molecular genotyping of Neisseria gonorrhoeae from a piece of clothing. J Med Microbiol 56:487-490.
28. Wang Y, LêCao K-A. 2019. Managing batch effects in microbiome data. Brief Bioinform 21:1954-1970.
29. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. 2015. The microbiome quality control project: baseline study design and future directions. Genome Biol 16:276.
30. Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, Abnet CC, Knight R, White O, Huttenhower C, The Microbiome Quality Control Project C. 2017. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. Nat Biotechnol 35:1077-1086.
31. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung F-E. 2017. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 35:1069-1076.
32. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA, Vaginal Microbiome C. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol 15:66.
33. Teng F, Darveekaran Nair SS, Zhu P, Li S, Huang S, Li X, Xu J, Yang F. 2018. Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. Sci Rep 8:16321.
34. Abusleme L, Hong B-Y, Dupuy AK, Strausbaugh LD, Diaz PI. 2014. Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing. J Oral Microbiol 6:23990.
35. Bjerre RD, Hugerth LW, Boulund F, Seifert M, Johansen JD, Engstrand L. 2019. Effects of sampling strategy and DNA extraction on human skin microbiome investigations. Sci Rep 9:17287.
36. Videnska P, Smerkova K, Zwinsova B, Popovici V, Micenkova L, Sedlar K, Budinska E. 2019. Stool sampling and DNA isolation kits affect DNA quality and bacterial composition following 16S rRNA gene sequencing using MiSeq Illumina platform. Sci Rep 9:13837.
37. Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The madness of microbiome: attempting to find consensus “best practice” for 16S microbiome studies. Appl Environ Microbiol 84:e02627-17.
38. Guo F, Zhang T. 2013. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. Appl Microbiol Biotechnol 97:4607-16.
39. Balato A, Cacciapuoti S, Di Caprio R, Marasca C, Masarà A, Raimondo A, Fabbrocini G. 2019. Human microbiome: composition and role in inflammatory skin diseases. Arch Immunol Ther Exp (Warsz) 67:1-18.
40. Segata N, Haake SK, Mannon P, Lemon KP, Waldron L, Gevers D, Huttenhower C, Izard J. 2012. Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13:R42.
41. Gerasimidis K, Bertz M, Quince C, Brunner K, Bruce A, Combet E, Calus S, Loman N, Ijaz UZ. 2016. The effect of DNA extraction methodology on gut microbiota research applications. BMC Res Notes 9:1-10.
42. Schrader C, Schielke A, Ellerbroek L, Johne R. 2012. PCR inhibitors–occurrence, properties and removal. J Appl Microbiol 113:1014-1026.
43. Rosenbaum J, Usyk M, Chen Z, Zolnik CP, Jones HE, Waldron L, Dowd JB, Thorpe LE, Burk RD. 2019. Evaluation of oral cavity DNA extraction methods on bacterial and fungal microbiota. Sci Rep 9:1531.
44. Lim Y, Totsika M, Morrison M, Punyadeera C. 2017. The saliva microbiome profiles are minimally affected by collection method or DNA extraction protocols. Sci Rep 7:8523.
45. Karstens L, Siddiqui NY, Zaza T, Barstad A, Amundsen CL, Sysoeva TA. 2021. Benchmarking DNA isolation kits used in analyses of the urinary microbiome. Sci Rep 11:6186.
46. Maukonen J, Simões C, Saarela M. 2012. The currently used commercial DNA-extraction methods give different results of clostridial and actinobacterial populations derived from human fecal samples. FEMS Microbiol Ecol 79:697-708.
47. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, Pamp SJ. 2016. Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. mSystems 1:e00095-16.
48. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87.
49. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. 2016. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog 8:1-12.

50. Lazarevic V, Gaïa N, Girard M, Schrenzel J. 2016. Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. BMC Microbiol 16:1-8.
51. Yang B, Wang Y, Qian P-Y. 2016. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. BMC Bioinformatics 17:135.
52. Kommedal O, Simmon K, Karaca D, Langeland N, Wiker HG. 2012. Dual priming oligonucleotides for broad-range amplification of the bacterial 16S rRNA gene directly from human clinical specimens. J Clin Microbiol 50:1289-1294.
53. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6:226.
54. Walker SP, Barrett M, Hogan G, Flores Bueso Y, Claesson MJ, Tangney M. 2020. Non-specific amplification of human DNA is a major challenge for 16S rRNA gene sequence analysis. Sci Rep 10:16356.
55. Sune D, Rydberg H, Augustinsson ÅN, Serrander L, Jungeström MB. 2020. Optimization of 16S rRNA gene analysis for use in the diagnostic clinical microbiology service. J Microbiol Methods 170:105854.
56. Pereira-Marques J, Hout A, Ferreira RM, Weber M, Pinto-Ribeiro I, van Doorn L-J, Knetsch CW, Figueiredo C. 2019. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. Front Microbiol 10:1277.
57. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. 2010. 16S 454 Sequencing Protocol HMP Consortium.
58. Ferretti P, Farina S, Cristofolini M, Girolomoni G, Tett A, Segata N. 2017. Experimental metagenomics and ribosomal profiling of the human skin microbiome. Exp Dermatol 26:211-219.
59. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621-1624.
60. Tremblay J, Singh K, Fern A, Kirton E, He S, Woyke T, Lee J, Chen F, Dangl J, Tringe S. 2015. Primer and platform effects on 16S rRNA tag sequencing. Front Microbiol 6:771.
61. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17:55.
62. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112-5120.
63. Schloss PD. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput Biol 6:e1000844.
64. Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl Environ Microbiol 71:7724-7736.
65. Kozlov AM, Zhang J, Yilmaz P, Glöckner FO, Stamatakis A. 2016. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Res 44:5022-5033.
66. Werner JJ, Koren O, Hugenholtz P, DeSantis TZ, Walters WA, Caporaso JG, Angenent LT, Knight R, Ley RE. 2012. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. ISME J 6:94-103.
67. Ritari J, Salojärvi J, Lahti L, de Vos WM. 2015. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. BMC Genomics 16:1-10.
68. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. 2018. New insights into human nostril microbiome from the expanded Human Oral Microbiome Database (eHOMD): a resource for the microbiome of the human aerodigestive tract. mSystems 3:e00187-18.
69. Cammarota G, Ianiro G, Ahern A, Carbone C, Temko A, Claesson MJ, Gasbarrini A, Tortora G. 2020. Gut microbiome, big data and machine learning to promote precision medicine for cancer. Nat Rev Gastroenterol Hepatol 17:635-648.
70. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, Chen M-X, Chen Z-H, Ji G-Y, Mujagond P. 2018. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 24:1532-1535.
71. Reitmeier S, Kiessling S, Clavel T, List M, Almeida EL, Ghosh TS, Neuhaus K, Grallert H, Linseisen J, Skurk T, Brandl B, Breuninger TA, Troll M, Rathmann W, Linkohr B, Hauner H, Laudes M, Franke A, Le Roy CI, Bell JT, Spector T, Baumbach J, O'Toole PW, Peters A, Haller D. 2020. Arrhythmic gut microbiome signatures predict risk of type 2 diabetes. Cell Host Microbe 28:258-272.e6.
72. Su X, Jing G, Sun Z, Liu L, Xu Z, McDonald D, Wang Z, Wang H, Gonzalez A, Zhang Y, Huang S, Huttley G, Knight R, Xu J. 2020. Multiple-disease detection and classification across cohorts via microbiome search. mSystems 5:e00150-20.

