Year : 2014

Human Identification Through Analysis of the Salivary Microbiome : Proof of Principle

Sarah Leake

Human Identification Through Analysis of the Salivary Microbiome: Proof of Principle

Sarah Leake

Ecole des sciences criminelles

Faculté de droit, des sciences criminelles et d'administration publique

Université de Lausanne

Thèse présentée pour l'obtention du grade de docteure ès Sciences en science forensique

Lausanne, Octobre 2014

UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de Mme Sarah Louise LEAKE, candidate au doctorat en science forensique, intitulée

« Human identification through analysis of the salivary microbiome »

Le Président du Jury

Professeur Pierre Margot

Lausanne, le 6 octobre 2014

Abstract

Human identification has played a prominent role in forensic science for the past three decades with identification based on unique genetic traits driving the field. However, this may have limitations, for instance, for twins or samples with low quality and/or low quantity of DNA. Therefore, there is need for a new technique of human identification which can exploit these samples. Moreover, high-throughput sequencing techniques are now available which may provide an unpreviously high amount of data likely useful in forensic science.

This thesis investigates the potential for bacteria found in the salivary microbiome to be used to differentiate individuals. Two different targets (16S rRNA and *rpoB*) were chosen to maximise coverage of the salivary microbiome and when combined, they increase the power of differentiation (identification). Paired-end Illumina high-throughput sequencing was used to analyse the bacterial composition of saliva from two different people at four different time points (t_1=0 and t_2=28 days and then one year later at t_3=365 and t_4=393 days). Five major phyla dominate the samples: *Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes* and *Fusobacteria*. *Streptococcus*, a firmicute, is one of the most abundant aerobic genera found in saliva and targeting *Streptococcus rpoB* has enabled a deeper characterisation of the different streptococci species, which cannot be differentiated using 16S rRNA alone. We have observed that samples from the same person group together regardless of time of sampling. The results indicate that it is possible to distinguish two people using the bacterial microbiota present in their saliva.

This is the first study to investigate the analysis of the salivary microbiome for forensic purposes, previous studies were limited to analysing only streptococci species from saliva. Secondly, this thesis demonstrates the advantages of targeting two genes and not using only the 'gold standard' gene, 16S rRNA, for bacterial community analysis.

Resumé

L'identification humaine a joué un rôle de premier plan dans les sciences criminelles ces trois dernières décennies, en se focalisant principalement sur les traits génétiques dits 'uniques'. Cependant, une telle approche présente des limites, comme dans les cas où il faut identifier des jumeaux ou analyser des échantillons ayant une faible qualité et/ou une faible quantité d'ADN. Pour exploiter ces types d'échantillons, une nouvelle méthode d'identification est explorée, dans le cadre de la présente recherche, en se basant sur les techniques de séquenage à haut débit, pouvant fournir une quantité élevée de données potentiellement utiles pour les sciences criminelles.

Cette thèse étudie des bactéries présentes dans le microbiome salivaire et leur potentiel de différenciation entre individus. Deux cibles différentes (16S ARNr et *rpoB*) ont été choisies pour maximiser la couverture des caractéristiques du microbiome salivaire, augmentant ainsi le pouvoir de différenciation (identification) par leur combinaison dans l'analyse. Le séquenage à haut debit de Illumina de type 'paired-end a été utilisé pour analyser la composition bactérienne de la salive de deux personnes différentes à quatre temps différents ($t_1 = 0$ et $t_2 = 28$ jours, puis un an plus tard à $t_3 = 0$ et $t_4 = 28$ jours). Suite aux analyses, cinq phylums majeurs ressortent comme étant dominants dans les échantillons testés: *Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes* et *Fusobacteria*. *Streptococcus*; un firmicute, est l'un des genres aérobiques les plus abondants dans la salive. Le ciblage *Streptococcus rpoB* a permis une caractérisation plus approfondie des différentes espèces de streptocoques, lesquelles ne peuvent être différenciées qu'en utilisant uniquement le séquenage du gène 16S ARNr. Les résultats finaux indiquent qu'il est possible de caractériser deux personnes en utilisant le microbiote bactérien présent dans leur salive et ceci indépendamment du moment collecte de l'échantillon.

Cette recherche représente la première étude focalisée sur l'analyse du microbiome salivaire à des fins forensiques; les études antérieures se limitaient à la seule analyse des espèces de streptocoques dans la salive. Cette recherche observe qu'il est plus avantageux de cibler deux gènes pour l'analyse de la

communauté bactérienne, plutôt que de se focaliser uniquement sur le gène dit standard 16S ARNr.

To Steven

Acknowledgements

I would like to take this opportunity to thank all those who have supported me in some way throughout the past five years:

My thesis supervisor Prof. Franco Taroni from the School of Criminal Justice at the University of Lausanne, who allowed me to pursue such a random topic which was outside of the field of anybody at the School. Thank you for believing in me, and for all the help and support you have provided throughout my entire PhD.

My thesis co-supervisor Prof. Gilbert Greub from the Institute of Microbiology at the University Hospital of Lausanne (CHUV), who agreed to co-supervise this project after just one meeting. Thank you for also believing in me and providing me with the support I needed to undertake and finish this project.

Dr Laurent Falquet for all the help he provided from the computational side, without his insight and patience I would have been lost in the vast world of computer programming.

Dr Marco Pagni for his advice throughout the whole project regarding project progression and data analysis.

All the members of the thesis committee for reading the manuscript and their insightful feedback.

Sebastien Aeby at the Institute of Microbiology for welcoming me to the lab and providing me with all the help I needed to perform my experiments.

The group of Keith Harshman at the Genomic Facility of the University of Lausanne for performing the sample preparation and sequencing.

I must thank the School of Criminal Justice of the University of Lausanne for funding the project and the University of Lausanne's Foundation of the 450th anniversary and the Equal Opportunity Office for providing funding to go to conferences so I could disseminate my work.

Anne Marville for her administrative support and answering all of my random questions.

All my colleagues from offices 6409 and 6410 who have over the past five years provided much moral support and entertainment.

I would specifically like to thank Aline, Jenny and Durdica for their moral support both at and outside of work.

All of my friends near and far who have been there when I have needed them.

My family for all of their moral support.

Last, but definitely not least, my husband Steven, for his constant support, motivation and amazing cooking.

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Introduction

Current methods of human identification in forensic science rely heavily upon the analysis of human DNA. This has proven to be very successful often yielding full DNA profiles, which can then be compared to DNA databases for individual selection. However, there have been many cases where analysing human DNA has proven difficult. One major problem is contact DNA which occurs when a persons skin comes into contact with another person or an object either accidentally or deliberately. Two scenarios arise whereby the host can react with the transferred DNA or upon analysis the DNA profile of the host could interfere and prevent a determination of the donor's DNA profile. In the first case, bacteria and enzymes found on skin can aid the degradation of DNA making profiling more difficult. In the second case, the amount of human DNA transferred is low(0-225pg/μl) (1) therefore, the host's DNA can dominate, yielding either a mixed or unexploitable profile and the extraction process is more demanding. Low template techniques are designed to overcome this through the use of stricter DNA extraction protocols and a greater number of PCR cycles. These techniques have been scrutinised for their reliability and accuracy (2, 3, 4, 5). An independent review commissioned in the United Kingdom by the Home Office Regulation Unit states that low template DNA analysis "has been validated in accordance with scientific principles" (3). However, these techniques are still only accepted on the same terms as standard DNA typing in courts in the UK, New York[1] and New Zealand and have been used in trial evidence in Australia and Sweden (6). The issue of low DNA quantity will continue to persist in forensic science regardless of the target/technique used. This is due to a number of problems; risk of contamination, mixture analysis and transfer through mechanisms other than those associated with the

[1]This technique has only been accepted in New York and not the whole of the USA.

crime in question. Therefore, the technique I propose aims to concentrate on the problem of degraded DNA samples.

An extension of these problems occurs with the transfer of saliva through either spitting, licking, kissing or biting to skin. In this case both the quantity and quality of the transfer are low, often rendering no result. This is in part due to enzymes in saliva breaking down human DNA, as described above for contact DNA. However, as bacteria can thrive in this environment (7), due to their increased protection from degradation, they can also be sequenced and thus have the potential for identification. A new technique is therefore required to exploit these types of samples. I proposed investigating the composition of bacteria, known as the microbiome, found in saliva by analysing bacterial DNA with the aim of producing an innovative, robust, alternate identification method.

This technique could be important for offering results or potential agressors in sexual assault investigations, as kissing, biting, licking and spitting are often encountered. Sexual assault cases rarely result in a conviction due to either lack of evidence or women being too scared to prosecute[1] (8). In those cases, the main form of evidence derives from DNA found in sperm; however, if no sperm is left behind or a person was sexually assaulted but not raped then it is very difficult to support that an assault has occurred. A recent report (9) shows that in the UK the number of cases being referred to the Crown Prosecution Service (CPS) has dropped despite an increase in the number of rapes reported to the police. If this technique is successful and well communicated to the general public then victims maybe more willing to report the sexual assault and the police more likely to refer it to the CPS[2].

1.1 Human Identification

The ability to (probabilistically) identify a person through the analysis of their DNA became possible in the 1980's due to two independent breakthroughs. The first, Alec Jeffreys discovered a method for human identification through DNA - Restriction Fragment Length Polymorphism (RFLP), a technique subsequently referred to as DNA fingerprinting (10). The second, Kary Mullis discovered the Polymerase Chain Reaction (PCR) a technique used to amplify specific regions of DNA (11). Combined, these two techniques proved revolutionary for human identification and forensic science. DNA fingerprinting was discovered in 1984 and the first conviction using this technique occurred just three years

[1] Statistics from the UK show that the rate of conviction for rape cases is 6 percent.

[2] This criminological aspect is very interesting however, it is outside the scope of the current research.

later (12). In 1989 in the USA a case used DNA evidence to overturn the conviction. This shows the power DNA evidence had in Court.

With advances in technology two new breakthroughs, in the late 80's to early 90's, changed the techniques used for DNA analysis. The first, a new marker for DNA analysis, the micro-satellite or Short Tandem Repeat (STR) (13). The second, a new method of visualisation based on fluorescent labelling which when combined with PCR increased the sensitivity of the technique enabling low quantities of DNA to be analysed (13). Gradually in the past decade this technique has evolved from four STR markers to seventeen and more, with increased sensitivity. However this analytical technique still has its limitations, such as analysing degraded DNA, therefore there is need for a new technique which can exploit these limitations and can focus on different genomes.

1.2 Concept of the human microbiome

A microbiome comprises the genomes of the bacteria found at a specific site. Specifically, the human microbiome describes all the individual microbiota found within and across the human body (14). With bacterial cells outnumbering somatic and germ line cells by a factor of ten their contribution to human life should not be underestimated. Furthermore, it is thought that a human should be viewed as a combination of microbial and human cells (14). Each distinct area of the human body, for example; the oral cavity, forearm, hand and gut have their own individual microbiome. Each microbiome consists of different combinations of bacteria, with, in theory, each person having a slightly different ratio or combination of bacteria at each site.

The phylogenies of bacteria have been studied for many years, however, the depth of analysis now available has shown that the level of microbial diversity has been greatly underestimated. This has lead many scientists to investigate the level of diversity within a species (15, 16). Tettelin *et al.* (16) proposed the concept of the pan-genome, within which all the genomes of one species are contained. The pan-genome can be divided into three distinct sections: 1) core genome - shared by all strains, 2) set of dispensable genes - shared by some isolates and 3) set of strain-specific genes - unique to each isolate, highlighting that the microbiome must differ within species as well as between species. A study by Hiller *et al.* (17) of 17 *Streptococcus pneumoniae* genomes has shown that the core-genome consists of 1454 genes whereas the pan-genome contains approximately 5000 genes. Hiller *et al.* (17) conclude that 142 genomes would need to be sequenced in order to have the complete *S. pneumoniae* genome. Therefore, this demonstrates that it is not

possible to characterise a species from a single genomic sequence (18). With the arrival of high-throughput sequencing many species can now be analysed at the same time, to an adequate depth for characterisation. Moreover, the differences within and between species can now be exploited for the purpose of forensic science.

1.3 Aims

The goal of this thesis is to produce a method for analysing the salivary microbiome for the purpose of human characterisation. As mentioned above the analysis of human microbiomes has already been undertaken by other domains, namely medicine. However, their goals are different as they aim to either characterise the microbiota present and/or find which bacteria cause disease. Whereas, the aim of this study is to find out which bacteria or combination of are potentially unique to a person. The following aims are proposed to address this problem and they are approached in this research:

1. Develop a method for analysing the salivary microbiome.
    - Selection of sampling method;
    - Selection of extraction method;
    - Selection of targets and optimisation of primers;
    - Selection of sequencing method;
    - Selection and optimisation of sequence processing programs.
2. Analyse the salivary microbiome of two individuals.
    - What bacteria are present?;
    - What are the most abundant bacteria and why?
3. Analyse the differentiation of individuals through analysis of their salivary microbiome.
    - How to compare individuals?;
    - How to combine target genes?;
    - What is the minimum number of sequences required for differentiation?

1.4 Relationship of the proposed work to existing literature and its originality

At a laboratory level, high-throughput sequencing techniques have only been developed over the past few years and therefore are still in their infancy. This has lead many groups to try to understand and improve the different analytical methods (19, 20, 21, 22, 23, 24). During this time these techniques have come to the forefront of many domains of biological research. Many researchers have used this technique to investigate the bacterial diversity in different environmental communities and habitats on the human body (14, 25, 26, 27, 28, 29). For saliva, the emphasis has been in the medical environment, notably on oral diseases, essentially in how different bacteria can cause different diseases and whether the detection of certain bacteria can be used as a diagnostic tool (30, 31, 32, 33, 34, 35, 36). However, there are very few articles linking these techniques with forensic science and its (forensic) value has not yet been established. In the domain of forensic science, Fierer *et al.* (37) investigated the use of bacteria for human identification concentrating on the potential of analysing skin bacterial communities. They suggested that the bacteria left behind after touching a surface could be used to identify the person. These results are promising, indicating that it could be possible to use microbiome analysis for forensic identification. However, the use of bacteria found in saliva for human identification has only been investigated in relation to bite-mark analysis, more specifically the analysis of streptococcal DNA (38, 39, 40, 41). This paves the way for a complete technique for human identification based on bacteria found in saliva.

In summary, high-throughput sequencing techniques are still being developed and improved and human identification methods rely heavily on human DNA. Therefore, there is the market for a new method for human identification in forensic science which can exploit where human DNA fails. This thesis explores whether the analysis of the salivary microbiome can be used to differentiate two individuals and therefore, be a potential new method for human identification.

1.5 Content

To respond to the aims exposed above this thesis will be laid out as follows: first the background to saliva and it's use will be presented followed by the materials and methods which will explain all techniques used. Next, the results concerning each of the three principle aims will be presented in individual chapters, followed by a discussion chapter

which will discuss each aim separately and then bring everything together presenting a global discussion of the problem and solution.

2

Background

This chapter will introduce saliva, describing it's bacterial composition and the stability and variability of the bacteria. The concept of metagenomics will also be presented.

2.1 Saliva

Saliva is the fluid tissue, found in the mouth of humans and other animals, which is produced and secreted by the salivary glands. Glantz (42) describes saliva as a fluid tissue, not just a solution, which can be divided into four levels of organisation: 1. continuous phase of electrolytes in water 2. a scaffold-like network structure 3. less water-soluble proteins and salivary micelles contained in the network filaments 4. lipoid material, bacterial and epithelial cells. 99.5% of human saliva is water whilst the remaining 0.5% contains many enzymes including α-amylase and lingual lipase, antibacterial compounds such as lysozyme and lactoferrin, electrolytes including sodium and potassium and mucus which contains mucopolysaccharides and glycoproteins. The different concentrations of electrolytes makes saliva a hypotonic solution (43). Saliva has many functions: it wets food to help swallowing, enzymes in saliva initiate digestion, it helps tasting by wetting the tongue aiding it to differentiate between flavours and washing of saliva over teeth helps keep them clean and protected from bacteria that cause decay.

Estimates show that on average a healthy human being will produce between 0.5 and 1.5 litres (L) of saliva per day (43). Saliva is specifically produced by the contra-lateral major glands and the minor salivary glands. The different glands vary in the types of secretions produced, these differences are due to the ratio of serous to glandular cells (44). Serous cells produce a watery fluid and the secretion of which is strongly activated by stimuli

whereas mucous cells produce a mucus-rich fluid. Other factors contribute to the composition of whole saliva including; blood, oral tissues and microorganisms (45, 46). Many factors influence the amount and composition of secreted human saliva; circadian rhythm (47), flow rate, type and size of salivary gland (48), duration and type of stimulus, diet, drugs, age, gender and blood type (49). With such a large variation in the composition of saliva a standardised method for collection is needed to minimise, as much as possible, any differences. Using a non-standardised method can result in high-variability in obtained data. When analysing whole saliva three factors need to be taken into account. Primarily, saliva contains components which originate from cells and bacteria therefore it can be difficult to determine which parts truly come from saliva. Secondly, over time bacterial metabolism changes the composition of saliva. Finally, cellular debris can inhibit some analytical techniques. These three factors can be dealt with through pre-treatment of whole saliva, for example, centrifugation and correct storage. However, this can also cause the composition of saliva to change. Centrifugation minimises bacterial action, however for this project the main focus is to investigate the bacteria found in saliva therefore centrifugation would not be appropriate. Saliva is inhomogeneous as it can simultaneously consist of liquid, gas and gel phases a property which can subsequently affect the reliability of measurements (50). To prevent changes in the number and proportion of bacteria present in saliva, samples should be handled in the cold with minimal time between collection and analysis. When analysing α-amylase the storage of the sample is very important as at +4°C the enzyme remains stable for a few days, however, if frozen the protein can precipitate (51). When analysing the data all of the above factors need to be taken into consideration.

Saliva is useful to forensic scientists for many reasons: it is easy to sample and less invasive than sampling blood or urine, traces of drugs can be detected and the levels of α-amylase are exploited in presumptive tests for saliva.

2.1.1 Bacteria

Saliva has its own microbiome consisting of specific bacteria and as many as 500 million bacterial cells can be found in one millilitre (mL) of saliva. It has been shown there are at least 700 bacterial species found in the mouth (52). One of the principle functions of saliva is to wet the mouth and therefore, saliva covers all surfaces found in the mouth except for deep cracks and periodontal pockets (34, 53). The oral cavity is the ideal habitat for aerobic and especially anaerobic bacteria thanks to a constantly high temperature, humidity and the regular arrival of food (54). The oral cavity contains two distinct surface

types; teeth which have a solid, mineralised and irregular surface and mucosal membranes such as gums and the tongue which are soft, stratified and regularly moult epithelial cells. Due to these differences, both surfaces are colonised by different bacteria and/or different concentrations of bacteria (30, 55). The bacteria colonising these surfaces form biofilms, which consist of layers of different bacteria (30, 34). In actual fact, the bacteria found in saliva come from the oral cavity or external sources. Specifically, Marsh and Martin state that the majority of bacteria in saliva comes from the tongue (56) and Hamilton and Bowden state that the concentration of certain bacteria in saliva reflects their respective concentration in dental plaque (53). The most commonly found bacteria (by both culture and sequencing) in the oral cavity and saliva will be presented below.

2.1.1.1 Gram-positive[1] cocci

Streptococcus is the most commonly found genus in the oral cavity and has been isolated from all oral sites (57). The genus *Streptococcus* is separated into four groups: mutans, salivarius, anginosus and mitis (41). From the mutans group *S. mutans* and *S. sobrinus* are the most common along with *S. criceti* and *S. ratti*. These species have specifically been found attached to teeth. The salivarius group contains *S. salivarius*, *S. thermophilus* and *S. vestibularis* all of which have been found on all surfaces of the oral cavity but they prefer to colonise mucosal membranes, specifically the tongue. The anginosus group contains *S. constellatus*, *S. intermedius* and *S. anginosus* all of which have been isolated from dental plaque and mucosal membranes. The final group contains many species including *S. mitis*, *S. oralis* and *S. sanguinis*. The first two are amongst the most commonly found bacteria in the oral cavity. Studies have shown that all three are initial colonisers of teeth with *S. mitis* being the most predominant (58, 59). Furthermore, previous studies have also shown that humans have many different strains of the same *Streptococcus* species with many strains being unique to individuals (60, 61).

Other Gram-positive cocci found in the oral cavity are *Granulicatella adiacens*, a bacterium which colonises all oral surfaces, *Abiotrophia defectiva* and other species from the genus *Gemella*. It is also possible to detect *Peptostreptococcus stomati* and more rarely species from the genus *Enterococcus* of which the commonly detected is *Enterococcus faecalis*. Species from the genera *Staphylococcus* and *Micrococcus* are rarely isolated from the oral cavity despite their abundances in neighbouring sites such as the skin and nasal mucosal membranes (62).

[1]Bacteria which colour blue-violet after application of Gram staining are classified as Gram-positive

2.1.1.2 Gram-positive bacilli

Species of the genus *Actinomyces* form a large portion of the dental plaque microbiome. Of the numerous *Actinomyces* species *A. naeslundii* and *A. oris* are the most common. A number of other genera including *Eubacterium*, *Mogibacterium*, *Pseudoramibacter*, *Slackia*, *Cryptobacterium*, *Shuttleworthia*, *Solobacterium* and *Bulleidia* have been detected in the buccal microbiome, however they are more commonly associated with peridontal diseases.

The genus *Lactobacillus* is often found in the oral cavity. However, it constitutes less than 1% of the cultivable oral microbiota. Other bacteria found include *Propionibacteria*, which are obligate anaerobes, *Corybacterium matruchotti*, *Rothia dentocariosa* and *Bifibacterium dentium* which are often isolated from dental plaque. Along with *Rothia dentocariosa*, *B. dentium* is also isolated from the tongue. Furthermore, species from the genera *Arcanobacterium* and *Actinobaculum* along with *Alloscardovia omnicolens* also inhabit the oral cavity (62).

2.1.1.3 Gram-negative[1] cocci

The genus *Neisseria* which consists of both facultative aerobic and anaerobic cocci is found in nearly all oral sites in low quantities. The most common species found are *N. subflava*, *N. mucosa*, *N. flavescens* and *N. pharyngis*. Species of the genus *Veillonella*, which are strict anaerobes, including *V. parvula*, *V. dispar*, *V. atypica*, *V. denticariosi* and *V. rogosae* constitute an important part of the oral microbiome. Another anaerobe *Megasphaera* is occasionally detected in dental plaque (62).

2.1.1.4 Gram-negative bacilli

The majority of Gram-negative bacilli are found in dental plaque, for example, *Haemophilus parainfluenzae*, the only species of the genus *Haemophilus* to have been detected in the oral cavity. Other examples of facultative anaerobes found in the oral cavity are species from the genus *Capnocytophaga*. Also found is *Kingella oralis*, a coccobacillus found in many sites throughout the mouth, *Aggregatibacter actinomycetemcomitans* and species from the genus *Simonsiella* which colonise the epithelial layer of the buccal cavity and *Eikenella* (62).

[1]Bacteria which colour red or pink after application of Gram staining are classified as Gram-negative

A large part of the bacterial flora found in dental plaque and on the tongue are obligate anaerobes, of which the most prevalent come from the genera *Prevotella* and *Porphyromonas*. Another important obligate anaerobe is *Fusobacterium*. *Leptorichia buccalis* has also been detected along with species from the genera *Campylobacter* and *Selenomonas*. Other species from the following genera *Centipeda*, *Johnsonella*, *Catonella* have been isolated but only from individuals with an oral disease. Furthermore, species from the genera *Dialister*, *Flavobacterium*, *Tannerella*, *Desulfomicrobium*, *Desulfovibrio* and *Methanobrevibacter* have been detected along with oral spirochaetes which are classified under the genus *Treponema* (62).

2.1.1.5 Mycoplasma

Mycoplasma is a genus of bacteria which have no cell wall and therefore cannot be classed as either Gram-negative or Gram-positive. Between 6 and 32% of the population carry this genus which includes species such as, *Mycoplasma buccale*, *M. orale*, *M. pneumoniae*, *M. salivarium* and *M. hominis*. Specifically, the last three have been found in human saliva (56).

2.1.1.6 Bacterial interactions

Having presented which species are found in saliva, it is worth noting that certain genera have been shown to be positively correlated (63). Specifically, Li *et al.* demonstrated that the following genus pairs were positively correlated: *Fusobacterium/Porphyromonas, Fusobacterium/Prevotella, Prevotella/Veillonella, Streptococcus/Actinomyces* and *Veillonella/Actinomyces*. These correlations correspond with the spatio-temporal model of oral bacterial colonisation (64, 65). They also found that genera from the same phylum tended to correlate positively with each other, especially for *Proteobacteria* and *Firmicutes*, the two most popular phyla in saliva. In another study (66) co-occurrence of bacteria was investigated, showing that the highly abundant genera such as *Streptococcus, Neisseria* and *Haemophilus* are always found together, whereas other genera, such as, *Abiotrophia* and *Dialister* are unlikely to be found together. When comparing taxa within a single phylum they found that only taxa within Firmicutes scored highly suggesting segregation of species and possible competitive species interactions. This is interesting as *Streptococcus* is the most abundant genus found in saliva and belongs to Firmicutes, indicating that different people may have different *Streptococcus* species due to this interaction. Bacterial interactions could have an impact on the interpretation of data as each bacterium does

not have the same chance of being present, so the data may have to be interpreted as a whole and not species by species.

2.1.1.7 Variability

It is evident that all the different bacteria presented above are not found in everybody. This is mainly due to different lifestyle habits of individuals, for example, smokers will have some bacteria that non-smokers do not have (see section 7.4.3 for more details). Furthermore, the abundances of bacteria will vary, especially through neglected oral hygiene and periodontal diseases (32, 67, 68, 69). Nevertheless, it has been possible to define the genera most commonly found in the oral cavity and present in everybody as: *Streptococcus*, *Neisseria*, *Haemophilus*, *Campylobacter*, *Veillonella*, *Fusobacterium*, *Rothia*, *Actinomyces*, *Prevotella*, *Corynebacterium*, *Capnocytophaga*, *Atopobium*, *Granulicatella* and *Bergeyella* of which the most abundant are *Streptococcus, Prevotella, Veillonella, Neisseria, Haemophilus* and *Rothia* (66). Specifically the salivary microbiome consists of eight genera which account for more than 70% of the population: *Streptococcus, Prevotella, Veillonella, Neisseria, Haemophilus, Rothia, Porphyromonas* and *Fusobacterium* (70). In addition, Nasidze *et al.* (70) showed the number of genera per individual ranged from six to thirty.

There are two types of variability; the variability between different individuals (inter-individual variability) and the variability within the same individual (intra-individual variability). For this idea to progress the intra-individual variability needed to be small and the inter-individual variability as large as possible. Due to the dynamic nature of the salivary microbiome individual variability is extremely likely. A study by Nasidze *et al.* (70) investigated the global diversity of the human salivary microbiome. They observed that the level of variation was significantly higher between individuals than within the same individual. Specifically, about 13.5% of variation in the distribution of genera was due to differences between individuals. In addition, they found that sequencing 120 clones from each individual was sufficient for analysing variation at the genus level. Another study by Zaura *et al.* (31) has shown that when comparing the oral microbiome of three different people 11-20% of the sequence-reads corresponded to unique sequences which were not shared. Furthermore, Bik *et al.* analysed the bacterial diversity in the oral cavity of ten healthy individuals and demonstrated that large inter-individual differences were present (66).

Further studies have shown microbial communities to be highly variable within and between people. However, for this thesis, it was essential to find out whether the variabil-

ity within a person was significantly lower than the variability between different people. Costello *et al.* (26) undertook an experiment which investigated 'bacterial community variation in human body habitats across space and time'. They analysed variable region 2 (V2) of the bacterial 16S ribosomal RNA encoding gene. Their results showed the range of bacteria found in the oral cavity was significantly less variable both within and between people. This could be partly due to the method of analysis as *Streptococcus* is one of the most common bacteria found in the mouth, is very abundant and the technique used might not have been accurate enough to separate all the different *Streptococcus* species. However, the results showed the size of the 'core' bacteria, the bacteria common to every individual, is likely to be larger in the oral cavity than, for example, skin. Subsequently, the oral microbiome differed the most from the skin microbiome. They also showed that when a forearm was inoculated with tongue bacteria the transferred bacteria were more similar to the tongue bacteria than the forearm bacteria. These two factors will be very important when investigating the transfer of saliva to the forearm (26).

2.1.1.8 Stability

The stability of the salivary microbiome is of utmost importance. The advancement of this project depended on demonstrating the stability of the salivary microbiome over time. The salivary microbiome needs to be stable enough for samples from the same person to group together. If this was not the case then the technique proposed could not be used as a means of identification. Firstly, unlike fully internal microbiomes, the salivary microbiome regularly comes into contact with the external environment through talking, breathing and eating, thus producing a dynamic microbiome (52). The following question was subsequently raised; does each individual create their own dynamic but stable microbiome? Costello *et al.* (26) concluded that each microbiome remains relatively stable over time and Lazarevic *et al.* demonstrated that the salivary microbiome is stable over one month (71).

It has been shown that the composition of oral bacterial microflora changes with age. At birth the oral cavity is sterile however, after six hours bacterial colonisation starts with species from *Streptococcus* (including *S. pneumoniae*), *Micrococcus, Enterococcus, Staphylococcus, Veillonella* and many others. The most commonly found species in infants is *Streptococcus salivarius* which mainly colonises the tongue. As teeth start to push through new species can proliferate and dental plaque starts to form with the colonisation of *Actinomyces* and diverse *Streptococci* species (72, 73). From the age of five, the composition of the oral microflora is comparable to that of an adult with the exception of

Spirochaetes and *Prevotella melaninogenica*, which are often absent (53). From here on, little change is observed in the oral microbiome, however if teeth are lost then a marked reduction in *Spirochaetes*, *Lactobacillus, Streptococcus mutans* and *Streptococcus sanguis* has been observed (72). Stahringer *et al.* showed that between the ages of 12-24 *Streptococcus* and *Actinomyces* increased with age and *Veillonella* decreased with age (74). However, an in depth study over a long time period, concentrating on adult saliva, is required to ascertain the stability of the salivary microbiome and whether it is stable enough for use in forensic science.

2.1.1.9 Persistence

Bacteria in saliva could persist for longer than human DNA because bacterial DNA is better protected than human DNA for two main reasons: 1. bacteria are prokaryotes and prokaryotic cells have a cell wall which protects the cell (75), whereas animal eukaryotic cells only have a cell membrane 2. in general, prokaryotic DNA is circular making it harder to break down whereas animal DNA is linear and can be more easily attacked at each end. This demonstrates a distinct advantage of bacterial DNA. Traces at crime scenes are often found in sub-optimal conditions, meaning that they are quite quickly subject to degradation. As bacterial DNA is innately better protected than human DNA it should provide a more robust target for forensic analysis. Therefore, persistence can be viewed as even more important than stability.

2.2 Metagenomics

Metagenomics is a new domain to have stemmed from novel technologies such as high-throughput sequencing. Metagenomics is the study of the genomes of all the organisms in a particular environment which, investigates the environment as a whole and includes all types of analysis from sequence-based to product or function based methods. The Human Microbiome Project (HMP), an extension of the Human Genome Project, consists of a number of metagenomics projects including studies on the oral cavity (30, 31, 52). The aim of the project is to characterise the human microbiome, including trying to understand the factors which influence the distribution and evolution of the bacteria, in order to see the effect the bacteria have on health and pre-disposition to various diseases (14). All of the data produced by these projects is uploaded into freely available databases, for example, at www.hmpdacc.org where currently around 3000 bacterial genome sequences are available. Most medically relevant human microbiomes have been analysed, for example, the gut

and the oral cavity. However, metagenomics is in its infancy, as are the techniques used, therefore there is still a lot of room for exploration. A study has shown that most of the bacteria found in the gut are rare, making them difficult to detect (76). If these rare bacteria are important then the level of characterisation of the microbiome needs to be very deep. If a similar pattern is seen in the salivary microbiome then more detail will be required from each sample to get an accurate representation of an individual's salivary microbiome. Caporaso *et al.* (77) investigated different human microbiomes, including the tongue, at a depth of millions of sequences per sample. Upon sub-sampling the data they found that only 2000 sequences were required to represent the same relationships found using the whole dataset. This shows that, in fact, very few sequences may be required to describe a specific human microbiome. However, it is important to elucidate the minimum number of sequences required to accurately characterise the salivary microbiome for forensic purposes, as more sequences may be required to differentiate two individuals. One way to characterise a microbiome at greater depth is to sequence more than one gene fragment in order to obtain better coverage of the microbiome. A second option could be to perform shotgun sequencing. Furthermore, if the diversity is high then a large number of sequences would still be required to attain the necessary coverage (78). As proposed in section 1.3, this thesis aims to work out how many sequences are required to differentiate two individuals.

3

Materials & methods

This chapter deals with the materials and methods used from the sampling of saliva through to the processing and analysis of the sequencing data. Statistical methods used to interpret the data are also briefly presented.

3.1 Sampling

The exploratory nature of this thesis meant that healthy individuals were used to limit the effects of any additional factors such as the use of antibiotics or smoking. Due to the cost of analysis only two individuals, a male and a female, were sampled. At the outset of this project the allocated budget was 16,000CHF which was sufficient to undertake one experiment (4 samples) however, by the time the first experiment was performed the cost of analysis had halved enabling the second experiment to be performed within the same budget. The chosen individuals were asked to brush their teeth in the morning and then not consume any food for one hour prior to sample collection. Saliva was collected by spitting into a sterile tube. Samples were stored at $-20\,^{\circ}\mathrm{C}$ until analysis to avoid any change in the bacterial flora. If samples are stored at $4\,^{\circ}\mathrm{C}$ they need to be processed within 24 hours. Samples were taken at four time points t_1=0, t_2=28 days and one year later at t_3=0 and t_4=28 days. Saliva was sampled as the pure fluid to provide the best characterisation possible.

