DNA Based Methods in Intelligence - Moving Towards Metagenomics

Gabriella Mason-Buck[1#], Eran Elhaik[2#], Alexandra Graf[3#], Jake Robinson[4], Ewelina Pośpiech[5], Manuela Oliveira[6, 7, 8], Josef Moser[3], Patrick K. H. Lee[9], Dedan Githae[5], David Ballard[1], Tae-Hyuk Ahn[10], Yana Bromberg[11], Carlos S. Casimiro-Soriguer[12], Eliza Dhungel[10], Jolanta Kawulok[13], Carlos Loucera[12], Feargal Ryan[14], Alejandro R. Walker[15], Chengsheng Zhu[11], Christopher E. Mason[16,17,18], António Amorim[6, 7, 8], Denise Syndercombe Court[1], Wojciech Branicki[5,19]*, Paweł P Łabaj[5]*

[1] King's Forensics, King's College London, London, United Kingdom
[2] Department of Biology University of Lund, Lund, Sweden
[3] Department of Applied Life Sciences, University of Applied Sciences, FH Campus Wien, Austria
[4] Department of Landscape, The University of Sheffield, Sheffield, United Kingdom
[5] Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[6] Instituto de Investigação e Inovação em Saúde (i3S), Universidade do Porto, Porto, Portugal
[7] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Porto, Portugal
[8] Faculdade de Ciências, Universidade do Porto. Porto, Portugal
[9] School of Energy and Environment, City University of Hong Kong, Hong Kong
[10] Saint Louis University, Saint Louis, USA
[11] Department of Microbiology and Biochemistry and Department of Genetics, Rutgers University, USA
[12] Clinical Bioinformatics Area, Fundación Progreso y Salud (FPS), Sevilla, Spain
[13] Silesian University of Technology, Gliwice, Poland
[14] South Australian Health and Medical Research Institute, Adelaide, Australia
[15] Department of Oral Biology, University of Florida, USA
[16] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
[17] The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA
[18] The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA
[19] Central Forensic Laboratory of the Police, Warsaw, Poland
[#] The first three authors contributed equally
* Corresponding authors

Abstract:

Advancements in DNA methods and biotechnology have enabled forensic scientists to explore the DNA evidence found as part of a criminal investigation on a much more comprehensive and predictive level. This has led to a rise in research into DNA intelligence tools such as phenotypic prediction (i.e., eye and hair colour) and inference of biogeographical ancestry. Both of which can be applied to gain further insights about a scene or sample in question. Although microorganisms have played a role in forensics for decades, investigations were focused on the pathogenicity aspect, mainly to determine the cause and time of death. Recent progress in studying the human microbiome has implicated the potential use of this data in forensics. Since each individual, place, or item has its own microbial pattern, a new suite of tools are now available to be exploited in criminal investigations. Although there is much interest and potential for these emerging metagenomic and microbial forensic tools, best practices and reference ranges need to be established before they are implemented. Here,

we discuss existing DNA intelligence tools applied to forensic science, the application of microbial forensics and metagenomics along with the challenges and concerns that future developments entail.

1. Introduction

DNA profiling of microsatellite markers (short tandem repeats, STRs) is the gold standard in human identification studies for criminal casework, as well as for the identification of human remains[1–3]. The methods used for DNA profiling have significantly improved in the last decade, allowing simultaneous analyses of multiple STR loci, while still maintaining functionality on both low quantity and reduced quality DNA. Analysis time has also been reduced and the use of simple capillary electrophoresis (CE) is still favoured[4]. Moreover, multiplex STR human identification methods currently on the market can provide information about the quality of the sample and include additional complementary markers for reliable sex determination[5,6]. An STR-based analysis relies on the ability to compare results obtained from a crime scene sample (casework sample) to a profile obtained from a known individual (reference sample)[7,8]. If the reference sample is unavailable and direct comparison cannot be made, the profile from the casework sample can be searched against a DNA database to look for direct matches or matches to potential relatives (familial searching).

In a number of serious criminal cases, DNA mass screenings using these technologies have been arranged. This method was first applied in 1987 to a case in the U.K., where 4,000 men aged between 17 and 34 provided either a blood or saliva sample with the aim of identifying the perpetrator of two sexual assaults and murders of young women. In 2018, the Dutch Police sampled 21,500 men with the aim of solving a 20-year-old murder[9]. However, mass screening is not considered a standard approach, since it is both laborious and expensive, and it creates ethical issues due to its coercive nature and the use of familial matching. When no direct or familial match can be assigned in the database, the case becomes “cold.” The work is then halted until either new evidence is found, or a decision is made to apply additional investigative tools to the case that may aid in identifying an individual.

However, the dawn of new DNA intelligence tools has given forensics the prospect to narrow down the pool of potential suspects based on particular characteristics of the perpetrator. DNA intelligence methods include, but are not limited to, the inference of biogeographical ancestry, age estimation and the prediction of externally visible characteristic such as eye, hair and skin colour[10]. The development of new investigative tools remains an important research focus in forensic genetics and has prompted the forensic community to change the technology from CE based analysis to massively parallel sequencing (MPS), which enables the simultaneous sequencing of many genetic markers[11,12]. The current forensic DNA intelligence tools rely on predicting information from human DNA, but emerging evidence has shown that this can be significantly strengthened by including non-human DNA analysis. Specifically, metagenomic profiles (Bacteria, Archaea and Eukaryota - fungi, plant, animal DNA) obtained

from environmental or casework samples can provide independent and complementary intelligence to facilitate an investigation[13,14]. It has already been shown that it is possible to harness these profiles to infer information on the activity of individuals, the location where someone or something has been and what they were in contact with[15,16].

In 2007, the National Research Council (US) Committee on Metagenomics identified the usefulness of metagenomics in the field of forensic science 'to precisely identify and characterise microbes that have played a role in war, acts of terrorism, and crime events, thus contributing to discovering the source of the microbes and the party responsible for their use'[17]. Sequencing techniques for non-human DNA analysis have significantly advanced over the past five years, and the cost of using such tools become affordable[18]. However, proper implementation of metagenomics in forensics requires further basic research from sample selection to data analysis and it remains fraught with challenges (e.g., the effect of background contamination from a kit - the 'kitome' - from contamination in consumables and reagents) and uncertainties that constraint the process. Additionally, it is yet unclear which features of microbial communities are of primary interest in an intelligence or forensic perspective. Is it the selection of key species with predictive power in terms of geolocalisation, or marker panels of bacterial regions that will help to identify perpatrators of crime or bring new insights in to a crime scene? In the past decade, there has been a renewed interest in utilising genomic and metagenomic data to increase the predictive capabilities of intelligence tools, primarily using machine learning and artificial intelligence-based approaches, all of which can fundamentally change how we leverage these molecular methods in forensic science.

2. Current DNA Intelligence tools

Any personally identifying characteristics, even the most elementary ones (e.g., sex or biogeographical ancestry), inferred from DNA found at a crime scene can provide important leads to an ongoing investigation. Forensic DNA intelligence has advanced considerably in predicting human appearance traits. The prediction accuracy of a genomic trait is often limited by the amount of explained heritability, such as the proportion of known genes and DNA variants involved in determining that trait. Other important factors to consider are the statistical methods used in prediction modelling and the size of the datasets available[19]. The availability of large genome-wide association study (GWAS) data for physical traits (such as the UK Biobank) has boosted scientific investigations in the field of DNA phenotyping and has been used to leverage immediate implementations to forensic genetics[10]. In the following, we review the leading state-of-the-art methods.

2.1 Human pigmentation (eye, hair and skin colour)

Human pigmentation, which comprises eye, hair and skin colour, is highly heritable and genetically less complex than many other physical traits[20]. Prediction accuracy is generally high for extreme pigmentation categories like blue and brown eyes, red hair or white and dark skin, where single genes and polymorphisms play a significant role, but lower for

intermediate phenotype categories where modifying genes and gene-gene interactions gain significance[10,21]. Hence, prediction of blue and brown eye colour (depending mostly on the rs12913832 polymorphism in HERC2) and red hair colour (depending mostly on variation in the *MC1R* gene) are exceptionally accurate with a high Area Under the Curve of prediction, (AUC) of ~0.9 (where 1.0 is a perfect predictive ability). However, intermediate eye colour and other hair colour categories are predicted with lower accuracies (AUC<0.85)[22,23]. The selection of suitable DNA variants allowed the development of the first prediction models and tools of practical value in forensics. The HIrisPlex-S method based on targeted MPS and the use of an online prediction calculator is the most advanced tool for simultaneous prediction of eye, hair and skin colour to date[24]. Recent papers have shown that the next step in prediction of pigmentation traits may require genomics solutions and more advanced machine learning approaches[25,26].

2.2 Hair morphology and baldness

The usefulness in predicting other head-hair features, such as hair shape and hair loss, has long been recognised in the field of forensic genetics. A recent large study demonstrated that 32 DNA variants can predict straight hair with AUC=0.66 in Europeans and AUC=0.79 in non-Europeans[27]. Though the importance of sex and age for hair loss prediction was already known, this study has also demonstrated their role in predicting hair shape. For men aged 50 and above, hair loss can be predicted using only 20 single nucleotide polymorphisms (SNP) (AUC=0.76)[28]. A recent GWAS study of 70,000 males reported that 71 genetic loci for hair loss are responsible for ~38% of the variation observed[29]. These findings provide a good basis for improving hair loss prediction systems in the future.

2.3 Height

As one of the most recognisable and stratifying physical traits – height – is highly heritable (up to 80%)[30]. Yet, while inferring height from DNA is a primary goal of forensic DNA phenotyping, the efforts put into the identification of genes that determine an individual's height were fruitless for forensics. The height of an individual exhibits both high heritability and high variation, and it is controlled by the additive effect of thousands of DNA variants that increase the complexity and uncertainty in height prediction. Initial studies suggested that 180 variants can explain only about 10% of phenotypic variance[31]. Even when the number of predictors was increased to ~ 9500, less than 30% of height variance was explained[32], which is insufficient for forensics. Another research indicated that tall stature predictions may be more practical in forensics[33,34]. Height predictions are expected to improve with more data, as shown with a recent genomic approach that used ~20k DNA variants and reported a better accuracy of height prediction (correlation between actual and predicted age ~0.61)[30].

2.4 Facial Appearance

The face is one of the most outwardly visible and recognisable features of an individual, and creating a DNA-based 'photofit' would be of indisputable value for forensics. The human

ability to identify individuals based on their faces is particularly powerful and has already become a ubiquitous feature of smart phone security access, but this ability is often restricted to specific continental groups, or confounded by genetic admixture from cross-regional ancestry[35]. From a genomic perspective, facial appearance is a complex combination of a large number of traits and its genetics is still poorly understood. In recent years, there was increased use of deep learning methods to identify the genes responsible for shaping facial features; however, the first method for facial shape prediction proposed by Claes et al. [36] relies mostly on non-genetic information. The method uses genomic ancestry and gender to create a first sketch that the authors call "base-face," which is then polished using information obtained via 24 SNPs associated with facial variation. Recently, Lippert and co-workers[37] employed complex methods including 3D face morphology applied to whole-genome sequencing data of 1,061 individuals for the prediction of facial appearance and claimed to be able to re-identify individuals from their cohort. The study was heavily criticised[38] on various grounds, primarily that the results did not differ from a simpler method that employed age, sex, and self-reported ethnicity and failed to demonstrate actual identification. An alternative available approach predicts DNA characteristics from 3 Dimensional (3D) facial images, using a face-to-DNA classifier provided the facial image database is on site[39].

