Applications of massively parallel sequencing in forensic genetics

Eirik Natås Hanssen

25th January 2018

Thesis submitted for the degree of Philosophiae Doctor

Institute of clinical medicine, University of Oslo
Department of forensic sciences, Oslo university hospital

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-256-2

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgements

This thesis is a result of work done during the period from 2014 to 2018. Although the main focus has been forensic genetics, the project has been highly dependent on the knowledge of experts from other fields of science such as medical and microbial genetics, informatics and statistics. These experts represent different Norwegian research institutions, and their expertise and hospitality have been exceptional. I'm grateful to many.

First, I would like to thank my primary supervisor Peter Gill at University of Oslo (UiO)/Oslo university hospital (OUS) for sharing his great expertise, for thoughtful guidance and his patience during numerous rounds of proofreading.

Thanks to my prior boss at the OUS forensic department Bente Mevåg for the opportunity to leave my casework duties part-time and for her approval of internal funding. I would also thank my current boss Solveig Jacobsen for continuing the support. I'm also thankful for my colleges' extra efforts in my absence, especially during periods of high workload.

Thanks to my co-supervisor Thore Egeland who has made essential contributions to this project. With his large contact network, he has been my main door opener. I am especially thankful for the opportunity to be associated with his biostatistics group at Norwegian university of life science (NMBU) and for his contribution on paper I.

Thanks to my co-supervisor Per Hoff-Olsen at UiO/OUS for sharing his expertise in both forensic genetics and pathology, and for all proofreading.

Thanks to Robert Lyle at UiO/OUS for introducing me to massively parallel sequencing, bioinformatics and for sharing his great expertise in genetics. Robert made essential efforts during the experimental phase, the writing process and during proofreading of paper I.

Thanks to Knut Rudi for giving me access to his lab at NMBU and Ekaterina Avershina for teaching me microbiome sequencing. Especially thanks for the contributions to the experimental design, labwork guidance, writing and proofreading of article II and for being so positive and helpful.

Paper II and III would not have been possible without the contribution from

my co-supervisor Lars Snipen. Thanks to Lars for welcoming me to NMBU and for sharing his great range of knowledge in biology, microbiology, informatics and statistics. Thanks for being inspirational.

Thanks to the bioinformatics group members at NMBU for you warm inclusion. Especially thanks to Kristian Hovde Liland for his contributions on paper III.

Last but not least I would like to thank my parents, my wife Kjersti and the kids Adrian and Julie for their love and support. I am really looking forward to our next vacation, kids!

Oslo January 2018,
Eirik Natås Hanssen

List of papers

[1] Hanssen, E. N., Lyle, R., Egeland, T. and Gill, P. “Degradation in forensic trace DNA samples explored by massively parallel sequencing”. In: *Forensic Sci Int Genet* 27 (Mar. 2017), pp. 160–166. DOI: 10.1016/j.fsigen.2017.01.002.

[2] Hanssen, E. N., Avershina, E., Rudi, K., Gill, P. and Snipen, L. “Body fluid prediction from microbial patterns for forensic application”. In: *Forensic Sci Int Genet* (June 2017). DOI: 10.1016/j.fsigen.2017.05.009.

[3] Hanssen, E. N., Liland, K., Gill, P. and Snipen, L. “Optimizing body fluid recognition from microbial taxonomic profiles”. In: *Manuscript submitted to BMC Bioinformatics* (5th Nov. 2017).

Summary

In forensic genetics, the main purpose has been to support the identification of biological trace samples through DNA analysis. This has been done by using polymerase chain reaction to target and amplify certain short tandem repeat markers and then separate the different amplified fragments by length using capillary electrophoresis. The method has been the gold standard for decades and has been used for generating practically all DNA-profiles stored in national databases around the world. Because of the high level of standardization necessary, the old technology will probably still be used for many years to come. However, massively parallel sequencing platforms have become a promising alternative to the capillary electrophoresis, by having the potential to both improve the current forensic routine analysis and to provide information beyond identification. During the work with this thesis, we have investigated these new possibilities and made contributions in two important and challenging fields of forensic genetics.

DNA degradation is a key obstacle for a successful analysis. During degradation, the DNA molecules are cleaved into shorter fragments, and the more the DNA is affected the less efficient the polymerase chain reaction will be. In the worst cases, the short tandem repeat markers will not be sufficiently amplified to be detected. In living cells, DNA associated with proteins or DNA present in higher ordered structure is shielded against degradation. We performed whole genome sequencing on 4 degraded samples to investigate if this also applies to biological trace material. The sequencing coverage data were adjusted and filtered for GC-effect and low mappability regions respectively, and then used as an expression for the relative amount of DNA present at any genomic region. High abundant regions would be interpreted as regions resistant to degradation and vice versa. However, we found the coverage data to be evenly distributed at the genomic level, the chromosomal level and the sequence level and concluded that for biological trace material, DNA degrades at an even rate throughout the genome. The lack of certain robust DNA regions put a stop to our intention to target such regions in order to develop a superior performing method for analysing degraded trace samples. However, the fact that the degradation rate seems even throughout the genome is

still highly relevant information when developing new MPS based methods.

Information on type of body fluid might be valuable in some cases. Testing for body fluids has traditionally been done by detecting enzyme activity or immunoaffinity. However, these tests can be inaccurate and some have high false positive rates. Alternatively, new gene expression based methods have been developed. These show higher accuracy by measuring body fluid specific mRNAs and miRNAs but have yet not found a wide-spread use. As accurate body fluid prediction is still challenging, we have developed another genetic-based method, primarily meant as a supplement to the gene expression methods. Our method takes advantage of the knowledge generated by health-related studies where it has been shown that bacteria-rich body fluids have a reasonable steady bacterial composition across individuals. These studies have also developed standard laboratory protocols and data handling workflows. In the laboratory, the bacteria in every sample is detected by sequencing different regions in the 16S mRNA gene. The subsequent data handling workflow starts with the building of taxonomic profiles which each represent the bacterial composition of a sample. Then, the dimension of the data is typically reduced by principal component analysis and used as input for a mathematical model such as linear discriminant analysis.

For our initial experimental setup, we used saliva on skin as a study model and sampled 6 different samples from 6 individuals. We used the mentioned standard procedures and tailored the design to measure method performance and the effect from what we regarded as critical factors. Variance analysis of the results confirmed the strong association between bacterial composition and body fluid, but also a weaker effect from person was observed. Other factors such as PCR technique (conventional and digital droplet PCR), sampling technique (tape and synthetic swab) or technical replicates (parallel 1 and 2), had no significant effect. A cross-validation using the experimental data gave an accuracy of 94%, but there was a clear bias when comparing the experimental data to data from the Human microbiome project. However, by changing from the standard to a customized data handling workflow, we were able to remove this bias. The new data handling workflow comprised of a combination of partial least square regression and linear discriminant analysis. In addition, the taxonomic profiles were build using direct binning to taxa instead of the standard binning to taxonomic operational units. When using data from the Human microbiome project for training the linear discriminant regression model and data from the American gut project for testing, we achieved an accuracy of 96%. Microbial data for feces, saliva, nasal and vaginal body fluids were included in these data sets.

Although our method for body fluid prediction is still not ready for casework, we have shown that it has the potential to provide high accuracy and that it seems robust enough to be implemented without excessive intra-laboratory validation ef-

forts. Further work is still needed to find the optimal calculation settings for highest possible accuracy and to develop an interpretation tool for mixtures of body fluids. In addition, a larger inter-laboratory validation study needs to be done.

Contents

Abbreviations

A	Adenine
ANOVA	Analysis of variance
BAM	Binary alignment map
BLAST	Basic local alignment search tool
bp	base pair
BWA	Burrows wheeler aligner
C	Cytosine
CE	Capillary electrophoresis
CNV	Copy number variation
COI	Mitochondrial cytochrome oxydase 1
ddNTP	Dideoxynucleotidestriphosphates
ddPCR	Droplet digital PCR
DNA	Deoxyribonucleic acid
DNASeqEx	DNA-STR massive sequencing & international information exchange
dNTP	Deoxynucleosidetriphosphate
EDNAP	The European DNA profiling group
EMP	Earth microbiome project
FFPE	Formalin-fixed paraffin-embedded
FDP	Forensic DNA phenotyping
G	Guanine
H3K9me3	Histone H3 trimethylation of lysine 9
HMP	Human microbiome project
indels	Insertions and deletions
ISFG	International society of forensic genetics
ITS	Internal transcribed spacer
LDA	Linear discriminant analysis
LINE	Long interspersed nuclear element
LOD	Limit of detection
MPS	Massively parallel sequencing
NCBI	National centre for biotechnology
NN	Nearest neighbour algorithm
OTU	Operational taxonomic units
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLS	Partial least square regression
RDP	Ribosomal database project
REK	Regional committees for medical and health research ethics
RFLP	Restriction fragment length polymorphism
SAM	Sequence alignment map
SCD	Sudden cardiac death
SINE	Short interspersed nuclear elements
SMRT	Single-molecule real-time
SNP	Single nucleotide polymorphism
STR	Short tandem repeats
STRSeq	STR sequencing project

T	Thymine
tDMSs	Twin-differentially methylated sites
VNTR	Variable number of tandem repeats
WGS	Whole genome sequencing
ZMW	Zero-mode waveguide

Chapter 1

Introduction

1.1 Background

The main application of genetics in forensics is to identify donors of biological traces. A typical example would be if a blood stain was found on the suspect's shirt in a violent crime case. Through analysing the blood, a DNA-profile could be deduced and if this matched the victim's profile, he or she would be identified as the donor. However, not all trace samples are this trivial. One sample might be so degraded that the analysis is resultless. Another sample might contain DNA from so many donors that the result is too complex to interpret. For a third sample, the challenge might not lie in identifying the donor, but to link the DNA-profile to activity or type of body fluid. While some of these problems will continue to be insoluble, others might find a solution by the support of new technology.

The inspiration for this work has been the rapid development in DNA sequencing during the last decade. The new technique, often referred to as massively parallel sequencing (MPS), has led to affordable sequencing and is now accessible to the general forensic laboratory. The obvious advantage of MPS over the currently used capillary electrophoresis technology is the high resolution of data and superior capacity, and within this lies the potential for further development of the forensic DNA analysis. This thesis presents two MPS based contributions to support this development.

1.1.1 The DNA molecule

Deoxyribonucleic acid (DNA) is a long-chained molecule. It consists of two anti-parallel DNA strands twisted into a α-helix structure. Each DNA strand is assembled from 4 different building blocks called nucleotides.

DNA molecules are associated with proteins to form chromosomes, and these are organized differently in different organisms. In bacteria, there is typically one

large circular chromosome. In an animal or a plant cell, there are several different chromosomes. These are tightly packed by being associated with histone proteins. In the human cell, there are in total 23 pairs of chromosomes. Each pair has two homologous chromosomes, one inherited from the mother and one from the father. The first 22 pairs are called the autosomal chromosomes, and the last 23rd pair is the sex chromosomes. In addition to nuclear DNA, the human cell also has shorter circular stretches of DNA in the mitochondria. The total DNA in a cell is referred to as the genome and this holds all the genetic information of the organism. In a multicellular organism such as a human, the genome is identical from cell to cell.

The 4 different nucleotide building blocks each consists of a ribose molecule, a triphosphate group and a nucleobase. The difference between the nucleotides lies in the nucleobases. In the DNA strand, the phosphate groups link the ribose molecules together in an alternating chain-like fashion to build the 'DNA backbone'. Each ribose molecule also binds to one of the 4 nucleobases. Thus, a single DNA strand will have bases sticking out from the 'backbone', and these will associate with the bases on the antiparallel DNA strand to form the α-helix. The base called adenine (A) associates with thymine (T) and guanine (G) associates with cytosine (C). For each strand, the order of the bases defines the DNA sequence. The ends of a strand are labelled as 5'-end and 3'-end respectively depending on which carbon in the ribose ring of the terminating base that has the free -OH group attached, and the sequence of the strand is read from 5' to 3' end. As the two strands are antiparallel to each other, sequences are read in opposite directions.

A gene is a stretch of sequence or successive parts of a sequence which codes for a molecule that has a function. The sequence of a gene is read from the coding strand. In the cell, a gene sequence is transcribed into mRNA which is then translated into proteins [1]. The protein-coding sequences of genes together with non-protein-coding genes and regulatory sequences are the only genetic regions known to be function related, and these constitute only a minor proportion of the genome [2]. The larger part of the genome is non-coding and composed of repetitive sequences (such as LINEs, SINEs and tandem repeats), introns (non-coding part of genes), retroviral elements (might originate from retrovirus), pseudogenes (gene-like elements having lost functionality) etc. Whether these regions play a role in cell physiology is highly debated [3, 4].

The DNA sequence is near identical from human to human with only $\sim 0.1\%$ being different [5]. These differences can appear as single nucleotide polymorphisms (SNPs), which are nucleotide differences at one base pair (bp) position, or as indels, which are either insertion or deletion of a sequence. SNPs and indels are found throughout the whole genome but are less frequent in coding regions because of evolutionary pressure [6]. Another form of variation is found in the mini- and microsatellite DNA positioned in and around the chromosomal centromeres

and telomeres [7]. These noncoding regions consist of repetitive sequence where the number of successively repeated subunits differs between individuals. The minisatellites are sometimes referred to as variable number of tandem repeats (VNTRs) and have a subunit length of $\sim 8 - 100$bp. Likewise, the microsatellites are named short tandem repeats (STRs) and have a subunit length of $\sim 1 - 7$bp. There is also other forms of genomic variation such as copy number variation (CNV), which has repetition of longer segments of sequence, and Alu elements, which is transposable and can vary in frequency. However, these are peripheral to or beyond the scope of this thesis.

For further reading on the topic of general genetics see the textbook Genetics by Meneely et al [8].

1.1.2 Human identification

The field of forensic genetics started with the VNTR markers in the mid-eighties [9]. By measuring a combination of these from different parts of the genome, a DNA profile could be deduced. As the number of combined VNTR markers got larger, one was able to identify people from their DNA. This principle is still the basis for determining paternity and other kinship, identifying bodies and remains and to solve criminal cases by identifying biological traces. However, the applied DNA typing methods have been adjusted to rapid technological and scientific development.

The initial analysis technique was restriction fragment length polymorphism (RFLP). This used restriction enzymes to cut the DNA strand close to the VNTR, which were then labelled with a homologous probe and separated by gel electrophoresis. Radioactively labelled multi-locus probes were used for detection. These created a complex pattern with a high power of discrimination, but their use was labour intensive and they were difficult to apply for mixed samples with DNA from more than one person. By the mid-eighties, they were replaced with the more efficient single-locus probes [10].

From the early nineties, the VNTRs were gradually overtaken by STRs [11, 12]. The chosen STRs were composed of 3-4bp subunits, and dependent on the STR, these subunits could all have the same sequence or form a pattern of different sequences. Based on the composition of subunits, the STRs were categorized into simple, compound and complex [13]. The new STR method used Polymerase chain reaction (PCR) to increase method sensitivity, and several STR markers were amplified simultaneously by using a multiplex of different primer pairs. The amplification product was separated by capillary electrophoresis (CE), and the primers were labelled with fluorescent dye to facilitate detection. This method was also more suitable for degraded DNA and was far less labour intensive then RFLP. In

addition, the statistical calculations and interpretation were simplified with the shift to automated methods of analysis [14]. Since its introduction, the STR method has been continuously optimized and is still the gold standard in forensic genetics. In addition to the advantages already mentioned, a significant reason for the method's success is the large national and international DNA databases that have been built based on STR profiles. Because of the considerable investment, it is difficult to imagine the introduction of a new alternative method unless it is compatible with the standard STR markers.

1.2 Limitations of human identification

The STR method outperformed RFLP when analysing degraded DNA, but it is still not optimal. DNA is fragmented when degraded, and if the STR marker region is broken, the PCR amplification will be disrupted for that particular DNA molecule. The greater the degradation the more evident this problem will be when analyzing a trace sample. As a result, the STR method can in the worst case fail completely, and no result will be obtained. A marker that is extended as a long stretch of DNA will be more vulnerable than a shorter marker. As an alternative, shorter markers such as SNPs will be a good choice when analysing degraded DNA. An obstacle is that SNPs are not compatible with the STR profiles registered in the DNA-databases. A pure SNP based method will therefore only be useful in cases where both trace and reference samples are analysed using the same markers. To compensate for the high selectivity of the STRs, more SNP markers have to be included in the analysis panel [15].

The interpretation of complex mixtures is dependent on large amounts of data beyond the capacity of the CE. Using an alternative high capacity analysis platform to add more STR markers is an obvious solution. Another feature of the STRs that can be exploited in this respect is the sequence variation that is found both in the flanking regions and in the repetitive regions [16, 17, 18, 19, 20]. By using this increased STR polymorphism, 30% of the homozygous markers in CE generated DNA profiles were heterozygous when derived from sequence data [19]. This will also help to some extent with identifying stutters and other artefacts. Another limitation for mixtures when using CE is detecting minor components. When the ratio between minor and major component is around 1:20, the minor often has too many allelic dropouts to be identified [21, 22].

1.3 Beyond human identification

The forensic scientist's main contribution to a criminal case is to assist in the identification of biological trace material through STR Profiling. A pure trace profile has an extremely high discrimination power, and when matching a reference profile representing a known person it will give an overwhelming supporting evidence if the trace is from that person rather than from another unknown individual. With an identified trace sample, the police may be able to solve the case. However, in other cases, additional information beyond identification might be necessary to give a certain biological trace any evidential value, or alternatively, to help the police investigation. The scientist's toolbox is still not sufficiently equipped to provide such additional information, but there is a large potential to exploit genetic information beyond STRs [23]. There are numerous examples of how this could be beneficial.

The court seeks to link biological traces to the criminal act in order to answer the ultimate question, "what actually happened?". Towards this effort, information on the type of biological material (eg blood or semen) might be essential. The STR profile does not provide such information, and in many cases, nothing can be deduced from the sampling position. In addition, the alternative proposition of the defence might be that the biological material was accidentally or innocently transferred [24, 25, 26]. However, if there was information that the DNA was associated with vaginal cells, the evidential value might increase. Presumptive testing for blood, saliva and semen has been used in forensics for decades. Typically an enzyme specific to a body fluid is being detected by a chemical reaction. In addition to the presumptive tests, there are also some lateral flow immunoassays available (`http://www.ifi-test.com/`). Although these tests have different degrees of accuracy, all can provide false results, and their selectivity and specificity are typically not given. Promising alternative methods, most based on gene expression measurements, have been reported [27, 28]. The European DNA profiling group (EDNAP) has for example performed collaborative studies on mRNA tests for blood [29], saliva and semen [30], menstrual blood and vaginal secretion [31] and finally on skin [32].

If the perpetrator is unknown and the biological trace sample is still unidentified after a database search, the police investigation might be in need of additional information for a quick solution to the case. The perpetrator's characteristics would obviously be beneficial in these cases. While the STR profile provides no such information, some characteristics can be derived from the genetic code. Even though many traits can still not be derived from gene sequence, there have been several successful studies on predicting hair, eye and skin colour [33], ancestry [34, 35] and age [36, 37]. A few commercial forensic analysis kits have started to become available (`http://www.Illumina.comandhttp://www.thermofisher.com`).

Monozygotic twins have identical STR profiles, but by performing extended analysis it has also been possible to separate twins. Two alternative approaches have been used to achieve this. The most resource demanding is to identify private mutations in the two twins by doing whole genome sequencing (WGS) of their reference samples and then target these loci in the trace sample [38]. A promising alternative method is to analyse twin-differentially methylated sites (tDMSs) [39, 40]. This is obviously a more affordable approach, but future studies are needed to prove if the method is applicable for forensic purposes.

A molecular autopsy is often performed on sudden cardiac death (SCD) cases where the deceased is below 40 years [41]. These genetic analyses are performed in cases with negative toxicology and pathology analyses. Different gene panels are typically used, but in some cases, all genes have been sequenced by exome sequencing. Lately, there has been a rapid advancement and discovery of novel disease-related genetic markers, and this development is accelerating [42]. In addition to SCD, there are also tests for different lethal infections [43] and genetic metabolic disorders which can cause poisoning in connection with medication and drug abuse [44, 45].

Non-human DNA analysis can also be useful in forensics. Microbial forensics is a newly emerging field, and several studies have applied microbial sequencing. The potential to predict post-mortem intervals has been demonstrated by using microbial composition data from human skin [46], human gut [47, 48] and mouse models [49]. It has also been shown that microbial composition data can be used to separate between samples taken from two different locations (phones or shoes) [50], and that such data even has the potential to identify the donor if samples are taken from touched objects [51, 52, 53, 54]. Microbial sequencing can also be used to identify hazardous or infectious microbes in connection with bioterrorism [55, 56] and infectious disease transmitted during a criminal act [57, 58].

Wildlife forensics and forensic botany are other fields where non-human DNA analysis has been beneficial. For animals, the Barcode for Life Consortium has defined the mitochondrial cytochrome oxidase 1 (COI) gene as the standard locus for identification at the species level [59]. For presumptive identification of individual animals or for pedigree assignment, different STR- and SNP-panels have been used dependent on species [60]. For plants, it has been more challenging to define a standard barcode sequence. However, there seems to be general consensus on that a combination of rbcL, matK, trnH-psbA spacer and the internal transcribed spacer (ITS) sequences should be used for identification at the species level [61].

1.4 DNA sequencing

CE has been the workhorse in the forensic laboratory for decades and still is. The currently used instrumentation, such as the 3500 Series (Thermo Fisher Scientific), will produce a DNA profile consisting of nearly 30 markers. The new Spectrum CE (Promega) will make it possible to include even more markers. However, the CE technology is near its maximum capacity limits, and cannot offer what is needed to improve the current methods and to bring new applications into forensic genetics. Hence, there has to be a transition to new technology with higher capacity, and Massively parallel sequencing (MPS) seems to be the obvious candidate [62].

Sanger sequencing has traditionally been the prime method for DNA sequencing [63]. In its mass production form it was even used for sequencing the human genome [64]. However, the Human genome project revealed the need for more advanced sequencing technologies and was driven by the need for lower costs, the first truly MPS platform was launched in the mid-2000s (eg 454 sequencer, 454 Life Sciences). Today high throughput platforms (eg NovaSeq, Illumina) can each sequence several thousand human genomes a year to under 1000$ per genome [65].

There are several different MPS technologies. However, they share the basic principle of parallel sequencing of a huge number of DNA fragments (typically several million of DNA fragments are sequenced simultaneously by the same instrumentation). In that respect, it is an up-scaling of Sanger sequencing which uses CE and is therefore limited to the number of parallel capillaries (eg 3730 DNA Analyzer from Thermo Fisher is used for Sanger sequencing and has up until 96 parallel capillaries). With MPS the DNA regions of interest are typically sequenced several times to exclude read errors by consensus. The number of times they are sequenced is denoted as the sequencing coverage. For example, when sequencing is done with 30x coverage, the genetic regions have been sequenced 30 times on average. MPS technologies can be divided into two groups, dependent on the length of DNA sequences that can be read. Short read platforms sequence fragments usually between 50 and 400bp whereas long read platforms sequence fragments usually in the range from 10,000 up to 100,000bp [65, 66]. Short reads are relatively cheap, but cannot be used to derive the sequence of repetitive regions longer than the actual read length. Longer reads are more expensive, but can be used to explore these longer regions and are therefore also essential for genome assembling [67].

1.4.1 Sanger sequencing

The sequence of interest is typically first amplified by PCR and the amplified fragments denatured to give single-stranded DNA fragments. In the sequencing reaction mix itself, these single-stranded fragments are combined with several different chemical components. A sequencing primer will bind to a region near the one end of the fragments. Then for each fragment, a DNA polymerase enzyme attaches to the primer and starts incorporating deoxynucleoside triphosphates (dNTPs include dATP, dGTP, dCTP and dTTP) in a growing homologous strand. Beside these 4 ordinary dNTPs, there are also present 4 dideoxynucleotides triphosphates (ddNTPs) which will stop the DNA synthesis if incorporated. The ddNTPs are added in low amounts relative to the dNTPs to facilitate a reasonable long read length. Since the reaction mix contains many single-stranded DNA fragments and the ddNTPs are incorporated randomly, there will be many different fragment lengths created at the end of the extension process. Nevertheless, the process is repeated several times by PCR so that fragments corresponding to each position in the sequence have been produced in sufficient amounts to be detected. Detection is possible as the ddNTPs are labelled with individual fluorescent dyes. The fragments are separated by size with CE and the sequence read directly from a fluorescent detector.

With Sanger sequencing, the read length is typically around 700bp and not above 1000bp. Beyond this length, the CE platform will have problems separating individual bases. Today Sanger sequencing is performed in smaller projects with a limited amount of samples. For larger studies, it cannot compete with MPS technology [68]. BigDye (Thermo Fisher) is one of many commercial kits available for Sanger sequencing.

1.4.2 Massively parallel sequencing

Library preparation

Sample preparation for MPS sequencing is extensive (see figure 1.1 for an example). In this process, the purified DNA extract is converted into a library consisting of DNA fragments ready to be sequenced. Each of these fragments consists of a portion of the sequence in question, often named insert, flanked by primers, indexes and adaptors needed for sequencing. Dependent on technology, the library fragments might also have other units incorporated. The length of each insert has to fit the applied read length, but together the inserts represent all DNA that is to be sequenced.

Initially, after DNA extraction and cleanup, there are different approaches for selecting the genomic regions to be sequenced. When sequencing whole genomes or longer DNA fragments, fragmentation is carried out directly by sonication [69]

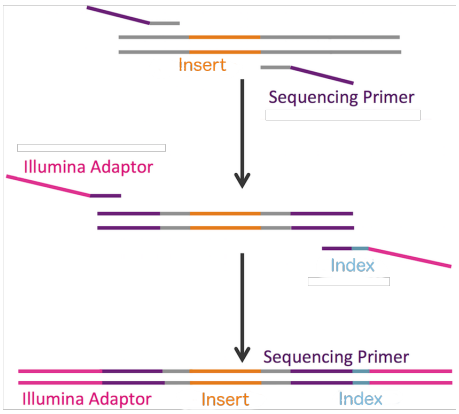

Figure 1.1: Illumina's TrueSeq library preparation workflow as a generic example. In the finished library fragment (bottom) the sequence in question/insert (orange) is surrounded by sequencing primers (purple), index for sample identification (grey) and adaptors (pink). Source: Kowalsky et al 2015 [74].

(eg by Focused-ultrasonicator, Covaris). If only shorter stretches of DNA are of interest, target enrichment strategies are the obvious choice [70]. Here PCR and hybrid capture are two frequently used techniques. Many of the commercial forensic kits use PCR amplification to target specific STRs, SNPs or indels (eg ForenSeq, Illumina or Precision ID GlobalFiler, Thermo Fisher). As the majority of forensic samples have low DNA levels, the PCR amplification is needed to enable detection. The available forensic kits use shorter fragments, but longer stretches of DNA could also be targeted by PCR. However, fragment size should be kept below 10kb [70]. If the fragment size is too long, the PCR product should be fragmented to fit the chosen read length. Large PCR multiplexes are also a possibility, and by digital PCR several thousand sites could be amplified at the same time (Digital PCR solutions, RainDance Technologies). Alternatively to PCR, hybrid capture is another target enrichment strategy. The DNA extract is then first fragmented and the fragments of interest are "fished out" using array-based capture [71] or in-solution capture [72]. For details and information of other target enrichment strategies, see Kozarewa et al [73].

With the wanted DNA fragments enriched, different synthetic fragments have to be attached to their ends dependent on sequencing technology. Generally, sequencing primer fragments will facilitate the binding of the sequencing primers and adaptor fragments will help to anchor the fragments while sequencing. For sample identification, indexes with unique sequence are typically ligated together with the adaptors. Finally, DNA concentration of the library is measured so that equimolar aliquots can be pooled and samples sequenced together.

Short read sequencing

Before starting sequencing on a short read platform, the fragments in the library needs to be amplified to give clusters of identical clones. This process is called clonal amplification and is done to enable detection when reading each of the original fragments. The sequencing platforms rely on different technologies for clonal amplification and sequence reading.

Illumina platforms such as HiSeq, NovaSeq and MiSeq are the most used platforms, and they all use the same principle for sequencing [75, 65]. The read length is typically between 150 and 300bp per read [65]. It is common to do paired-end sequencing where inserts are read from both ends. The sequencing itself takes place on a slide placed in a flow cell. The original library fragments attach to slide-bound adaptors before being clonally amplified by so-called bride amplification into a "lawn" of clusters (each cluster having fragments of identical sequence). The sequence is read by flushing all 4 fluorescently-labelled 3'-O-azidomethyl-dNTPs simultaneously over the cell. Similar to the ddNTPs used for Sanger sequencing, the azidomethyl-dNTPs stops the extension, but in this case for each base incorporated. During incorporation, a light with a wavelength dependent on the base is emitted. Hence, the sequence can be read by taking a photo of the flow cell for each flushing cycle. To enable incorporation of a new base, the fluorescent moiety and the 3' block are removed just before the next flush cycle. The sequencing error rate is typical $\sim 0.1\%$ for the Illumina technology [65].

For Ion torrent platforms such as Ion S5 and PGM, individual library fragments are attached to beads and amplified by clonal amplification using emulsion PCR. For one bead this results in clones of the initial fragment covering the whole sphere. Thus, the bead becomes equivalent to a cluster on the Illumina flow cell. The beads are then distributed to individual wells on a sequencing chip where each well has a pH sensor. The different dNTPs are flushed sequentially over the chip, and if incorporated, H+ ions are released. The DNA sequence can then be read from detection the pH shift. The number of H+ ions released is proportional to the number of dNTPs incorporated simultaneously, and this is used to read homo-polymer stretches of sequence (stretches that have the same base throughout the whole sequence). The Ion torrent technology uses single-end sequencing where the insert is read from one side only. The read length is typical 200 or 400bp, and the error rate is $\sim 1\%$, mainly caused by difficulties in reading homo-polymer stretches [65].

In addition to the mentioned technologies, there are alternative short read platforms such as the relatively new GeneReader (Qiagen) system.

Long read sequencing

The most common long read sequencing platforms are based upon single molecule detection where the detector optics is sensitive enough to read the incorporation of single dNTPs. Consequently, no PCR amplification is needed for cluster generation as for the short read technologies. However, the long read technology still demands a relatively large amount of input DNA (250–5000ng dependent on technology) [62].

The Single-molecule real-time (SMRT) technology applied by Pacific Biosystems is the most used technology for long read sequencing [65]. The sequencing adaptors have a hairpin structure making the original double-stranded fragments into a single-stranded circular molecule. The original fragment length can be up to 40kb. The sequencing reaction takes place in a zero-mode waveguide (ZMW) well where an active polymerase complex is bound to the bottom [76]. The sequence is read in real time as wavelengths corresponding to the incorporated fluorescent dNTPs are emitted [77]. The circular shape facilitates reading the original fragment in both directions and multiple times. By this, the random sequencing error is reduced from 13% to 0.001% by consensus [65]. The SMRT technology can by measuring polymerase kinetics also detect DNA methylation [78].

MinIon (Oxford nanopore technologies) is based on a technology where protein nanopores are inserted into an electrically resistant polymer membrane [79]. Leader and hairpin adaptors are ligated on each side of the double-stranded DNA fragment. The leader adaptor helps the positioning of the fragment into the current leading pore, and a motor protein pulls one of the DNA strands through the pore [80]. The voltage across the pore is modulated according to the k-mer sequence positioned in the pore at any given time, and these changes can be used to derive the sequence. The signal outcome has more than 1000 levels, one for each type of k-mer, and hence information on modification of bases in native DNA can also be extracted. When the whole length of the fragment has been read through the pore, the hairpin structure at the end will help the second complementary strand being pulled into the pore and then read. Alternatively, if the hairpin adaptor is not applied, only single strands will be read. The nanopore technology has a large potential because of the long read lengths (up to 200kb), easy library preparation and high mobility of the equipment, but the use has been limited by high error rates ($\sim 12\%$) [65]) and low robustness [66].

An alternative to the single molecule sequencing technologies is synthetic long reads [81]. This technology uses the normal short read platforms, but the difference lies in the library preparation. One or a few longer DNA fragments are captured in small reaction chambers (wells or emulsion droplets), and here they are fragmented and labelled with a certain index. After short read sequencing, the

fragments which originate from an original larger fragment can be isolated through the common index and assembled into a local sequence. In this way, even longer repetitive sequences can be assembled. The 10X Genomics emulsion-based system can handle fragments up to 100kb [65].

For further information on MPS platforms and technologies there are several complementary review papers [82, 83, 75, 84, 65, 66].

1.5 Bioinformatics

Bioinformatics is applied in many biological disciplines, but only sequencing related topics fall within the scope of this thesis. MPS produce massive amounts of data and it would be impossible to interpret these without the support of informatics. The need for extreme computing power has been so demanding that it has pushed the evolution of large computer cluster systems [85]. From the start of the MPS era scientists have developed their own software to fit their needs, but also to help others by making the software available as freeware or open-source. Even though much software has been short-lived, some tools have become standard. Most of the work has been done in the Unix environment as this is extremely fast and efficient. The disadvantage is that Unix is relatively inaccessible to the average Windows user and requires some effort in the beginning. As a consequence, semi-automatic platforms have surfaced, such as Galaxy (`https://galaxyproject.org/`) where the user can get access to universal workflows based on standard software. However, there is still no easy solution and some customization is always needed.

The scope of sequence-related bioinformatics is large, but some main fields of application are resequencing, de novo assembly and RNA-seq. Resequencing is done for example to measure variation between individuals, identify certain species or genotype individuals. De novo assembly is done for example when no reference sequence exists and the reads have to be fitted together to derive one. RNA-seq is used for example, to measure the level of mRNAs in tissues or individuals for genetic expression studies. De novo assembly and RNA sequencing lie beyond the scope of this thesis, and comprehensive information can be found elsewhere [86, 87]. As resequencing has been applied throughout the work with this thesis, it is used below as an example to illustrate a generic workflow and the use of the most important bioinformatics tools.