3.2 DNA extraction and amplification

To analyse the bacterial composition of the salivary microbiome first the bacterial DNA was extracted. In order to standardise the extraction process and eliminate as much contamination as possible the MagNA Pure 96 DNA extraction system was used. This system uses the MagNA Pure Magnetic Glass Particle Technology, the main steps are as follows (79) :

1. Sample material lysed releasing nucleic acids and nucleases denatured.
2. Due to the chaotropic salt conditions and high ionic strength of the lysis/binding buffer the nucleic acids bind to the silica surface of the added magnetic glass particles (MGPs).
3. Nucleic acid bound MGPs are magnetically separated from the residual lysed sample.
4. Several washing steps used to remove unbound substances (e.g. PCR inhibitors)
5. Purified nucleic acids are eluted from the MGPs.

All samples were extracted using the above technique specifically using the MagNA Pure 96 DNA and viral nucleic acid small volume kit following the pathogen universal 200 v2.0 protocol (79) and then specific targets described below were amplified using the polymerase chain reaction (PCR).

3.2.1 Targets

The choice of target for high-throughput sequencing is very important as different targets result in different depths of sequencing. Currently most published results are derived from sequencing one or more of the variable regions of the 16S rRNA gene including V2, V4 and V6 (see Figure 3.1). As shown in Figure 3.1, the 16S rRNA gene is composed of nine conserved and hypervariable regions. The hypervariable regions are exploited for taxonomy and the conserved regions are used for primer binding. The conserved regions are called such as they are conserved across nearly all species of bacteria enabling the use of universal primers to amplify the hypervariable region(s) of almost all bacteria simultaneously. The 16S rRNA gene is highly conserved as it is essential for bacterial life. It has a structural role in which it provides a scaffold which defines the position of the ribosomal protein. In addition it stabilises correct codon-anticodon pairing in the A site along with initiation of protein synthesis. Regions V2 and V4 have been shown to give the lowest error rates when assigning taxonomy (80) and can be used for community clustering

(81). Primer design is another important step which can influence results (80, 82). Over- or under-representation of specific taxa can be caused by primer bias along with some taxa being missed completely (83). A recent study by Lazarevic *et al.* (84) explored the use of Illumina sequencing technology for a metagenomic study of the oral microbiota. They targeted the V5 hypervariable region of 16S rRNA which is smaller than some of the other hypervariable regions of 16S rRNA. They found there were a number of advantages to targeting a smaller region, including; a reduction in the chance of producing chimera and an increase in the probability of detecting low-abundance taxa. Liu *et al.* (81) demonstrated that with carefully chosen primers short sequences can yield equally as accurate microbial community analysis as longer sequences. Due to the use of different primers by different groups direct comparison of data between studies is difficult and caution should be taken.

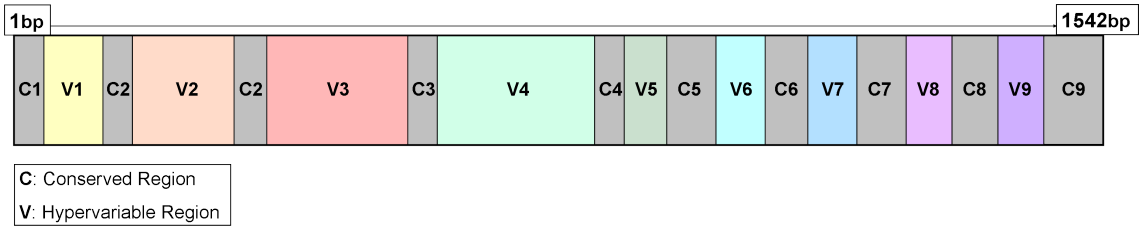

Figure 3.1: Schematic representation of 16S rRNA gene highlighting the conserved and hypervariable regions.

There are limitations to using 16S rRNA; some organisms have polymorphisms in the primer binding region and therefore these species are either poorly detected or completely lost during amplification (85). It is also poor at resolving sub-populations within species; as it is based on a single locus it lacks the necessary resolution (86). Many bacterial genomes contain multiple copies of 16S rRNA and the number of copies varies between species and can even vary within the same species (87), therefore it is difficult to make accurate abundance estimates. Other studies have shown that using a housekeeping gene such as *rpoB*, an RNA polymerase, can give better depth of analysis. Thus, sequences which cannot be separated using the 16S rRNA gene can be separated by *rpoB*, giving a better idea of the composition of a particular microbiome. *rpoB* tends to occur in single copy making it ideal for abundance estimates (88).

rpoB encodes the β-subunit of RNA polymerase, a very important enzyme, which is highly conserved throughout bacteria. The main function of RNA polymerase is the synthesis of mRNA, rRNA and tRNA, specifically the β-subunit provides most of the catalytic function (89). It has been shown that like the 16S rRNA gene the *rpoB* gene contains alternating variable and conserved regions (90). The hypervariable region of

rpoB has shown promise for bacterial identification down to the species and subspecies levels (91, 92). However, *rpoB* is in general less conserved and therefore more specific primers are necessary. Combined with 16S rRNA a deeper level of identification should be obtained. To enhance this identification a third gene can be added, in this case 23S rRNA was selected.

23S rRNA forms part of the large prokaryotic subunit 50S which contains the ribosomal peptidyl transferase. The main enzymatic function of this protein is the formation of peptide bonds between adjacent amino acids during protein biosynthesis. This is an essential process and thus 23S is highly conserved. 23S rRNA shows a similar pattern of variable and conserved regions as 16S rRNA and can therefore be used for biodiversity analysis (93). 23S rRNA is larger than 16S rRNA and thus could offer deeper phylogenetic analysis due to larger sequence variation (94).

Studies have shown that it is unrealistic to try and characterise all known species from a single genome sequence (15, 16). Therefore an adequate depth of coverage is required for DNA sequencing for the accurate characterisation of the bacterial composition of a sample.

3.2.2 Primer Design

With the targets chosen the subsequent step was to design suitable primers for each target. For each target a set of species were chosen to base the primer design on (see Table 3.1). For 16S/23S rRNA the species were chosen as they come from the top four phyla (Firmicutes, Bacteroidetes, Proteobacteria and Actinobacteria) previously found in the saliva microbiome (84). With these two pairs the broadest coverage possible was desired which is why species representing the top four phyla were chosen and not just the top phylum. Specifically, *Staphylococcus aureus* is a Firmicute found in the human respiratory tract (95), *Escherichia coli* is a proteobacterium found in the human gut (96), *Mycobacterium* is an actinobacterium which was chosen as it is different and if successfully included would add more diversity (97), *Bacteroides fragilis* is a bacteroidete and the most common species found in clinical specimens (98) and *Afipia broomeae* is a proteobacterium found in human sputum (99).

For *rpoB*1 all of the chosen species come from the phylum Firmicutes, the top phylum found in the saliva microbiome. *Streptococcus pyogenes* and *Streptococcus bovis* were chosen as *Streptococcus* is one of the most common bacteria found in the mouth (100). Along with *Staphylococcus aureus*, described above and *Enterococcus faecalis* which is

16S/23S rRNA	RpoB1	RpoB2
Staphylococcus aureus	*Staphylococcus aureus*	*Escherichia coli*
Escherichia coli	*Streptococcus pyogenes*	*Klebsiella oxytoca*
Mycobacterium	*Streptococcus bovis*	*Serratia marcescens*
Bacteroides fragilis	*Enterococcus faecalis*	*Enterobacter cloacae*
Afipia broomeae		

Table 3.1: List of species used for primer design by target

frequently found in root canal-treated teeth (101). For *rpoB*2 all of the chosen species come from the phylum proteobacteria, the third most common phylum found in the saliva microbiome. Specifically, *Klebsiella oxytoca* is associated with nosocomial infections and can colonise different areas of the human body (102), *Serratia marcescens* is also linked to nosocomial infections and is commonly found in the respiratory tract (103), *Enterobacter cloacae* is found in the human gut and has been associated with respiratory tract infections (104) and *Escherichia coli* described above. For both *rpoB* primer pairs the aim was to have a deeper coverage of species only covered at the genus or higher level with 16S rRNA/23S rRNA, for example, streptococcus, hence the use of only one phylum.

Before starting from scratch a literature search was carried out to see whether any suitable primers had already been designed and tested. Adekambi and Drancourt (105) published a table of commonly used primers for 16S rRNA and *rpoB* however, due to the small insert size required for Illumina sequencing, the distances between the primers was too large to be applicable to this study. For 23S rRNA no commonly used primers were found.

The first stage of primer design invovled extracting the sequence for each of the chosen species, see Table 3.1, from the NCBI nucleotide database in 'fasta' format. The sequences along with their GI number, taxon name and gene description were pasted into a word processor. The sequences were subsequently uploaded into Multalin, an online tool for aligning multiple sequences (106). As described in the previous section 3.2.1 the structure of the genes allows for primers to be designed in conserved regions and amplify hypervariable regions, however it is not always possible to design primers in a completely conserved region, this can be due to a number of reasons. Firstly, as Illumina was the chosen method of sequencing the size of the amplified region and primers was, at the time of design, limited to 110 base pairs (bp). Secondly, it depends on the taxa used for the alignment, the broader the range of taxa the harder it is to find completely conserved regions. For 16S rRNA a mis-match of one base pair was necessary to be able to include Bacteroidetes. As the mis-match is in the middle the primer should be bound before the polymerase reaches the mis-match, however the stringency of the primer is lowered and sequences other than

those intended may also be amplified. In this case as the majority of bacteria do not have the mis-match they will be amplified normally.

For both 16S rRNA and 23S rRNA the alignment was re-done using species specific to saliva to see whether the designed primers fitted these species. As the *rpoB*1 species showed no conserved regions large enough for primer design, a new set of species specific to *Streptococcus* were chosen instead as they are the most abundant genus in saliva and 16S rRNA cannot resolve this genus to the species level.

Ideally primers should be as short as possible to maximise the region of analysis, however a minimum melting temperature (Tm) of 50 °C is needed. For multiplexing primers should all have similar if not the same Tm. The Promega BioMath calculator was used to calculate the expected Tm of each primer (107). All primers were tested for self-complementarity using Oligo Calc (108) and all showed none. See Table 3.2 for the list of proposed primers.

3.2.3 Virtual Simulation

In order to decide whether the chosen primers would yield the best results, virtual simulation was used to simulate PCR *in silico*. Two different simulations were needed to first check that the primers would only amplify bacterial DNA and second to discover which level of classification could be reached (see Figure 3.2) and the number of different organisms within that classification. For the first simulation all primers were compared against the human genome using BLAST (basic local alignment search tool) (109). For the second simulation (*in silico* PCR) ecoPCR, a tool designed specifically to test primers for assessing biodiversity in environmental and taxonomic studies on large databases was used (110). The NCBI nucleotide database was used to act as the sample. To analyse the results a dataset of the same size as what was expected from sequencing was used to avoid any up-scaling problems later on.

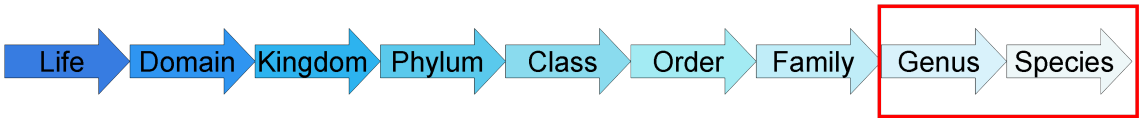

Figure 3.2: Flow diagram of the hierarchy of taxonomic classifications Highlighted by the red box, genus and species are the two classifications that are the most important to this research.

ecoPCR simulates a PCR reaction and outputs a table with the following information: accession Number, sequence length, taxonomic (tax) ID, rank, species tax ID, scientific

Primers	Difference between primers (bp)	no. mis-matches	Primer	Sequence (5'–3')	Tm ($^{\circ}C$)
16s rRNA					
783F-891R	114	46	792F	AGGATTAGATACCCTGGTAG	49
			891R	CGTACTCCCCAGGCGG	57
1097F-1175R	97	37	1097F	CCGCAACGAGCGCAACCC	63
			1175R	GAGGAAGGTGGGGATGAC	54
23s rRNA					
1831F-1924R	115	31	1831F	GCCTGCCCGGTGC	55
			1924R	GGAATTTCGCTACCTTAGG	50
RpoB1					
130F-220R	108	12	130F	GGACCTGGTGGTTTGAC	52
			220R	CGATGTTAGGGTCCTTCAGG	52
340F2-439R	115	33	340F2	GAAATCGTTTGGTTGACAGC	53
			439R	CCTTGATGACGTCCCAT	51
RpoB2					
340F-434R	110	13	340F	GGACCAGAACAACCCG	52
			434R	GGGTGTCCGTCTCGAAC	52

Table 3.2: Suggested Primers - Overview of primers designed for each gene target. Primer name for 16S rRNA, 23S rRNA and *rpoB2* corresponds to the *E.coli* positions and for *rpoB1* to *S.bovis* positions.

name, genus tax ID, genus name, family tax ID, family name, super kingdom tax ID, super kingdom name, strand, first oligonucleotide, number of errors first strand, second oligonucleotide, number of errors second strand, amplification length, sequence description. From this information the most popular genera/species amplified were compiled and checked to see whether they corresponded with bacteria commonly found in saliva.

In 2010, when this project started, Illumina sequencing produced about 30 million reads in total, 4 samples were multiplexed, hence producing 7.5 million reads per sample. As previously stated, studies have shown that it is impossible to characterise a species from a single genome sequence, (15, 16) therefore, the more reads per sample, the better the coverage. If the virtual simulation proved successful with all genes tested yielding positive results then all 3 genes with two sets of primers for both 16S rRNA and *rpoB* would be used. With 7.5 million reads per sample and the analysis of 5 regions, 1.5 million sequences per region will potentially be produced. By using five different regions the results can be combined to increase the power of identification, this means to increase the probability of identifying an individual.

3.2.4 Primer optimisation

The next step after testing the primers *in silico* was to test and optimise the primers in the laboratory. Firstly, the primers were used to amplify human DNA, *Escherichia coli* and *Streptococcus mitis*. Both *Escherichia coli* and *Streptococcus mitis* were purchased from the American Type Culture Collection (ATCC). The Phusion® Hot Start II polymerase was used as it is a high fidelity enzyme which limits amplification errors. To check whether the amplification had worked the samples were run on an acrylamide gel (see appendix B (section 8.4) for protocols). Initially the Tm used was the one calculated using Promega BioMath (see above), however the results were not optimum. The Phusion enzyme has its own formula for calculating Tm, this formula was used and the result averaged with the Promega BioMath calculator result to give a more accurate Tm. To experimentally find the optimum Tm the reaction was carried out using a temperature gradient of 56 °C to 66 °C in increments of 2 °C. Once the primers had been optimised using human DNA, *E. coli* and *S. mitis* they were tested on pure saliva samples.

3.3 Sequencing Methods

This section presents both types of sequencing; traditional and high-throughput, along with their limitations for the analysis of bacteria. All of these are presented to show where

the analysis of bacteria is at the moment and why Illumina was the chosen technique.

3.3.1 Traditional methods

The principle traditional method for analysing bacteria is culturing and the original method of sequencing is Sanger sequencing, so both of these are presented below.

3.3.1.1 Culturing

The traditional way to investigate the types of bacteria found in a sample involves culturing the bacteria on different media. The shape, colour and size of the colonies can be used to identify each type of bacteria. This technique is mainly used in diagnostics for determining the cause of an infectious disease. Many diseases are site specific so when undertaking diagnostic tests a doctor knows, in general, which bacteria to look for so picks the media necessary to grow the specific family/genus. Whereas, for this project, identifying as many bacteria as possible is key, therefore, a technique which does not rely on choosing the right types of media would be necessary. In addition, this method is not suitable for this project as only the more common types of bacteria are identified not the rarer ones. It has also been estimated that >99% of bacteria found in the environment cannot be cultured (111), more specifically this value is around 30-50% for bacteria found in the oral microbiome, these figures reiterate the boundaries of this technique.

3.3.1.2 Sanger sequencing

With the arrival of DNA sequencing bacteria could be typed to a whole new level. New species were and are still being discovered today. Up until a few years ago the favoured method for sequencing was Sanger sequencing, developed by Fred Sanger in the late 1970s (112) and further developed in the 1980s (113, 114, 115) to produce longer read lengths of 450 to 850bp. It is otherwise known as the chain-termination method. This technique uses dideoxynucleotide triphosphates (ddNTPs) as the DNA chain terminators. These lack the 3′-OH group necessary for forming the phosophodiester bond between two nucleotides and hence terminates the DNA strand extension. Each reaction contains a single-stranded DNA template, a DNA polymerase, a DNA primer, deoxynucleotides (dATP, dCTP, dGTP and dTTP) and ddNTPs. Four separate reactions are set up each containing the aforementioned products however, only one ddNTP is included. Therefore one reaction contains ddATP, one ddCTP and so on. The newly formed fragments are heat denatured and

separated by size using gel electrophoresis, with each reaction run in a different lane according to the ddNTP. The gel is visualised using autoradiography or UV light. The sequence can be read directly off the gel. This method was effective but time-consuming, it is mainly for this reason that the technique has been developed further. Hood *et al.* (116) developed fluorescently labelled ddNTPs and primers which meant that only one reaction was needed instead of four, a technique now known as dye-terminator sequencing. Each ddNTP is labelled with a different fluorescent dye with a specific wavelength of fluorescence and emission. It is this technique which is exploited by high-throughput DNA sequence analysers. The main reason Sanger sequencing is no longer favoured is the cost of analysis. With only 1000 bases being read per run the cost of sequencing a whole genome is enormous, for example, the sequencing of the human genome cost in the region of 300 million dollars.

3.3.2 High-throughput sequencing

High-throughput sequencing is a technique for sequencing a large number of DNA sequences at the same time at a lower cost. This technique has enabled DNA to be sequenced to a depth which was not previously possible. In 2004, the National Institute of Health set the challenge of the " $1000 human genome " to be achieved by 2015. This has pushed the scientific world to improve high-throughput sequencing technologies with break-throughs still being made (117). A number of different companies have all produced high-throughput sequencing machines which work in slightly different ways, the two most developed at the time of choosing which technique to use, are described below.

High-throughput sequencing has only been developed over the last few years. It is therefore still a relatively new technique which is subject to certain limitations. These limitations include; short fragments being difficult to assemble into whole genomes and PCR amplification being required before sequencing; a process which can introduce errors decreasing the accuracy of sequencing. With millions of sequences being produced with every run, powerful computers are required to analyse the data along with computer programs capable of processing large volumes of data. Statistical data analysis software needs to be developed and perfected along with software specific to every application of the technique to ensure valid conclusions can be drawn. With the popularity of this technique rising every day the cost of analysis will eventually drop to the point where it could be used for routine analysis (118). Since the start of this project the cost of analysis has nearly halved, demonstrating that as the technique advances and becomes more popular the cost of analysis decreases. One current technique which can be applied to high-throughput

sequencing to make it cheaper is barcoding. Barcodes can be added to each sample enabling them to be analysed in the same lane of one run. The number of reads per sample decreases as the number of barcoded samples increases, however for many applications one sample per run produces more sequences than necessary.

3.3.2.1 454 Sequencing

454 sequencing is based on the technique of pyrosequencing and around 400-700 megabases of DNA can be sequenced per twenty-three hour run, with each sequence read ranging between 500 to 1000 base pairs. 454 sequencing is a two step process: initially the DNA is fragmented and oligonucleotide adaptors are attached. These adaptors are used for purification, amplification and sequencing steps. Then each fragment is attached to a bead with each bead carrying a unique single-stranded DNA fragment (see Figure 3.3 a(i)). The beads are then amplified via emulsion PCR (emPCR), where each bead is emulsified with amplification reagents in an oil-water mixture, hence producing multiple copies of the same sequence on each bead. Subsequently each bead is captured in a picolitre-sized well on a fabricated substrate containing the sequencing enzymes (see Figure 3.3 a(ii)). The size of the wells in the picotiter plate only allow for one bead per well. Finally pyrosequencing is performed on all the beads in one plate. The four nucleotides; adenine, guanine, cytosine and thymine, are washed over the plate one at a time. If a nucleotide is incorporated then an inorganic phosphate (PPi) is released and converted to ATP. The ATP is then used by luciferase to generate light which is detected and indicates that the base has been incorporated. This process is repeated about 200 times producing read lengths of 500-1000bp. The main advantage of this technique is the read lengths produced. The longer the read length, the easier it is to assemble the DNA fragments. However, there are also disadvantages to this technique; prone to errors when estimating fragment length, little ability to detect insertions or deletions of single base pairs and unreliable for sequencing homopolymers. This unreliability is caused by a lack of reversible terminator nucleotides enabling more than one base to be incorporated at each cycle (119). It is also possible that some of the errors found with 454 sequencing could lead to perceived changes in genetic diversity. When analysing 16S rRNA, about 1000 sequences per sample are required to get a good depth of coverage. At this level the relative frequencies of species at 1% abundance can be fairly accurately inferred, however, many rare species will be missed (78).

3.3.2.2 Illumina

Illumina sequencing technology is based on the technique of sequencing by synthesis and about 20-500 Gb of paired-end data per run is produced with sequence reads now up to 300 base pairs. The initial step involves fragmenting the DNA and attaching adaptors. The fragmented DNA is then attached to a planar, optically transparent solid surface and amplified using bridge PCR and anchored primers (see Figure 3.3b(i)). This solid-phase amplification undergoes multiple cycles producing an ultra-high density sequencing flow cell which contains clusters, with each one consisting of about 1000 copies of the same single-stranded DNA template. Subsequently, sequencing is performed using a combination of primers, DNA polymerase and four fluorophore-labelled, reversibly terminating nucleotides (see Figure 3.3b(ii)). When a nucleotide is incorporated an image is taken and the identity of the base noted. The fluorophores and terminators are removed and the incorporation, detection and identification steps are repeated. The whole process is repeated again from amplification to sequence the reverse strand. This technique can produce many more sequence reads than 454 sequencing, however the reads are much shorter. It is thought that due to the increased number of sequences Illumina is a better technique for analysing genetic diversity. The possibility of paired-end sequencing enables analysis of sequences from both ends producing overlapping reads. Hence, combating the problem of low-quality ends produced by single reads as the ends can be overlapped to form high-quality consensus sequences. The purpose of this study is to compare the genetic diversity of samples between people, therefore, Illumina is the chosen technique as it is more accurate at measuring genetic diversity (120).

3.3.2.3 Newer techniques

Since the start of this project many other companies have produced high-throughput sequencing machines which are competing with both 454 and Illumina. The three most prevalent are: Ion Torrent, Pacific Biosciences and Oxford Nanopore. Ion Torrent uses semiconductor sequencing technology, which is based on hydrogen ion detection (121). This technology has been evolving quite fast and can now offer a variety of sequencing options including: single or paired-end reads, read length from 35 to 400 bp and runs as short as 90 minutes (122). However, it suffers from similar problems as 454 in terms of sequencing errors concerning homopolymers and therefore, it's error rate is higher than Illumina. The sample preparation is also much more labour intensive than for Illumina. The cost of sequencing is about $1000/Gb which is much higher than Illumina ($41/Gb for the HiSeq 2000 and $502/Gb for the MiSeq) (123). Pacific Biosciences have produced a

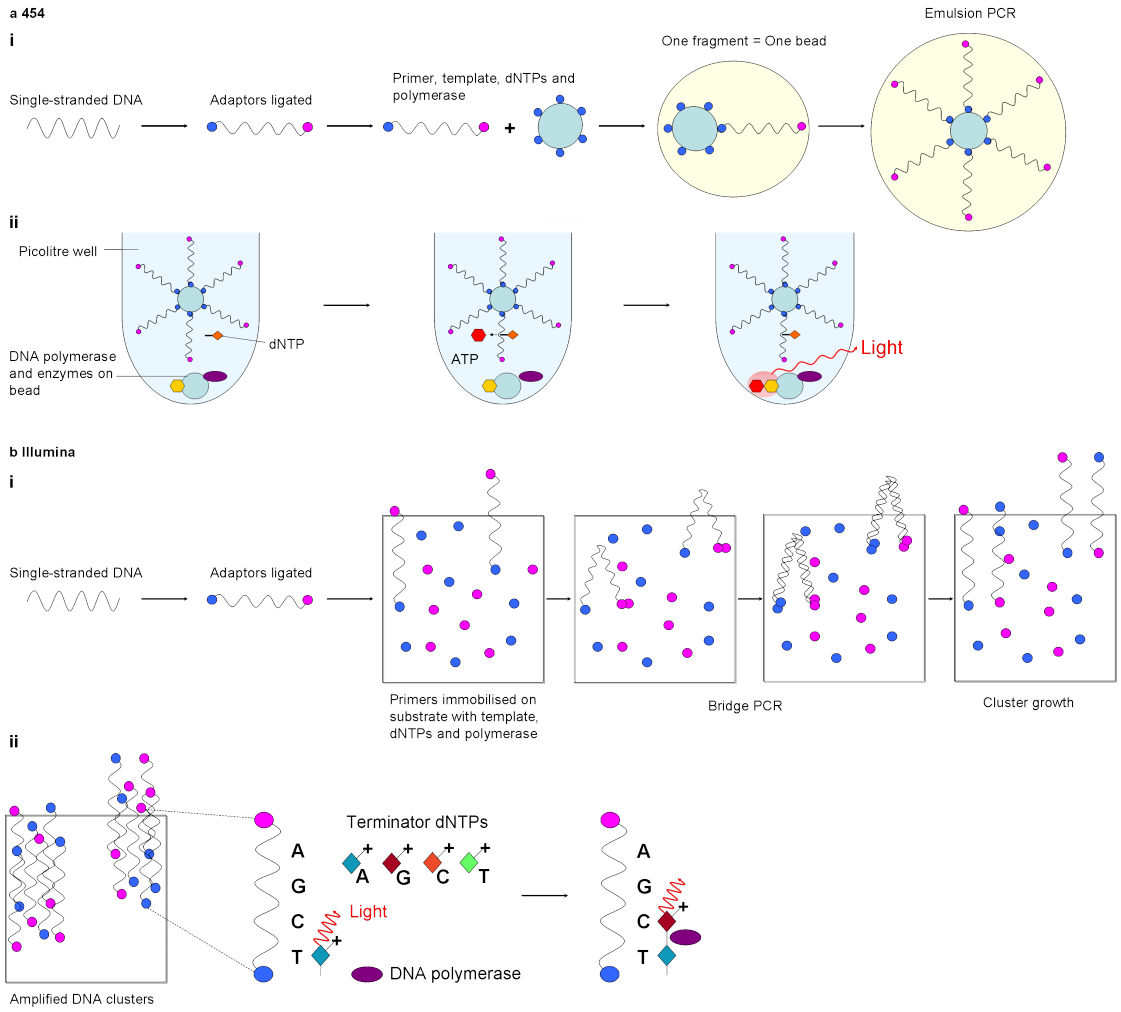

Figure 3.3: Overview of 454 and Illumina sequencing techonologies. (**a**) 454 pyrosequencing. Oligonucleotide adaptors are attached to fragmented DNA, subsequently the DNA is immobilised on a bead and amplified in a water-oil emulsion. For sequencing, each bead is placed in a picolitre well with dNTPs, DNA polymerase and the enzymes needed for the chemiluminescent reaction. Each base is washed over the wells one at a time, if the base is incorporated then pyrophosphate is released which then helps produce ATP and yields enough energy to generate light. The light emitted is recorded enabling the sequence to be deduced. (**b**) Illumina sequencing. Oligonucleotide adaptors are ligated to fragmented DNA and used to attach the fragments to a prepared substrate densely populated by complementary primers. Bridge amplification is performed to amplify the immobilised template using immediately adjacent primers to form clusters. DNA polymerase and terminator nucleotides are subsequently added to create the complementary DNA strand. An image is taken at the end of each cycle to enable the identification of each added base. [Adapted from Figure 1 (119)]

single molecule real-time sequencer, enabling the observation of DNA synthesis in real-time (124). This technology can sequence the longest read lengths of any platform at around 8.5kb however, the number of reads sequenced in one run is limited to around 50,000. The run time is also quite quick ranging from 30 minutes to 3 hours. This technique is good for characterising plasmids, viruses, mitochondrial DNA and microbial pathogens (125) but not so good for bacterial community analysis. This technology is still being developed, has high error rates and the highest cost of all platforms at about $2000/Gb. Oxford Nanopore technology is based on nanopore sensing and provides single molecule real-time sequencing (126), much like Pacific Biosciences. Unlike Pacific Biosciences Oxford Nanopore have developed a way to parallelise the sequencing making it possible to analyse many sequences/samples at the same time. This technology is even newer than that of Pacific Biosciences and is therefore still being developed and not much data is available to compare it with other more established techniques. One advantage of single molecule sequencing is that amplification is not required, hence removing a major source of bias and lower DNA input concentrations are possible. These techniques all show promise, however when I was choosing which sequencing technique to use these were not well established enough. Furthermore, Illumina still stands up to these new technologies producing high quality reads at a great depth.

3.3.2.4 Sample preparation

Before the amplified samples could be sequenced they needed to be prepared. First, the samples were quantified using a Qubit fluorometer (Life Technologies) and then run on a Fragment Analyzer (Advanced Analytical) to check purity and concentration. To limit the number of barcodes used all amplified targets from the same sample were pooled together and the pooled sample barcoded. To pool samples, equal molar amounts of each sample are necessary, in this case approximately ten picomoles of each was used. The samples were then purified using Agencourt AMPure XP PCR purification (Beckman Coulter) to remove the PCR reagents as they interfere with the sequencing library preparation. The purified products were then separated on an agarose gel (E-gel®SizeSelect^{TM}) and the band corresponding to the target size (120bp) excised. Finally, the sequencing libraries were prepared using the TruSeq DNA sample preparation kit (Illumina), following standard protocol (127) and then re-run on the bioanalyser to check concentration and purity. Additionally the libraries were checked to see if they were balanced, i.e. an even number of each base at each point in the sequence. If unbalanced, i.e. one base dominates, it can be difficult to read the base in that position as the flow cell will look like a smear of a single

colour. If the libraries are unbalanced there are two ways to get around it: 1) load half as much to dilute the effect and 2) mix the library with a known well-balanced library to give the necessary contrast. In this case the libraries were balanced. Finally the libraries were loaded onto the HiSeq 2000 and sequenced using the paired-end reads 100 cycles run type.

3.4 Data Processing

After sequencing the data produced is in the form of sequences of bases. For Illumina sequencing the CASAVA 1.82 pipeline is used for base calling and links to FastQC for quality control (see Figure 3.4 for overview of initial stages of sequence processing). Casava also separates the samples by barcode producing, in this case, four different samples with each sample divided between up to twenty eight files, as paired end sequencing was used there are up to fourteen files for read one and up to fourteen for read two.

3.4.1 FastQC

Before any data analysis could take place the sequences were quality controlled using FastQC a program which gives an overview of the sequence quality from all aspects. This system is based on what FastQC classes as a 'normal' sample which is a random and diverse sample. If a library is known to be bias then that has to be taken into account when using the quality assessment of FastQC. The following measures are given: basic statistics, sequence diversity, sequence identification, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence duplication levels and overrepresented sequences. For each different measure there is a set threshold for issuing a warning or a fail. Running FastQC in Casava mode removes sequences flagged for filtering and therefore the total sequence count does not include the filtered sequences (128).

3.4.2 Flash

The second step was to overlap all the paired reads to produce the consensus sequences for each sample. FLASH (Fast Length Adjustment of SHort reads) was the chosen software to correctly overlap the reads (130), specifically version 1.0.2 was used. FLASH 1.0.2 was run locally using the following command line and parameters (GG4_R1 and GG4_R2 are the two individual reads that are to be paired together):

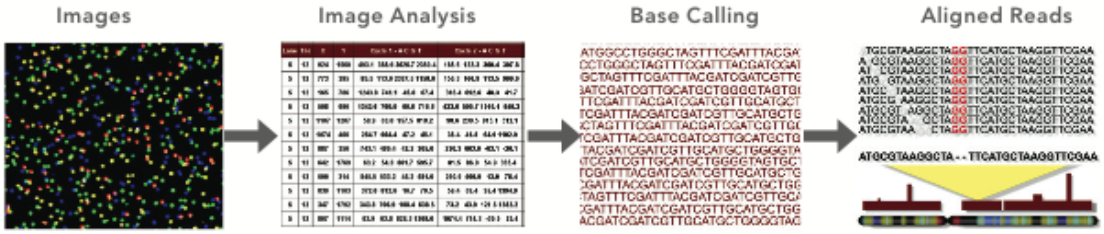

Figure 3.4: Schematic representation of the initial stages of sequence processing from the Illumina platform (129) - (Reproduced with permission from Illumina)

```
../FLASH_v1.0.2/flash GG4_R1_001.fastq GG4_R2_001.fastq -m 20 -M 110
 -f 120 -d GG4_001_flash
```

-m minimum overlap (minimum required overlap length between two reads to provide a confident overlap, default 10bp), 20bp was used.

-M maximum overlap (maximum overlap length expected in approximately 90% of read pairs, default 70bp) 110bp was used.

-f average fragment length (default 180bp) 120bp was used.

-d defines output directory

3.4.3 Barcode splitter

After the reads had been paired the next step was to split each sample into the individual targets; 16S rRNA, *rpoB*1 and *rpoB*2. To carry out this task barcode splitter was used. Barcode splitter comes from the FASTX-tool kit, a set of tools for pre-processing FASTA/Q sequence reads (131). Before barcode splitter was employed all the paired reads from one individual were concatenated to produce one file containing all the consensus sequences. Due to each primer having a different length each target was split separately. Barcode splitter was run locally using the following command line and parameters:

```
cat SL4_all_out.extendedfrags.fastq | /Users/admin/Desktop
/fastx_toolkit-0.0.13.2/scripts/fastx_barcode_splitter.pl
--bcfile strep.txt --bol --exact --prefix SL3_004_
```

–**bcfile** barcode file

–**bol** try to match barcode at the beginning of sequences

–**exact** no mismatches allowed

–**prefix** prefix for output file

3.4.4 Clustering

The next step after the samples had been split into the three different targets was to cluster the sequences. There are two types of methods which can be used; supervised and unsupervised. Supervised techniques involve directly comparing the sequences with a database of known sequences mainly employing BLAST and constructing a phylogenetic tree, this is otherwise known as a homology-based approach (132). The main problem with this type of method is it will only cluster sequences which are in the database whereas all others will be classified as unassigned. This makes it difficult to characterise novel sequences. The classification of each sequence is dependent on the completeness of the database. Often these databases are only classified to the genus level or higher, not species level. For this study the more taxonomic assignments at species level the greater detail that should give for characterising the microbiome. Obviously at some stage in the analysis process taxonomic assignment is important to be able to analyse which taxa are present, however it is best to do this as far downstream as possible to avoid any bias in the data. The second problem is the choice of distance measure used to create the phylogenetic tree. Many distance measures exist which all use different calculations, so each one will represent the data in a slightly different way.