2.5 Age estimation

Forensic DNA Phenotyping (FDP) in a broader context also includes age estimation and biogeographic ancestry inference. Ageing plays an essential role in many physical traits, including height, hair colour, hair loss or skin wrinkles, and thus is valuable and, in some cases, even critical for accurate prediction of human appearance traits. But even without these correlations, it has a tremendous impact on the profile of an individual. A real breakthrough in age prediction research was achieved through the identification of robust epigenetic markers in epigenome-wide association studies (EWAS) [40,41]. DNA methylation has since been shown to predict age with far greater accuracy (age correlation = 0.96 and error = 3.6 years) compared to other methods, including determination of telomere length, measurement of the 4,977 bp deletion of mitochondrial DNA, age-dependent accumulation of advanced glycation end-products, racemization of aspartic acid and measurement of signal joint T-cell receptor rearrangement excision circles (sjTRECs) accumulation[42]. Discovery of age-associated DNA methylation sites was followed by the development of the DNA methylation-based age prediction models. The most well-known methods involve analysis of large numbers of differentially methylated sites using microarray technology, which is not practical in forensic investigations[40,43]. However, epigenetic tests based on several most powerful markers that predict age with a mean absolute error of less than 4 years and allowing low-input analyses may be soon implemented in forensics[41,44]. Considering that DNA methylation is tissue-specific not all markers are equally useful in predicting age in different forensic samples. Therefore, dedicated predictive models have been specifically developed for body fluids like blood and saliva. A different set of markers for predicting the epigenetic age in

semen was proposed, demonstrating a completely different pattern of DNA methylation in sperm cells compared to somatic cells[45].

2.6 Forensics genealogy

The Council for the Advancement of Forensic Genealogy defines *forensic genealogy* as 'genealogical research, analysis and reporting in cases with legal implications' [46] and although the usage of DNA for forensic genealogy is not new, up until recently, it was limited to identify unclaimed decedents and military repatriation[47]. The identification of Joseph James DeAngelo (age 73), known as the "Golden State Killer," who was found guilty with 26 counts of murder kidnapping[48] committed during the 70s and the 80s was unfeasible prior to the genetic genealogy era, although the FBI had his DNA samples from 1980[49]. A DNA match with DeAngelo's family members was only made possible when millions of people shared their DNA data and family trees through online platforms. This case demonstrated the limitation of state-owned National Criminal Intelligence DNA databases, which usually collect DNA only from accused and/or convicted felons and explore a limited set of STR markers that allow only a direct or familial searching[50,51]. Following this case, over 50 other cases were resolved using a similar approach[52], raising hope to solve additional "cold cases" and at the same time, raising big expectations in the forensic community[53]. However, a wide array of ethical issues and privacy concerns have also emerged[54,55]. In May 2019, after these concerns were ignored by law enforcements, GedMatch, one of the major websites of genetic genealogy, restricted law enforcements access to their databases[52]. Due to the high potential of genetic genealogy to resolve cases, it is hoped that law enforcement and genealogy service providers should work together to alleviate privacy concerns while retaining access to commercial DNA databases. Concerns should also be made to "false positive" cases, for example, where a DNA match with an individual is due to receiving tissues from another individual[56] rather than their actual involvement in the case.

3. Predicting ancestry and the upcoming possibility to infer a sample origin

3.1 Ancestry prediction

Ancestry inference using ancestry informative markers (AIMs) can be a very effective intelligence tool by narrowing down the number of potential suspects in criminal investigations. AIMs are usually SNPs but can also be Insertion-Deletion Polymorphisms (InDels) on their own or an Indel associated with an STR marker (Indel-STRs)[57,58]. These exhibit large variation in minor allele frequencies (MAF) among populations and can amplify the ancestry signature. STRs on the Y-chromosome (Y-STR) and variation in mitochondrial DNA (mtDNA) can also be useful in ancestry inference studies, due to known differences in haplogroup frequencies in various population groups. Unsurprisingly, forensics makes extensive use of AIMs. Even in more international regions, like the US or European capital

cities, such information is invaluable due to the tendency of migrants to segregate. These methods utilise between a few dozen[57,59,60] to several thousand[61] AIMs. For instance, the T allele of the SNP rs316598 is very rare in Africans (3-14%) but common (>37-77%) elsewhere[62] and can be used to differentiate Africans from non-Africans. AIMs are typically validated by showing that: 1) samples of similar origins clustered together and apart from samples of different localities; and 2) samples exhibit a unique allele frequency pattern, calculated via STRUCTURE[63] or ADMIXTURE[64]. Due to the cost of typing AIMs along with the ability to type poor quality and low quantity DNA samples, they are preferred in forensic analyses and forensic scientists have developed tools that can facilitate ancestry inference[65,66]. Some kits have also become commercially available for forensics (e.g., the ForenSeq DNA Signature Prep Kit[67], Precision ID Ancestry Panel[68]). While AIMs were validated to classify individuals into distinct subcontinental populations, this does not answer the question of whether these populations are indeed genetically separated. Since the said methods rely on a limited set of reference populations and tools that were not designed for biogeography, the classification may also be an artefact inherent to the method[69].

Notably, AIMs-based methods better predict the ancestry of one's closest ancestors (e.g., parents) than their distant ones and do not predict the actual geolocation of the investigated individual directly. The geographical localisation of biological samples to their site or region of origin based on human DNA is assumed from the ancestry. For example, if one has two Italian parents, they are assumed to be Italians (from Italy). This assumption does not always hold, certainly when the prediction is inaccurate or the individual is highly admixed; however, providing hints on the previous location or even ancient origins of individuals connected to the crime scene (or items) may provide intelligence information complementary to the ancestry inference.

3.2 Sample origin

Despite its importance for forensic investigations, biogeographical capabilities matured slowly and received relatively little attention compared to ancestry. One of the reasons being that initial approaches did not properly translate the genetic distances between populations into geographical distances and suffered from many difficulties (Figure 1)[70,71], such as modelling admixed individuals and being cohort-dependent, that is, an individual will be predicted to different regions based on the other individuals in the cohort. Spatial Ancestry Analysis (SPA), for example, failed to assign 98% of the individuals to their countries[70]. Their reliance on a large number of markers (entire SNP set) and low accuracy even for a handful of European countries, rendered these approaches inapplicable for research or forensic applications.

Biogeographical methods typically capitalise on the strong relationship between genetic and geographic distances in worldwide human populations to predict geographical origins. Although major deviations explicable by admixture, extreme isolation, or recent migrations exist – these approaches can be expected to work well in many parts of the world where these

deviations are not the norm and otherwise provide clues as to the culture and heritage of individual of interest. Biogeographical analyses rest on three pillars: 1) the choice of markers, which should be common and ancestry informative, 2) the comprehensiveness of the reference populations, and 3) a model of the genetic-geographic relationships. Past attempts focused on each of these pillars separately with limited success, but a more recent method tried to conquer all three biogeography pillars at once. Relying on 40,000 – 130,000 AIMs, a comprehensive and global reference panel, and directly modelling the genetic-geographic relationships, the Geographic Population Structure (GPS) achieved a high prediction accuracy (Figure 1)[70]. However, localisation was not always corresponding with modern residency. For example, GPS tracked Ashkenazic Jews to Northeastern Turkey, a region dubbed "ancient Ashkenaz," which their ancestors inhabited in the early centuries A.D., and Israeli Druze to Syria and Turkey, whence they emerged in the 11th century[72]. GPS Origins[73] addressed GPS's limitation in modelling two-way admixture, yet the modelling of more complex ancestries remained unaddressed and the cost of using a microarray, along with the DNA amount requirements, may prohibit forensic applications.

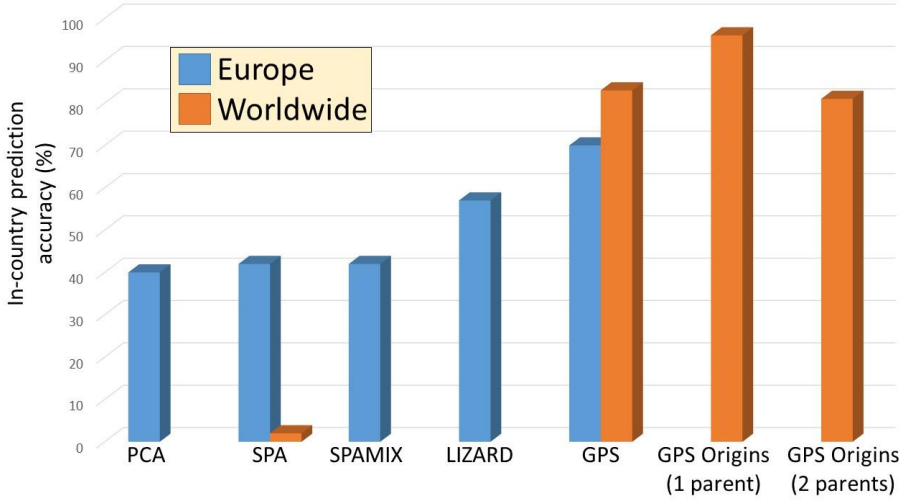

Figure 1. In-country prediction accuracy of biogeographical tools of the second and third generations. Results were curated from the literature (PCA[74], SPA[70,71], SPAMIX[71], LIZARD[71], GPS[70,75], and GPS Origins[73])

To overcome this critical limitation, the latest attempts in predicting biogeographical localisation no longer rely on human DNA, which does not change with geography, but rather the microbiome (Figure 2) [76]. Participants of the CAMDA MetaSUB challenge (section 5) have already demonstrated that the structure of bacterial communities is geographically informative. The advantage of this approach to forensics is its reduced limitation on human DNA, which can be scarce in the crime scene. However, further work is needed to develop forensic-oriented genotyping platforms (e.g., DREAM[77]), refine biogeographical methods, calibrate them to the conditions of the crime scene, and test their reliability after various time periods.