73. Zheng Y, Fang Z, Xue Y, Zhang J, Zhu J, Gao R, Yao S, Ye Y, Wang S, Lin C, Chen S, Huang H, Hu L, Jiang GN, Qin H, Zhang P, Chen J, Ji H. 2020. Specific gut microbiome signature predicts the early-stage lung cancer. Gut Microbes 11:1030-1042.
74. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, McKay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579:567-574.
75. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun 8:1784.
76. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. 2016. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol 12:e1004977.
77. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, Le Roy CI, Raygoza Garay JA, Finnicum CT, Liu X, Zhernakova DV, Bonder MJ, Hansen TH, Frost F, Rühlemann MC, Turpin W, Moon J-Y, Kim H-N, Lüll K, Barkan E, Shah SA, Fornage M, Szopinska-Tokov J, Wallen ZD, Borisevich D, Agreus L, Andreasson A, Bang C, Bedrani L, Bell JT, Bisgaard H, Boehnke M, Boomsma DI, Burk RD, Claringbould A, Croitoru K, Davies GE, van Duijn CM, Duijts L, Falony G, Fu J, van der Graaf A, Hansen T, Homuth G, Hughes DA, Ijzerman RG, Jackson MA, Jaddoe VWV, Joossens M, Jørgensen T, et al. 2021. Large-scale association analyses identify host factors influencing human gut microbiome composition. Nature Genet 53:156-165.
78. Blaser MJ. 2010. Harnessing the power of the human microbiome. Proc Natl Acad Sci U S A 107:6125-6126.
79. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. 2020. A framework for effective application of machine learning to microbiome-based classification problems. mBio 11:e00434-20.
80. Human Microbiome Project C. 2012. A framework for human microbiome research. Nature 486:215-21.
81. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, DeRight Goldasich L, Dorrestein PC, Dunn RR, Fahimipour AK, Gaffney J, Gilbert JA, Gogul G, Green JL, Hugenholtz P, Humphrey G, Huttenhower C, Jackson MA, Janssen S, Jeste DV, Jiang L, Kelley ST, Knights D, Kosciolek T, Ladau J, Leach J, Marotz C, Meleshko D, Melnik AV, Metcalf JL, Mohimani H, Montassier E, Navas-Molina J, Nguyen TT, Peddada S, Pevzner P, Pollard KS, Rahnavard G, Robbins-Pianka A, Sangwan N, Shorenstein J, Smarr L, Song SJ, Spector T, Swafford AD, Thackray VG, et al. 2018. American Gut: an open platform for citizen science microbiome research. mSystems 3:e00031-18.
82. Dave M, Higgins PD, Middha S, Rioux KP. 2012. The human gut microbiome: current knowledge, challenges, and future directions. Transl Res 160:246-257.
83. Jumpertz R, Le DS, Turnbaugh PJ, Trinidad C, Bogardus C, Gordon JI, Krakoff J. 2011. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. Am J Clin Nutr 94:58-65.
84. Gaboriau-Routhiau V, Rakotobe S, Lécuyer E, Mulder I, Lan A, Bridonneau C, Rochet V, Pisi A, De Paepe M, Brandi G. 2009. The key role of segmented filamentous bacteria in the coordinated maturation of gut helper T cell responses. Immunity 31:677-689.
85. Cryan JF, O'Riordan KJ, Cowan CSM, Sandhu KV, Bastiaanssen TFS, Boehme M, Codagnone MG, Cussotto S, Fulling C, Golubeva AV. 2019. The microbiota-gut-brain axis. Physiol Rev 99:1877-2013.
86. Proctor L, LoTempio J, Marquitz A, Daschner P, Xi D, Flores R, Brown L, Ranallo R, Maruvada P, Regan K, Dwayne Lunsford R, Reddy M, Caler L, Team NIHHMPA. 2019. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome 7:31.
87. Brewster R, Tamburini FB, Asiimwe E, Oduaran O, Hazelhurst S, Bhatt AS. 2019. Surveying gut microbiome research in Africans: Toward improved diversity and representation. Trends Microbiol 27:824-835.
88. Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. 2015. Genome-wide association studies of the human gut microbiota. PLoS One 10:e0140301.
89. Si J, Lee S, Park JM, Sung J, Ko G. 2015. Genetic associations and shared environmental effects on the skin microbiome of Korean twins. BMC Genomics 16:992.
90. Brooks AW, Priya S, Blekhman R, Bordenstein SR. 2018. Gut microbiota diversity across ethnicities in the United States. PLoS Biol 16:e2006842.
91. Gupta VK, Paul S, Dutta C. 2017. Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. Front Microbiol 8:1162-1162.
92. Levin AM, Sitarik AR, Havstad SL, Fujimura KE, Wegienka G, Cassidy-Bushrow AE, Kim H, Zoratti EM, Lukacs NW, Boushey HA. 2016. Joint effects of pregnancy, sociocultural, and environmental factors on early life gut microbiome structure and diversity. Sci Rep 6:1-16.
93. Lokmer A, Aflalo S, Amougou N, Lafosse S, Froment A, Tabe FE, Poyet M, Groussin M, Said-Mohamed R, Ségurel L. 2020. Response of the human gut and saliva microbiome to urbanization in Cameroon. Sci Rep 10:1-15.