Unsupervised techniques cluster the sequences into Operational Taxonomic Units (OTUs) based on their similarity, with no prior information. An OTU is defined as a group of organisms with target sequences that show a certain level of identity (133). In the case of 16S ribosomal RNA at least 97% identity is required for the OTU to be considered equivalent to the working definition of species. For *rpoB* the value is lower at about 95% (134). Many programmes exist to cluster sequences into OTUs with three different methods available; hierarchical clustering, heuristic clustering and model-based clustering, all of which have their advantages and disadvantages (135). For hierarchical clustering initially a distance matrix is calculated using the difference between each pair of sequences then a specific similarity level is set and standard hierarchical clustering is used to define the OTUs. This method is intrinsically computationally complex and therefore is not well suited to large datasets. Heuristic clustering functions by first defining a similarity threshold then it takes a sequence and uses that as a seed for the first cluster then each sequence is analysed sequentially. If the distance between the representative sequences of the existing clusters and the query sequence falls within the pre-defined threshold then the query sequence is added to the corresponding cluster, if not then it becomes the seed sequence for a new cluster. Both of these methods use strict threshold values which has caused debate as some studies have shown that it can be difficult using a strict threshold

as there is overlap in the maximum intra-taxon distance between taxonomic levels (136). Model-based clustering was proposed to overcome this problem. This method uses an unsupervised probabilistic Bayesian clustering algorithm with a soft threshold avoiding setting a strict threshold (137).

For this project CD-HIT a heuristic clustering algorithm was chosen due to its capacity to deal with very large datasets quickly (138). The speed of analysis comes from the use of short word filtering which takes short substrings (words) and compares them between sequences, the more the sequences have in common the more similar those sequences are. This calculates the similarity between sequences without carrying out a sequence alignment, hence making it faster. This does not provide an exact sequence comparison however it is effective at estimating whether two sequences are below a certain threshold. Before this program could be used the fasta header for every sequence was changed to contain the sample name and make it smaller as in the CD-HIT output file each sequence header can only contain a restricted number of characters. Specifically CD-HIT-EST 4.5.4, the algorithm used to cluster nucleotide sequences (139), was run locally using the following command line and parameters:

```
/Users/admin/Desktop/cd-hit-v4.5.4-2011-03-07/cd-hit-est -i all_16S2.fa
 -o 16S2_97 -c 0.97 -n 9 -M 9000
```

-i input file (fasta file containing all sequences from one target from one sample)

-o output filename

-c clustering threshold (here it is 97% for 16S rRNA and 95% for *rpoB*)

-n word length

-M maximum available memory (Mbyte)

CD-HIT-EST ouputs two files, the first is a fasta file containing representative sequences for each cluster and the second a text file containing a list of clusters.

3.4.5 Data filtering

The final step of data processing involves transforming the list of clusters into a table containing the abundance of each taxa per sample. Firstly, the CD-HIT-EST output cluster file was filtered along with the fasta file of representative sequences. Cluster filtering is necessary due to errors introduced during the amplification and sequencing processes.

An error in amplification or sequencing can lead to a cluster appearing as novel when in reality it is an error. These errors mainly affect low abundant taxa. It is possible that some singletons (taxa represented by one sequence) are real however it is impossible to differentiate these from errors. A threshold is set to remove any clusters with less than the specified number of sequences from the dataset. Choosing this threshold has caused much debate, with most groups deciding upon three, i.e. they keep all clusters containing three or more sequences (84). However, this threshold is data dependent and can vary for different studies. For this study any clusters containing less than twenty sequences were filtered out. This threshold was chosen to ensure that as many errors as possible were removed whilst keeping enough data. For a forensic application the removal of error is of utmost importance so it is best to be over cautious. The choice of threshold was calculated by performing the analysis through to hierarchical clustering using the same clustering parameters (shown above) and comparing the relative distances and separation of samples. A threshold of 10 was also tested to see the effect of using a lower threshold, however the results were not that different. Therefore, I chose to use a threshold of 20 to help ensure removal of errors. To filter the cluster file the python script filter_cluster.py was used (see appendix B (section 8.5.1)). This script produces a filtered fasta file and a summary table containing the abundances of each sample per cluster, however at this point the taxa have not yet been assigned to each cluster.

3.4.5.1 BLAST

To assign the taxonomy to each cluster BLAST was used. The filtered fasta file containing the representative sequences for each cluster containing twenty sequences or more was inputted into BLAST and compared against the entire nucleotide database. BLAST functions by finding regions of local similarity between query sequences and a sequence database and calculates the statistical significance of the match (109). The basic output of BLAST contains all the possible matches however for this study only the best match is required. To produce this BLAST has a Best-Hit algorithm which filters out the best hit using the expect value (E-value) and bit score. The bit score indicates whether the alignment is good or not, the higher the score the better the alignment. The E-value indicates how statistically significant an alignment is, the lower the E-value the more significant the hit is. For example, an E-value of 0.05 means that the similarity has a 0.05 probability of occurring by chance alone. In essence, the E-value is considered as a measure of the random background noise. It takes into account the size of the database so calculates the chance of a sequence occurring by chance in that particular database. For short sequences

the E-value will naturally be higher as the E-value calculation takes the sequence length into account and small sequences have a higher probability of occurring in the database by chance. The Best-Hit algorithm uses a large bit score (best_hit_overhang) and a low E-value (best_hit_score_edge) to filter out the best alignment for each cluster(140).

BLAST nucleotide (blastn) was run locally using the following command line:

```
/usr/local/ncbi/blast/bin/blastn -query 16S_filter.fa -db nt_db
-evalue 1e-10 -best_hit_score_edge 0.05 -best_hit_overhang 0.25
-num_descriptions 1 -num_alignments 1 -out 16S_97_blast_best_hit
```

By setting both num_descriptions and num_alignments to one the output file contained only the top hit for each sequence.

3.4.6 Table production

The next step involves combining the blast best hit with the filtered fasta file to produce a table containing the abundance of each cluster with the taxon assignment for each sample. To achieve this the python script sort_cluster.py was used (see appendix B (section 8.5.2)). As expected some of the clusters represent the same taxon so to get a better idea of how many of each taxon are present a second table was produced summing together any clusters with an identical taxon assignment. The python script adjust_table.py was used to produce this table (see appendix B (section8.5.3)). There is a risk with combining the abundance data in that the taxon assignment could be wrong as BLAST outputs the best hit however this might not be the correct hit. As BLAST uses a database it relies on the annotation in the database being correct. For this study, the NCBI nucleotide database was used as it is the main database containing all annotated nucleotide sequences and as three targets were used, to keep as many parameters as possible the same, a generic database was necessary. In order to more easily manage the data the merging of the abundances was necessary. To minimise any error the abundances were only combined if the taxon names matched exactly, meaning that different strains of the same species were classed as being different.

3.5 Data interpretation

With the data in the form of an abundance table for each species in each sample, the next step was to work out which taxa are the best at separating out different individuals and which ones are common between all.

3.5.1 Normalisation

In order to compare the datasets from both experiments (t_1,t_2 and t_3,t_4) the abundance counts were normalised to avoid bias effects due to differences in sequencing of the samples. Normalisation puts all of the datasets on the same scale enabling accurate comparison. This data consists of counts of sequences and this needs to be taken into consideration when choosing which method of normalisation to use. This data cannot be normalised by total read count as a few highly abundant OTUs may have a strong influence on the total read count and therefore the ratio of total read counts will not be a good estimate for the ratio of expected counts.

For this study DESeq, a package designed to estimate variance-mean dependence in count data from high throughput sequencing experiments was used. Specifically DESeq was designed for use with RNA-seq differential gene expression assays (141) however, the principal behind its use is the same. The type of data produced i.e. containing a few high abundant species is similar and the constraints caused by high throughput sequencing are the same, it is only the downstream analysis which differs. Normalisation is only one of the applications offered by DESeq. For the normalisation process instead of using total read count the median of the ratios of observed counts is used.

DESeq uses statistical testing to see whether, for a given OTU, the difference in read counts can be considered as significant under a list of assumptions i.e. is it greater than what would be observed through natural random variation (141). Previous studies have modelled count data using a multinomial distribution (approximated by a Poisson distribution (142, 143)). However, this would only work if the reads were independently sampled from a population containing a fixed fraction of OTUs. The assumption of the Poisson distribution is too restrictive as it predicts smaller variation than what is observed in the data. To address this problem, a negative binomial (NB) distribution could be used (141) which has parameters that are defined by the mean (μ) and variance (σ^2). However, often the number of replicates is too small to accurately estimate both mean and variance. Yet if it is assumed that the mean and variance are related by equation 3.1 with α being a single proportionality constant that remains the same throughout the experiment and can be estimated from the data, then only one parameter needs to be estimated for each OTU. Hence, this method can be applied to experiments with small numbers of replicates. This technique only works with datasets containing a few very abundant OTUs and the rest being a lot less abundant otherwise could normalise away any variance.

$$\sigma^2 = \mu + \alpha\mu^2 \tag{3.1}$$

In order to normalise the two experiments together, the two abundance tables were combined using the python script concatenate.py (see appendix B (section 8.5.4)). This script combines both tables and inputs zero abundance for OTUs not found in the other experiment. A debug file is also outputted containing the taxon name from both input files for non-exact matches so they can be verified.

To use DESeq the DESeq library was loaded into R and the following steps taken (144):

1. the combined abundance table was uploaded in csv format.
2. conditions vector created - this informs the program which samples belong together
3. newCountDataSet created
4. size factors estimated - size factors are used to render counts from different samples, which may have been sequenced to different depths, comparable
5. normalised counts visualised

3.5.2 Data transformation

After data normalisation the next step was to transform the data to enhance the potential differences in the data. As mentionned above the data contains a few highly abundant taxa with the rest being a lot less abundant. To minimise the effect of the highly abundant taxa the data was transformed by taking the $\log_{10}(x+1)$ of each count (x). One was added to the log transformation to avoid problems with zero counts as $\log_{10}(0)$ is undefined and therefore cannot be calculated, whereas $\log_{10}(1)$ is zero. This transformation brings the data to a more manageable size by reducing the range. To analyse the experiments individually normalisation is not required, only log transformation. After either both normalisation and log transformation or just log transformation the data is in the form of a table as presented above in section 3.4.6, see supplementary files for data tables (https://independent.academia.edu/SarahLeake/Papers).

3.5.3 Significant taxa

In order to work out which taxa were significantly different between individuals some descriptive and inferential statistics were carried out. The first was to calculate the mean abundance between the samples from one individual, first for each experiment separately then for both experiments combined. To be able to support whether there was a difference between the means both Frequentist and Bayesian methods were applied. Frequentist F-

and t-tests were performed as suggested by large biological literature (145). Specifically, an unpaired, 2-tailed t-test was performed using excel. However, this methodology (notably through the use of p-values) often causes pitfalls of intuition (in fact, usually scientists believe that if the p-value is greater than a given threshold, say 0.05, then there is no significant difference between, say, the means of the two populations of interest).

A frequentist p-value answers the following question 'how frequently would I observe a result at least as extreme as the one obtained if the null hypothesis were true?' This represents a statement about the plausibility of the data given the hypothesis. It is out of the scope of this Ph.D research to debate on pros and cons of statistical methods. Anyway, let us note that one of the appealing features of Bayesian methods is that they allow one to overcome the definitional difficulties that arise with frequentist hypothesis testing where users may tend to view the p-value as the probability of the null hypothesis. As previously expressed, when $p = 0.05$, it is tempting to state that there is only a 5% probability that the null hypothesis is true. The p-value cannot, however, measure the probability of the truth of the null hypothesis because its calculation assumes the null hypothesis is true. Bayesian analysis allows the definition of a (prior) probability (i.e., a probability that is evaluated, usually subjectively, prior to observation of the data) for each hypothesis (null and alternative), that is an expression of the personal degree of uncertainty about a hypothesis truthfulness, on the basis of which a posterior distribution (i.e., a probability that is evaluated posterior to observation of the data) for the hypotheses can be inferred.

A Bayesian approach allows one to calculate the probability that the two populations means are (are not) equal and so express how confident one can be on the hypothesis of interest (say, that the two populations have different means).

An ingredient of the Bayesian model, the so-called Bayes factor (BF for short), is used in this research data treatment. The BF value supports one or the other of the hypotheses of interest (same/different means). A value greater than 1 supports the first hypothesis, a value less than 1 supports the alternative hypothesis. A value equal to 1 does not allow one to discriminate between the hypotheses. Note that larger the value, greater is its support of the hypothesis.

So, from a Bayesian point of view, by considering two groups of observations, it is of interest to test the equality of the means. More specifically, a scientist may intend to test the null hypothesis, say the difference between the two means equals 0 versus the alternative hypothesis that the difference does not equal 0. So, a BF greater than 1 supports the main hypothesis (there is no difference between the two means) and a value

below 1 supports the alternative hypothesis that there is a difference between the two means[1].

The above calculations were performed on a species by species basis meaning that only abundances from the same species were compared between each other. The p-values and BF values for all species comparisons are found in supplementary spreadsheet ' data ' (https://independent.academia.edu/SarahLeake/Papers).

3.5.3.1 Hierarchical clustering

The data was ordered based on the p-values and BF values with the smallest p-value/BF value first, representing the most significant taxon. To visualise how the most significant taxa separated the samples hierarchical clustering was performed. This method is recommended when comparing a small number of samples, which is the case here. Hierarchical clustering shows how the samples group together not precisely which taxa are best at separating the groups. To perform this technique the hclust algorithm in R was used (147). Firstly a distance matrix was calculated using the Euclidean distance. hclust provides a number of different clustering methods described below:

complete linkage	the distance between the clusters is determined by the objects which are furthest apart, this finds similar clusters.
single linkage	uses the 'friends of friends' strategy often producing one big cluster with a few small cluster
average linkage	uses the average distance between the members of the clusters and is half-way between complete and single linkage.
Ward	takes into account the number of members in each cluster, producing clusters similar to complete linkage.

The hierarchical clustering was visualised in the form of a dendrogram produced using the as.dendrogram function in R (148).

3.6 Further analysis

Once the basic analysis of which taxa were significant and whether it was possible to separate samples from different individuals using these taxa, further analysis was required

[1]Details of the calculations are presented in (146), chapter 6.

to identify the minimum number of sequences per target required to achieve adequate separation and which clustering threshold was best.

3.6.1 Minimum number of sequences

In order to work out the minimum number of sequences required to separate samples from different individuals the data was sub-sampled at different levels. This was performed on the sequences from the first experiment only and on each target separately. The combined sequences from each target where split up into each sample using the python script file_splitter.py (see appendix B). Subsequently, a second script rand.py (see appendix B (section 8.5.5)) was used to randomly sample sequences from each file. The sub-sampled files were then concatenated and the analysis process followed as described above (3.4.4 onwards).

3.6.2 Clustering threshold

As above only the first experiment was used to test different clustering thresholds. As described in section 3.4.4 CD-HIT-EST was used to cluster the samples into OTUs. Exactly the same process was followed here except the clustering threshold was changed. For 16S rRNA 80%, 90%, 97%, 98% and 100% similarity were tested and for both *rpoB* targets 80%, 90%, 95%, 96% and 100% were tested. The rest of the protocol was followed as described above (3.4.5 onwards).

4

Results - Primer design and virtual simulation

This chapter presents the results of the primer optimisation and the virtual simulation of the optimised primers. This stage is important to ensure the primers used are efficient and amplify the desired bacteria.

4.1 Primer Design

The primers were designed by aligning the 16S rRNA, 23S rRNA or *rpoB* gene sequences from target species, see Table 3.1, and finding the best overall region in terms of both conserved regions for primer binding and hypervariable regions for maximum taxa differentiation within the size limit for Illumina sequencing. Once the primers had been designed using generic species they were then checked against species known to be found in saliva (see Table 4.1). For 16S rRNA and 23S rRNA a mix of species was chosen to cover the principle expected phyla; Firmicutes, Proteobacteria, Actinobacteria and Bacteroidetes. Figure 4.1 shows that for 16S rRNA pair one the forward primer falls in a highly conserved region with only two mismatches for *Veillonella parvula*. The reverse primer also falls in a conserved region however there are a few more mismatches; three for *Mycoplasma buccale* and one each for *Porphyromonas gingivalis* and *Prevotella oralis*. No sequence is visible for *N. mucosa* as that sequence could not be aligned with the others. For the second pair a similar pattern is seen (see Figure 4.2), with the forward primer having only one mismatch for *M. buccale* and two for *P.gingivalis* and *P. oralis*. The reverse primer has one mismatch for both *L. salivarius* and *P. gingivalis*. There is no sequence visible for *N. mucosa* and *V. parvula* as neither could be aligned with the others. As 16S rRNA is

a gene essential for life and therefore found in all taxa it is evident that not all taxa will have the same sequences in the conserved regions due to natural genetic variation. This shows that even though the primers were designed using a more general set of species they can still be applied to a more specific set of species. The specific region amplified for 16S rRNA is the V5 region, this choice is corroborated by Lazarevic *et al.* (84) where very similar primers were designed.

The alignment for 23S rRNA with species known to be found in saliva does not yield good results (see Figure 4.3). Firstly, only four out of the ten chosen species could be aligned with the primers and of those four only two match exactly with the forward primer and three with the reverse primer. Even with this result these primers were still carried through to the initial stage of primer optimisation to see whether the target region could be amplified.

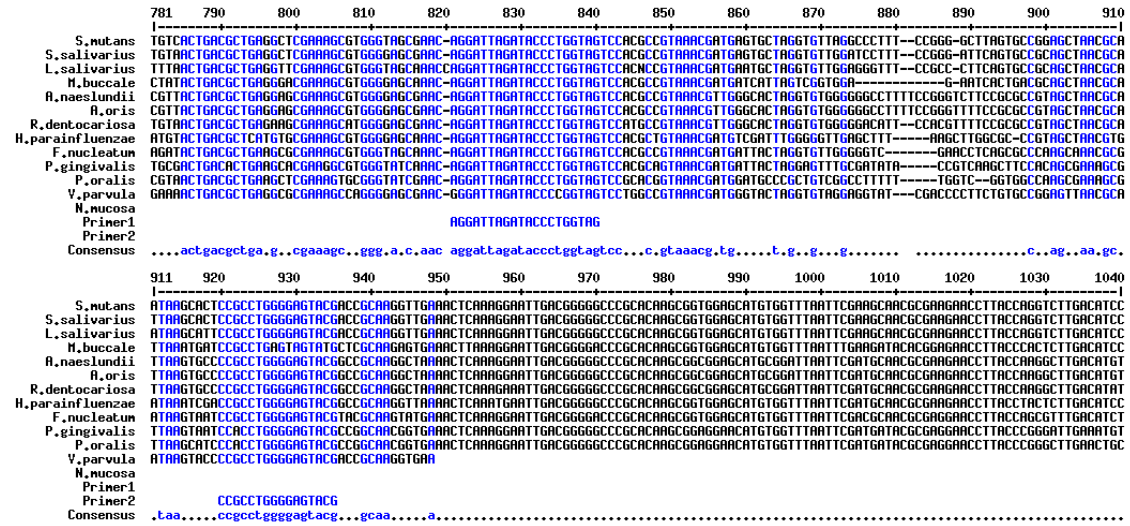

Figure 4.1: 16S rRNA first primer pair (783F-878R) alignment with species found in saliva - Bases in blue are identical and form the consensus sequence whereas bases in black differ from the consensus sequence. Primer1 corresponds to the forward primer and primer2 to the reverse primer. (Alignment performed using Multalin (106))

Figure 4.4 shows the alignment for *rpoB*1 pair one with species known to be found in saliva. For the forward primer the sequence matches exactly indicating that the chosen region is very conserved. For the reverse primer the sequence matches exactly except for *S. anginosus* which has two mismatches. As for 16S rRNA this shows that primers designed using more general species can be applied to more specific species. The alignment for the second pair of primers for *rpoB*1 is not as good. Figure 4.5 shows that the sequence for *S. pyogenes* is not as conserved as the others and does not really match either the forward or reverse primer. The forward primer matches the rest exactly except for *S. bovis* for which,

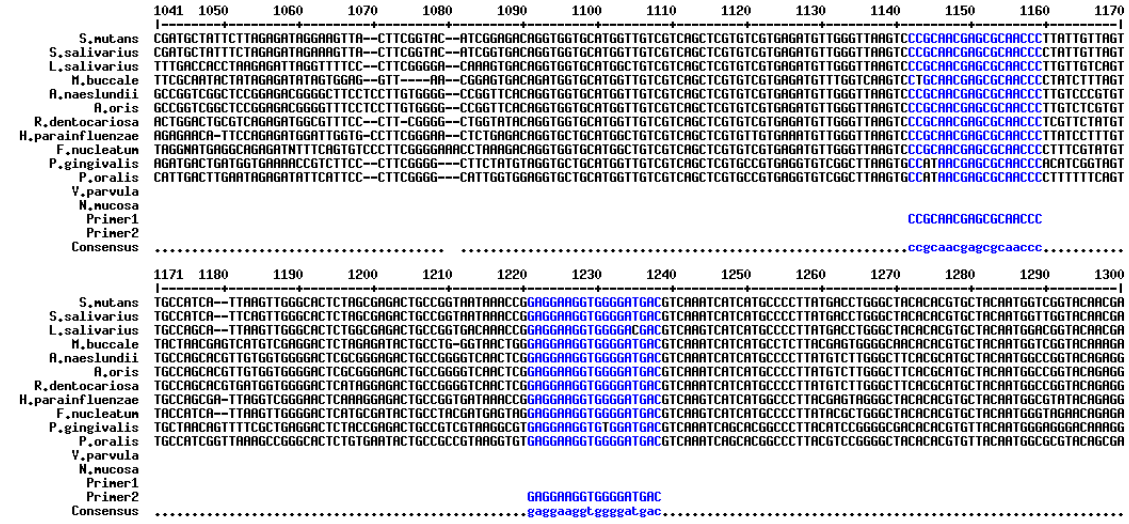

Figure 4.2: 16S rRNA second primer pair (1097F-1175R) alignment with species found in saliva - Bases in blue are identical and form the consensus sequence whereas bases in black differ from the consensus sequence. Primer1 corresponds to the forward primer and primer2 to the reverse primer. (Alignment performed using Multalin (106))

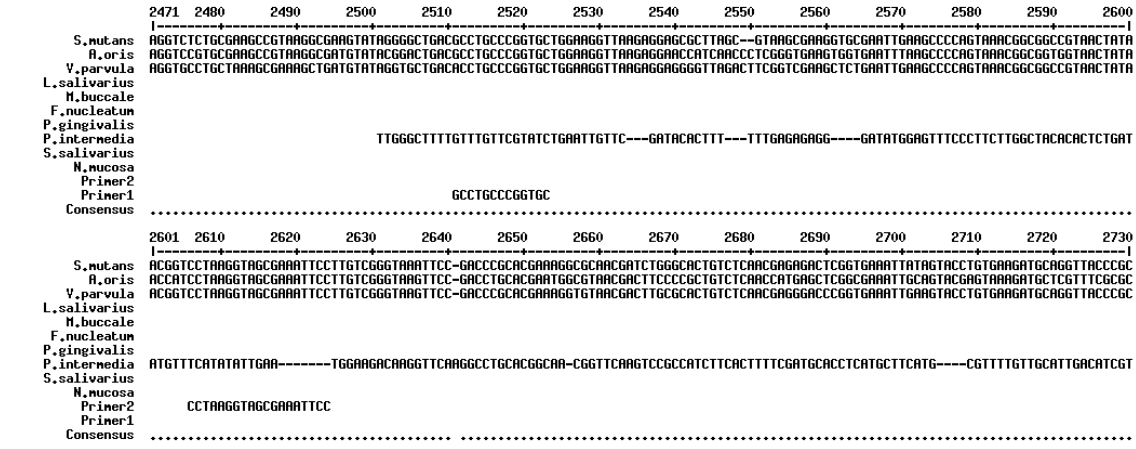

Figure 4.3: 23S rRNA alignment with species found in saliva - All bases are in black as there is no consensus sequence. Primer1 corresponds to the forward primer and primer2 to the reverse primer. (Alignment performed using Multalin (106))

it has two mismatches. The reverse primer matches both *S. bovis* and *S. oralis* exactly and *S. anginosus* and *S. mitis* with one mismatch. This indicates that the first pair of primers for *rpoB*1 would yield the best results. For *rpoB*2, both the forward and reverse primers do not match exactly with any of the saliva specific species. This indicates that this region is not as conserved however, enough bases are in common that the amplification should work, therefore this primer pair was kept and optimised. For both *rpoB* targets the V1 region was chosen.

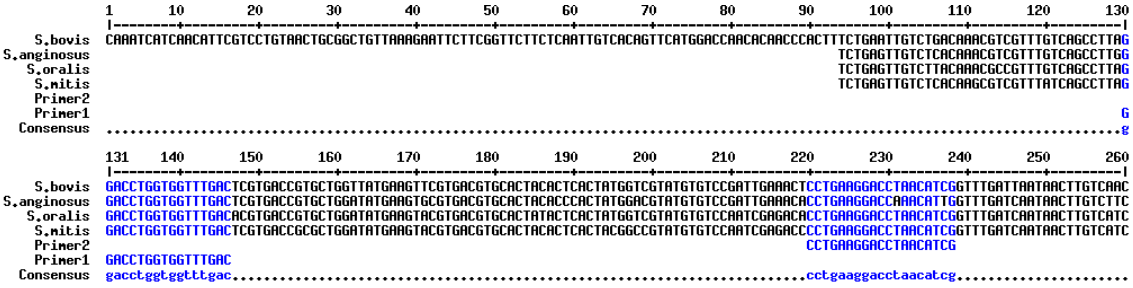

Figure 4.4: *rpoB*1 first primer pair (130F-220R) alignment with species found in saliva - The blue bases show where the primers bind and the black bases represent the rest of the sequences. Primer1 corresponds to the forward primer and primer2 to the reverse primer. (Alignment performed using Multalin (106))

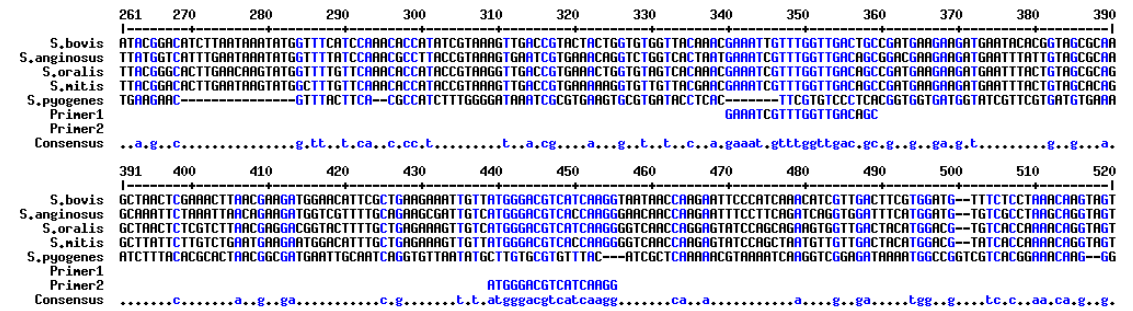

Figure 4.5: *rpoB*1 second primer pair (340F2-439R) alignment with species found in saliva - Bases in blue are identical and form the consensus sequence whereas bases in black differ from the consensus sequence. Primer1 corresponds to the forward primer and primer2 to the reverse primer. (Alignment performed using Multalin (106))

All primers were then tested *in silico* using EcoPCR (110).

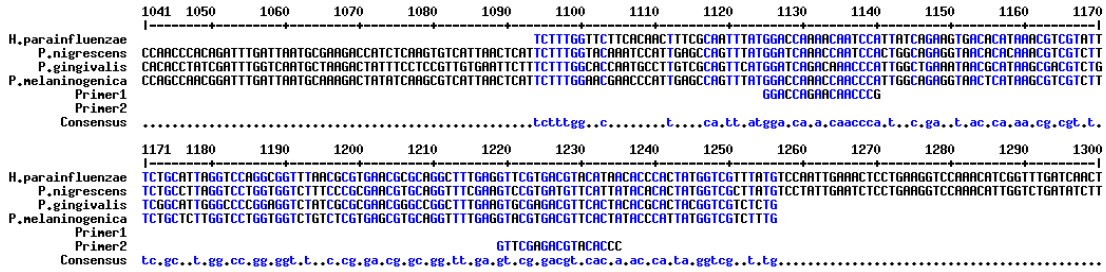

Figure 4.6: *rpoB*2 alignment with species found in saliva - Bases in blue are identical and form the consensus sequence whereas bases in black differ from the consensus sequence. Primer1 corresponds to the forward primer and primer2 to the reverse primer. (Alignment performed using Multalin (106))

16S/23S rRNA	***RpoB*1**	***RpoB*2**
Streptococcus mutans	*Streptococcus bovis*	*Haemophilus parainfluenzae*
Streptococcus salivarius	*Streptococcus anginosus*	*Porphyromonas gingivalis*
Actinomyces naeslundii	*Streptococcus oralis*	*Prevotella nigrescens*
Actinomyces oris	*Streptococcus mitis*	*Prevotella melaninogenica*
Lactobacillus salivarius		
Rothia dentocariosa		
Neisseria mucosa		
Veillonella parvula		
Fusobacterium nucleatum		
Haemophilus parainfluenzae		
Porphyromonas gingivalis		
Prevotella oralis		
Mycoplasma buccale		

Table 4.1: List of species used for primer design check by target

4.2 Virtual Simulation

For the first simulation all primers were compared against the human genome and the NCBI nucleotide (nt) database using BLAST (basic local alignment search tool) (109) (see Table 4.2 for results). All primers showed no 100% match with the human genome however, smaller portions of the primers did match. Due to the size of the human genome and the short length of the primer sequences it is unsurprising that parts of the primers match. It is possible that a small amount of human DNA is amplified with these primers, however it is unlikely this will impede the amplification of the target regions. This is very important due to the high amount of human DNA in the samples. Results presented in section 5.1 confirm this experimentally as sequences from *Homo* and *Pan* make up about 1% of total sequences. The comparison against the nt database shows that all primer

pairs are target specific. In addition, for both 16S rRNA and 23S rRNA the primers were compared against target specific databases using their probe match tools. For 16S rRNA the Ribosomal Database Project (RDP) (149) was used and for 23S rRNA the Silva comprehensive ribosomal RNA database (150) was used. For 16S rRNA the first primer pair (783F-878R) matches about 55% of the bacterial sequences with the second primer pair (1097F-1175R) matching only 44% of the sequences. For 23S rRNA even though the primers are specific they only match about 6% of the sequences in the database, indicating that this primer pair would not provide adequate results.

Target primers	**Primer**	**BLAST human**	**BLAST bacteria**	**Probe match bacteria**
16s rRNA				
783F-878R	783F	zero	specific	886955/1498677
	878R	zero	specific	704827/1498677
1097F-1175R	1097F	zero	specific	700023/1498677
	1175R	zero	specific	598964/1498677
23s rRNA				
1831F-1924R	1831F	zero	specific	11266/180344
	1924R	zero	specific	11285/180344
***RpoB*1**				
130F-220R	130F	zero	specific	n/a
	220R	zero	specific	n/a
340F2-439R	340F2	zero	specific	n/a
	439R	zero	specific	n/a
***RpoB*2**				
340F-434R	340F	zero	specific	n/a
	434R	zero	specific	n/a

Table 4.2: BLAST results for all primer pairs - BLAST human result refers to percent of sequences which match the primer sequence completely. BLAST bacteria refers to specificity of result to target region. For 16S rRNA probe match was performed against the Ribosomal Database Project (RDP) and for 23S rRNA the Silva comprehensive ribosomal RNA database was used. No probe match tool was available for *rpoB*. Primer name for 16S rRNA, 23S rRNA and *rpoB*2 corresponds to the *E.coli* positions and for *rpoB*1 to *S.bovis* positions.

For the second simulation the nt database was used as the sample to test all targets as it contains all nucleotide sequences, not a subsection. For 16S rRNA the Human Oral Microbiome Database (HOMD) (151) exists however this is limited to sequences which have been associated to the oral microbiome and not those which could still be associated. Therefore, the nucleotide database was used to include as many species as possible and not exclude any just because they have not already been classified as being associated to the oral microbiome. By using the nt database the same database was used for all primer pairs from all target regions, standardising the approach. As described in section 3.2.3

ecoPCR was used to simulate a PCR amplification in order to see which bacteria would theoretically be amplified by the primers.

For 16S_1, 16S_2 and 23S rRNA primer pairs, 23, 24 and 23 phyla respectively were amplified. All of the phyla are the same except for two which are different; Methanobacteria for 23S rRNA and Verrucomicrobia for 16S_2. However, these two phyla are not important as only one Methanobacteria species has been found thus far in the oral microbiome and no Verrucomicrobia species. Figure 4.7 shows the relative abundance of the top five phyla commonly found in saliva. The proportions of the top five phyla for 16S_1, 16S_2 and 23S rRNA are about the same, indicating that at the phylum level of classification only one of these three pairs of primers is required.