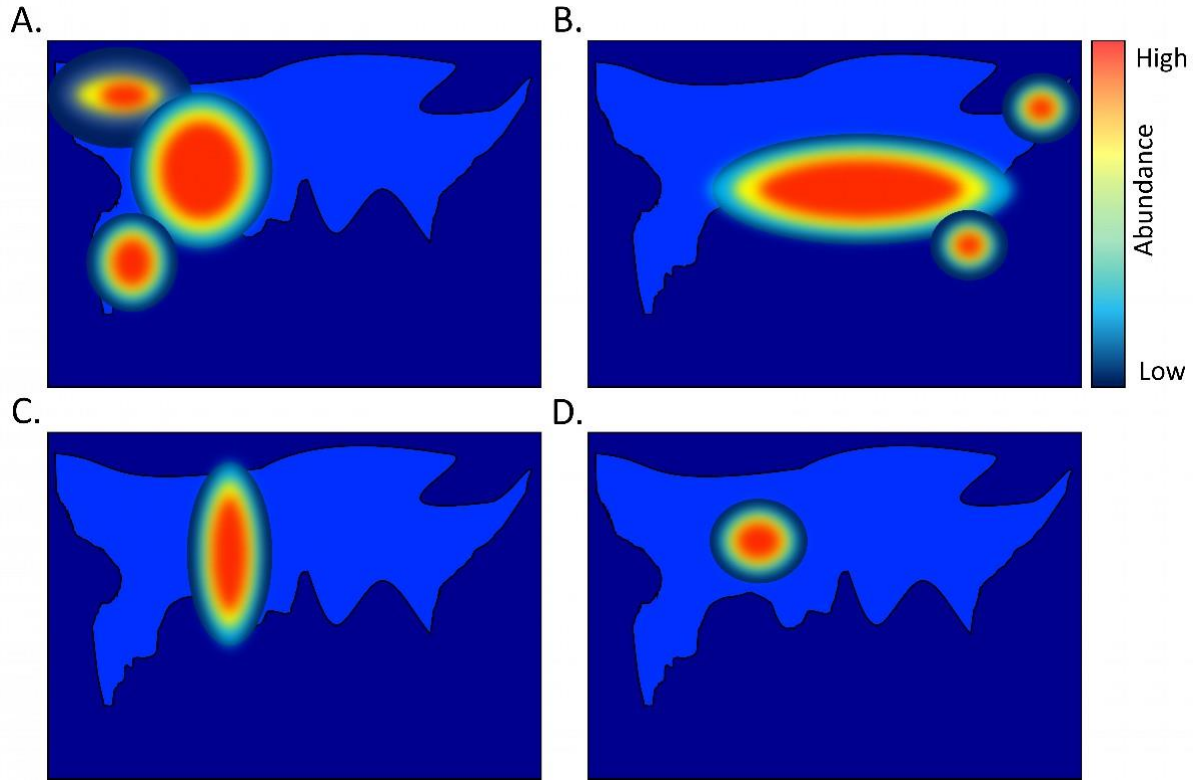

Figure 2. An illustration of how microbiome data can be used for biogeographical applications. Heat maps (A-C) represent the abundance of three hypothetical bacteria as calculated from multiple sampling sites. All three bacteria were found in an object retrieved from a crime scene believed to belong to the suspect. To identify the geographical origin of the object and suspect (which could be different to their ancestry), on overlay of the heat maps (D) identifies the most likely geographical region (darkest red). This example underlines the potential of using global abundance data for thousands of bacteria.

Overall, the continuous improvement in the accuracy and resolution of biogeographical applications, either using genomics or metagenomics, can have major implications for forensics. The successful translation of genomic methods would provide more accurate information both at the country-wide level and for multiple countries whereas metagenomic applications can provide the most recent whereabouts of suspects. Therefore, biogeographical progress must be accompanied by a proper translation for forensics purposes. The combined information on gender, ancestry, physical appearance, biogeography, and whereabouts can provide invaluable intelligence to help progress forensic casework.

4. Current use of metagenomic/microbial forensics

Microbial forensics was first coined back in 2003 as "a scientific discipline dedicated to analysing evidence from an act of terrorism, terrorism crime, or inadvertent microorganism/toxin release for attribution purposes"[78]. This was as a direct result of the need to identify the strain used in the 2001 anthrax attacks in the United States[79]. This

definition did not contain or even recognise the potential of the microbiome for DNA intelligence. Since then, microorganisms have aided in answering some of the most basic forensic questions, such as identifying individuals[80], body fluid or body site prediction[81,82], the estimation of Post Mortem Interval (PMI) [83–85], and determining the probable cause of death[86–88] (Figure 3).

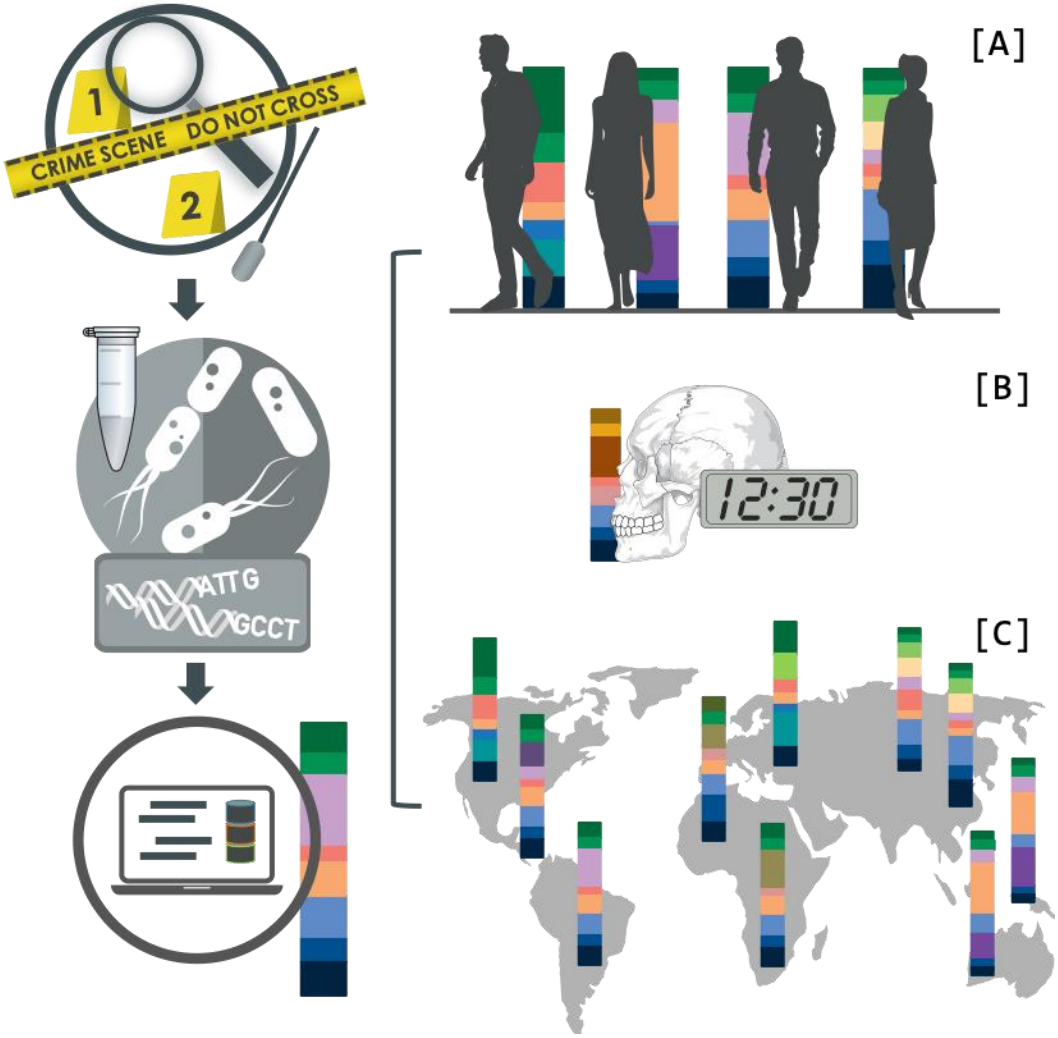

Figure 3. Illustration of the information that can be gained from the microbiome at a crime site, which could be used to identify individuals if no human DNA based identification is possible by e.g., reducing the suspect pool (A). It can help to accurately predict the time of death and help to elucidate the cause of death (B) and give hints toward geographical localisation of individuals or items (C).

4.1 Identifying individuals

Identifying individuals (both the victims and/or perpetrators) is of utmost importance in any criminal investigation. Due to the inter-individual variability of the human microbiome, several studies have attempted to identify individuals by analysing the microbial communities that colonise the human body. A recent study showed that individuals could be identified with an accuracy of 85% by comparing their skin microbiome over a period of two years (*n*=11, 3

time points per year)[89]. The human microbiome, especially skin associated species, are readily transferred to surfaces we come in contact with[90,91]. Several studies have shown that individuals can be detected and identified through items they came in contact with (e.g., as shown by Fierer et al.[92] and Goga et al.[93]) by assigning shoes to their owners with 79% accuracy when samples from shoe insoles were compared with samples taken from plantar skin (*n*=14).

Other preliminary studies suggest that microbiome-based trace evidence could play an important role in providing intelligence in crime investigations. One such project is the Burglary Microbiome Project, which uses mock crime scenes and scientific sampling of homes, residents and intruders. The project employs the microbiome as a forensic tool using unique markers to identify the intruders[94]. Objects such as smoking pipes, medical devices, and telephones have been matched to individuals via skin microbiome sampling (targeted 16S rRNA) with a high degree of accuracy (75-100%)[95]. Microbiome data paired with machine learning approaches harbours great potential. One study reported matching all individuals with their microbiome samples (*n*=12)[96]. Schmedes et al.[97] recently presented the hidSkinPlex, a targeted sequencing panel comprised of 286 skin microbiome markers aimed for forensic human identification. In their study, they tested the hidSkinPlex on the three body-sites of eight individuals (*n*=72) and reported an average accuracy of 94% (92% (foot), 96% (manubrium), 100% (hand)). Body site origin could be predicted with up to 86% accuracy. Microarrays have also proven to be useful in differentiating between faecal samples[98,99]. Microbiome studies on indoor air have shown that, in addition to the microbiome transferred through touch, a distinct microbial “cloud” surrounds the human body[100]. Therefore, under certain circumstances, indirect and less-intrusive sampling methods could be deployed. For example, an analysis of the airborne microbiome (aerobiome) has been able to identify the sex of residential occupants with an accuracy of 79% (*n*=91)[101].

In addition to bacteria, whole metagenome sequencing (WMS) analyses include viruses, archaea, and eukaryotic DNA found in plants, fungi, and animals. The observation of the mycological profile of fungal spores has already proved useful in forensic casework. In one example, it was used to match clothing to a crime scene in a sexual assault case in England[102]. Additionally, transmission of viruses such as human immunodeficiency (HIV)[103] or hepatitis C (HCV)[104] can also be studied in this manner. Animals and their DNA can be found in casework and are often overlooked for their evidential value. Skin microbiome uniqueness has been studied in a number of animals: estrildid finches[105], amphibians[106], bats[107], cetaceans[108,109] and dogs[110]. Since humans were shown to share microbial communities with their dogs[111], the study of animal DNA can offer additional intelligence, especially of those kept by humans as pets due to the close proximity and contact of owner and pet. Linking a person to a crime (via the environment, occupation, or pet ownership) could thereby be achieved based on inter-species microbiome sharing or from trace microbial profiles from other species.

4.2 Post-mortem interval (PMI) and cause of death

The microbial composition of a body goes through specific and predictable changes after death. These changes can, for example, be translated into a microbial reference "clock," with which samples can be compared to determine the time of death[112]. Studies have shown that the use of microbial data can improve prediction accuracy, especially in the early decay stage[113,114]. Other studies have tested targeted sequencing approaches in combination with machine learning methods and found that the human skin microbiome together with soil microbiome markers provided the lowest error rates[115]. Zhang et al.[116] predicted PMI, manner of death, and death event location using three machine learning approaches on targeted sequencing data (16S rRNA) derived from swabs taken during routine death investigation (*n*=188). The predictions showed accuracies ranging from 71% to 88%. Pechal et al. [117] also highlighted the potential of post-mortem microbiome analysis by reporting that it can be used as a predictor for the host's antemortem health condition. There is abundant evidence that the composition of the human microbiome can be linked to disease phenotypes, psychological[118] and physical[119].