For resequencing, there has to be a reference sequence for comparison, and such are typically available through online services. Whole genome reference sequences are available for many organisms. A resequencing workflow typically starts with the output file from the sequencing platform. This file is nor-

mally in a fastq format and includes reads from all the samples sequenced in the same batch. For each read, the fastq format includes the sequence and the corresponding base call quality. First, the data has to be demultiplexed into individual samples, which is made possible by the unique indexes used in the library preparation. Demultiplexing is often supported by the sequencing platform software. Overall read quality can be evaluated by tools like FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then the raw reads are filtered based on base call quality and typically trimmed for regions of poor quality base calls and adaptor sequence. Filtering and trimming are performed by tools such as Trimmomatic [88] or Cutadapt [89]. The reads are then mapped to their original genomic position by using the reference genome and software such as Burrows wheeler aligner software (BWA) [90] or Bowtie [91]. The output file from the mapping tool is typically in a Sequence alignment map (SAM) format, which in addition to the information in the fastq file, also includes several output parameters from the mapping process. This SAM file is often compressed to the more efficient Binary alignment map (BAM) format by using a tool such as SAMtools [92], and the resulting BAM file is then used by many downstream applications. This is also the first point where the mapped reads can be visually inspected by the use of tools like Integrative Genomics Viewer [93]. If the aim is calling variants, the BAM file first has to be prepared by sorting, adding metadata and removing duplicates. This is typically done by using software such as SAMtools and Picard (http://broadinstitute.github.io/picard/). In the variant calling process, there is often first a realignment step, where local miss-alignments are corrected, before the actual variant calling. These final steps are performed by tools like GATK [94]. Comprehensive literature on resequencing can be found elsewhere [95].

Sequencing studies demand solid funding, as sequencing is still relatively expensive. Luckily, the extent of sequencing can often be limited by using public data available through online resources. The latest version of the human reference genome is GRCh38, and this can be downloaded through the National centre for biotechnology (NCBI) web page (https://www.ncbi.nlm.nih.gov/). For microbial resequencing, 16S reference sequences can be found in large data repositories such as the Silva database (https://www.arb-silva.de/), the Ribosomal database project (RDP, https://rdp.cme.msu.edu/) and the Greengenes database (http://greengenes.lbl.gov/). Beyond this, whole genome reference sequences for a large variety of organisms can be accessed through large genomic browsers like Ensembl (http://www.ensembl.org/index.html) and UCSC genome browser (https://genome.ucsc.edu/). In addition to reference sequences, there is also available data from large consortiums on human diversity like 1000 genomes project (http://www.internationalgenome.org/) and

on microbial diversity like Human microbiome project (HMP - https://hmpdacc.org/) and Earth microbiome project (EMP - http://www.earthmicrobiome.org/). Identifying sequence of unknown origin is typically done by using Basic local alignment search tool (BLAST - https://blast.ncbi.nlm.nih.gov/Blast.cgi) which will search against reference databases such as NCBI genebank (https://www.ncbi.nlm.nih.gov/genbank/), DNA dataBank of Japan (http://www.ddbj.nig.ac.jp/) and the European nucleotide archive (https://www.ebi.ac.uk/ena).

1.6 Biostatistics

Interpretation of MPS data also needs knowledge of biostatistics. Statistical calculations are often done in the R environment (R Development Core Team, https://www.r-project.org/) or in Python (Python Software Foundation, https://www.python.org/) and can be supported by the add-on modules like Bioconductor [96] and Biopython [97], respectively.

The statistical platforms provide tools for basic calculation such as statistical testing, regression and analysis of variance (ANOVA), and by combining these with available open source packages, the individual scientist can create scripts for customized data interpretation workflows. In the work with this thesis different statistical tools and methods have been used, but pattern recognition has been essential and will be discussed in more detail.

Pattern recognition is a part of machine learning or statistics more generally, where regularities in a training dataset are used to predict characteristics of samples in a new test dataset. The training and test datasets need to have the same format. Linear discriminant analysis (LDA) and nearest neighbour (NN) are two alternative models for pattern recognition. LDA is a linear model where a linear hyperplane is used to separate groups of samples, whereas NN uses the nearest data points in the training data set for classification of a new test sample. NN demands a large training data set to map the space of all possible outcomes. Hence, NN models have the potential for high accuracy, but may be unstable and overfitted. LDA demands fewer data, and to compensate for this LDA makes huge assumptions. As a consequence, LDA is stable, but without the potential for extremely accurate predictions in very large data. Despite this, LDA is a popular method for pattern recognition, much due to its simplicity, and the relatively low level of resources needed for data collection.

LDA needs input data of full rank which means that all columns in the input data matrix have to be independent of each other. If this is not the case, it is possible to remove these dependencies before the LDA step by reducing the dimensions in

the data, eg by using Principal component analysis (PCA) or Partial least square regression (PLS). PCA transforms the dataset into a space of orthogonal principal components where each component is chosen to include most possible of the remaining variance in the data. PLS finds the relationship between the independent X data and the dependent Y data by calculating the direction in X-space which explains the largest possible part of the variation in Y-space. As PCA does not use the dependent data it is defined as an unsupervised method. PLS, on the other hand, uses the dependent data and is therefore defined as a supervised method.

For further reading on the topic of pattern recognition, the reader is recommended to read Hastie et al [98].

1.7 Current status of MPS in forensics

MPS has had much attention in forensic research for the last few years and has been among the main topics at major conferences, lately at the International society of forensic genetics (ISFG) conference in 2017 (`http://www.isfg2017.org/`). However, implementation of new technology into the forensic routine laboratories is naturally a long and consuming process. Among 33 European laboratories, 20 have already invested in MPS instrumentation or will do so in the coming few years [99]. Most of the European laboratories are reporting that they are currently evaluating MPS protocols for typing autosomal STRs and SNPs in addition to Y-STRs. This is not surprising as many of these markers are included in the standard CE based identity panels. To the author's knowledge, only a few laboratories have already implemented MPS as a routine method in casework. According to the same survey, the laboratories view the largest hurdles for implementation of MPS as lack of reporting standards, lack of DNA database compatibility, insufficient population data and no adequate legislation.

The first sequencing studies on forensic relevant autosomal STRs was performed from the beginning of the decade [100, 101, 19]. These were performed with the 454 Genome Sequencer platforms (Roche), and Van Neste et al [101], who used the Profiler Plus (Applied Biosystems) for amplification, reported of difficulties with a low level of full length reads and homopolymer sequencing errors. Since then production of the 454 platforms has been terminated, and the most relevant studies have been done on the PGM/S5 platforms (Thermo Fisher) or the MiSeq platform (Illumina). A large majority of these studies have been done to evaluate performance of different STR panels such as the commercially available ForenSeq kit (Illumina) [102, 103, 104, 105, 106, 107, 108, 109] and prototype versions of STR panels from Promega [110, 111, 112] and Thermo Fisher [113, 114]. There has also been performance studies on customized STR panels [115]

and Y-STR panels [116, 117]. All these kits use PCR target enrichment as the alternative hybrid capture strategies are not yet sensitive enough [62]. Summarized, the performance equals that of the CE based STR kits when comparing the standard validation parameters such as repeatability, concordance, inter- and intra-locus balance and stutters percentage. The analytical threshold is reported to lie in the region between $10 - 50$pg of input DNA, and for 2 person mixtures, the minor component is identified down to 1:20 ratio. Performance is also similar in the presence of PCR inhibitors and for real case samples, and MPS even outperforms CE for degraded samples. The latter is caused by MPSs independence on fragment size separation, and when the STRs can be reduced to their actual sizes (mostly below 260bp [111]), valuable partial DNA profiles can be obtained even for samples where mean fragment lengths are $\sim$ 200bp [114]. In connection with these studies, there has also been pointed out some potentially underperforming markers [104, 105] and raised concern about the limitations of interpretation software and the relatively high cost of MPS forensic analysis [62]. In addition, MPS also has longer runtime compared to CE [106]. Others have expressed the need for joint standards on databasing, data storage and nomenclature [109].

Another important condition for seamless implementation of MPS in forensics is representative population databases, and frequency data has been reported for several populations such as Korean [118], Spanish [119], Greenlandic [120], Basque [121], Dutch [111], Chinese [122] and US populations [112, 123, 104]. As mentioned above, there are isoalleles that have the same length but differ in sequence. Isoalleles are mainly observed for the compound and complex STRs, and for 9 STR loci, the increase in numbers of alleles is $> 30\%$ [20]. In order to quantify the lowest allele frequencies, the frequency databases have to be large (include several thousand samples). Another issue is backwards compatibility towards CE based DNA profiles and the ambiguity that can occur in some cases. For example, if the flanking of a repetitive region contains an indel, the allele call derived from counting numbers of repeats will be different from that obtained from measuring STR length. Finally, STR variants uncovered by sequencing could potentially be associated with disease. FGA and SE33 both contain exons in the flanking regions, and one SNP in the flanking region of FGA is known to be associated with a rare blood coagulation defect [124].

With the advent of MPS, the identity SNP markers have had increased attention because these can now be co-analysed with the STRs. The SNPs can be a valuable supplement when more data is needed for mixture interpretation, or as an alternative marker set for degraded DNA samples. There have been several performance studies on the ForenSeq kit (Illumina) [102, 103, 104, 105, 106, 108, 109] and the AmpliSeq/Precision Identity kits (Thermo Fisher) [125, 126, 127, 128], in addition to a new 140 SNP panel (Qiagen) [129] and a customized 273

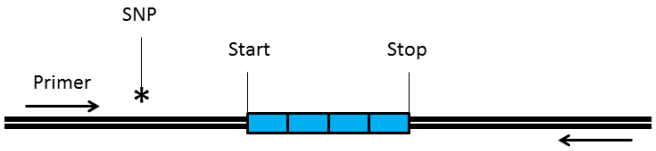

Figure 1.2: The DNA commission of ISFG's proposition on nomenclature for STR sequence data where assignment is based on forward strand only. The start and stop coordinates of the repeat region (blue) is proposed as anchor points. The STR in the figure would be assigned as *D13S317[CE12]-Chr13-GRCh38 82148025-82148068 [TATC]$_{12}$ 82148001-A* where the different parts can be explained as: *D13S317[CE12]* is locus name and CE allele name, *Chr13-GRCh38* is chromosome and version of reference genome, *82148025-82148068 [TATC]$_{12}$* is start and stop coordinates and repeat motif and *82148001-A* is the location of sequence variant (SNP) in the flanking region. Source: Parson et al 2016 [133].

SNP panel [130]. Overall, the SNPs perform similarly to the STRs for the standard validation parameters, including analytical threshold and detection of the minor contributor in mixtures. As for the STRs, there has also been reported on a few poor performing SNPs markers, especially when the samples have low DNA levels [125, 104, 105, 128]. Those who have evaluated degraded DNA analysis report on improved performance for the SNPs compared to the STRs. Guided by the degradation parameter of the latest quantification kits (Quantifiler Trio (Thermo Fisher) or PowerQuant (Promega) it then becomes possible to choose SNP based analysis exclusively for challenging degraded trace samples [126].

SNP panels for biogeographical ancestry are also commercially available, and performance has been evaluated for the Forenseq kit (Illumina) [104, 105, 106, 131, 108] and the Precision ID Ancestry kit (Thermo Fisher) [120, 132]. In general, the technical performance is similar to that of the identity STR and SNP panels mentioned above. The kits separate easily between individuals from the large population groups roughly divided by continents [132, 108], but despite this, the Forenseq panel has been found useful even in a society of multiple populations [131]. For the Precision ID Ancestry kit, difficulties have been reported when sequencing a few markers with homopolymeric sequence [132]. It has also been shown that it is essential for accuracy to have representative data in the applied population databases [120]. SNPs for phenotypic traits like eye, hair and skin colour provide similar information as the biogeographical SNPs [33]. The phenotypic SNPs have been included in the ForenSeq kit, and from the same studies as mentioned above, performance is similar to the other SNP panels evaluated.

The sequence STR data have to be compatible with the millions of CE generated DNA profiles stored in the national databases. To facilitate this, a common nomenclature for sequencing data has to be established, and the DNA commission of ISFG has already published minimal requirements [133]. They propose that sequencing data should be exported and stored as text strings to capture all information and that only forward strand sequence should be given. A common reference such as GRCh38 should be used, and the coordinates for start and stop points of the repeat region are proposed as anchor points. To allow communication of results, the simple STR nomenclature of the CE base DNA profiles could be used, but the nomenclature should also include information on sequence variation (see figure 1.2). However, the Commission believes that future software could remove the need for nomenclature by calculating the strength-of-evidence directly from string based frequency databases. In addition to the commission's recommendations, there has also been published guidelines for publication of genetic population data [134]. According to these, a minimum of 50 individuals should be included per publication, and only high-quality full genotype profiles should be submitted in string format. Quality control can be done by the already established central curator system of STRidER (`http://strider.online/`) [135].

In addition to the commercial software provided by the vendors of the sequencing platforms, there is also free community software available to support handling of sequence data. Tools such as STRait Razor [136], STRinNGS [137], SEQ Mapper [138] and ToaSTR [139] assign STRs from sequencing data, and SEQ Mapper assign even SNP markers. FDSTools is a software for recognition and removal of stutters and other analytical noise in order to facilitate detection of low-level minor mixture components [140]. NOMAUT is a software under development by the EU supported DNA-STR massive sequencing & international information exchange (DNASeqEx) project which is planned to be a STR nomenclature web service for sequence queries. In the ongoing STR sequencing project (STRSeq) STR data will be maintained as GenBank records at NCBI, and tools will be developed to facilitate interaction with the mentioned STRidER web portal. Even though much effort has already been invested, it is evident that significant resources have to be put into building new software in the coming years [62, 133].

1.8 Ethical and legal issues

The development of forensic genetics has always been accompanied by ethical considerations and legal implication [141]. Currently, most countries have legislation that prohibits deriving forensic genetic information for any other purpose than identification. The extended MPS analysis of the standard STRs and SNPs

for identification should not be in conflict with these restrictions. However, there has been raised some concern regarding a few of these markers having sequence variants that might be associated with ancestry or disease [112]. Another aspect is that increased discrimination power could facilitate the use of extended familial searching in the national databases [142].

Forensic DNA phenotyping (FDP), which includes biogeographical ancestry or visible traits, has been more controversial. Several countries prohibit the use of coding markers in forensics (eg Germany). To our knowledge, the Netherlands is the only country which explicitly allows determination of biogeographical ancestry, while the United Kingdom allow FDP without dedicated legislation [23, 143]. The general critical view is that these analyses may reinforce existing prejudice and racist generalizations [144], and that the outcome is too broad and will stigmatize large groups of people [145]. Mass screening of reference samples from such large groups should also be avoided. Others have questioned the relatively high chance of over-interpretation the outcome due to the probabilistic nature of these analysis [146]. On the other hand, it has been argued that the visual appearance of a person cannot be hidden and therefore cannot be considered as private data [143]. However, there seems to be a consensus that only forensic relevant information should be obtained [142], and that no personality traits or disease associated information should be reported [147]. However, there may be instances where these two considerations might conflict [143].

For the studies included in this thesis, only a small number of anonymized samples have been used. Where human whole genome sequencing has been done, only coverage data has been relevant, and variant calling information, from which personal trait information could have been inferred, has not been derived. The studies have been approved by the local Data protection official for research where this has been required. The same local authority has also considered the studies not to fall within the responsibility of the Regional Committees for medical and health research ethics (REK).

1.9 Selected topic 1: DNA degradation

The DNA molecule is stabilized by the double helix structure but has some weak spots that can be targeted leading to DNA damage. In living cells, the DNA repair mechanisms counterbalance this, but after death, the DNA damage accumulates. Hydrolysis causes depurination and deamination leading to strand breakage and base conversions respectively [148]. Oxidation also causes base and deoxyribose lesions with the same consequences. Different reaction agents cause DNA cross-linking which hinder DNA polymerase extension [149, 150, 151, 152]. DNA is

also damaged by UV-radiation, extreme pH conditions, microbial growth and enzymes as nucleases. The speed of the degradation processes can be reduced by dry state surroundings or low temperature [153].

In living cells DNA degradation is not random [154, 155]. This is due to the nucleosome core particles where DNA is associated with histones. Each nucleosome has an octamer of histone proteins encircled by 147bp of DNA. In this fundamental form, the DNA strand has nucleosomes with linker DNA in between, like "beads on a string" [156]. The positioning of the nucleosomes in a certain region can be static or vary between cells [157]. This "bead on a string" structure is defined as euchromatin and can be further wrapped into higher order structures called heterochromatin. The heterochromatin structure has been shown to give additional protection against DNA damage when studied in vitro [158]. The euchromatin structure is prevalent in the genome with only $\sim 6\%$ having the heterochromatin structure [64]. The chromatin structure is also associated with gene regulation [159]. The "open" euchromatin is found where genes are expressed and the "closed" heterochromatin structure where genes are silenced.

The protective features of the DNA structure have been studied using in vitro conditions or living cells and do not need to apply to DNA in biological trace material. Biological traces contain dried and dead cell material and the DNA has often been influenced by rough environmental conditions. A few forensic studies have been done on nucleosome protection for both STR [160] and SNP markers [161]. For the latter, there was no significant improvement in performance compared to the most robust established forensic SNP multiplex. Ancient DNA is a field related to forensic genetics. During a sequencing study of an old hair sample, it was observed a distinct coverage pattern claimed to be a result of DNA being protected in the nucleosomes [162].

1.10 Selected topic 2: Microbiome

The study of the bacterial world, previously confined to a small minority of species that could be cultivated in a lab, suddenly broke free by the advent of MPS. Now, potentially all bacteria could be detected. The microbiome is defined as all microorganisms in a particular environment, and MPS has been used to study microbiome diversity, shifts in microbiome composition, discover novel organisms and more [163]. Large studies organized by the HMP and EMP consortiums have alone produced enormous amounts of sequence data which is publically available on their respective websites. In addition, there are large data repositories specifically devoted to 16S rRNA gene data such as the Silva database (`https://www.arb-silva.de/`), the Ribosomal Database Project (RDP, `https://rdp.cme.`

msu.edu/) and the Greengenes database (http://greengenes.lbl.gov/).

The healthy human microbiome has been studied extensively. It has been found to vary among individuals, but even more among various body sites or body fluids [164, 165]. For a specific location though, the bacterial composition is relative stable over time [166], or might vary slightly between community state types [167]. Others factors that may influence the bacterial composition is medication, diet and the geographical and ethnical origin of the individual [168].

Microbiomes have mainly been investigated by barcode sequencing as opposed to whole genome sequencing [169]. The prokaryotic 16S rRNA gene has by far been the preferred barcode. This sequence is found in all bacterial species and is roughly 1500bp long which is sufficient for bioinformatics methods. It is evolutionarily preserved but has 9 hypervariable regions designated V1-V9, all with high discrimination power [170]. Due to the limited read lengths of MPS, only a subset of the regions can be selected for sequencing. Regions from between V2-V6 are typically chosen.

Several biases can arise in the laboratory, and the most significant are introduced in the extraction and PCR amplification steps [171]. In the extraction, the main bias is a skewed bacterial composition which occurs as a result of unequal extraction efficiency between different bacteria. However, this can be marginalized by using bead-beating [172, 173, 174, 175, 176]. There is no standard extraction protocol, but MoBio Kits are used by many studies, including the ones organized by HMP (https://hmpdacc.org/) and EMP (http://www.earthmicrobiome.org/). In the PCR amplification step, the biases are more complex. One bias is skewed distributions of PCR products resulting in false bacterial composition. This can be minimized by using high ramp rates between the denaturation and annealing steps, low annealing temperature, and by avoiding longer extension times [177, 178]. Another PCR bias is chimera formation. These are artificial PCR products that form when shorter PCR fragments from aborted amplifications act as primers hybridizing to heterologous fragments in subsequent PCR cycles [179]. If chimeras are not detected and removed in the data handling workflow, they can lead to false detection of novel bacteria species. Digital droplet PCR (ddPCR) is a bias-free alternative to conventional PCR (see Droplet Digital PCR Applications Guide at www.bio- rad.com). Bias is removed by using microdroplets as reaction chambers with just one or a few fragments in each droplet.

The data handling of sequence data is comprehensive [180] (see figure 1.3). First, the sequencing data are prepared for downstream analysis. This involves merging of paired-end reads, quality filtering, removal of indexes and primer sequence and de-multiplexing of reads into individual samples. For each sample, the sequencing data are typically clustered into operational taxonomic units (OTUs) to build a taxonomic profile [181, 182]. This process involves first a de-replication

step where all identical reads are grouped and sorted after abundance. Singletons are removed before the remaining reads are clustered based on typically 97% similarity to find the centroids sequences. The centroid sequences tend to be among the most abundant sequences and minimizes the sum of distances to all other sequences in the cluster. The centroid sequences are filtered for chimeras and all reads are clustered once more towards the centroid sequences using typically 97% similarity. The clusters are defined as OTUs and the assembly of these and their associated read counts make up the taxonomic profile of the sample. The taxonomic profiles for several samples are usually stacked on top of each other into an OTU table for smoother data handling in the downstream analysis.

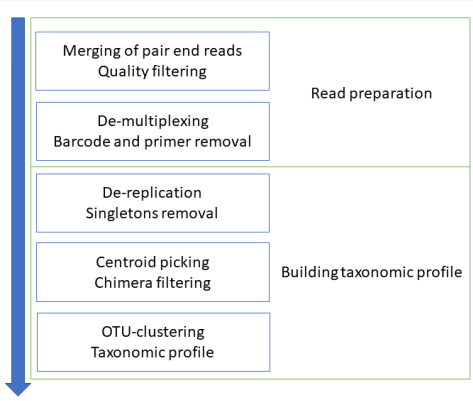

Figure 1.3: Typical microbiome data handling workflow from rawdata to taxonomic profile.

Chapter 2

Paper summaries

2.1 Paper I - Degradation in forensic trace DNA samples explored by massively parallel sequencing

The original idea behind this paper was to investigate if certain DNA regions are more resistant to degradation than others. A forensic panel with markers positioned in such regions could improve analysis performance for degraded DNA and might provide results where none has been obtained using the current panels. The chosen markers for such a panel could be STRs for compatibility towards CE based STR profiles, but SNPs would be preferable as these will maximize performance.

It is known that certain DNA regions in living cells are protected from degrading by being associated with proteins in the nucleosomes and by being present in higher order structures. If this was also the case for degraded trace material, DNA in regions susceptible to degradation should be found to a lesser degree than DNA in regions resistant to degradation. Hence, by whole genome sequencing degraded samples, we should literally be able to detect each fragment present at any genomic region, and the coverage data could be used to map the robustness status at a very high resolution.

For the experimental setup, the DNA was extracted from 25-30 years old blood and semen stains. Based on the conventional forensic STR typing and Bioanalyzer results, 4 samples with different degradation levels were chosen for whole genome sequencing. In order for remove bias, the raw coverage data had to be adjusted for GC effect and filtered for low mappability sequences before the final data interpretation.

In summary, our experimental data indicated that DNA in the different genomic regions degrades at the same rate, and showed no indication of regions more susceptible or resistant towards degradation than others. The lack of regions robust to degradation removes the possibility of using such regions to further improve

analysis performance for degraded DNA. On the other hand, there are no regions susceptible to degradation that should be avoided in this context.

2.2 Paper II - Body fluid prediction from microbial patterns for forensic application

With this study, we evaluated if microbiome sequencing could potentially be used for predicting bacteria-rich body fluids in a forensic setting. As a starting point, we decided to use standard laboratory protocols and data interpretation workflows from the health and environmental related microbiome studies. As a study model, we mimicked a scenario often seen in rape cases where the perpetrator has left saliva on the skin of the victim.

Initially, we had to find a suitable sampling media. The conventional cotton swab had to be abandoned due to the low recovery of bacteria from the cotton for the applied soaking volume. However, forensic tape and synthetic swabs proved to be satisfactory. Saliva from 6 donors were deposited on the hands of 6 individuals (one donor per individual), and from each individual we collected in total 6 different sample types: pure saliva as reference, pure saliva on skin sampled with tape, pure saliva on skin sampled with synthetic swab, diluted saliva on skin sampled with tape, skin only sampled with tape and skin only sampled with synthetic swab. The bacterial DNA of the samples were extracted, amplified by conventional PCR or ddPCR and sequenced. In addition, 2 technical parallels were used for each sample. In total 144 samples were analysed, exclusive the positive and negative controls (36 original samples, 2 parallel PCR techniques, and 2 technical parallels). For the data interpretation workflow, we built OTU based taxonomic profiles, used PCA for visualization and dimension reduction and finally used LDA as model for body fluid prediction.

We also used ddPCR for quantification of bacterial DNA, and the results showed that the content of skin only samples was significantly lower compared to saliva containing samples. This was convenient as our aim was to recognize saliva and not skin. ANOVA showed that type of body site (saliva vs skin) had the main effect on the taxonomic profiles. In addition, there was a weaker significant effect from person (individual 1-6) and no effect from PCR technique (conventional PCR or ddPCR), sampling technique (tape or synthetic swab) or technical replicates (parallel 1 and 2). From a 6-fold cross-validation, we achieved 94% accuracy, but a clear bias was observed between our experimental data and data from the HMP consortium.

In conclusion, the method is not ready for casework, but we showed that the standard laboratory protocols can be applied without large adjustments and that the

data interpretation workflow needs customization in an effort to reduce or remove the observed bias between different datasets.

2.3 Paper III - Optimizing the body fluid recognition from microbial taxonomic profiles

This article was a continuation of the work with the method for body fluid prediction (*paper II*). From that last paper, it was evident that removal of the observed bias between datasets was the key condition for a reasonable calibration effort demand and thus a widespread use of the method. We expected the main solution to lie in the data preparation workflow and therefore decided to customize this rather than to search for the optimal pattern recognition model.

Initially, we settled on a model where PLS was used in combination with LDA. PLS uses supervised learning and should be ideal for pattern recognition. In a cross-validation experiment with HMP data, different combinations of calculation settings were tested in order to optimize accuracy. The optimal combination of these factors was: Taxonomic profiles based on OTUs with 0.98 identity threshold, Aitchisons simplex transform with $C = 1$ pseudo-count and no regularization ($r = 1$) in the PLS step. Accuracy when using these settings was $\sim 98\%$.

In the data preparation workflow, the use of standard OTU based taxonomic profiles was compared to an alternative approach where taxonomic profiles were built from direct assigning of reads to taxa. When using the optimal calculation settings from the cross-validation and training on HMP data and predicting on AGP data, the accuracy collapsed when using standard taxonomic profiles. However, the high accuracy levels from the cross-validation were nearly maintained when the alternative taxonomic profiles were used (accuracy 96%).

The method is still not ready for casework, but by this work, we have taken a promising step toward this aim. Our findings will be implemented in an R-package for microbial forensics that we are currently developing.

Chapter 3

Discussion

The forensic community is currently in the middle of a transition from CE to MPS based methods [183, 62, 20]. The big leap in analysis capacity will improve the routine analysis for identifying individuals, provide extended intelligence information beyond identification and more. Even though earlier research results have been actualized by the introduction of MPS, there is still a big knowledge gap to be filled. MPS will also help speed up new research. This thesis includes contributions in two central areas where MPS has provided new possibilities. Firstly, DNA degradation is the main obstacle for a successful DNA analysis, and to overcome this will provide useful results where no results have been obtained before. Secondly, a reliable method for body fluid recognition to replace the currently used presumptive tests will significantly increase confidence in the evidence evaluation process.

3.1 Improving analysis of degraded trace samples

Earlier studies have shown that nucleosomes and higher order chromatin structures protect the associated DNA from degradation in living cells under different conditions [154, 155, 158]. With this study, we wanted to investigate if this also applies to forensic trace material. If so, a forensic panel targeting markers in the robust regions could improve analysis performance for degraded DNA (*paper I*).

For the experimental setup, we sequenced 2 degraded semen samples and 2 heavily degraded blood samples and downloaded sequence data for 2 undegraded control samples from the 1000 genomes project's webpage. We used the sequencing coverage data adjusted for GC bias and filtered for low mappability regions, as a measurement for DNA concentration in different genetic regions. We postulated that for degraded samples the DNA concentration would correlate with robustness at any given region. After evaluating the coverage data from the genomic level

down to the level of a few bases, our conclusion is that DNA degrades at an even rate throughout the genome. The data do not support the existence of specific regions being more susceptible or resistant towards degradation than others. We were therefore not able to fulfil our ambition of using markers positioned in robust regions to customize a superior performing panel for degraded trace samples. However, our findings are highly relevant in the forensic community's effort to develop new MPS based SNP applications to improve analysis performance for degraded DNA.

At the genome level, the variation of coverage was mainly random and not largely dependent on external factors. The coverage also showed uniform and symmetric density distribution which ruled out that large or many genomic regions had extreme levels of coverage. At the chromosomal level, the coverage was similar between the chromosomes, even for the heterochromatic inactivated female X chromosome. This was a major finding, as the coverage for this chromosome should have been significantly higher compared to the others if the heterochromatin structure protected against DNA degradation. The comparison of condensed heterochromatic regions (H3K9me3 sites) and open euchromatic regions (promoter sites) showed no difference in coverage levels either. At the base pair level, regions of strongly-positioned nucleosomes did not show a repetitive pattern as expected if nucleosome-associated DNA was protected against degradation. From a visual inspection, the variation of coverage at this level seemed random.

The idea to use coverage as a measurement of the amount of DNA present in different genetic regions is not new. It is a standard method in copy number variation (CNV) research and has been shown to give usable results even for shallow whole genome sequencing ($\sim$ 0.1x coverage) [184]. It is also established that the coverage data are mainly affected by GC bias and low mappability regions and should be corrected for these [185, 186]. Scheinin et al [184] performed the two corrections simultaneous as the formalin-fixed paraffin-embedded (FFPE) samples produced poor quality data. However, as this is a special case, we decided to follow the general approach where the corrections are done independently. Another potential bias is non-random DNA shearing in the library preparation step [187]. However, as we did not use size selection and each fragment in the library was completely sequenced this should not be a significant bias in our case.

The 4 degraded samples included in the study where sequenced in the same batch to produce similar numbers of reads per sample. However, we observed a difference in the final coverage levels between the degraded ($\sim$ 3x coverage) and the heavily degraded ($\sim$ 0.5x coverage) samples. The samples in both categories lost similar amounts of read sequence during adapter trimming and merging of the paired-end reads. This can mainly be explained by the short fragment size in the library. However, the heavily degraded samples lost half of the reads in the map-

ping to the reference genome, and then half of the mapped reads in the subsequent quality filtering steps. The degraded samples lost only a fraction of the reads during this part of the workflow. This difference is striking and might be caused by a larger proportion of non-human DNA in the heavily degraded samples. For ancient DNA, post-mortem DNA damage will complicate mapping [151, 188], but we regard the level of such damage to be marginal in our case due to the different storage conditions. The low coverage level for the heavily degraded samples should still be fit for purpose as even lower levels have been used for CNV studies [184]. As the final coverage data was evenly distributed in the genome, there is also no reason to claim that this loss of reads introduced skewness. We also regard the 2 undegraded control samples suitable for purpose, even if these were not sequenced in the same batch as the degraded samples. This is supported by these samples being sequenced using the same sequencing technology and with the same coverage level as we used for the degraded samples. By being fresh and optimal for sequencing, it is also reasonable that the control samples had the highest final coverage level ($\sim$ 5x coverage).

It is also not necessarily straightforward to compare our experimental samples with ancient DNA samples. After sequencing hair shafts of a 4000-yr-old Paleo-Eskimo, Pedersen et al found coverage patterns corresponding to expected nucleosome positions, and this was interpreted as a consequence of nucleosome shielding of the DNA strand [162]. However, ancient DNA has been exposed to significantly different conditions than the typically biological trace sample. In addition to the obvious differences in timespan and environment conditions, ancient DNA samples are often taken from human or animal remains that have gone through post-mortem apoptosis [189]. Apoptosis is also a part of the hair differentiation process [190]. As mentioned before, the nucleosome-associated DNA escapes enzymatic degradation in apoptosis. Biological trace material will presumably not be affected by this process because of immediate temperature drop after deposition that will lower DNase activity [191] and a short drying period which will remove nearly all water and stop the biological reactions. Hence, Pedersen's findings for ancient DNA does not need to be contradictory to our conclusion for forensic traces. If the observed shielding effect has occurred during apoptosis only, Pedersen's findings could still fit our proposition that the nucleosome shielding effect is removed by the possible dissociation of DNA and histones due to denaturation in a near water-free environment.

As highly repetitive sequence in the centromeres and telomeres have not yet been sequenced [64], our conclusion does not apply to these regions. Most centromeres and telomeres regions have highly condensed heterochromatic structure, but with no shielding effect observed in the other genomic regions, it is doubtful that this effect should be present exclusively in these regions. As the forensic markers also

are positioned elsewhere in the genome, it is difficult to see at present that these regions will have any forensic relevance.

Our conclusion is based on sequence data for a limited amount of degraded samples. Even though these 4 samples were chosen to cover a large span in degradation level, they might not be representative for all samples. Some samples will have degradation levels beyond the experimental samples. A significant proportion of samples have been exposed to outdoor conditions, whereas the experimental samples were stored indoors. There will probably also be special cases, for example of active apoptosis, which, as mentioned above, might prevent degradation of nucleosome-associated DNA. Although it is clear that a larger sample set is needed to extensively map the variety of degraded trace samples, we regard our study provides valuable information on the degradation process, and that our conclusion will apply to a large proportion of forensically relevant trace samples.