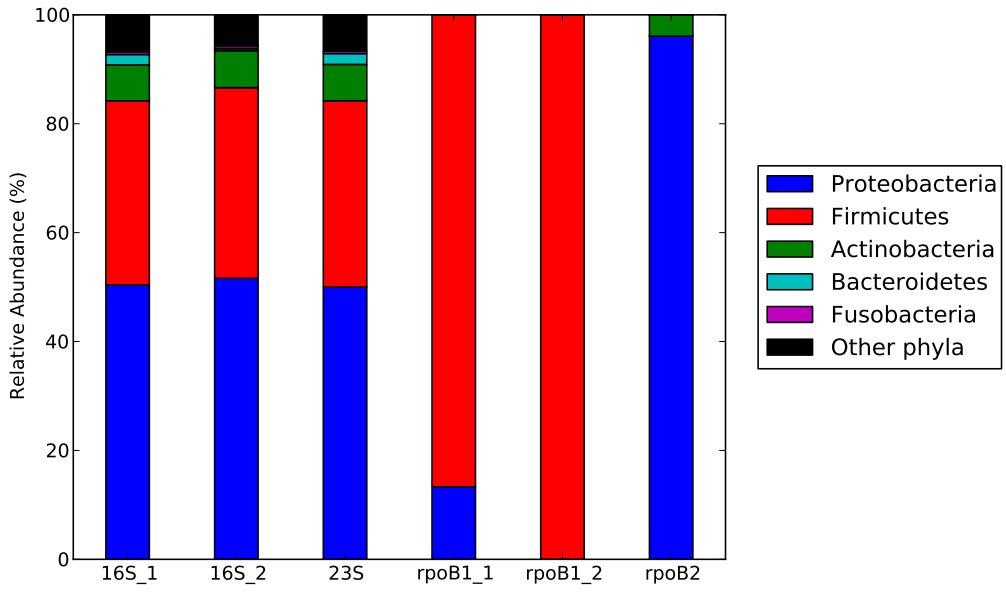

Figure 4.7: Relative abundance of the top five phyla commonly found in saliva by primer pair - 16S_1 corresponds to primer pair 783F-878R, 16S_2 to 1097F-1175R, *rpoB*1_1 to 130F-220R and *rpoB*1_2 to 340F2-439R.

All *rpoB* primer pairs amplify a less diverse range of phyla, however this is expected as the primers were designed to be more specific. Both *rpoB*1 primer pairs were designed to principally target *Streptococcus*, a Firmicutes. *rpoB*1_2 amplifies only Firmicutes whereas *rpoB*1_2 amplifies mainly Firmicutes with a small amount of Proteobacteria. *rpoB*1_1 is more useful as it amplifies a slightly more diverse range of bacteria compared to *rpoB*1_2. *rpoB*2 amplifies mainly Proteobacteria with a small proportion of Actinobacteria. This shows the benefit of targeting two different regions of *rpoB*. These results concern all of the possible bacteria amplified by the primers however, the proportions will differ when

amplifying saliva as not all bacteria are found in saliva. The principle two phyla Proteobacteria and Firmicutes are the same for both the whole database and saliva samples indicating that all the primers would be suitable for analysing the bacterial composition of saliva.

The genera found by virtual simulation were then compared with the genera from the HOMD to see how many could be found in the oral microbiome (see Table 4.3). For both 16S rRNA primer pairs and 23S rRNA the simulation produced between 350 and 400 genera of which around 70 were found in HOMD. Most of them are the same for the three primer pairs indicating that only one pair is required. When the number of genera found in common with HOMD is compared to the total number of genera in HOMD the ' % in common ' increases to around 50%. This shows that whilst the primers amplify many more genera than those found in the oral cavity, one set of primers alone can amplify half of the genera present in the oral cavity. Based on this percentage 16S_1 (783F-878R) would amplify the most oral specific genera. The number of genera amplified by the *rpoB* primers is a lot lower and therefore the ' % in common with all HOMD genera ' is very low, all under 5%. As stated above these primers were designed to target specific genera and hence will not amplify as many species as 16S or 23S rRNA. *RpoB*1_2 (340F2-439R) primer pair only amplified one genus, *Streptococcus*, however it was designed to do so. Whereas, *rpoB*1_1 (130F-220R) which was designed in the same way, amplified 11 genera of which 4 were found in HOMD. Of the 4 in common *Streptococcus* was the most abundant indicating that this primer pair is still specific just not as specific as *rpoB*1_2. Concerning *rpoB*2, the seven genera in common with HOMD are completely different to those found by *rpoB*1 primers, confirming the benefit of targeting different regions of *rpoB*. By combining one set of primers from a generic target (16S or 23S rRNA) with one or two sets of primers targeting a more specific gene (*rpoB*) the bacterial composition of saliva can be analysed to a level unattainable by one target alone.

Target primers	no. in common with HOMD/total simulated genera	% in common with HOMD/total simulated genera	% in common with HOMD/total HOMD genera
16s rRNA			
783F-878R	73/398	18.3	50.3
1097F-1175R	66/350	18.9	45.5
23s rRNA			
1831F-1924R	71/369	19.2	49
***RpoB*1**			
130F-220R	4/11	36.4	2.8
340F2-439R	1/1	100	0.7
***RpoB*2**			
340F-434R	7/21	33.3	4.8

Table 4.3: Comparison of simulated genera with Human Oral Microbiome Database (HOMD) genera

4.3 Primer optimisation

The first phase of primer optimisation involved testing all primer pairs with two generic species; *Escherichia coli* and *Streptococcus mitis*, to see whether they successfully amplified the targeted regions. For 23S rRNA no band was visible on the gel, indicating that the amplification of that target was unsuccessful, a positive control was run to verify that the gel migration had been successful. As shown above, the taxa theoretically amplified by 16S rRNA and 23S rRNA are similar therefore, both targets are not required. As the amplification of 23S rRNA was unsuccessful, it was removed from further optimisation. However, had the theoretically amplified taxa been different then the 23S rRNA primers would have been redesigned and optimised. All other primer pairs produced a band on the gel around 100bp, which corresponds to the targeted region and therefore they were all kept for further optimisation.

For the next stage of primer optimisation a temperature gradient was used to find out the best annealing temperature for each primer pair (see section 3.2.4). Figures 4.8 and 4.9 show the resulting gels for each primer pair at every temperature point (A-F). The optimum result as visualised on a gel is a single band corresponding to the target size, in this case all targets are around 100bp. The more extra bands present the less specific the amplification is. Non-specific amplification reduces the number of target sequences amplified by using up some of the reagents in the mix which would otherwise be used to amplify the target sequences. Therefore, at the end of the amplification process only some of the DNA comes from the target sequence. Subsequently, if this sample was sequenced directly there would be noise from the non-target sequences making it more difficult to

select the target sequences. Therefore, it is important to optimise the primers to get the most specific amplification possible.

Figures 4.8 and 4.9 show the amplification of *Streptococcus mitis* (odd numbers) and *Escherichia coli* (even numbers) with all primer pairs. Figure 4.8A shows the amplification result for the *rpoB*2 primer pair. For A1-F1 there is no band around 100bp indicating that for this species these primers do not amplify the targeted region. However, for A2-F2 there is a very visible band around 100bp, with a band growing in intensity at the top as the annealing temperature decreases. This larger band corresponds to non-specific amplification, therefore for this primer pair the chosen annealing temperature was 64.3 °C. Figure 4.8A and B shows the amplification results for the *rpoB*1_2 primers. A3-F3 show non-specific amplification with four bands visible for all but A3 and B3 as the latter have high annealing temperatures which, in this case impede amplification. For C3-F3 a band is visible around 100bp however due to the amount of non-specific amplification these primers are not ideal. This is corroborated by the results for A4-F4 which show no bands around 100bp except for H4 which has a very faint band. The results for the other *rpoB*1 primer pair (*rpoB*1_1) are presented in Figure 4.8B and C. For A5-F5 there is a distinct band around 100bp, however as the temperature decreases the amount of non-specific amplification increases with E5 and F5 showing bands between 900bp and 1500bp. For A6-F6 a band around 100bp can only be seen for the lower temperatures (D6-F6), however, as seen before with lower temperatures, non-specific amplification is observed. The results for this primer pair are much better than the first pair, therefore this pair is kept. The more intense a band is the higher the quantity of DNA associated with the band. E5 shows the greatest intensity with no non-specific amplification therefore, 59.9 °C is the chosen temperature for *rpoB*1. Figure 4.8C and Figure 4.9D and E show the results for 16S rRNA_1, with A9-F9 being a repeat of A7-F7 and A8-F8 the repeat of A10-F10. For the first set (A7-F7 and A9-F9) there is virtually no amplification with only a small amount of non-specific amplification at the lower temperatures. Whereas, for the second set (A8-F8 and A10-F10) there are lots of bands however most correspond to non-specific amplification, therefore this primer pair is not optimal for this study. Figure 4.9E and 3F show the results for the second 16S rRNA primer pair. For A11-E11 there are no bands with F11 showing a very faint band around 100bp. For A12-C12 there are also no bands however, D12-F12 all have a band around 100bp with F12 having the most intense band. F12 shows a very small amount of non-specific amplification however, to ensure enough sequences are amplified for sequencing, 56 °C is the chosen temperature.

Figure 4.8: Acrylamide gels of the amplification of *Streptococcus mitis* and *Escherichia coli* with *rpoB*2, *rpoB*1_2, *rpoB*1_1 and 16S_2 at different annealing temperatures

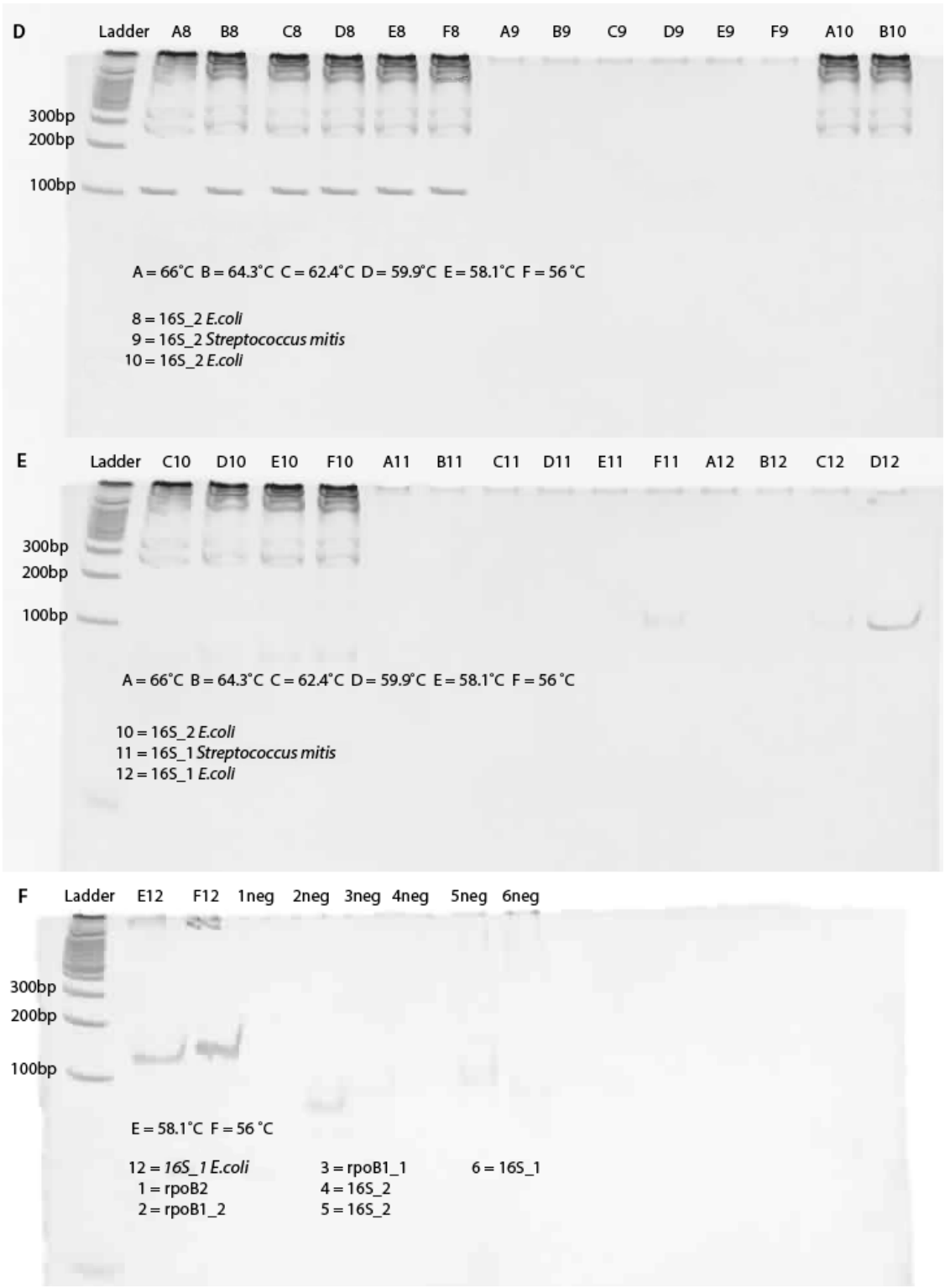

Figure 4.9: Acrylamide gels of the amplification of *Streptococcus mitis* and *Escherichia coli* with 16S_2 and 16S_1 at different annealing temperatures

The final stage of primer optimisation involved checking that the three chosen primer pairs; *rpoB*1, *rpoB*2 and 16S rRNA (see table 4.4 for final primer sequences), could successfully amplify the target regions from the saliva samples. Figure 4.10 shows the saliva samples (A1, A2, B1 and B2) from experiment one (t_1, t_2) amplified using the final three primer pairs. All samples show a band around 100bp which corresponds to the targeted region. This band is least intense for *rpoB*2, this corresponds to the fact that there are a number of mismatches in the primer binding site (see Figure 4.6) so less sequences will be amplified. There is also a small amount of non-specific amplification. This does not pose too much of a problem, as described in section 3.3.2.4 the pooled samples were separated on a gel and the band corresponding to the target region excised, hence removing the non-specific amplicons.

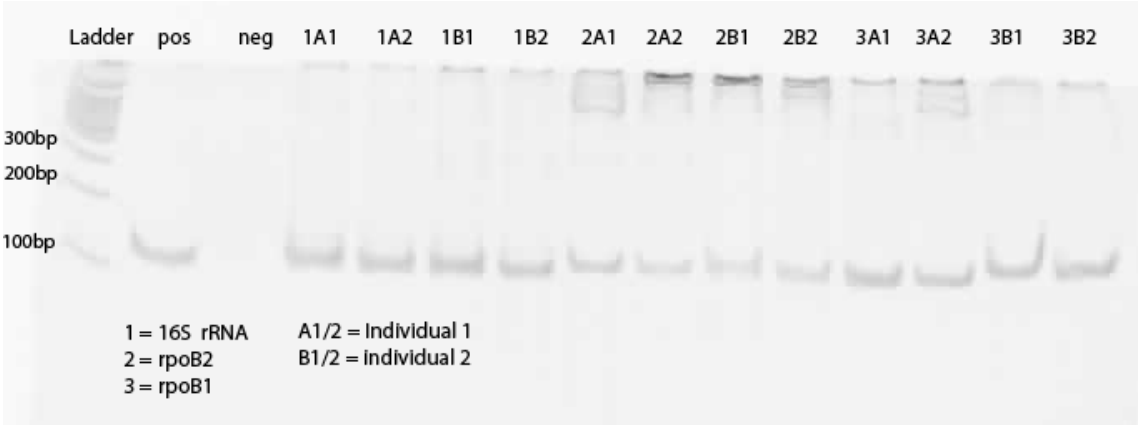

Figure 4.10: Acrylamide gel of the amplification of saliva samples with the final primers - 16S rRNA corresponds to primer pair 783F-878R, *rpoB*1 to 130F-220R and *rpoB*2 to 340F-439R.

Gene	Primer name	Primer Sequence (5' - 3')
16S rRNA	783F	AGGATTAGATACCCTGGTAG
	878R	CGTACTCCCCAGGCGG
*rpoB*1	130F	GGACCTGGTGGTTTGAC
	220R	CGATGTTAGGTCCTTCAGG
*rpoB*2	340F	GGACCAGAACAACCCG
	434R	GGGTGTCCGTCTCGAAC

Table 4.4: Final Primers - Overview of primers chosen for each gene target. Primer name for 16S rRNA and *rpoB*2 corresponds to the *E.coli* positions and for *rpoB*1 corresponds to *S.bovis* positions.

5

Results - Characterisation of the salivary microbiome

This chapter presents the results of the sequencing of all samples from both experiments (t_1, t_2 and t_3, t_4). Subsequently, the analysis of the sequences is presented in terms of which bacteria are present in which samples and which parameters were used to characterise the bacteria.

5.1 Illumina sequencing results

The saliva microbiome composition of 2 individuals was explored at 4 different time points. The samples were split into two sequencing runs with samples taken one month apart being sequenced together. Therefore, each run contained two samples per individual making 4 samples in total, per run. Run one corresponds to experiment one and was was performed one year before run two which corresponds to experiment two. This section presents the raw data in terms of the number of sequences produced by each sequencing run along with some basic sequence analysis.

In total, run one produced 193,221,302 reads and run two 201,692,619 reads. After quality control, pairing and filtering 59,971,947 and 56,762,234 sequences respectively, were used for analysis (sequences available in the European Nucleotide Archive under accession number PRJEB6052). Table 5.1 shows the summary statistics for both experiments broken down by target. The number of sequences for *rpoB*2 is a lot lower than for *rpoB*1 and 16S rRNA because, as mentioned in the previous chapter 4.1, the primers used have some base pair differences and therefore do not bind to as many sequences. All sequences assigned to the genera *Homo* or *Pan* were removed as these sequences are classed as contamination

from human DNA or the corresponding bacteria has the same sequence as humans and therefore the two cannot be differentiated. As both *Homo* and *Pan* genomes have been very well studied the database contains many accurate sequences and therefore if a sequence corresponds to both, the alignment with *Homo* or *Pan* is likely to be selected first due to the BLAST parameters used. As can be seen in table 5.1 the percent of contamination from *Homo* or *Pan* is very low at around 1% and therefore does not really impact the number of sequences available for analysis.

The percent unknown is calculated using all sequences where BLAST outputs a no hit. For both *rpoB* targets this percentage is very low and for 16S rRNA is zero or very close to zero. The percentage is lower for 16S rRNA as 16S rRNA is the standard target for metagenomic analyses and therefore it has been better characterised than *rpoB* and there are a lot more sequences in the database. However, the percent uncultured is very high for 16S rRNA at around 67% for experiment one and 58% for experiment two. As mentioned above 16S rRNA is the most common target used and therefore there are a high number of sequences in the database which have no taxon assignment other than uncultured bacterium/organism. With the arrival of high throughput sequencing bacterial communities could be analysed to a greater depth by targeted sequencing of 16S rRNA. Through rules for taxa delimitation using 16S rRNA (133) potential new species have been discovered however, many of these have not been cultured and cannot be cultured, making it difficult to prove their existence. This explains the high percent of uncultured for 16S rRNA. As the same database was used for both experiments these percentages remain consistent. The percent uncultured for both *rpoB* targets is very low, around 1% for experiment one and 0.01% for experiment two. This is because *rpoB* has a higher genetic resolution than 16S rRNA (152) meaning that universal primers cannot be used, so one pair of primers amplifies less species but the species which are detected are well defined so both the percent uncultured and unknown are very low.

Target	no. sequences after filtering	no. sequences with *Homo* & *Pan* removed	% *Homo* & *Pan*	% unknown	% uncultured	no. OTUs pre-filtering	no. OTUs	no. different OTUs
Experiment 1								
***rpoB*1**	29,693,058	29,560,125	0.448	0.008	0.084	72,926	891	150
***rpoB*2**	8,744,686	8,723,854	0.238	0.001	0.169	13,742	142	34
16S rRNA	21,534,203	21,270,245	1.226	0	66.873	165,587	1962	847
Experiment 2								
***rpoB*1**	17,007,924	16,813,642	1.142	0.349	0.010	61,732	1452	187
***rpoB*2**	9,149,974	9,099,842	0.548	1.940	0.002	28,482	313	54
16S rRNA	30,604,336	30,112,806	1.606	0.001	58.389	290,911	4665	1307

Table 5.1: Sequencing summary statistics for experiments 1 and 2

5.2 OTU count

Table 5.1 shows species-level OTU counts at different stages. The number of OTUs corresponds to the number outputted by the clustering algorithm. As described in section 3.4.5 any clusters containing less than twenty sequences were filtered out to remove the majority of sequencing errors. As can be seen in table 5.1 the number of OTUs pre-filtering is significantly greater than the number post filtering with about 99% of OTUs being filtered out. This does not mean that 99% of the data consists of sequencing errors, however distinguishing between sequencing errors and rare OTUs is extremely difficult. As stated above the percent of uncultured for 16S rRNA is very high and this corresponds to the much higher number of OTUs found for this target. The number of different OTUs is calculated by combining any OTUs with identical taxon name and Table 5.1 shows that the number of different OTUs is a lot lower than the number of OTUs. For 16S rRNA this number is not as low as it could be due to each unknown bacterium/organism having a different number and therefore none of the unknowns can be combined. Combining the OTUs enables a more accurate estimation of abundance per OTU, a feature which is used in downstream analysis for separating individuals.

Tables 5.2, 5.3 and 5.4 show the number of OTUs found in each sample, along with the percent in common between firstly, both samples from the same individual from each experiment and secondly, all samples from one experiment, for *rpoB*1, *rpoB*2 and 16S rRNA, respectively. For an OTU to be classed as in common between all samples from one experiment it must appear in at least one sample from each individual (e.g. A1 and B1). For the ' % OTUs in common ' between all samples from one individual, to be classed as in common an OTU must be present in at least one sample per experiment (e.g. A1 and A3 not A1 and A2). The number of different OTUs is consistent between samples in one experiment for all targets, with all samples in the the second experiment having more. *rpoB*2 contains the least with between 20 and 48 OTUs, *rpoB*1 is second with 145 to 185 OTUs and 16S rRNA has, by far, the most with between 793 and 1291. This correlates with the overall OTU count per target presented in Table 5.6. For *rpoB*1 97% of OTUs are the same between samples from the same individual for experiment one and about 94% for experiment two (see Table 5.2). About 97% of OTUs are in common between all samples in each experiment meaning that the differences between individuals are due to variation in abundances of bacteria and not different individuals having different bacteria. For *rpoB*2 these percentages are lower (see Table 5.3). For experiment one, about 73% are in common between samples from the same individual and 68% between all samples. For experiment two, about 90% are in common between samples from the same individual and

78% between all samples. The lower percentage in common between all samples indicates that some of the differences between individuals comes from different individuals having different bacteria. For 16S rRNA the percentages are similar to *rpoB*1 (see Table 5.4), indicating that the differences between individuals are due to different abundances of the same bacteria. The above tables demonstrate that samples sequenced in the same run have very similar populations. Table 5.5 shows the ' % OTUs in common ' between all samples from each individual separately. It can be seen that the ' % in common ' is a lot lower and comparable to the ' % OTUs in common ' per target for all samples combined, see table 5.6. This implies that in terms of OTUs present, both individuals have very similar populations. When the OTUs in common with all samples are compared to those in common in each individual they are nearly all the same. For 16S rRNA there is one species per individual which is different. For *rpoB*1 all of the species in common are the same and for *rpoB*2 there are two species per individual which are different, see Table 5.5 for the corresponding percentages. This reiterates the idea that inter-individual variation comes from differences in abundances of bacteria rather than biodiversity.

Sample	**Different OTUs**	**% OTUs in common 1**	**% OTUs in common 2**
Experiment 1			
A1	145		
A2	147	97	
B1	149		
B2	144	97	98
Experiment 2			
A3	182		
A4	185	97	
B3	169		
B4	171	91	95

Table 5.2: Comparison of species-level OTUs between all samples for *rpoB*1 - % OTUs in common 1 describes the number of OTUs in common between the two samples from the same individual from one experiment (e.g. A1 and A2) and % OTUs in common 2 describes the number of OTUs in common between all the samples from one experiment.

Sample	Different OTUs	% OTUs in common 1	% OTUs in common 2
Experiment 1			
A1	20		
A2	23	72	
B1	25		
B2	29	75	68
Experiment 2			
A3	46		
A4	44	88	
B3	44		
B4	48	92	78

Table 5.3: Comparison of species-level OTUs between all samples for *rpoB*2 - % OTUs in common 1 describes the number of OTUs in common between the two samples from the same individual from one experiment (e.g. A1 and A2) and % OTUs in common 2 describes the number of OTUs in common between all the samples from one experiment.

Sample	Different OTUs	% OTUs in common 1	% OTUs in common 2
Experiment 1			
A1	810		
A2	793	94	
B1	839		
B2	828	97	98
Experiment 2			
A3	1273		
A4	1267	97	
B3	1291		
B4	1283	99	98

Table 5.4: Comparison of species-level OTUs between all samples for 16S rRNA - % OTUs in common 1 describes the number of OTUs in common between the two samples from the same individual from one experiment (e.g. A1 and A2) and % OTUs in common 2 describes the number of OTUs in common between all the samples from one experiment.

		% OTUs in common	
Individual	***rpoB*1**	***rpoB*2**	**16S rRNA**
A	50	26	37
B	52	25	38
A+B	100	89	99.5

Table 5.5: Comparison of species-level OTUs between individuals for all targets - the percentages for A+B combined refers to the percent OTUs in common between those in common for each individual separately.

5.2.1 Both experiments combined

By combining the two experiments a number of species which are not in common between the two experiments are lost. By removing these species the technique remains conservative. It is possible that some of these species are real however, some of them could be sequencing artefacts so to be sure to only include real and present species those not found in both experiments were removed. The chance of a species being observed in both samples from an individual and then not at all in the other two samples is unlikely, one would expect it to be present in at least one of the other two samples. However, it is not impossible as it could be a transient species detected on a short term basis due to a particular lifestyle habit. Even less likely is both samples from both individuals having a species in common and then that species not being present in either of their samples in the second experiment. Table 5.6 shows the comparison of species-level OTU count between the two experiments. Systematically experiment two has more OTUs than experiment one, however the proportions of each target remains the same. For *rpoB*2 and 16S rRNA about one third of the OTUs are in common between the two experiments whereas for *rpoB*1 this figure is higher at 50%. These percentages seem quite low indicating that many OTUs are not detected in both experiments. However, when looking at the percentage of sequences allocated to ' OTUs in common between both experiments ' (see table 5.6), the values are a lot larger. In fact, nearly all of the sequences are assigned to OTUs found in both experiments. This indicates that the OTUs not in common most likely correspond to rare taxa which are harder to sequence and therefore, are not detected in both experiments.

Target	OTUs exp 1	OTUs exp 2	OTUs in common	% OTUs in common	Average % sequence allocation in common
***rpoB*1**	150	187	113	50	99.40
***rpoB*2**	34	54	19	28	99.48
16S rRNA	847	1307	590	38	88.42

Table 5.6: Comparison of species-level OTUs between experiments for all targets - (exp = experiment), the average % sequence allocation in common describes the percentage of sequences allocated to OTUs in common between both experiments.

5.3 Microbiome composition

The use of three targets enables the microbiome composition to be analysed to a greater depth. Figure 5.1 shows the proportions of the top five phyla per individual, per target.

The first thing to note is that, at the phylum level, both individuals have very similar abundances. However, this is logical as phylum is very high in taxonomic classification (see Figure 3.2) and therefore any differences are more likely to be at lower levels of classification, such as species. It can be seen that for both *rpoB*1 and 16S Firmicutes is the most common phylum constituting over 90% and 70% of the population respectively. For *rpoB*2 the population is composed of over 90% Actinobacteria. Previous studies (71, 74, 84) have shown that the most common phlya found in saliva are: Firmicutes, Proteobacteria, Actinobacteria, Bacteroidetes and Fusobacteria and this study concurs with these findings, however the abundances differ slightly. Stahringer *et al.* analysed 264 saliva samples and showed that bacteria abundances varied greatly, this study falls within the observed variation. In the same study they defined a genus-level core microbiome containing eight genera (74) (see section 1.2 for definition of core microbiome). By combining three targets in this study, through merging of abundance tables, a genus-level core microbiome of 58 genera was observed, see table 5.7 for the breakdown of genera per target. This high number of genera covers about 95% of the population of each individual implying that most differences come from the species/strain level.

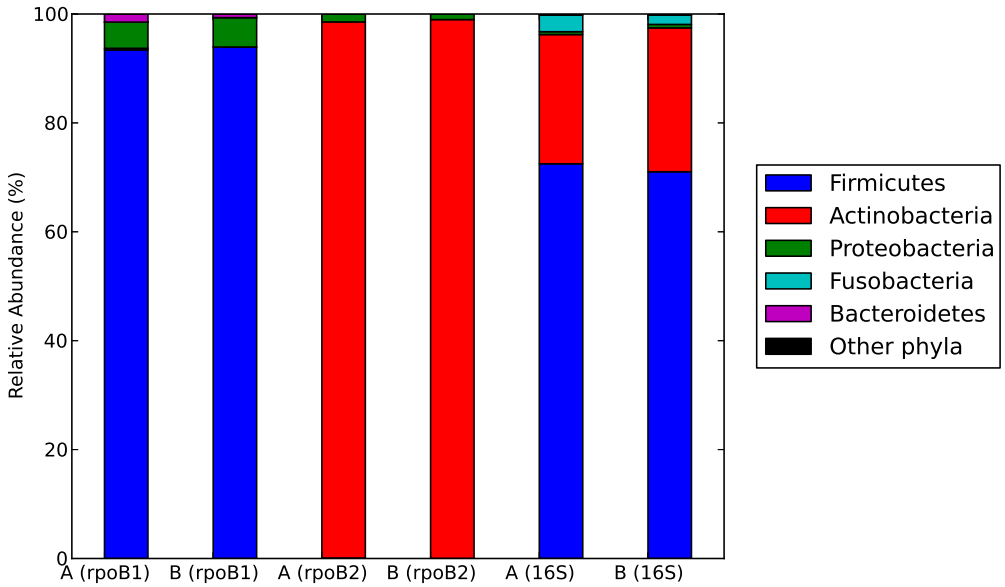

Figure 5.1: Relative abundance of the top five phyla per individual per target for both experiments combined - A and B are different individuals and the target is in brackets.

The addition of *rpoB* enables certain genera to be analysed down to the species and even strain level. Specifically, with 16S *Streptococcus* can be detected at the genus level and

occasionally the species level (9 different OTUs) however, with *rpoB* it can be detected to the species/strain level (53 different OTUs) enabling a deeper characterisation of this part of the saliva microbiome (see appendix A for list of species/strains per target (section 8.1)). This is important as *Streptococcus* makes up about 80% of Firmicutes, the most abundant phylum.

16S rRNA	***RpoB* 1**	***RpoB* 2**
*A*ctinomyces	*A*biotrophia	*A*ctinobacillus
*A*erococcus	*A*cinetobacter	*A*rthrobacter
*A*naerovorax	*A*ctinobacillus	*B*ifidobacterium
*A*rthrobacter	*A*eromonas	*E*scherichia
*A*topobium	*A*rthrobacter	*R*othia
*B*acillus	*A*topobium	
*B*revibacterium	*B*acteroides	
*C*ampylobacter	*B*artonella	
*C*itricoccus	*E*nterococcus	
*C*lostridium	*E*xiguobacterium	
*C*orynebacterium	*G*allibacterium	
*D*ermabacter	*H*aemophilus	
*D*ietzia	*L*actobacillus	
*E*nteroactinococcus	*L*actococcus	
*E*nterococcus	*L*eadbetterella	
*E*ubacterium	*L*isteria	
*K*ocuria	*M*arinomonas	
*L*actobacillus	*M*ethylotenera	
*M*icrococcus	*N*eisseria	
*M*obiluncus	*P*aenibacillus	
*N*egativicoccus	*P*asteurella	
*N*eisseria	*P*edobacter	
*N*ocardiopsis	*P*revotella	
*P*elotomaculum	*P*seudomonas	
*P*eptostreptococcus	*R*othia	
*P*revotella	*S*accharophagus	
*P*ropionibacterium	*S*hewanella	
*R*othia	*S*taphylococcus	
*S*elenomonas	*S*treptococcus	
*S*treptococcus	*V*eillonella	
*T*reponema	*V*ibrio	
*T*richococcus	*W*eissella	

Table 5.7: Core genera per target - list of core genera per target in alphabetical order.

Table 5.8 shows the most common genera found in all samples along with the percent abundance each genus represents per target. For *rpoB*1, *Streptococcus* is the most abundant, confirming the results presented above in Figure 5.1 showing Firmicutes as the most abundant phylum. As for 16S rRNA there are many more unclassified/uncultured or-

ganisms than anything else and the second most common genus is *Streptococcus*. Both *rpoB*1 (designed to target streptococci species) and 16S rRNA reveal that *Streptococcus* is the most commonly detected genus in saliva. These results reiterate the advantage of sequencing more than one target as without *rpoB*1, many of the streptococcus species would remain hidden. For *rpoB*2 *Rothia* is the most abundant genus at 93%, moreover it is barely detected by the other two targets (see annex for complete list of species and abundances per target). As described in section 1.4, all of the above mentioned genera, are commonly found in saliva and therefore these results agree with previously published work (30, 70, 153). Even though there is a core microbiome of 58 genera, most of them are in very low abundance with only a few genera making up the majority of the population. The species varying most in abundance between individuals will be presented in the next chapter.

***rpoB*1**	***rpoB*2**	**16S rRNA**
*S*treptococcus (75%)	*R*othia (93.13%)	Unclassified (66.12%)
*L*actobacillus (10.84%)	*A*rthrobacter (5.43%)	*S*treptococcus (22.85%)
*A*biotrophia (7.55%)		*A*rthrobacter (8.73%)
*N*eisseria (2.23%)		*A*ctinomyces (1.52%)
*S*accharophagus (1.76%)		
*P*revotella (1.21%)		

Table 5.8: Most common genera in all samples - this table shows the genera which constitute about 99% of the population for each target, the average percentage, across all samples, represented by each genus is in brackets.