During an autopsy, the presence of specific microorganisms may act as bioindicators, providing useful insights concerning the cause of death[18]. For instance, in cases of drowning, bacterioplankton composition (freshwater, e.g.: *Aeromonas* spp., *Pseudomonas* spp. and *Shewanella* spp.; seawater e.g.: *Vibrio* spp., *Photobacterium* spp. and *Listonella* spp.) and salinity levels in the victim's blood act as bioindicators[120,121]; in cases of medical malpractice, microorganisms from the intestinal flora (e.g.: *Enterococcus faecalis*, *E. faecium*, *E. casseliflavus*, *E. thailandicus* and *E. cloacae*) in the peritoneal fluid and cardiac blood indicates possible colon perforation during colonoscopy[122], while HIV[103] and HCV[104] can be linked to cross-contamination or intentional spread of the disease; in cases of child abuse or negligence, abnormal infection agents (e.g.: *Staphylococcus aureus*, *Escherichia coli*, enterovirus, respiratory syncytial virus, rotavirus, adenovirus) in several tissue specimens, coupled with elevated values of C-reactive protein (CRP), may point out to transient inflammation, pain, and bacteraemia followed by toxaemia, potentially resulting in sudden infant death syndrome (SIDS) [123,124].

5. Forensic microbiome intelligence - predicting sample origin using metagenomic data

Man-made, as well as natural environments, differ in their microbial composition. These distinguishable microbial communities can and have been utilised to infer sample origin or location. In one example, the microbiome of nine offices across three cities was studied over the course of one year. The results showed a higher degree of similarity between offices within the same city and an accuracy of 85% when predicting the city of origin of a sample[125]. Another study reported a prediction accuracy of 83% when matching foot skin samples of study participants to their corresponding houses (18 participants in 10 houses; *n*=1625 microbial samples)[16]. It has also been shown that skin microbiomes differ between humans

living in high and low altitudes[126], suggesting an ability to infer geolocation based on altitudinal parameters.

The ubiquity, heterogeneity, transferability and soil composition (geoforensics) makes it especially useful as evidence in criminal investigations[127]. The ability to pinpoint the source of a sample based on the soil microbiome recently been demonstrated[128]. Habtom et al.[15] found that the geographic location was more important than soil type in determining microbial community composition (25m - 1000m; 5 sites, 2-4 soil types, 5 replicates). The potential of soil DNA metabarcoding to provide geolocation evidence has already spurred the interest in forensics. In the past few years, Ribosomal Intergenic Spacer Analysis (RISA) and 16S rRNA gene sequencing were used to discriminate between soil samples originating from a mixture of different substrates[129]. The first results showed that 18S rRNA gene sequencing provides greater discriminatory power over traditional Mid Infrared (MIR) spectroscopy at fine scales[130].

5.1. The MetaSUB International Consortium and Critical Assessment of Massive Data Analysis

The forensic potential in metagenomic profiles was recognised in the early stages of the MetaSUB International Consortium[131]. The consortium aims to build molecular portraits of cities, with one branch focusing on biogeographical analyses and prediction[132]. Samples are collected in public transport systems during an annual sample collection event and subsequently processed for WMS. This growing unique and rich can be exploited for a number of applications[76]. In 2017 MetaSUB started a collaboration with the Critical Assessment of Massive Data Analysis conference (CAMDA)[133], to develop, improve, and benchmark classification tools for metagenomic data in open-science based contest. The potential of this approach was shown at the CAMDA conference in 2017[134–137] where a metagenomic classification challenge was posed based on MetaSUB data. Since then, the challenge has moved beyond constructing city-specific metagenomic profiles towards the applicability of predictive models in the: i) classification of a new sample from a known location (CAMDA 2018); and ii) prediction of sample origin from an unknown location (CAMDA 2019). Summarising, the presented approaches can be divided into three conceptual groups (see Supplement for a detailed description of methods):

Taxonomy/species centric[138–140]. In this “classical” approach, data is analysed either by read-based or assembly-based taxonomy profiling to identify the presence and abundance of a given species. By this, the *structure* of a given metagenomic community is described, and methods from ecology such as “forbidden species” that never appear can be utilised. Here, both the qualitative as well as quantitative factors, can be used to build a classifier. Among multiple machine learning approaches tested, the random forest approach turned out to be most successful. Considering the size and complexity of the provided datasets this choice seems to be optimal. Neuronal networks (or deep networks) usually require a larger size of

training data, which hampers their performance for the moment, but further data collection efforts are ongoing[76].

Functional approach[141,142]. Several studies have shown that species composition is less important than the presence or absence of certain functions required by a microbial community to settle a specific ecological niche[143]. The second strategy takes this finding into account and looks at the functional profile of a sample without prior taxonomic assignment[144]. More specifically, the analysis uses read data to predict genes or domains that encode for a specific function(s). These can sometimes be grouped into pathways for which a pathway coverage score can be calculated[145]. In the classification step, the functional signatures are used as input features. Notably, these have a clear biological meaning and can account for city-specific functional properties. A further development of this approach, especially with the use of advanced feature selection approaches, can also provide a novel perspective for microbiota-based biogeography prediction.

K-mer based approach[146]. In microbiome studies, about half of the data cannot be assigned to any known species[76,147]. The third approach attempts to overcome this limitation by using all the data (from known and unknown species) directly to classify samples after reads are "cleaned" to remove adapters or any sequencing errors. Machine learning is then applied based on the characteristic of the environmental patterns of the DNA fragments (*k*-mers) or reads. Considering the potential size of the feature space, feature selection is crucial. Also, it must be noted that a *k*-mer based approach uses unstructured data, where the biological meaning of a sequence (e.g., synteny) is not well preserved.

Interestingly, all three approaches achieve a high classification accuracy [>90%] on the test data sets. However, results from the 2018 and 2019 challenges highlight the shortcomings in the original prediction of previously unknown samples[138–142,146]. An analysis of the results and feedback from the participants made it clear that for a proper assessment - a bigger data set [better geo-resolution] and richer meta information would be beneficial. This will be further investigated in close collaboration with the CAMDA and MetaSUB communities and then applied in the upcoming CAMDA challenge(s).

6. Challenges and Considerations for the Introduction of Metagenomics for Intelligence

In the previous sections, we argued that metagenomics could be a powerful tool for intelligence purposes. However, before it can be integrated into the forensic toolbox, certain challenges have to be addressed. Sampling, sample treatment, and DNA extraction procedures strongly influence the quantity and quality of recovered metagenomic DNA[148]. Therefore, these processes could have a larger impact on the usefulness and comparability of results over the choice of the bioinfromatics analysis tools[149]. This highlights the need for validated experimental protocols and shared standard operating procedures (SOPs). To

guarantee not only cross-laboratory, but batch comparability, both the sensitivity and reproducibility of the methods must be understood. Protocols should also reflect the necessity for consistent sampling procedures to be performed by forensic specialists at the crime site. Since the microbiome is ubiquitous, it is fundamental to minimise and control for background DNA from unwanted sources like the laboratory environment, or contamination through consumables and reagents ('kitome')[150]. Another consideration being made by laboratories implementing this type of work is the DNA extraction protocols to be used. If the sample collected is for the purpose of microbial and metagenomic DNA recovery, then a specific extraction protocol can be created and applied to maximise the recovery of DNA. It is difficult to extract DNA from some taxa due to their make-up (i.e. gram-stain, membrane composition). If this method is to be applied to a sample post-DNA extraction for human DNA using one of the standard extraction kits (e.g., QIAamp® DNA Investigator Kit, Qiagen) then consideration has to be made that some of the taxa present in the original sample will not be present in the DNA extract.

Another critical point for the practical implementation of metagenomics in forensic genetic investigations is the availability of instruments for metagenomic analyses. Although many forensic laboratories have made a step forward and incorporated medium-throughput DNA sequencers like the MiSeq™ (Illumina, CA) or the Ion S5™ (ThermoFisher Scientific, UK), most are more familiar with capillary electrophoresis (CE) techniques, methods and instruments. Current examples where targeted MPS solutions were employed by forensic DNA laboratories show promising results that lead to the conclusion that the field can benefit from this technology[151–156].

6.1. Sample collection and wet-lab challenges

Determining the best methodology of sampling DNA at a crime scene is not a straightforward choice. Not only is the sampling tool important (swab, tape, spatula, etc.), but it is also important to identify the correct sampling location, number of replicates, suitable environmental negative controls and the optimial storage and transportation conditions (temperature, humidity). Previous experiences in recovering human DNA can be applied to metagenomic sample collection, yet the broad spectrum of metagenomic sampling options will require additional work. For example, although the majority of metagenomic studies report using cotton swabs to collect samples, the type and coating material varies, as does the swabbing solution[148,157]. Further research is needed to determine efficient swab or lift types, coating material or methods for metagenomic sample collection. The chosen method will depend on the surface or material, and the intended analysis approach.

Wet lab techniques, previously used for human DNA analysis, have likewise, been adapted and transformed for use on non-human samples[14]. But since different techniques are to be used for different sample types, the choice of method will be determined by the scientist at the submission stage of a forensic sample. In many cases, human DNA testing is paramount

and thus, universal DNA extraction methods would allow the whole spectrum of forensically relevant markers to be analysed. Depending on the nature of the samples obtained for metagenomic analysis, the matrix can be complex, containing inhibitors or other materials that may influence downstream analysis. Surface samples (including human skin) often result in very low DNA yields, hampering downstream analysis and resulting in a limited taxonomic and functional range[158,159]. Methods that include an amplification step, either a random priming strategy or targeted metagenomics such as a specific marker panel set are better suited and may be the only option for extremely low yield samples, but these too can introduce bias [160].

DNA contamination in laboratory kits and reagents (kitome) is a widespread phenomenon that is well known for its confounding effect on metagenome analysis, especially for low yield samples[161,162]. Several strategies have been proposed to eradicate DNA contamination, including UV irradiation, DNAse treatment or enzymatic digestion[163], but all of these have disadvantages. To evaluate contamination, a comprehensive set of positive and negative controls is of utmost importance. It has also been proposed to establish a local database of common lab contaminants and exclude or downweigh these species in the downstream analysis[164]. Eisenhofer et al. [162] propose a checklist covering each step from experimental design to data analysis to keep contamination under control. This list has been developed with medical applications in mind but could be adapted for metagenomics in forensics as well.

6.2. Marker selection

As metagenomics encompasses the analysis of DNA from a range of taxa, the methodology used will be dependent on the question being asked. Analysis of the eukaryotic content of a sample can be supported and complemented by a bacterial profile. Depending on the availability of genetic material, a targeted approach can limit the information recovered from the sample and which marker sets to use will strongly depend on the insight the taxonomic group can provide for the specific case.