94. Bai J, Hu Y, Bruner DW. 2019. Composition of gut microbiota and its association with body mass index and lifestyle factors in a cohort of 7–18 years old children from the American Gut Project. Pediatr Obes 14:e12480.
95. Dimitriu PA, Iker B, Malik K, Leung H, Mohn WW, Hillebrand GG. 2019. New Insights into the Intrinsic and Extrinsic Factors That Shape the Human Skin Microbiome. mBio 10:e00839-19.
96. Scepanovic P, Hodel F, Mondot S, Partula V, Byrd A, Hammer C, Alanio C, Bergstedt J, Patin E, Touvier M, Lantz O, Albert ML, Duffy D, Quintana-Murci L, Fellay J, Abel L, Alcover A, Aschard H, Astrom K, Bousso P, Bruhns P, Cumano A, Demangel C, Deriano L, Di Santo J, Dromer F, Duffy D, Eberl G, Enninga J, Fellay J, Gelpi O, Gomperts-Boneca I, Hasan M, Hercberg S, Lantz O, Leclerc C, Mouquet H, Pellegrini S, Pol S, Rausell A, Rogge L, Sakuntabhai A, Schwartz O, Schwikowski B, Shorte S, Soumelis V, Tangy F, Tartour E, Toubert A, Touvier M, et al. 2019. A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. Microbiome 7:130.
97. Radjabzadeh D, Boer CG, Beth SA, van der Wal P, Kiefte-De Jong JC, Jansen MAE, Konstantinov SR, Peppelenbosch MP, Hays JP, Jaddoe VWV, Ikram MA, Rivadeneira F, van Meurs JBJ, Uitterlinden AG, Medina-Gomez C, Moll HA, Kraaij R. 2020. Diversity, compositional and functional differences between gut microbiota of children and adults. Sci Rep 10:1040.
98. Santoro A, Pini E, Scurti M, Palmas G, Berendsen A, Brzozowska A, Pietruszka B, Szczecinska A, Cano N, Meunier N. 2014. Combating inflammaging through a Mediterranean whole diet approach: the NU-AGE project's conceptual framework and design. Mech Ageing Dev 136:3-13.
99. Kasmanas JC, Bartholomäus A, Corrêa FB, Tal T, Jehmlich N, Herberth G, von Bergen M, Stadler PF, Carvalho André Carlos Ponce de Leon Ferreira d, Nunes da Rocha U. 2020. HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Res 49:D743-D750.
100. Vangay P, Burgin J, Johnston A, Beck KL, Berrios DC, Blumberg K, Canon S, Chain P, Chandonia J-M, Christianson D, Costes SV, Damerow J, Duncan WD, Dundore-Arias JP, Fagnan K, Galazka JM, Gibbons SM, Hays D, Hervey J, Hu B, Hurwitz BL, Jaiswal P, Joachimiak MP, Kinkel L, Ladau J, Martin SL, McCue LA, Miller K, Mouncey N, Mungall C, Pafilis E, Reddy TBK, Richardson L, Roux S, Schriml LM, Shaffer JP, Sundaramurthi JC, Thompson LR, Timme RE, Zheng J, Wood-Charlson EM, Eloe-Fadrosh EA. 2021. Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities. mSystems 6:e01194-20.
101. Zaura E, Pappalardo VY, Buijs MJ, Volgenant CMC, Brandt BW. 2021. Optimizing the quality of clinical studies on oral microbiome: A practical guide for planning, performing, and reporting. Periodontol 2000 85:210-236.
102. Kachroo N, Lange D, Penniston KL, Stern J, Tasian G, Bajic P, Wolfe AJ, Suryavanshi M, Ticinesi A, Meschi T, Monga M, Miller AW. 2021. Standardization of microbiome studies for urolithiasis: an international consensus agreement. Nat Rev Urol 18:303-311.
103. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, Christian L, Grant B, Olendorf R, Sandusky RJ. 2020. Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLoS One 15:e0229003.
104. Kessler R. 2018. Whitepaper: Practical challenges for researchers in data sharing. Learn Publ 31.
105. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, Leff JW, Vázquez-Baeza Y, Gonzalez A, Knight R. 2014. Temporal variability is a personalized feature of the human microbiome. Genome Biol 15:1-13.
106. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci U S A 112:E2930-E2938.
107. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH, McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550:61-66.
108. Ding T, Schloss PD. 2014. Dynamics and associations of microbial community types across the human body. Nature 509:357-360.
109. Kong HH, Segre JA. 2012. Skin microbiome: Looking back to move forward. J Invest Dermatol 132:933-939.
110. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, Karlebach S, Gorle R, Russell J, Tacket CO, Brotman RM, Davis CC, Ault K, Peralta L, Forney LJ. 2011. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci U S A 108 Suppl 1:4680-4687.
111. Hyman RW, Fukushima M, Diamond L, Kumm J, Giudice LC, Davis RW. 2005. Microbes on the human vaginal epithelium. Proc Natl Acad Sci U S A 102:7952-7.

112. Zhou X, Bent SJ, Schneider MG, Davis CC, Islam MR, Forney LJ. 2004. Characterization of vaginal microbial communities in adult healthy women using cultivation-independent methods. Microbiology (Reading) 150:2565-2573.
113. Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, Strauss JF, Jefferson KK, Buck GA, Consortium VM. 2014. Differences in vaginal microbiome in African American women versus women of European ancestry. Microbiology (Reading) 160:2272-2282.
114. Oh J, Byrd AL, Park M, Kong HH, Segre JA, Sequencing NC. 2016. Temporal stability of the human skin microbiome. Cell 165:854-866.
115. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, Neufeld JD, Beiko RG, Senadheera DB. 2017. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. NPJ Biofilms Microbiomes 3:1-7.
116. Hansen TH, Kern T, Bak EG, Kashani A, Allin KH, Nielsen T, Hansen T, Pedersen O. 2018. Impact of a vegan diet on the human salivary microbiota. Sci Rep 8:5847.
117. Clarke TH, Gomez A, Singh H, Nelson KE, Brinkac LM. 2017. Integrating the microbiome as a resource in the forensics toolkit. Forensic Sci Int Genet 30:141-147.
118. Mason C, Afshinnekoo E, Ahsannudin S, Ghedin E, Read T, Fraser C, Dudley J, Hernandez M, Bowler C, Stolovitzky G, Chernonetz A, Gray A, Darling A, Burke C, Łabaj PP, Graf A, Noushmehr H, Moraes s, Dias-Neto E, Ugalde J, Guo Y, Zhou Y, Xie Z, Zheng D, Zhou H, Shi L, Zhu S, Tang A, Ivanković T, Siam R, Rascovan N, Richard H, Lafontaine I, Baron C, Nedunuri N, Prithiviraj B, Hyat S, Mehr S, Banihashemi K, Segata N, Suzuki H, Alpuche Aranda CM, Martinez J, Christopher Dada A, Osuolale O, Oguntoyinbo F, Dybwad M, Oliveira M, Fernandes A, Oliveira M, et al. 2016. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome 4:24.
119. Takayasu L, Suda W, Takanashi K, Iioka E, Kurokawa R, Shindo C, Hattori Y, Yamashita N, Nishijima S, Oshima K, Hattori M. 2017. Circadian oscillations of microbial and functional composition in the human salivary microbiome. DNA Res 24:261-270.
120. Dobay A, Haas C, Fucile G, Downey N, Morrison HG, Kratzer A, Arora N. 2019. Microbiome-based body fluid identification of samples exposed to indoor conditions. Forensic Sci Int Genet 40:105-113.
121. Diana W. Williams GG. 2017. Individualization of pubic hair bacterial communities and the effects of storage time and temperature. Forensic Sci Int Genet 26:9.
122. Salzmann AP, Arora N, Russo G, Kreutzer S, Snipen L, Haas C. 2021. Assessing time dependent changes in microbial composition of biological crime scene traces using microbial RNA markers. Forensic Sci Int Genet: 102537.
123. Javan GT, Finley SJ, Can I, Wilkinson JE, Hanson JD, Tarone AM. 2016. Human thanatomicrobiome succession and time since death. Sci Rep 6:29598.
124. Can I, Javan GT, Pozhitkov AE, Noble PA. 2014. Distinctive thanatomicrobiome signatures found in the blood and internal organs of humans. J Microbiol Methods 106:1-7.
125. Kodama WA, Xu Z, Metcalf JL, Song SJ, Harrison N, Knight R, Carter DO, Happy CB. 2019. Trace evidence potential in postmortem skin microbiomes: from death scene to morgue. J Forensic Sci 64:791-798.
126. Neckovic A, van Oorschot RAH, Szkuta B, Durdle A. 2019. Investigation of direct and indirect transfer of microbiomes between individuals. Forensic Sci Int Genet 45:102212.
127. Lee S-Y, Woo S-K, Lee S-M, Eom Y-B. 2016. Forensic analysis using microbial community between skin bacteria and fabrics. Toxicol Environ Health 8:263-270.
128. Pamela T, D'Angiolella G, Brun P, Castagliuolo I, Gino S, Caenazzo L. 2020. Skin microbiome analysis for forensic human identification: What do we know so far? Microorganisms 8:873.
129. Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J, Rolf B. 2014. Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic Sci Int Genet 9:42-46.
130. Singh H, Clarke T, Brinkac L, Greco C, Nelson KE. 2021. Forensic microbiome database: A tool for gorensic geolocation meta-analysis using publicly available 16S rRNA microbiome sequencing. Front Microbiol 12:644861.
131. Hoisington AJ, Brenner LA, Kinney KA, Postolache TT, Lowry CA. 2015. The microbiome of the built environment and mental health. Microbiome 3:60.
132. Tu P, Chi L, Bodnar W, Zhang Z, Gao B, Bian X, Stewart J, Fry R, Lu K. 2020. Gut microbiome toxicity: connecting the environment and gut microbiome-associated diseases. Toxics 8:19.
133. Beghini F, Renson A, Zolnik C, Geistlinger L, Usyk M, Moody TU, Thorpe L, Dowd JB, Burk R, Segata N, Jones HE, Waldron LD. 2019. Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. Ann Epidemiol 34:18-25.

134. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559-563.
135. Williams DW, Gibson G. Classification of individuals and the potential to detect sexual contact using the microbiome of the pubic region. Forensic Sci Int Genet 41:177-187.
136. Song SJ, Lauber C, Costello EK, Lozupone CA, Humphrey G, Berg-Lyons D, Caporaso JG, Knights D, Clemente JC, Nakielny S. 2013. Cohabiting family members share microbiota with one another and with their dogs. Elife 2:e00458.
137. Choo JM, Leong LEX, Rogers GB. 2015. Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5:1-10.
138. Roesch LFW, Casella G, Simell O, Krischer J, Wasserfall CH, Schatz D, Atkinson MA, Neu J, Triplett EW. 2009. Influence of fecal sample storage on bacterial community diversity. Open Microbiol J 3:40-46.
139. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. 2007. The Human Microbiome Project. Nature 449:804-810.
140. Shafquat A, Joice R, Simmons SL, Huttenhower C. 2014. Functional and phylogenetic assembly of microbial communities in the human microbiome. Trends Microbiol 22:261-266.
141. Bashiardes S, Zilberman-Schapira G, Elinav E. 2016. Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 10:19-25.
142. Kleiner M. 2019. Metaproteomics: much more than measuring gene expression in microbial communities. mSystems 4:e00115-19.
143. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, Rodriguez-Mueller B, Zucker J, Thiagarajan M, Henrissat B. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol 8:e1002358.
144. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016. Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ 4:e1869.

Addendum

SUMMARY

In the last few years, the human microbiome has attracted the attention of the forensic research community as a promising tool to help answering questions in crime scene investigation. Prerequisites were the previous developments and improvements of the experimental and analytical tools for human microbiome analysis, the human microbiome findings in the clinical field, and the increasing availability of microbiome data from different locations of the human body in public repositories. The studies presented in this thesis were some of the first to address forensic questions with the help of human microbiome analysis. The questions addressed with human microbiome work in this thesis include 'What happened at the crime scene?' (*Part I*), 'When did the crime happened?' (*Part II*) and 'Who is the perpetrator of the crime?' (*Part III*).

In ***Chapter 1*** the overall topic and scope of this thesis is introduced. The chapter starts with describing the concept of the human microbiome and its features relevant for the work presented in this thesis. The chapter continues with the description, advantages and limitations of the two most-widely used approaches to generate human microbiome data, namely targeted amplicon sequencing and whole genome shotgun sequencing. Subsequently, particular characteristics of human microbiome data are described, as well as ways for analysing and interpreting such data. Lastly, I outline the applications of the human microbiome in forensic science proposed prior to starting this thesis work and discuss some caveats with previous studies that I aim to improve with my thesis work.

Part I introduces microbiome variation *in vivo* across human body sites for inferring the body site of origin of forensically relevant, single-source human biological samples. When applied to forensic practice in the future, this can help to evaluate the crime relevance of a biological trace found at the crime scene as well as to reconstruct the events taken place at the crime scene. ***Chapter 2*** describes the body site of origin classification of epithelial materials, namely saliva, skin and vaginal secretions. Our proposed approach is based on the use of reference 16S rRNA gene sequencing data from the Human Microbiome Project (HMP) of oral, skin and vaginal origin (N=1,636). These data covered the V1-V5 region of the 16S rRNA gene from which we identified 240 nucleotide positions presenting high inertia among the three reference body sites. The genomic information at those positions served to train 50 independent deep learning (DL) artificial neural networks (ANN) following a taxonomy-independent approach. The output of the DL consisted of assigned classification probabilities to each body site category (i.e. oral, skin, vagina) by means of averaging the output from the 50 DL networks. We validated our approach using 110 test samples from which we generated *de novo* 16S rRNA gene sequencing data using the Ion Torrent™ Personal Genome Machine (PGM). Body site classification accuracy was very high for these test samples with AUC≥0.99. Additionally, we performed forensic validation in 41 mock casework samples mimicking realistic and often encountered forensic scenarios (maximum sample age of 7.6 years). Our taxonomy-independent DL approach also performed well in classifying these mock samples to their corresponding body site of origin with AUC≥0.99.

Chapter 3 describes the body site of origin classification of blood materials, namely venous blood, menstrual blood, nasal blood and fingerprick blood. In this case, the HMP data repository lacked reference data for these fluids. Consequently, we used as reference HMP microbiome data from the body sites from which forensic blood stains could originate i.e.

vaginal epithelium (for menstrual blood), nasal mucosa (for nasal blood) and skin epithelium (for fingerprick blood) (N=623). Moreover, we generated *de novo* data for venous blood (N=150) due to the absence of publicly available data suitable to be used as reference. The Ion S5™ sequencing platform was used for the *de novo* sequencing data generation. We newly selected 920 nucleotide positions in the V1-V5 region of the 16S rRNA gene for DL ANN training. We further validated our approach using 94 test samples from which high classification accuracy was obtained for all the body site categories with AUC≥0.98. We also proved that the day of menses had no effect on the correct classification of menstrual blood samples to the vaginal category. Last, we performed forensic validation in 45 mock casework samples (maximum sample age of 21 years), from which correct body site classification was also revealed for many of the samples with AUC=0.79-0.93.

Part II introduces human microbiome variation *ex vivo* (i.e. once it is out of the human body in the form of biological stains) for estimating the time since deposition of human biological samples. When applied to forensic practice in the future, this can be valuable for providing with clues on when a DNA-identified stain donor left the biological trace behind at the crime scene. Such information can also be useful to find unknown perpetrators and can help the police to assess alibis given by known suspects. Moreover, when multiple stains are found at the scene belonging to different donors, information on their time since deposition may aid investigators select the ones with the highest investigative value for further analysis, if the time of the crime is known. In ***Chapter 4*** we chose saliva as the human biological trace for estimating the time since deposition based on its forensic relevance, its high loads of bacteria and because it harbours one of the most stable microbiome communities in the human body. Firstly, we identified the most abundant and prevalent bacterial species in human saliva using publicly available 16S rRNA gene sequencing data from 1,848 samples. Next, we assessed time-dependent changes in 15 identified species using *de novo* generated 16S rRNA gene sequencing data (Illumina Miseq) in saliva stains of two individuals exposed to indoor conditions for different time intervals up to 1 year. From this data, we selected four bacterial species i.e. *Fusobacterium periodonticum, Haemophilus parainfluenzae, Veillonella dispar,* and *Veillonella parvula* showing statistically significant time-dependent changes and developed a 4-plex qPCR assay for their targeted analysis. Then, we applied the assay to saliva stains of 15 individuals exposed to indoor conditions for different time intervals up to 1 month. From this independent and enlarged dataset, we confirmed that the four targeted bacteria presented significant time-dependent changes. Thereafter, based on the high inter-individual variation observed, we built individual-specific linear regression models for predicting the time since stain deposition, resulting in an average mean absolute error (MAE) of 5 days. Overall, the deposition time of 181 (82%) stains was correctly predicted within one week. Prediction models were also assessed in stains collected from the same individuals seven months later and exposed to similar conditions, resulting in an average MAE of 9 days. With this study, we demonstrated the potential of targeting human bacteria for estimating the time since deposition of saliva stains in the forensic scenario, which in the future might be expanded to human stains originating from other forensically relevant body sites.