5.4 Clustering threshold

Unlike previous studies the main aim of this thesis was to investigate whether the bacteria found in saliva could be used to separate samples from different individuals and not just characterise the microbiome. Different clustering thresholds, used with CD-HIT (see section 3.6.2), were tested to see which one gave the best separation taking into account analysis time (see spreadsheet ' clustering threshold ' on accompanying CD for raw and processed data). Figure 5.2 shows that as the percent identity[1] increases so does the relative distance between the two individuals. The results for both *rpoB* targets are shown in Figure 5.2A where the dashed line indicates the chosen threshold of 95%. In Figure 5.2B the dashed line highlights the chosen threshold for 16S rRNA of 97%. These percentages correspond to previously published studies for species level characterisation for *rpoB* and 16S rRNA, respectively (134, 154). For both targets 100% identity provides

[1]the percent similarity between DNA sequences required to combine them as one cluster

the best separation however the analysis time, for 16S rRNA especially, is very long and therefore it is not the most efficient solution. The majority of the analysis time is taken up by the clustering and BLAST stages. The higher the percent identity the longer the analysis takes. At 100% identity for two sequences to be placed in the same cluster every base pair must match. With species level characterisation for *rpoB* and 16S rRNA being 95% and 97% respectively, 100% identity would correspond to strain level, making many more clusters. The more clusters there are the longer BLAST takes as there are more representative sequences to compare to the database. The analysis of 16S rRNA at 100% identity took about 3 weeks whereas for at 97% identity it took about 2 weeks. For rpoB the analysis time is a lot less (hours-days) as there are fewer sequences, however for rpoB1 at 100% identity the analysis still took about 6 days, whereas at 95% it took about 4 days. As strain level identification is not necessary for the separation of individuals neither is a clustering threshold of 100%. In order to make this technique as time efficient as possible the clustering threshold used is the lowest one which enables separation of individuals whilst following standard taxonomic classification rules.

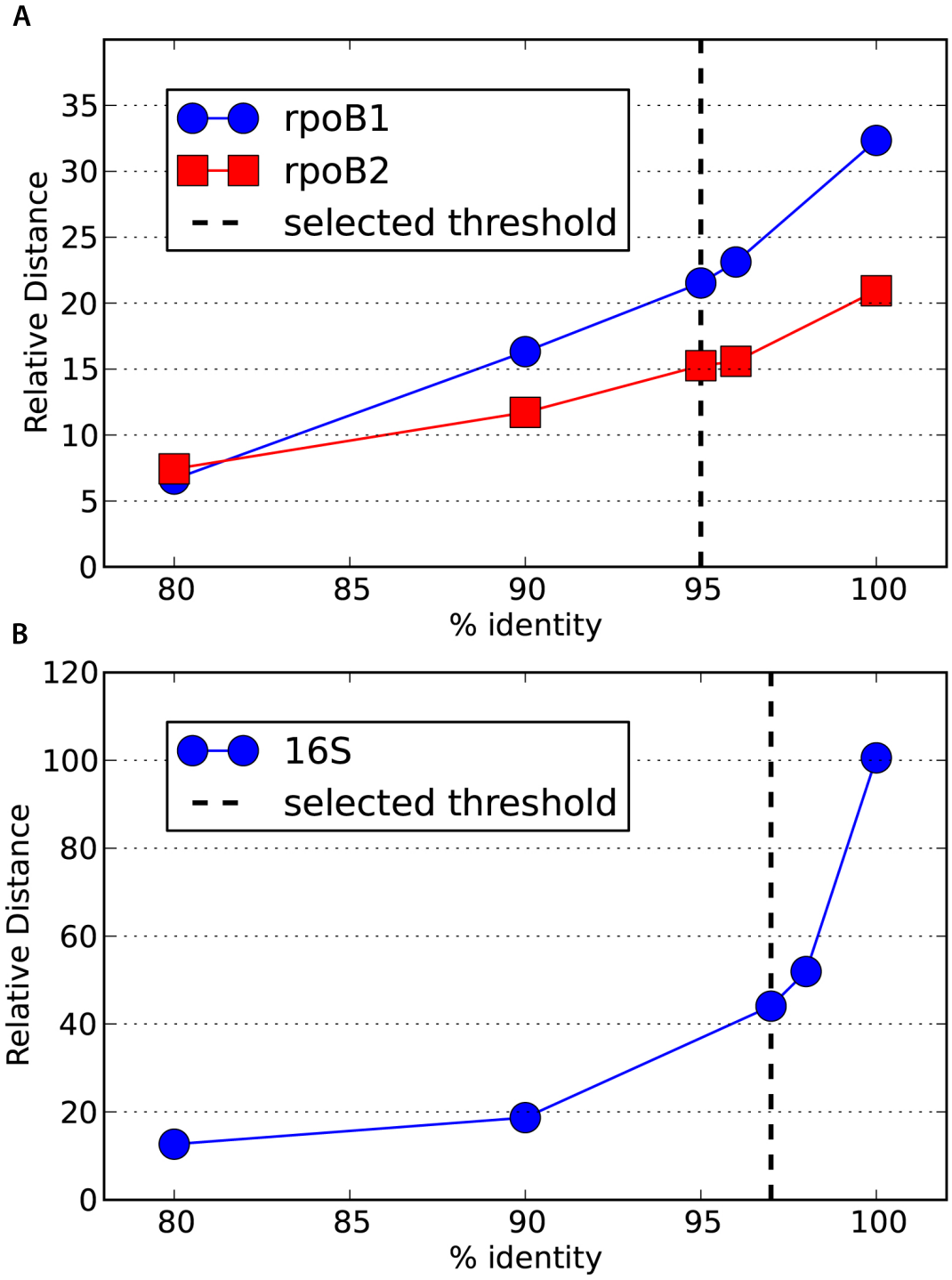

Figure 5.2: Comparison of clustering thresholds for the separation of individuals - The percent identity is that used for clustering the sequences into OTUs with CD-HIT. The relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.1 from a t-test (a Bayesian analysis) between the samples from each individual or a BF <1 were used. A = both *rpoB* targets and B = 16S rRNA. The dashed line highlights the chosen threshold.

6

Results - Comparison of two salivary microbiomes

This chapter describes how the composition of the salivary microbiome, as presented in the previous chapter 5.3, can be used to separate samples from one individual from those from a second individual. As detailed in the materials and methods chapter 3 the data was transformed in order to analyse it and reveal the potential differences between the two individuals. Section 5.2.1 describes how by combining both experiments a certain number of species are lost, however this is necessary to avoid including any sequencing errors. In this chapter, when all eight samples are being compared it is implicit that only the species in common with both experiments are used.

6.1 Normalisation

As two different sequencing runs were used to analyse the samples, the data from both were normalised to make them comparable. As can be seen from the count tables found in the supplementary files (https://independent.academia.edu/SarahLeake/Papers), the counts between sequencing runs for the same species can differ a lot, indicating that it is necessary to normalise the data between runs. If normalisation is not used then differences seen may only be due to differences between the sequencing runs and not real differences between individuals. The choice of normalisation algorithm is important, as described in section 3.5.1 the structure of the data has an impact on which algorithm is used. In this case a few species have high abundance with most having very low abundance. Table 6.1 shows the number of high abundance species compared with the total number of species in common between both experiments and what percentage their sequences represent of

the total in common. Species were classed as high abundant if there were more than 200 occurrences in everyone of the eight samples. It can be seen that for both *rpoB* targets about 16% of the species have a high abundance and they account for about 98% of the sequences assigned to species in common between both experiments. For 16S rRNA only about 6% of the species have high abundance however they account for about 97% of the assigned sequences. This data demonstrates why a normalisation algorithm, like DESeq (141), which takes into account the uneven distribution of sequences is important.

Target	**no. high abundance species/total in common**	**% high abundance species sequences**
***RpoB*1**	18/113	98
***RpoB*2**	3/19	99
16S rRNA	38/590	97

Table 6.1: Comparison of high abundance species between targets - the percent of high abundance species sequences refers to the percentage of sequences assigned to a high abundant species out of all of the assigned sequences for species in common between both experiments.

6.2 Data filtering

After the data was normalised and log transformed (see section 3.5.2) it was filtered to first remove any sequences assigned to the genera *Homo* or *Pan* (see section 5.1) and secondly to keep only the OTUs which were calculated as being significantly different between the two individuals. As described in section 3.5.3 the data was filtered using two different approaches to show that both approaches produce the same results, ensuring the robustness of the analysis. Figure 6.1 shows the relative distance calculated between samples from each experiment separately for both unfiltered and filtered data per target. It can be seen that the relative distance is higher for the filtered data than the unfiltered data. The larger the relative distance the better the separation between individuals, hence demonstrating the advantage of filtering the data. It is also worth noting that the pattern is the same for both experiments, so even when the number of sequences differs the effect of filtering remains the same.

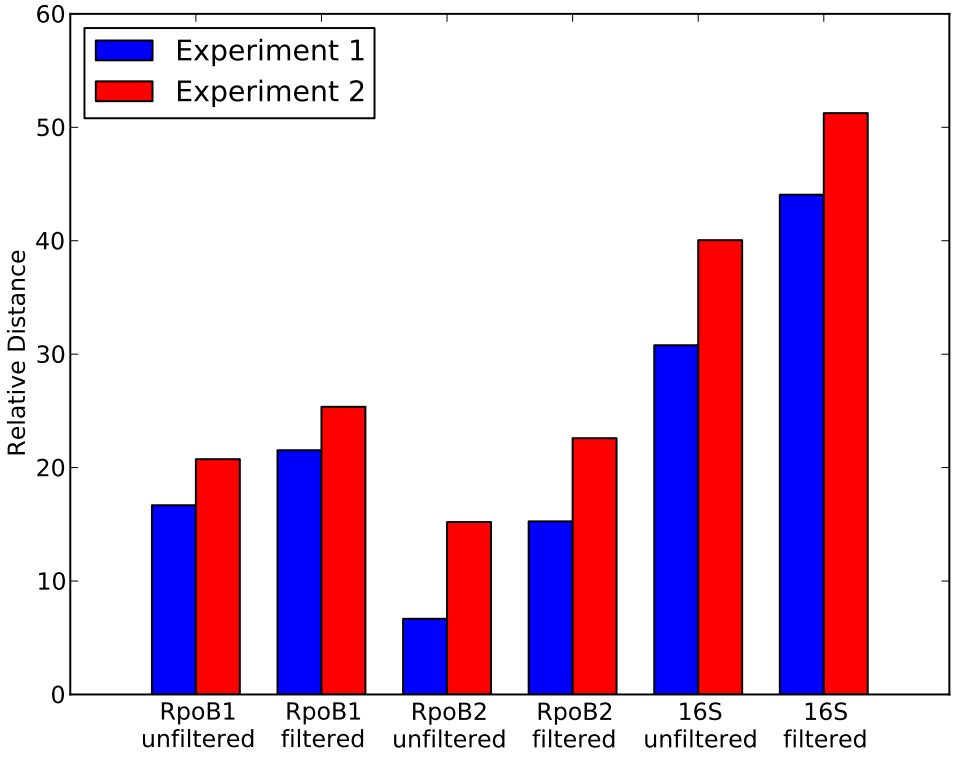

Figure 6.1: Comparison of unfiltered and filtered data for both experiments per target - the relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. For the filtered samples only species with a p-value <0.1 from a t-test between the samples from each individual or a BF <1 were used.

When combining the two experiments, as mentioned previously in section 5.2.1, a number of species are lost as only those in common to both experiments are kept. Due to this, if the data is left unfiltered then for *rpoB*1 and 16S rRNA the samples group by experiment and not by individual (see Figure 6.2A and C) as there are enough experiment specific species that the sample grouping is skewed. For *rpoB*2 the samples still group by individual (see Figure 6.2B) however, the separation between individuals is not as large as for the filtered data. All the data analysis in the rest of this chapter is performed on the filtered data only.

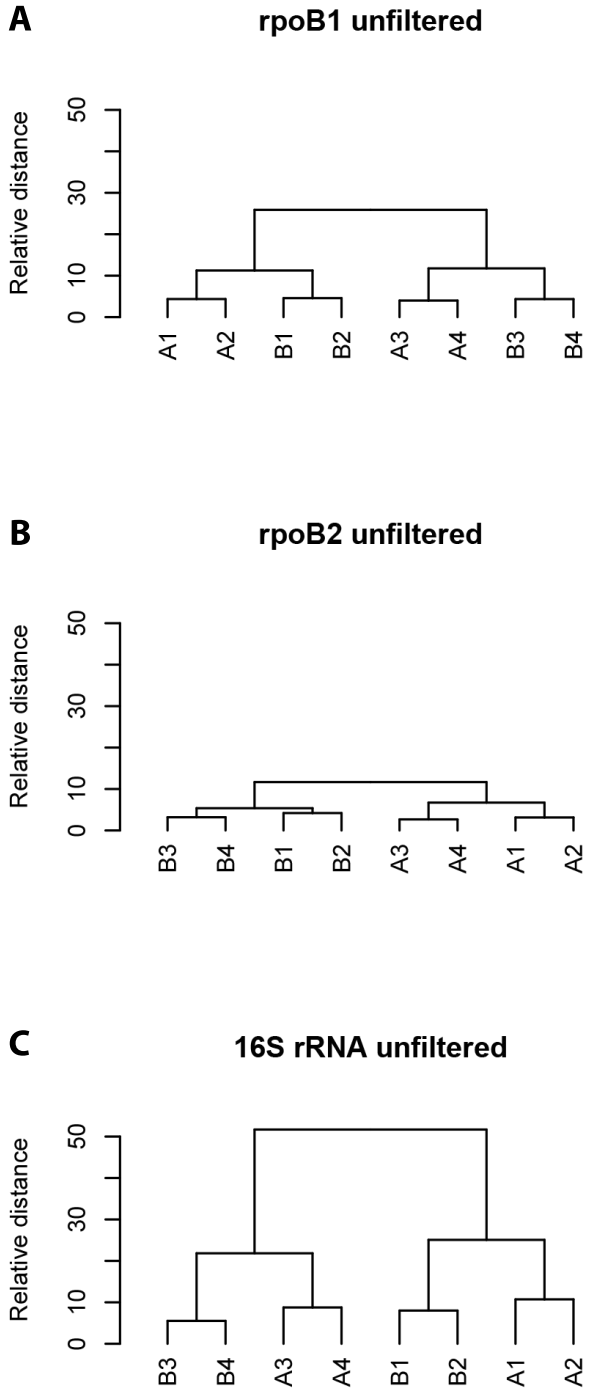

Figure 6.2: Hierarchical clustering of both experiments combined using unfiltered data - the relative distance corresponds to the distance between two individuals (A and B) calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.01 from a t-test between the samples from each individual or a BF <1 were used.

6.2.1 Choice of p-value

As described in section 3.5.3 and under assumptions expressed there, the data was filtered by performing a t-test between each species and then ordering the data by p-value. The BF values were also calculated, however the purpose of this was to corroborate the p-values. Therefore, only the choice of p-value is presented here. The next step was to decide which p-value separates the samples best. Table 6.2 shows the relative distance, as calculated through hierarchical clustering, between samples from both experiments individually and then combined at different p-values. For the individual experiments as the p-value decreases so does the relative distance, therefore a p-value <0.1 produces the greatest separation between individuals. Only 16S rRNA has distances for all p-values, this is because the other two target genes are more specific so fewer OTUs are detected (see Table 6.3), therefore as the data is filtered less OTUs are available and the chance of one of the few being highly significant is low. Also, with such a low number of OTUs some fall into the same category, for example, for *rpoB*2 in experiment two the distances for p-values <0.05 and <0.01 are the same because between the two p-values no OTUs are filtered out. When the two experiments are combined the best p-value for *rpoB*1 and 16S rRNA is <0.01 and <0.05 for *rpoB*2. By combining the experiments eight samples are being clustered instead of four therefore OTUs which are more significant produce a better separation. However, when filtered at a too high significance the distance decreases as not enough OTUs are included. The p-value chosen for the analysis of individual experiments is <0.1 and <0.01 for the combined experiments. When the individual experiments are compared to the combined experiments both a p-value of <0.1 and <0.01 will be presented for the combined experiments.

Target/p-value	<0.1	<0.05	<0.01	<0.005	<0.001	<0.0005
Experiment 1						
***RpoB*1**	21.53	19.94	10.69	8.78	4.20	-
***RpoB*2**	15.26	15.06	13.70	8.05	-	-
16S rRNA	44.06	40.94	27.20	21.47	12.72	12.72
Experiment 2						
***RpoB*1**	25.36	22.02	14.76	11.47	9.29	9.07
***RpoB*2**	22.59	22.24	14.83	14.83	-	-
16S rRNA	51.25	48.24	35.57	26.68	13.51	9.13
Combined						
***RpoB*1**	22.44	23.79	25.46	25.38	22.12	20.65
***RpoB*2**	15.13	15.96	13.81	13.81	9.30	9.30
16S rRNA	54.22	56.20	57.36	55.83	51.75	49.40

Table 6.2: Comparison of relative distance between individuals for all targets for experiment 1, experiment 2 and both experiments combined at different t-test p-values - (dashes mean that there were no significant species under that p-value).

6.3 Hierarchical clustering method

The aim of this study was to see whether the bacteria found in saliva could be used to separate samples from two individuals. In order to achieve this the abundances of each species per sample were clustered using hierarchical clustering to see how the samples group together. As the aim is to group samples coming from the same individual together, the clustering method used should reflect this. Section 3.5.3.1 lists the clustering methods available for hierarchical clustering. Initially two different methods were tested; complete linkage and Ward, as they both find similar clusters. Figure 6.3 shows the relative distance calculated between samples from each experiment separately using both the complete linkage (comp) and Ward methods of clustering. The relative distance for the Ward method is always larger than for complete linkage for all targets and both experiments. Furthermore, the difference between the two methods is fairly substantial with the Ward method producing distances about 50% larger than complete linkage for all targets. Therefore, the Ward method of hierarchical clustering is the chosen method for the data analysis in this thesis.

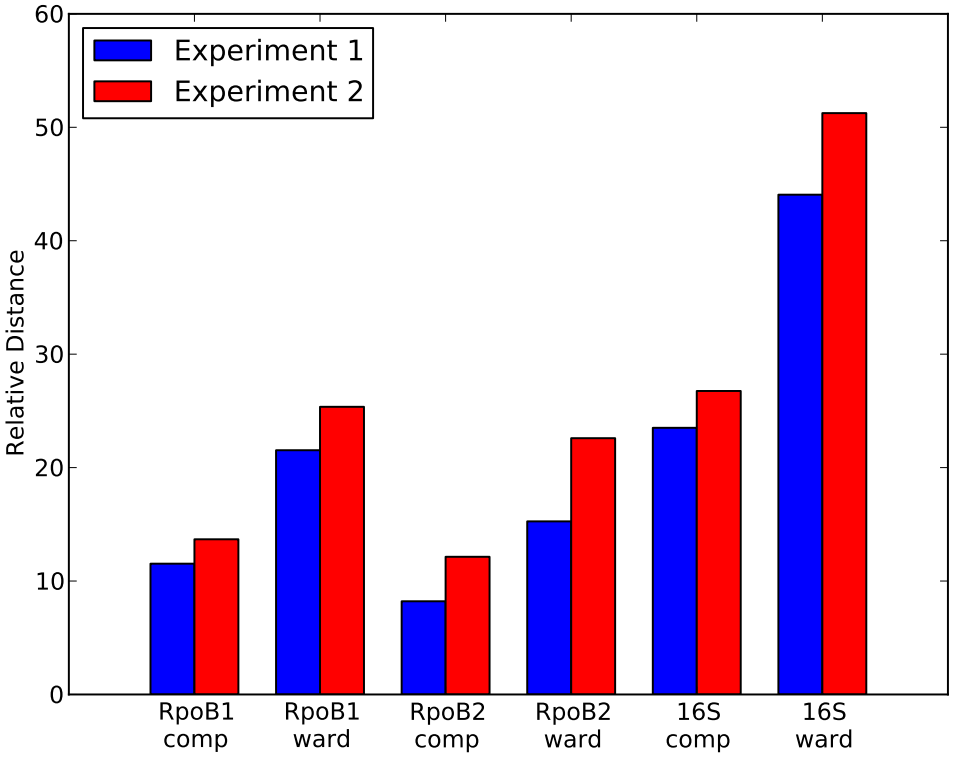

Figure 6.3: Comparison of two different hierarchical clustering methods - the relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and either the complete linkage (comp) or Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.1 from a t-test between the samples from each individual or a BF <1 were used.

6.4 Clustering of individual experiments

Before comparing the two experiments combined, the samples from each individual experiment were clustered separately. This was important in order to show that it was possible to initially group samples from the same individual together, if this was not the case, then combining them would very unlikely provide better results. Figure 6.4 shows the hierarchical clustering of the samples from each experiment separately, per target. For all targets, samples from the same individual group together and are separated from the other individual. However, each target separates the samples differently with 16S rRNA providing the greatest separation with an average total relative distance of 46 compared to 23 and 19 for *rpoB*1 and *rpoB*2, respectively. This correlates to the number of OTUs available for analysis per target with 16S rRNA having the most and *rpoB*2 the least (see section 5.2). It is also worth noting that the relative distance is larger for all targets in experiment two (see Figure 6.4B, D and C) than experiment one, this can also be explained by the number of OTUs available for analysis, with experiment two having more (see Table 6.3). Table 6.3 also shows that of all the OTUs that pass the filtering stages only about 25% and 32% for experiment one and experiment two respectively, are classed as significant. For both *rpoB*1 and 16S rRNA the significant OTUs account for about 16% of the total number of sequences after filtering, this figure drops to 1% for *rpoB*2. This indicates that the majority of the OTUs are similar between individuals and that the significant OTUs are in low abundance.

The relative distance between samples from the same individual is not equal to zero due to intra-individual variation. This variation can come from a number of factors, of which principally, diet and daily routine are the main contributors. As described in section 3.1 the participants brushed their teeth in the morning and did not consume any food one hour prior to sampling. However, this did not stop them from eating after they brushed their teeth and before the one hour prior to sampling. Therefore, on the days of sampling they could have eaten different foods and/or had contact with different environments. In terms of the applicability of this technique it is important to keep the samples as real as possible. If everything was controlled then it would be hard to draw conclusions concerning the natural variation of samples. With the small sample size available concrete conclusions cannot be drawn regarding intra-individual variation, however all of these samples show that the intra-individual variation is much lower than the inter-individual variation. For *rpoB*1 (Figure 6.4A and B) the intra-individual variation is about 18% of the inter-individual variation, this figure is around 11% for *rpoB*2 (Figure 6.4C and D) and 17% for 16S rRNA (Figure 6.4E and F). Furthermore, the intra-individual variation

is about the same for both individuals indicating that the variation within a person is stable over time. For all targets the intra-individual variation is less than 20% of the inter-individual variation.

Target	no. significant OTUs	% significant OTUs	Average % sequence allocation of significant OTUs
Experiment 1			
***RpoB*1**	40	26.67	14.71
***RpoB*2**	8	23.53	0.19
16S rRNA	190	22.43	14.61
Experiment 2			
***RpoB*1**	72	38.50	13.54
***RpoB*2**	15	27.78	1.78
16S rRNA	370	28.31	22.83
Combined			
***RpoB*1**	19 / 8	16.81 / 7.08	2.38 / 1.61
***RpoB*2**	10 / 2	50 / 10	0.30 / 0.1
16S rRNA	159 / 68	26.95 / 11.53	20.66 / 5.95

Table 6.3: Comparison of significant OTUs between all targets for experiment 1, experiment 2 and both experiments combined - significant OTUs are those with a p-value <0.1 from a t-test between the samples from each individual or BF <1. The last column shows the % of sequences assigned to significant OTUs out of the total number of sequences used for analysis after filtering. For the experiments combined species significant at a p-value <0.01 are also presented (the number after the /).

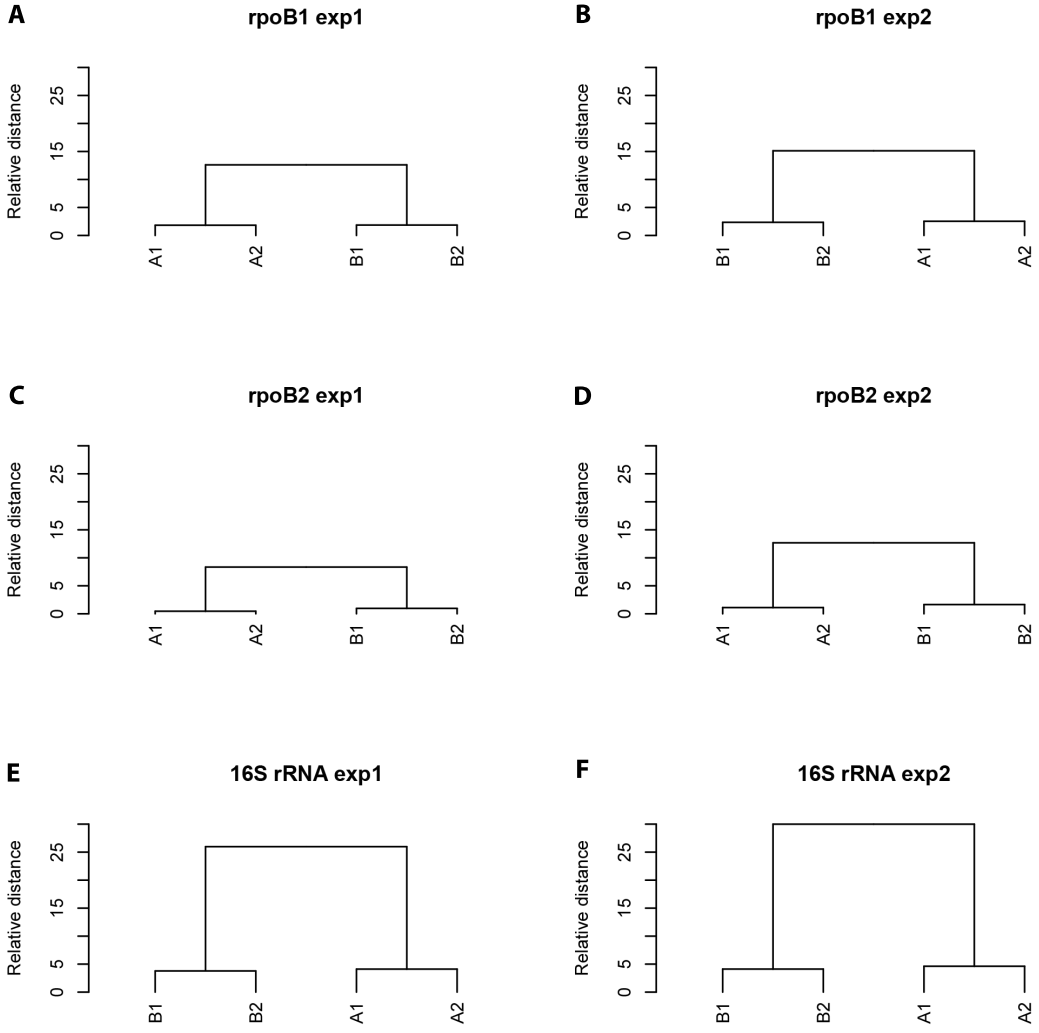

Figure 6.4: Hierarchical clustering for each target for both experiments individually - the relative distance corresponds to the distance between two individuals (A and B) calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.1 from a t-test between the samples from each individual or a BF <1 were used. Exp1 = experiment 1 and exp2 = experiment 2.

6.5 Clustering of combined experiments

Section 6.4 demonstrates that when each experiment is analysed separately, samples from one individual can be separated from those of another. The next step is to see whether, once the two experiments have been combined, this separation is still possible. Figure 6.5 shows the hierarchical clustering of the samples from both experiments combined, per target. For all targets, samples from the same individual group together and are separated from the other individual. As seen for the clustering of the individual experiments each target separates the samples with varying distances (see Table 6.2). 16S rRNA provides the greatest separation with total relative distance of 57 compared to 25 and 14 for *rpoB*1 and *rpoB*2, respectively. When these values are compared to those from the single experiments, for *rpoB*2, they are lower, however for *rpoB*1 and 16S rRNA the values are higher. This could be because *rpoB*1 and 16S rRNA target more species hence, the number of species found is greater and therefore, more are available for clustering the samples (see Table 6.3). Table 6.3 shows that the number of significant OTUs for the combined experiments is lower than for each experiment individually. By combining the two experiments and filtering out only those in common to both some of the significant OTUs in the individual experiments could be filtered out. Therefore, less OTUs are available for clustering. The % significant OTUs changes slightly for the combined data. For *rpoB*1 the value is lower than for either experiment individually implying that by combining the samples some of the significant taxa are lost. Whereas, for *rpoB*2 the value nearly doubles to 50% for a p-value of <0.1, indicating that some of the lost taxa are not significant and more of the significant taxa are kept. However, at a p-value of <0.01 this percentage drops to 10 indicating that in comparison with the individual experiments there are fewer significant OTUs. For 16S rRNA the value remains similar at a p-value of <0.1, however these values decrease at a p-value of <0.01 as less OTUs are kept. Concerning the average % sequence allocation of the significant OTUs for both *rpoB*1 and *rpoB*2 the values are lower than for the individual experiments at 2.4% and 0.3% respectively. This reiterates the above results that the significant OTUs are in low abundance. For 16S rRNA the value remains the same implying that the significant OTUs represent the same proportion of the population.

To further investigate the effect of combining the experiments the significant OTUs from each individual experiment were compared to those from the two experiments combined. Table 6.4 shows that per experiment under 50% of the significant OTUs are still significant when the two experiments are combined. However, when the significant OTUs in common to each experiment are summed together and compared with the significant OTUs from

the two experiments combined, at a p-value of <0.1, this percentage increases to around 70% for both *rpoB*1 and 16S rRNA and 50% for *rpoB*2. The main reason for this is that an OTU might be classed as significant in an individual experiment but when combined with the second experiment is no longer significant and vice versa. Additionally, when the experiments are combined the total number of OTUs decreases (see Table 5.6) in comparison to the individual experiments and OTUs which are found in one experiment may not be present in both. Furthermore, there are some OTUs which are not classed as significant in either of the individual experiments but when the two experiments are combined are classed as significant. For example, for *rpoB*1, at a p-value of <0.01, 0% of the significant OTUs from both experiments are in common with the combined experiments, indicating that the significant OTUs for the combined experiments are different to those for the individual experiments. This explains why when the significant OTUs from each individual experiment are combined the % in common is not 100.

When the significant OTUs are compared at a p-value of <0.01 all the percentages decrease as the number of significant OTUs in common decreases, this is because for the individual experiments a p-value of <0.01 removes too many OTUs (see section 6.2.1). By increasing the significance there is even more chance that OTUs found to be significant in an individual experiment will not be as significant when the two experiments are combined. However, when the OTUs significant at a p-value <0.1 from the individual experiments are compared to OTUs significant at a p-value <0.01 from both experiments combined the percentage of significant OTUs from both experiments in common with the experiments combined increases to between 75% and 100% (see Table 6.4. This implies that a large number of OTUs are in fact significant to both individual and combined experiments.

As described above for the individual experiments the relative distance between samples from the same individual is not zero due to intra-individual variation. With the combined experiments this variation increases for *rpoB*2 but remains smaller than the inter-individual variation and decreases for *rpoB*1 and 16S rRNA. For *rpoB*1 (Figure 6.5A) the intra-individual variation is about 15% of the inter-individual variation, this value is around 30% for *rpoB*2 (Figure 6.5B) and 14% for 16S rRNA (Figure 6.5C). Furthermore, the intra-individual variation differs slightly between individuals and targets in terms of separation of samples from the same individual. It could be expected that samples sequenced in the same run would be grouped together as they were subjected to identical conditions. This is the case for both individuals with *rpoB*2 but only for one individual for *rpoB*1 and 16S rRNA. This implies that some of the differences transcend the two

experiments and have more of an effect than the sequencing conditions. As the data has been filtered and only those OTUs in common to both experiments are kept then it is also possible that the samples from the same individual group arbitrarily. The most important point to make is that all the samples from one individual group together and the amount of intra-individual variation is comparable between individuals, implying that the variation within a person is stable over time. *rpoB*2 has the smallest difference between inter and intra-individual variation making it the least suitable target gene. Both *rpoB*1 and 16S rRNA have intra-individual variation of less than 20% correlating with the intra-individual variation found in the individual experiments (see section 6.4).

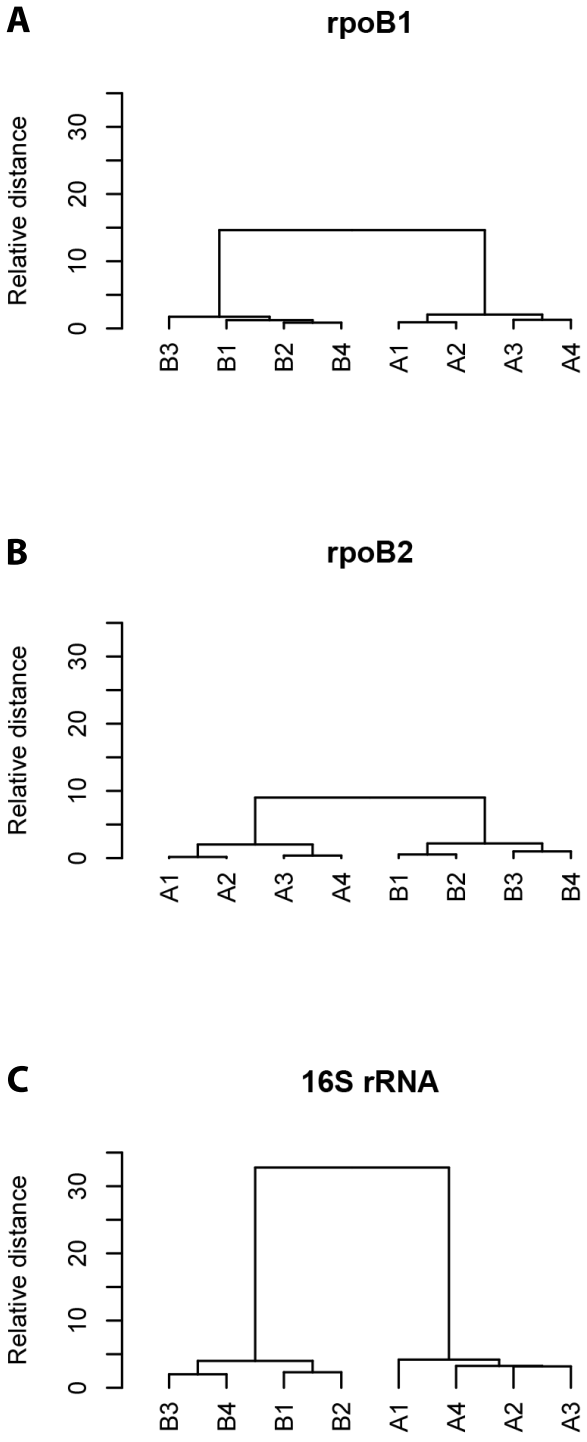

Figure 6.5: Hierarchical clustering of both experiments combined, per target - the relative distance corresponds to the distance between two individuals (A and B) calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.01 from a t-test between the samples from each individual or a BF <1 were used.