In animal identification, the most frequently used molecular markers are cytochrome b (Cyt b) and cytochrome c oxidase I (COI) gene[165]. For green plants, the genes ribulose-bisphosphate carboxylase large subunit (rbcL), chloroplast maturase K (matK), the intergenic spacer regions trnH-psbA and internal transcribed spacer (ITS) serve as markers. But even within the same domain of life, there is much divergence with respect to the markers to be used[166]. Several STR panels have been formulated for animal species Identification of domestic animals (e.g., cats and dogs), livestock (e.g., cattle and pigs), wild animals and endangered species. In the microbial field, the use of the 16S rRNA gene is well established for the identification of bacteria. The 18S rRNA gene and the ITS region are studied for fungi. Occasionally, it is necessary to resort to other genes to increase the resolution capacity[167,168]. Although targeted marker amplification can overcome the obstacle of low DNA yield, targeted metagenomics has its own set of challenges. Taxonomic biases associated with the choice of primers[169], or the amplified region[170], are well known. Furthermore, the amplification process

can result in chimeric sequences[171] that contain 16S rRNA pieces from more than one organism. Recent advances in MPS have enabled scientists to sequence longer genomic regions. The application of this technique would negate the aforementioned challenges but would incur a higher cost and require higher yield and quality of DNA.

6.3. Computational and Bioinformatics challenges

It is expected that metagenomic sequencing data for forensic application will show varying levels of complexity, based on the origin of the sample, and characteristics caused by technical biases, based on the sample type and preparation methods. Furthermore, the large number of parameters describing the properties of the samples results in requirements that exceed the possibilities of conventional statistical and bioinformatic methods. As discussed above, machine-learning algorithms are, in principle, able to tackle these issues. Ideally, this should be done by the application of unsupervised or reinforcement learning approaches that involve the requirement of a significant amount of input data and powerful hardware. The high number of partly unknown influencing factors (e.g., unknown species, interactions in microbial communities, inaccuracies in sample preparation) will likely limit the possibilities of accurate predictions based on models that are trained solely from existing sample data. Therefore, and for the time being, the classification of taxonomic features and/or functional traits in metagenomes is still essential.

Both WMS and targeted sequencing have advantages and disadvantages in their applicability for intelligence purposes. Key requirements like fast processing time, ease of handling, and reproducibility have to be complemented with an analysis that correctly answers the investigative question. Targeted metagenomics, especially 16S rRNA, has a long history and, therefore, well-established protocols coupled with known biases. Extensive databases with reference data (e.g., Greengenes, SILVA, RDP) exist, although not all analysis tools use the most recent versions, and not all databases are maintained equally well[172]. Bioinformatics analysis is well established with QIIME[173] and Mothur[174] being widely accepted and commonly applied tools. For all of the above reasons, as well as the larger reference databases, targeted metagenomics is currently more robust than WMS[175]. It can also potentially be more sensitive, but only if a set of several target molecules is used to overcome the inherent biases of the method, and despite still relying on some form of sequencing, it cannot reach the breadth of a WMS approach[96]. Especially in cases where the available amount of DNA is limited, collective analysis of bacterial and eukaryotic (e.g., human) DNA enhances the information extracted from the sample. WMS also provides multi-layered data that can offer a long-term, incomparable advantage when applying machine learning, especially in *k-mer* and functional-based approaches.

For WMS, several databases exist that contain genomic reference materials, which vary in quality and completeness, with large public databases being more comprehensive but also more prone to assembly errors and misclassification. Well-curated databases, on the other

hand, tend to be small and specialised towards certain organisms or fields[150]. In the past years, efforts have been made to increase the quality of data in public databases, but misclassifications are still not uncommon[176–179], and recently it has been shown that many species in metagenomic samples will result in a taxonomic best hit assignment, even though the actual species are novel and not yet present in the database[137,180,181]. Furthermore, new microbial genomes are deposited in public databases at a high rate, influencing the results by providing new and differing taxonomic assignments[182]. Regardless of the chosen analysis method, the database version and how often it is updated can have a paramount effect on the results.

In intelligence, a precise taxonomic assignment may not even be needed to provide useful information. More general microbial sequence profiles can be used for the geo-localisation of a sample or in predictive modelling. To extract a metagenomic profile from a sample and compare it to other profiles, it is more important that a method would consistently give robust results than for it to be highly sensitive. Marker gene panels offer faster processing times and are easier to handle, but the selection of species, genes or regions that should be used for optimal discrimination is still under active investigation[97]. Although metagenomics analysis is common in a variety of fields now, large datasets with a comprehensive geographical distribution, controlled protocols and enough replicates are still rare.

6.4. Metagenome Analysis Benchmarking and Standards

In recent years, the need for benchmarking and standardisation in the field has been recognised. The human microbiome[183] and earth microbiome projects[184] were crucial drivers in the standardisation of experimental design and protocols, especially for targeted microbiome analysis. In other areas of metagenomics analysis, international standards are also being developed, and the quality of results is expected to improve considerably[185]. The Critical Assessment of the Metagenome Interpretation (CAMI) initiative was founded to provide an independent evaluation of metagenomics software-based on comparable measures and simulated datasets[186]. These simulated datasets and real case study benchmark datasets are now available[187] to evaluate the performance and accuracy of new software and pipelines. Since the choice of software combinations and parameters is often dependent on the sample features the analysis can also benefit from the use of in-silico standards[137], simulated data that are spiked into the raw data to measure how well the selected pipeline is able to recover the expected taxonomic or functional assignment. This spike-ins can also be used to detect epigenetic states of the organisms, adding another layer of specificity and forensic application.

6.5. Ethical considerations

The importance of ethical, legal and social implications (ELSI) when analysing human DNA has been the subject of many publications and debates since scientists initiated the Human Genome Project (HGP) in 1990. The National Institutes of Health (NIH) in the US established

and funded the ELSI Research Program as an integral part of the HGP (National Human Genome Research Institute - NHGRI 2019), which was essential to its success. Traditionally, the microbial communities have been viewed as environmental factors[188], but recent views call to consider humans as 'superorganisms' incorporating multiple symbiotic cell species[189], or what is often called the holobiont. Such a view raise philosophical questions as to how integral the microbiome is to our conception as a human being[190].

When the NIH Human Microbiome Project (HMP) was launched in 2007, one of its aims was to characterise the microbiomes of healthy human subjects at five major body sites to answer the questions of whether humans have a core microbiome, whether it remains stable throughout life, and whether there are predictive similarities within communities and environments[191]. Like the HGP, human microbiome research raises important ethical considerations in this relatively uncharted ethical landscape[192].

Beyond the important issues around health, forensic scientists are also interested in the information gained from knowledge of the microbiome of our surroundings. Many commentators raise issues of privacy associated with the potentially unique personal microbial fingerprint or 'cloud'[193] that may, in addition, reveal our past exposures or movements that can infer group affiliation, ancestral origins, and socio-economic status. The forensic use of this information[194] may be harmful for the individuals involved if, for example, the knowledge is linked to practices such as deportation or incarceration. The forensic potential of the microbiome is considerable and proactive and collaborative consideration of ELSI in forensic microbiome research is vital if we are to be able to make use of this private intelligence for community good and avoid personal harm.

7. Conclusions

During the last decades, the advances in massively parallel sequencing (MPS) and genomics technologies, fuelled by the improvements in the field of machine learning, have resulted in copious quantities of high-quality sequences for the rapid analysis of extensive microbial communities from environmental samples[195,196]. Explorations into whole-genome and targeted metagenomics by forensic scientists have shown the benefits and potentials the method harbours. As discussed above, however, there are still a number of challenges associated with the use of metagenomics as a forensic investigative tool that needs to be addressed[18].

The first challenge concerns the need for standard operating procedures (SOPs) along with ensuring forensically robust documentation (including chain-of-custody) for specimen collection, handling and transportation of the samples. Some recommendations can be found in the NATO "Handbook for Sampling and Identification of Biological and Chemical Agents (SIBCA)"[197]. While a blood spot or a cigarette butt found at a crime scene are easily identified as evidence, microorganisms are difficult to detect and cannot be observed with the naked eye. As such, all personnel involved in a criminal case from law enforcement agents to judges

and juries should receive further training to deal with this type of evidence - from sample collection through to analysis in the lab and finally, to the presentation of such evidence in written reports and in court.

The second challenge is the rigorous validation that each analytical method and step of the process (from sampling through to analysis) needs to go through for quality assurance. The key components required are: sensitivity, specificity, precision, accuracy, reproducibility, repeatability, the limits of detection, reportable range, false positive and negative ranges and robustness[194,196]. It should be emphasised that some of the protocols applied in non-human forensics have been adapted from techniques used for human forensics. Nevertheless, in the case of microbial communities, an additional level of complexity needs to be addressed. Especially, the effects of the choice in analysis methods on the statistical result, as well as the confidence/uncertainty of an analysis-based interpretation, need to be explored to decide on a reliable microbial forensic analysis protocol. There are also biological factors to account for when adapting protocols, as, contrary to humans, microorganisms are mostly haploid, reproduce asexually (clonal genetically indistinguishable individuals) and frequently are involved in horizontal gene transfer events[198].

The third and last challenge relates to the creation of reliable and comprehensive databases for the accurate interpretation of the obtained results[199]. More work needs to be carried out to produce and curate reliable reference databases, genomic reference sequences, metadata and high coverage maps of the microbiome diversity across human populations[200] as well as the environment. The temporal stability of the microbiome is an aspect that is essential when examining the databases. Some research has been carried out on the stability of the human microbiome – with conflicting results –, especially in relation to environmental samples. Indeed, temporal population dynamics within microbial communities needs to be addressed to discern how long a sample can be reliably detected after it was deposited at a crime scene. This parameter is likely to be dependent on the environment (temperature, humidity, traffic), the species, and also the local microbial community. Still, current applications of MPS in forensics show that targeted microbiome analysis, either based on 16S rRNA analysis or through new targeted DNA panels, would surely be welcomed by the forensic community. Using a targeted approach should also reduce the bioinformatics burden for forensic DNA laboratories by offering validated and easy to use solutions.

Up to now, the major contributions of metagenomics in the field of Microbial Forensics have been associated with both epidemiological studies and the investigation of bioterrorism attacks. In both scenarios, a periodic monitoring system would enable the trace and detection of outbreaks of both natural (accidental) or intentional. Of particular interest would be the source of the outbreak, the transmission route, and if possible, the identification of the individual responsible. Analysis of the molecular variations between closely related strains is essential in this scenario. As a result of this strategy, the information can contribute to the

construction of effective response plans and would also act as a deterrent for terrorists. Metagenomics circumvents two of the major limitations associated with the "classical" studies of microbial communities: it is culture-independent; therefore information on the true diversity of a given ecosystem can be obtained (85-99% of microorganisms in nature that have not yet been cultivated)[201] and it provides insights to the complex metabolic pathways associated with these microorganisms (e.g., antimicrobial resistance or virulence factors).

Apart from this more classical application, Microbial Forensics has much greater development potential. The usage of information within metagenomics data combined with the predictive power of artificial intelligence could be a game-changer for the application of metagenomics for biogeography. The CAMDA participants and the International MetaSUB Consortium have already demonstrated that the structure of bacterial communities is geographically informative. Further work will focus on the integration of knowledge inferred from multiple sources of complementary data (poor but dense 16S-based data, rich but sparse WGM-based data, global climate meta-data). This would allow forensic science to match metagenomic-based knowledge with environmental metadata, as the composition of every microbial community is linked to the conditions of a specific ecological niche. Based on the progress so far, adding the metagenomic factor to the human genomic one will improve the accuracy and resolution of biogeographical applications, and thus will have major implications for intelligence and forensics. Technological progress in applications for the whole genome and whole metagenome sequencing along with advances in deep learning approaches, indicate a bright future to forensic DNA intelligence.