3.2 Body fluid prediction from microbial composition patterns

Our aim with this study was to provide a reliable method for body fluids recognition to replace the currently used presumptive tests and to be a supplement for the mRNA based methods. The developmental process was divided into identifying a robust laboratory protocol and to make a suitable data interpretation workflow. To make such a comprehensive process efficient the problem had to be divided into several steps. As a starting point we used a conventional laboratory protocol for microbiome sequencing and a standard data interpretation workflow [180] (*paper II*). The experimental setup was designed to evaluate the most critical factors in the laboratory protocol. We used a real case scenario where saliva was deposited on skin as a study model. In addition, we stress tested the data interpretation workflow by training and testing on different datasets. Based on the results of this initial experiment the second development step became customizing the data interpretation workflow to optimize performance in a real case scenario (*paper III*).

First of all, we had to find the appropriate sampling method (*paper II*). The standard cotton swab failed to release enough of the sampled bacteria to the soaking solution. This was probably caused by the low liquid volume applied (200μL). However, synthetic swabs and tape proved to be suitable as sampling media. That the cotton swabs had to be abundant was disappointing as these are commonly used in forensics, but both synthetic swabs and tape have been shown to perform equally as good in a forensic setting [192, 193].

To evaluate the laboratory protocol, ANOVA was used to identify the significant experimental factors influencing the taxonomic profiles. As the taxonomic

profiles are multi-dimensional with many OTUs, PCA was used to project each profile into the first, second and third principal axis (*paper II*). The score values of each of these dimensions were sequentially used as the dependent variable in the ANOVA. In conclusion, type of body site (saliva vs skin) had the main significant effect, with person having a less, but still significant effect. This was expected as it is known that eg diet can impact the personal microbial composition [168]. The effect was strongest for skin samples, and to neutralize the effect it might be beneficial to collect a reference sample from the skin nearby the sampling site. The ANOVA results showed no significant effect on the taxonomic profiles from PCR technique (conventional PCR or ddPCR), sampling technique (tape or synthetic swab) or technical replicates (parallel 1 and 2). The two latter results are reassuring as these state the stability of the sampling and laboratory protocol, but the lack of effect from the PCR technique was surprising. Beforehand we viewed the PCR step to be the main critical factor in the laboratory due to several potential PCR biases. To investigate this further we used ANOVA to test if PCR had effect on chimera formation and microbial diversity, but still, no significant effects were found. However, ddPCR tended to have a higher correlation between numbers of reads and initial DNA inputs and more reproducible results with less variation between technical replicates. This near trouble-free PCR amplification might be explained by a dilution effect leading to lower probability for artefact formation. This is encouraging as most forensic samples are low level.

Our initial data interpretation workflow used PCA in combination with LDA *paper II*. This model was based on a standard workflow for microbiome analysis where unsupervised learning is typically used to support explorative studies [180]. We also used the standard approach to make OTU based taxonomic profiles. A cross-validation of the experimental data gave an accuracy of 94%. However, when mimicking a real case scenario by training on the HMP dataset and testing on the experimental data, a bias was evident for both skin and saliva samples, and the accuracy collapsed as a consequence. The skin samples showed a lower degree of similarity than the saliva samples between the two datasets. One obvious explanation for this was that the HMP skin samples were taken from other body sites than our samples (elbow cleavage and behind the ear compared to between fingers). These body sites are known to have different microbial compositions [194]. Another explanation could be that the HMP and our samples were collected at different geographical locations [165]. Factors such as extraction, library preparation and sequencing also probably contributed to the bias, but if so, the bias would equally influence the skin and saliva samples. The observed bias between the two datasets could have been overcome by strict standardization of the methods used by different laboratories or by internal calibration in each laboratory, but both these alternatives would have been inefficient and hindered a wide-spread use of

the method. Without being able to identify critical steps with great improvement potential in the laboratory protocol, we considered a larger part of the solution to lie in a customized data handling workflow optimized for pattern recognition instead of exploration. We pursued this in the next step of the method development process.

Intuitively, a supervised model would fit our problem of pattern recognition best. However, instead of searching for the optimal supervised model, we chose to focus on one model and to find the calculation settings that would optimize performance (*paper III*). We chose a model based on a combination of PLS and LDA to investigate this closer, and still used the OTU based taxonomic profiles. In a cross-validation using the HMP dataset, we obtained an optimal accuracy of $\sim 98\%$ by using the following calculation settings: taxonomic profiles based on OTUs with 0.98 identity threshold (OTU98), Aitchisons simplex transform with $C = 1$ pseudo-count and no regularization ($r = 1$) in the PLS step. As for the initial model, the accuracy collapsed when testing on a foreign dataset (training on HMP data and testing on AGP data). The collapse was a consequence of overfitting the model. A probable explanation was that when the OTU centroid sequences from the training data were used to build taxonomic profiles for the test data, a skewness was introduced in the profiles as a proportion of the reads were categorized into wrong OTUs. This skewness would be worse for higher resolution. In an attempt to remove the bias we assigned reads in both the training and test datasets directly to genus when building the taxonomic profiles. The same database of reference reads was used for both datasets, and by introducing this common anchor point we nearly managed to keep the high accuracy levels from the cross-validation ($\sim 96\%$ accuracy). In the process, we used a lower taxonomic resolution than what was found optimal during cross-validation (1640 different genera compared to ~ 25000 OTU98 bins), and this was a consequence of genus being the highest taxonomic resolution supported by the applied taxMachine tool [195]. We will evaluate accuracy at higher taxonomic resolutions when this is supported by a new version of taxMachine. The other optimal calculation settings from the cross-validation were still optimal. Other advantages of assigning read directly to taxa are that it is magnitudes faster than any OTU finding workflow and the training and test datasets no longer have to represent overlapping regions of the 16S gene.

The performance of the customized model was evaluated beyond accuracy (*paper III*). The sensitivity is defined as the proportion of positives that are correctly identified. From the cross-validation, the fecal, oral and vaginal samples all had sensitivities ≥ 0.99. For the skin and nasal samples, the sensitivity was 0.97 and 0.84 respectively. Most of the misclassified nasal samples were predicted as skin samples, probably as a result of contamination from skin microbiota surrounding the nostrils. Specificity is defined as the proportion of negatives that are correctly

identified, and this is the most important quality parameter in a forensic context as a false positive result can lead to an incorrect verdict. The specificities were 0.99 for the vaginal, oral, fecal and nasal samples and 0.98 for the skin samples. When training on HMP samples and testing on AGP samples and using direct assigned taxonomic profiles the sensitivities were 0.94 for the oral samples, 0.88 for the skin samples and 0.97 for the fecal samples. The respective specificity values were 0.99, 0.99 and 0.98. The numbers of nasal and vaginal samples in the APG data were very low and thus no sensitivity and specificity values could be calculated for these. We will repeat the sensitivity and specificity calculation at a higher resolution taxonomy when this is supported by a new version of taxMachine.

To get an estimate of the limit of detection (LOD) we evaluated if the number of reads had effect on the taxonomic profiles (*paper II*). However, the ANOVA showed no significance for such an effect. Of the 10 samples with the lowest number of reads, only 2 samples were classified incorrectly and 1 of these had below 100 reads. However, to determine the final LOD a larger sample set is needed.

The skin samples were included in this study as many forensic traces with body fluids are sampled from skin or from objects where skin microbiota can be present. Hence, we wanted to study what influence skin microbiota would have on data interpretation. First, we investigated the influence on mixtures, and found that skin microbiota only contributed to a small proportion of the total bacterial DNA in the saliva samples taken from skin (*paper II*). It is reasonable to believe that the same quantity ratio will apply to other bacteria-rich body fluids such as vaginal secretion, even though this has to be supported by empirical studies [196]. Our results showed that the low-level contribution from skin microbiota only had a minor impact on accuracy, but that accuracy could be further increased by controlling the mentioned small significant effect from person on the taxonomic profiles. We continued our investigation of skin samples by studying the large proportion of pure skin samples in the HMP dataset (*paper III*). The skin samples had a relatively low sensitivity at 0.97 and were confused with all other types of body fluids. This finding is not surprising since it is known that skin samples have a large variation in bacterial composition [197, 198, 165], but underline the importance of having security mechanisms in the final data interpretation tool to avoid false conclusions. Such a mechanism could be eg outlier detection.

We investigated if variable selection would remove noise and thereby improve method performance, but variable selection had little impact on accuracy (*paper III*). Another motivation to study variable selection was to identify important bacteria that could be potentially be included in a PCR multiplex for a fast and cheap qPCR based application. However, considered the low impact of variable selection on accuracy, the additional effort needed to define a stable panel of selected

bacteria and the decreasing sequencing costs, we would not invest much effort into developing a strong variable selection for the purpose of PCR multiplexing.

3.2.1 Future perspectives

It is challenging to predict body fluids at the low detection limits and the high accuracy level demanded by a forensic method. Our method is still not ready for casework, but the results are promising, and we see a clear roadmap leading to a final method ready for casework.

To increase accuracy further the use of taxonomic profiles with higher resolution than genus should be tested. This needs the support from a new version of the taxMachine tool. In addition, finding the optimal model based on supervised learning could contribute to higher accuracy, but it is unclear how much this will increase an already high accuracy.

It is also evident that the method needs support from a few additional features that must be developed. As a large proportion of trace samples are mixtures, it will be essential to have a deconvolution tool to separate different microbiota. A tool to detect outliers will help remove samples with taxonomic profiles that do not resemble any of the known body fluid patterns. These samples could for example be detected from the posterior probabilities and end up in an unclassified category. Alternatively, there are statistical tools designed to detect outliers. The final method should also support the use of reference samples from skin nearby the sampling site as these samples would help to remove the observed effect from person on especially the skin taxonomic profiles. The final method should also be validated by an inter-laboratory collaborative study.

The future general use of MPS in forensics might consist of a basic panel for identifying trace material and then special additional panels, dependent on case circumstances. If information on body fluid should be important for a specific sample, it is reasonable to propose a combined use of mRNA and microbiome analysis to optimize accuracy. This will demand the use of different sample preparation protocols, but the resulting sequencing libraries can be pooled and sequenced together. Although sample consumption always should be kept to a minimum in forensics, using extra aliquotes of sample in a rational fashion could be defended.

Beside body fluid prediction, there are also other research possibilities within microbial forensics. To the author's knowledge little is known of bacterial growth in deposited biological trace material, and novel knowledge might lead to a method for time since deposition assessments. As mentioned above, there is also some research activity of microbial activity in connection with decomposition of cadavers. These studies aim to develop a more reliant method for post-mortem interval estimations and a screening tool for lethal infections detection.

Chapter 4

Conclusion

MPS has created many new possibilities in forensics, both in research and eventually in routine casework. The work with this thesis has resulted in contributions in two selected fields.

Our first study was the investigation of DNA degradation in biological trace material by using coverage data from whole genome sequencing. Although the sample size of 4 samples was limited, our results supported that DNA degrades at an even rate throughout the genome and that there are no specific genomic regions more susceptible or resistant to degradation than others. This put a stop to our initial ambition of using robust DNA regions to develop a superior performing method for analysing degraded trace samples. However, our conclusion is still highly relevant information in the forensic community's effort to develop new MPS based methods.

The other contribution was a novel method for predicting bacteria-rich body fluids based on microbial composition patterns. With this study, we introduced a promising new tool to solve a very complex forensic problem. Our proposed method follows a standard laboratory protocol for microbiome sequencing and then uses a PLS/LDA based data handling workflow where the taxonomic profiles are built from direct assignment of reads to taxa. By using optimal calculation settings when training and testing on different datasets, we obtained an accuracy of $\sim 96\%$. Although this is a promising result, the method is not ready for casework. Until then, there is still potential for accuracy optimization, a mixture deconvolution tool needs to be developed and an inter-laboratory validation study needs to be performed. We see the final method as a robust method without the need of excessive intra-laboratory validation efforts, and as a valuable supplement to the mRNA based method for a highly reliable forensic body fluid prediction.

Bibliography

[1] Crick, F. H. “On protein synthesis”. In: *Symp. Soc. Exp. Biol.* 12 (1958), pp. 138–163.

[2] Lander, E. S. et al. “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822 (Feb. 2001), pp. 860–921. DOI: 10.1038/35057062.

[3] Pennisi, E. “ENCODE Project Writes Eulogy for Junk DNA”. In: *Science* 337.6099 (7th Sept. 2012), pp. 1159–1161. DOI: 10.1126/science.337.6099.1159.

[4] Doolittle, W. F. “Is junk DNA bunk? A critique of ENCODE”. In: *PNAS* 110.14 (2nd Apr. 2013), pp. 5294–5300. DOI: 10.1073/pnas.1221376110.

[5] 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. and McVean, G. A. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (Nov. 2012), pp. 56–65. DOI: 10.1038/nature11632.

[6] Ellegren, H., Smith, N. G. C. and Webster, M. T. “Mutation rate variation in the mammalian genome”. In: *Curr. Opin. Genet. Dev.* 13.6 (Dec. 2003), pp. 562–568.

[7] Butler, J. M. *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, 2011. 700 pp.

[8] Meneely, P., Hoang, R. D., Okeke, I. N. and Heston, K. *Genetics: Genes, Genomes, and Evolution*. Oxford University Press, 2017. 775 pp.

[9] Jeffreys, A. J., Wilson, V. and Thein, S. L. “Hypervariable ’minisatellite’ regions in human DNA”. In: *Nature* 314.6006 (Mar. 1985), pp. 67–73.

[10] Budowle, B. *DNA Typing Protocols: Molecular Biology and Forensic Analysis*. Eaton Pub., 2000. 304 pp.

[11] Puers, C., Hammond, H. A., Jin, L., Caskey, C. T. and Schumm, J. W. "Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder." In: *Am J Hum Genet* 53.4 (Oct. 1993), pp. 953–958.

[12] Gill, P., Kimpton, C., D'Aloja, E., Andersen, J. F., Bar, W., Brinkmann, B., Holgersson, S., Johnsson, V., Kloosterman, A. D. and Lareu, M. V. "Report of the European DNA profiling group (EDNAP)–towards standardisation of short tandem repeat (STR) loci". In: *Forensic Sci. Int.* 65.1 (Mar. 1994), pp. 51–59.

[13] Urquhart, A., Kimpton, C. P., Downes, T. J. and Gill, P. "Variation in short tandem repeat sequences–a survey of twelve microsatellite loci for use as forensic identification markers". In: *Int. J. Legal Med.* 107.1 (1994), pp. 13–20.

[14] Butler, J. M. *Fundamentals of Forensic DNA Typing*. Academic Press, Sept. 2009. 519 pp.

[15] Gill, P. "An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes". In: *Int. J. Legal Med.* 114.4 (2001), pp. 204–210.

[16] Pitterl, F., Schmidt, K., Huber, G., Zimmermann, B., Delport, R., Amory, S., Ludes, B., Oberacher, H. and Parson, W. "Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations". In: *Int J Legal Med* 124.6 (Nov. 2010), pp. 551–558. DOI: 10.1007/s00414-009-0408-x.

[17] Kline, M. C., Hill, C. R., Decker, A. E. and Butler, J. M. "STR sequence analysis for characterizing normal, variant, and null alleles". In: *Forensic Sci Int Genet* 5.4 (Aug. 2011), pp. 329–332. DOI: 10.1016/j.fsigen.2010.09.005.

[18] Rockenbauer, E., Hansen, S., Mikkelsen, M., Børsting, C. and Morling, N. "Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing". In: *Forensic Sci Int Genet* 8.1 (Jan. 2014), pp. 68–72. DOI: 10.1016/j.fsigen.2013.06.011.

[19] Gelardi, C., Rockenbauer, E., Dalsgaard, S., Børsting, C. and Morling, N. "Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles". In: *Forensic Sci Int Genet* 12 (Supplement C Sept. 2014), pp. 38–41. DOI: 10.1016/j.fsigen.2014.04.016.

[20] Gettings, K. B., Aponte, R. A., Vallone, P. M. and Butler, J. M. "STR allele sequence variation: Current knowledge and future issues". In: *Forensic Sci Int Genet* 18 (Sept. 2015), pp. 118–130. DOI: 10.1016/j.fsigen.2015.06.005.

[21] Tucker, V. C., Hopwood, A. J., Sprecher, C. J., McLaren, R. S., Rabbach, D. R., Ensenberger, M. G., Thompson, J. M. and Storts, D. R. "Developmental validation of the PowerPlex® ESX 16 and PowerPlex® ESX 17 Systems". In: *Forensic Sci Int Genet* 6.1 (Jan. 2012), pp. 124–131. DOI: 10.1016/j.fsigen.2011.03.009.

[22] Green, R. L., Lagacé, R. E., Oldroyd, N. J., Hennessy, L. K. and Mulero, J. J. "Developmental validation of the AmpFlSTR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus". In: *Forensic Sci Int Genet* 7.1 (Jan. 2013), pp. 41–51. DOI: 10.1016/j.fsigen.2012.05.012.

[23] Kayser, M. and Knijff, P. de. "Improving human forensics through advances in genetics, genomics and molecular biology". In: *Nat Rev Genet* 12.3 (Mar. 2011), pp. 179–192. DOI: 10.1038/nrg2952.

[24] Oorschot, R. A. van, Ballantyne, K. N. and Mitchell, R. J. "Forensic trace DNA: a review". In: *Investig Genet* 1.1 (Dec. 2010), p. 14. DOI: 10.1186/2041-2223-1-14.

[25] Goray, M., Eken, E., Mitchell, R. J. and Oorschot, R. A. H. van. "Secondary DNA transfer of biological substances under varying test conditions". In: *Forensic Sci Int Genet* 4.2 (Feb. 2010), pp. 62–67. DOI: 10.1016/j.fsigen.2009.05.001.

[26] Fonneløp, A. E., Egeland, T. and Gill, P. "Secondary and subsequent DNA transfer during criminal investigation". In: *Forensic Sci Int Genet* 17 (July 2015), pp. 155–162. DOI: 10.1016/j.fsigen.2015.05.009.

[27] Sijen, T. "Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers". In: *Forensic Sci Int Genet* 18 (Sept. 2015), pp. 21–32. DOI: 10.1016/j.fsigen.2014.11.015.

[28] Harbison, S. and Fleming, R. "Forensic body fluid identification: state of the art". In: *Research and Reports in Forensic Medical Science* (Feb. 2016), p. 11. DOI: 10.2147/RRFMS.S57994.

[29] Haas, C., Hanson, E., Anjos, M. J., Bär, W., Banemann, R., Berti, A., Borges, E., Bouakaze, C., Carracedo, A., Carvalho, M., Castella, V., Choma, A., Cock, G. D., Dötsch, M., Hoff-Olsen, P., Johansen, P., Kohlmeier, F., Lindenbergh, P. A., Ludes, B., Maroñas, O., Moore, D., Morerod, M.-L., Morling, N., Niederstätter, H., Noel, F., Parson, W., Patel, G., Popielarz, C., Salata, E., Schneider, P. M., Sijen, T., Sviežena, B., Turanská, M., Zatkalíková, L. and Ballantyne, J. "RNA/DNA co-analysis from blood stains - Results of a second collaborative EDNAP exercise". In: *Forensic Science International: Genetics* 6.1 (1st Jan. 2012), pp. 70–80. DOI: 10.1016/j.fsigen.2011.02.004.

[30] Haas, C., Hanson, E., Anjos, M. J., Banemann, R., Berti, A., Borges, E., Carracedo, A., Carvalho, M., Courts, C., Cock, G. D., Dötsch, M., Flynn, S., Gomes, I., Hollard, C., Hjort, B., Hoff-Olsen, P., Hríbiková, K., Lindenbergh, A., Ludes, B., Maroñas, O., McCallum, N., Moore, D., Morling, N., Niederstätter, H., Noel, F., Parson, W., Popielarz, C., Rapone, C., Roeder, A. D., Ruiz, Y., Sauer, E., Schneider, P. M., Sijen, T., Court, D. S., Sviežená, B., Turanská, M., Vidaki, A., Zatkalíková, L. and Ballantyne, J. "RNA/DNA co-analysis from human saliva and semen stains - Results of a third collaborative EDNAP exercise". In: *Forensic Science International: Genetics* 7.2 (1st Feb. 2013), pp. 230–239. DOI: 10.1016/j.fsigen.2012.10.011.

[31] Haas, C., Hanson, E., Anjos, M. J., Ballantyne, K. N., Banemann, R., Bhoelai, B., Borges, E., Carvalho, M., Courts, C., Cock, G. D., Drobnic, K., Dötsch, M., Fleming, R., Franchi, C., Gomes, I., Hadzic, G., Harbison, S. A., Harteveld, J., Hjort, B., Hollard, C., Hoff-Olsen, P., Hüls, C., Keyser, C., Maroñas, O., McCallum, N., Moore, D., Morling, N., Niederstätter, H., Noël, F., Parson, W., Phillips, C., Popielarz, C., Roeder, A. D., Salvaderi, L., Sauer, E., Schneider, P. M., Shanthan, G., Court, D. S., Turanská, M., Oorschot, R. A. H. v., Vennemann, M., Vidaki, A., Zatkalíková, L. and Ballantyne, J. "RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: Results of a fourth and fifth collaborative EDNAP exercise". In: *Forensic Science International: Genetics* 8.1 (1st Jan. 2014), pp. 203–212. DOI: 10.1016/j.fsigen.2013.09.009.

[32] Haas, C., Hanson, E., Banemann, R., Bento, A. M., Berti, A., Carracedo, Á., Courts, C., De Cock, G., Drobnic, K., Fleming, R., Franchi, C., Gomes, I., Hadzic, G., Harbison, S. A., Hjort, B., Hollard, C., Hoff-Olsen, P., Keyser, C., Kondili, A., Maroñas, O., McCallum, N., Miniati, P., Morling, N., Niederstätter, H., Noël, F., Parson, W., Porto, M. J., Roeder, A. D., Sauer, E., Schneider, P. M., Shanthan, G., Sijen, T., Syndercombe Court, D., Turanská, M., Berge, M. van den, Vennemann, M., Vidaki, A., Zatkalíková, L.

and Ballantyne, J. “RNA/DNA co-analysis from human skin and contact traces - Results of a sixth collaborative EDNAP exercise”. In: *Forensic Sci Int Genet* 16 (May 2015), pp. 139–147. DOI: 10.1016/j.fsigen.2015.01.002.

[33] Walsh, S., Liu, F., Wollstein, A., Kovatsi, L., Ralf, A., Kosiniak-Kamysz, A., Branicki, W. and Kayser, M. “The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA”. In: *Forensic Sci Int Genet* 7.1 (Jan. 2013), pp. 98–115. DOI: 10.1016/j.fsigen.2012.07.005.

[34] Kidd, K. K., Speed, W. C., Pakstis, A. J., Furtado, M. R., Fang, R., Madbouly, A., Maiers, M., Middha, M., Friedlaender, F. R. and Kidd, J. R. “Progress toward an efficient panel of SNPs for ancestry inference”. In: *Forensic Sci Int Genet* 10 (May 2014), pp. 23–32. DOI: 10.1016/j.fsigen.2014.01.002.

[35] Phillips, C. “Forensic genetic analysis of bio-geographical ancestry”. In: *Forensic Sci Int Genet* 18 (Sept. 2015), pp. 49–65. DOI: 10.1016/j.fsigen.2015.05.012.

[36] Bocklandt, S., Lin, W., Sehl, M. E., Sánchez, F. J., Sinsheimer, J. S., Horvath, S. and Vilain, E. “Epigenetic predictor of age”. In: *PLoS ONE* 6.6 (2011), e14821. DOI: 10.1371/journal.pone.0014821.

[37] Horvath, S. “DNA methylation age of human tissues and cell types”. In: *Genome Biol.* 14.10 (2013), R115. DOI: 10.1186/gb-2013-14-10-r115.

[38] Weber-Lehmann, J., Schilling, E., Gradl, G., Richter, D. C., Wiehler, J. and Rolf, B. “Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing”. In: *Forensic Sci Int Genet* 9 (Mar. 2014), pp. 42–46. DOI: 10.1016/j.fsigen.2013.10.015.

[39] Li, C., Zhao, S., Zhang, N., Zhang, S. and Hou, Y. “Differences of DNA methylation profiles between monozygotic twins’ blood samples”. In: *Mol. Biol. Rep.* 40.9 (Sept. 2013), pp. 5275–5280. DOI: 10.1007/s11033-013-2627-y.

[40] Vidaki, A., Díez López, C., Carnero-Montoro, E., Ralf, A., Ward, K., Spector, T., Bell, J. T. and Kayser, M. “Epigenetic discrimination of identical twins from blood under the forensic scenario”. In: *Forensic Sci Int Genet* 31 (Aug. 2017), pp. 67–80. DOI: 10.1016/j.fsigen.2017.07.014.

[41] Lahrouchi, N., Behr, E. R. and Bezzina, C. R. “Next-Generation Sequencing in Post-mortem Genetic Testing of Young Sudden Cardiac Death Cases”. In: *Front Cardiovasc Med* 3 (May 2016). DOI: 10.3389/fcvm.2016.00013.

[42] MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., Leal, S. M., Pennacchio, L. A., Stamatoyannopoulos, J. A., Sunyaev, S. R., Valle, D., Voight, B. F., Winckler, W. and Gunter, C. “Guidelines for investigating causality of sequence variants in human disease”. In: *Nature* 508.7497 (Apr. 2014), pp. 469–476. DOI: 10.1038/nature13127.

[43] Frey, K. G., Herrera-Galeano, J. E., Redden, C. L., Luu, T. V., Servetas, S. L., Mateczun, A. J., Mokashi, V. P. and Bishop-Lilly, K. A. “Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood”. In: *BMC Genomics* 15 (Feb. 2014), p. 96. DOI: 10.1186/1471-2164-15-96.

[44] Visser, L. E., Schaik, R. H. N. van, Vliet, M. van, Trienekens, P. H., De Smet, P. A. G. M., Vulto, A. G., Hofman, A., Duijn, C. M. van and Stricker, B. H. C. “Allelic variants of cytochrome P450 2C9 modify the interaction between nonsteroidal anti-inflammatory drugs and coumarin anticoagulants”. In: *Clin. Pharmacol. Ther.* 77.6 (June 2005), pp. 479–485. DOI: 10.1016/j.clpt.2005.02.009.

[45] Sullivan, D., Pinsonneault, J. K., Papp, A. C., Zhu, H., Lemeshow, S., Mash, D. C. and Sadee, W. “Dopamine transporter DAT and receptor DRD2 variants affect risk of lethal cocaine abuse: a gene-gene-environment interaction”. In: *Transl Psychiatry* 3 (Jan. 2013), e222. DOI: 10.1038/tp.2012.146.

[46] Johnson, H. R., Trinidad, D. D., Guzman, S., Khan, Z., Parziale, J. V., DeBruyn, J. M. and Lents, N. H. “A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval”. In: *PLoS ONE* 11.12 (2016), e0167370. DOI: 10.1371/journal.pone.0167370.

[47] Hauther, K. A., Cobaugh, K. L., Jantz, L. M., Sparer, T. E. and DeBruyn, J. M. “Estimating Time Since Death from Postmortem Human Gut Microbial Communities”. In: *J. Forensic Sci.* 60.5 (Sept. 2015), pp. 1234–1240. DOI: 10.1111/1556-4029.12828.

[48] DeBruyn, J. M. and Hauther, K. A. “Postmortem succession of gut microbial communities in deceased human subjects”. In: *PeerJ* 5 (June 2017). DOI: 10.7717/peerj.3437.

[49] Metcalf, J. L., Wegener Parfrey, L., Gonzalez, A., Lauber, C. L., Knights, D., Ackermann, G., Humphrey, G. C., Gebert, M. J., Van Treuren, W., Berg-Lyons, D., Keepers, K., Guo, Y., Bullard, J., Fierer, N., Carter, D. O. and Knight, R. “A microbial clock provides an accurate estimate of the postmortem interval in a mouse model system”. In: *eLife* 2 (Oct. 2013). DOI: 10.7554/eLife.01104.

[50] Lax, S., Hampton-Marcell, J. T., Gibbons, S. M., Colares, G. B., Smith, D., Eisen, J. A. and Gilbert, J. A. “Forensic analysis of the microbiome of phones and shoes”. In: *Microbiome* 3 (2015), p. 21. DOI: 10.1186/s40168-015-0082-9.

[51] Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K. and Knight, R. “Forensic identification using skin bacterial communities”. In: *Proc. Natl. Acad. Sci. U.S.A.* 107.14 (Apr. 2010), pp. 6477–6481. DOI: 10.1073/pnas.1000162107.

[52] Schmedes, S. E., Woerner, A. E. and Budowle, B. “Forensic human identification using skin microbiomes”. In: *Appl. Environ. Microbiol.* (Sept. 2017). DOI: 10.1128/AEM.01672-17.

[53] Schmedes, S. E., Woerner, A. E., Novroski, N. M. M., Wendt, F. R., King, J. L., Stephens, K. M. and Budowle, B. “Targeted sequencing of clade-specific markers from skin microbiomes for forensic human identification”. In: *Forensic Sci Int Genet* 32 (Oct. 2017), pp. 50–61. DOI: 10.1016/j.fsigen.2017.10.004.

[54] Ross, A. A., Doxey, A. C. and Neufeld, J. D. “The Skin Microbiome of Cohabiting Couples”. In: *mSystems* 2.4 (Aug. 2017). DOI: 10.1128/mSystems.00043-17.

[55] Budowle, B., Schutzer, S. E., Einseln, A., Kelley, L. C., Walsh, A. C., Smith, J. A. L., Marrone, B. L., Robertson, J. and Campos, J. “Public health. Building microbial forensics as a response to bioterrorism”. In: *Science* 301.5641 (Sept. 2003), pp. 1852–1853. DOI: 10.1126/science.1090083.

[56] Schmedes, S. E., Sajantila, A. and Budowle, B. “Expansion of Microbial Forensics”. In: *J. Clin. Microbiol.* 54.8 (Jan. 2016), pp. 1964–1974. DOI: 10.1128/JCM.00046-16.

[57] Scaduto, D. I., Brown, J. M., Haaland, W. C., Zwickl, D. J., Hillis, D. M. and Metzker, M. L. "Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences". In: *PNAS* 107.50 (Dec. 2010), pp. 21242–21247. DOI: 10.1073/pnas.1015673107.

[58] González-Candelas, F., Bracho, M. A., Wróbel, B. and Moya, A. "Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source". In: *BMC Biol.* 11 (July 2013), p. 76. DOI: 10.1186/1741-7007-11-76.

[59] Hebert, P. D. N., Ratnasingham, S. and deWaard, J. R. "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species." In: *Proc Biol Sci* 270 (Suppl 1 Aug. 2003), S96–S99. DOI: 10.1098/rsbl.2003.0025.

[60] Johnson, R. N., Wilson-Wilde, L. and Linacre, A. "Current and future directions of DNA in wildlife forensic science". In: *Forensic Sci Int Genet* 10 (May 2014), pp. 1–11. DOI: 10.1016/j.fsigen.2013.12.007.

[61] Roy, S., Tyagi, A., Shukla, V., Kumar, A., Singh, U. M., Chaudhary, L. B., Datt, B., Bag, S. K., Singh, P. K., Nair, N. K., Husain, T. and Tuli, R. "Universal Plant DNA Barcode Loci May Not Work in Complex Groups: A Case Study with Indian Berberis Species". In: *PLOS ONE* 5.10 (2010), e13674. DOI: 10.1371/journal.pone.0013674.

[62] Børsting, C. and Morling, N. "Next generation sequencing and its applications in forensic genetics". In: *Forensic Sci Int Genet* 18 (Sept. 2015), pp. 78–89. DOI: 10.1016/j.fsigen.2015.02.002.

[63] Sanger, F., Nicklen, S. and Coulson, A. R. "DNA sequencing with chain-terminating inhibitors". In: *Proc Natl Acad Sci U S A* 74.12 (Dec. 1977), pp. 5463–5467.

[64] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (Oct. 2004), pp. 931–945. DOI: 10.1038/nature03001.

[65] Goodwin, S., McPherson, J. D. and McCombie, W. R. "Coming of age: ten years of next-generation sequencing technologies". In: *Nat. Rev. Genet.* 17.6 (2016), pp. 333–351. DOI: 10.1038/nrg.2016.49.

[66] Mardis, E. R. "DNA sequencing technologies: 2006-2016". In: *Nat Protoc* 12.2 (Feb. 2017), pp. 213–218. DOI: 10.1038/nprot.2016.182.

[67] Miller, J. R., Koren, S. and Sutton, G. "Assembly algorithms for next-generation sequencing data". In: *Genomics* 95.6 (June 2010), pp. 315–327. DOI: 10.1016/j.ygeno.2010.03.001.

[68] Karger, B. L. and Guttman, A. “DNA Sequencing by Capillary Electrophoresis”. In: *Electrophoresis* 30 (Suppl 1 June 2009), S196–S202. DOI: 10.1002/elps.200900218.

[69] Knierim, E., Lucke, B., Schwarz, J. M., Schuelke, M. and Seelow, D. “Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing”. In: *PLoS ONE* 6.11 (2011), e28240. DOI: 10.1371/journal.pone.0028240.

[70] Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J. and Turner, D. J. “Target-enrichment strategies for next-generation sequencing”. In: *Nat Meth* 7.2 (Feb. 2010), pp. 111–118. DOI: 10.1038/nmeth.1419.

[71] Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J. and McCombie, W. R. “Genome-wide in situ exon capture for selective resequencing”. In: *Nat. Genet.* 39.12 (Dec. 2007), pp. 1522–1527. DOI: 10.1038/ng.2007.42.

[72] Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S. and Nusbaum, C. “Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing”. In: *Nat Biotechnol* 27.2 (Feb. 2009), pp. 182–189. DOI: 10.1038/nbt.1523.

[73] Kozarewa, I., Armisen, J., Gardner, A. F., Slatko, B. E. and Hendrickson, C. L. “Overview of Target Enrichment Strategies”. In: *Curr Protoc Mol Biol* 112 (Oct. 2015), pp. 7.21.1–23. DOI: 10.1002/0471142727.mb0721s112.

[74] Kowalsky, C. A., Klesmith, J. R., Stapleton, J. A., Kelly, V., Reichkitzer, N. and Whitehead, T. A. “High-Resolution Sequence-Function Mapping of Full-Length Proteins”. In: *PLOS ONE* 10.3 (19th Mar. 2015), e0118193. DOI: 10.1371/journal.pone.0118193.