Part III introduces the use of the human microbiome for inferring host's lifestyle habits from human biological traces, with the example of cigarette smoking. Typically, smoking data are collected for medical/research purposes via self-reported questionnaires that might be unreliable, which could be overcome by microbiome-based prediction of smoking habits. In

addition, such knowledge can add valuable information on the long-term lifestyle habits of a patient in a diagnostic setting. In a forensic setup, it could help finding unknown perpetrators of a crime by providing an investigative lead; hence, reducing the suspect pool to those that habitually smoke. In ***Chapter 5*** we first introduced a novel analytical approach based on data augmentation and machine learning (ML) to address the class imbalance problem in microbiome-based classification and applied it to the question of smoking habits prediction as example. For that, we collected publicly available 16S rRNA gene sequencing data and associated smoking habits metadata from two studies where class imbalance was an unavoidable issue i.e. 175 smokers *vs.* 1,070 non-smokers. Three different data augmentation techniques: synthetic minority oversampling technique (SMOTE), adaptive synthetic (ADASYN) and tree-based associative data augmentation (TADA) were applied together with seven ML methods: logistic regression, k nearest neighbors, support vector machine with linear and radial kernels, decision trees, random forest, extreme gradient boosting. K-fold nested cross-validation was used with the different augmented data types and baseline non-augmented data to validate the prediction outcome. We found that combining data augmentation with ML generally outperformed baseline methods and that combining TADA and support vector machine with linear kernel achieved the best prediction outcome. In this study, we successfully addressed the issue of class imbalance in microbiome data for reliable classification and proposed a generalizable approach to other data and classification problems. Moreover, we presented the first-of-its-kind solution for microbiome-based prediction of smoking habits from saliva 16S rRNA gene sequencing data despite of the issue of class-imbalanced data.

Chapter 6 provides a general discussion of the results presented in *Chapters 2-5* of this thesis. I discuss the technical side of the microbiome research, including the lack of consensus in the field, potential biases introduced throughout different steps of the microbiome pipeline and how to avoid or reduce them during microbiome analysis. I also consider current hurdles in the microbiome research that stem from the underrepresentation of certain microbiomes in the literature and the absence of metadata standardization. This chapter also focuses on the effect of inter- and intra-individual variation in the microbiome results. Finally, I discuss the specific forensic considerations that need to be further investigated before the microbiome work performed as part of this thesis can be included in the forensic toolkit for practical casework analysis, like sample collection and deposition and stability of the microbiome in forensic traces.

With the research described in this thesis, I, together with my colleagues and co-authors of the respective scientific publications, have proved the large potential of the human microbiome to answer forensically relevant questions. Although the use of the human microbiome in forensics is still in its infancy, the research presented in this thesis, and more to come in the future, holds great promises that eventually human microbiome analysis will become an integral part of forensic analysis, especially whenever other approaches show limitations.

SAMENVATTING

De laatste jaren heeft het humane microbioom de aandacht getrokken van de forensische onderzoeksgemeenschap als een veelbelovend hulpmiddel dat een rol kan spelen bij het beantwoorden van vragen in forensisch zaakonderzoek. Dit was alleen mogelijk door eerdere ontwikkelingen en verbeteringen van de experimentele en analytische hulpmiddelen voor humane microbioom analyse, uitkomsten van klinisch humaan microbioom onderzoek, en de toenemende beschikbaarheid van openbare microbioom data van verschillende locaties in en op het menselijke lichaam ('body sites'). De studies die in deze thesis worden gepresenteerd behoren tot de eerste die forensische vraagstellingen het hoofd bieden met behulp van microbioom analyse. De vraagstellingen die in deze thesis worden behandeld zijn: 'Wat is er gebeurd op het plaats delict?' (*Deel I*), 'Wanneer is de misdaad gepleegd?' (*Deel II*) en 'Wie is de dader van het misdrijf?' (*Deel III*).

In ***Hoofdstuk 1*** worden het onderwerp in brede zin en de reikwijdte van deze thesis geïntroduceerd. Het hoofdstuk begint met een beschrijving van het concept van het humane microbioom en haar relevante eigenschappen voor het werk gepresenteerd in deze thesis. Dit hoofdstuk gaat verder met een beschrijving van de twee meest gebruikte benaderingen voor het genereren van humaan microbioom data, namelijk: 'targeted amplicon sequencing' en 'whole genome shotgun sequencing'; van beide technieken worden ook de voordelen en de beperkingen besproken. Vervolgens, worden bijzondere eigenschappen van microbioom data beschreven, evenals manieren om dergelijke data te analyseren en te interpreteren. Tenslotte schets ik de contouren van toepassingen van het humane microbioom in forensische wetenschappen zoals beoogd voorafgaande aan de start van het onderzoek voor deze thesis. Tevens bespreek ik enkele tekortkomingen van eerdere studies die ik tracht te ondervangen met het werk gepresenteerd in mijn thesis.

Deel I introduceert microbioom variaties, *in vivo*, tussen verschillende plaatsen van het lichaam welke kunnen worden gebruikt voor het afleiden van de oorsprong van forensisch relevante, humane biologische monsters vanuit een enkelvoudige bron. Wanneer toegepast in toekomstig forensisch zaakonderzoek, dan kan dit helpen om te relevantie van een biologisch spoor gevonden op het plaats delict te beoordelen, alsmede helpen om een reconstructie te maken van de gebeurtenissen die zich daar af hebben gespeeld. ***Hoofdstuk 2*** beschrijft de classificatie van de oorspronkelijke body sites van epitheel materiaal, namelijk: speeksel, huid en vaginale afscheiding. Onze voorgestelde aanpak is gebaseerd op het gebruiken van referentie 16S rRNA gen sequentie data van het 'Human Microbiome Project' (HMP) van orale, huid en vaginale origine (N=1,636). Deze data omvatten de V1-V5 regio van het 16S rRNA gen waaruit we 240 nucleotide posities hebben geïdentificeerd die veelal verschillen tussen de drie onderzochte plaatsen van het lichaam. De genomische informatie op deze posities werd gebruikt om 50 onafhankelijke 'deep learning' (DL) artificiële neurale netwerken (ANN) te trainen, gebruikmakend van een taxonomie-onafhankelijke aanpak. De output van het DL algoritme bestond uit toegewezen classificatie waarschijnlijkheden voor elke body site categorie (d.w.z. speeksel, huid, en vaginaal), deze waarden werden bepaald door het gemiddelde van de 50 DL netwerken te nemen. We hebben onze aanpak gevalideerd gebruikmakend van 110 monsters waarvan we *de novo* 16S rRNA gen sequentie data hebben gegenereerd met behulp van de Ion Torrent™ Personal Genome Machine (PGM). De nauwkeurigheid van de body site classificatie was hoog voor deze monsters met aan

AUC≥0.99. Daarnaast hebben we een forensische validatie uitgevoerd op 41 nagemaakte forensische zaakmonsters, welke realistische en vaak voorkomende forensische scenario's vertegenwoordigden (maximaal 7.6 jaar oud). Onze taxonomie onafhankelijke DL aanpak presteerde ook goed in de classificatie van deze samples aan hun respectievelijke body site origine met aan AUC≥0.99.