Targeted (amplicon) analysis: In this well-established method a specific part of the DNA (gene) is amplified and mapped against a reference database. The most common target is part of the small ribosomal subunit, 16S rRNA for bacteria[204]*, the ITS region for fungi*[205]*, and the 18S rRNA for other eukaryotes. The ribosomal region has the advantage that it contains highly conserved stretches, serving as primer regions, with several variable regions that allow discrimination between microbial groups. There are certain known shortcomings and challenges with targeted analysis of the microbiome*[206]*. Apart from the biases associated with the amplification process, the 16S rRNA does not always allow a species or even strain resolution*[207]*, which might be necessary to distinguish between individuals or locations. Other markers have been employed differentiate between species and strains, not discernible through 16S rRNA analysis, most prominently, multilocus sequence typing (MLST)*[208]*. The advantages of this method are its cost-effectiveness and low computational load.*
Advantages*: works for low DNA yield, does not need high sequencing depth, kingdom specific, smaller curated reference databases*
Disadvantages*: amplification bias, poor species/strain resolution in some groups*

Whole-metagenome analysis: In whole genome metagenomics, DNA is extracted from a sample without selection of a specific region and without amplification prior to sequencing. In contrast to targeted metagenomics, taxonomic as well as functional information can be derived directly from the data. It also carries the potential of de-novo assembly of microbial genomes, thereby investigating yet unknown species. Prokaryotic and eukaryotic DNA can be analysed in one process.
Advantages*: higher resolution, information on biochemical pathways without the need for taxonomic assignment, works also for unknown species*
Disadvantages*: higher demand on resources, especially for genome assembly; more complicated sample processing*

References

1. Parsons, T. J., Huel, R. M. L., Bajunović, Z. & Rizvić, A. Large scale DNA identification: The ICMP experience. *Forensic Science International: Genetics* **38,** 236–244 (2019).
2. Watherston, J., McNevin, D., Gahan, M. E., Bruce, D. & Ward, J. Current and emerging tools for the recovery of genetic information from post mortem samples: New directions for disaster victim identification. *Forensic Science International: Genetics* **37,** 270–282 (2018).
3. Linacre, A. M. T. in *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Elsevier, 2018). doi:10.1016/B978-0-12-409547-2.14203-9
4. Prinz, M. *et al.* DNA Commission of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Science International: Genetics* **1,** 3–12 (2007).
5. Parys-Proszek, A., Wróbel, M., Marcińska, M. & Kupiec, T. Dual amplification strategy for improved efficiency of forensic DNA analysis using NGM Detect™ NGM™ or Globalfiler™ kits. *Forensic Sci. Int. Genet.* **35,** 46–49 (2018).
6. Elwick, K., Mayes, C. & Hughes-Stamm, S. Comparative sensitivity and inhibitor tolerance of GlobalFiler® PCR Amplification and Investigator® 24plex QS kits for

challenging samples. *Leg. Med.* **32,** 31–36 (2018).
7. Jobim, M. R., Gamio, F., Ewald, G., Jobim, M. & Jobim, L. F. Human identification using DNA purified from residues in used toothbrushes. *Int. Congr. Ser.* **1261,** 491–493 (2004).
8. Tanaka, M. *et al.* Usefulness of a Toothbrush as a Source of Evidential DNA for Typing. *J. Forensic Sci.* **45,** 14746J (2000).
9. Dutch police launch biggest-ever DNA hunt for boy's killer. *Phys.Org.* (2018). Available at: https://phys.org/news/2018-01-dutch-police-biggest-ever-dna-boy.html.
10. Kayser, M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci. Int. Genet.* **18,** 33–48 (2015).
11. Børsting, C. & Morling, N. Next generation sequencing and its applications in forensic genetics. *Forensic Sci. Int. Genet.* **18,** 78–89 (2015).
12. Devesse, L. *et al.* Concordance of the ForenSeq™ system and characterisation of sequence-specific autosomal STR alleles across two major population groups. *Forensic Sci. Int. Genet.* **34,** 57–61 (2018).
13. Oliveira M, Arenas M, Pinto N, A. A. in *Genética Forense: Del laboratorio a los Tribunales.* 291–318 (2019).
14. Arenas, M. *et al.* Forensic genetics and genomics: Much more than just a human affair. *PLoS Genetics* **13,** (2017).
15. Habtom, H. *et al.* Applying microbial biogeography in soil forensics. *Forensic Sci. Int. Genet.* **38,** 195–203 (2019).
16. Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science (80-.).* **345,** 1048–1052 (2014).
17. *The new science of metagenomics: Revealing the secrets of our microbial planet*. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet* (National Academies Press, 2007). doi:10.17226/11902
18. Oliveira, M. & Amorim, A. *Microbial forensics: new breakthroughs and future prospects*. *Applied Microbiology and Biotechnology* **102,** 10377–10391 (Springer Berlin Heidelberg, 2018).
19. de los Campos, G., Vazquez, A. I., Hsu, S. & Lello, L. Complex-Trait Prediction in the Era of Big Data. *Trends in Genetics* **34,** 746–754 (2018).
20. Sturm, R. A. & Larsson, M. Genetics of human iris colour and patterns. *Pigment Cell and Melanoma Research* **22,** 544–562 (2009).
21. Pośpiech, E. *et al.* The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci. Int. Genet.* **11,** 64–72 (2014).
22. Eiberg, H. *et al.* Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum. Genet.* **123,** 177–187 (2008).
23. Branicki, W. *et al.* Model-based prediction of human hair color using DNA variants. *Hum. Genet.* **129,** 443–454 (2011).
24. Breslin, K. *et al.* HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. *Forensic Sci. Int. Genet.* **43,** 102152 (2019).
25. Hysi, P. G. *et al.* Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat. Genet.* **50,** 652–656 (2018).

26. Morgan, M. D. *et al.* Genome-wide study of hair colour in UK Biobank explains most of the SNP heritability. *Nat. Commun.* **9,** (2018).
27. Pośpiech, E. *et al.* Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA. *Forensic Sci. Int. Genet.* **37,** 241–251 (2018).
28. Marcińska, M. *et al.* Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness. *PLoS One* **10,** (2015).
29. Pirastu, N. *et al.* GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. *Nat. Commun.* **8,** (2017).
30. Lello, L. *et al.* Accurate genomic prediction of human height. *Genetics* **210,** 477–497 (2018).
31. Allen, H. L. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).
32. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
33. Liu, F. *et al.* Common DNA variants predict tall stature in Europeans. *Hum. Genet.* **133,** 587–597 (2014).
34. Liu, F. *et al.* Update on the predictability of tall stature from DNA markers in Europeans. *Forensic Sci. Int. Genet.* **42,** 8–13 (2019).
35. Behrman, B. W. & Davey, S. L. Eyewitness identification in actual criminal cases: An archival analysis. *Law Hum. Behav.* **25,** 475–491 (2001).
36. Claes, P. *et al.* Modeling 3D Facial Shape from DNA. *PLoS Genet.* **10,** (2014).
37. Lippert, C. *et al.* Identification of individuals by trait prediction using whole-genome sequencing data. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 10166–10171 (2017).
38. Erlich, Y. Major flaws in 'identification of individuals by trait prediction using whole genome'. *bioRxiv* 1–5 (2017). doi:dx.doi.org/10.1101/185330
39. Sero, D. *et al.* Facial recognition from DNA using face-to-DNA classifiers. *Nat. Commun.* **10,** (2019).
40. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14,** (2013).
41. Zbieć-Piekarska, R. *et al.* Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci. Int. Genet.* **17,** 173–179 (2015).
42. Yi, S. H., Jia, Y. S., Mei, K., Yang, R. Z. & Huang, D. X. Age-related DNA methylation changes for forensic age-prediction. *Int. J. Legal Med.* **129,** 237–244 (2015).
43. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49,** 359–367 (2013).
44. Naue, J. *et al.* Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. *Forensic Sci. Int. Genet.* **31,** 19–28 (2017).
45. Lee, H. Y. *et al.* Epigenetic age signatures in the forensically relevant body fluid of semen: A preliminary study. *Forensic Sci. Int. Genet.* **19,** 28–34 (2015).
46. Council for the Advancement of Forensic Genealogy (CAFG). Available at: http://www.forensicgenealogists.org/. (Accessed: 25th September 2019)
47. https://heinonline.org. Available at: https://heinonline.org/HOL/Page?handle=hein.journals/miprobeslnj33&id=61&collection=journals&index=. (Accessed: 25th September 2019)
48. Associated Press. DNA clears accused Golden State Killer Joseph DeAngelo of 1975

murder. Available at: https://www.nbcnews.com/news/us-news/dna-clears-accused-golden-state-killer-joseph-deangelo-1975-murder-n956566. (Accessed: 25th September 2019)

49. Arango, T., Goldman, A. & Fuller, T. To Catch a Killer: A Fake Profile on a DNA Site and a Pristine Sample - The New York Times. *The New York Times* (2018). Available at: https://www.nytimes.com/2018/04/27/us/golden-state-killer-case-joseph-deangelo.html. (Accessed: 25th September 2019)
50. Karantzali, E., Rosmaraki, P., Kotsakis, A., Le Roux-Le Pajolec, M. G. & Fitsialos, G. The effect of FBI CODIS Core STR Loci expansion on familial DNA database searching. *Forensic Sci. Int. Genet.* **43,** 102129 (2019).
51. Scudder, N., Robertson, J., Kelty, S. F., Walsh, S. J. & McNevin, D. Crowdsourced and crowdfunded: the future of forensic DNA? *Aust. J. Forensic Sci.* 1–7 (2018). doi:10.1080/00450618.2018.1486456
52. Shapiro, E. How a DNA database's new policy is changing police access and could hinder solving cold cases. *ABCNews* (2019).
53. Phillips, C. The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Sci. Int. Genet.* **36,** 186–188 (2018).
54. Syndercombe Court, D. Forensic genealogy: Some serious concerns. *Forensic Sci. Int. Genet.* **36,** 203–204 (2018).
55. Berkman, B. E., Miller, W. K. & Grady, C. Is it ethical to use genealogy data to solve crimes? *Annals of Internal Medicine* **169,** 333–334 (2018).
56. Zhang, S. A Woman's Ancestry DNA Test Revealed a Medical Secret. (2019). Available at: https://www.theatlantic.com/science/archive/2019/09/woman-cord-blood-donor-dna-test/597928/. (Accessed: 25th September 2019)
57. Phillips, C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int. Genet.* **18,** 49–65 (2015).
58. Moriot, A., Santos, C., Freire-Aradas, A., Phillips, C. & Hall, D. Inferring biogeographic ancestry with compound markers of slow and fast evolving polymorphisms. *Eur. J. Hum. Genet.* **26,** 1697–1707 (2018).
59. Huckins, L. M. *et al.* Using ancestry-informative markers to identify fine structure across 15 populations of European origin. *Eur. J. Hum. Genet.* **22,** 1190–1200 (2014).
60. Reiner, A. P. *et al.* Population structure, admixture, and aging-related phenotypes in African American adults: The cardiovascular health study. *Am. J. Hum. Genet.* **76,** 463–477 (2005).
61. Xu, S. & Jin, L. A Genome-wide Analysis of Admixture in Uyghurs and a High-Density Admixture Map for Disease-Gene Discovery. *Am. J. Hum. Genet.* **83,** 322–336 (2008).
62. Consortium, T. 1000 G. P. *et al.* A global reference for human genetic variation. *Nature* **526,** 68 (2015).
63. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164,** 1567–1587 (2003).
64. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
65. Binary AIM classification of individuals. (2019). Available at: http://mathgene.usc.es/snipper/.
66. Rajeevan, H., Soundararajan, U., Pakstis, A. J. & Kidd, K. K. Introducing the Forensic Research/Reference on Genetics knowledge base, FROG-kb. *Investigative Genetics* **3,**