[75] Reuter, J. A., Spacek, D. and Snyder, M. P. “High-Throughput Sequencing Technologies”. In: *Mol Cell* 58.4 (May 2015), pp. 586–597. DOI: 10.1016/j.molcel.2015.05.004.

[76] Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G. and Webb, W. W. “Zero-mode waveguides for single-molecule analysis at high concentrations”. In: *Science* 299.5607 (Jan. 2003), pp. 682–686. DOI: 10.1126/science.1079700.

[77] Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J. and Turner, S. "Real-time DNA sequencing from single polymerase molecules". In: *Science* 323.5910 (Jan. 2009), pp. 133–138. DOI: 10.1126/science.1162986.

[78] Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J. and Turner, S. W. "Direct detection of DNA methylation during single-molecule, real-time sequencing". In: *Nat. Methods* 7.6 (June 2010), pp. 461–465. DOI: 10.1038/nmeth.1459.

[79] Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. "Continuous base identification for single-molecule nanopore DNA sequencing". In: *Nat Nanotechnol* 4.4 (Apr. 2009), pp. 265–270. DOI: 10.1038/nnano.2009.12.

[80] Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S. and Turner, S. W. "A flexible and efficient template format for circular consensus sequencing and SNP detection". In: *Nucleic Acids Res.* 38.15 (Aug. 2010), e159. DOI: 10.1093/nar/gkq543.

[81] Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H. C., Mantalas, G. L., Palmeri, K. J., Ishizuka, K. J., Gissi, C., Griggio, F., Ben-Shlomo, R., Corey, D. M., Penland, L., White, R. A., Weissman, I. L. and Quake, S. R. "The genome sequence of the colonial chordate, Botryllus schlosseri". In: *eLife* 2 (July 2013), e00569. DOI: 10.7554/eLife.00569.

[82] Buermans, H. P. J. and Dunnen, J. T. den. "Next generation sequencing technology: Advances and applications". In: *Biochim. Biophys. Acta* 1842.10 (Oct. 2014), pp. 1932–1941. DOI: 10.1016/j.bbadis.2014.06.015.

[83] Dijk, E. L. van, Auger, H., Jaszczyszyn, Y. and Thermes, C. "Ten years of next-generation sequencing technology". In: *Trends Genet.* 30.9 (Sept. 2014), pp. 418–426. DOI: 10.1016/j.tig.2014.07.001.

[84] Levy, S. E. and Myers, R. M. "Advancements in Next-Generation Sequencing". In: *Annu Rev Genomics Hum Genet* 17 (Aug. 2016), pp. 95–115. DOI: 10.1146/annurev-genom-083115-022413.

[85] Ocaña, K. and Oliveira, D. de. “Parallel computing in genomic research: advances and applications”. In: *Adv Appl Bioinform Chem* 8 (13th Nov. 2015), pp. 23–35. DOI: 10.2147/AABC.S64482.

[86] Simpson, J. T. and Pop, M. “The Theory and Practice of Genome Sequence Assembly”. In: *Annu Rev Genomics Hum Genet* 16 (2015), pp. 153–172. DOI: 10.1146/annurev-genom-090314-050032.

[87] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. and Mortazavi, A. “A survey of best practices for RNA-seq data analysis”. In: *Genome Biol.* 17 (Jan. 2016), p. 13. DOI: 10.1186/s13059-016-0881-8.

[88] Bolger, A. M., Lohse, M. and Usadel, B. “Trimmomatic: A flexible trimmer for Illumina Sequence Data”. In: *Bioinformatics* (Apr. 2014), btu170. DOI: 10.1093/bioinformatics/btu170.

[89] Martin, M. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12. DOI: 10.14806/ej.17.1.200.

[90] Li, H. and Durbin, R. “Fast and accurate short read alignment with Burrows-Wheeler transform”. In: *Bioinformatics* 25.14 (July 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[91] Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10 (Mar. 2009), R25. DOI: 10.1186/gb-2009-10-3-r25.

[92] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup. “The Sequence Alignment/Map format and SAMtools”. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[93] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. and Mesirov, J. P. “Integrative Genomics Viewer”. In: *Nat Biotechnol* 29.1 (Jan. 2011), pp. 24–26. DOI: 10.1038/nbt.1754.

[94] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. In: *Genome Res.* 20.9 (Sept. 2010), pp. 1297–1303. DOI: 10.1101/gr.107524.110.

[95] Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, Y. S. "Genotype and SNP calling from next-generation sequencing data". In: *Nat. Rev. Genet.* 12.6 (June 2011), pp. 443–451. DOI: 10.1038/nrg2986.

[96] Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L. and Morgan, M. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nat Meth* 12.2 (Feb. 2015), pp. 115–121. DOI: 10.1038/nmeth.3252.

[97] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., Hoon, D. and L, M. J. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11 (June 2009), pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

[98] Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, Aug. 2009. 756 pp.

[99] Alonso, A., Müller, P., Roewer, L., Willuweit, S., Budowle, B. and Parson, W. "European survey on forensic applications of massively parallel sequencing". In: *Forensic Sci Int Genet* (Apr. 2017). DOI: 10.1016/j.fsigen.2017.04.017.

[100] Fordyce, S. L., Ávila-Arcos, M. C., Rockenbauer, E., Børsting, C., Frank-Hansen, R., Petersen, F. T., Willerslev, E., Hansen, A. J., Morling, N. and Gilbert, M. T. P. "High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform". In: *BioTechniques* 51.2 (Aug. 2011), pp. 127–133. DOI: 10.2144/000113721.

[101] Van Neste, C., Van Nieuwerburgh, F., Van Hoofstat, D. and Deforce, D. "Forensic STR analysis using massive parallel sequencing". In: *Forensic Sci Int Genet* 6.6 (Dec. 2012), pp. 810–818. DOI: 10.1016/j.fsigen.2012.03.004.

[102] Iozzi, S., Carboni, I., Contini, E., Pescucci, C., Frusconi, S., Nutini, A. L., Torricelli, F. and Ricci, U. "Forensic genetics in NGS era: New frontiers for massively parallel typing". In: *Forensic Sci Int Genet Suppl Ser* 5 (Dec. 2015), e418–e419. DOI: 10.1016/j.fsigss.2015.09.166.

[103] Churchill, J. D., Schmedes, S. E., King, J. L. and Budowle, B. “Evaluation of the Illumina(®) Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling”. In: *Forensic Sci Int Genet* 20 (Jan. 2016), pp. 20–29. DOI: 10.1016/j.fsigen.2015.09.009.

[104] Churchill, J. D., Novroski, N. M. M., King, J. L., Seah, L. H. and Budowle, B. “Population and performance analyses of four major populations with Illumina’s FGx Forensic Genomics System”. In: *Forensic Sci Int Genet* 30 (Sept. 2017), pp. 81–92. DOI: 10.1016/j.fsigen.2017.06.004.

[105] Guo, F., Yu, J., Zhang, L. and Li, J. “Massively parallel sequencing of forensic STRs and SNPs using the Illumina(®) ForenSeq™ DNA Signature Prep Kit on the MiSeq FGx™ Forensic Genomics System”. In: *Forensic Sci Int Genet* 31 (Sept. 2017), pp. 135–148. DOI: 10.1016/j.fsigen.2017.09.003.

[106] Jäger, A. C., Alvarez, M. L., Davis, C. P., Guzmán, E., Han, Y., Way, L., Walichiewicz, P., Silva, D., Pham, N., Caves, G., Bruand, J., Schlesinger, F., Pond, S. J. K., Varlaro, J., Stephens, K. M. and Holt, C. L. “Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories”. In: *Forensic Sci Int Genet* 28 (May 2017), pp. 52–70. DOI: 10.1016/j.fsigen.2017.01.011.

[107] Just, R. S., Moreno, L. I., Smerick, J. B. and Irwin, J. A. “Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens”. In: *Forensic Sci Int Genet* 28 (May 2017), pp. 1–9. DOI: 10.1016/j.fsigen.2017.01.001.

[108] Silvia, A. L., Shugarts, N. and Smith, J. “A preliminary assessment of the ForenSeq™ FGx System: next generation sequencing of an STR and SNP multiplex”. In: *Int. J. Legal Med.* 131.1 (Jan. 2017), pp. 73–86. DOI: 10.1007/s00414-016-1457-6.

[109] Xavier, C. and Parson, W. “Evaluation of the Illumina ForenSeq™ DNA Signature Prep Kit - MPS forensic application for the MiSeq FGx™ benchtop sequencer”. In: *Forensic Sci Int Genet* 28 (May 2017), pp. 188–194. DOI: 10.1016/j.fsigen.2017.02.018.

[110] Zeng, X., King, J., Hermanson, S., Patel, J., Storts, D. R. and Budowle, B. “An evaluation of the PowerSeq™ Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing”. In: *Forensic Sci Int Genet* 19 (Nov. 2015), pp. 172–179. DOI: 10.1016/j.fsigen.2015.07.015.

[111] Gaag, K. J. van der, Leeuw, R. H. de, Hoogenboom, J., Patel, J., Storts, D. R., Laros, J. F. J. and Knijff, P. de. "Massively parallel sequencing of short tandem repeats-Population data and mixture analysis results for the PowerSeq™ system". In: *Forensic Sci Int Genet* 24 (Sept. 2016), pp. 86–96. DOI: 10.1016/j.fsigen.2016.05.016.

[112] Gettings, K. B., Kiesler, K. M., Faith, S. A., Montano, E., Baker, C. H., Young, B. A., Guerrieri, R. A. and Vallone, P. M. "Sequence variation of 22 autosomal STR loci detected by next generation sequencing". In: *Forensic Sci Int Genet* 21 (Mar. 2016), pp. 15–21. DOI: 10.1016/j.fsigen.2015.11.005.

[113] Fordyce, S. L., Mogensen, H. S., Børsting, C., Lagacé, R. E., Chang, C.-W., Rajagopalan, N. and Morling, N. "Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM™". In: *Forensic Sci Int Genet* 14 (Jan. 2015), pp. 132–140. DOI: 10.1016/j.fsigen.2014.09.020.

[114] Guo, F., Zhou, Y., Liu, F., Yu, J., Song, H., Shen, H., Zhao, B., Jia, F., Hou, G. and Jiang, X. "Evaluation of the Early Access STR Kit v1 on the Ion Torrent PGM™ platform". In: *Forensic Sci Int Genet* 23 (July 2016), pp. 111–120. DOI: 10.1016/j.fsigen.2016.04.004.

[115] Kim, E. H., Lee, H. Y., Yang, I. S., Jung, S.-E., Yang, W. I. and Shin, K.-J. "Massively parallel sequencing of 17 commonly used forensic autosomal STRs and amelogenin with small amplicons". In: *Forensic Sci Int Genet* 22 (May 2016), pp. 1–7. DOI: 10.1016/j.fsigen.2016.01.001.

[116] Kwon, S. Y., Lee, H. Y., Kim, E. H., Lee, E. Y. and Shin, K.-J. "Investigation into the sequence structure of 23 Y chromosomal STR loci using massively parallel sequencing". In: *Forensic Sci Int Genet* 25 (Nov. 2016), pp. 132–141. DOI: 10.1016/j.fsigen.2016.08.010.

[117] Sathirapatya, T., Sukawutthiya, P. and Vongpaisarnsin, K. "Massively parallel sequencing of 24 Y-STR loci in Thai population". In: *Forensic Sci Int Genet Suppl Ser* (Sept. 2017). DOI: 10.1016/j.fsigss.2017.09.129.

[118] Kim, E. H., Lee, H. Y., Kwon, S. Y., Lee, E. Y., Yang, W. I. and Shin, K.-J. "Sequence-based diversity of 23 autosomal STR loci in Koreans investigated using an in-house massively parallel sequencing panel". In: *Forensic Sci Int Genet* 30 (Sept. 2017), pp. 134–140. DOI: 10.1016/j.fsigen.2017.07.001.

[119] Casals, F., Anglada, R., Bonet, N., Rasal, R., Gaag, K. J. van der, Hoogenboom, J., Solé-Morata, N., Comas, D. and Calafell, F. "Length and repeat-sequence variation in 58 STRs and 94 SNPs in two Spanish populations". In: *Forensic Sci Int Genet* 30 (Sept. 2017), pp. 66–70. DOI: 10.1016/j.fsigen.2017.06.006.

[120] Espregueira Themudo, G., Smidt Mogensen, H., Børsting, C. and Morling, N. "Frequencies of HID-ion ampliseq ancestry panel markers among greenlanders". In: *Forensic Sci Int Genet* 24 (Supplement C Sept. 2016), pp. 60–64. DOI: 10.1016/j.fsigen.2016.06.001.

[121] García, O., Soto, A. and Yurrebaso, I. "Allele frequencies and other forensic parameters of the HID-Ion AmpliSeq™ Identity Panel markers in Basques using the Ion Torrent PGM™ platform". In: *Forensic Sci Int Genet* 28 (May 2017), e8–e10. DOI: 10.1016/j.fsigen.2017.03.010.

[122] Zhao, X., Li, H., Wang, Z., Ma, K., Cao, Y. and Liu, W. "Massively parallel sequencing of 10 autosomal STRs in Chinese using the ion torrent personal genome machine (PGM)". In: *Forensic Sci Int Genet* 25 (Nov. 2016), pp. 34–38. DOI: 10.1016/j.fsigen.2016.07.014.

[123] Novroski, N. M. M., King, J. L., Churchill, J. D., Seah, L. H. and Budowle, B. "Characterization of genetic sequence variation of 58 STR loci in four major population groups". In: *Forensic Sci Int Genet* 25 (Nov. 2016), pp. 214–226. DOI: 10.1016/j.fsigen.2016.09.007.

[124] Lefebvre, P., Velasco, P. T., Dear, A., Lounes, K. C., Lord, S. T., Brennan, S. O., Green, D. and Lorand, L. "Severe hypodysfibrinogenemia in compound heterozygotes of the fibrinogen AalphaIVS4 + 1G>T mutation and an AalphaGln328 truncation (fibrinogen Keokuk)". In: *Blood* 103.7 (Apr. 2004), pp. 2571–2576. DOI: 10.1182/blood-2003-07-2316.

[125] Eduardoff, M., Santos, C., Puente, M. de la, Gross, T. E., Fondevila, M., Strobl, C., Sobrino, B., Ballard, D., Schneider, P. M., Carracedo, Á., Lareu, M. V., Parson, W. and Phillips, C. "Inter-laboratory evaluation of SNP-based forensic identification by massively parallel sequencing using the Ion PGM™". In: *Forensic Sci Int Genet* 17 (July 2015), pp. 110–121. DOI: 10.1016/j.fsigen.2015.04.007.

[126] Elena, S., Alessandro, A., Ignazio, C., Sharon, W., Luigi, R. and Andrea, B. "Revealing the challenges of low template DNA analysis with the prototype Ion AmpliSeq™ Identity panel v2.3 on the PGM™ Sequencer". In: *Forensic Sci Int Genet* 22 (May 2016), pp. 25–36. DOI: 10.1016/j.fsigen.2015.07.011.

[127] Guo, F., Zhou, Y., Song, H., Zhao, J., Shen, H., Zhao, B., Liu, F. and Jiang, X. “Next generation sequencing of SNPs using the HID-Ion AmpliSeq™ Identity Panel on the Ion Torrent PGM™ platform”. In: *Forensic Sci Int Genet* 25 (Nov. 2016), pp. 73–84. DOI: 10.1016/j.fsigen.2016.07.021.

[128] Meiklejohn, K. A. and Robertson, J. M. “Evaluation of the Precision ID Identity Panel for the Ion Torrent(™) PGM(™) sequencer”. In: *Forensic Sci Int Genet* 31 (Aug. 2017), pp. 48–56. DOI: 10.1016/j.fsigen.2017.08.009.

[129] Puente, M. de la, Phillips, C., Santos, C., Fondevila, M., Carracedo, Á. and Lareu, M. V. “Evaluation of the Qiagen 140-SNP forensic identification multiplex for massively parallel sequencing”. In: *Forensic Sci Int Genet* 28 (May 2017), pp. 35–43. DOI: 10.1016/j.fsigen.2017.01.012.

[130] Zhang, S., Bian, Y., Chen, A., Zheng, H., Gao, Y., Hou, Y. and Li, C. “Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM”. In: *Forensic Sci Int Genet* 27 (Mar. 2017), pp. 50–57. DOI: 10.1016/j.fsigen.2016.12.003.

[131] Ramani, A., Wong, Y., Tan, S. Z., Shue, B. H. and Syn, C. “Ancestry prediction in Singapore population samples using the Illumina ForenSeq kit”. In: *Forensic Sci Int Genet* 31 (Aug. 2017), pp. 171–179. DOI: 10.1016/j.fsigen.2017.08.013.

[132] Pereira, V., Mogensen, H. S., Børsting, C. and Morling, N. “Evaluation of the Precision ID Ancestry Panel for crime case work: A SNP typing assay developed for typing of 165 ancestral informative markers”. In: *Forensic Sci Int Genet* 28 (May 2017), pp. 138–145. DOI: 10.1016/j.fsigen.2017.02.013.

[133] Parson, W., Ballard, D., Budowle, B., Butler, J. M., Gettings, K. B., Gill, P., Gusmão, L., Hares, D. R., Irwin, J. A., King, J. L., Knijff, P. d., Morling, N., Prinz, M., Schneider, P. M., Neste, C. V., Willuweit, S. and Phillips, C. “Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements”. In: *Forensic Sci Int Genet* 22 (May 2016), pp. 54–63. DOI: 10.1016/j.fsigen.2016.01.009.

[134] Gusmão, L., Butler, J. M., Linacre, A., Parson, W., Roewer, L., Schneider, P. M. and Carracedo, A. “Revised guidelines for the publication of genetic population data”. In: *Forensic Sci Int Genet* 30 (Supplement C Sept. 2017), pp. 160–163. DOI: 10.1016/j.fsigen.2017.06.007.

[135] Bodner, M., Bastisch, I., Butler, J. M., Fimmers, R., Gill, P., Gusmão, L., Morling, N., Phillips, C., Prinz, M., Schneider, P. M. and Parson, W. "Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER)". In: *Forensic Sci Int Genet* 24 (Sept. 2016), pp. 97–102. DOI: 10.1016/j.fsigen.2016.06.008.

[136] Woerner, A. E., King, J. L. and Budowle, B. "Fast STR allele identification with STRait Razor 3.0". In: *Forensic Sci Int Genet* 30 (Sept. 2017), pp. 18–23. DOI: 10.1016/j.fsigen.2017.05.008.

[137] Friis, S. L., Buchard, A., Rockenbauer, E., Børsting, C. and Morling, N. "Introduction of the Python script STRinNGS for analysis of STR regions in FASTQ or BAM files and expansion of the Danish STR sequence database to 11 STRs". In: *Forensic Sci Int Genet* 21 (Mar. 2016), pp. 68–75. DOI: 10.1016/j.fsigen.2015.12.006.

[138] Lee, J. C.-I., Tseng, B., Chang, L.-K. and Linacre, A. "SEQ Mapper: A DNA sequence searching tool for massively parallel sequencing data". In: *Forensic Sci Int Genet* 26 (Jan. 2017), pp. 66–69. DOI: 10.1016/j.fsigen.2016.10.006.

[139] Ganschow, S., Wiegand, P. and Tiemann, C. "toaSTR: A web-based forensic tool for the analysis of short tandem repeats in massively parallel sequencing data". In: *Forensic Sci Int Genet Supplement Series* (Sept. 2017). DOI: 10.1016/j.fsigss.2017.09.034.

[140] Hoogenboom, J., Gaag, K. J. van der, Leeuw, R. H. de, Sijen, T., Knijff, P. de and Laros, J. F. J. "FDSTools: A software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise". In: *Forensic Sci Int Genet* 27 (Mar. 2017), pp. 27–40. DOI: 10.1016/j.fsigen.2016.11.007.

[141] Wienroth, M., Morling, N. and Williams, R. "Technological innovations in forensic genetics: social, legal and ethical aspects". In: *Recent Adv DNA Gene Seq* 8.2 (2014), pp. 98–103.

[142] Ethics Group of the National DNA Database. *Next generation sequencing technologies: ethical considerations - GOV.UK*. URL: https://www.gov.uk/government/publications/next-generation-sequencing-technologies-ethical-considerations.

[143] Kayser, M. "Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes". In: *Forensic Sci Int Genet* 18 (Sept. 2015), pp. 33–48. DOI: 10.1016/j.fsigen.2015.02.003.

[144] Bioethics, N. C. on. *The forensic use of bioinformation: ethical issues*. Nuffield Bioethics. Sept. 2007. URL: http://nuffieldbioethics.org/project/bioinformation.

[145] M'charek, A., Toom, V. and Prainsack, B. "Bracketing off population does not advance ethical reflection on EVCs: a reply to Kayser and Schneider". In: *Forensic Sci Int Genet* 6.1 (Jan. 2012), e16–17, author reply e18–19. DOI: 10.1016/j.fsigen.2010.12.012.

[146] Sankar, P., Cho, M. K., Condit, C. M., Hunt, L. M., Koenig, B., Marshall, P., Lee, S. S.-J. and Spicer, P. "Genetic research and health disparities". In: *JAMA* 291.24 (June 2004), pp. 2985–2989. DOI: 10.1001/jama.291.24.2985.

[147] Williams, R. and Wienroth, M. "Social and ethical aspects of forensic genetics: A critical review". In: *Forensic Sci Rev* 29.2 (July 2017), pp. 145–169.

[148] Lindahl, T. "Instability and decay of the primary structure of DNA". In: *Nature* 362.6422 (Apr. 1993), pp. 709–715. DOI: 10.1038/362709a0.

[149] Lindahl, T. and Andersson, A. "Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid". In: *Biochemistry* 11.19 (Sept. 1972), pp. 3618–3623. DOI: 10.1021/bi00769a019.

[150] Lindahl, T. and Nyberg, B. "Rate of depurination of native deoxyribonucleic acid". In: *Biochemistry* 11.19 (Sept. 1972), pp. 3610–3618. DOI: 10.1021/bi00769a018.

[151] Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L. and Hofreiter, M. "Genetic analyses from ancient DNA". In: *Annu. Rev. Genet.* 38 (2004), pp. 645–679. DOI: 10.1146/annurev.genet.37.110801.143214.

[152] Hansen, A. J., Mitchell, D. L., Wiuf, C., Paniker, L., Brand, T. B., Binladen, J., Gilichinsky, D. A., Rønn, R. and Willerslev, E. "Crosslinks Rather Than Strand Breaks Determine Access to Ancient DNA Sequences From Frozen Sediments". In: *Genetics* 173.2 (June 2006), pp. 1175–1179. DOI: 10.1534/genetics.106.057349.

[153] Baust, J. "Strategies for the storage of DNA". In: *Biopreserv Biobank* 6.4 (Dec. 2008), pp. 251–252. DOI: 10.1089/bio.2008.0604.lett.

[154] Wyllie, A. H. "Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation". In: *Nature* 284.5756 (Apr. 1980), pp. 555–556.

[155] Barone, F., Belli, M., Pazzaglia, S., Sapora, O. and Tabocchini, M. A. "Radiation damage and chromatin structure." In: *Ann Ist Super Sanita* 25.1 (1989), pp. 59–67.

[156] Hansen, J. C. "Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions". In: *Annu Rev Biophys Biomol Struct* 31 (2002), pp. 361–392. DOI: 10.1146/annurev.biophys.31.101101.140858.

[157] Gaffney, D. J., McVicker, G., Pai, A. A., Fondufe-Mittendorf, Y. N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y. and Pritchard, J. K. "Controls of Nucleosome Positioning in the Human Genome". In: *PLOS Genetics* 8.11 (Nov. 2012), e1003036. DOI: 10.1371/journal.pgen.1003036.

[158] Ljungman, M. and Hanawalt, P. C. "Efficient protection against oxidative DNA damage in chromatin". In: *Mol. Carcinog.* 5.4 (1992), pp. 264–269.

[159] Voss, T. C. and Hager, G. L. "Dynamic regulation of transcriptional states by chromatin and transcription factors". In: *Nat. Rev. Genet.* 15.2 (Feb. 2014), pp. 69–81. DOI: 10.1038/nrg3623.

[160] Thanakiatkrai, P. and Welch, L. "Evaluation of nucleosome forming potentials (NFPs) of forensically important STRs". In: *Forensic Sci Int Genet* 5.4 (Aug. 2011), pp. 285–290. DOI: 10.1016/j.fsigen.2010.05.002.

[161] Freire-Aradas, A., Fondevila, M., Kriegel, A.-K., Phillips, C., Gill, P., Prieto, L., Schneider, P. M., Carracedo, A. and Lareu, M. V. "A new SNP assay for identification of highly degraded human DNA". In: *Forensic Sci Int Genet* 6.3 (May 2012), pp. 341–349. DOI: 10.1016/j.fsigen.2011.07.010.

[162] Pedersen, J. S., Valen, E., Velazquez, A. M. V., Parker, B. J., Rasmussen, M., Lindgreen, S., Lilje, B., Tobin, D. J., Kelly, T. K., Vang, S., Andersson, R., Jones, P. A., Hoover, C. A., Tikhonov, A., Prokhortchouk, E., Rubin, E. M., Sandelin, A., Gilbert, M. T. P., Krogh, A., Willerslev, E. and Orlando, L. "Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome". In: *Genome Research* 24.3 (Mar. 2014), pp. 454–466. DOI: 10.1101/gr.163592.113.

[163] Bishop (ed), O. T. *Bioinformatics and Data Analysis in Microbiology*. Caister Academic Press, July 2014.

[164] HMP Consortium, 2012b. "Structure, function and diversity of the healthy human microbiome". In: *Nature* 486.7402 (June 2012), pp. 207–214. DOI: 10.1038/nature11234.

[165] Lloyd-Price, J., Abu-Ali, G. and Huttenhower, C. “The healthy human microbiome”. In: *Genome Med* 8 (27th Apr. 2016). DOI: 10.1186/s13073-016-0307-y.

[166] Dethlefsen, L., Huse, S., Sogin, M. L. and Relman, D. A. “The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing”. In: *PLoS Biol.* 6.11 (Nov. 2008), e280. DOI: 10.1371/journal.pbio.0060280.

[167] DiGiulio, D. B., Callahan, B. J., McMurdie, P. J., Costello, E. K., Lyell, D. J., Robaczewska, A., Sun, C. L., Goltsman, D. S. A., Wong, R. J., Shaw, G., Stevenson, D. K., Holmes, S. P. and Relman, D. A. “Temporal and spatial variation of the human microbiota during pregnancy”. In: *Proc. Natl. Acad. Sci. U.S.A.* 112.35 (Sept. 2015), pp. 11060–11065. DOI: 10.1073/pnas.1502875112.

[168] De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J. B., Massart, S., Collini, S., Pieraccini, G. and Lionetti, P. “Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa”. In: *Proc. Natl. Acad. Sci. U.S.A.* 107.33 (Aug. 2010), pp. 14691–14696. DOI: 10.1073/pnas.1005963107.

[169] Chakraborty, C., Doss, C. G. P., Patra, B. C. and Bandyopadhyay, S. “DNA barcoding to map the microbial communities: current advances and future directions”. In: *Appl. Microbiol. Biotechnol.* 98.8 (Apr. 2014), pp. 3425–3436. DOI: 10.1007/s00253-014-5550-9.

[170] Vinje, H., Almøy, T., Liland, K. H. and Snipen, L. “A systematic search for discriminating sites in the 16S ribosomal RNA gene”. In: *Microb Inform Exp* 4.1 (Jan. 2014), p. 2. DOI: 10.1186/2042-5783-4-2.

[171] Cuesta-Zuluaga, J. de la and Escobar, J. S. “Considerations For Optimizing Microbiome Analysis Using a Marker Gene”. In: *Front Nutr* 3 (2016), p. 26. DOI: 10.3389/fnut.2016.00026.

[172] Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z. and Forney, L. J. “Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome”. In: *PLoS One* 7.3 (Mar. 2012). DOI: 10.1371/journal.pone.0033865.

[173] Abusleme, L., Hong, B.-Y., Dupuy, A. K., Strausbaugh, L. D. and Diaz, P. I. “Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing”. In: *J Oral Microbiol* 6 (2014). DOI: 10.3402/jom.v6.23990.

[174] Rubin, B. E. R., Sanders, J. G., Hampton-Marcell, J., Owens, S. M., Gilbert, J. A. and Moreau, C. S. “DNA extraction protocols cause differences in 16S rRNA amplicon sequencing efficiency but not in community profile composition or structure”. In: *Microbiologyopen* 3.6 (Dec. 2014), pp. 910–921. DOI: 10.1002/mbo3.216.

[175] Wesolowska-Andersen, A., Bahl, M. I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R. and Licht, T. R. “Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis”. In: *Microbiome* 2 (June 2014), p. 19. DOI: 10.1186/2049-2618-2-19.

[176] Vesty, A., Biswas, K., Taylor, M. W., Gear, K. and Douglas, R. G. “Evaluating the Impact of DNA Extraction Method on the Representation of Human Oral Bacterial and Fungal Communities”. In: *PLoS ONE* 12.1 (2017), e0169877. DOI: 10.1371/journal.pone.0169877.

[177] Ishii, K. and Fukui, M. “Optimization of Annealing Temperature To Reduce Bias Caused by a Primer Mismatch in Multitemplate PCR”. In: *Appl Environ Microbiol* 67.8 (Aug. 2001), pp. 3753–3755. DOI: 10.1128/AEM.67.8.3753-3755.2001.

[178] Kurata, S., Kanagawa, T., Magariyama, Y., Takatsu, K., Yamada, K., Yokomaku, T. and Kamagata, Y. “Reevaluation and reduction of a PCR bias caused by reannealing of templates”. In: *Appl. Environ. Microbiol.* 70.12 (Dec. 2004), pp. 7545–7549. DOI: 10.1128/AEM.70.12.7545-7549.2004.

[179] Wang, G. C. and Wang, Y. “The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species”. In: *Microbiology (Reading, Engl.)* 142 (Pt 5) (May 1996), pp. 1107–1114. DOI: 10.1099/13500872-142-5-1107.

[180] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. “QIIME allows analysis of high-throughput community sequencing data”. In: *Nat. Methods* 7.5 (May 2010), pp. 335–336. DOI: 10.1038/nmeth.f.303.

[181] Edgar, R. C. “UPARSE: highly accurate OTU sequences from microbial amplicon reads”. In: *Nat. Methods* 10.10 (Oct. 2013), pp. 996–998. DOI: 10.1038/nmeth.2604.

[182] Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. "VSEARCH: a versatile open source tool for metagenomics". In: *PeerJ* 4 (2016), e2584. DOI: 10.7717/peerj.2584.

[183] Yang, Y., Xie, B. and Yan, J. "Application of next-generation sequencing technology in forensic science". In: *Genomics Proteomics Bioinformatics* 12.5 (Oct. 2014), pp. 190–197. DOI: 10.1016/j.gpb.2014.09.001.

[184] Scheinin, I., Sie, D., Bengtsson, H., Wiel, M. A. van de, Olshen, A. B., Thuijl, H. F. van, Essen, H. F. van, Eijk, P. P., Rustenburg, F., Meijer, G. A., Reijneveld, J. C., Wesseling, P., Pinkel, D., Albertson, D. G. and Ylstra, B. "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly". In: *Genome Res.* 24.12 (Dec. 2014), pp. 2022–2032. DOI: 10.1101/gr.175141.114.

[185] Benjamini, Y. and Speed, T. P. "Summarizing and correcting the GC content bias in high-throughput sequencing". In: *Nucl. Acids Res.* 40.10 (May 2012), e72–e72. DOI: 10.1093/nar/gks001.

[186] Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R. and Ribeca, P. "Fast Computation and Applications of Genome Mappability". In: *PLoS ONE* 7.1 (Jan. 2012), e30377. DOI: 10.1371/journal.pone.0030377.

[187] Nechipurenko, D. Y., Il'icheva, I. A., Panchenko, L. A., Poptsova, M. S., Khodikov, M. V., Oparina, N. Y., Polozov, R. V., Grokhovsky, S. L. and Nechipurenko, Y. D. "Non-random DNA fragmentation in next-generation sequencing". In: *Scientific Reports* 4 (Mar. 2014), p. 4532. DOI: 10.1038/srep04532.

[188] Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., AL-Rasheid, K. A., Willerslev, E., Krogh, A. and Orlando, L. "Improving ancient DNA read mapping against modern reference genomes". In: *BMC Genomics* 13 (May 2012), p. 178. DOI: 10.1186/1471-2164-13-178.

[189] Becila, S., Herrera-Mendez, C. H., Coulis, G., Labas, R., Astruc, T., Picard, B., Boudjellal, A., Pelissier, P., Bremaud, L. and Ouali, A. "Postmortem muscle cells die through apoptosis". In: *Eur Food Res Technol* 231.3 (July 2010), pp. 485–493. DOI: 10.1007/s00217-010-1296-5.

[190] Botchkareva, N. V., Ahluwalia, G. and Shander, D. "Apoptosis in the hair follicle". In: *J. Invest. Dermatol.* 126.2 (Feb. 2006), pp. 258–264. DOI: 10.1038/sj.jid.5700007.

[191] Pohl, F. M., Thomae, R. and Karst, A. “Temperature dependence of the activity of DNA-modifying enzymes: endonucleases and DNA ligase”. In: *Eur. J. Biochem.* 123.1 (Mar. 1982), pp. 141–152.

[192] Verdon, T. J., Mitchell, R. J. and Oorschot, R. A. H. van. “Swabs as DNA collection devices for sampling different biological materials from different substrates”. In: *J. Forensic Sci.* 59.4 (July 2014), pp. 1080–1089. DOI: 10.1111/1556-4029.12427.

[193] Hess, S. and Haas, C. “Recovery of Trace DNA on Clothing: A Comparison of Mini-tape Lifting and Three Other Forensic Evidence Collection Techniques”. In: *J. Forensic Sci.* 62.1 (Jan. 2017), pp. 187–191. DOI: 10.1111/1556-4029.13246.