Target	no. significant OTUs in common with combined	% significant OTUs in common with combined	% of combined significant OTUs	% of significant OTUs from both exps in common with combined
Experiment 1				**Combined**
***RpoB*1**	11 / 0 / 5	27.5 / 0 / 12	57.9 / 0 / 62.5	79 / 0 / 75
***RpoB*2**	3 / 1 / 1	37.5 / 20 / 20	30 / 100 / 100	50 / 100 / 100
16S rRNA	80 / 12 / 48	42.1 / 44.4 / 25.3	50.3 / 17.6 / 70.6	71.1 / 30.9 / 86.8
Experiment 2				
***RpoB*1**	9 / 0 / 5	12.5 / 0 / 7	47.4 / 0 / 62.5	
***RpoB*2**	3 / 0 / 1	20 / 0 / 6.7	30 / 0 / 100	
16S rRNA	85 / 13 / 48	23 / 21.7 / 12.9	53.5 / 19.1 / 70.6	

Table 6.4: Comparison of significant OTUs between the combined and individual experiments - This table compares the number of significant OTUs found in both an individual experiment and both experiments combined. Significant OTUs are those with a p-value <0.1 or <0.01 (or BF <1) from a t-test between the samples from each individual. The first and second numbers correspond to both individual and combined experiments filtered at a p-value <0.1 and <0.01, respectively. The third number corresponds to the individual experiments filtered at a p-value <0.1 and the combined experiments to a p-value <0.01.

6.5.1 Updated database

All BLAST analyses for this thesis were performed using the same database to standardise the analysis between experiments and avoid any differences being put down to the use of more than one database. However, the real world application of this technique must be taken into consideration. Realistically databases get updated and therefore for this technique to be usable it is required to function with different databases. For example, if a suspect reference sample was analysed for one case and subsequently this person was implicated in another case then it is highly probable that the reference sample would be analysed with a different database to a trace analysed in the second case. Therefore the technique is required to be robust enough to link samples from the same individual analysed with different databases. To test the robustness of this technique, experiment one was re-analysed using the most up-to-date nucleotide database available.

Table 6.5 shows the comparison of species-level OTUs for experiment one analysed with both the old database (database 1) and the new database (database 2), (see supplementary spreadsheet 'new database ' (https://independent.academia.edu/SarahLeake/Papers) for raw and processed data). For both *rpoB*2 and 16S rRNA the total number of sequences assigned to a taxon are the same and for *rpoB*1 the difference is negligible indicating that using a newer database does not increase the number of sequences assigned to a taxon, however in the future this could change. At the species-level a high proportion of the OTUs have identical species assignment and abundances for all targets (see Table 6.5). However, inevitably a certain number of the OTUs are different. This is most apparent for 16S rRNA as these sequences are the most likely ones to change and be updated and therefore OTUs which were assigned to one taxon in the first database could be assigned to a different one in the new database. This can also be seen in the total OTU count where for 16S rRNA there is a decrease of 45 (see Table 6.5), which is likely due to some sequences in the first database being reassigned to the same taxon, hence decreasing the total number of OTUs. For both *rpoB* targets the numbers are stable indicating that there are few differences between the two databases.

To compare how well the databases separate the samples they were clustered using the Ward method of hierarchical clustering (see section 3.5.3.1) and the relative distances between individuals are shown in table 6.5. For both *rpoB* targets the distance is identical and for 16S rRNA the difference is very small, implying that changing database does not affect the ability of this method to separate individuals. To further check this, experiment one analysed with the up-to-date database was combined with experiment two analysed with the old database and hierarchical clustering performed (see Figure 6.6). This also

emulates the real life situation mentionned above. The distances were compared to those from the hierarchical clustering of both experiments analysed with the old database. For both *rpoB* targets the relative distance between individuals (Figure 6.6A and B) is very similar to the respective distance in the other analysis (Figure 6.6B and D), with the latter providing slightly better separation. 16S rRNA follows the same pattern with the samples analysed with the same database producing better separation (Figure 6.6E) however the difference between the distances is larger. For all targets the intra-individual separation is very similar indicating that changing the database has very little effect on how the samples are separated.

Target	total sequences	no hits	identical OTUs	identical abundance OTUs	identical species OTUs	different OTUs	total OTUs	hierarchical clustering distance
Database 1								
***rpoB*1**	29560091	24838	90	5	30	25	150	21.53
***rpoB*2**	8723851	14730	26	3	0	5	34	15.26
16S rRNA	21270245	0	576	8	58	203	845	44.01
Database 2								
***rpoB*1**	29559993	11338	90	5	30	28	153	21.53
***rpoB*2**	8723851	14730	26	3	0	5	34	15.26
16S rRNA	21270245	0	576	8	58	158	800	44.99

Table 6.5: Comparison between species-level OTUs from experiment one analysed with two different BLAST databases for all three targets. Database 1 is the old database and database 2 the new one. Identical OTUs are those with the same taxon assignment and abundance, whereas identical abundance OTUs have the same abundance but a different taxon assignment. Identical species OTUs are assigned to the same species but have different abundances. The hierarchical clustering distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the logged species abundance. Only species with a p-value <0.1 from a t-test between the samples from each individual were used.

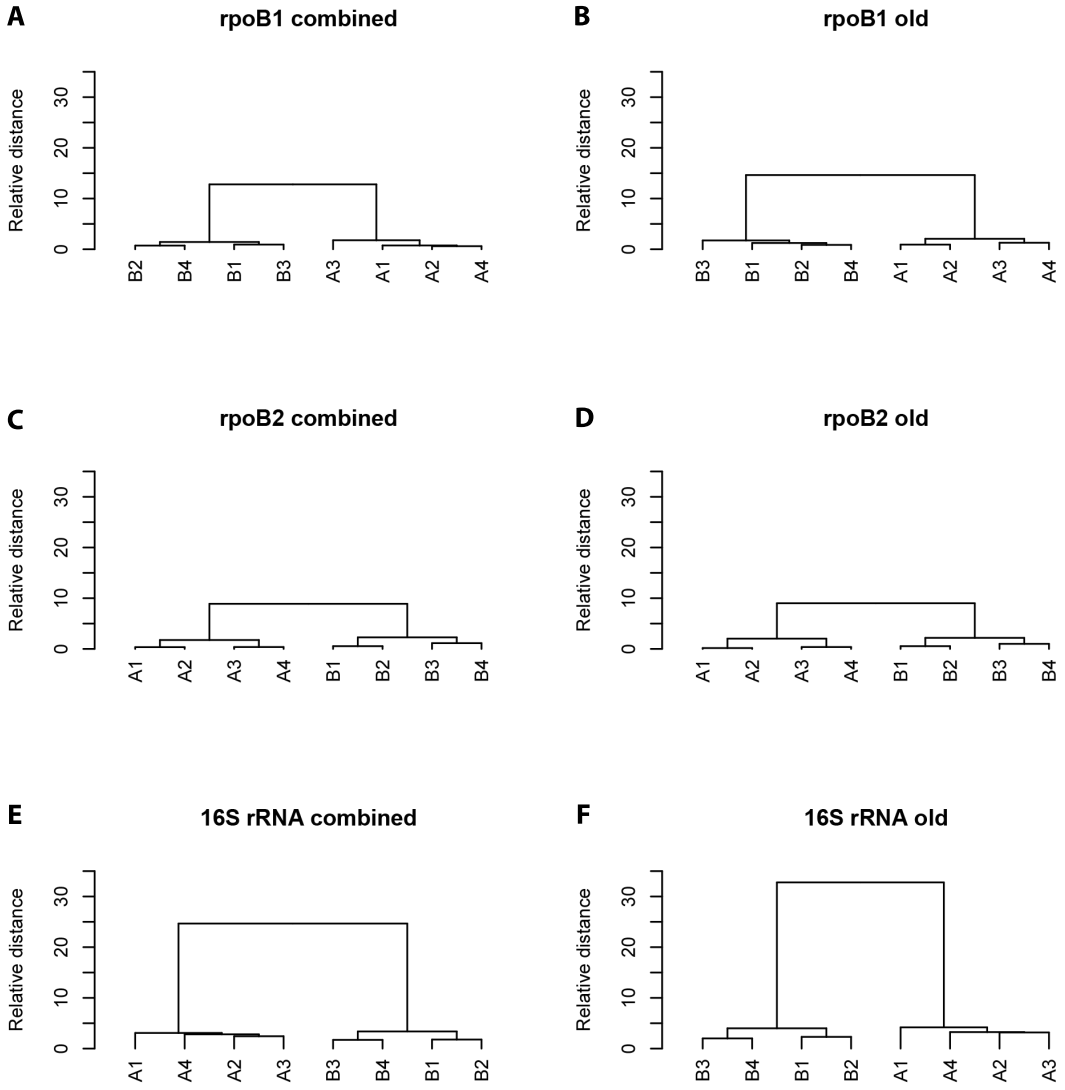

Figure 6.6: Hierarchical clustering of all eight samples with different databases for each target - A, C and E represent the clustering of experiment 1 analysed with the new database and experiment 2 analysed with the old database. B, D and F represent the clustering of both experiments with the old database. The relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.01 from a t-test between the samples from each individual or a BF <1 were used.

6.5.2 Combination of target genes

Section 6.5 supports that when both experiments are combined and analysed together, samples from one individual can be separated from those of another whilst maintaining low intra-individual variation. The next thing to look at is whether this separation can be improved by combining the target genes. Figure 6.7 shows the separation of individuals in terms of relative distance, calculated by hierarchical clustering, for each target gene individually and all permutations of combined target genes at different t-test p-values. Different p-values were analysed to check whether by combining the target genes a different p-value produced the greatest separation. Firstly, a p-value of 0.01 provides the greatest separation, justifying this choice of p-value for the combined experiments (see section 6.2.1). *rpoB*2 and *rpoB*1 produce the smallest and second smallest relative distances, respectively, correlating with the number of significant OTUs per target (see Table 6.3). The combination of both *rpoB* target genes produces a greater separation than either target gene individually however, the separation is still much lower than for 16S rRNA alone or combined. The combination of all three target genes produces the greatest separation regardless of the p-value used to filter the data. However, the difference between this and the combination of 16S rRNA with either *rpoB* target gene is not very big, implying that the majority of the separation comes from 16S rRNA. This is confirmed by 16S rRNA alone which falls just below the combination of 16S rRNA with either *rpoB* target genes.

From the hierarchical clustering (see Figure 6.8) the percent of intra-individual variation can be calculated for each combination; all - 14.5% (Figure 6.8A), *rpoB*1 + 16S rRNA - 14.2% (Figure 6.8B), *rpoB*2 + 16S rRNA - 14.9% (Figure 6.8C) and *rpoB*1 + *rpoB*2 - 16.9% (Figure 6.8D). The best separation is one which minimises intra-individual variation and maximises inter-individual variation. The smallest intra-individual variation is acheived by combining *rpoB*1 with 16S rRNA, a combination which also produces the second greatest inter-individual variation. For the combination of all target genes the intra-individual variation is third highest indicating that the addition of *rpoB*2 increases the intra-individual variation. These values must be interpreted with caution due to the small sample number, however they can still give an indication of which combination of target genes is the most effective at separating samples from different individuals.

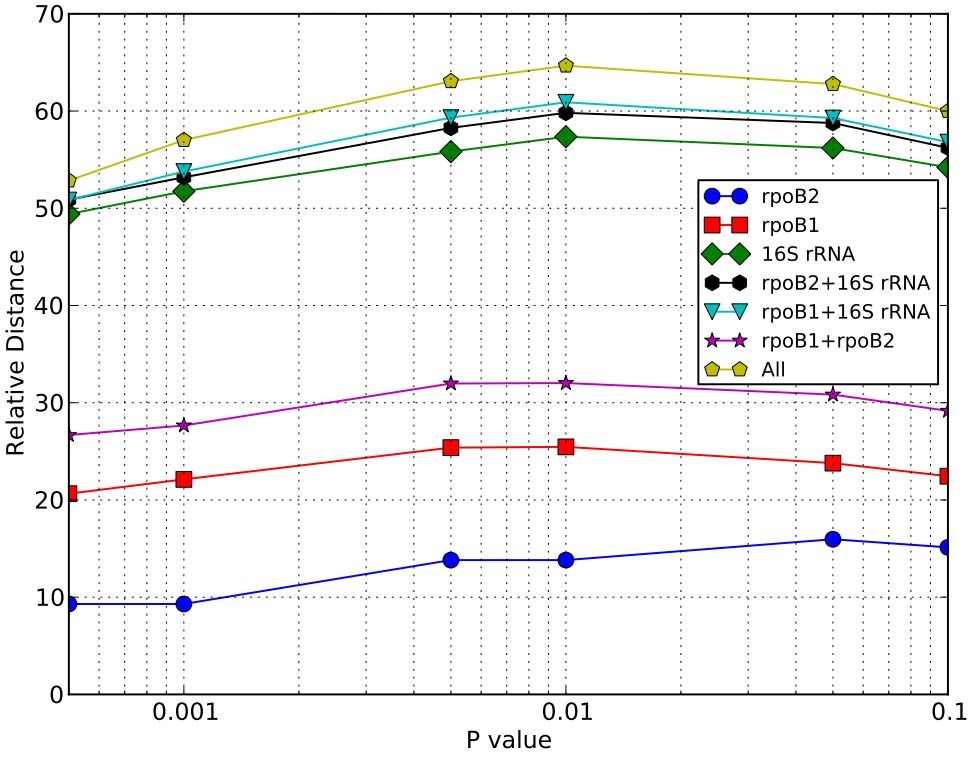

Figure 6.7: Hierarchical clustering of individual and combined target genes at different t-test p-values used for data filtering - the relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance.

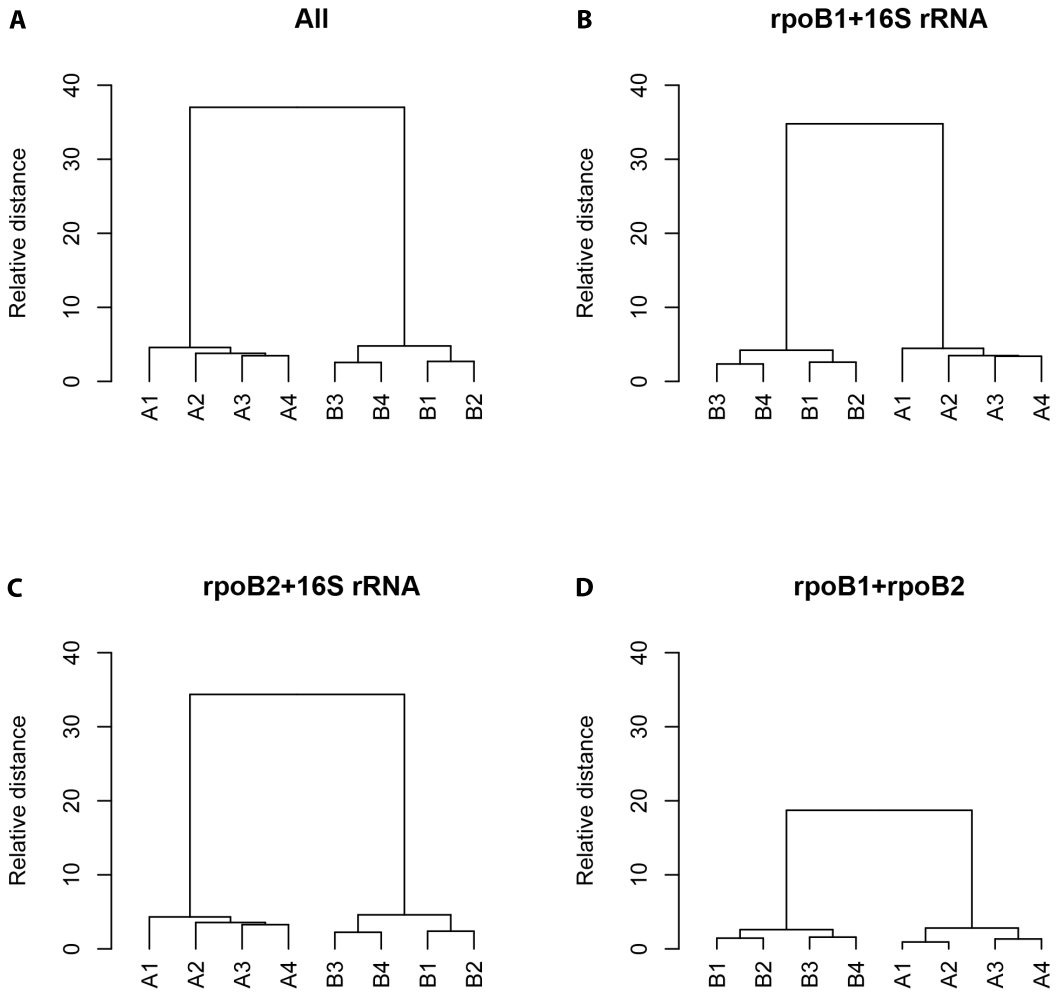

Figure 6.8: Hierarchical clustering for all combinations of target genes - the relative distance corresponds to the distance between two individuals (A and B) calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.01 from a t-test between the samples from each individual or a BF <1 were used.

6.6 Minimum sequences required

This study used the HiSeq2000 to analyse the samples, a machine which can produce over one billion reads, as at the outset of this study the number of sequences required to separate two individuals was unknown. To calculate the minimum number of sequences necessary the data from experiment one were randomly sub-sampled at different levels: 1000, 10000, 50000, 100000, 500000 and 1000000 sequences (see supplementary spreadsheet

'minimum sequences ' (https://independent.academia.edu/SarahLeake/Papers) for raw and processed data). The analysis was performed to the end (from OTU clustering to hierarchical clustering) and the relative distances calculated between the samples at all levels are shown in Figure 6.9. For *rpoB*2 and All there are no points before 50000 as separation was not achieved, *rpoB*2 also produces the smallest separation. 16S rRNA provides the best separation when looking at the targets individually. However, when 16S rRNA and *rpoB*1 are combined the separation is improved. Combining all three targets produces the best separation, however the addition of *rpoB*2 does not greatly improve the separation except at 50000 sequences where the separation is significantly improved. This correlates with the results presented above in section 6.5.2 confirming the best and most efficient separation is achieved by combining 16S rRNA and *rpoB*1.

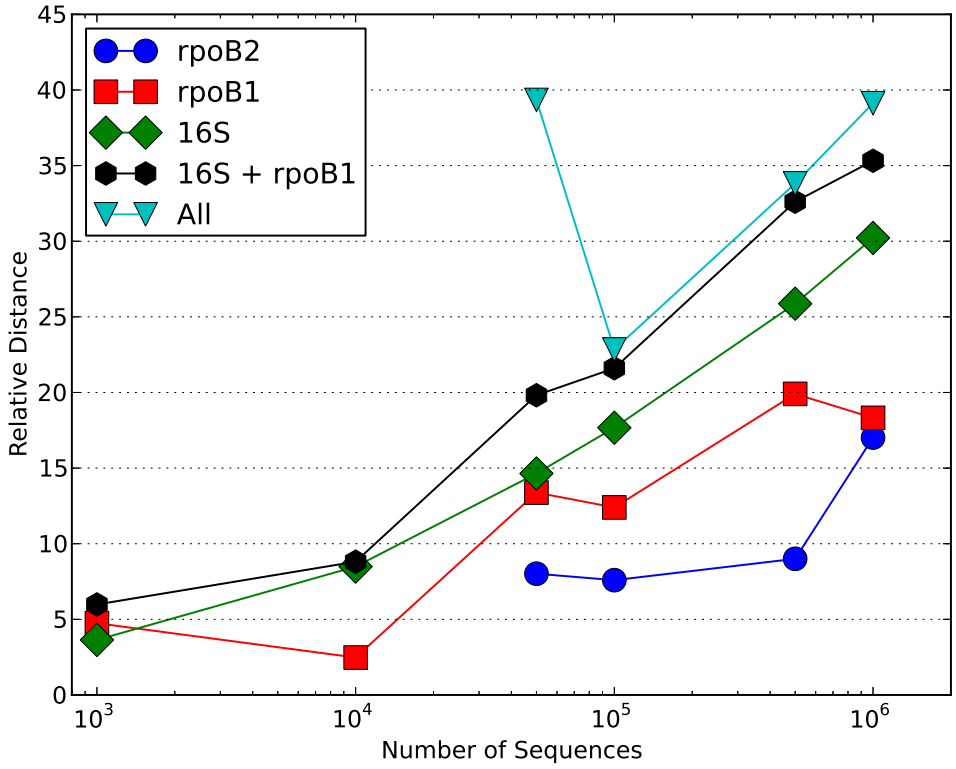

Figure 6.9: Number of sequences required for sample separation - the relative distance corresponds to the distance between two individuals calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.1 from a t-test between the samples from each individual or a BF < 1 were used.

Figure 6.9 suggests that to ensure the separation of individuals a minimum number of sequences of 100,000 would be best. From this value it is possible to calculate how many

samples could be analysed in one sequencing run. This is important as the economic value of a technique must be considered before it can be integrated into routine analysis. The sequencer used for this study was the Hiseq 2000 which is a very high-throughput machine producing up to two billion reads and taking around 10 days for one analysis. This machine would not be suitable for a forensic laboratory due to both cost and time of analysis. However, Illumina have produced a lower throughput benchtop sequencer called the Miseq which produces on average 25 million reads after quality filtering. The run time for a Miseq varies between 5 to 65 hours, which is significantly faster than the Hiseq, making it more suitable for forensic applications.

no. target genes	**min no. sequences**	**no. samples analysed Hiseq**	**no. samples analysed Miseq**
1	100,000	595	106
2	200,000	297	53
3	300,000	198	35
1	500,000	119	21
2	1,000,000	59	10
3	1,500,000	39	7

Table 6.6: Comparison of potential number of samples analysed per number of sequences for the Hiseq and Miseq

Table 6.6 compares the potential number of samples which could be analysed by both the Hiseq and Miseq. For the type of sample/analysis performed in this study, on average, the Hiseq produces 200 million reads. About 30% of reads are removed through quality filtering leaving 140 million reads. As paired-end sequencing is used 2 reads = 1 sequence therefore 140 million reads gets reduced to 70 million sequences. Due to read pairing and barcode splitting a further 15% of reads are lost leaving 59.5 million sequences. By using this as the starting value the potential number of samples which could be analysed can be calculated (see Table 6.6). This number varies depending on how many targets are analysed and the minimum number of sequences required. The same procedure is applied to the Miseq producing a start value of 10.6 million sequences. As mentioned above the combination of two targets, *rpoB*1 and 16S rRNA, provides the best results therefore, with 100,000 sequences per target 53 samples could be analysed using the Miseq and 297 with the Hiseq. This shows that the Miseq is a realistic option for analysing forensic samples. Even if a higher coverage of half a million sequences is required, with two targets, 10 samples could be analysed, which is the equivalent of 5 traces and 5 reference samples. For identification purposes this technique will only work with a reference sample therefore, this needs to be taken into consideration when planning how many samples can be sequenced together.

Another application of this technique; providing intelligence (linking specific bacteria to lifestyle traits), would be one reason for increasing the coverage as more species would be obtained, increasing the information provided by each sample. If the coverage is increased too high then not enough samples could be sequenced in one run making the technique less affordable. Table 6.7 shows the number of OTUs per target at 50,000, 100,000 and 500,000 sequences. As the number of sequences increases as does the number of OTUs and the number of significant OTUs. For intelligence purposes it is only the number of total OTUs that is important as they are being used to inform about a particular sample not be used for comparing two different samples. If it is known that the sample is only going to be analysed to provide intelligence and not be used for differentiating individuals, then a higher sequence coverage would be advised. Deciding the sequence coverage to use would also depend on the demand on the machine, if only a few samples needed to be sequenced then a higher coverage would always provide more information.

Target	**no. OTUs**	**no. significant OTUs**
50,000		
***rpoB*1**	33	14
***rpoB*2**	5	1
16S rRNA	80	23
100,000		
***rpoB*1**	41	17
***rpoB*2**	8	2
16S rRNA	113	33
500,000		
***rpoB*1**	66	27
***rpoB*2**	20	3
16S rRNA	256	87

Table 6.7: Comparison of OTUs between all targets for experiment 1 at different sequence coverage - significant OTUs are those with a p-value <0.1 from a t-test between the samples from each individual or a BF <1.

To check the robustness of the proposed p-value and stability of intra-individual variation for the two experiments combined, both experiments sub-sampled at 100,000 sequences were combined and analysed. Due to the low number of OTUs for *rpoB*2 (see Table 6.7) this target was not included in the analysis and only the combination of 16S rRNA and *rpoB*1 is presented as this has already been chosen as the best combination (see section 6.5.2). Figure 6.10A and B shows that *rpoB*1 and 16S rRNA sub-sampled at 100,000 sequences and filtered at a p-value of 0.01 can separate the samples from two different individuals whilst minimising the intra-individual variation, 15.4% and 13.4% respectively. The combination of the two target genes provides increased separation between individuals

and lower intra-individual variation (11.9%) (see Figure 6.10C). These results corroborate those found by both the individual and combined experiments. In fact, the intra-individual variation of 11.9% is the lowest achieved by any combination observed thus far, indicating that a coverage of 100,000 sequences is optimal for separating samples from different individuals.

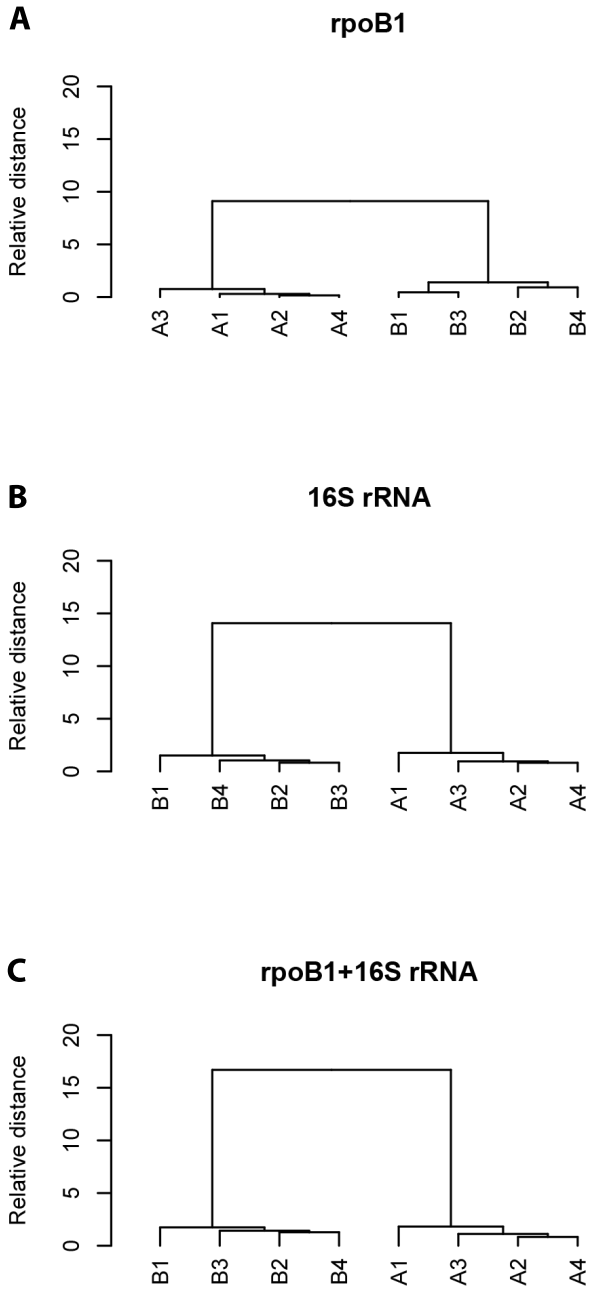

Figure 6.10: Hierarchical clustering of both experiments combined and sub-sampled at 100,000 sequences for *rpoB*1, 16S rRNA and the two target genes combined - the relative distance corresponds to the distance between two individuals (A and B) calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.01 from a t-test between the samples from each individual or a BF <1 were used.

7

Discussion

This chapter discusses each of the results chapters in order, starting with chapter 4 and ending with chapter 6. Subsequently, the factors which can influence the salivary microbiome are discussed along with ethical considerations and the scientific and forensic relevance of this thesis. Finally, future work is proposed in order to extend and develop the work presented in this thesis.

7.1 Method optimisation

The first aim of my thesis was to develop a method for analysing the salivary microbiome in order to see whether it was possible to differentiate two people based on the bacteria present in their saliva. For the most part this was achieved, however due to budget restraints not all parts could be optimised.

7.1.1 Sampling and extraction methods

Neither the sampling method nor the extraction method could be optimised due to budget restraints. In order to test the effect of different sampling and extraction methods the analysis would have to have been carried out in full, including sequencing and unfortunately the budget available did not permit this. The number of sequences required to differentiate two individuals was unknown and therefore the number of samples sequenced in one run was limited to four in order to have a large number of sequences per sample. This meant that in total only eight samples could be analysed. If the minimum number of sequences was known then more samples could have been sequenced at the same time and more parameters such as sampling/extraction method studied.

It was also unknown whether it was even possible to differentiate two people based on the bacteria present in their saliva, therefore it was important to demonstrate this before attempting to optimise the sampling/extraction methods. In order to try and prove this, saliva was sampled by spitting into a tube ensuring enough sample was present for sequencing. The samples were collected under identical conditions for both participants at all time points, in order to minimise any factors which could effect the amount and composition of saliva (see section 2.1). The goal was to have a new technique for human identification, therefore this was always considered when designing the method. However, the technique had to be proven to work before real casework type samples could be tested. The standard method in forensic science for sampling dried saliva on human skin is to use a cotton swab dampened with sterile water followed by a second dry swab (155). This method should still be applicable for sampling saliva for bacterial DNA analysis as both the water and cotton swabs are sterile and therefore, they should be free from contamination. Avoiding contamination is extremely important in forensic science as the results can be used in a court of law and therefore must be robust. As this technique would rely on reference samples to match the bacterial profile from a trace to a person, the method of sampling used in this thesis could be used for the reference samples. Reference samples are taken directly from the person, therefore spitting into a tube would provide adequate saliva to produce an accurate reference sample.

The samples were extracted using the automated MagNA Pure 96 DNA system to limit human error and minimise possible contamination. Many extraction methods exist for extracting bacterial DNA and a recent study concentrating on extracting bacterial DNA from oral samples has shown that results differ depending on the extraction method used (156). Four different extraction methods were tested, two of which are commonly used in oral microbiome studies (chemical/enzymatic lysis + DNeasy blood and tissue kit (Qiagen) and crude chemical/enzymatic lysis). They show that protocols which included lysozyme produced better results as it increased nucleic acid recovery from gram-positive bacteria by disrupting peptidoglycans in cell walls. However, no technique was without bias and therefore consistency is required between experiments. Otherwise, problems will arise when comparing results from experiments which use different extraction methods. Therefore, if this technique were to be implemented into the forensic analysis pipeline, all laboratories would be required to use the same extraction protocol. A study by Edelmann *et al.* tested the performance of the MagNA Pure system in a clinical setting and demonstrated that it produced reliable and reproducible results (157) and would hence be suitable for forensic analysis. However, for a technique to be used with real case samples it needs to be validated. Therefore, it would be interesting to test whether any already

validated DNA extraction techniques could be applied to bacterial DNA. Furthermore, to exploit a trace fully, it would be beneficial if human DNA and bacterial DNA could be co-extracted, enabling the production of a human DNA profile and a bacterial profile. Most methods of bacterial DNA extraction will also extract human DNA. Therefore, it would be interesting to see whether the amount of human DNA extracted is sufficient to produce a human DNA profile.

7.1.2 Target selection

Deciding which genes to target was important as different genes provide different levels of identification. 16S rRNA can be described as the standard gene used for molecular taxonomic assignment for bacteria. Therefore, many primers already exist covering one or more of the variable regions, along with databases specifically containing 16S rRNA sequences making it an essential target gene. However, as described in section 3.2.1 there are disadvantages to targeting 16S rRNA alone. Rajendhran and Gunasekaran (158) discuss these disadvantages, of which the principle ones are; intra-genomic heterogeneity, mosaicism and the lack of a universal threshold sequence identity value. They demonstrate the advantage of targeting more than one gene. Another problem with 16S rRNA is the presence of partial sequences in databases, which lead to ambiguous classification of sequences (159). This needs to be taken into consideration when assigning taxonomy to sequences. Whilst these problems exist, the aim of this thesis is not to comprehensively characterise the salivary microbiome but to use the bacteria found as a means of distinguishing two people. Therefore, as long as the same database is used for all samples, even if there is an annotation error that error would be applied to all samples and hence not cause a problem. The same is true for the other disadvantages, for example, if a species is known to show gene heterogeneity then it is likely to do so in all samples and as the abundances are compared on a species by species basis this should not pose too much of a problem. The only potential problem could be from genes which show divergent sequences for the same species and therefore the copy number for that species could vary between samples (160). This reiterates the need for a single copy target.

rpoB was chosen as the single copy target because it has already been proven to successfully identify many different bacterial species (91, 92). A description of *rpoB* can be found in section 3.2.1. One problem with *rpoB* is that, in comparison with 16S rRNA, it is less conserved and therefore, more than one pair of primers is required to cover the same bacterial population. However, this can also be seen as an advantage as the *rpoB* primers can be designed to amplify bacteria known to be found in a specific sample type. In this

case, one of the *rpoB* primer sets was designed to target *Streptococcus*, one of the main genera found in saliva. *rpoB* also has higher resolution enabling identification down to the strain level (89), something impossible with 16S rRNA. A recent study by Kraal *et al.* showed that using 16S rRNA alone can obscure strain-level detail (161). Previous studies have also shown that humans have many different strains of the same *Streptococcus* species with many strains being unique to individuals (60, 61). This confirms the importance of targeting *rpoB*. A second problem is that no specialised databases exist for *rpoB*, so large databases such as the NCBI nucleotide database must be used. A disadvantage of this is that the sequences might be less curated however, as discussed above if partial sequences are used then other problems arise. In this case, two different target genes were used and to keep the analysis as homogeneous as possible the same database was used for both and therefore a specialised database would not have been suitable. A recent study by Vos *et al.* compares the use of 16S rRNA and *rpoB* as markers in studies of bacterial diversity (152). They show that both markers give similar total diversity estimates, however *rpoB* reveals more species and requires less sequences to obtain 90% of the true diversity. This implies that *rpoB* is a more efficient marker when targeting specific species, however if a general overview is required with only one primer pair then 16S rRNA is better. Furthermore, by combining the two targets a more complete view of a microbial population can be achieved.