(2012).
67. Illumina. ForenSeq™ DNA Signature Prep Kit Data Sheet. 36 (2015). doi:# TG-450-9001DOC Material # 20000923 Document # 15049528 v01
68. Biosystems, T. A. Precision ID Ancestry Panel. *https://www.thermofisher.com/order/catalog/product/A25642* (2019).
69. Esposito, U., Das, R., Syed, S., Pirooznia, M. & Elhaik, E. Ancient ancestry informative markers for identifying fine-scale ancient population structure in eurasians. *Genes (Basel).* **9,** (2018).
70. Elhaik, E. *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat. Commun.* **5,** (2014).
71. Margalit, Y., Baran, Y. & Halperin, E. Multiple-ancestor localization for recently admixed individuals. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9289,** 121–135 (2015).
72. Marshall, S., Das, R., Pirooznia, M. & Elhaik, E. Reconstructing Druze population history. *Sci. Rep.* **6,** (2016).
73. HomeDNA GPS Origins Ancestry Tests. (2016). Available at: https://homedna.com/product/gps-origins.
74. Yang, W. Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet.* **44,** 725–731 (2012).
75. Das, R., Wexler, P., Pirooznia, M. & Elhaik, E. Localizing Ashkenazic Jews to primeval villages in the ancient Iranian lands of Ashkenaz. *Genome Biol. Evol.* **8,** 1132–1149 (2016).
76. Danko, D. C. *et al.* Global Genetic Cartography of Urban Metagenomes and Anti-Microbial Resistance. *bioRxiv* 724526 (2019). doi:10.1101/724526
77. Elhaik, E. *et al.* The Diversity of REcent and Ancient huMan (DREAM): A New Microarray for Genetic Anthropology and Genealogy, Forensics, and Personalized Medicine. *Genome Biol. Evol.* **9,** 3225–3237 (2017).
78. Budowle, B. *et al.* Building microbial forensics as a response to bioterrorism. *Science* **301,** 1852–1853 (2003).
79. Van Ert, M. N. *et al.* Strain-specific single-nucleotide polymorphism assays for the Bacillus anthracis Ames strain. *J. Clin. Microbiol.* **45,** 47–53 (2007).
80. Williams, D. W. & Gibson, G. Individualization of pubic hair bacterial communities and the effects of storage time and temperature. *Forensic Sci. Int. Genet.* **26,** 12–20 (2017).
81. Nakanishi, H. *et al.* A novel method for the identification of saliva by detecting oral streptococci using PCR. *Forensic Sci. Int.* **183,** 20–23 (2009).
82. Hanssen, E. N., Avershina, E., Rudi, K., Gill, P. & Snipen, L. Body fluid prediction from microbial patterns for forensic application. *Forensic Sci. Int. Genet.* **30,** 10–17 (2017).
83. Burcham, Z. M. *et al.* Bacterial community succession, transmigration, and differential gene transcription in a controlled vertebrate decomposition model. *Front. Microbiol.* **10,** (2019).
84. Javan, G. T. *et al.* Human Thanatomicrobiome Succession and Time since Death. *Sci. Rep.* **6,** (2016).
85. Metcalf, J. L. *et al.* A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system. *Elife* **2013,** (2013).
86. Aoyagi, M. *et al.* A novel method for the diagnosis of drowning by detection of

Aeromonas sobria with PCR method. *Leg. Med.* **11,** 257–259 (2009).
87. Rutty, G. N. *et al.* Detection of bacterioplankton using PCR probes as a diagnostic indicator for drowning; the Leicester experience. *Leg. Med.* **17,** 401–408 (2015).
88. Uchiyama, T. *et al.* A new molecular approach to help conclude drowning as a cause of death: Simultaneous detection of eight bacterioplankton species using real-time PCR assays with TaqMan probes. *Forensic Sci. Int.* **222,** 11–26 (2012).
89. Watanabe, H. *et al.* Minor taxa in human skin microbiome contribute to the personal identification. *PLoS One* **13,** (2018).
90. Wilkins, D., Leung, M. H. Y. & Lee, P. K. H. Microbiota fingerprints lose individually identifying features over time. *Microbiome* **5,** (2017).
91. Lax, S. *et al.* Bacterial colonization and succession in a newly opened hospital. *Sci. Transl. Med.* **9,** (2017).
92. Fierer, N. *et al.* Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6477–6481 (2010).
93. Goga, H. Comparison of bacterial DNA profiles of footwear insoles and soles of feet for the forensic discrimination of footwear owners. *Int. J. Legal Med.* **126,** 815–823 (2012).
94. Gilbert, P. J. The Human Microbiome – A New Potential Fingerprint in Forensic Evidence ? Thought Leaders. 1–8 (2018). Available at: https://www.news-medical.net/news/20180329/The-Human-Microbiome-e28093-A-New-Potential-Fingerprint-in-Forensic-Evidence.aspx.
95. Kodama, W. A. *et al.* Trace Evidence Potential in Postmortem Skin Microbiomes: From Death Scene to Morgue. *J. Forensic Sci.* **64,** 791–798 (2019).
96. Schmedes, S. E., Woerner, A. E. & Budowle, B. Forensic human identification using skin microbiomes. *Appl. Environ. Microbiol.* **83,** (2017).
97. Schmedes, S. E. *et al.* Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification. *Forensic Sci. Int. Genet.* **32,** 50–61 (2018).
98. Quaak, F. C. A., de Graaf, M. L. M., Weterings, R. & Kuiper, I. Microbial population analysis improves the evidential value of faecal traces in forensic investigations. *Int. J. Legal Med.* **131,** 45–51 (2017).
99. Quaak, F. C. A., van Duijn, T., Hoogenboom, J., Kloosterman, A. D. & Kuiper, I. Human-associated microbial populations as evidence in forensic casework. *Forensic Sci. Int. Genet.* **36,** 176–185 (2018).
100. Meadow, J. F. *et al.* Humans differ in their personal microbial cloud. *PeerJ* **2015,** (2015).
101. Luongo, J. C. *et al.* Microbial analyses of airborne dust collected from dormitory rooms predict the sex of occupants. *Indoor Air* **27,** 338–344 (2017).
102. Wiltshire, P. E. J., Hawksworth, D. L., Webb, J. A. & Edwards, K. J. Palynology and mycology provide separate classes of probative evidence from the same forensic samples: A rape case from southern England. *Forensic Sci. Int.* **244,** 186–195 (2014).
103. Ou, C. Y. *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science (80-.).* **256,** 1165–1171 (1992).
104. González-Candelas, F., Bracho, M. A. & Moya, A. Molecular Epidemiology and Forensic Genetics: Application to a Hepatitis C Virus Transmission Event at a Hemodialysis Unit. *J. Infect. Dis.* **187,** 352–358 (2003).
105. Engel, K. *et al.* Individual- and Species-Specific Skin Microbiomes in Three Different

Estrildid Finch Species Revealed by 16S Amplicon Sequencing. *Microb. Ecol.* **76,** 518–529 (2018).

106. McKenzie, V. J., Bowers, R. M., Fierer, N., Knight, R. & Lauber, C. L. Co-habiting amphibian species harbor unique skin bacterial communities in wild populations. *ISME J.* **6,** 588–596 (2012).
107. Avena, C. V. *et al.* Deconstructing the bat skin microbiome: Influences of the host and the environment. *Front. Microbiol.* **7,** (2016).
108. Erwin, P. M. *et al.* High diversity and unique composition of gut microbiomes in pygmy (Kogia breviceps) and dwarf (K. sima) sperm whales. *Sci. Rep.* **7,** (2017).
109. Russo, C. D. *et al.* Bacterial Species Identified on the Skin of Bottlenose Dolphins Off Southern California via Next Generation Sequencing Techniques. *Microb. Ecol.* **75,** 303–309 (2018).
110. Torres, S. *et al.* Diverse bacterial communities exist on canine skin and are impacted by cohabitation and time. *PeerJ* **2017,** (2017).
111. Song, S. J. *et al.* Cohabiting family members share microbiota with one another and with their dogs. *Elife* **2013,** (2013).
112. Metcalf, J. L. Estimating the postmortem interval using microbes: Knowledge gaps and a path to technology adoption. *Forensic Science International: Genetics* **38,** 211–218 (2019).
113. Metcalf, J. L., Carter, D. O. & Knight, R. Microbiology of death. *Current Biology* **26,** R561–R563 (2016).
114. Johnson, H. R. *et al.* A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval. *PLoS One* **11,** (2016).
115. Belk, A. *et al.* Microbiome data accurately predicts the postmortem interval using random forest regression models. *Genes (Basel).* **9,** (2018).
116. Zhang, Y. *et al.* Machine learning performance in a microbial molecular autopsy context: A cross-sectional postmortem human population study. *PLoS One* **14,** (2019).
117. Pechal, J. L., Schmidt, C. J., Jordan, H. R. & Benbow, M. E. A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. *Sci. Rep.* **8,** (2018).
118. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* **4,** 623–632 (2019).
119. Jackson, M. A. *et al.* Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat. Commun.* **9,** (2018).
120. Kakizaki, E. *et al.* In vitro study of possible microbial indicators for drowning: Salinity and types of bacterioplankton proliferating in blood. *Forensic Sci. Int.* **204,** 80–87 (2011).
121. Kakizaki, E. *et al.* Detection of marine and freshwater bacterioplankton in immersed victims: Post-mortem bacterial invasion does not readily occur. *Forensic Sci. Int.* **211,** 9–18 (2011).
122. Ventura Spagnolo, E. *et al.* Forensic microbiology applications: A systematic review. *Legal Medicine* **36,** 73–80 (2019).
123. Rambaud, C. *et al.* Microbiology in sudden infant death syndrome (SIDS) and other childhood deaths. *FEMS Immunol. Med. Microbiol.* **25,** 59–66 (1999).
124. Weber, M. *et al.* Infection and sudden unexpected death in infancy: a systematic retrospective case review. *Lancet* **371,** 1848–1853 (2008).
125. Chase, J. *et al.* Geography and Location Are the Primary Drivers of Office Microbiome