[194] Grice, E. A. and Segre, J. A. “The skin microbiome”. In: *Nat Rev Microbiol* 9.4 (Apr. 2011), pp. 244–253. DOI: 10.1038/nrmicro2537.

[195] Liland, K. H., Vinje, H. and Snipen, L. “microclass: an R-package for 16S taxonomy classification”. In: *BMC Bioinformatics* 18 (Mar. 2017). DOI: 10.1186/s12859-017-1583-2.

[196] Sender, R., Fuchs, S. and Milo, R. “Revised Estimates for the Number of Human and Bacteria Cells in the Body”. In: *PLoS Biol* 14.8 (Aug. 2016). DOI: 10.1371/journal.pbio.1002533.

[197] Grice, E. A., Kong, H. H., Conlan, S., Deming, C. B., Davis, J., Young, A. C., NISC Comparative Sequencing Program, Bouffard, G. G., Blakesley, R. W., Murray, P. R., Green, E. D., Turner, M. L. and Segre, J. A. “Topographical and temporal diversity of the human skin microbiome”. In: *Science* 324.5931 (May 2009), pp. 1190–1192. DOI: 10.1126/science.1171700.

[198] Kong, H. H. “Skin microbiome: genomics-based insights into the diversity and role of skin microbes”. In: *Trends Mol Med* 17.6 (June 2011), pp. 320–328. DOI: 10.1016/j.molmed.2011.01.013.

Paper I

Forensic Science International: Genetics 27 (2017) 160–166

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Degradation in forensic trace DNA samples explored by massively parallel sequencing

Eirik Nataas Hanssen [a,b,*], Robert Lyle [c,d], Thore Egeland [a,e], Peter Gill [a,b]

[a] Department of Forensic Biology, Norwegian Institute of Public Health, Oslo, Norway
[b] Department of Forensic Medicine, University of Oslo, Oslo, Norway
[c] Department of Medical Genetics, University of Oslo, Oslo, Norway
[d] Department of Medical Genetics, Oslo University Hospital, Oslo, Norway
[e] IKBM, Norwegian University of Life Sciences, Aas, Norway

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 29 November 2016
Accepted 2 January 2017
Available online 3 January 2017

Keywords:
DNA degradation
Massively parallel sequencing

ABSTRACT

Routine forensic analysis using STRs will fail if the DNA is too degraded. The DNA degradation process in biological stain material is not well understood. In this study we sequenced old semen and blood stains by massively parallel sequencing. The sequence data coverage was used to measure degradation across the genome. The results supported the contention that degradation is uniform across the genome, showing no evidence of regions with increased or decreased resistance towards degradation. Thus the lack of genetic regions robust to degradation removes the possibility of using such regions to further optimize analysis performance for degraded DNA.

1. Introduction

Standard forensic analysis using short tandem repeat markers (STRs) is limited by the size of the markers when highly degraded material is encountered. To overcome this problem, shorter markers (mini STRs) were introduced [1], and in the latest generation of routine analysis kits all fragment sizes are below ~350 bp [2,3]. This is close to the theoretical limitations for STRs. Recently, single nucleotide polymorphisms (SNPs) have been included in commercially available sequencing panels [4,5]. If performance is to be further improved, it is important to establish if there is variability in a marker's resistance to degradation. However, current knowledge on the robustness of genomic regions to degradation in biological trace material is sparse [6–8].

DNA associates with proteins to form a complex called chromatin. The basic chromatin structure has a repetitive pattern of nucleosomes each consisting of ~146 bp of DNA wrapped round the histone octamer with linker DNA in between. Nucleosome protection of the DNA strand in living cells has been observed in radioactive radiation [9]. In apoptosis, or programmed cell death, the nucleosomal associated DNA sequences escape enzymatic cleavage [10]. This relatively open structured chromatin is named euchromatin and constitutes most of the genome including the genes (~94%) [11,12]. The remaining ~6% of the genome has a higher ordered structure which is stabilized by the association between the nucleosomes and the linker histone H1. This structure is called heterochromatin and has been shown to give additional protection against DNA damage when studied in vitro [13]. Heterochromatin is further divided into two subgroups. Constitutive heterochromatin is stable with the same positioning between cell types, while facultative heterochromatin can also adopt the open euchromatic form [14,15]. Sperm has its own chromatin structure where less then 15% of the DNA is histone-bound and the vast majority is associated with protamines which is further condensed into toroids [16]. For living semen cells it has been shown that DNA is more protected against degradation when associated with protamines than histones [17].

In relation to forensic biology, few studies have been performed on potential nucleosome protected DNA sequences using standard STRs [18] and SNPs [19]. For the latter, the improvement in success rate compared to the most robust established forensic SNP multiplex was relatively small (~6%), but significantly higher compared to the mini-STR assay. To our knowledge no study has been performed to investigate DNA protection in biological trace material further.

Massively parallel sequencing technology (MPS) provides an opportunity to explore DNA degradation in much greater detail. Each DNA region is typically sequenced several times, and the number of times is expressed as the coverage. If one region has

* Corresponding author at: Department of Forensic Biology, Norwegian Institute of Public Health, Oslo, Norway.
E-mail address: kjeirik@gmail.com (E.N. Hanssen).

http://dx.doi.org/10.1016/j.fsigen.2017.01.002
1872-4973/

more DNA present compared to others, more DNA is available for sequencing, and the relative coverage will be elevated. Thus the coverage can be used as an expression of degradation level along the DNA strand. However, the coverage is influenced by several factors that must be adjusted for in order to use it quantitatively.

1. PCR used in the sequencing process causes a bias in the GC rich regions to be underrepresented because amplification is less efficient. In the AT rich regions the opposite effect is seen [20,21].
2. In regions with repetitive sequences the mapping can be ambiguous, and reads will map to different sites in the reference genome creating a bias [22].

The twin effects of GC bias and mappability bias, are the most significant causes of inaccurate coverage [23]. Non-random DNA shearing from the fragmentation step in the sequencing process will have minor impact and is not considered in this context [24]. See Covaris webpage for additional information (http://covarisinc.com/).

To explore DNA degradation, and its impact in forensics, we sequenced genomes of degraded DNA in semen and blood stains. The coverage, adjusted for GC bias and mappability, was used to measure the degradation level. In addition, we developed a bioinformatics workflow to support the data analysis.

2. Materials and methods

2.1. Sample selection and preparation

A limited number of samples were chosen from ten candidate samples. In order to optimize the possibility to detect differences in degradation level along the DNA strands, we selected samples of different cell types and different degrees of overall degradation. To optimize the sequencing process, DNA concentration and fragment size distribution in the samples were also important [25].

The candidate samples were collected from 25–30 year old semen and blood stains. The sample material had been applied onto cotton towels and dried. The towels were then put in separate plastic bags before storing in cardboard boxes under dry conditions at room temperature. The semen samples were extracted using differential extraction [26] and the MinElute protocol (Qiagen) for cleanup. Blood samples were extracted using the Chelex method (Bio-Rad) without SDS/proteinase K treatment or any additional cleanup stage. To evaluate the DNA quality the samples were quantified using both the Quantifiler Duo kit/7500 RT-PCR (Thermo Fisher) and the High Sensitivity DNA Kit/2100 Bioanalyzer (Agilent). In addition the fragment distributions were evaluated from the Bioanalyser data. The samples were also STR typed with ESX 17 (Promega). Based on the overall results, four samples were chosen from the ten candidate samples. Two were taken from the same semen stain (concentration ~0.5 ng human DNA/μL, moderately degraded and named Degraded1 and 2) and two from different bloodstains of male and female origin respectively (concentration ~0.01 ng human DNA/μL, heavily degraded and named HeavilyDeg1 and 2). For more details see "Choosing samples for sequencing" in supplement.

The Bioanalyzer results showed that the samples had a broad size distribution ranging from 35 to several 1000 bp. This range would lead to inefficient sequencing of the longest fragments. To avoid this bias, mechanical shearing was used, following the Covaris sonicator protocol (http://covarisinc.com/). The aim was to generate 300 bp fragments to fit the chosen 150 bp read lengths used for sequencing. In order to compensate for the low DNA quantity the library preparation was carried out with the MicroPlex kit (Diagenode) designed for the low DNA levels handled in chromatin immunoprecipitation analysis (ChIP). During the library process adapters and indexes were ligated to each end of the DNA fragments. Paired-end sequencing, where each fragment was sequenced from both ends, was used. A read length of 150 bp was performed on a single lane on a HiSeq 2500 platform (Illumina).

2.2. Bioinformatics

The evaluation of the paired-end sequencing data showed that the libraries had shorter fragment sizes than expected with a median at around 100 bp (see "Size distribution of fragments in library" in supplement for details). With a 150 bp read length from both sides, nearly all fragments would be completely sequenced. However those below 150 bp reads would include the adapter. To remove the adapter sequences from the reads we used the palindrome mode of Trimmomatic [27]. To remove the impact of reads overlapping by different degrees in the middle of the fragments, each pair was merged together by making a consensus sequence of the overlapping middle part using the FLASH tool [28]. A small fraction of the paired reads did not overlap and could therefore not be merged (around 6% of fragments from 285–425 bp). The regions in the middle of these fragments made up less than 1% of the total sequence, and were therefore not adjusted for. Consequently it is reasonable to regard every fragment completely sequenced and covered only once.

The merged and the non-merged datasets were separately mapped and filtered before being combined to create a comprehensive dataset for coverage statistics. In this process the Burrows-Wheeler Aligner software (BWA) with default settings was used for mapping against the hg19 reference genome (downloaded from the UCSC webpage (http://genome.ucsc.edu/index.html)). Single-end and paired-end approaches were used respectively [29]. The resulting alignment files containing the mapped reads information (BAM files) were filtered using the SAMtools software removing reads of mapping quality below 37 (probability of mapping error <1/10e3.7) [30]. In addition PCR duplicates were removed using the Picard tool (http://broadinstitute.github.io/picard/). GC bias was corrected using the deepTools package (computeGCbias and computeGCbias tools) [31]. The filtered merged and non-merged datasets were combined and sorted using SAMtools.

Whole genome coverage data in bedgraph format was derived from the combined dataset using Igvtools (used default settings: maximum zoom level to precompute ($-z$) = 7, window size over which coverage is averaged ($-w$) = 25 and window functions to use when reducing the data ($-f$) = mean) [32]. Regions with reduced mappability [33] were removed from the dataset by applying the intersect function of the Bedtools software [34]. The annotation tracks wgEncodeDacMapabilityConsensusExcludable.bed and wgEncodeCrgMapabilityAlign100mer.bigwig were used as template files (text files downloaded from the UCSC table browser tool: https://genome.ucsc.edu). The original wgEncodeCrgMapabilityAlign100mer.bigwig was modified by standard Unix commands cutting 200 bp from each side of the genomic segments and then removing segments shorter than 400 bp.

Datasets for two control samples sequenced with comparable parameters as the degraded samples were downloaded from the 1000 genomes project webpage (http://www.1000genomes.org/ – blood samples named ERR233219 (female) and ERR257960 (male)). These datasets had been marginally filtered in advance. Because of long library fragments (450 bp) compared to the read length (2×100) the read pairs would neither contain adapter sequence nor overlap each other as for the degraded samples. Consequently all fragments would have sequence in the middle not covered by the read pairs. These uncovered regions were left unadjusted as the fragmentation in the library preparation step was assumed to be random thereby giving these regions an even distribution across the genome for undegraded samples. Therefore,

the control sample was mapped to the reference genome as paired-end reads only using BWA, and then further corrected and filtered following the same workflow as for the degraded samples.

The filtered and corrected coverage datasets were used as a measure of degradation along the DNA strands. Several different approaches were followed. Integrative Genomics Viewer (IGV) [32] was used for visual evaluation. R software [35] was used for calculations and plots.

To compare chromosomes, the coverage of each chromosome was calculated by dividing the number of mapped bp by the total number of mappable bp in the whole chromosome [11].

When comparing regions with different chromatin structure, coverage of heterochromatic and promoter regions respectively were extracted using genomic annotation tracks as targets. The heterochromatin track was derived by merging six venous blood H3K9me3 tracks downloaded by connecting the Blueprint track hub to the UCSC table browser tool [36]. The tracks were converted from hg38 to hg19 coordinates using the liftover tool at the same website and further merged using Bedtools [37]. The promoter track was derived from whole blood expression data of the GTEx track downloaded from the UCSC table browser tool [38]. Active promoter regions were defined as 2 kb upstream of genes with a FPKM above 1 (RNA-seq data for a gene is typically normalized for exon length and sequencing depth and given as fragments per kilobase of exon per million fragments mapped (FPKM)). In addition, we used the random function of Bedtools to generate a track of randomly chosen genomic regions (10,000 × 1000 kb long regions). The three track files were filtered, as the samples and controls, for low mappability regions.

For a full list of software used in this study see "Software and versions" in supplement.

3. Results and discussion

Initial evaluation of the sequencing data was done using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and the data showed generally good quality. No overrepresented sequences such as adapter dimers was observed. Sequencing blocking and miscoding lesions are caused by oxidation, hydrolysis and free radicals in decomposing biological material, and these effects are often a problem when sequencing ancient DNA [39,8]. For this study the biological material had been stored under dry conditions at room temperature, and these events should therefore be marginalized [7]. Nor did the FastQC evaluation show indication of such.

The need for correcting the coverage data for mappability and GC bias was evaluated. The uncorrected coverage data were significant reduced in regions of low mappability when visually compared to the mappability annotated tracks using IGV. The need for GC bias correction was demonstrated by plots from deepTools as exemplified in Fig. 1.

In the final datasets coverage was ~0.5× for the heavily degraded samples, ~3× for the degraded samples and ~5× for the controls. From the original sequence data the optimal theoretical coverages were calculated to be 9×, 10× and 7× respectively. For the degraded and heavily degraded samples however ~2/3 of the reads were lost in the adapter trimming and merging of the pair-end reads. Adjusted for this, the degraded samples lost only a fraction of the reads in the mapping and subsequent data filtering. For the heavily degraded samples half of the reads did not map to the hg19 reference which could be caused by nonhuman DNA in the samples. In the subsequent filtering high mapping quality thresholds left only half of the mapped reads in the final dataset. However we regard the final datasets to be fit for purpose as even lower coverage data has been used to determine copy number variation (CNV) [40].

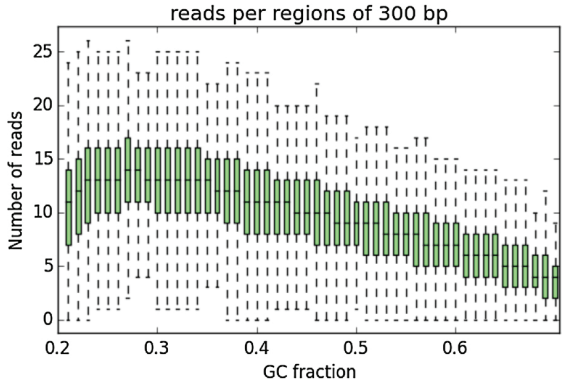

Fig. 1. As the hydrogen bonding between the GC bp is stronger than between the AT bp, the polymerases used in the sequencing process are less efficient in the GC rich regions. This results in GC bias with lower coverage in these regions. The samples were checked for GC bias by dividing the genome into regions of 300 basepairs and plotting number of reads as a function of the GC fraction. The resulting plot shows that the number of reads or coverage decreases as the GC content increases. As the coverage is used quantitatively the sequence data needed to be GC corrected.

By visually evaluation the final datasets using IGV the degraded and heavily degraded samples showed an even level of coverage across all chromosomes which was similar to that observed for the controls. The coverage at the STRs used in forensic routine work had the same level as the rest of the genomes. High coverage peaks were observed at positions around the centromeres and to a lesser extent in the telomeres for the majority of chromosomes. These peaks were observed in both the degraded samples and the controls and were caused by long arrays of tandem repeats in these regions.

3.1. Variation of coverage

The difference between each pair of adjacent coverage values was used to investigate the variation of coverage. The distribution of differences for each sample were compared to a normal distribution with the same standard deviation, as shown in Fig. 2. All genome data were used.

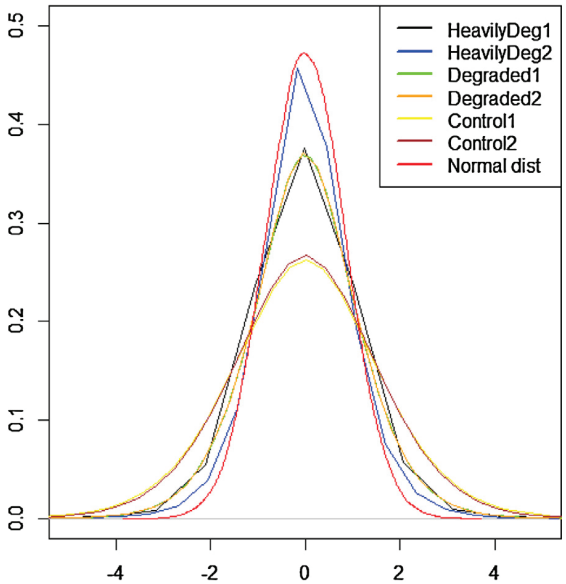

Fig. 2. The distributions of variation of coverage along the chromosomes for all samples including the controls were comparable to that of a normal distribution with the same standard deviation. Consequently the variation in coverage is principally random.

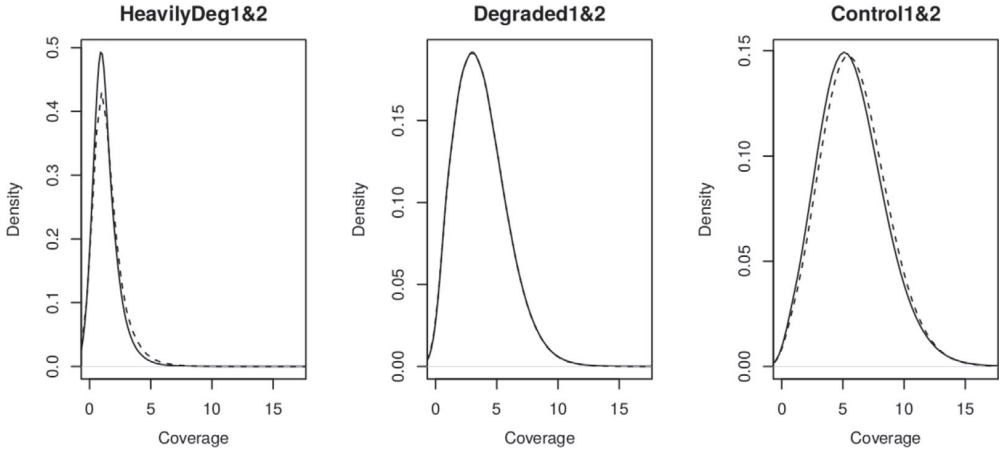

Fig. 3. Distribution of whole genome coverage data plotted using the density function in R. The plots show uniform shapes for all samples and controls. In addition no significant extreme values are observed. This indicates a reasonable constant level of DNA degradation across the whole genome. For each plot sample 1 is represented with a solid line and sample 2 with a dotted line. Sample Degraded1 and Degraded2 have overlapping plots.

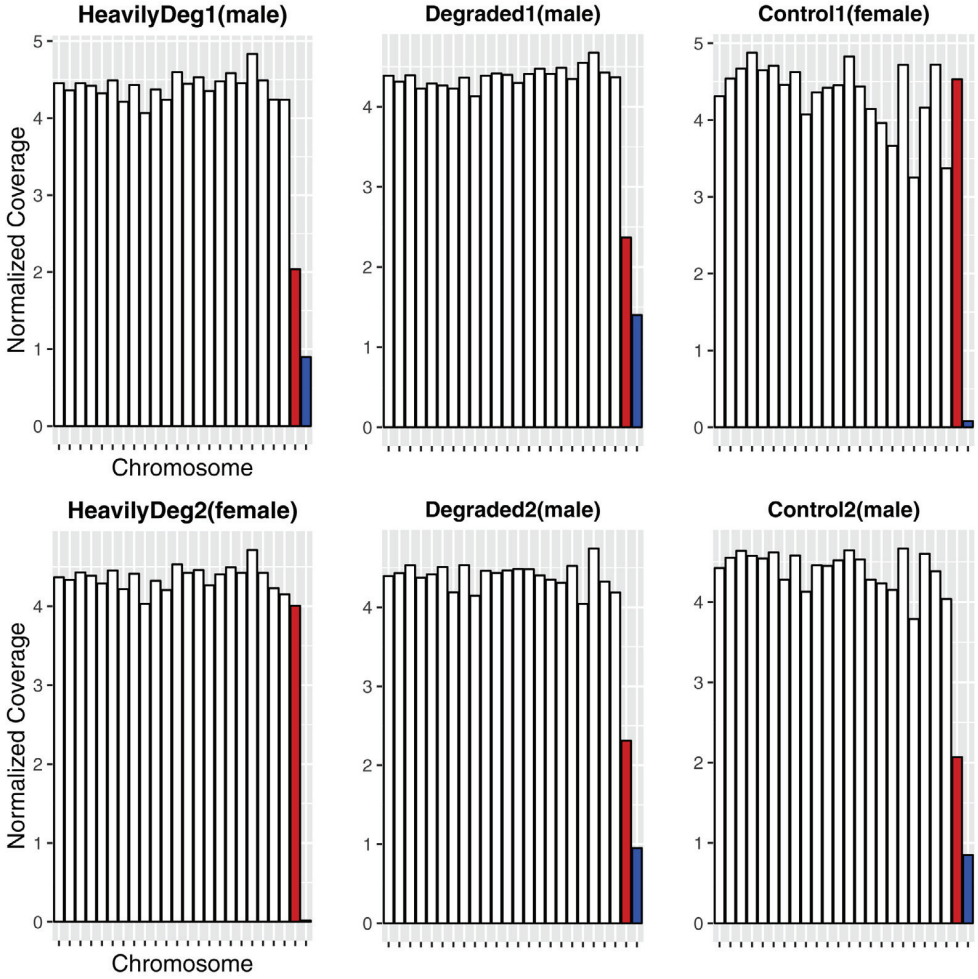

Fig. 4. The coverage values indicate that there is no difference in degradation level between the chromosomes. The controls show even more variation between the chromosomes than the samples. Most heterochromatic regions are not included in the sequencing data, but for the female sample (Heavily degraded 2) the second X-chromosome is silenced by heterochromatin and should give increased coverage values if heterochromatin protects against degradation. The lower than expected coverage values for the Y-chromosomes are presumed to be caused by an artifact rather than degradation, while low coverage value was also observed for the control (Control2). The X-chromosomes are colored red and Y-chromosomes colored blue.

All sample and control distributions were comparable to the normal distribution. Thus the variation in coverage is mainly random and not largely dependent on external factors.

3.2. Coverage distribution between samples

For each sample and control, the frequency of all coverage data were plotted as density plots (see Fig. 3).

Coverage distribution of the degraded samples is uniform and comparable in shape to the controls. Additional peaks, extreme values or significant skewed distributions would have indicated regions with accumulation of higher or lower coverage, but none of these are observed. Thus degradation seems to have been uniform throughout the genome.

3.3. Coverage distribution between chromosomes

The genome has in total only ~6% heterochromatin, but heterochromatin is not distributed evenly between and within the chromosomes. According to International Human Genome Sequencing Consortium (IHGSC) chromosomes 1, 9, 13, 14, 15, 16, 17, 21, 22 and Y have higher proportions of heterochromatin with the Y-chromosome having the highest at 50 % heterochromatin [11]. In order to allow direct comparison between samples each chromosomal coverage value was normalized. This was done by calculating the chromosomal coverage as a percentage of the sum of all chromosomal coverages in the same sample. See Fig. 4 for bar plots.

Fig. 4 shows that the coverage is evenly distributed between chromosomes. Most of the heterochromatin regions are positioned near the centromeres and telomeres and have long repetitive sequences which results in gaps in the reference genome. It is therefore not possible, to measure degradation directly from coverage in these areas. From the IHGSC data the remaining heterochromatin regions represent a relatively small part of the chromosomes and was not expected to show a significant effect in Fig. 4. In females, due to X-inactivation, we would expect a significant increase in X-chromosome coverage compared to the other chromosomes if heterochromatin protects against DNA degradation. However Fig. 4 shows that the coverage of chromosome X for the female sample (HeavilyDeg2) was equal to the other chromosomes in the sample and to the X-chromosome in the female control (Control1). Lower coverage than expected was observed for the Y-chromosome in the male samples (HeavilyDeg1, Degraded1 and Degraded2). As coverage values in non-repetitive regions of X- and Y-chromosomal are similar for these samples we expect the lower overall coverage for the Y-chromosome to be caused by the relatively complex structure of

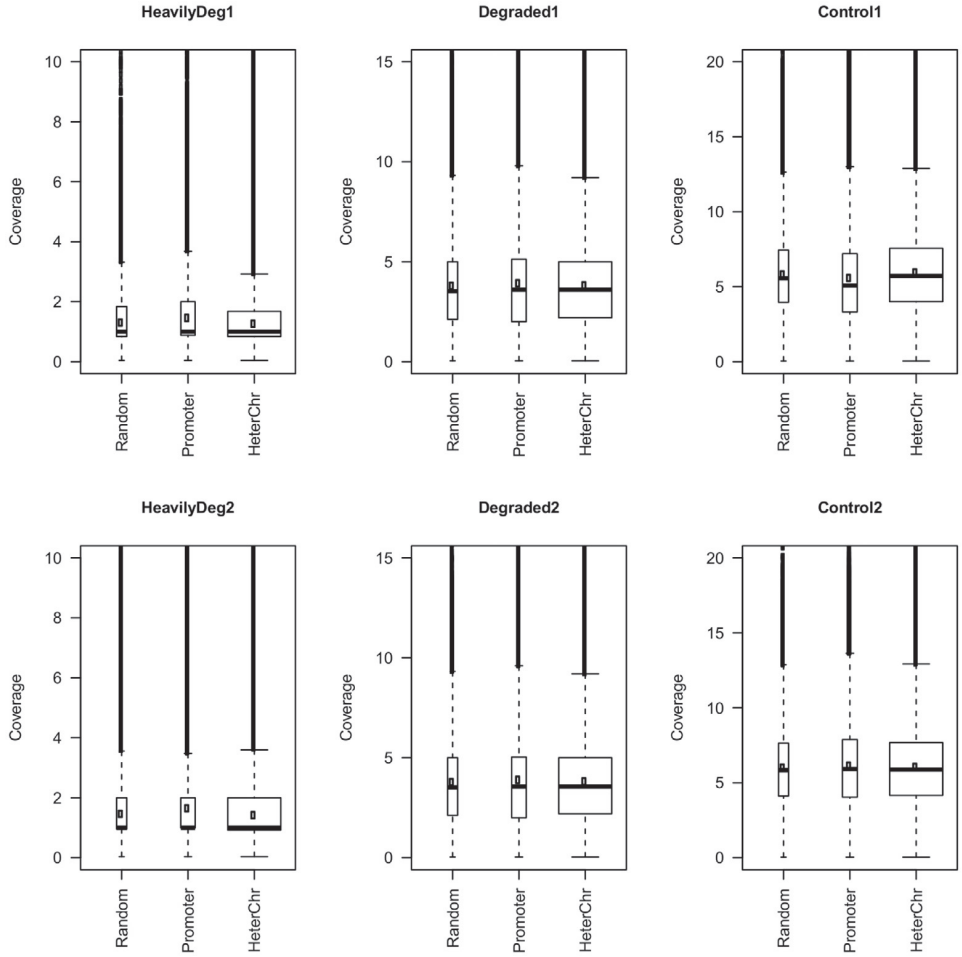

Fig. 5. For whole blood (see heavily degraded samples left) there is no change in coverage between regions with open chromatin structure (promoter), condensed chromatin structure (heterochrom) or randomly chosen regions. Sperm (see degraded samples in the middle) has a different chromatin structure than blood which is highly condensed in a large proportion of the genome. The steady trend over the different genetic regions is also seen for sperm. These results support the contention that DNA degradation is independent of higher order chromatin structure in biological trace material as opposed to DNA damage in living cells. The controls (see right) have an even coverage across the tracks as expected when the DNA is not degraded. For each box the dot represents the mean and the width the relative number of data points.

longer repeats in Y-chromosome [41]. The fact that the same trend is seen in the male control (Control2) also support that this has another cause than a higher degradation level of Y-chromosome. The female sample (HeavilyDeg2) and control (Control1) show small portions of reads mapped to the Y chromosome. These reads originate from the X-chromosome, but have mapped to homologous regions on the Y-chromosome instead. However their contribution is too small to have effect on the coverage values for the X-chromosomes.

3.4. Coverage of different genomic regions

The highly condensed constitutive heterochromatin is believed to occur at the same positions in all cell types and its position correlates with the trimethylation of histone H3 on lysine 9 (H3K9me3) [42]. The heterochromatic regions' counterpart is the promoter regions of expressed genes with an open chromatin structure to facilitate the transcriptional machinery's access to the DNA strand. It is expected that most promoter regions lie within the first 2 kb upstream of genes.

The boxplots in Fig. 5 indicate no difference in the coverage between the genomic regions investigated for all the degraded samples. No obvious difference is seen between samples and controls either. Due to the high number of data points (several 100,000 per boxplot) no statistical test was applied, but the distribution of the pairwise difference between coverage values of two and two tracks all had a mean close to zero. For blood (the heavily degraded samples) the heterochromatin and the promoter regions represent the two extremes of chromatin structure. Thus there is no evidence that more condensed chromatin structure of heterochromatin prevents DNA degradation. Sperm (the degraded samples) has a different and more condensed chromatin structure throughout the large proportion of the genome. Even though the annotation tracks are derived from whole blood data and do not represent different chromatin structures of semen, they still represent different genetic regions.

There is a trend in the boxplots of increasing skewness for the heavily degraded samples, but this might be caused by the relatively low coverage. However the distribution of the coverage for all tracks is narrow and comparable with the distributions seen when comparing the coverage between samples and controls above. The boxplots also show a number of outliers which may be a result of the high numbers of datapoints.

3.5. Coverage in regions of strongly-positioned nucleosomes

Nucleosomes are dynamic and move along the DNA strand, but around ~10% have been reported by Gaffney et al. to have moderate to strong positioning in seven human lymphoblastoid cell lines [43]. Pedersen et al. have sequenced DNA from ancient hair shafts and reported a pattern of coverage peaks corresponding to the strongly-positioned nucleosomes for a specific region in chromosome 12 [44]. The positioning in this region was postulated by Gaffney to be independent of cell type. Pedersen hypothesized that the coverage pattern was caused by nucleosomes protecting the DNA from degradation. The coverage at the same location for the four degraded samples were extracted using Igvtools with a 1 bp window size (position chr12:34443733-34453733 in hg19 converted from hg18 using the liftover tool in UCSC genome browser: https://genome.ucsc.edu). No obvious repetitive pattern was observed within or between the degraded samples. However the coverage might be too low to observe such a pattern with maximum ~3× compared to the 20× used in the Pedersen study. In addition semen, with its different chromatin structure, might not have strongly-positioned nucleosomes at the same locus.

4. Conclusion

In this study we have investigated DNA degradation in forensically relevant trace samples by massively parallel sequencing. The aim was to determine if parts of the genome are more resistant to degradation than others. The answer would be relevant for choosing future forensic markers. Although the number of samples was limited, several different approaches to measure degradation from sequencing data supports the hypothesis that degradation is uniform throughout the genome. A possible explanation is dissociation of protein and DNA in degraded forensic trace material. In addition, we demonstrate that whole-genome sequencing is applicable for forensic trace samples with minute amounts of degraded DNA. For the future, this indicates the applicability of a broader range of MPS applications beyond the targeted sequencing currently used in forensics.

Ethics

Only anonymized samples donated as research material were used in this study. No variant calling or personal trait information has been derived from the sequence data as only coverage data has been relevant. The study has been approved by the local Data Protection Official for Research.

Acknowledgement

This work was supported by the Norwegian Institute of Public Health.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.01.002.

References

[1] P. Martín, O. García, C. Albarrán, P. García, A. Alonso, Application of mini-STR loci to severely degraded casework samples, Int. Congr. Ser. 1288 (April) (2006) 522–525.
[2] V.C. Tucker, A.J. Hopwood, C.J. Sprecher, R.S. McLaren, D.R. Rabbach, M.G. Ensenberger, J.M. Thompson, D.R. Storts, Developmental validation of the PowerPlex® ESX 16 and PowerPlex® ESX 17 Systems, Forensic Sci. Int. Genet. 6 (January (1)) (2012) 124–131.
[3] R.L. Green, R.E. Lagacé, N.J. Oldroyd, L.K. Hennessy, J.J. Mulero, Developmental validation of the AmpFlSTR® NGM SElect™ PCR Amplification Kit: a next-generation STR multiplex with the SE33 locus, Forensic Sci. Int. Genet. 7 (January (1)) (2013) 41–51.
[4] C. Børsting, S.L. Fordyce, J. Olofsson, H.S. Mogensen, N. Morling, Evaluation of the Ion Torrent™ HID SNP 169-plex: a SNP typing assay developed for human identification by second generation sequencing, Forensic Sci. Int. Genet. 12 (September) (2014) 144–154.
[5] K.B. Gettings, K.M. Kiesler, P.M. Vallone, Performance of a next generation sequencing SNP assay on degraded DNA, Forensic Sci. Int. Genet. 19 (November) (2015) 1–9.
[6] M.E. Allentoft, M. Collins, D. Harker, J. Haile, C.L. Oskam, M.L. Hale, P.F. Campos, A. Jose, M. Samaniego, T.P. Gilbert, E. Willerslev, G. Zhang, R. Paul Scofield, R.N. Holdaway, M. Bunce, The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils, Proc. R. Soc. B 279 (December (1748)) (2012) 4724–4733.
[7] S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, S. Pääbo, Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA, PLoS ONE 7 (March (3)) (2012) e34131.
[8] J. Dabney, M. Meyer, S. Pääbo, Ancient DNA damage, Cold Spring Harb. Perspect. Biol. 5 (July(7)) (2013).
[9] B. Rydberg, Radiation-induced DNA damage and chromatin structure, Acta Oncol. 40 (6) (2001) 682–685.
[10] A.H. Wyllie, Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation, Nature 284 (5756) (1980) 555–556.
[11] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (October (7011)) (2004) 931–945.
[12] E. Eisenberg, E.Y. Levanon, Human housekeeping genes, revisited, Trends Genet. 29 (October (10)) (2013) 569–574.
[13] M. Ljungman, P.C. Hanawalt, Efficient protection against oxidative DNA damage in chromatin, Mol. Carcinog. 5 (4) (1992) 264–269.