Hoofdstuk 3 beschrijft de classificatie van de body site origine van bloedmaterialen, namelijk: veneus bloed, menstruatiebloed, nasaal bloed en vingerprik bloed. In dit geval waren er geen publieke data gegenereerd vanuit deze vloeistoffen beschikbaar in de HMP dataopslag. Daarom hebben we de HMP referentiedata van de body sites waar de forensische bloedsporen vandaan kunnen komen gebruikt, d.w.z. vaginaal epithelium voor menstruatiebloed, nasale slijmvliezen voor nasaal bloed en huid epitheel voor vingerprik bloed (N=623). Daarnaast hebben we *de novo* data gegenereerd voor veneus bloed (N=150) omdat geschikte referentiedata niet publiekelijk beschikbaar was. Het Ion S5™ sequencing platform is gebruikt om de *de novo* sequentie data te genereren. Deze keer hebben we 920 nucleotide posities geselecteerd in de V1-V5 regio van het 16s rRNA gen voor de DL ANN training. We hebben onze aanpak gevalideerd met behulp van 94 test monsters waaruit bleek dat een hoge nauwkeurigheid van classificatie kon worden behaald voor alle categorieën van body sites met een AUC≥0.98. We hebben ook aangetoond dat de dag van de menstruatie geen effect had op het correct classificeren van menstruatiebloed in de vaginale categorie. Tenslotte, hebben we forensische validatie uitgevoerd in 45 nagemaakte zaakmonsters (maximaal 21 jaar oude monsters), ook dit resulteerde in correcte classificatie van veel van de monsters met aan AUC=0.79-0.93.

Deel II introduceert humane microbioom variatie *ex vivo* (d.w.z. vanaf het moment dat het zich buiten het menselijke lichaam bevindt in de vorm van een biologisch spoor) om de tijd sinds de plaatsing van het humane biologische spoor in te schatten. Wanneer in de toekomst toegepast in de forensische praktijk dan kan deze bepaling waardevol zijn om zo aanwijzingen te geven over wanneer een d.m.v. DNA geïdentificeerde donor het biologische spoor heeft achtergelaten. Dergelijke informatie kan bruikbaar zijn om onbekende daders te vinden en kan de politie helpen om alibi's van bekende verdachten te verifiëren. Tevens, als meerdere sporen, die toebehoren aan verschillende donoren, zijn gevonden op het plaats delict dan kan informatie over de tijd sinds de plaatsing van het spoor de onderzoekers helpen om te beslissen welke van de sporen de meeste relevantie kunnen hebben voor het onderzoek en verdere analyse behoeven, mits de tijd van het misdrijf bekend is. In ***Hoofdstuk 4*** hebben we speeksel gekozen als het type humaan biologische spoor om de tijd sinds de plaatsing van het spoor te bepalen. Deze keuze is gemaakt gezien de forensische relevantie, de grote hoeveelheid aanwezige bacteriën, en de relatief stabiele microbioom compositie vergeleken met andere body sites, van dit type biologisch materiaal. Eerst hebben we de meest voorkomende bacteriële soorten in humaan speeksel geïdentificeerd gebruikmakend van publiekelijk beschikbare 16S rRNA gen sequentie data van 1,848 monsters. Daarna hebben we de tijdsafhankelijke verandering van 15 geïdentificeerde soorten beoordeeld m.b.v. *de novo* gegenereerde 16S rRNA gen sequentie data (Illumina Miseq) vanuit speeksel sporen van twee individuen die waren blootgesteld aan binnenshuise condities voor verschillende intervallen met een maximum van één jaar. Vanuit deze data hebben we vier bacteriële soorten geselecteerd, namelijk: *Fusobacterium periodonticum, Haemophilus parainfluenzae, Veillonella dispar,* en *Veillonella parvula*, die allen statistisch significante

tijdsafhankelijke veranderingen toonde en voor deze vier soorten hebben we vervolgens een 4-plex qPCR test ontwikkeld voor een gerichte kwantitatieve analyse van deze soorten. Daarna hebben we deze test toegepast op aan binnenshuise condities blootgestelde speeksel monsters van 15 individuen die met verschillende intervallen van maximaal één maand zijn verzameld. Met deze onafhankelijke en uitgebreidere dataset konden we bevestigen dat dat de vier geselecteerde bacteriesoorten inderdaad tijdsafhankelijke verschillen vertoonden. Hierna hebben we, gezien de hoge interindividuele variantie, individu-specifieke lineaire regressie modellen gebouwd om de tijd sinds de plaatsing van het spoor te voorspellen, dit resulteerde in een gemiddelde 'mean absolute error'(MAE) van vijf dagen. De tijd sinds de plaatsing van 181 sporen (82%) werden in de correcte week voorspeld. De voorspellingsmodellen werden ook getoetst met sporen van dezelfde individuen zeven maanden later en onder vergelijkbare condities bewaard, dit resulteerde in een gemiddelde MAE van 9 dagen. Met deze studie hebben we de potentie aangetoond van een op humane bacteriën gerichte aanpak voor het schatten van de tijd sinds de plaatsing van speeksel sporen in een forensisch scenario, hetgeen in de toekomst wellicht van worden uitgebreid naar humane sporen die hun oorsprong vinden in andere forensisch relevante body sites.

Deel III introduceert het gebruik van het humane microbioom voor het afleiden van menselijke leefstijl vanuit humane biologische sporen, met als voorbeeld het roken van sigaretten. Doorgaans wordt data omtrent rookgedrag verzameld voor medische/ onderzoeksdoeleinden d.m.v. zelf-gerapporteerde enquêtes die onbetrouwbaar zouden kunnen zijn, dit zou kunnen worden ondervangen door op het microbioom gebaseerde voorspelling van rookgedrag. Daarnaast zou dergelijke kennis waardevolle informatie kunnen bieden over de lange termijn leefstijl gewoonten van een patiënt in een diagnostische setting. In een forensisch scenario zou het kunnen helpen om onbekende daders van een misdrijf te vinden door een aanknopingspunt voor het onderzoek te bieden, dus door de groep van verdachten te beperken tot de rokers. In ***Hoofdstuk 5*** introduceren we een nieuwe analytische aanpak gebaseerd op data 'augmentation' en 'machine learning' (ML) om het 'class imbalance' probleem in op microbioom gebaseerde classificatie te adresseren en hebben we het toegepast op de vraag omtrent het voorspellen van rookgedrag, als voorbeeld. Daarvoor hebben we publiek toegankelijke 16S rRNA gen sequentie data en daaraan verbonden metadata van rookgedrag van twee studies verzameld waar 'class imbalance' een onontkoombaar probleem was, d.w.z. 175 rokers t.o.v. 1,070 niet-rokers. Drie verschillende data 'augmentation' technieken: 'synthetic minority oversampling technique' (SMOTE), 'adaptive synthetic' (ADASYN) en 'tree-based associative data augmentation' (TADA) zijn toegepast in combinatie met zeven ML methoden: 'logistic regression', 'k nearest neighbors', 'support vector machine with linear and radial kernels', 'decision trees', 'random forest', en 'extreme gradient boosting'. 'K-fold nested cross-validation', is gebruikt om met de verschillende typen 'augmented data' en als nulmeting de 'non-augmented data' om de predictie uitkomsten te valideren. We hebben gevonden dat het combineren van 'augmented data' met ML over het algemeen de beter presteerde dan de nulpuntsmeting methoden en dat TADA en 'support vector machine with linear kernel' de beste voorspelling opleverde. In deze studie hebben we met succes het probleem van 'class imbalance' in microbioom data voor betrouwbare classificatie geadresseerd en hebben we een voorstel gedaan voor een gegeneraliseerde aanpak voor andere data en classificatie problemen. Hiernaast hebben we een oplossing voor microbioom gebaseerde voorspelling van rookgedrag vanuit speeksel met

16S rRNA gen sequentie data ondanks het probleem met ’class imbalance data’, deze oplossing in de eerste van zijn soort.