Composition. *mSystems* **1,** (2016).
126. Zeng, B. *et al.* High-altitude living shapes the skin microbiome in humans and pigs. *Front. Microbiol.* **8,** 1929 (2017).
127. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science (80-.).* **359,** 320–325 (2018).
128. Macdonald, C. A., Ang, R., Cordiner, S. J. & Horswell, J. Discrimination of Soils at Regional and Local Levels Using Bacterial and Fungal T-RFLP Profiling. *J. Forensic Sci.* **56,** 61–69 (2011).
129. Demanèche, S., Schauser, L., Dawson, L., Franqueville, L. & Simonet, P. Microbial soil community analyses for forensic science: Application to a blind test. *Forensic Sci. Int.* **270,** 153–158 (2017).
130. Young, J. M., Weyrich, L. S., Breen, J., Macdonald, L. M. & Cooper, A. Predicting the origin of soil evidence: High throughput eukaryote sequencing and MIR spectroscopy applied to a crime scene scenario. *Forensic Sci. Int.* **251,** 22–31 (2015).
131. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* **4,** 24 (2016).
132. Shamarina, D., Stoyantcheva, I., Mason, C. E., Bibby, K. & Elhaik, E. Communicating the promise, risks, and ethics of large-scale, open space microbiome and metagenome research. *Microbiome* **5,** (2017).
133. CAMDA. Available at: http://camda.info.
134. Qiao, Y. *et al.* MetaBinG2: A fast and accurate metagenomic sequence classification system for samples with many unknown organisms. *Biol. Direct* **13,** (2018).
135. Zolfo, M. *et al.* Profiling microbial strains in urban environments using metagenomic sequencing data. *Biol. Direct* **13,** (2018).
136. Walker, A. R., Grimes, T. L., Datta, S. & Datta, S. Unraveling bacterial fingerprints of city subways from microbiome 16S gene profiles. *Biol. Direct* **13,** 1–16 (2018).
137. Gerner, S. M., Rattei, T. & Graf, A. B. Assessment of urban microbiome assemblies with the help of targeted in silico gold standards. *Biol. Direct* **13,** (2018).
138. Harris, Z. N., Dhungel, E., Mosior, M. & Ahn, T. H. Massive metagenomic data analysis using abundance-based machine learning. *Biol. Direct* **14,** 12 (2019).
139. Walker, A. R. & Datta, S. Identification of city specific important bacterial signature for the MetaSUB CAMDA challenge microbiome data. *Biol. Direct* **14,** 11 (2019).
140. Ryan, F. J. Application of machine learning techniques for creating urban microbial fingerprints. *Biol. Direct* **14,** (2019).
141. Casimiro-Soriguer, C. S., Loucera, C., Perez Florido, J., López-López, D. & Dopazo, J. Antibiotic resistance and metabolic profiles as functional biomarkers that accurately predict the geographic origin of city metagenomics samples. *Biol. Direct* **14,** (2019).
142. Zhu, C. *et al.* Fingerprinting cities: differentiating subway microbiome functionality. *Biol. Direct* **14,** 19 (2019).
143. Heintz-Buschart, A. & Wilmes, P. Human Gut Microbiome: Function Matters. *Trends in Microbiology* **26,** 563–574 (2018).
144. Alves, L. D. F. *et al.* Metagenomic Approaches for Understanding New Concepts in Microbial Science. *International Journal of Genomics* **2018,** (2018).
145. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15,** 962–968 (2018).
146. Kawulok, J., Kawulok, M. & Deorowicz, S. Environmental metagenome classification for constructing a microbiome fingerprint. *Biol. Direct* **14,** (2019).

147. Afshinnekoo, E. *et al.* Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1,** 72–87 (2015).
148. You, H. S. *et al.* Influence of swabbing solution and swab type on DNA recovery from rigid environmental surfaces. *J. Microbiol. Methods* **161,** 12–17 (2019).
149. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35,** 833–844 (2017).
150. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. *Nature Reviews Genetics* **20,** 341–355 (2019).
151. Van Neste, C., Vandewoestyne, M., Van Criekinge, W., Deforce, D. & Van Nieuwerburgh, F. My-Forensic-Loci-queries (MyFLq) framework for analysis of forensic STR data generated by massive parallel sequencing. *Forensic Sci. Int. Genet.* **9,** 1–8 (2014).
152. Zhao, X. *et al.* Multiplex Y-STRs analysis using the ion torrent personal genome machine (PGM). *Forensic Sci. Int. Genet.* **19,** 192–196 (2015).
153. Ralf, A. *et al.* Forensic Y-SNP analysis beyond SNaPshot: High-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing. *Forensic Sci. Int. Genet.* **41,** 93–106 (2019).
154. Bulbul, O. & Filoglu, G. Development of a SNP panel for predicting biogeographical ancestry and phenotype using massively parallel sequencing. *Electrophoresis* **39,** 2743–2751 (2018).
155. Gallimore, J. M., McElhoe, J. A. & Holland, M. M. Assessing heteroplasmic variant drift in the mtDNA control region of human hairs using an MPS approach. *Forensic Sci. Int. Genet.* **32,** 7–17 (2018).
156. Holland, M. M., Makova, K. D. & McElhoe, J. A. Deep-coverage MPS analysis of heteroplasmic variants within the mtgenome allows for frequent differentiation of maternal relatives. *Genes (Basel).* **9,** (2018).
157. Pechal, J. L. *et al.* The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing. *Int. J. Legal Med.* **128,** 193–205 (2014).
158. Zaheer, R. *et al.* Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.* **8,** (2018).
159. Pereira-Marques, J. *et al.* Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* **10,** (2019).
160. Ahsanuddin, S. *et al.* Assessment of REPLI-g multiple displacement whole genome amplification (WGA) techniques for metagenomic applications. *J. Biomol. Tech.* **28,** 46–55 (2017).
161. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** (2014).
162. Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology* **27,** 105–117 (2019).
163. Stinson, L. F., Keelan, J. A. & Payne, M. S. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Lett. Appl. Microbiol.* **68,** 2–8 (2019).
164. Gu, W., Miller, S. & Chiu, C. Y. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. *Annu. Rev. Pathol. Mech. Dis.* **14,** 319–338 (2019).

165. Cooper, J. E. & Cooper, M. E. Forensic veterinary medicine: A rapidly evolving discipline. *Forensic Sci. Med. Pathol.* **4,** 75–82 (2008).
166. Clark, G. ; & Byrd, J. H. Wildlife DNA: Veterinary Forensic Lab Examines Crimes Against Animals. *Forensic Magazine* (2015). Available at: https://www.forensicmag.com/article/2015/12/wildlife-dna-veterinary-forensic-lab-examines-crimes-against-animals. (Accessed: 25th September 2019)
167. Andy, N. Man Sentenced in Cat Torture Case Involving DNA. *City Room; New York Times* (2011).
168. Milheiras, S. & Hodge, I. Attitudes towards compensation for wolf damage to livestock in Viana do Castelo, North of Portugal. *Innovation* **24,** 333–351 (2011).
169. Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41,** (2013).
170. Meisel, J. S. *et al.* Skin Microbiome Surveys Are Strongly Influenced by Experimental Design. *J. Invest. Dermatol.* **136,** 947–956 (2016).
171. Wang, G. C. Y. & Wang, Y. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142,** 1107–1114 (1996).
172. Balvočiute, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC Genomics* **18,** (2017).
173. Bolyen, E. *et al.* QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ* (2018). doi:10.7287/peerj.preprints.27295
174. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
175. Tessler, M. *et al.* Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.* **7,** (2017).
176. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178,** 779–794 (2019).
177. McIntyre, A. B. R. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18,** 1–19 (2017).
178. Edgar, R. Taxonomy annotation and guide tree errors in 16S rRNA databases. *PeerJ* **2018,** (2018).
179. Lydon, K. A. & Lipp, E. K. Taxonomic annotation errors incorrectly assign the family Pseudoalteromonadaceae to the order Vibrionales in Greengenes: Implications for microbial community assessments. *PeerJ* **2018,** (2018).
180. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176,** 649–662.e20 (2019).
181. Almeida, O. G. G. & De Martinis, E. C. P. Bioinformatics tools to assess metagenomic data for applied microbiology. *Applied Microbiology and Biotechnology* **103,** 69–82 (2019).
182. Nasko, D. J., Koren, S., Phillippy, A. M. & Treangen, T. J. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* **19,** (2018).
183. Methé, B. A. *et al.* A framework for human microbiome research. *Nature* **486,** 215–221 (2012).

184. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017).
185. Mason, C. E., Afshinnekoo, E., Tighe, S., Wu, S. & Levy, S. International standards for genomes, transcriptomes, and metagenomes. *J. Biomol. Tech.* **28,** 8–18 (2017).
186. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* **14,** 1063–1071 (2017).
187. Fritz, A. *et al.* CAMISIM: Simulating metagenomes and microbial communities. *Microbiome* **7,** (2019).
188. Ma, Y., Chen, H., Lan, C. & Ren, J. Help, hope and hype: ethical considerations of human microbiome research and applications. *Protein and Cell* **9,** 404–415 (2018).
189. Juengst, E. T. in *New Visions of Nature: Complexity and Authenticity* 129–145 (2009). doi:10.1007/978-90-481-2611-8_10
190. Hawkins, A. K. & O'Doherty, K. C. 'Who owns your poop?': Insights regarding the intersection of human microbiome research and the ELSI aspects of biobanking and related studies. *BMC Med. Genomics* **4,** (2011).
191. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5,** 1556–1573 (2007).
192. McGuire, A. L. *et al.* Ethical, legal, and social considerations in conducting the Human Microbiome Project. *Genome Research* **18,** 1861–1864 (2008).
193. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E2930–E2938 (2015).
194. Schmedes, S. E., Sajantila, A. & Budowle, B. Expansion of microbial forensics. *Journal of Clinical Microbiology* **54,** 1964–1974 (2016).
195. Ditzler, M. A. *et al.* High-throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase. *Nucleic Acids Res.* **41,** 1873–1884 (2013).
196. Budowle, B. *et al.* Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics* **5,** (2014).
197. Budowle, B. *et al.* Quality sample collection, handling, and preservation for an effective microbial forensics program. *Applied and Environmental Microbiology* **72,** 6431–6438 (2006).
198. Budowle, B. & Chakraborty, R. Genetic considerations for interpreting molecular microbial forensic evidence. *Int. Congr. Ser.* **1261,** 56–58 (2004).
199. Skopp, G. Postmortem toxicology. *Forensic Science, Medicine, and Pathology* **6,** 314–325 (2010).
200. Yang, R. & Keim, P. Microbial Forensics: A Powerful Tool for Pursuing Bioterrorism Perpetrators and the Need for an International Database. *J. Bioterror. Biodef.* **2,** (2011).
201. Leadbetter, J. R. Cultivation of recalcitrant microbes: Cells are alive, well and revealing their secrets in the 21st century laboratory. *Current Opinion in Microbiology* **6,** 274–281 (2003).
202. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7,** (2016).
203. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *mSystems* **3,** (2018).
204. Ward, D. M., Weller, R. & Bateson, M. M. 16S rRNA sequences reveal numerous

uncultured microorganisms in a natural community. *Nature* **345,** 63–65 (1990).
205. De Filippis, F., Laiola, M., Blaiotta, G. & Ercolini, D. Different amplicon targets for sequencing-based studies of fungal diversity. *Appl. Environ. Microbiol.* **83,** (2017).
206. Bonk, F., Popp, D., Harms, H. & Centler, F. PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls. *Journal of Microbiological Methods* **153,** 139–147 (2018).
207. Kalia, V. C., Kumar, R., Kumar, P. & Koul, S. A Genome-Wide Profiling Strategy as an Aid for Searching Unique Identification Biomarkers for Streptococcus. *Indian J. Microbiol.* **56,** 46–58 (2016).
208. Maiden, M. C. J. Multilocus Sequence Typing of Bacteria. *Annu. Rev. Microbiol.* **60,** 561–588 (2006).