[14] P. Trojer, R. Danny, Facultative heterochromatin: is there a distinctive molecular signature? Mol. Cell 28 (October (1)) (2007) 1–13.
[15] J.W.K. Ho, Y.L. Jung, T. Liu, B.H. Alver, S. Lee, K. Ikegami, K.-A. Sohn, A. Minoda, M.Y. Tolstorukov, A. Appert, S.C.J. Parker, T. Gu, A. Kundaje, N.C. Riddle, E. Bishop, T.A. Egelhofer, S.S. Hu, A.A. Alekseyenko, A. Rechtsteiner, D. Asker, J.A. Belsky, K. Sarah, Q. Bowman, B. Chen, R.A.-J. Chen, D.S. Day, Y. Dong, A.C. Dose, X. Duan, C.B. Epstein, S. Ercan, E.A. Feingold, F. Ferrari, J.M. Garrigues, N. Gehlenborg, P.J. Good, P. Haseley, D. He, M. Herrmann, M.M. Hoffman, T.E. Jeffers, P.V. Kharchenko, P. Kolasinska-Zwierz, C.V. Kotwaliwale, N. Kumar, S.A. Langley, E.N. Larschan, I. Latorre, M.W. Libbrecht, X. Lin, R. Park, M.J. Pazin, H.N. Pham, A. Plachetka, B. Qin, Y.B. Schwartz, N. Shoresh, P. Stempor, A. Vielle, C. Wang, C.M. Whittle, H. Xue, R.E. Kingston, J.H. Kim, B.E. Bernstein, A.F. Dernburg, V. Pirrotta, M.I. Kuroda, W.S. Noble, T.D. Tullius, M. Kellis, D.M. MacAlpine, S. Strome, S.C.R. Elgin, X.S. Liu, J.D. Lieb, J. Ahringer, G.H. Karpen, P.J. Park, Comparative analysis of metazoan chromatin organization, Nature 512 (7515) (2014) 449–452.
[16] S.W. Ward, Function of sperm chromatin structural elements in fertilization and development, Mol. Hum. Reprod. 16 (January (1)) (2010) 30–36.
[17] S. Kuretake, Y. Kimura, K. Hoshi, R. Yanagimachi, Fertilization and development of mouse oocytes injected with isolated sperm heads, Biol. Reprod. 55 (January (4)) (1996) 789–795.
[18] P. Thanakiatkrai, L. Welch, Evaluation of nucleosome forming potentials (NFPs) of forensically important STRs, Forensic Sci. Int. Genet. 5 (4) (2011) 285–290.
[19] A. Freire-Aradas, M. Fondevila, A.-K. Kriegel, C. Phillips, P. Gill, L. Prieto, P.M. Schneider, A. Carracedo, M.V. Lareu, A new SNP assay for identification of highly degraded human DNA, Forensic Sci. Int. Genet. 6 (3) (2012) 341–349.
[20] J.C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, Nucleic Acids Res. 36 (September (16)) (2008) e105.
[21] Y. Benjamini, T.P. Speed, Summarizing and correcting the GC content bias in high-throughput sequencing, Nucleic Acids Res. 40 (January (10)) (2012) e72.
[22] T. Derrien, J. Estellé, S.M. Sola, D.G. Knowles, E. Raineri, R. Guigó, P. Ribeca, Fast computation and applications of genome mappability, PLoS ONE 7 (January (1)) (2012) e30377.
[23] D. Sims, I. Sudbery, N.E. Ilott, A. Heger, C.P. Ponting, Sequencing depth and coverage: key considerations in genomic analyses, Nat. Rev. Genet. 15 (February (2)) (2014) 121–132.
[24] L. Deleye, D. De Coninck, C. Christodoulou, T. Sante, A. Dheedene, B. Heindryckx, E. Van den Abbeel, P. De Sutter, B. Menten, D. Deforce, F. Van Nieuwerburgh, Whole genome amplification with SurePlex results in better copy number alteration detection using sequencing data compared to the MALBAC method, Sci. Rep. 5 (June) (2015).
[25] H.P.J. Buermans, J.T. den Dunnen, Next generation sequencing technology: advances and applications, Biochim. Biophys. Acta 1842 (October (10)) (2014) 1932–1941.
[26] P. Gill, A.J. Jeffreys, D.J. Werrett, Forensic application of DNA 'fingerprints', Nature 318 (December (6046)) (1985) 577–579.
[27] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data, Bioinformatics (April) (2014) btu170.
[28] T. Magoč, S.L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies, Bioinformatics 27 (November (21)) (2011) 2957–2963.
[29] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.
[30] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.
[31] F. Ramírez, F. Dündar, S. Diehl, B.A. Grüning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data, Nucleic Acids Res. 42 (July (Web Server issue)) (2014) W187–W191.
[32] J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, Integrative genomics viewer, Nat. Biotechnol. 29 (January (1)) (2011) 24–26.
[33] ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (September (7414)) (2012) 57–74.
[34] M.R. Breese, Y. Liu, NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets, Bioinformatics 29 (February (4)) (2013) 494–496.
[35] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.
[36] D. Adams, L. Altucci, S.E. Antonarakis, J. Ballesteros, S. Beck, A. Bird, C. Bock, B. Boehm, E. Campo, A. Caricasole, F. Dahl, E.T. Dermitzakis, T. Enver, M. Esteller, X. Estivill, A. Ferguson-Smith, J. Fitzgibbon, P. Flicek, C. Giehl, T. Graf, F. Grosveld, R. Guigo, I. Gut, K. Helin, J. Jarvius, R. Küppers, H. Lehrach, T. Lengauer, Åke Lernmark, D. Leslie, M. Loeffler, E. Macintyre, Mai FAntonello, J.H.A. Martens, S. Minucci, W.H. Ouwehand, P.G. Pelicci, H. Pendeville, B. Porse, V. Rakyan, W. Reik, M. Schrappe, D. Schübeler, M. Seifert, R. Siebert, D. Simmons, N. Soranzo, S. Spicuglia, M. Stratton, H.G. Stunnenberg, A. Tanay, D. Torrents, A. Valencia, E. Vellenga, M. Vingron, J. Walter, S. Willcocks, BLUEPRINT to decode the epigenetic signature written in blood, Nat. Biotechnol. 30 (3) (2012) 224–226.
[37] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (6) (2010) 841–842.
[38] M. Melé, P.G. Ferreira, F. Reverter, D.S. DeLuca, J. Monlong, M. Sammeth, T.R. Young, J.M. Goldmann, D.D. Pervouchine, T.J. Sullivan, R. Johnson, A.V. Segré, S. Djebali, A. Niarchou, The GTEx Consortium, F.A. Wright, T. Lappalainen, M. Calvo, G. Getz, E.T. Dermitzakis, K.G. Ardlie, R. Guigó, The human transcriptome across tissues and individuals, Science 348 (6235) (2015) 660–665.
[39] S. Pääbo, H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, M. Kuch, J. Krause, L. Vigilant, M. Hofreiter, Genetic analyses from ancient DNA, Annu. Rev. Genet. 38 (2004) 645–679.
[40] I. Scheinin, D. Sie, H. Bengtsson, M.A. van de Wiel, A.B. Olshen, H.F. van Thuijl, H.F. van Essen, P.P. Eijk, F. Rustenburg, G.A. Meijer, J.C. Reijneveld, P. Wesseling, D. Pinkel, D.G. Albertson, B. Ylstra, DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly, Genome Res. 24 (December (12)) (2014) 2022–2032.
[41] T.J. Treangen, S.L. Salzberg, Repetitive DNA and next-generation sequencing: computational challenges and solutions, Nat. Rev. Genet. 13 (November (1)) (2011) 36–46.
[42] N. Saksouk, E. Simboeck, J. Déjardin, Constitutive heterochromatin formation and transcription in mammals, Epigenetics Chromatin 8 (January) (2015).
[43] D.J. Gaffney, G. McVicker, A.A. Pai, Y.N. Fondufe-Mittendorf, N. Lewellen, K. Michelini, J. Widom, Y. Gilad, J.K. Pritchard, Controls of nucleosome positioning in the human genome, PLoS Genet. 8 (November (11)) (2012) e1003036.
[44] J.S. Pedersen, E. Valen, A.M. Vargas Velazquez, B.J. Parker, M. Rasmussen, S. Lindgreen, B. Lilje, D.J. Tobin, T.K. Kelly, S. Vang, R. Andersson, P.A. Jones, C.A. Hoover, A. Tikhonov, E. Prokhortchouk, E.M. Rubin, M. Albin Sandelin, T.P. Gilbert, A. Krogh, E. Willerslev, L. Orlando, Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome, Genome Res. 24 (January (3)) (2014) 454–466.

Paper II

Forensic Science International: Genetics 30 (2017) 10–17

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Body fluid prediction from microbial patterns for forensic application

Eirik Nataas Hanssen[a,b,*], Ekaterina Avershina[c], Knut Rudi[c], Peter Gill[a,b], Lars Snipen[c,**]

[a] Department of Forensic Biology, Oslo University Hospital, Oslo, Norway
[b] Department of Forensic Medicine, University of Oslo, Oslo, Norway
[c] IKBM, Norwegian University of Life Sciences, Aas, Norway

ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 28 May 2017
Accepted 29 May 2017
Available online 1 June 2017

Keywords:
Body fluid prediction
Microbiome
Massively parallel sequencing
Principal component analysis
Linear discriminant analysis

ABSTRACT

The association of a DNA profile with a certain body fluid can be of essential importance in the evaluation of biological evidence. Several alternative methods for body fluid prediction have been proposed to improve the currently used presumptive tests. Most of them measure gene expression. Here we present a novel approach based on microbial taxonomic profiles obtained by standard 16S rRNA gene sequencing. We used saliva deposited on skin as a forensically relevant study model, but the same principle can be applied for predicting other bacteria rich body fluids. For classification we used standard pattern recognition based on principal component analysis in combination with linear discriminant analysis. A cross-validation of the experimental data shows that the new method is able to successfully classify samples from saliva deposited on skin and samples from pure skin in 94% of the cases. We found that there is a person-effect influencing the result, especially from skin, indicating that a reference sample of pure skin microbiota from the same person could improve accuracy. In addition the pattern recognition methods could be further optimized. Although there is room for improvement, this study shows the potential of microbial profiles as a new forensic tool for body fluid prediction.

1. Introduction

The short tandem repeat DNA-profile is used to identify the person from which a biological trace originates. However, the DNA-profile provides no information on how the trace was deposited. In this context, information on type of body fluid can be of crucial importance when evaluating biological evidence.

Classical presumptive tests are still preferred methods for body fluid prediction in many laboratories [1]. Generally, a body fluid specific enzyme catalyzes a chemical reaction, and the result is visually detected, often as a color change. The tests are fast and easy to use, but have high error rates as the target enzyme is also present in low quantities in other body fluids [2,3]. In addition, common household items and chemicals can give false results [4]. A few immunochromatographic lateral flow strip tests are commercially available as an alternative [5,6], but even if these tests are more specific, the presence of enzymes in other body fluids still give false positive results for some tests. Traditionally, no probabilistic statements have been associated with a presumptive test result.

Lately several alternative detection technologies have been reported to be applicable for body fluid prediction [7,8]. Most of these measure gene expression using mRNA, miRNA or epigenetic markers. The European DNA Profiling Group (EDNAP) has performed collaborative studies on mRNA tests for blood [9], saliva and semen [10] menstrual blood and vaginal secretion [11] and finally on skin [12]. The chemically more stable miRNA markers can also be used to differentiate between body fluids [13–15]. However, both RNA methods currently lack a reliable quantification method. In addition, there is often a large difference in abundance between the RNA makers within a sample. This is especially challenging for minor components in mixed samples where it can be difficult to separate between real and background signal. Another approach is to measure degree of methylation at GpC islands. Although this has been promising [13,14], it is not yet ready for implementation in casework as methylation levels can differ between individuals, tissues and exhibit age or environmental dependency [8]. Since none of the aforementioned methods will detect and separate all body fluids, it has been proposed to combine their use, but even then some mixtures might be challenging [7].

* Corresponding author at: Department of Forensic Biology, Oslo University Hospital, Oslo, Norway.
** Principal corresponding author.
E-mail addresses: kjeirik@gmail.com (E.N. Hanssen), lars.snipen@nmbu.no (L. Snipen).

http://dx.doi.org/10.1016/j.fsigen.2017.05.009
1872-4973/

As an alternative to gene expression measurements, microbial markers have been proposed as a way to discriminate between various body fluids [7,8]. The main idea is to look for the taxonomic composition of bacteria in the various body fluids, and recognize them based on specific patterns in this composition. The standard genetic marker for taxonomic profiling of microbial communities is the small subunit ribosomal RNA gene, also known as the 16S gene. Large data repositories specifically devoted to 16S rRNA gene data exist, e.g. the Silva database (https://www.arb-silva.de/), the Ribosomal Database Project (RDP, https://rdp.cme.msu.edu/) and the Greengenes database (http://greengenes.lbl.gov/). Microbiota-based body fluid recognition is most likely best suited for bacteria-rich body fluids such as saliva, vaginal secretion, feces and menstrual blood, while sterile or nearly sterile body fluids such as blood, semen and tears are probably more problematic to recognize [16]. Other limitations may be geographical variation [17] and drug use [18], but for a large proportion of cases such limitations can be ruled out.

Since inception of the Human Microbiome Project (HMP) [19], many efforts have been made to study the human microbiota by amplicon sequencing of the 16S rRNA gene. While most such studies have been health related and targeted the human gut, there are also some studies with a forensic focus. In [20,21] a search for body fluid-specific taxonomic markers was conducted, but with a negative result. Although it would be convenient to have unique markers to identify a specific body fluid, this may be unrealistic. The *taxonomic profile* of a given body fluid is a vector of quantitative values describing the bacterial composition. Provided that there is sufficient specificity, body fluids can be identified. A large variety of multivariate pattern recognition approaches are already available for the data analysis part of this problem. Such methods have already been used to separate microbiota from phones and shoes [22] and could even potentially be used to identify persons based on skin samples [23].

In microbiome sequencing two major sources of bias have been thoroughly discussed in the literature. First, different DNA extraction method may have an impact on microbial community profiling [24–28]. However there seems to be consensus on that a bead-beating step increases the yield, and that the same extraction protocol should be used throughout a study to ensure reproducibility. The other source of bias is PCR amplification which can result in artifacts such as chimeras [29] and skewed fragment distributions [30–32]. Chimeras form when short aborted extension products function as primers in later PCR cycles to create full length artificial fragments. Chimeras and other PCR artifacts are problematic for bacteria rich samples, but little is known about artifact formation in samples with low levels of bacterial DNA. Digital droplet PCR (ddPCR) use micro droplets as reaction chambers with just one or a few fragments in each droplet. This results in unbiased amplification (see Droplet Digital PCR Applications Guide at www.bio-rad.com).

Health related microbiota studies have investigated pure body fluids sampled directly from the human body. In a forensic context the conditions will be different and care should be taken when adopting standardized lab protocols and bioinformatics workflows. Biological traces are typically collected with cotton swabs [33] and stored in dry state until analysis [34]. Most trace samples will have relatively low bacterial levels and require highly sensitive methods [35] and appropriate routines to prevent contamination [31]. Low bacterial levels might also enhance different biases e.g. in the sampling [36,37] and PCR amplification [29,30,32] steps. Trace samples are rarely single source, but often mixtures of different body fluids. In addition the data interpretation should not be exploratory as in many health studies, but based on pattern recognition.

In this paper we present a study where we have investigated potential effects of sampling and lab-protocols on the detection and recognition of saliva deposited on human skin. This is a typical example of a biological trace from a crime scene, and to our knowledge the first study to demonstrate the identification of body fluids from microbiota data in this context.

2. Materials and methods

2.1. Experimental setup

Six healthy persons participated in this study. They were told not to eat or wash hands during a period of 1 h before the experiment. Traces of both pure and diluted saliva were deposited between the base of the fingers on the back of each participants hands. The liquid was smeared in the sampling areas using the pipette tip and then dried for 10 min before sampling. The experiment was designed so that each of the six individuals had saliva donated from another participant deposited onto their hands (one donation per participant). All experiments were performed on the same day.

The following samples were collected from each participant: (1) Pure saliva sampled directly from the mouth (to be deposited on another participant), (2) the trace consisting of 20 μL pure saliva deposited between fingers, (3) the trace consisting of 20 μL saliva diluted in PCR water (1:10) deposited between fingers and (4) a sample from pure skin between fingers.

Initially three different sampling techniques were evaluated. 20 μL saliva were applied onto cotton swabs (Medical Wire), synthetic swab (DNA Genotek) and tape (Scenesafe) and processed in parallel with 20 μL pure saliva. Bacterial DNA extraction was performed as described below, and recovery was measured for all three techniques. The use of the cotton swab was discontinued based on the results. Tape was used when sampling the left hand and synthetic swabs when sampling the right hand. Diluted saliva was only collected by tape from the left hand. Thus, we define 6 different types of samples:

1. Pure saliva from mouth.
2. Saliva deposited on skin, collected with tape.
3. Saliva deposited on skin, collected with swab.
4. Diluted saliva on skin, collected with tape.
5. Pure skin, collected with tape.
6. Pure skin, collected with swab.

One droplet of PCR grade water was added to the swab before sampling to mimic standard procedure [38].

2.2. Soaking, extraction and quantification

The samples were first soaked to release the sample material. The tape was cut with a sterile razor before being transferred to a 1.5 mL Eppendorf tube with 200 μL S.T.A.R. buffer (Roche Diagnostics). The synthetic swab was placed in the associated tube containing 1 mL soaking solution. The Eppendorf tubes were put on a horizontal shaker at 1400 rpm and 56 °C for 30 min while the tubes with the synthetic swab was briefly vortexed according to producers recommendations. For each sample 150 μL soaking solution was transferred to a 2 mL conical tube (Sarstedt) with approximate 0.24 g acid-washed glass beads (<106 μm; Sigma Aldrich). The samples were homogenized at 1800 rpm for 2 × 30 s using FastPrep96 (MPBio) and then centrifuged at 13,000 *g* for 5 min. DNA was extracted using LGC mag midi kit (LGC Genomic) following the manufacturer's recommendations. The resulting DNA extracts were quantified by digital droplet PCR (Bio-Rad

QX200) using PRK341F and PRK806R primers targeting the V3–V4 region of the prokaryotic 16S rRNA gene.

2.3. Library preparation and sequencing

16S rRNA gene was sequenced using Illumina-modified PRK341F and PRK806R primer pair following the "Illumina 16S metagenomic sequencing library preparation" protocol. The protocol is available at http://illumina.com.

To evaluate the effect of PCR technique, we used both conventional PCR (Applied Biosystems 2720) and ddPCR (Bio-Rad QX200) techniques for the first step of 16S rRNA gene enrichment. 30 and 40 cycles were used respectively. The amplification product from the ddPCR was recovered by breaking the emulsion according to the Droplet Digital PCR Applications Guide at www.bio-rad.com (page 70). All PCR products were amplified a second time with Illumina-modified primers. Finally the samples were pooled in equimolar volumes and paired-end sequenced on a Miseq platform using v3 chemistry. In total we sequenced 144 samples; 36 extracts amplified in duplicates for each of the two PCR techniques. We define the duplicates as technical replicates. In addition 5 positive (*Escherichia coli*) and 13 negative controls was sequenced. The experimental setup is illustrated in supplementary figure S1.

2.4. Taxonomic profiles

Conventional bioinformatics workflows were used to cluster all reads into operational taxonomic units (OTUs). Our workflow was built around the open source software VSEARCH [39]. First, read-pairs were merged, including the filtering of low quality reads (maxee = 1, see [40]). The de-multiplexing step included removal of non-biological barcode and primer sequences. After de-replication, all singleton sequences were removed before clustering using the standard 97% identity value. The resulting OTU-centroid sequences were chimera-filtered using the gold database of the ChimeraSlayer utility [41]. OTU centroids were given taxonomic assignments using the taxMachine tool [42].

For each sample, a read-count vector was found by searching with all reads against the OTU centroids, using the standard 97% identity threshold. The taxonomic profile of a sample was found by dividing this vector by the total read-count for the sample, producing relative read-counts for each OTU. Finally all taxonomic profiles were stacked into an OTU-matrix, one row for each sample.

2.5. Data analysis

Data analysis was performed in the R computing environment [43] and in MATLAB 2016b (The MathWorks Inc., Natick, MA, 2000). All software used is listed in supplementary table S1.

Chimera were detected using VSEARCH as described above. The proportion of chimera in each sample was used in an Analysis of Variance (ANOVA) to investigate if there was significant difference in chimera formation between sample types or between the two PCR techniques.

Alpha-diversity for each sample was calculated as Simpson's index of diversity (1-D) using the Vegan R-package [44]. Diversity is dependent on the number of reads, and rarefaction is typically used to remove this bias [45]. We randomly picked 1000 reads from each sample using the "sample" function in R, and then employed these reads as input for the "diversity" function of the Vegan package. ANOVA was again used to investigate if there was significant difference in diversity between sample types, the two PCR techniques or the two sampling techniques.

Principal component analysis (PCA) was used to project the high-dimensional taxonomic profiles onto a lower dimensional space, used for graphical display or further data analysis. Principal components are ordered by variance in the data set, i.e. the first component is the linear combination of taxa with most variation. Thus, the first components are most likely to contain any systematic variation in the data. We used the first principal component scores as response in an ANOVA to evaluate the effect of person, sampling technique, sample type, technical replicate and PCR technique. Main effects were included, but also the interaction between technical replicate and PCR technique. This ANOVA was also repeated for the second and third principal component.

We also investigated the reproducibility of the data across persons with respect to pattern recognition of the taxonomic profiles. Samples were categorized into 2 classes, either Saliva (sample types 1–4) or Skin (sample types 5–6) – see Section 2.1 for definition of sample types. Data were split by person, and a cross-validation was performed by training a classifier on data from 5 persons, and then predicting the class of the samples from the left out person. As classifier we used the linear discriminant analysis (LDA) implemented in the MASS-package in R. Due to the large number of OTUs we have many more predictor variables than samples in this data set, which permits the straightforward use of LDA. This was performed by using PCA on the training data first, and then truncated the data to 6 principal components, using these scores when training the LDA classifier. The exact same centering/rotation was used on the test-data during prediction. For further reading on the subject of statistical learning and prediction we refer to Hastie et al. [46].

We also compared our taxonomic profiles to those of similar samples from the HMP. From the public data at http://hmpdacc.org/ we downloaded reads for all samples. Only samples amplifying the V3–V5 region of the 16S gene were used, as these overlap with our data (V3–V4 region). We considered only samples taken from the mouth or the skin, and categorized these as Saliva and Skin. In total, this resulted in data from 2465 samples. Next, we used our VSEARCH pipeline to find OTUs in these data, and constructed an OTU-matrix in a similar way as described above. Then these centroid sequences were used as a database, and we constructed a new OTU-matrix for our experimental data, matching the reads against the HMP centroids and assigning them to the same OTUs. Finally, the OTU-matrices from HMP and our experiments were assembled into one matrix, and a PCA was conducted on this joint matrix.

3. Results

In these experiments we investigated if pure and diluted saliva can be recognized by the microbiota taxonomic profiles in traces collected from the hands of persons. We also tested various ways of sample collection and PCR amplification, and compared the obtained profiles to those in the public data sets from the HMP.

3.1. Microbiota raw data quality

It is reasonable to expect that it is more difficult to collect sufficient quantities of microbial DNA from a deposited trace compared to a sample taken directly from a body fluid. If the trace is deposited on human skin, it is also infested by the natural skin-microbiota, potentially masking or distorting the body fluid profile.

First, we tested three ways of sampling biological traces (cotton swab, synthetic swab and tape). Median recovery for synthetic swab and tape was determined from ddPCR quantification results to be ~90% and ~60%, respectively (swab recovery had to be adjusted for a 5× larger soaking volume than tape). See Fig. 1 for details. In contrast, the cotton swab only had a few percent recovery and was therefore not suitable for purpose. This was

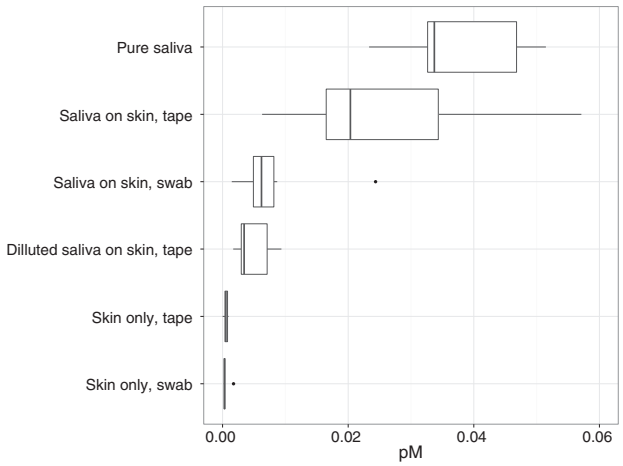

Fig. 1. Bacterial DNA concentration (pM) in extracts for the different sample types. The swab samples were soaked in a 5× larger volume than the tape samples and this has to be adjusted for when evaluating recovery. Quantification was done by ddPCR, and data for all 6 participants are included in the boxplot ($n = 144$).

disappointing as cotton swabs have a widespread used in general casework. However, in that respect, both synthetic swabs and tape have been proven to perform equivalent to cotton swabs [47,48].

Fig. 1 shows the bacterial DNA concentration of the extracts for the six different sample types. Data from each of the six participants is included. Pure saliva sampled directly from the mouth clearly contains the most, and even the diluted saliva samples have considerable higher bacterial density than the samples from skin only.

The OTU-finding workflows will discard a proportion of reads based on various quality criteria. In our workflow, on average, 18% of the read-pairs were impossible to merge, and thereby lost. Next, on average 7% and 16% of the reads had no detectable barcode or primer sequences, and were discarded. In total an average of 64% of the original reads ended up in the final data set, with a small variation between the various fastq-files. These numbers are in line with other studies [49]. For details see supplementary table S2.

The chimera content was lower than ~9% for all samples, with the pure saliva samples having a tendency of a larger chimera proportion than the other sample types ($p = 0.0006$). See supplementary figure S5 for details. There was a large variation in the number of reads between the samples in the final OTU table. 22 samples had a read count below 1000, and 2 of these were below 100. At the other end of the scale 10 samples had over 100,000 reads. See supplementary figure S2. The effect of the two PCR techniques was evaluated using both chimera proportion and diversity as response. ANOVA showed that the PCR techniques had no significant effect on either chimera proportion or the Simpson's index of diversity ($p > 0.5$). However ddPCR tended to have higher correlation between number of reads and initial DNA inputs and more reproducible results with less variation between technical replicates ($p = 0.002$). See supplementary figures S3 and S4, respectively.

3.2. Community characteristics

In total the OTU-finding workflow suggested 849 OTUs in the data set. Of these, 607 were detected in skin-samples and 735 in saliva. Assigning OTUs to a genus resulted in the skin samples containing 285 and the saliva samples 288 different genera. Only 17 (6.0%) and 19 genera (6.6%), respectively had a relative abundance more than 1%. See Fig. 2 for details.

The samples had a fairly high diversity with a median Simpson's index of diversity at 0.92. ANOVA showed no significant difference in diversity between the 6 sample types. See supplementary figure S6. The sampling technique had no significant main effect on

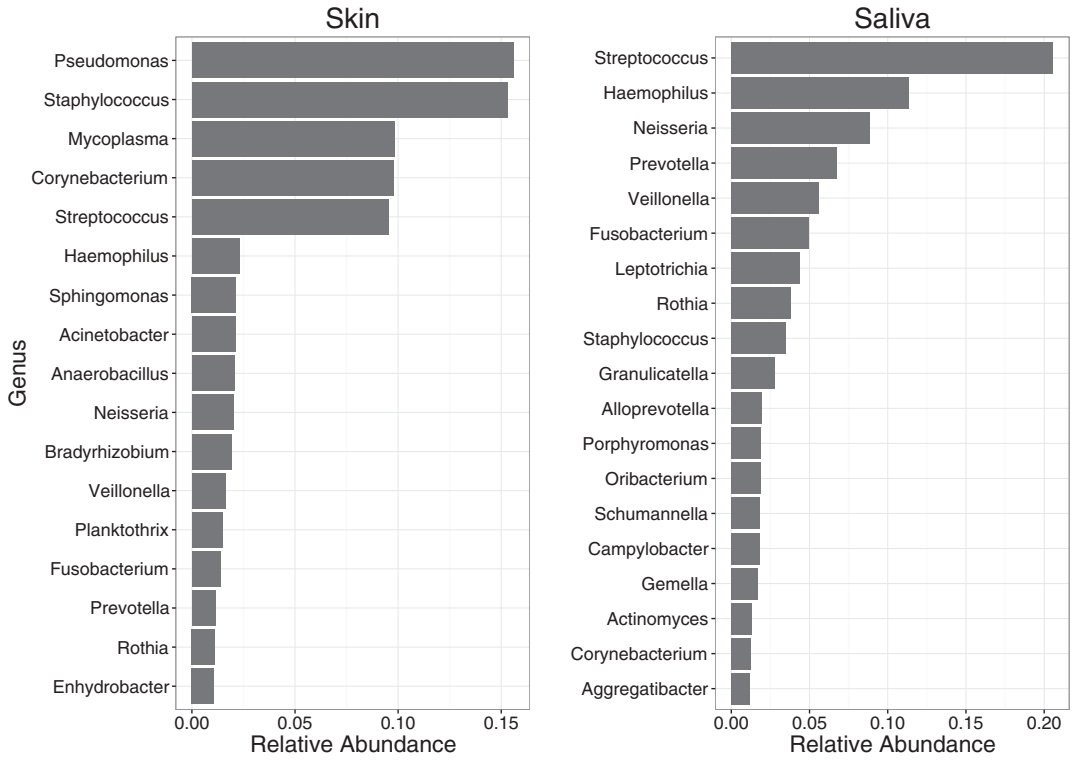

Fig. 2. Genera with an abundance over 1% in the skin and the saliva samples.

diversity, but tape showed a slight tendency to give higher diversity than synthetic swabs when sampled from skin ($p = 0.04$). See supplementary figure S7.

3.3. Taxonomic profiles

Taxonomic profiles are high-dimensional, and in order to analyze for differences in profiles between various samples, we first subjected the OTU-matrix to a PCA. The first principal components will contain the largest data variations, most likely the systematic differences between taxonomic profiles. The first three components contained 44%, 19% and 12% of the total variation in taxonomic profiles.

We successively used the scores of the first, second and third principal component as a response in an ANOVA, testing the main effects of different experimental factors. Using the first component the most significant difference was between pure saliva and skin ($p < 10^{-16}$). In addition, diluted saliva on skin was also significantly different from pure saliva ($p < 10^{-3}$). We also found that two of the persons were significantly different from the others ($p < 10^{-3}$). Sampling technique, PCR technique and technical replicate (including the interaction effect of the two last factors) had no effect on any of the first principal components taxonomic profiles. For the second component we observed the same clear effects as for the first component. In addition there was a weak significant effect from PCR technique ($p = 0.05$). For the third component the picture was a bit different. There was a weaker effect from sample type ($p = 0.009$), new persons had an effect and PCR technique had no effect. An overview of the ANOVA results is given in supplementary table S3.

In Fig. 3 the samples are plotted in a PCA-plot using the two first principal components. We first notice that saliva-samples (blue and cyan) tend to group with a small variation (beta-diversity) in the left part of the panel (negative scores on first principal component), while skin-samples (tan) are scattered much more, but mostly with positive scores along both axes. There are 4 notable exceptions, the cyan markers in the lower right corner. There is a weak tendency for diluted saliva on skin (cyan) to be located on the border between pure saliva (blue) and skin (tan). The samples amplified with ddPCR (triangles) and conventional PCR (circles) are randomly dispersed throughout. In addition supplementary figure S8 shows that control samples separate well from the experimental samples.

Ultimately the differences in taxonomic profiles are only relevant if they make it possible to recognize a body fluid in a trace. To investigate this, we trained the LDA classifier on data from 5 of the persons, and used this to classify samples from the last person. PCA scores for the 6 first components were used to train the LDA model. This was repeated in a 6-fold cross-validation, leading to body fluid predicted for each sample once. In Table 1 we show the results. In total, 135 out of 144 samples were correctly recognized as either saliva (on skin) or skin.

Table 1
Prediction results from cross-validation.

Person	Incorrect	Correct
Person 1	4	20
Person 2	1	23
Person 3	0	24
Person 4	0	24
Person 5	0	24
Person 6	4	20
Sum	9	135

Of the erroneously classified samples there were 5 false positives (skin classified as saliva) and 4 false negatives (saliva classified as skin). Only 2 of these were among the 10 samples with fewest reads. The false negatives were the cyan samples positioned in the lower right corner of Fig. 3. These are all from one person. The false positives were located on the border between saliva and skin. The posterior probabilities assigned by LDA, were for 4 of these samples weakly in favour of the wrong class (see supplementary figure S10 for details).