Hoofdstuk 6 geeft een algehele discussie van de resultaten getoond in *Hoofdstuk 2-5* van deze thesis. Ik bediscussieer de technische kant van microbiome onderzoek, inclusief het gebrek van overeenstemming in het veld, potentiële vooringenomenheden geïntroduceerd gedurende de verschillende stappen van de microbioom pipeline en hoe deze kunnen worden voorkomen of verminderd tijdens microbioom analyse. Ik neem ook de huidige obstakels bij microbioom analyse die voortkomen uit de onderrepresentatie van bepaalde microbiomen in de literatuur en het gebrek aan standaardisatie van metadata in ogenschouw. Dit hoofdstuk richt zich ook op het effect van inter- en intra-individuele variatie in microbioom resultaten. Tot slotte, bediscussieer ik de specifiek forensische overwegingen die nader onderzoek behoeven alvorens microbioom werk, zoals gepresenteerd in deze thesis, kan worden opgenomen in de forensische gereedschapskist voor analyses in zaakwerk, zoals bemonstering, opslag en stabiliteit van het microbioom in forensische sporen.

Met het onderzoek beschreven in deze thesis heb ik, samen met mijn collega’s en coauteurs van de verschillende wetenschappelijke publicaties, bewezen dat humane microbioom analyse een grote potentie heeft om forensisch relevante vragen te beantwoorden. Ondanks het feit dat humane microbioom analyse in het forensische werkveld nog in zijn kinderschoenen staat, zijn de onderzoeken in deze thesis en het onderzoek dat hierop zal volgen veelbelovend. In het bijzonder waar andere typen analyse gebreken vertonen, daar zal uiteindelijk humane microbioom analyse een integraal onderdeel van forensische analyse kunnen worden.

AUTHOR'S AFFILIATIONS

In order of appearance in this thesis.

Erasmus MC, University Medical Center Rotterdam, Department of Genetic Identification, Rotterdam, The Netherlands
Athina Vidaki, Arwin Ralf, Diego Montiel González, Manfred Kayser

Erasmus MC, University Medical Center Rotterdam, Department of Internal Medicine, Rotterdam, The Netherlands
Djawad Radjabzadeh, Robert Kraaij, André G. Uitterlinden

Erasmus MC, University Medical Center Rotterdam, Department of Epidemiology, Rotterdam, The Netherlands
Robert Kraaij, André G. Uitterlinden

Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland
Cordula Haas

CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain
Oscar Lao

Universitat Pompeu Fabra (UPF), Barcelona, Spain
Oscar Lao

PUBLICATION LIST

(1) **Díez López C**, Vidaki A, Kayser M. Integrating the human microbiome in the forensic toolkit: current bottlenecks and future solutions. *Submitted*

(2) **Díez López C**, Montiel González D, Vidaki A, Kayser M. Addressing class imbalance in microbiome-based classification using data augmentation and machine learning with smoking habit prediction from saliva as example. *Submitted*

(3) **Díez López C**, Kayser M, Vidaki A. 2021. Estimating the time since deposition of saliva stains with a targeted bacterial DNA approach: a proof-of-principle study. Front Microbiol 12: 647933.

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

(5) **Díez López C**, Vidaki A, Ralf A, Montiel González D, Radjabzadeh D, Kraaij R, Uitterlinden A G, Haas C, Lao O, Kayser M. 2019. Novel taxonomy-independent deep learning microbiome approach allows for accurate classification of different forensically relevant human epithelial materials. Forensic Sci Int Genet 41:72-82.

(6) Vidaki A, **Díez López C**, Carnero-Montoro E, Ralf A, Ward K, Spector T, Bell JT, Kayser M. 2017. Epigenetic discrimination of identical twins from blood under the forensic scenario. Forensic Sci Int Genet 31:67-80

PHD PORTFOLIO

Name PhD candidate: Celia Díez López
Department: Genetic identification
PhD period: 01-05-2017 until 31-07-2021
Promotor: Prof. dr. M. Kayser
Co-promotor: Assist.-Prof. dr. A. Vidaki

1. PhD training		
	Year	**Workload (ECTS)**
General academic skills		
Safely Working in the Laboratory (MGC)	2018	0.3 ECTS
Scientific Integrity (MGC)	2018	0.3 ECTS
Career Development for PhD Candidates (MolMed)	2018	0.3 ECTS
Presenting Skills for PhD students and Post Docs (MolMed)	2019	1 ECTS
Publication: Reading, Writing and Reviewing (ISFG)	2019	0.3 ECTS
Personal Leadership and Communication for PhD Students and Post Docs (MolMed)	2020	1 ECTS
Write your own Career Story for the Coming Year and Focus on your Career (EUR)	2020	0.15 ECTS
Biomedical English Writing (MolMed)	2020	2 ECTS
Research skills		
Basic Course on R (MolMed)	2017	2 ECTS
Practical Linux (MGC)	2017	0.4 ECTS
Genetics for PhD Students in Biomedical Sciences (MGC)	2017	3 ECTS
Molecular Microbiology: Evolution of Diagnostics and Forensic Microbiology, Molecular Diagnostics XI (MolMed)	2017	0.3 ECTS
Basic Introduction on SPSS (MolMed)	2017	1 ECTS
Special Topics on Optogenetics (MGC)	2018	2 ECTS
Biostatistical Methods I: Basic Principles Part A (NIHES)	2018	2 ECTS
Microbiomics I (MolMed)	2018	2 ECTS
Data Analysis in Python (MolMed)	2019	1.7 ECTS
Microsoft Power Query (MolMed)	2020	0.3 ECTS
Machine Learning with Python (MolMed)	2020	0.8 ECTS
Presentations (including presentations on international conferences)		
Oral presentations on Paper Club, Department of Genetic Identification, Erasmus Medical Center Rotterdam	2017-2020	0.5 ECTS
Poster presentation on 22nd Molecular Medicine Day, Rotterdam, the Netherlands	2018	1.3 ECTS
Slam session presentation on 25th MGC Workshop, Texel, the Netherlands	2018	1 ECTS

Poster presentation on 7th International Human Microbiome Consortium Congress, Killarney, Ireland	2018	2 ECTS
Oral presentation on International Commission on Missing Persons visit, Rotterdam, the Netherlands	2018	0.3 ECTS
Oral presentation on 28th MGC Symposium, Leiden, the Netherlands	2018	1.3 ECTS
Poster presentation on 26th MGC Workshop, Maastricht, the Netherlands	2019	2 ECTS
Poster presentation on 8th Congress of European Microbiologists, Glasgow, Scotland, United Kingdom	2019	2 ECTS
Oral presentation on 28th Congress of the International Society for Forensic Genetics, Prague, Czech Republic	2019	2 ECTS
Oral presentation on Virtual Winter Seminars, Students and Young Researchers in Forensic DNA	2020	1.3 ECTS
Oral presentation on 27th MGC Workshop, Leiden-Rotterdam, the Netherlands	2021	2 ECTS
Poster presentation on 8th International Human Microbiome Consortium Congress, Barcelona, Spain	2021	2 ECTS
National and international conferences (without presentations)		
BIG Biomedical Sciences Theme Day, Rotterdam, the Netherlands	2019	0.3 ECTS
23rd Molecular Medicine Day, Rotterdam, the Netherlands	2019	0.3 ECTS
The Barcelona Debates on the Human Microbiome 2020, Barcelona, Spain	2020	0.25 ECTS
World of Microbiome: Pregnancy, Birth & Infancy 2020, Vienna, Austria	2020	0.7 ECTS
Seminars, workshops and symposia		
27th MGC Symposium, Rotterdam, the Netherlands	2017	0.3 ECTS
4th CLHC Forensic PhD Symposium, Amsterdam, the Netherlands	2017	0.3 ECTS
The Erasmus University Research Integrity Week, Rotterdam, the Netherlands	2017	0.3 ECTS
CHAMALEON - Connecting Academia, Medicine, Entrepreneurs, Life Sciences and Education, Rotterdam, the Netherlands	2018	0.3 ECTS
2. Teaching activities		
	Year	**Workload (ECTS)**
Supervising Bachelors thesis		
BSc. student Liese Boonstra, Leiden University of Applied Sciences '*Prediction of individual's smoking habits from human oral microbial signatures'*	2019-2020	6 ECTS
Lecture		
Minor *Genetics in the society*, Department of Clinical Genetics, Erasmus University Medical Center	2020	0.7 ECTS