7.1.3 Primer design and optimisation

The pre-amplification of DNA with specific primers is important for two reasons:

1. to amplify specific regions capable of identifying which bacteria are present
2. to have enough DNA for sequencing

In order to amplify the DNA, primers specific to the chosen targets were designed and optimised, without which identification of bacteria present in the samples would be very difficult. Point two is especially important in forensic science as the average concentration of DNA found in forensic traces is quite low and hence amplification is required to be able to further analyse the samples. Chandler *et al.* (162) demonstrated that the quality and quantity of DNA influences the microbial community structure, for this reason it is important to optimise the primers to ensure as much of the sample is amplified as efficiently as possible. Further work is required to analyse forensic traces to see whether enough DNA is available to give an adequate representation of the microbial communities present.

Concerning 16S and 23S rRNA designing primers to amplify all bacterial species present in a saliva sample is impossible due to natural genetic variation present in the conserved regions. To design the best primers possible species representing the four principle expected phyla were used for initial primer design, followed by a check against species known to be found in saliva. This enabled primers to be designed which covered a large variety of bacteria whilst ensuring saliva specific bacteria were amplified. However, for 23S rRNA this was not the case as only four out of ten of the saliva specific species could be aligned. This shows that even though the initial alignment provided suitable regions for primer design that this does not always translate to a good alignment with more specific species. As shown by the virtual simulation of primers 16S rRNA amplified very similar species to 23S rRNA and therefore 23S rRNA was redundant. Had this not been the case the 23S rRNA primers would have been redesigned.

The *rpoB* primers were designed using two different species sets, however the results show that they amplified more than the designed species. This is to be expected as not all species could be covered during primer design, Vos *et al.* found the same when using *rpoB* (152).

All primer pairs were blasted against the human genome to check whether human DNA would be amplified. This is important as saliva samples contain a lot of human DNA and if this was to be amplified the amplification of bacterial DNA would be impeded due to the preferential amplification of human DNA. Results show that only 1% of sequences could be associated to the genera *Homo* and *Pan*. *Pan* is included as these sequences do not naturally occur in saliva but must be homologous to *Homo* sequences which is why occasionally they are assigned to *Pan*. Secondly, all primer pairs were blasted against the NCBI nucleotide database to test for specificity to target region. This is important to ensure the efficiency of the reaction, if the primers were not specific then other genes sequences may get amplified creating noise for the desired target sequences. They would also use up reagents in the reaction meaning less target sequences could be amplified. Thirdly, all primer pairs were compared against the HOMD to ensure that they would amplify taxa which have already been associated to the oral microbiome. The chosen 16S rRNA primer pair amplified about 50% of the genera in the HOMD demonstrating that one primer pair is not sufficient to cover all potential genera.

The virtual simulation of all primers using the nt database as the sample showed that both 16S rRNA primers and the 23S rRNA primers reveal very similar phylum level classification, indicating that only one of these primer pairs is required. This demonstrates

the benefit of testing the primers *in silico*, otherwise it would not be known before carrying out the full experiment that these three primers all amplify very similar species and therefore it is not necessary to use all three. From these results 16S_1 would be the chosen primer pair as Bacteroidetes are more present and previous results showed that 23S rRNA primers only amplified 6% of possible bacteria (see section 4.2). The results for the *rpoB* primers show the advantage of targeting more than one area of the gene and as *rpoB* has a higher resolution the bacteria can be identified down to the species/strain level. When the primer pairs were compared against the HOMD to see how many of the genera amplified have already been found in the oral microbiome about 50% were amplified by the designed 16S rRNA/23S rRNA primers. This figure is a lot lower for *rpoB*, as *rpoB* does not target as many genera. However, of the genera potentially amplified by the *rpoB* primers, about 35% of them are found in HOMD. As these primers can classify to the species/strain level, they should still amplify a good number of bacteria.

Primer optimisation is important to ensure the reaction is as specific and efficient as possible. A temperature gradient was used to find the best annealing temperature for each primer pair. For all primer pairs the target band is around 100bp, the other distinct bands on the gels correspond to non-specific amplification and the bands at the top of the gels to high molecular weight DNA. The darker the band the greater the quantity of DNA amplified at that size. An ideal amplification would show one dark band for the target region and no other bands, however this is not always possible. Now, there is a technique which enables DNA to be purified from a band in a gel, avoiding the problem of non-specific amplification. However, non-specific amplification still uses reagents which would otherwise be used to amplify the target region therefore, when choosing the best annealing temperature both band colour and number of bands are taken into consideration. Finally to check the chosen primers with the chosen annealing temperature, saliva samples were amplified. All samples show a distinct band at around 100bp indicating that the target region was amplified however, some non-specific amplification is present. Therefore, to have a pure sample containing only the target region the samples are required to be excised from the gel and purified.

To improve the streamlined analysis of samples multiplexing of primers would be desired. However, this would require the primers to be optimised together as only one annealing temperature could be used. This will not be an easy task as each primer pair has different annealing temperatures ranging from 64.3 °C to 56 °C. As this range is quite large, trying to find one annealing temperature will likely involve sacrificing some specificity. If too much specificity is lost then it would be better to amplify each primer set separately and

pool afterwards. The only problem with this for forensic samples is the amount of DNA required. With multiplexed primers only a certain amount of DNA is required, however if each primer pair is amplified separately three times as much DNA will be required. Forensic samples are known to not always contain much DNA and therefore the sample may not contain enough to amplify three different targets. This will be discussed further in section 7.6.

The one major drawback of PCR amplification is the introduction of errors. If an error is introduced in the first few cycles then it will be amplified along with all the other sequences and subsequently be sequenced. If an error occurs in the later cycles it will have less of an impact as not as many sequences containing the error will be produced. One specific type of error is the production of chimeric sequences, this happens when incomplete PCR products act as primers amplifying related fragments (163). Chimeras have been found to make up between 5 to 45% of a sample (164) and could therefore pose quite a big problem. During data analysis the data was split up into each individual target by matching the primer sequence. To reduce chimeras, sequences were only kept that matched the primer 100%. However, this does not avoid the problem of part of the sequence containing the wrong base. To combat this a high fidelity polymerase is used, in this case the Phusion® Hot Start II, as it is known to reduce amplification error. Furthermore, as discussed above, amplification is very important and cannot be avoided therefore, measures can only be taken to minimise errors. A recent study has proposed shotgun metagenomic sequencing as an alternative to target amplification followed by sequencing as this technique fragments the extracted DNA and directly uses this to create the libraries, hence removing the pre-amplification step (165). Their results show very similar bacterial composition to previous studies, however for this application DNA quantity is an issue and therefore pre-amplification is required to produce enough DNA for sequencing. Nevertheless, these techniques are developing rapidly and library preparation can now be performed with as little as 0.01ng of DNA (166), making this a possible option for a forensic application.

7.1.4 Sequencing method

The decision was taken to use Illumina sequencing over 454 due to the shorter read length enabling a greater depth of sequencing and the capability of paired-end sequencing producing high confidence consensus sequences. Furthermore, there are errors associated with 454 sequencing such as problems with reading homopolymers and chimeras which are less apparent with Illumina.

An article published by Chengwei *et al.* in 2012 compared the bacterial composition of the same sample using both 454 and Illumina sequencing and found that overall the results were comparable. About 90% of the 454 unique contig sequences overlapped with the Illumina contig sequences and concerning gene/genome abundance both techniques provided similar estimates. However, Illumina produced longer and more accurate contigs. They found more sequencing errors with 454 especially coming from A- and T-rich homopolymers, however Illumina did produce some homopolymer and non-homopolymer associated sequencing errors (167). This article shows that whilst 454 and Illumina can produce similar results, there are less sequencing errors with Illumina and combined with the much greater number of sequences produced it is the more favourable technique for metagenomic studies. In terms of costs for the large high-throughput machines the Illumina HiSeq 2000 costs \$0.07/Mb whereas the 454 GS FLX costs \$10/Mb (168), for the benchtop sequencers the Illumina MiSeq costs \$0.5/Mb whereas the 454 GS Junior costs \$31/Mb (169), further justifying the choice of Illumina.

As mentioned above errors occur during the amplification process a problem which also effects sequencing as the first step uses amplification to attach the library adaptors. For Illumina chimeras may be less frequent due to a shorter amplicon length when compared to 454 sequencing.

Regardless of the method chosen as only clusters containing 20 or more sequences were kept for further analysis it is highly unlikely that any sequencing errors would be found in the filtered dataset. The sequences were initially quality tested and any that did not pass were removed. Therefore, most errors should have been removed before clustering into OTUs and those that were not would be removed in the subsequent step (removing clusters containing less than 20 sequences). An advantage to choosing a method which produces less errors is that more sequences should pass the quality control and hence be available for analysis. As exactly the same methodology was applied to all samples, regardless of what errors can occur, the conclusions drawn are still valid as all samples should be affected in the same manner.

7.2 Characterisation of the salivary microbiome

The second aim of my thesis was to characterise the microbiome of two individuals to see which bacteria were present and in what abundance.

7.2.1 OTU clustering and BLAST

As described in chapter three, the sequences were processed in order to identify which bacterial species were present in the sample (see section 3.4). When I started this project the goal was to have a technique which could be used in a standard forensic laboratory, therefore when I was deciding which programs to use to process the data, this was taken into consideration. Scientists who work in forensic laboratories do not necessarily have bioinformatics training therefore the programs used needed to be easy to implement. Another important factor with forensic analysis is time. The faster a result can be produced the better as suspects can only be held for a certain amount of time after which, if no evidence is produced, they are released. The programs chosen for quality filtering, read pairing and barcode splitting were simple as they perform basic tasks and therefore do not take much time. The choice of clustering algorithm was more complicated as there are many options available (see section 3.4.4). CD-HIT was chosen for it's speed and ease of use. A recent study by Chen *et al.* compared methods for clustering 16S rRNA sequences into OTUs. They showed that CD-HIT inferred the true (or closest to true) number of OTUs and outperformed hierarchical clustering algorithms (135). This article demonstrates that as well as being fast CD-HIT is also accurate and a good choice for clustering sequences into OTUs. However, there are a couple of disadvantages to this method, the first comes from the short word filtering (170). If the mismatches are evenly spread across the sequences then the number of k-mers (words) can be artificially low. However, real biological sequences tend to have motifs so evenly distributed mismatches are rare. The second problem comes from the greedy incremental algorithm which compares the sequences by length with the longest sequence first. Therefore, if two sequences pass the threshold with the longest sequence then they will be grouped with that one regardless of whether they actually match a shorter sequence better (170). This could explain why more than one cluster is assigned to the same taxon. To overcome this problem all clusters assigned to the same taxon were combined to give a more accurate estimate of OTU abundance. However, this raises another concern; errors in the database used for taxon assignment. A second method for clustering sequences, known as de novo clustering, could help overcome the above-mentioned problems with OTU clustering. De novo clustering clusters reads against each other without using a reference database. However, this technique is very time consuming and not suitable for large datasets. Therefore, due to this constraint, this technique would not be suitable for a forensic application as speed of analysis is very important.

Sequences in the NCBI nucleotide (nt) database are manually inputted hence errors can occur impacting upon taxon assignment. However, the use of databases cannot be avoided when assigning taxa to a large number of sequences. To combat this problem the same version of the nt database was used for all samples. Therefore, an incorrect taxon assignment would be used for all samples and therefore the abundance comparison between the two individuals would still be valid. Errors will impact upon the bacterial characterisation of the samples leading to under or over representation of certain taxa. The number of errors in the nt database is kept as low as possible by containing some curated sequences (refseq (171)) and relying on the user to supply accurate sequences and report any errors (140). However, due to the number of sequences it is impossible for the NCBI to curate them all. As two target genes were used, to standardise the analysis a database which contains sequences for both targets was required, hence the use of the nt database. If a manually compiled database specific to 16S rRNA and *rpoB* was used then it would have to be manually updated, a task which would take a lot of time, whereas the nt database is constantly updated. As shown in section 6.5.1 using a more up to date database does not affect the separation of individuals and hence samples analysed with different databases can still be compared. This is important as in real casework it is not uncommon for suspects to be involved in more than one crime and therefore due to a time gap between analyses different databases could be used for taxon assignment. In terms of characterising which bacteria are present it is important to use the most up to date database as it will contain the most current knowledge. As more studies are performed these targets get better characterised and the sequences in the databases should be more accurate and previously uncharacterised sequences get characterised. Other than possible errors the only other problem with using the nt database is the time required to perform the BLAST. The nt database is very large and therefore depending on how many sequences are to be compared the analysis can take about 1 week, which was the case for 16S rRNA. One way to combat this problem is to split the data into smaller chunks and analyse each chunk separately. It is here that having a lot of computing power is useful as the BLASTs can be run in parallel speeding up analysis.

7.2.2 OTUs

The analysis of the samples in terms of species-level OTUs enables an overview of how many OTUs are in common with samples from one individual and both individuals combined. As described in section 5.2 99% of the sequences are filtered out, which seems very high. However, stringent filtering has been used to try and ensure no errors of any kind

have been included, this includes removing all sequences which appear 19 times or less, of which singletons make up a large proportion. Even with such stringent filtering a decent number of OTUs are kept showing that a lot of information can be obtained with a low percentage of sequences.

For 16S rRNA and *rpoB*1 a very high percentage of OTUs are found in all samples indicating that most differences between individuals comes from variation in bacterial abundance. However, for *rpoB*2 this percentage is lower implying that for this target different individuals can have different bacteria. Yet, *rpoB*2 contributes a much smaller percentage of OTUs than *rpoB*1 or 16S rRNA and therefore, the one or two different bacteria do not impact much upon the separation of individuals. These results show the importance of including *rpoB*, a single copy gene, in order to more accurately assess abundance estimates. When combining the experiments a large number of OTUs are removed, however when the percentage of sequences allocated to OTUs in common between both experiments is calculated the values are very high. This implies that the OTUs not in common belong to the rare microbiome (bacteria represented by few sequences) and are therefore not always detected (31). This shows that after filtering the sequences, nearly all those left are common to both experiments and can therefore be used to separate out the two individuals.

7.2.3 Bacteria

As presented in section 5.3 the bacterial composition of all eight saliva samples concurs with previous studies, indicating that the primers designed for this study are robust. Differences in abundances of the principle taxa were observed however, due to the differences in primers and sequencing technology used in other studies this is expected. This study showed a core genus-level microbiome of 58 genera covering about 95% of the reads. Of the 58 genera, 24 are unique to both 16S rRNA and *rpoB*1 respectively and 2 to *rpoB*2 with 2 being in common with all three targets, 1 in common with *rpoB*1 and *rpoB*2 and 6 in common with 16S rRNA and *rpoB*1. This shows that the addition of *rpoB*1 generates 24 core genera which would not have been detected with 16S rRNA alone, reiterating the benefit of using more than one target gene. The inefficiency of *rpoB*2 is also demonstrated here with only five core genera detected of which three are found by the other two targets, confirming the choice to remove this target from analysis. Previous studies have suggested the existence of a saliva core microbiome however, they all differ slightly indicating that larger-scale studies are required to properly define the core microbiome, if one exists (31, 71, 172, 173). Huse *et al.* defined a genus-level core microbiome of 22 OTUs,

with OTUs occurring in 95% of samples. Of the 22 OTUs 14 are found in the core genera of this study from 16S rRNA and *rpoB*1 combined, including *Streptococcus*, *Prevotella*, *Rothia* and *Veillonella* with 10 coming from 16S rRNA alone. They also show that of all the body sites tested saliva shows the largest core microbiome (173). A study specific to defining the healthy core microbiome of oral microbial communities finds similar results with the predominant taxa belonging to Firmicutes (*Streptococcus*, *Veillonellaceae* and *Granulicatella*), Actinobacteria (*Rothia*), Bacteroidetes (*Porphyromonas*) and Fusobacteria (*Fusobacterium*). They found that 99.8% of the reads belonged to shared higher taxa (genus-level and up) (31). This concurs with the results of this thesis which showed that the core genera covered about 95% of reads. Along with the suggested core microbiome Li *et al.* (172) propose a 'minor' microbiome which consists of taxa with low but stable abundances that appear in the majority of samples. This idea fits well with the results presented in table 5.8 which show that of the 58 core genera only a few make up about 99% of the reads for each target. Therefore, the majority of core genera are in low abundance. Overall, the microbial composition of the saliva samples analysed in this thesis corresponds with the literature. Defining the core microbiome is less important for the goal of this thesis as, if true, the core microbiome in saliva is quite large and therefore most of the differences between individuals come from varying abundances of bacteria. However, defining the core microbiome of healthy individuals will help to better understand the function of the bacteria (30). As, if a bacteria is found in everybody then it could be assumed that it is essential for the healthy functioning of the mouth.

As described in section 2.1.1.1 *Streptococcus* colonises all surfaces of the oral cavity, especially the tongue. As a large proportion of bacteria found in saliva come from the tongue it is logical that *Streptococcus* is found in high concentrations in saliva.

7.3 Comparison of two salivary microbiomes

The third aim of this thesis was to see whether the differentiation of two individuals through the analysis of the salivary microbiome is possible.

7.3.1 Hierarchical clustering

To perform the hierarchical clustering the data was first normalised to enable the comparison of both experiments, otherwise it would have been impossible to tell whether differences in the data were real or sequencing artefacts. For the analysis of the individual

experiments normalisation was not required. Subsequently the data was filtered to only use the taxa found to be significant by a 2-tailed unpaired t-test (and BF < 1 meaning a support, generally with values that very strongly support the hypothesis H_1). A t-test was chosen as only two individuals were being compared and through the log transformation of data a parametric test was suitable, if more individuals were to be compared an ANOVA would have to be used. Different significances were used for the individual and combined experiments (see section 6.2.1). A p-value of <0.1 is quite high compared to the standard use of p-values, however for this study it was only used as an indication of which taxa were more significant than others, not a definitive measure of significance. This seemed the most efficient way to filter the data. To reduce analysis complexity, only OTUs found in both sequencing runs were kept as they could be more accurately attributed to an individual and techniques used in forensic science are required to be as robust as possible. As described in section 6.2 for the combined experiments the unfiltered data groups by experiment not by individual. This indicates that samples sequenced in the same run exhibit run specific artefacts or DNA extraction/PCR-specific artefacts, which if not removed, skew the data.

Cluster analysis was developed for biological classification in 1963 (174) and has been successfully used since then, therefore it is well suited to this study. Specifically, hierarchical clustering was the chosen method for comparing samples as it has been shown to work well with two datasets (175). Hands and Everitt showed that the Ward method was the best overall method when compared with single linkage, complete linkage, average and centroid. Hierarchical clustering can easily be represented as a dendrogram enabling simple visualisation of results. As a dendrogram is a graphical representation of a cophenetic matrix, dendrograms can be compared using the cophenetic correlation coefficient or cophenetic distance (176). This coefficient can be used to test how faithfully a dendrogram preserves the pairwise distances between the original unmodelled data points (177). The closer the value is to one the more accurately the dendrogram represents the data (178). For all the individual experiments, for each target, the cophenetic distance was greater than 0.99 indicating that the dendrograms accurately represent the data. For the combined experiments both 16S rRNA alone and 16S rRNA combined with *rpoB*1 produced a cophenetic distance greater than 0.99, however *rpoB*1 and *rpoB*2 alone produced distances of 0.97 and 0.80 respectively. This supports the decision made to remove *rpoB*2 from the analysis as the results are not as reliable, whereas both *rpoB*1 and 16S rRNA produce dendrograms which accurately represent the data.

7.3.2 Individual differentiation

The results from the individual and combined experiments demonstrate that it is possible to group samples from one individual and separate them from samples from a second individual (see sections 6.4 and 6.5). For both the individual and combined experiments less than one third of the OTUs are classed as significant, indicating that a large portion of the bacterial communities are similar in both individuals. This is supported by a study by Nasidze *et al.* which showed that 86% of variation was shared between all datasets/individuals, however it also showed that the variation between different individuals was greater than variation within the same individual (179). This also shows why it is necessary to filter out the most significantly different bacteria and use them to differentiate the individuals. As explained in section 6.5, for the combined experiments the significant bacteria are not all the same as those found in the individual experiments. However, when those significant at $p < 0.1$ for the individual experiments are compared to those significant at $p < 0.01$ for the combined experiments the percent of significant OTUs in common increases to between 75% and 100%. This could be because in the individual experiments there are only two samples per individual so to be significant at $p < 0.01$ involves the abundances being very different so it is more likely that the abundances will differ slightly and therefore only be significant at $p < 0.1$. Whereas, for the combined experiments there are four samples per individual and therefore differences in abundances are more likely, classifying certain OTUs significant at $p < 0.01$ when in the individual experiments they were only significant at $p < 0.1$.

For both the individual and combined experiments intra-individual variation is observed. Inevitably there is some natural variation in saliva microbiota due to it being a dynamic fluid and certain bacteria will not always be detected, being either absent or in too few numbers. This explains the existence of both intra- and inter-individual variation and specifically why they vary, as not all bacteria are detected in every sample from one individual. Therefore, even when comparing samples from the same individual variation is present. Furthermore, even though this study only investigated two individuals, it shows that intra-individual variation is a lot smaller than inter-individual variation. However, more samples will need to be analysed in order to confirm this pattern. Lazarevic *et al.* also investigated the inter- and intra-individual variations in the salivary microbiome over the period of one month. They found that samples from the same individual clustered together indicating that the salivary microbiome is quite stable (71). They also found that within the same individual samples taken closer together did not group more closely than those taken further apart, agreeing with the results presented in section 6.5.

The combination of 16S rRNA and *rpoB*1 provides the smallest percentage of intra-individual variation and the second highest inter-individual variation (see section 6.5.2). Even though the best separation in terms of largest inter-individual variation is achieved by combining all three targets the addition of *rpoB*2 does not provide much more separation than 16S rRNA and *rpoB*1. As discussed above, *rpoB*1 identifies a number of bacteria undetectable by 16S rRNA and it is therefore essential to include it. In terms of both identification of bacteria and separation of individuals the combination of 16S rRNA and *rpoB*1 is the most effective.

7.3.2.1 Evaluative framework

Developing an evaluative framework is beyond the scope of this thesis, however the subject deserves a mention as without a means for evaluating the results they could not reliably be presented in a court of law. I proposed two potential methods for evaluating microbiome data (180). The first is a simple counting method which can be used to tally how many of each sequence appears per sample, however this can prove difficult. As described above, the copy number of 16S rRNA can vary greatly between bacterial species and PCR-induced bias can both skew estimations of biodiversity (181). This method would also require a comparison database and statistics on the proportions of bacteria in different populations. Due to the cost of analysis and the number of samples required to produce such a database and statistics this method is not feasible for the near-future. The second method uses population data to help associate a sample to a cluster. The population data is required to give accurate estimations for intra- and inter-individual variation. The next question posed is: how to compare the intra- and inter-individual variation? From the results presented in this thesis I suggest that a ratio between the relative distances calculated through hierarchical clustering could be used. As presented in chapter 6 the intra-individual variation is never more than 20% of the inter-individual variation, therefore a threshold could be set defining a maximum level of intra-individual variation, below which all samples are classed as belonging to one individual. However, as discussed above, hierarchical clustering was chosen as it is well adapted to small sample size. Therefore, if many more samples were analysed a different method, better suited to a large sample size, would be required. Examples of such techniques would be principal coordinates analysis (PCoA) and principal component analysis (PCA). These techniques have been applied by many undertaking bacterial community analysis with a larger sample size (26, 37, 67, 71, 182). These techniques explore and visualise similarities and dissimilarities in a dataset using a distance matrix as its base. Each sample is represented as a

point in a low-dimensional space. The more similar samples are the closer they will be. Ideally all samples from one individual would group together and distinct groups could be visualised.

7.3.3 Minimum number of sequences

The results show that the minimum number of sequences for this type of analysis is 100,000 as this provides good separation between individuals whilst minimising the intra-individual variation. The minimum number of sequences required to differentiate two individuals was calculated in order to propose the maximum number of samples which can be sequenced in one run. As explained in section 6.6, the cost of analysis per sample needs to be considered before this technique could be integrated into routine analysis. By maximising the number of samples analysed in one run the cost per sample dramatically decreases. By using the Miseq the time required for one run is significantly lower and hence more suitable for a forensic laboratory where speed is essential. It is important to remember that if two target genes are used each one requires 100,000 sequences and therefore halves the number of samples per run. As reference samples are essential for the differentiation of individuals, if possible to minimise the effect of sequencing artefacts, I suggest analysing trace and reference samples together. If the sole purpose for analysis is to provide intelligence then the more sequences per sample the better as this will provide a more complete overview of the bacteria present.

7.3.4 What next?

Having shown that two individuals can be differentiated using the proposed method the next step is to see how the separation differs when 1: more samples from the same individual are added and 2: samples from different individuals are added. For point one, two outcomes are possible; either the additional samples fit with the existing ones or they provide even more variation. A study by Lazarevic *et al.* compared the number of species-level phylotypes shared as a function of the number of samples compared (183). They showed that the higher the number of samples from one individual the smaller the number of shared phylotypes, however they only compared three samples per individual. This study has already shown that four samples from one individual can be grouped together, indicating that even though they might share fewer phylotypes they are still more similar than samples from another individual. Regarding point two, the addition of more individuals will indicate how similar different individuals are. For example, if one more

individual is added will the individuals be separated by similar relative distances to those calculated for two individuals or will greater depth of analysis be required. In the same paper Lazarevic *et al.* also show that the more samples which are combined the smaller the number of shared species-level phylotypes. This indicates that the core microbiome decreases with increasing sample number. However, this will not necessarily negatively impact upon the differentiation of individuals. As described above the separation of individuals is mainly based on differences in abundances of bacteria and not presence/absence of bacteria. In actual fact, the fewer similarities there are between individuals the easier the differentiation will be.

Following on from point two, the next question to ask is how many individuals can be differentiated at the same time i.e. what is the limit of this technique? Realistically, how many samples would need to be differentiated at the same time? In a real case scenario, if a person was sexually assaulted by one person then the expected number of individuals contributing to a trace would be two (victim and suspect) or maybe three if the victim has a partner who might have left residual traces. It is unlikely but not impossible that many individuals could contribute to a trace, in which case, it would be beneficial if this technique could differentiate them all. With current DNA profiling methods complex mixtures pose a problem as it is difficult to deconvolute them even with reference profiles, therefore a method which could successfully analyse multi-contributor mixtures would be welcomed (184). However, without reference profiles this technique would not be able to deconvolute mixtures. Moreover, this technique will rely on reference samples to identify any number of individuals from one onwards. Without a reference profile it will be impossible to associate the bacteria to a particular person.

It would also be interesting to perform serial dilutions to see whether DNA concentration influences clustering. This would provide a lower limit in terms of DNA concentration for the use of this technique.

7.4 Influencing factors

Saliva is a dynamic fluid, which means that through different mechanisms such as talking and eating bacteria can enter and exit contributing to the bacterial flora. It is mainly for this reason that intra- and inter-individual variation is observed. There are many different factors which can influence the salivary microbiome:

7.4.1 Genetics

One of the first questions asked is: what influence does genetics have? To date most studies have concentrated on characterising the salivary microbiome in healthy and diseased states. However, a few studies have analysed the oral/salivary microbiome of twins, common subjects for studying the effects of genetics. Corby *et al.* published two papers; the first demonstrates that genetic factors contribute to the salivary levels of mutans streptococci, a major caries pathogen, in preschool twins (185). The second investigates the heritability of oral microbial species in caries-active and caries-free twins (186). They found that the relative abundance of bacteria associated with caries-free twins was in part determined by genetics and that there was a distinct difference between the caries-free and caries-active microbiomes. They conclude that genetic and/or familial factors contribute significantly to the colonisation of oral bacteria in twins. The first study indicates that genetics may play a part in forming the diseased oral microbiome. However, the second study shows that the role of genetics is unclear with no concrete conclusion being drawn. The main disadvantage of both these studies is that they use twins aged about four years old and therefore, it is difficult to extrapolate these results to adults. A more recent study investigated the salivary microbiome of identical twins. Stahringer *et al.* showed that for twins aged between 12-24 years their salivary microbiomes were not statistically more similar than for any other pair (74). This indicates that overall there is very little or no genetic influence on salivary microbiome composition and that the differences observed between twins mainly come from environmental factors. To confirm this pattern more studies need to be carried out on a larger age range starting from 24 years. However, the results presented in this article show promise for the use of the salivary microbiome for differentiating twins. The one conclusion which can be drawn from these articles is that much more work is required to elucidate the link between genetics and the composition of the salivary microbiome.

7.4.2 Antibiotics

One major factor which can effect the salivary microbiome is antibiotics. Thus far, very few studies have broached the subject (187, 188, 189) with only one concentrating on saliva (190). Lazarevic *et al.* described the effects of amoxicillin treatment on the salivary microbiota in children with acute otitis media. They showed that directly after treatment there was a change in the microbiota in terms of both species richness and diversity. However, three weeks after the end of treatment the microbiota had mainly recovered back to pre-antibiotic diversity. This, would only impact cases where the saliva was deposited on a crime scene whilst the perpetrator was taking antibiotics. In such cases, presence of

antibiotics in the sample might be determined and an additional sample might then be obtained upon treatment with the same antimicrobial substance. In the case where the perpetrator is taking antibiotics when apprehended a reference sample could be taken at a later date once the salivary microbiome had recovered. This study only investigated children, therefore similar studies concentrating on adults are required to see if the same pattern is observed. This is extremely important as the majority of crime is committed by adults so it will be essential to show that antibiotics only have a short term effect on the composition of the adult salivary microbiome.

7.4.3 Environmental factors

Many studies mention that environmental factors such as diet, oral hygiene, smoking, alcohol and drug consumption may influence the salivary microbiome (70, 71, 74, 165, 183, 191) however, only one study thus far has directly approached the subject (192). Belstrom *et al.* investigated whether diet, lifestyle and socio-economic status had an effect on the salivary microbiome of 292 participants. They detected two bacteria (*Streptococcus sobrinus* and *Eubacterium brachy*) in smokers which were not detected in non-smokers and when former smokers were compared with never smokers there was no statistical difference. This suggests that the two bacteria are associated with smoking and could potentially be used for intelligence purposes to indicate whether a person smokes or not. They also found statistical differences between high and low socio-economic status with 20 bacteria having different abundances. It would be more difficult to use this for intelligence purposes as it would involve calculating ratios between the statistically different bacteria and making an inference about socio-economic status. Quite surprisingly they found that diet did not produce any statistical differences and neither did age, gender, BMI or alcohol consumption. However, as they discuss in their paper, participants of medical studies tend to be healthier than the general population and this may limit the amount of variation and the detection of significant associations. Stahringer *et al.* briefly mention that their study included data relating to personal preferences and characteristics and they also found no effect from weight, gender or diet (74), yet they conclude that environmental factors provide the greatest influence on the composition of the oral microbiome. A recent study by Benitez-Paez *et al.* which investigated microbiota diversity in oral biofilms showed that bacteria which changed activity during biofilm formation and after meal ingestion were person-specific and after meal ingestion some individuals showed no changes in the active bacterial population (193). These results indicate that the oral microbiome is quite

stable, person specific and the effect of meal ingestion is minimal. Furthermore, some of the bacteria will be found in saliva and therefore may follow the same pattern.

Subsequently, even if diet does not provide great variation in the salivary microbiome, certain bacteria could still be associated with a particular diet or foodstuff. For example, in both individuals analysed for this thesis *Malus x domestica* was detected, which is otherwise known as the common eating apple. This could imply that both individuals eat apples. One could hypothesize that a vegetarian would have different bacteria to a carnivore or vegan. A recent study has shown that children with celiac disease who therefore follow a gluten-free diet have a different salivary microbiome to healthy children with a decreased number of *Streptococcaceae* (194). However much more research is required to link specific bacteria to specific foodstuffs or food regimes.

Song *et al.* published a paper comparing oral, skin and gut microbiota of cohabiting family members and their pet dogs (195). They found that for the oral microbiome, age had an influence with a large increase in diversity happening between 0-3 years. However, neither gender nor dogs seemed to affect the oral or gut microbiome but they did significantly affect skin microbiota, whereas cohabitation influenced all three but affected the skin microbiota more. This could be because even though saliva is a dynamic fluid the mouth will provide a certain level of protection, whereas skin is always exposed to environmental contact. This paper also estimates that oral bacteria make up about eleven percent of palm microbiota indicating that close physical contact can affect the taxonomic composition of the skin. To date this is the first study to investigate the effect of cohabitation and dogs on the oral microbiome, therefore further studies are required to confirm the results. Another study by Jung-Gyu Kang *et al.* investigated bacterial diversity in human saliva from different ages (196). They found that young adults (32 and 35 years) had more species coming from *Streptococcus* and *Prevotella*, whereas individuals aged 5 and 65 had more species belonging to *Rothia* and the latter showed higher bacterial diversity. This study indicates that age may influence the composition of the salivary microbiome. Furthermore, Nasidze *et al.* investigated global diversity in the salivary microbiome and they found that geographical location did not have an influence. However, they do state that this conclusion is limited to the pool of 16S rRNA sequences identified (70).