To compare our experimental data with a large data set of "pure" body samples, we assigned our reads to the OTUs based on 2465 HMP samples. In Fig. 4 we have again plotted all samples in a PCA plot. Samples from skin and saliva separate well in the HMP data, even if there is a huge diversity within each group. Skin samples tend to have a negative score along the first principal axis, while saliva have positive scores. Our experimental data have a notable shift in location compared to the HMP data, and are all located near the border between skin and saliva. The saliva samples are in most cases on the "correct" side, but skin samples are not.

4. Discussion

The aim of this study was to investigate how well we can recognize human saliva in a trace deposited on human skin, based on the taxonomic profile from amplicon sequencing of the 16S rRNA gene. We investigated the effect of several potential critical experimental factors, to identify some bottlenecks.

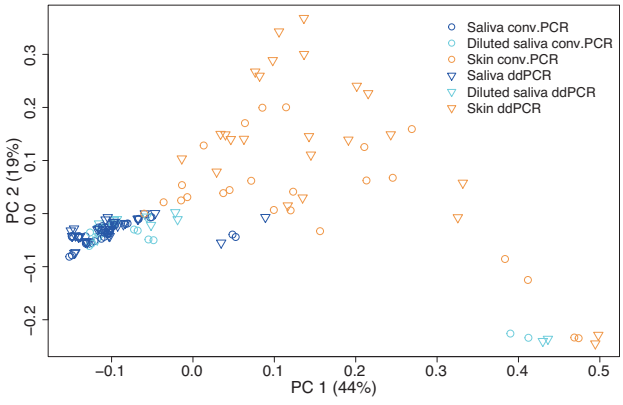

Fig. 3. The scores of the taxonomic profiles along the two first principal components. Each marker corresponds to a sample. The coloring and marker types indicate different sample types, as explained in the figure legend.

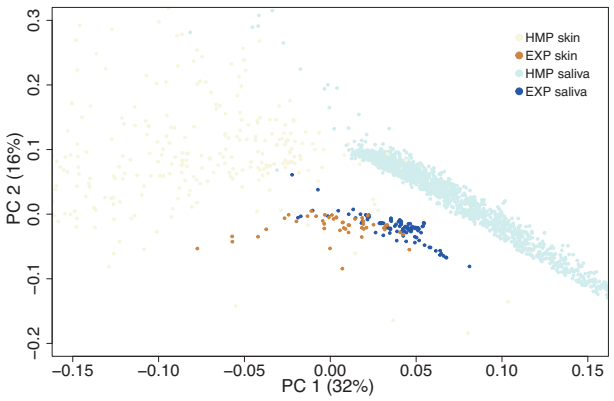

Fig. 4. PCA plot with both the experimental data and HMP data (zoomed in on experimental data). The saliva samples are more similar than the pure skin samples.

In order to discriminate saliva from skin, these sources need to have taxonomic profiles which are distinctly different and also sufficiently stable within each group. Since taxonomic profiles are multi-dimensional (many taxa), we used PCA to project each profile onto either the first, second or third principal axis, thereby obtaining a single value (score) for each profile along each axis. We used these values in an ANOVA to see which factors explain differences in taxonomic profiles best.

Of all factors analyzed in our ANOVA, the sample type (saliva or skin) had by far the largest significant effect. This is in line with earlier studies [19,50] and confirms the potential for using 16S profiles for this purpose. The largest difference was between pure saliva and pure skin, with saliva deposited on skin as an intermediate. This illustrates nicely how saliva deposited on skin shows a profile which is a mixture of the two pure sources, and therein lies the potential difficulty in recognizing saliva when contaminated with skin bacteria.

The only other really significant effect found was between persons. This means there is a bias due to person in the taxonomic profiles. This has also been observed earlier [19,50]. The reason for this is not fully understood, but diet could be one of several factors [51]. Hence the effect from person is clearly an important factor when evaluating method accuracy, but our results show that the effect from person in most cases will not influence the prediction result.

All other factors investigated (sampling technique, PCR-technique, parallels), resulted in small or no effects on the taxonomic profiles.

Sampling traces with either tape or synthetic swab seems to have no effect, and the cotton swab was deemed useless even before sequencing, since it did not collect sufficient amounts of DNA.

PCR artifacts might influence the profiles, and we included two types of PCR in our study. Even though PCR technique had a modest effect along the second principal component, the over-all picture is that it has little impact on the taxonomic profiles, and surprisingly little effect on chimera formation and alpha-diversity. This observation suggests that chimera formation and other PCR artifacts are less influential for samples with little DNA [30]. This might be explained by a dilution effect leading to lower probability for formation of PCR artifacts.

A reasonable high accuracy is supported by the cross-validation results. An actual pattern recognition was performed by fitting an LDA model using the PCA scores (component 1–6) as predictors and saliva or skin as categorical responses. A per-person cross-validation was used, i.e. all samples from one person were predicted based on training data from the other 5 persons. This resulted in ~94% of the samples which where correctly classified as either saliva or skin. In the interests of conservativeness, it is more important to avoid false positive results (detecting saliva when saliva is not present) rather than false negative results (not detecting saliva when saliva is present). 4 of the 5 false positive samples had intermediate posterior probabilities (0.7–0.9), indicating rather uncertain predictions. In a real case, we would probably include a third “inconclusive-category”, where uncertain predictions would be notified. Positioning of a threshold for this category could be guided on the basis of the confidence interval of posteriors for correctly predicted samples, but should be set to ensure a conservative prediction. The false negative samples were all from one single person, and were all samples of saliva on skin from this person. These samples had a profile where saliva was not at all recognized, based on how it looks in the other persons. Also, the skin-microbiota of this person was rather distinct, and the mixture (saliva on skin) was clearly dominated by this skin microbiota.

It is reasonable to expect that low read-counts would lead to low resolution in the taxonomic profile thereby potentially causing a prediction error. However this is not clearly supported by our results. Two of the experimental samples had very low read-counts, less than 100 reads. One of these was miss-classified, but most of the other low read-count samples were correctly classified (see supplementary figure S9). The ANOVA showed that read-counts did not have a significant effect on the profiles. Thus, for now read-counts as low as 100 gave similar results to much higher numbers of reads, but larger data sets should be investigated to verify that such low read-counts could be used for body fluid prediction in casework.

There are a number of other factors that we did not investigate and that might influence the data quality for the intended purpose. One factor is the extraction procedure. However, as we use a bead-beating step to optimize extraction performance, it is uncertain how large this effect could be. Another factor is sampling site on the skin. Other studies have shown that the microbiota differs considerably between even nearby body sites [36,52]. It is also worth noticing that even our own experimental skin samples, which are taken from between the fingers, are more diverse and do not cluster as tightly as the saliva samples in the PCA plot (see Fig. 3). This implies that skin might be relative challenging to predict, and especially hands. However the low level of bacterial background from skin (see Fig. 1) is an advantage for recognizing a deposited body fluid.

When comparing the experimental data to HMP data there is an obvious bias for both sample types (see Fig. 4), but the skin samples show a lower degree of similarity than the saliva samples. One obvious explanation for the skewed bias is that the HMP skin samples are taken from other body sites than our samples (elbow cleavage and behind the ear compared to between fingers). These body sites are known to have different microbial compositions [36,52]. The sampling technique is also different, but the deeper skin layers sampled for the HMP studies should not differentiate significantly in composition from our surface samples [53]. Another explanation for the observed bias could be that the HMP and our samples are collected at different geographical locations [50]. Factors such as extraction, library preparation and sequencing have also probably contributed to the bias, but if so, these would have influenced skin and saliva samples equally.

As mentioned above, bias between datasets from different studies is expected. Our results confirm this and emphasize the need for harmonization of protocols and data interpretation. However, these results also suggest the use of reference samples is something to consider for future method-development. By this, we mean samples taken from pure sources wherever it is possible. If we suspect saliva has been deposited on the skin of some person, much is gained by collecting samples from the same person's pure skin as close to site as possible. Even smallish amounts of saliva can be detected on a non-typical skin if we have the profiles of this non-typical skin.

For this study we have adopted standard bioinformatics workflows that have already been incorporated in health and environmental studies. Although accepted as suitable for such use, these methods can be further optimized to fit the problem of pattern recognition which is our purpose. Also, at present massively parallel sequencing is needed to survey all bacteria in the samples. In a longer perspective this does not need to be the case. As only bacteria relevant for body fluid prediction need to be detected, a customized multiplex in combination with quantitative PCR (qPCR) might be the most efficient approach. This will, however, require a systematic search for the discriminating taxa across a huge set of samples.

5. Conclusion

For this study we have used saliva deposited on skin as a study model, but the principle can be used for other bacteria rich body

fluids. We demonstrate that forensic trace samples can be sequenced by following a standard microbial 16S gene sequencing protocol, and that the resulting microbial composition profiles are applicable for discriminating saliva from skin. Experimental factors, such as PCR technology and sampling technique did not have any significant effect on the taxonomic profiles. A cross-validation of the experimental data gave correct classification in 94% of the cases. Although our results are promising, there are several aspects that needs to be investigated before the method can be used in casework. As we find a clear bias between our experimental data and the HMP database, it is evident that the final method needs standardized protocols and training datasets for calibration. We find that there is a person-effect influencing the taxonomic profiles, especially from skin, indicating that a method for detecting a body-fluid deposited on skin should always be accompanied by a reference-sample of pure skin microbiota from the same person. In addition the pattern-recognition methods could be optimized. Despite this, our results show the potential of microbial taxonomic profiles as a new forensic tool for body fluid prediction. The method will probably have limitations when it comes to classifying bacteria poor body fluids, but could, dependent on the final accuracy levels obtained, be a valuable supplement to other body fluid determination techniques.

Ethics

This study has been approved by the local Data Protection Official for Research at Norwegian Institute of Public Health.

Acknowledgements

This work was supported by Norwegian Institute of Public Health, Oslo University Hospital, University of Oslo and Norwegian University of Life Sciences.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2017.05.009.

References

[1] L. Gefrides, K. Welch, Forensic biology: serology and DNA, in: A. Mozayani, C. Noziglia (Eds.), The Forensic Laboratory Handbook Procedures and Practice, Humana Press, 2011, pp. 15–50.
[2] P.H. Whitehead, A.E. Kipps, A test paper for detecting saliva stains, J. Forensic Sci. Soc. 15 (1975) 39–42.
[3] F. Mannello, L. Condemi, A. Cardinali, G. Bianchi, G. Gazzanelli, High concentrations of prostate-specific antigen in urine of women receiving oral contraceptives, Clin. Chem. 44 (1998) 181–183.
[4] S.S. Tobe, N. Watson, N.N. Daid, Evaluation of six presumptive tests for blood, their specificity, sensitivity, and effect on high molecular-weight DNA, J. Forensic Sci. 52 (2007) 102–109.
[5] M.N. Hochmeister, B. Budowle, O. Rudin, C. Gehrig, U. Borer, M. Thali, R. Dirnhofer, Evaluation of prostate-specific antigen (PSA) membrane test assays for the forensic identification of seminal fluid, J. Forensic Sci. 44 (1999) 1057–1060.
[6] B.C.M. Pang, B.K.K. Cheung, Identification of human semenogelin in membrane strip test as an alternative method for the detection of semen, Forensic Sci. Int. 169 (2007) 27–31.
[7] T. Sijen, Molecular approaches for forensic cell type identification: on mRNA, miRNA, DNA methylation and microbial markers, Forensic Sci. Int. Genet. 18 (2015) 21–32.
[8] S. Harbison, R. Fleming, Forensic body fluid identification: state of the art, Res. Rep. Forensic Med. Sci. (2016) 11.
[9] C. Haas, E. Hanson, M.J. Anjos, W. Br, R. Banemann, A. Berti, E. Borges, C. Bouakaze, A. Carracedo, M. Carvalho, V. Castella, A. Choma, G.D. Cock, M. Dtsch, P. Hoff-Olsen, P. Johansen, F. Kohlmeier, P.A. Lindenbergh, B. Ludes, O. Maroas, D. Moore, M.-L. Morerod, N. Morling, H. Niedersttter, F. Noel, W. Parson, G. Patel, C. Popielarz, E. Salata, P.M. Schneider, T. Sijen, B. Svieena, M. Turansk, L. Zatkalkov, J. Ballantyne, RNA/DNA co-analysis from blood stains. Results of a second collaborative EDNAP exercise, Forensic Sci. Int. Genet. 6 (2012) 70–80.
[10] C. Haas, E. Hanson, M.J. Anjos, R. Banemann, A. Berti, E. Borges, A. Carracedo, M. Carvalho, C. Courts, G.D. Cock, M. Dtsch, S. Flynn, I. Gomes, C. Hollard, B. Hjort, P. Hoff-Olsen, K. Hrbikov, A. Lindenbergh, B. Ludes, O. Maroas, N. McCallum, D. Moore, N. Morling, H. Niedersttter, F. Noel, W. Parson, C. Popielarz, C. Rapone, A.D. Roeder, Y. Ruiz, E. Sauer, P.M. Schneider, T. Sijen, D.S. Court, B. Svieen, M. Turansk, A. Vidaki, L. Zatkalkov, J. Ballantyne, RNA/DNA co-analysis from human saliva and semen stains. Results of a third collaborative EDNAP exercise, Forensic Sci. Int. Genet. 7 (2013) 230–239.
[11] C. Haas, E. Hanson, M.J. Anjos, K.N. Ballantyne, R. Banemann, B. Bhoelai, E. Borges, M. Carvalho, C. Courts, G.D. Cock, K. Drobnic, M. Dtsch, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, J. Harteveld, B. Hjort, C. Hollard, P. Hoff-Olsen, C. Hls, C. Keyser, O. Maroas, N. McCallum, D. Moore, N. Morling, H. Niedersttter, F. Nol, W. Parson, C. Phillips, C. Popielarz, A.D. Roeder, L. Salvaderi, E. Sauer, P.M. Schneider, G. Shanthan, D.S. Court, M. Turansk, R.A. van Oorschot, M. Vennemann, A. Vidaki, L. Zatkalkov, J. Ballantyne, RNA/DNA co-analysis from human menstrual blood and vaginal secretion stains: results of a fourth and fifth collaborative EDNAP exercise, Forensic Sci. Int. Genet. 8 (2014) 203–212.
[12] C. Haas, E. Hanson, R. Banemann, A.M. Bento, A. Berti, Á Carracedo, C. Courts, G. De Cock, K. Drobnic, R. Fleming, C. Franchi, I. Gomes, G. Hadzic, S.A. Harbison, B. Hjort, C. Hollard, P. Hoff-Olsen, C. Keyser, A. Kondili, O. Maroas, N. McCallum, P. Miniati, N. Morling, H. Niedersttter, F. Nol, W. Parson, M.J. Porto, A.D. Roeder, E. Sauer, P.M. Schneider, G. Shanthan, T. Sijen, D. Syndercombe Court, M. Turansk, M. van den Berge, M. Vennemann, A. Vidaki, L. Zatkalkov, J. Ballantyne, RNA/DNA co-analysis from human skin and contact traces-results of a sixth collaborative EDNAP exercise, Forensic Sci. Int. Genet. 16 (2015) 139–147.
[13] E.K. Hanson, H. Lubenow, J. Ballantyne, Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs, Anal. Biochem. 387 (2009) 303–314.
[14] D. Zubakov, A.W.M. Boersma, Y. Choi, P.F. van Kuijk, E.A.C. Wiemer, M. Kayser, MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation, Int. J. Leg. Med. 124 (2010) 217–226.
[15] E. Sauer, A.-K. Reinke, C. Courts, Differentiation of five body fluids from forensic samples by expression analysis of four microRNAs using quantitative PCR, Forensic Sci. Int. Genet. 22 (2016) 89–99.
[16] K.M. Hunt, J.A. Foster, L.J. Forney, U.M.E. Schtte, D.L. Beck, Z. Abdo, L.K. Fox, J.E. Williams, M.K. McGuire, M.A. McGuire, Characterization of the diversity and temporal stability of bacterial communities in human milk, PLoS ONE 6 (2011).
[17] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J.R. Kultima, E. Prifti, T. Nielsen, A.S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J.Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H.B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Dor, S.D. Ehrlich, MetaHIT Consortium, P. Bork, J. Wang, MetaHIT Consortium, An integrated catalog of reference genes in the human gut microbiome, Nat. Biotechnol. 32 (2014) 834–841.
[18] L. Dethlefsen, D.A. Relman, Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation, Proc. Natl. Acad. Sci. U. S. A. 108 (Suppl 1) (2011) 4554–4561.
[19] HMP Consortium, Structure, function and diversity of the healthy human microbiome, Nature 486 (2012) 207–214.
[20] R.I. Fleming, S. Harbison, The use of bacteria for the identification of vaginal secretions, Forensic Sci. Int. Genet. 4 (2010) 311–315.
[21] C.C.G. Benschop, F.C.A. Quaak, M.E. Boon, T. Sijen, I. Kuiper, Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context? Int. J. Leg. Med. 126 (2012) 303–310.
[22] S. Lax, J.T. Hampton-Marcell, S.M. Gibbons, G.B. Colares, D. Smith, J.A. Eisen, J.A. Gilbert, Forensic analysis of the microbiome of phones and shoes, Microbiome 3 (2015) 21.
[23] N. Fierer, C.L. Lauber, N. Zhou, D. McDonald, E.K. Costello, R. Knight, Forensic identification using skin bacterial communities, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 6477–6481.
[24] S. Yuan, D.B. Cohen, J. Ravel, Z. Abdo, L.J. Forney, Evaluation of methods for the extraction and purification of DNA from the human microbiome, PLoS ONE 7 (2012).
[25] L. Abusleme, B.-Y. Hong, A.K. Dupuy, L.D. Strausbaugh, P.I. Diaz, Influence of DNA extraction on oral microbial profiles obtained via 16s rRNA gene sequencing, J. Oral Microbiol. 6 (2014).
[26] A. Wesolowska-Andersen, M.I. Bahl, V. Carvalho, K. Kristiansen, T. Sicheritz-Pontn, R. Gupta, T.R. Licht, Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis, Microbiome 2 (2014) 19.
[27] H.C. Veb, M.K. Karlsson, E. Avershina, L. Finnby, K. Rudi, Bead-beating artefacts in the Bacteroidetes to Firmicutes ratio of the human stool metagenome, J. Microbiol. Methods 129 (2016) 78–80.
[28] A. Vesty, K. Biswas, M.W. Taylor, K. Gear, R.G. Douglas, Evaluating the impact of DNA extraction method on the representation of human oral bacterial and fungal communities, PLOS ONE 12 (2017).
[29] G.C. Wang, Y. Wang, The frequency of chimeric molecules as a consequence of PCR co-amplification of 16s rRNA genes from different bacterial species, Microbiology (Reading, Engl.) 142 (Pt 5) (1996) 1107–1114.
[30] M.T. Suzuki, S.J. Giovannoni, Bias caused by template annealing in the amplification of mixtures of 16s rRNA genes by PCR, Appl. Environ. Microbiol. 62 (1996) 625–630.
[31] F. von Wintzingerode, U.B. Gbel, E. Stackebrandt, Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis, FEMS Microbiol. Rev. 21 (1997) 213–229.

[32] S.G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, M.F. Polz, PCR-induced sequence artifacts and bias: insights from comparison of two 16s rRNA clone libraries constructed from the same sample, Appl. Environ. Microbiol. 71 (2005) 8966–8969.
[33] J.B. Morrow, A.S. Downey, Challenges in Microbial Sampling in the Indoor Environment, Technical Note (NIST TN) – 1737, (2012) .
[34] S.J. Song, A. Amir, J.L. Metcalf, K.R. Amato, Z.Z. Xu, G. Humphrey, R. Knight, Preservation methods differ in fecal microbiome stability, affecting suitability for field studies, mSystems 1 (2016).
[35] B.E.R. Rubin, J.G. Sanders, J. Hampton-Marcell, S.M. Owens, J.A. Gilbert, C.S. Moreau, DNA extraction protocols cause differences in 16s rRNA amplicon sequencing efficiency but not in community profile composition or structure, Microbiologyopen 3 (2014) 910–921.
[36] E.A. Grice, H.H. Kong, G. Renaud, A.C. Young, G.G. Bouffard, R.W. Blakesley, T.G. Wolfsberg, M.L. Turner, J.A. Segre, A diversity profile of the human skin microbiota, Genome Res. 18 (2008) 1043–1050.
[37] S.M. Huse, V.B. Young, H.G. Morrison, D.A. Antonopoulos, J. Kwon, S. Dalal, R. Arrieta, N.A. Hubert, L. Shen, J.H. Vineis, J.C. Koval, M.L. Sogin, E.B. Chang, L.E. Raffals, Comparison of brush and biopsy sampling methods of the ileal pouch for assessment of mucosa-associated microbiota of human subjects, Microbiome 2 (2014) 5.
[38] R.A. van Oorschot, K.N. Ballantyne, R.J. Mitchell, Forensic trace DNA: a review, Investig. Genet. 1 (2010) 14.
[39] T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mah, VSEARCH: a versatile open source tool for metagenomics, PeerJ 4 (2016).
[40] R.C. Edgar, H. Flyvbjerg, Error filtering, pair assembly and error correction for next-generation sequencing reads, Bioinformatics 31 (2015) 3476–3482.
[41] B.J. Haas, D. Gevers, A.M. Earl, M. Feldgarden, D.V. Ward, G. Giannoukos, D. Ciulla, D. Tabbaa, S.K. Highlander, E. Sodergren, B. Meth, T.Z. DeSantis, Human Microbiome Consortium, J.F. Petrosino, R. Knight, B.W. Birren, Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons, Genome Res. 21 (2011) 494–504.
[42] K.H. Liland, H. Vinje, L. Snipen, microclass: an R-package for 16s taxonomy classification, BMC Bioinform. 18 (2017).
[43] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.
[44] J. Oksanen, F.G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, E. Szoecs, H. Wagner, vegan: Community Ecology Package, (2017) .
[45] J. Kuczynski, J. Stombaugh, W.A. Walters, A. Gonzlez, J.G. Caporaso, R. Knight, Using QIIME to analyze 16s rRNA gene sequences from microbial communities, Curr. Protoc. Bioinform. 10 (2011) Chapter, Unit 10.7.
[46] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer Science & Business Media, 2009.
[47] R.J. Brownlow, K.E. Dagnall, C.E. Ames, A comparison of DNA collection and retrieval from two swab types (cotton and nylon flocked swab) when processed using three QIAGEN extraction methods, J. Forensic Sci. 57 (2012) 713–717.
[48] T.J. Verdon, R.J. Mitchell, R.A.H. van Oorschot, Evaluation of tapelifting as a collection method for touch DNA, Forensic Sci. Int. Genet. 8 (2014) 179–186.
[49] T. Unno, Bioinformatic suggestions on MiSeq-based microbial community analysis, J. Microbiol. Biotechnol. 25 (2015) 765–770.
[50] J. Lloyd-Price, G. Abu-Ali, C. Huttenhower, The healthy human microbiome, Genome Med. 8 (2016).
[51] C. De Filippo, D. Cavalieri, M. Di Paola, M. Ramazzotti, J.B. Poullet, S. Massart, S. Collini, G. Pieraccini, P. Lionetti, Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 14691–14696.
[52] G.I. Perez Perez, Z. Gao, R. Jourdain, J. Ramirez, F. Gany, C. Clavaud, J. Demaude, L. Breton, M.J. Blaser, Body site is a more determinant factor than human population diversity in the healthy skin microbiome, PLOS One 11 (2016).
[53] J. Oh, A.L. Byrd, C. Deming, S. Conlan, NISC Comparative Sequencing Program, H.H. Kong, J.A. Segre, Biogeography and individuality shape function in the human skin metagenome, Nature 514 (2014) 59–64.

Paper III

Optimizing body fluid recognition from microbial taxonomic profiles

Eirik Nataas Hanssen[1,2], Kristian Hovde Liland[3,4], Peter Gill[1,2], Lars Snipen[*4]

[1]Department of Forensic Biology, Oslo University Hospital , P.O. Box 4950 Nydalen, N-0424 Oslo, Norway
[2]Department of Forensic Medicine, University of Oslo, P.O. Box 4950 Nydalen, N-0424 Oslo, Norway
[3]Faculty of Science and Technology, Norwegian University of Life Sciences, P.O.Box 5003, N-1432 Ås, Norway
[4]Faculty of Chemistry, Biotechnology and Food Sciences, Norwegian University of Life Sciences, P.O.Box 5003, N-1432 Ås, Norway

Email: Eirik Natås Hanssen - kjeirik@gmail.com; Kristian Hovde Liland - kristian.liland@nmbu.no; Peter Gill - peterd.gill@gmail.com; Lars Snipen*- lars.snipen@nmbu.no;

*Corresponding author

Abstract

Background: In forensics the DNA-profile is used to identify the person who left a biological trace, but information on body fluid can also be essential in the evidence evaluation process. Microbial composition data could potentially be used for body fluid recognition as an improved alternative to the currently used presumptive tests. We have developed a customized workflow for interpretation of bacterial 16S sequence data based on a model composed of Partial Least Squares (PLS) in combination with Linear Discriminant Analysis (LDA). Large data sets from the Human Microbiome Project (HMP) and the American Gut Project (AGP) were used to test different settings in order to optimize performance.

Results: From the initial cross-validation of body fluid recognition within the HMP data the optimal overall accuracy was close to 98%. Sensitivity values for the fecal, oral and vaginal samples were all ≥ 0.99, and for the skin and nasal samples 0.97 and 0.84 respectively. Specificity values were high for all 5 categories, mostly > 0.99. This optimal performance was achieved by using the following settings: Taxonomic profiles based on operational taxonomic units (OTUs) with 0.98 identity (OTU98), Aitchisons simplex transform with $C = 1$ pseudo-count and no regularization ($r = 1$) in the PLS step. Variable selection did not improve the performance further. To test for robustness across sequencing platforms, we also trained the classifier on HMP data and tested on the AGP data set. In this case, the standard OTU based approach showed a severe decline in accuracy. However, by using taxonomic profiles made by direct assignment of reads to a genus, we were able to nearly maintain the high accuracy levels. The optimal combination of settings was still used, except the taxonomic level been genus instead of OTU98.

Conclusions: We present an optimized workflow for recognizing body fluids based on 16S sequence data. The method was customized for pattern recognition, and shows high accuracy, comparable to the alternative mRNA based methods. In addition, the method was proven to be robust across data sets from different studies, which is a condition for reasonable standardization needs and a wide-spread use. A major finding was that to achieve this robustness, the taxonomic profiles should be based on a supervised (classify reads directly to pre-determined bins) rather than an unsupervised (clustering into OTUs) method. The performance may be improved even further by using higher resolution taxonomic bins. The methods resulting from this study makes up the core of an R-package for the recognition of human body fluids from 16S sequence data.

Keywords: Forensics, massive parallel sequencing, microbiome, PLS, discriminants.

Background

DNA analysis is used to identify a person from a biological trace found at a crime scene. However, a trace might not be crime related and could be a result of 'innocent' activity or contamination [1–5]. In order to establish a picture of how the trace was deposited, it is helpful to determine from which part of the body the trace originates. Traditionally, presumptive tests have been used for body fluid recognition, but immunochromatographic lateral flow strip tests are also commonly used in forensic routine work [6, 7]. In addition some laboratories have also implemented gene expression based methods using mRNA and miRNA markers [8, 9]. Although not yet ready for casework, a novel microbiota-based recognition method is a promising alternative [10].

Microbiota sequencing was made possible with the introduction of massively parallel sequencing (MPS). In the last few years this field has had exceptional interest, and both human [11, 12] and environmental microbiota [13] have been thoroughly explored. For the majority of these studies, hypervariable regions of the 16S ribosomal gene have been used for taxonomic identification. The microbiota refers to the taxonomic composition of a microbial community, and is typically what we identify by 16S amplicon sequencing. Large amounts of raw data (reads) from such samples are now publicly available, e.g. in the NCBI/SRA database (https://www.ncbi.nlm.nih.gov/sra).

Tools for handling, preparation and analysis of 16S amplicon data are numerous, e.g. [14–17]. We showed earlier that body fluids can be recognized using standard methods for raw data processing, using Principal Component Analysis (PCA) in combination with Linear Discriminant Analysis (LDA) for pattern recognition [10]. However, the potential of this approach can only be revealed by a systematic evaluation over larger data sets exploring various settings in the data analysis.

Microbiota sequencing has primarily been conducted to explore new microbial communities. The data processing pipelines for discovering and exploring the ecology are not necessarily optimal when it comes to recognizing already characterized communities. To recognize a body fluid from microbiota data is a classical pattern recognition problem. Pattern recognition is a branch of machine learning where a model utilizes regularities in a training data set to classify samples in a new test data set. The training and test data sets need to have the same format, with the same predictor variables. In the case of using microbiota data for body fluid recognition, the predictor variables are the taxonomic bins that the reads are assigned to. The taxonomic resolution determines the number of predictor variables. The standard pipelines will in

general cluster reads into Operational Taxonomic Units (OTUs) using an identity threshold of 97% [18]. Since we cannot hope to find taxa which are easily detected and unique to any body fluid [19, 20], we must rely on fairly stable patterns of high or low abundances of several taxa. The optimal taxonomic resolution must be fine-grained enough to produce abundance patterns that permit discrimination between body fluids, but still coarse enough to yield reproducible results over many samples. In addition to the taxonomic resolution, there are several choices for data transformation that will affect the precision of a method for microbiota-based body fluid recognition [21, 22].

We have performed a systematic study on data processing and pattern recognition approaches, and quantified their effects on microbiota-based body fluid recognition. The findings will be implemented in an R-package for microbial forensics that we are developing.

Methods

Data

Public data from the Human Microbiome Project (HMP) [23] were downloaded from http://hmpdacc.org/. The HMP sequenced 16S amplicons from various body sites of hundreds of people, and we assembled these into four body fluids (oral, nasal, vaginal, fecal). The data set also include samples from human skin. Such samples may also be relevant in a forensic setting, and are included as a fifth category in addition to the four body fluids. In table 1 we show how our categories include the original body sites annotated by HMP. The 16S gene has 9 hypervariable regions designated V1-V9, and the HMP amplicons are from two distinct regions, V1-V3 and V3-V5. Reads for 5035 samples were downloaded, but only 4936 samples had above 100 reads and were used in this analysis. These public data have been subject to a careful preprocessing (de-multiplexing, removing contaminants, etc), see protocols at http://hmpdacc.org/ for all details. Each sample was downloaded as a FASTA file of reads, using the SRA toolkit (https://www.ncbi.nlm.nih.gov/sra). Table 1 shows a summary of the data.

To test the body fluid recognition performance with these data, we used a 10-fold cross-validation. The data were split into 10 non-overlapping subsets or segments, where each segment in turn was used as a test-set and the remaining data as a training-set. The samples were first sorted by body fluid and then by person within each body fluid. Next, they were given a segment-number from 1 to 10 repeatedly throughout the data set to achieve maximum spread of body fluids across segments. However, within each body fluid all samples from the same person were given identical segment-number, to ensure data from the same person and body fluid was found within one segment.

As an external test-set we also downloaded public data from the American Gut Project (AGP) [24], again using the SRA toolkit. These data differ from HMP in several ways. Different sequencing technologies (Illumina, while HMP uses Roche 454) were used to obtain short reads (120 bp) from 9500 samples from almost as many persons. The vast majority of these samples are from feces, as suggested by the project title, but the data set also includes some samples from all the other categories in our study. Another difference is the extraction protocols used by the two projects (AGP uses the Mobio MagAttract PowerSoil kit while HMP uses the Mobio DNeasy PowerSoil Kit). The AGP data are from region V4 of the 16S gene.

Taxon read-counts

The first question we addressed was how to group the reads in a sample into a set of taxa. We focused on two distinct approaches, one of clustering into OTUs and one of direct taxonomic classification of all reads. We also explored different taxonomic resolution for both approaches.

A standardized pipeline for OTU-finding was set up using the VSEARCH software [17]. For a training data set, the reads were filtered for chimera, clustered into OTUs, and singleton clusters discarded. We used three different clustering identity-thresholds, $0.97, 0.98$ and 0.99, where the first is the standard. Next, all reads in each training set sample were assigned to OTUs by searching against the OTU centroids, using same similarity threshold for clustering. For each test set sample the steps were repeated, except for the clustering, where the OTUs found from the training set were used.

We also tested an alternative approach where the OTU-clustering step was omitted by using the `taxMachine` function of the `microclass` package [25] which is available for installation in the R computing environment [26]. The `taxMachine` is a pre-trained 16S classifier which uses transformed K-mer counts (8-mers) in a Naïve Bayes classifier to assign 16S reads into one of 1774 genera. It has been trained on 38 781 complete 16S sequences of high quality, producing a quick and accurate classifier with additional statistics indicating if classifications are outliers (typically new genera or bad quality reads) or close to the border between two known genera (low confidence in classified genus). See [25] for details. We explored two thresholds, where the recognition-probability was set to $10e-4$ (weak filtering) and $10e-10$ (strict filtering). The `taxMachine` only classify reads to a genus, but we also considered lower resolution taxonomic profiles, by assembling genus-classifications into family, order, class and phylum. Thus, the reads were divided into fewer taxa, with larger read-counts in each taxon.

From reads to profiles

When we refer to a *predictor* in the text below, it could mean an OTU, a genus or any other taxonomic bin that we group the reads into by one of the methods mentioned above.

The read-count for predictor j in sample i is denoted $c_{i,j}$ and all read-counts from sample i were arranged in a vector $\boldsymbol{c}_i = (c_{i,1}, ..., c_{i,p})$ where the predictors are sorted alphabetically. The total read-count for a sample has no information about body fluid, and we consider only the relative read-counts

$$r_{i,j} = \frac{c_{i,j}}{m_i} \tag{1}$$

where $m_i = \sum_{j=1}^{p} c_{i,j}$, i.e. the total read-count from sample i. The vector $\boldsymbol{r}_i = (r_{i,1}, ..., r_{i,p})$ is what we denote the *raw taxonomic profile* for sample i. The data type of a raw taxonomic profile is known as compositional data, i.e. the elements are relative abundances that always sum to 1.0 [22].