Webinar		
'*Using the microbiome to unlock the secrets of forensic evidence: classification of the body source of human traces'* for Qiagen	2020	0.5 ECTS
Peer-review for scientific journals		
International Journal of Legal Medicine (x2), *Applied Environmental Microbiology* (x1), *Frontiers in Microbiology* (x1), *Frontiers in Cellular and Infection Microbiology* (x1), *Microorganisms* (x1), *Microbiome* (x1)	2018-2021	0.5 ECTS
3. Other activities		
	Year	**Workload (ECTS)**
Organizer/presenter '*Crime scene investigation in a hotel suite*' session, Science Hotel, Science Meets the City event, Rotterdam, the Netherlands	2018	0.3 ECTS
Project finance and administration '*I-Lab-U: speed up a scientist*' science communication project, Department of Genetic Identification, Erasmus University Medical Center	2019-2020	5 ECTS
Interview '*Novel uses of the microbiome in forensic genetics: an interview with Celia Díez López*', Biomarker Insights, Qiagen	2019	0.15 ECTS
Organizer 27th MGC Workshop	2020-2021	0.5 ECTS
Interview 2020 Insights Magazine, Forensic Methods Section, Qiagen	2020	0.3 ECTS
'*Using the human microbiome to decipher crime scene traces*', Qiagen Customer Stories	2020	0.15 ECTS

ABOUT THE AUTHOR

Celia Díez López was born on July 2nd, 1993 in Burgos, Spain. In 2011 she completed high school at the IES Cardenal López de Mendoza in Burgos with specialization in science and technology. In the same year, she started her studies in Biotechnology at the University of Salamanca in Salamanca, Spain. In 2014, during her third Bachelor's year, she spent two months in the summer as a research intern on the 'Characterization of quantitative trait locus (QTLs) involved in the morphology and domestication of melon' at the Department of Genomics in Plant Breeding in the Molecular and Cellular Plant Biology Institute in Valencia, Spain (Supervisor: dr. Antonio J. Monforte Gilabert). During her last Bachelor's year, she spent five months as an intern for her final thesis on the 'Analysis of genetic variants using RT-PCR associated with the development of brain aneurysms' at the Department of Molecular Medicine in the University of Salamanca (Supervisor: dr. Rogelio González Sarmiento), after which she obtained her Bachelor's degree in 2015. In the same year, she moved to London, United Kingdom, to study a Master's degree in Forensic Science at King's College London, from which she graduated with distinction in 2017. For her 8-month Master's research internship she moved to Rotterdam, the Netherlands, for 'Investigating the potential of a methylation-specific qPCR-based method to differentiate between monozygotic twins for forensic applications' at the Department of Genetic Identification in Erasmus MC University Medical Center Rotterdam, the Netherlands (Supervisors: Prof. dr. Manfred Kayser, Assist.-Prof. dr. Athina Vidaki), which resulted in a publication at the leading forensic genetics journal. From May 2017 onwards, she became a PhD candidate at the same department with the same supervisors. Her research focused on 'Novel applications of the human microbiome in forensic science'. During her PhD she was four times awarded with bursaries for her participation at international scientific meetings. Additionally, she was one of the top ten finalists in the 2018 Microbiome Awards, PhD Student Category organized by Qiagen. Finally, together with other colleagues from her department, she won the 250th Jubilee Scientific Competition for the experimental connection between the scientists and the public, where she was in charge of the project finance and administration.

ACKNOWLEDGEMENTS

After four years, my PhD journey comes to an end. Like everything in life, every ending is a new beginning and a new opportunity. If there is something I have to highlight, it is my personal growth as I was progressing on my PhD. As the Spanish poet Antonio Machado wrote '*Walker, there is no path, the path is made by walking*' (*Caminante, no hay camino, se hace camino al andar*; in Spanish). The PhD period has been a remarkable time in my life and thus, I would like to express my gratitude to all the people who have had a positive impact on it.

First, I would like to thank my supervisor Prof. dr. Manfred Kayser for giving me the opportunity to pursue a PhD at the Department of Genetic Identification. Thank you so much for your support, good advice and flexibility through the years.

Special thanks to my co-promotor Assit.-Prof dr. Athina Vidaki. My dearest Athina, there are not enough words to express how marvellous your guidance has been, both as a supervisor and a friend. I can say loud and proud that you are a true role model, and I am sure that, as it happened with me, you will continue inspiring young researchers over the years. Thanks for your immense empathy, kindness and optimism. I feel lucky that life gave me the opportunity to meet you in London and follow your steps to Rotterdam.

To all my incredible colleagues and friends, thank you for all the wonderful moments we have spent together in and outside the lab. You are without a doubt one of the best things I get out of the PhD journey. Diego, my friend, I will be eternally grateful for all your valuable input in my PhD research. Thank you so much for your brilliant work and your goodwill. Dear Gabriela, I could have not asked for a better person to spend my PhD days side by side; thanks for always listening and giving good advice. Dear Arwin, simply your presence fills the lab with joy. Thanks a lot for everything you have taught me in the lab, for always being available when I needed something and for your help with the samenvatting of this thesis. Dear Benjamin, discussions with you are always a source of inspiration and learning; thank you a lot for all your valuable advice. Dear Lucie, thank you for all the interesting conversations and for your incredible sense of humour. Dear Roy, thanks for double-checking the samenvatting of this thesis. Thank you to all the colleagues who I was lucky to work with, but left the lab already: Nefeli, Silvana, Irma, Vivian, Leroy, Delano, Hedayat, Rochelle, Faidra, Dion, Alex, Liese and Floor.

My sincere thanks to my reading committee, Prof. dr. A.G. Uitterlinden, Prof. dr. T. Sijen and Assoc.-Prof. dr. J.P. Hays. Thank you for taking the time to read my thesis. I would also like to extend my thanks to the members of the defence committee Prof. dr. P. de Knijff and Assoc.-Prof. dr. C. Belzer; thank you a lot for being there for my defence.

Thank you to all the friends who either in closeness or distance made my PhD journey smoother. Thank you for being there to celebrate the good moments and for your support in the difficult ones.

Infinitas gracias a mis padres por siempre confiar en mí y por apoyarme en las decisiones que tomo. Gracias por educarme en la libertad y por estar siempre cerca de mí, cuidándome, aún en la distancia. Papá, gracias por ser mi referente de esfuerzo, constancia, atención al detalle y eficacia. Mamá, gracias por ser mi ejemplo de fortaleza, de lucha por lo que a uno le apasiona y de reivindicación.

Gracias Eze por tu infinita paciencia, tu bondad, tu buen humor. Gracias por aceptarme, por dejarme ser y ayudarme a crecer. Gracias por tanta alegría y admiración en mis éxitos, y por tanto apoyo y comprensión en mis caídas.