A number of pattern recognition methods may benefit from a transformation of the data prior to their use for training a classifier. A commonly used transformation for compositional data is known as the Aitchisons simplex transform [27]:

$$x_{i,j} = \log_2 \left(\frac{r_{i,j}}{(\prod_{j=1}^{p} r_{i,j})^{1/p}} \right) = \log_2(r_{i,j}) - \frac{1}{p} \sum_{j=1}^{p} \log_2(r_{i,j}) \tag{2}$$

i.e. the logarithm of the read-counts divided by their geometric mean in the sample. Such transformed values will always sum to 0 in each sample.

This transform requires only nonzero read-counts. For body fluid recognition some taxa are prevalent in some body fluids, but absent in other, and such taxa are among the most valuable discriminating variables in the entire data set. To discard them is not an option. A simple way around this is to add a given value as a pseudo-count to all read-counts. Adding C pseudo-counts to all taxa means we get

$$r'_{i,j} = \frac{c_{i,j} + C}{m_i + pC} \tag{3}$$

as the smoothed read-count for genus j in sample i. Next, $r_{i,j}$ is replaced by $r'_{i,j}$ in (2). The vector $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})$ is the *transformed taxonomic profile* for sample i. The adding of various numbers of pseudo-counts may have an impact on the performance of the pattern recognition algorithm, and in this analysis we tried out pseudo-counts C over several magnitudes to investigate this effect (C between 0.001 and 100).

Pattern recognition

The pattern recognition method we have used in this study is a combination of two well-known and much used supervised learning methods, Partial Least Squares (PLS) [28] and LDA [29]. PLS can in this context be viewed as a dimension reduction method which finds the linear combinations of predictors that best explain the difference between pairs of body fluids. A single parameter is tuned, the number of linear combinations, balancing the dimension reduction between too little explained variation and too much focus on local details. LDA is a linear classifier which assigns objects into groups based on their Mahalanobis distance to group centres. It estimates a common covariance for the included groups, making it robust, and uses Bayes rule for calculation of posterior probabilities of group affiliations.

For the first step of the pattern recognition method taxonomic profiles for all samples in the training-set were assembled, as rows, into a matrix $\boldsymbol{X}$. The dimensions of this matrix is $(n \times p)$ where n is the number of samples and p is the number of predictors. For each row of $\boldsymbol{X}$ there is a corresponding body fluid label in the $(n \times 1)$ vector $\boldsymbol{y}$.

We have in this case $N = 5$ categories (body fluids). We could have opted for one universal model recognising all five categories, but for reasons discussed later, we have taken a different approach. We split the problem into $N(N-1)/2 = 10$ different two-category problems. This means ten separate submodels, each discriminating between two body fluids. In each submodel the two involved body fluids are dummy coded as 0 and 1, i.e. the label vector $\boldsymbol{y}$ has a corresponding numeric vector $\boldsymbol{y}_d$ of zeros and ones.

The training of a submodel starts by fitting a PLS-model to the training-set $(\boldsymbol{y}_d, \boldsymbol{X})$. The reason for the PLS-step is that $\boldsymbol{X}$ has many columns (p predictors), with collinearity, making $\boldsymbol{X}'\boldsymbol{X}$ (close to) singular. This makes any subsequent LDA model fitting impossible. The PLS-step simplifies the problem. This is achieved by replacing the original $(n \times p)$ data matrix $\boldsymbol{X}$ by the $(n \times q)$ scores matrix $\boldsymbol{Z}$ from the PLS-model, where $q << p$. The latter matrix has orthogonal columns, and the number of dimensions q to include is found using the McNemar-test procedure described in [30], seeking the smallest dimension giving not significantly poorer prediction than the best possible. The stringency of this test can be adjusted, producing various degrees of dimension-reduction, and in this study we tried out several stringencies (stringency between 0.1 and 1). We refer to this as *regularization* below.

Next, the reduced subset $(\boldsymbol{y}, \boldsymbol{Z})$ is used to train the LDA classifier. This means fitting a multivariate gaussian density to the scores $\boldsymbol{Z}$ of each category, assuming equal variances. We used flat priors in all cases, i.e. equal prior probability of both body fluids in all models.

When a new sample taxonomic profile $\boldsymbol{x}$ is considered, it is classified by all the ten submodels and in each

case assigned by LDA a posterior probability for the two categories of the submodel. All categories are involved in four submodels, and by averaging the posterior probabilities over these four outcomes for each category, the new sample is assigned to the category with the largest average posterior probability.

Variable selection

It is more than likely that many of the predictors are not informative for discriminating between two specific body fluids. A variable selection procedure was implemented to see if a reduced set of taxa could improve the results. This variable selection was done independently for each of the ten submodels.
We used the backward elimination algorithm described in [31], with some minor modifications. This means we start out using all predictors, and then gradually discard the least informative in each iteration until only one is left. The cross-validated classification error is monitored for each iteration, and the selected subset of predictors is the smallest subset producing not significantly poorer results than the optimum along the elimination path.
The importance of each predictor is decided in the PLS-step of our pattern recognition method, using the VIP-criterion [31]. This criterion is defined only for two-category problems, and this is an important argument for splitting the entire problem into many two-category problems instead of one multi-category problem.

Results and Discussion

Optimization within HMP data

Based on cross-validation inside the HMP data set, we explored how various factors influence the accuracy of classification. All combinations of factors were tested in a 10-fold cross-validation within the HMP data set to compute body fluid recognition accuracy. Thus, for each factor combination 10 accuracy values were computed. We tested the effect of each factor by using these accuracy-values as the dependent variable in an Analysis of Variance (ANOVA). The combination giving the best (largest) accuracy was found, and all other combinations were tested against this to see if they produced significantly poorer results. All factors and their levels are given in Table 2.
First, we made a preliminary comparison within the coarse-level taxonomic profiles. In this case all reads from each sample were classified into known taxonomic bins at levels phylum, class, order, family or genus, and in this step a recognition-probability threshold was used to decide how well a read must be recognized to be assigned to a specific bin. We tried out two thresholds, 10^{-4} (weak) and 10^{-10} (strict). This analysis

produced two distinct results: First, the genus- and family-level taxonomic profiles produced much better results than the coarser resolutions. Second, using the weak threshold produced slightly better accuracies compared to the strict regime ($p = 0.1$). Thus, from this preliminary study we discarded the phylum, class and order taxonomic levels from further analysis, and used only the weak threshold (10^{-4}) for the family and genus read classification.

The full comparison of all remaining factors is displayed in Table 2. This revealed that the best combination produced an overall accuracy close to 98%, i.e. in 98 out of 100 samples we correctly recognized the body fluid. This indicates that body fluids can be recognized from microbiota data with a very high precision, and that method performance is comparable to that of mRNA based methods [32–34]. In Figure 1 we display the details in the recognition of the different body fluids under optimal settings for both region V1-V3 and V3-V5 datasets. The vaginal, oral and fecal body fluids were extremely well recognized, having sensitivities ≥ 0.99. This is promising, since we regard these highly relevant in a forensic perspective. In addition, a reliable forensic test for feces is to our knowledge non-existing. The errors were predominately made for the nasal and skin samples. Nasal samples had the lowest sensitivity at 0.84 and were typically miss-classified as skin (in 14.8% of the cases). As the samples are from the nostrils, the most obvious explanation for this is a contamination from the nearby skin microbiota. The skin samples were also sometimes difficult to recognize (sensitivity at 0.97), and were confused with all other types of body fluids. It is known that skin samples have a large variation in bacterial composition. Not only does the composition differ between body sites, but also between individuals for the same body site [12, 35, 36]. Recognizing skin samples based on microbiota is bound to be difficult, and it is not surprising that such samples are confused with all the different body fluids. For this study our main focus was to recognize body fluids, and skin was included as body fluid samples are often collected from skin. We have previously shown that skin microbiota has a relatively low DNA content compared to saliva when sampling with tape or synthetic swabs from saliva deposited on skin [10]. It is reasonable to believe that this quantitative ratio also applies for feces and vaginal secretion [37]. If empirical studies should verify this assumption, the impact of skin microbiota on body fluid samples collected from skin should be less problematic. Anyhow, a final tool applicable in casework should to be able to deconvolute mixtures as many trace samples will contain more than one body fluid. Such a tool should also improve accuracy for the nasal samples by separating between skin and nasal microbiota.

Sensitivity and specificity are much used parameters to evaluate method performance (see table 3). The sensitivity is the proportion of positives that are correctly identified, mentioned above. The specificity is

the proportion of negatives that are correctly identified. Specificity is the most important parameter in a forensic context as a false positive result can lead to a wrong conviction. Hence, the specificity value needs to be high. In our case the specificity values were > 0.99 for the vaginal, oral and nasal samples and > 0.98 for the skin and fecal samples, and we regard this as promising.

The optimal accuracy was achieved using the combination: Taxonomic profiles based on OTUs with 0.98 identity (OTU98), Aitchisons simplex transform with $C = 1$ pseudo-count and no regularization ($r = 1$) in the PLS step. This is the Reference-combination in Table 2. First, we notice that it was optimal with OTUs of a finer resolution than the standard 0.97 identity. However, from Table 2 the drop in accuracy to OTU97 or OTU99 was insignificant. The genus level model showed a significantly poorer accuracy of 0.01 compared to the reference. This means that there was some variation in very specific species, or even strains, typical for some body fluids, and that this was lost when assigning reads to the genus level. Next, we notice that the Aitchison simplex transform seems to be beneficial for the pattern-recognition methods. It was optimal using this transform with $C = 1$ pseudo-count, but using $C = 0.1$ or $C = 100$ gave nearly the same result. The important difference was to $C = 0$, i.e. no transformation at all, which produced poorer results. The regularization in the PLS-step had little impact, and using no regularization was the best choice. The HMP reads are sequenced from two distinct regions of the 16S gene, and we observe that data from the V3-V5 region produced slightly better accuracies than those from V1-V3. However, since there are approximately twice as many samples from the V3-V5 region, this might explain the difference. Larger training data sets in general means better classification accuracies. Previous studies indicate that the discriminative power between various 16S regions is very small [38].

Unless otherwise stated, we used the optimal combination of factors (reference-combination) in the analysis below.

Variable selection

A variable selection procedure was included to evaluate if selecting only a smallish number of OTUs would improve the classification results from the optimal results achieved above. Again, we used the HMP data and 10-fold cross-validation. Variable selection was performed separately for each pair-model, see Methods for details. Also, in the cross-validation there were 10 (slightly) different training data sets, hence 10 (slightly) different sets of OTUs were found. Thus, variable selection had to be performed separately for each segment, producing 10 (slightly) different selections. Variable selection was performed using the optimal factor combination from above, but we used data for both regions V1-V3 and V3-V5 separately.

The overall result was that variable selection had little impact on the total classification accuracy. For V1-V3 region the accuracy improved from 0.97 to 0.98 after variable selection, and for V3-V5 region the difference was in reality zero. The mean number of selected OTUs for each model (body fluid pair) is given in supplementary table S1. The number of selected OTUs ranged from 9 ("nasal vs fecal") to several hundred (e.g. "nasal vs skin"). To investigate the selected OTUs further, their centroid sequences were classified to genera using `taxMachine`. Each body fluid is involved in 4 pair-models, and the selected genera for all these were summed to give the genera most frequently associated with a body fluid. An overview is given in figure 2. Note that the association of a genus with a body fluid does not imply that this genus is (highly) present in that body fluid, it could just as well be that its absence is important for body fluid recognition. In figure 2 the text size used for a genus is proportional to the expected relative abundance in the actual body fluid [39]. Overall, the most abundant genera of a body fluid seemed to be important for selection as these were all found among the top ranking associations. In addition, many were important for multiple body fluids. This was expected as many of the selected genera have a uniquely high abundance in only one or a few of the body fluids. Consequently, they will also be important as low level genera when identifying the other body fluids.

Even if classification was not improved by variable selection, a reduced taxa model may still be of some interest. Currently, PCR multiplexing of a smallish number of taxa would be a fast and cheap way to obtain the required microbiota data for body fluid recognition. In addition PCR multiplexing is well-known in all forensic labs. However, we should remember that any selected variable in a multivariate problem means that we select some out of many other highly correlated variables, and unless we have huge training data sets we may end up with unstable results if we base them on too few taxa. Given the fast developments of sequencing technologies and decreasing sequencing costs, we would not invest much effort into developing a strong variable selection for the purpose of PCR multiplexing.

Predicting AGP samples

The results achieved by cross-validation within the HMP data can be seen as a best-case scenario, where both test- and training-set raw data have been obtained by the same protocols and sequencing technology. Since the HMP and AGP are two independent studies, using models trained on HMP data to recognize body fluids from the AGP data, is a more realistic scenario with respect to actual casework [40]. From the HMP data, models were trained on both region V1-V3 and V3-V5, using the optimal settings described above for all other parameters (no variable selection). Next, all AGP samples were classified, resulting in

accuracies of 0.73 and 0.76 for regions V1-V3 and V3-V5, respectively. For region V1-V3 the accuracy was expected to be low, since the AGP reads were from the V4 region. Assigning these reads to the OTUs found in the V1-V3 region must result in many errors. However, the model trained on the HMP data from region V3-V5 also showed considerably poorer accuracy than in the previous cross-validation (0.76 compared to 0.98 previously). Figure 3 shows detailed results for all body fluids. The numbers of nasal and vaginal AGP samples were very low, and their respective accuracies are unreliable. The specificity values are comparable with the specificity values seen from the HMP cross-validation (the lowest specificity is still > 0.98). See table 3 for details.

The severe loss in classification accuracy compared to the cross-validation within the HMP data illustrates that there are effects of protocols and sequencing technologies between HMP and AGP that influence how the taxonomic profiles will look like. A very high-resolution taxonomic profile, like OTUs with 0.98 identity, will tend to overfit the model to the data in the training set, and is correspondingly easy to mislead once we have slightly different reads. For this reason we again tested the coarser approach, where reads are assigned directly to a genus instead of finding new OTUs from every training set. Reads from both HMP and AGP data sets were assigned directly to a genus using the `taxMachine` described above. Models were trained on the genus-profiles from the HMP data, and again we used separate models for the V1-V3 and V3-V5 regions. The previous optimal settings were used, except that models were now based on genus rather than OTU98. Samples from the AGP data set were classified according to their genus-profiles. The resulting overall accuracies were now 0.96 for both the V1-V3 and V3-V5 trained models.

This, most astonishing result, shows that the accuracy for predicting AGP-samples is almost as high as for cross-validation within the HMP samples. Assigning reads to pre-defined genera results in a much higher reproducibility across experiments, and the profiles from HMP-data and AGP-data become very similar for the same body fluids. OTU-finding procedures have in general been designed to find all kinds of potentially interesting taxa when probing new microbial communities. We must expect several of these OTUs to be artefacts or at least not very robust to a change in sequencing protocols and technologies [41, 42]. Assigning reads directly to pre-defined genera is much more robust in this perspective. The resolution is poorer, but also more stable, since the same number of reads is assigned to much fewer taxa, giving larger counts for each. In the HMP data sets, we find that reads are assigned to 1640 different genera, while OTU97, OTU98 and OTU99 produce a resolution of ~ 14000, ~ 25000 and ~ 65000 taxa, respectively. Another point is that the direct binning of reads with `taxMachine` is magnitudes faster than any OTU-finding pipeline.

We also noticed that when we trained a model on the V1-V3 data, and predicted AGP-samples taken from

the V4-region, the results were just as good as if we trained on the V3-V5 region (accuracy 0.96). Direct binning of reads does not require the reads to be from any specific region of the 16S gene. This is obviously a huge advantage. As an illustration, we merged HMP-samples from both the V1-V3 and the V3-V5 regions into one large training data set. We trained a model on this, and obtained an overall accuracy for the AGP samples of 0.96.

From the cross-validation within HMP data, we saw that higher resolution profiles have a potential to improve body fluid recognition compared to genus-profiles. For this reason, we propose that in the future we should investigate the use of direct binning of reads, but into pre-defined taxa based on a finer resolution than genus. Currently, taxonomic classifiers like `taxMachine` or the RDP-classifier [43], will only make use of generic taxa and stop at the genus level. It is possible to re-train such tools on taxonomic bins purpose made for recognizing body fluids, and this could show to be the best choice.

We have earlier proposed that the expected bias between forensic labs could be solved by standardization [10]. However, using the direct binning approach will probably reduce the standardization effort. This might also pave the way for using universal training data sets that can be shared between laboratories.

For all experiments where AGP data were used for testing, we have confirmed that the used optimal settings still holds by repeating the set-up from the initial optimization experiment.

Conclusions

We present a customized workflow for recognizing body fluids from 16S sequence data by using a model composed of PLS in combination with LDA. For method development and evaluation we used large data sets from the HMP and AGP consortiums which each were categorized into fecal, nasal, oral, skin and vaginal samples.

We used the standard approach to build taxonomic profiles where reads were assigned to OTUs. Method optimization was performed by testing combinations of different calculation settings in a cross-validation setup using HMP data. Method performance was generally high for the majority of combinations with only a few leading to substantial decrease in performance. Nevertheless we were able to identify optimal combination of settings as: Taxonomic profiles based on operational taxonomic units (OTUs) with 0.98 identity (OTU98), Aitchisons simplex transform with C = 1 pseudo-count and no regularization (r = 1) in the PLS step. By using these settings, the fecal, oral and vaginal samples had sensitivities ≥ 0.99 and specificities > 0.99. This is promising as we regard these body fluids as highly relevant in a forensic setting.

Variable selection did not improve method performance significantly.

When mimicking a real case scenario by training on HMP data and testing on AGP data, the performance using the OTU98 based model declined severely. However, this was overcome by using the alternative approach to build the taxonomic profiles. Instead of searching for OTUs in each training set, reads were assigned directly to pre-defined genera with the `taxmachine` tool. By using this approach we obtain a much more stable performance. The overall accuracy was now ~ 0.96 and specificities ≥ 0.98 which is quite similar to what is achieved using the alternative mRNA based methods. Performance was best for the fecal and oral samples. To our knowledge, a forensic method for fecal recognition is non-existing. We confirmed that when replacing OTU98 by genus, all other optimal settings from the initial cross-validation still held when testing on the AGP data. Re-training a tool like `taxMachine` on a purpose-designed set of sequences for human body fluid recognition is probably the ultimate solution in this direction.

In this study we have demonstrated the power of microbiota based body fluid recognition for forensic use. We have made no attempts to compare different pattern-recognition methods, and our choice of PLS in combination with LDA could very well be improved upon, however, not by much given the good accuracies already obtained. The method is, however, still not ready for casework as this demands inter-laboratory validation studies. Biological trace samples are often mixtures of several body fluids, and any casework-ready method should be able to deconvolute mixtures, or at least indicate which body fluids are present. Also, in forensics some kind of statement of reliability is needed. Currently, it is far from obvious how such statements should be computed and presented from the outcome of a pattern-recognition method. However, the methods and results presented here forms a core of an R-package we are developing for use in a forensic setting.

List of abbreviations

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data used in this study are publicly available through the Human Microbiome Project and American Gut Project. If requested, the authors can assist in accessing these data.

Competing interests

The authors declare that they have no competing interests.

Funding

This project has been financed by Oslo University Hospital and Norwegian University of Life Sciences.

Authors' contributions

All authors have contributed significantly to all programming, documentation and preparation of this manuscript.

Acknowledgements

Not applicable.

References

1. Gill P: *Misleading DNA Evidence: Reasons for Miscarriages of Justice*. Elsevier 2014.
2. van Oorschot RAH, Glavich G, Mitchell RJ: **Persistence of DNA deposited by the original user on objects after subsequent use by a second person**. *Forensic Sci Int Genet.* 2014, **8**:219–225.
3. Goray M, Oorschot RAHv: **The complexities of DNA transfer during a social setting**. *Legal Medicine.* 2015, **17**(2):82–91.
4. van den Berge M, Ozcanhan G, Zijlstra S, Lindenbergh A, Sijen T: **Prevalence of human cell material: DNA and RNA profiling of public and private objects and after activity scenarios**. *Forensic Sci Int: Genetics.* 2016, **21**:81–89.
5. Fonneløp AE, Ramse M, Egeland T, Gill P: **The implications of shedder status and background DNA on direct and secondary transfer in an attack scenario**. *Forensic Sci Int Genet.* 2017, **29**:48–60.
6. Hochmeister MN, Budowle B, Rudin O, Gehrig C, Borer U, Thali M, Dirnhofer R: **Evaluation of prostate-specific antigen (PSA) membrane test assays for the forensic identification of seminal fluid**. *J Forensic Sci.* 1999, **44**(5):1057–1060.
7. Pang BCM, Cheung BKK: **Identification of human semenogelin in membrane strip test as an alternative method for the detection of semen**. *Forensic Sci Int.* 2007, **169**:27–31.
8. Sijen T: **Molecular approaches for forensic cell type identification: On mRNA, miRNA, DNA methylation and microbial markers**. *Forensic Sci Int Genet.* 2015, **18**:21–32.
9. Harbison S, Fleming R: **Forensic body fluid identification: state of the art**. *Research and Reports in Forensic Medical Science.* 2016, **6**:11–23.
10. Hanssen EN, Avershina E, Rudi K, Gill P, Snipen L: **Body fluid prediction from microbial patterns for forensic application**. *Forensic Sci Int Genet.* 2017, **30**:10–17.

11. Human Microbiome Project Consortium: **A framework for human microbiome research**. *Nature.* 2012, **486**(7402):215–221.
12. Lloyd-Price J, Abu-Ali G, Huttenhower C: **The healthy human microbiome**. *Genome Med.* 2016, **8**.
13. Gilbert JA, Jansson JK, Knight R: **The Earth Microbiome project: successes and aspirations**. *BMC Biol.* 2014, **12**:69.
14. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities**. *Appl Environ Microbiol.* 2009, **75**(23):7537–7541.
15. Caporaso J, Kuczynski J, Stombaugh J, Bittinger K, Bushman F, Costello E, Fiere N, Pena A, Goodrich J, Gordon J, Huttley S GA andKelley, Knights D, Koenig J, Lozupone C, McDonald D, Muegge B, Pirrung M, Reeder J, Sevinsky J, Turnbaugh P, Walters W, Widmann J, Yatsunenko T, Zaneveld J, Knigh R: **QIIME allows analysis of high-throughput community sequencing data**. *Nat Methods.* 2010.
16. Edgar RC: **Search and clustering orders of magnitude faster than BLAST**. *Bioinformatics (Oxford, England).* 2010, **26**(19):2460–2461.
17. Rognes T, Flouri T, Nichols B, Quince C, Mahé F: **VSEARCH: a versatile open source tool for metagenomics**. *PeerJ.* 2016, **4**:e2584.
18. Edgar RC: **UPARSE: highly accurate OTU sequences from microbial amplicon reads**. *Nat Methods.* 2013, **10**(10):996–998.
19. Fleming RI, Harbison S: **The use of bacteria for the identification of vaginal secretions**. *Forensic Sci Int: Genetics.* 2010, **4**(5):311–315.
20. Benschop CCG, Quaak FCA, Boon ME, Sijen T, Kuiper I: **Vaginal microbial flora analysis by next generation sequencing and microarrays; can microbes indicate vaginal origin in a forensic context?** *Int J Legal Med.* 2012, **126**(2):303–310.
21. McMurdie PJ, Holmes S: **Waste not, want not: why rarefying microbiome data is inadmissible**. *PLoS Comput Biol.* 2014, **10**(4):e1003531.
22. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ: **It's all relative: analyzing microbiome data as compositions**. *Ann Epidemiol.* 2016, **26**(5):322–329.
23. NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M: **The NIH Human Microbiome Project**. *Genome Res.* 2009, **19**(12):2317–2323.
24. McDonald D, Birmingham A, Knight R: **Context and the human microbiome**. *Microbiome.* 2015, **3**:52.
25. Liland KH, Vinje H, Snipen L: **microclass: an R-package for 16S taxonomy classification**. *BMC Bioinformatics.* 2017, **18**.
26. R Development Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria 2008, [[http://www.R-project.org]].
27. Aitchison J: *The statistical analysis of compositional data.* Chapman and Hall 1986.
28. Wold H: *Estimation of Principal Components and Related Models by Iterative Least squares.* New York: Academic Press 1966.
29. Fisher RA: **The Use of Multiple Measurements in Taxonomic Problems**. *Ann Eugen.* 1936, **7**(7):179–188.
30. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L: **Mining for genotype-phenotype relations in Saccharomyces using partial least squares**. *BMC Bioinformatics.* 2011, **12**(318):318.
31. Mehmood T, Warringer J, Snipen L, Sæbø S: **Improving stability and understandability of genotype-phenotype mapping in Saccharomyces using regularized variable selection in L-PLS regression**. *BMC Bioinformatics.* 2012, **13**:327.

32. Roeder AD, Haas C: **mRNA profiling using a minimum of five mRNA markers per body fluid and a novel scoring method for body fluid identification**. *Int J Legal Med.* 2013, **127**(4):707–721.
33. Hanson EK, Ballantyne J: **Rapid and inexpensive body fluid identification by RNA profiling-based multiplex High Resolution Melt (HRM) analysis**. *F1000Res.* 2013, **2**:281.
34. van den Berge M, Carracedo A, Gomes I, Graham EaM, Haas C, Hjort B, Hoff-Olsen P, Maroñas O, Mevåg B, Morling N, Niederstätter H, Parson W, Schneider PM, Court DS, Vidaki A, Sijen T: **A collaborative European exercise on mRNA-based body fluid/skin typing and interpretation of DNA and RNA results**. *Forensic Sci Int: Genetics.* 2014, **10**:40–48.
35. Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, NISC Comparative Sequencing Program, Bouffard GG, Blakesley RW, Murray PR, Green ED, Turner ML, Segre JA: **Topographical and temporal diversity of the human skin microbiome**. *Science (New York, N.Y.).* 2009, **324**(5931):1190–1192.
36. Kong HH: **Skin microbiome: genomics-based insights into the diversity and role of skin microbes**. *Trends Mol Med.* 2011, **17**(6):320–328.
37. Sender R, Fuchs S, Milo R: **Revised Estimates for the Number of Human and Bacteria Cells in the Body**. *PLoS Biol.* 2016, **14**(8).
38. Vinje H, Almøy T, Liland KH, Snipen L: **A systematic search for discriminating sites in the 16S ribosomal RNA gene**. *Microb Inform Exp.* 2014, **4**:2.
39. Consortium THMP: **Structure, Function and Diversity of the Healthy Human Microbiome**. *Nature.* 2012, **486**(7402):207–214.
40. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Strauss JF, Jefferson KK, Buck GA: **The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies**. *BMC Microbiol.* 2015, **15**.
41. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF: **PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample**. *Appl Environ Microbiol.* 2005, **71**(12):8966–8969.
42. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG: **Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling**. *Appl Environ Microbiol.* 2014, **80**(24):7583–7591.
43. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Appl Environ Microbiol.* 2007, **73**(16):5261–5267.

Figures

Figure 1 - Samples classified when cross-validating on HMP data

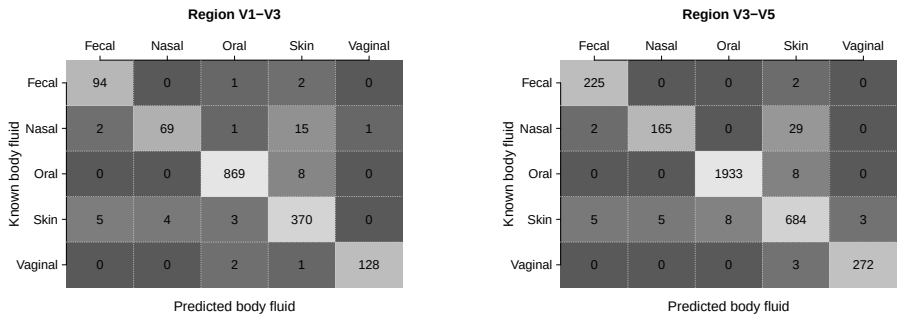

Figure 1: The number of samples assigned to the various body fluid categories after 10-fold cross-validation in the HMP data. The rows indicate the true category of the samples, while the columns are the predicted categories. The diagonal elements indicate the number of correctly classified samples for each body fluid.

Figure 2 - Selected genus important for body fluid recognition

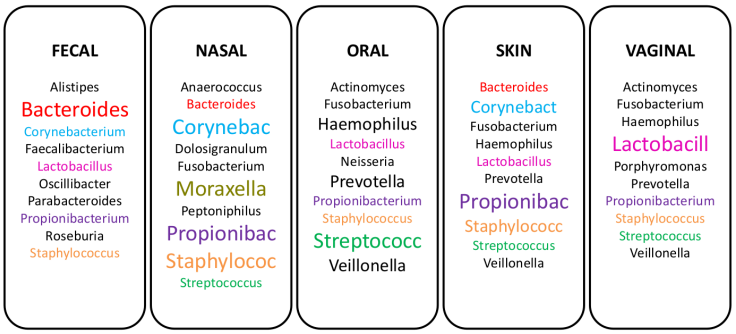

Figure 2: The 10 most frequent genera selected for each body fluid. Text size is proportional with expected abundance of a genus in a specific body fluid. For convenience, high abundance genera have been given individual colors in the figure.

Figure 3 - AGP samples classified from models trained on HMP data

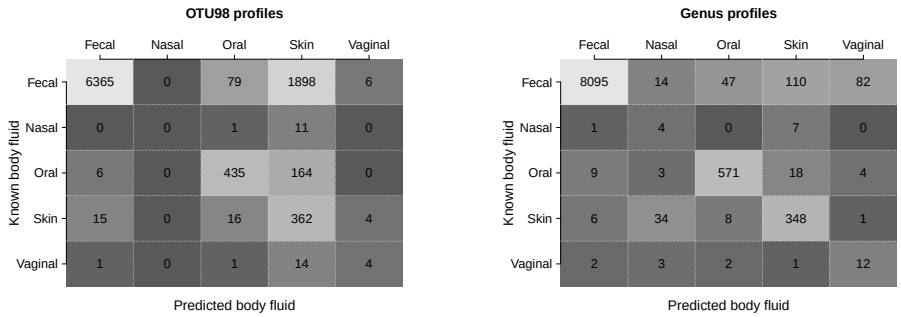

Figure 3: The number of samples assigned to the various body fluid categories after training on the HMP data and tested on the AGP data. Only results for the V3-V5 region are shown. The left panel uses the OTU98 based taxonomic profiles that turned out optimal in the cross-validation procedure (see Table 2), and the right panel uses the taxonomic profiles given by direct assignment of reads to genus.

Tables

Table 1 - Summary of HMP data used

Table 1: A summary of the data. Each sample belongs to a body fluid indicated in the left column. Then we list, for each body fluid, the number of persons contributing, the total number of samples, median number of reads per sample, average read-length (bases) in a sample and finally the original body site annotations given by the Human Microbiome Project.

Body fluid	Persons	Samples	Reads	Length	Original body site
Fecal	222	324	13980	416	Stool
Nasal	204	281	10471	425	Anterior nares
Oral	218	2827	11904	437	Attached/Keratinized gingiva, Buccal mucosa, Hard palate, Palatine tonsils, Saliva, Subgivingal plaque, Supragingivl plaque, Throat, Tongue dorsum
Vaginal	103	406	12711	431	Mid vagina, Posterior fornix, Vaginal introitus
Skin	229	1040	10747	424	Antecubital fossa (left or right), Retroauricular crease (left or right)

Table 2 - Optimizing accuracy supported by ANOVA

Table 2: ANOVA results to investigate the different factors' effect on accuracy when strictness is set to weak and initial predictor levels phylum, class and order have been removed from the analysis. Reference corresponds to Predictor=OTU98, Region=V3-V5, Pseudo-count C=1 and Regularization=1.0. In the estimate column the accuracy for the reference settings is given at the top, and the negative differences relative to this are given for the other combinations below. The p-value column gives the significant levels for the comparison of accuracy obtained with the respective combinations of setting and the reference accuracy.

Test	Estimate (accuracy)	p-value
Reference	0.9808	$< 2e-16$
Predictor=otu99	-0.0004	0.76
Predictor=otu97	-0.0003	0.82
Predictor=genus	-0.0097	$3e-12$
Predictor=family	-0.0169	$< 2e-16$
Region=V1-V3	-0.0081	$< 2e-16$
Pseudo-count C=0	-0.0122	$< 2e-16$
Pseudo-count C=0.01	-0.0028	0.02
Pseudo-count C=100	-0.0050	$6e-05$
Regularization=0.1	-0.0042	$8e-05$
Regularization=0.5	-0.0017	0.12

Table 3 - Sensitivity and Specificity

Table 3: Comparison of sensitivity and specificity for different models. Left: Cross-validation within the HMP region V3-V5 dataset. Right: Reads assigned directly to genus, training done with HMP region V3-V5 data and testing done on AGP region V3-V5 data. For both models the other optimal settings from table 2 were used. The numbers of nasal and vaginal samples were extremely low in the AGP data set, and corresponding sensitivity and specificity values are not given in the table.

	HMP region V3-V5		AGP region V3-V5	
Body fluid	Sensitivity	Specificity	Sensitivity	Specificity
Nasal	0.842	0.998	-	-
Oral	0.996	0.994	0.944	0.994
Skin	0.970	0.984	0.877	0.985
Vaginal	0.989	0.999	-	-
Fecal	0.991	0.998	0.970	0.983

Additional Files

Supplementary Table 1 - Average number of variables selected

Table S1: Average number of variable selected across selection matrices for the 10 data subsets.

Pair of Body fluid	Region V1-V3	Region V3-V5
Fecal vs Nasal	9	9
Fecal vs Oral	27	21
Fecal vs Skin	88	85
Fecal vs Vaginal	17	16
Nasal vs Oral	49	44
Nasal vs Skin	379	454
Nasal vs Vaginal	107	54
Oral vs Skin	171	1039
Oral vs Vaginal	51	220
Skin vs Vaginal	19	136