Despite the fact that the oral microbiome has been studied extensively in dentistry there are very few studies which discuss the effect of oral hygiene on the salivary microbiome, most articles concentrate on the effect the bacteria have on oral health. The assumption is often made that oral hygiene does influence the composition of the salivary microbiome. A Brazilian study investigated bacterial diversity in the saliva of patients with different oral

hygiene indexes. They found that indviduals with good oral hygiene had greater bacterial diversity than those with poor oral hygiene and the two communities were significantly different (197). Another study by Matsui *et al.* investigated the effect of tongue cleaning on bacterial flora in tongue coating and dental plaque (198). They found that tongue cleaning reduced the amount of bacteria in tongue coating and that as the total bacteria recovered an increase in *Fusobacterium nucleatum* was observed. Both of these studies indicate that oral hygiene influences the composition of the oral/salivary microbiome. However, further studies are required to fully elucidate the effect of oral hygiene on the composition of the salivary microbiome.

These studies provide a good start to investigating the effect of environmental factors however, more studies are required to reveal the real impact of these factors on the composition of the salivary microbiome. They also indicate that a persons microbiome could be used as intelligence to inform about their lifestyle.

7.5 Ethical considerations

When any new technique is proposed for use in the field of forensic science the ethical implications of the technique have to be taken into consideration. As the technique will be used for law enforcement it cannot break any human rights laws or any laws specific to the country where it is being used. For example, in Switzerland it is illegal to use any technique which is deemed to reveal anything other than the sex of a person. Therefore, for example, phenotyping cannot be used. DNA profiling is acceptable, as the only personal trait of the person revealed is the sex and the targeted STRs are in non-coding regions. With human DNA there have been many debates over whether a DNA profile can be kept and if so, where and how long for and subsequently, who can have access to the profiles. For a technique to be used on a real case, it has to be scrutinised and validated, (199) because it could be the one piece of evidence that indicates whether a person is innocent or guilty.

With advances in technology, analysis of the human microbiome has reached new levels, enabling scientists to fully characterise which bacteria are present, where and what their function are. This means that, much more is now being learnt about the connection between health and the microbiome. With this link to health, individuals will become responsible for maintaining the health of their microbiome (200). Subsequently, the general public will become more interested in the microbiome and this will influence whether they accept the analysis of specific microbiomes for forensic purposes.

The major difference between standard human identification techniques and one based on microbiome analysis is the latter uses bacterial DNA not human DNA. Analysis of human DNA directly connects a sample to a person and a person perceives their DNA as being part of them. Therefore, in general, people do not like giving samples of their DNA, especially when they are unsure what analysis will be carried out. As more is discovered about the human microbiome peoples' perception of it will develop and potentially change. It has already been referred to as the second genome (201). Therefore, will people start to perceive their microbiome as part of them? This idea was developed by Gli *et al.* in (200), where they discuss the human microbiome and conceptions of self. They propose that, if our microbiota are unique and therefore, could be used to identify us then this should encourage us to identify with our microbiome. Moreover, we could see our unique microbial mark as a unique expression of self, in the same way we do with fingerprints. As discussed above, environmental factors strongly influence our microbiota and this could lead to a realisation that our identity is partly determined by our environment. They conclude that, in the future, we might view a person as a human and bacterial hybrid, a superorganism. It is therefore, very important that research on the human microbiome is accurately communicated to the general public so they can understand it's importance and make informed decisions regarding it's use.

In section 7.3.2.1 I propose a method for evaluating a comparison between microbiomes for the purpose of human identification which would involve using databases containing the proportions of bacteria in different populations. At the moment, the only feasible way of acquiring such information would be to use data from clinical studies and combine that with forensic studies. Furthermore, in the future, the analysis of the salivary microbiome may become standard dental practice, in which case, it may be possible for the forensic domain to have access to this data for comparison purposes. However, both of these possibilities may pose ethical problems due to confidentiality agreements with participants/patients. A recent book discussing the ethical, legal and social concerns of human microbiome research states that access to private patient information for research might be justified if the research is potentially beneficial to society (202). Therefore, the general public may agree to allowing access to data concerning their microbiomes if it helps catch criminals and the data is used for comparison purposes only. However, this will involve properly informing the general public of exactly who will have access to their data and what it will be used for.

7.6 Scientific (forensic) relevance

The work presented in this thesis is original and interesting to the scientific community as it takes a technique which is already in use, adapts it, and applies it to a completely different field. Instead of approaching the analysis of the salivary microbiome from the pure characterisation aspect, it uses the characterisation to differentiate samples from different individuals. Results support the fact that the saliva microbiome is stable over time and distinct in each individual, however a core microbiome does seem to exist, the extent of which is currently undetermined. Furthermore, the data presented will contribute to defining which bacteria are present in the healthy salivary microbiome and how they change over time. Unlike most other studies which use only 16S rRNA this study combined 16S rRNA with *rpoB* proving that the combination of the two provides a deeper characterisation of the bacteria present. Without *rpoB* many of the streptococci species present would go undetected due to the lack of genetic resolution of 16S rRNA. As *Streptococcus* is the most abundant genus in saliva it is important to characterise it in as much detail as possible. This approach would be beneficial for studies aiming to characterise bacteria to a level unattainable with 16S rRNA alone whilst still getting a good overview of which bacteria are present. As more is learnt about which bacteria are present in health and disease and their function more can be understood about the influencing factors. By combining the potential uses of forensic science with the more standard goals of microbiology a better understanding of the salivary microbiome can be achieved.

The work presented in this thesis is also of interest to the forensic community as it provides a potential new method for human identification. As described in section 1.1 the standard technique of STR typing used for human identification has its limitations and currently there is no real alternative, only other less sensitive techniques based on human DNA. The technique proposed in this thesis will provide a complimentary method for saliva traces. As saliva is often found in sexual assault cases which more often than not do not result in a conviction (see section 1), this technique could become indispensable. This thesis provides the basis for future studies into the application of high-throughput sequencing of bacterial DNA to saliva samples. It has been shown that the differentiation of two individuals is possible and that samples from the same individual group together. The application of this technique to case-like samples was beyond the scope of this thesis, therefore this needs to be tested before the technique could be used on a real case.

One issue with forensic samples is the quantity of DNA is often quite low and this could pose a problem for high-throughput sequencing. Currently the desired amount of DNA for paired-end library preparation is about one microgram, whereas forensic traces can

contain as little as a few picograms of DNA. PCR amplification is used to increase the amount of DNA however, a minimum amount of DNA is required to overcome stochastic effects (203). The average bacterial genome has a genome mass of a few femtograms which is a lot smaller than the diploid human genome at 6.3 picograms. Therefore, from the same forensic trace much more bacterial DNA can be amplified than human DNA. As described in section 2.1.1 500 million bacterial cells can be found in one millilitre of saliva which equates to 500 thousand bacterial cells in one microlitre of saliva. Most saliva traces found are larger than one microlitre therefore, in theory, the DNA of more than 500 thousand bacterial cells could be amplified. However, it must be considered that not all bacterial DNA will be amplified by the chosen primers and due to amplification bias differences may be seen between traces and reference samples as the latter will contain much more DNA. More work is required to reveal any differences and whether they impact upon associating traces to reference samples. A recent study has presented a method for preparing high-quality Illumina sequencing libraries from picogram quantities of DNA (204). Currently, this method does not work with pre-amplified sequences but it shows promise that in the future a method could be developed, enabling forensic samples very low in DNA to be analysed.

Another important point to consider with regards to forensic traces is how resistant the traces (i.e here the bacterial DNA) are to external factors. Indeed, human DNA can be degraded by UV light, heat and humidity, environmental conditions which are often found at crime scenes. Specifically nucleases, such as deoxyribonuclease I, found in saliva degrade exposed human DNA making it difficult to obtain sufficient quantity and quality of DNA to produce a DNA profile (41). One advantage of microbiota based forensic investigation is that bacterial DNA is better protected from enzymatic degradation than human DNA as bacterial DNA is circular often highly condensed as “nucleoid”. Moreover, prokaryotic cells have a cell wall which is chemically complex with a peptidoglycan matrix that better protects the contents of the cell compared to the cell membrane of eukaryotic cells. Therefore bacterial DNA should be more resistant than eukaryotic DNA to external factors taking longer to be degraded.

A further point to consider is the effect of background bacteria and mechanisms of transfer. When a saliva trace is left on a surface there will invariably be some bacteria already present on the surface. The question becomes, can the bacteria in the saliva be differentiated from the background. This is still to be investigated, however I would propose that a second sample of only the background be taken from as close to the trace as possible in order to determine which bacteria are present on the surface. These bacteria could

subsequently be subtracted from the trace sample revealing the bacteria in the saliva. The second issue of mechanisms of transfer is important for determining activity level, i.e. how the saliva got there and not just who the saliva belongs to. Would it be possible to find out from the saliva trace whether it was transfered innocently or during a crime. I think this will be very difficult, already linking the bacteria to a specific person will be an achievement. However, it is definitely worth investigating further and simulating crimes to see if any differences can be observed.

The application of this technique to sexual assault cases would involve analysing mixed skin and saliva samples as the majority of saliva traces are likely to be found on the skin through either biting, kissing or licking. For this to be successful further work is required to see whether the saliva microbiome can be separated from the skin microbiome and subsequently associated to a reference sample. A study by Costello *et al.* into bacterial community variation in human body habitats across space and time showed that skin bacterial communities varied both within and between individuals and were most divergent from oral bacteria (26). They also found that different skin sites (e.g. forearm, palm, forehead and foot) showed different levels of diversity with the forearm having high diversity. However, as discussed previously the skin has more contact with the external environment so greater variation is expected. They also tested how oral bacteria reacts when placed on the forearm and they discovered that, over a couple of hours, it stays the same and does not adapt to the forearm microbiome. This work is very promising as it indicates that if a saliva sample is left on skin that it could be associated to the corresponding reference profile. A few studies have investigated the use of streptococcal DNA in bite-mark analysis (38, 39, 40, 41). The first two studies show that after a few hours streptococcal DNA could be recovered and strain level identification used to link the bite-mark to the suspect. Rahimi *et al.* (40) used arbitrarily primed PCR to amplify streptococcal DNA, identifying 106 genotypes of which, at least 8 distinct strains were found in each participant. They used the amplicon profile to match the biter to the sampled bite-mark. Most recently Kennedy *et al.* (41) used targeted sequencing of 16S rRNA, 16S-23S intergenic spacer and *rpoB* aimed specifically at streptococcus species/strains. They found that for all targeted regions it was possible to match the biter to the bite-mark, however *rpoB* was by far the most effective matching 100% of bite-marks. They conclude that this technique could be used to corroborate other evidence for the identification of assailants. These results are very promising as they show that streptococcal DNA can be detected on skin and matched to a reference sample. Therefore, in theory this should apply to all bacteria and enable the technique presented in this thesis to be used for human identification and not just corroborate other evidence.

When a new technique is proposed, laboratories must consider the economical impact of integrating the technique into routine analysis. Currently, high-throughput sequencing is quite expensive however, since it's introduction about eight years ago costs have gradually decreased making the technique more and more affordable. With the arrival of benchtop sequencers, smaller laboratories are now able to buy one. In the past couple of years groups have started investigating the use of high-throughput sequencing for analysing STRs (205, 206) and developing specific programs to analyse the data (207). Concurrently, Illumina have been developing a sequencing strategy aimed at forensic science (208). With these advances, it is possible that in the next few years high-throughput sequencing will replace capillary electrophoresis, as a result most forensic laboratories will own a high-throughput sequencer. If this is the case, then the technique proposed in this thesis could easily slot into routine analysis without incurring large costs.

7.7 Future work

As discussed throughout this chapter there are many different factors left to study. Having shown that this technique works with the DNA extraction method used it would be interesting to look into other methods to see whether there is a big difference and whether samples extracted using different methods could be compared. Furthermore, the possibility to co-extract human and bacterial DNA is of great interest and should be investigated. For the applicability to forensic science this technique should be tested on the Illumina benchtop sequencer, the MiSeq, to check that the results are comparable to those from the HiSeq. However, one would expect the results to be comparable as the technology is the same. The next major step is to analyse samples from different people to test the limits of the technique and thus offer a picture of the abundance (occurrence) of the characteristics of interest. It is also important to analyse more samples from the two individuals used in the current research to see which taxa are found and whether they correspond to either of the first two experiments. To be able to use the data for intelligence purposes the effect of external factors needs to be investigated, in order to make links between specific bacteria and certain lifestyle choices. As forensic traces are often found on exposed surfaces the persistence of bacteria on these surfaces should be studied, in order to investigate the effect of background bacteria and how the bacteria interact with different surface types. Finally, from all this new data an evaluative framework can be developed so that the results can be presented in a court of law.

8

Conclusion

This thesis presented the first study into the use of the salivary microbiome for human identification. It has shown that the salivary microbiome exhibits a significant biodiversity and by using a PCR-based metagenomic approach the discrimination of two unrelated individuals was possible. The biodiversity revealed in all samples was similar to that found by previous studies, showing that the designed primers are robust. However, the abundances do differ but this has been observed previously (74).

The goal of this technique is not to replace current methods used for human identification but to be complimentary. When these methods do not produce satisfactory results there is no other option from a biological identification stand-point. By analysing the salivary microbiome, new options become available that previously were not possible. There are two potential applications of this technique in forensic science: human identification and intelligence. The first, presented in this thesis, will only be possible if a reference sample is available. The second putative application, presented in the discussion, uses the same data but looks at the presence of specific bacteria which could indicate a certain lifestyle. This information might be used to help guide an investigation. If an identification is not possible then the data acquired could still provide valuable information to a case. However, much more work is required to relate given species to given lifestyle habits. It is possible to extract both human and bacterial DNA from the same sample. This will enable both methods of human identification to be applied to the same sample, avoiding having to chose which method to use. The advantage of this is that if the human DNA analysis produces a full profile but no match in a database, the bacterial DNA could potentially be analysed for intelligence purposes.

In conclusion, Illumina high-throughput sequencing of the salivary microbiome can be used to identify saliva samples from two different individuals. Altogether, this technique proved

to be highly robust and is innovative not only for its putative application in forensic science, but also by using a combination of a highly discriminative gene (*rpoB*) with the 16S rRNA target generally used for PCR-based metagenomics. Furthermore, this technique shows promise for human identification, specifically for twins and other cases where standard DNA typing does not provide satisfactory results due to degradation of human DNA. Further work is required to investigate the benefit and limitations of this technique.

Bibliography

[1] L. Bontadelli. *Study of DNA shedder quality*. Master's thesis, Ecole de Sciences Criminelles, University of Lausanne, 2009. 1

[2] S. Cowen, P. Debenham, A. Dixon, S. Kutranov, J. Thomson, and K. Way. **An investigation of the robustness of the consensus method of interpreting Low-Template DNA profiles**. *Forensic Science International: Genetics*, **5**(5):400 – 406, 2011. 1

[3] B. Caddy, G.R. Taylor, and A. Linacre. **A Review of the Science of Low Template DNA Analysis**. *Home Office Regulation Unit*, 2008. 1

[4] B. Budowle, A. J. Eisenberg, and A. van Daal. **Low Copy Number typing has yet to achieve general acceptance**. *Forensic Science International: Genetics Supplement Series*, **2**(1):551–552, 2009. 1

[5] B. Budowle. **Low Copy Number Typing Still Lacks Robustness and Reliability.** *Profiles in DNA*, **13**(2), 2010. 1

[6] A. Rennison. **Making the case for Low-Template DNA analysis**. *Nature*, **465**:157, 2010. 1

[7] N. Fierer, M. Hamady, C. L. Lauber, and R. Knight. **The influence of sex, handedness, and washing on the diversity of hand surface bacteria**. *Proceedings of the National Academy of Sciences*, **105**(46):17994–17999, 2008. 2

[8] A. Gentleman. **Inside a Sexual Assault Referral Centre**. *The Guardian*, **25th November**, 2010. 2

[9] **CPS/Home Office Roundtable briefing**, 26 September 2013. 2

[10] A. J. Jeffreys, V. Wilson, and S. L. Thein. **Individual-specific 'fingerprints' of human DNA**. *Nature*, **316**(6023):76–79, 07 1985. 2

[11] K. B. Mullis and F. A. Faloona. **Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction**. In Ray Wu, editor, *Recombinant DNA Part F*, **155** of *Methods in Enzymology*, pages 335 – 350. Academic Press, 1987. 2

[12] J. Wambaugh. *The Blooding*. Perigord Press, New York, 1989. 3

[13] J. M. Butler. *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press, London, 2001. 3

[14] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. **The Human Microbiome Project**. *Nature*, **449**(7164):804–810, 2007. 3, 5, 14

[15] M. Brochet, E. Couvé, M. Zouine, T. Vallaeys, C. Rusniok, M-C. Lamy, C. Buchrieser, P. Trieu-Cuot, F. Kunst, C. Poyart, and P. Glaser. **Genomic diversity and evolution within the species Streptococcus agalactiae**. *Microbes and Infection*, **8**(5):1227–1243, 2006. 3, 20, 24

[16] H. Tettelin, V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli, and C. M. Fraser. **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome"**. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(39):13950–13955, 2005. 3, 20, 24

[17] N. L. Hiller, B. Janto, J. S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N. E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S. D. Bentley, J. C. Post, G. D. Ehrlich, and F. Z. Hu. **Comparative Genomic Analyses of Seventeen Streptococcus pneumoniae Strains: Insights into the Pneumococcal Supragenome**. *Journal of Bacteriology*, **189**(22):8186–8195, 2007. 3

[18] D. Medini, D. Serruto, J. Parkhill, D. A. Relman, C. Donati, R. Moxon, S. Falkow, and R. Rappuoli. **Microbiology in the post-genomic era**. *Nature Reviews Microbiology*, **6**(6):419–430, 2008. 4

[19] M. Meyer, U. Stenzel, and M. Hofreiter. **Parallel tagged sequencing on the 454 platform**. *Nature Protocols*, **3**(2):267–278, 2008. 5

[20] T. Maricic and S. Paabo. **Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands**. *Biotechniques*, **46**(1):51, 2009. 5

[21] N. Lennon, R. Lintner, S. Anderson, P. Alvarez, A. Barry, W. Brockman, R. Daza, R. Erlich, G. Giannoukos, L. Green, A. Hollinger, C. Hoover, D. Jaffe, F. Juhn, D. McCarthy, D. Perrin, K. Ponchner, T. Powers, K. Rizzolo, D. Robbins, E. Ryan, C. Russ, T. Sparrow, J. Stalker, S. Steelman, M. Weiand, A. Zimmer, M. Henn, C. Nusbaum, and R. Nicol. **A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454**. *Genome Biology*, **11**(2):R15, 2010. 5

[22] G. B. Gloor, R. Hummelen, J. M. Macklaim, R. J. Dickson, A. D. Fernandes, R. MacPhee, and G. Reid. **Microbiome Profiling by Illumina Sequencing of Combinatorial Sequence-Tagged PCR Products**. *Plos One*, **5**(10):e15406, 2010. 5

[23] A. K. Bartram, M. D. J. Lynch, J. C. Stearns, G. Moreno-Hagelsieb, and J. D. Neufeld. **Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads**. *Applied and Environmental Microbiology*, **77**(11):3846–3852, 2011. 5

[24] K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara, and S. Kanaya. **Sequence-specific error profile of Illumina sequencers**. *Nucleic Acids Research*, **39**(13):e90, 2011. 5

[25] M. J. Friedrich. **Microbiome Project Seeks to Understand Human Body's Microscopic Residents**. *JAMA*, **300**(7):777–778, 2008. 5

BIBLIOGRAPHY

[26] E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. **Bacterial Community Variation in Human Body Habitats Across Space and Time**. *Science*, **326**(5960):1694–1697, 2009. 5, 13, 111, 121

[27] E. A. Grice, H. H. Kong, S. Conlan, C. B. Deming, J. Davis, A. C. Young, NISC Comparative Sequencing Program, G. G. Bouffard, R. W. Blakesley, P. R. Murray, E. D. Green, M. L. Turner, and J. A. Segre. **Topographical and Temporal Diversity of the Human Skin Microbiome**. *Science*, **324**(5931):1190–1192, 2009. 5

[28] M. J. Blaser. **Harnessing the power of the human microbiome**. *Proceedings of the National Academy of Sciences*, **107**(14):6125–6126, 2010. 5

[29] G. J. Caporaso, C. Lauber, E. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. Gordon, and R. Knight. **Moving pictures of the human microbiome**. *Genome Biology*, **12**(5):R50, 2011. 5

[30] J. A. Aas, B. J. Paster, L. N. Stokes, I. Olsen, and F. E. Dewhirst. **Defining the Normal Bacterial Flora of the Oral Cavity**. *Journal of Clinical Microbiology*, **43**(11):5721–5732, 2005. 5, 9, 14, 66, 108

[31] E. Zaura, B. J. F. Keijser, S. M. Huse, and W. Crielaard. **Defining the healthy core microbiome of oral microbial communities**. *BMC Microbiology*, **9**, 2009. 5, 12, 14, 107, 108

[32] M. A. Curtis, C. Zenobia, and R. P. Darveau. **The Relationship of the Oral Microbiotia to Periodontal Health and Disease**. *Cell Host and Microbe*, **10**(4):302 – 306, 2011. 5, 12

[33] P. Belda-Ferre, L. D. Alcaraz, R. Cabrera-Rubio, H. Romero, A. Simon-Soro, M. Pignatelli, and A. Mira. **The oral metagenome in health and disease**. *ISME J*, **6**(1):46–56, 2012. 5

[34] M. F. A. Syed and T. Farzeen. **Oral microbial habitat a dynamic entity**. *Journal of Oral Biology and Craniofacial Research*, (0), 2012. 5, 8, 9

[35] W. Yamanaka, T. Takeshita, Y. Shibata, K. Matsuo, N. Eshima, T. Yokoyama, and Y. Yamashita. **Compositional Stability of a Salivary Bacterial Population against Supragingival Microbiota Shift following Periodontal Therapy**. *PLoS ONE*, **7**(8):e42806, 2012. 5

[36] W. G. Wade. **Characterisation of the human oral microbiome**. *Journal of Oral Biosciences*, **55**(3):143–148, 2013. 5

[37] N. Fierer, C. L. Lauber, N. Zhou, D. McDonald, E. K. Costello, and R. Knight. **Forensic identification using skin bacterial communities**. *Proceedings of the National Academy of Sciences of the United States of America*, **107**(14):6477–6481, 2010. 5, 111

[38] K. A. Brown, T. R. Elliot, A. H. Rogers, and J. C. Thonard. **The survival of oral streptococci on human skin and its implication in bite-mark investigation**. *Forensic Science International*, **26**(3):193–197, 1984. 5, 121

[39] T. R. Elliot, A. H. Rogers, J. R. Haverkamp, and D. Groothuis. **Analytical pyrolysis of Streptococcus salivarius as and aid to identification in bite-mark investigation**. *Forensic Science International*, **26**(2):131–137, 1984. 5, 121

[40] M. Rahimi, N. Heng, J. Kieser, and G. Tompkins. **Genotypic comparison of bacteria recovered from human bite marks and teeth using arbitrarily primed PCR**. *Journal of Applied Microbiology*, **99**(5):1265–1270, 2005. 5, 121

[41] D. M. Kennedy, J-A. L. Stanton, J. A. García, C. Mason, C. J. Rand, J. A. Kieser, and G. R. Tompkins. **Microbial Analysis of Bite Marks by Sequence Comparison of Streptococcal DNA**. *PLoS ONE*, **7**(12):e51757, 2012. 5, 9, 120, 121

[42] P-O. Glantz. **Interfacial phenomena in the oral cavity**. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, **123-124**:657–670, 1997. 7

[43] S. P. Humphrey and R. T. Williamson. **A review of saliva: Normal composition, flow, and function**. *The Journal of Prosthetic Dentistry*, **85**(2):162–169, 2001. 7

[44] A. R. Silvers and P. M. Som. **Salivary Glands**. *Radiologic Clinics of North America*, **36**(5):941–966, 1998. 7

[45] C-M. Huang. **Comparative proteomic analysis of human whole saliva**. *Archives of Oral Biology*, **49**(12):951–962, 2004. 8

[46] D.J. Macarthur and N.A. Jacques. **Proteome Analysis of Oral Pathogens**. *Journal of Dental Research*, **82**(11):870–876, 2003. 8

[47] D. B. Ferguson and C. A. Botchway. **A comparison of circadian variation in the flow rate and composition of stimulated human parotid, submandibular and whole salivas from the same individuals**. *Archives of Oral Biology*, **25**(8-9):559–568, 1980. 8

[48] E. C. I. Veerman, P. A. M. van den Keybus, A. Vissink, and A. V. N. Amerongen. **Human glandular salivas: their separate collection and analysis**. *European Journal of Oral Sciences*, **104**(4):346–352, 1996. 8

[49] R. G. Schipper, E. Silletti, and M. H. Vingerhoeds. **Saliva as research material: Biochemical, physicochemical and practical aspects**. *Archives of Oral Biology*, **52**(12):1114–1135, 2007. 8

[50] H. A. Waterman, C. Blom, H. J. Holterman, E. J. s Gravenmade, and J. Mellema. **Rheological properties of human saliva**. *Archives of Oral Biology*, **33**(8):589–596, 1988. 8

[51] C. A. Francis, M. P. Hector, and G. B. Proctor. **Precipitation of specific proteins by freeze-thawing of human saliva**. *Archives of Oral Biology*, **45**(7):601–606, 2000. 8

[52] N. B. Parahitiyawa, C. Scully, W. K. Leung, W. C. Yam, L. J. Jin, and L. P. Samaranayake. **Exploring the oral bacterial flora: current status and future directions**. *Oral Diseases*, **16**(2):136–145, 2010. 8, 13, 14

[53] I R. Hamilton and G H. Bowden. *Chapter - Oral Microbiology*, pages 739–753. Elsevier Academic, Amsterdam, 2004. 8, 9, 14

[54] A. V. N. Amerongen and E. C. I. Veerman. **Saliva – the defender of the oral cavity**. *Oral Diseases*, **8**(1):12–22, 2002. 8

[55] D. L. Mager, L. A. Ximenez-Fyvie, A. D. Haffajee, and S. S. Socransky. **Distribution of selected bacterial species on intraoral surfaces**. *Journal of Clinical Periodontology*, **30**(7):644–654, 2003. 9

[56] P D. Marsh and M V. Martin. *Oral Microbiology*, pages 1–59. Churchill Livingstone, Edinburgh, 5th edition, 2009. 9, 11

[57] R. A. Whiley and D. Beighton. **Current classification of the oral streptococci**. *Oral Microbiology and Immunology*, **13**(4):195–216, 1998. 9

[58] N. Bente and K. Mogens. **Microbiology of the early colonization of human enamel and root surfaces in**

vivo. *European Journal of Oral Sciences*, **95**(5):369–380, 1987. 9

[59] C. Pearce, G. H. Bowden, M. Evans, S. P. Fitzsimmons, J. Johnson, M. J. Sheridan, R. Wientzen, and M. F. Cole. **Identification of pioneer viridans streptococci in the oral cavity of human neonates**. *Journal of Medical Microbiology*, **42**(1):67–72, 1995. 9

[60] J. D. Rudney and C. J. Larson. **Use of restriction fragment polymorphism analysis of rRNA genes to assign species to unknown clinical isolates of oral viridans streptococci.** *Journal of Clinical Microbiology*, **32**(2):437–443, 1994. 9, 100

[61] H. Wisplinghoff, R. R. Reinert, O. Cornely, and H. Seifert. **Molecular Relationships and Antimicrobial Susceptibilities of Viridans Group Streptococci Isolated from Blood of Neutropenic Cancer Patients**. *Journal of Clinical Microbiology*, **37**(6):1876–1880, 06 1999. 9, 100

[62] H. Shaw. *Etude des bactéries présentes dans la salive. Un potentiel pour de nouveaux tests indicatifs?* Master's thesis, Ecole de Sciences Criminelles, University of Lausanne, Switzerland, 2011. 9, 10, 11

[63] J. Li, I. Nasidze, D. Quinque, M. Li, H-P. Horz, C. Andre, R. Garriga, M. Halbwax, A. Fischer, and M. Stoneking. **The saliva microbiome of Pan and Homo**. *BMC Microbiology*, **13**(1):204, 2013. 11

[64] P. E. Kolenbrander, R. J. Palmer, S. Periasamy, and N. S. Jakubovics. **Oral multispecies biofilm development and the key role of cell–cell distance**. *Nat Rev Micro*, **8**(7):471–480, 2010. 11

[65] P. E Kolenbrander. **Multispecies communities: interspecies interactions influence growth on saliva as sole nutritional source**. *Int J Oral Sci*, **3**(2):49–54, 04 2011. 11

[66] E. M. Bik, C. D. Long, G. C. Armitage, P. Loomer, J. Emerson, E. F. Mongodin, K. E. Nelson, S. R. Gill, C. M. Fraser-Liggett, and D. A. Relman. **Bacterial diversity in the oral cavity of 10 healthy individuals**. *ISME J*, **4**(8):962–974, 08 2010. 11, 12

[67] D. Belstrøm, N. E. Fiehn, C. H. Nielsen, P. Holmstrup, N. Kirkby, V. Klepac-Ceraj, B. J. Paster, and S. Twetman. **Altered Bacterial Profiles in Saliva from Adults with Caries Lesions: A Case-Cohort Study**. *Caries Research*, **48**(5):368–375, 2014. 12, 111

[68] X. Ge, R. Rodriguez, M. Trinh, J. Gunsolley, and P. Xu. **Oral Microbiome of Deep and Shallow Dental Pockets In Chronic Periodontitis**. *PLoS ONE*, **8**(6):e65520, 2013. 12

[69] B. Liu, Li. L. Faller, N. Klitgord, V. Mazumdar, M. Ghodsi, D. D. Sommer, T. R. Gibbons, T. J. Treangen, Y-C. Chang, S. Li, O. C. Stine, H. Hasturk, S. Kasif, D. Segrè, M. Pop, and S. Amar. **Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease**. *PLoS ONE*, **7**(6):e37919, 2012. 12

[70] I. Nasidze, J. Li, D. Quinque, K. Tang, and M. Stoneking. **Global diversity in the human salivary microbiome**. *Genome Research*, **19**(4):636–643, 2009. 12, 66, 115, 116

[71] V. Lazarevic, K. Whiteson, D. Hernandez, P. Francois, and Ja. Schrenzel. **Study of inter- and intra-individual variations in the salivary microbiota**. *BMC Genomics*, **11**(1):523, 2010. 13, 64, 107, 110, 111, 115

[72] S. S. Socransky and S. D. Manganiello. **The Oral Microbiota of Man From Birth to Senility**. *Journal of Periodontology*, **42**(8):485–496, 1971. 13, 14

[73] V. Zijnge, M. B. M. van Leeuwen, J. E. Degener, F. Abbas, T. Thurnheer, R. Gmür, and H. J. M. Harmsen. **Oral Biofilm Architecture on Natural Teeth**. *PLoS ONE*, **5**(2):e9321, 2010. 13

[74] S. S. Stahringer, J. C. Clemente, R. P. Corley, J. Hewitt, D. Knights, W. A. Walters, R. Knight, and K. S. Krauter. **Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood**. *Genome Research*, **22**(11):2146–2152, 2012. 14, 64, 114, 115, 123

[75] C. Paradis-Bleau, G. Kritikos, K. Orlova, A. Typas, and T. G. Bernhardt. **A Genome-Wide Screen for Bacterial Envelope Biogenesis Mutants Identifies a Novel Factor Involved in Cell Wall Precursor Metabolism**. *PLoS Genet*, **10**(1):e1004056, 2014. 14

[76] L. Dethlefsen, M. McFall-Ngai, and D. A. Relman. **An ecological and evolutionary perspective on human–microbe mutualism and disease**. *Nature*, **449**(7164):811–818, 2007. 15

[77] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight. **Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample**. *Proceedings of the National Academy of Sciences*, 2010. 15

[78] M. Hamady and R. Knight. **Microbial community profiling for human microbiome projects: Tools, techniques, and challenges**. *Genome Research*, **19**(7):1141–1152, 2009. 15, 27

[79] Roche Diagnostics GmbH. **MagNA Pure 96 DNA and Viral NA Small Volume Kit**. *Version 04*, 2012. 18

[80] Z. Liu, T. Z. DeSantis, G. L. Andersen, and R. Knight. **Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers**. *Nucleic Acids Research*, **36**(18):e120, 2008. 18, 19

[81] Z. Liu, C. Lozupone, M. Hamady, F. D. Bushman, and R. Knight. **Short pyrosequencing reads suffice for accurate microbial community analysis**. *Nucleic Acids Research*, **35**(18):e120, 2007. 19

[82] A. F. Andersson, M. Lindberg, H. Jakobsson, F. Backhed, P. Nyren, and L. Engstrand. **Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing**. *Plos One*, **3**(7), 2008. 19

[83] T. Kanagawa. **Bias and artifacts in multitemplate polymerase chain reactions (PCR)**. *Journal of Bioscience and Bioengineering*, **96**(4):317–323, 2003. 19

[84] V. Lazarevic, K. Whiteson, S. Huse, D. Hernandez, L. Farinelli, M. Østerås, J. Schrenzel, and P. François. **Metagenomic study of the oral microbiota by Illumina high-throughput sequencing**. *Journal of Microbiological Methods*, **79**(3):266–271, 2009. 19, 20, 35, 44, 64

[85] H. Huber, M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter. **A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont**. *Nature*, **417**(6884):63–67, 2002. 19

[86] G. E. Fox, J. D. Wisotzkey, and P. Jurtshuk. **How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity**. *International Journal of Systematic Bacteriology*, **42**(1):166–170, 01 1992. 19

[87] A. Y. Pei, W. E. Oberdorf, C. W. Nossa, A. Agarwal, P. Chokshi, E. A. Gerz, Z. Jin, P. Lee, L. Yang, M. Poles, S. M. Brown, S. Sotero, T. DeSantis, E. Brodie, K. Nelson, and Z. Pei. **Diversity of 16S rRNA Genes within**

BIBLIOGRAPHY

Individual Prokaryotic Genomes. *Applied and Environmental Microbiology*, **76**(12):3886–3897, 06 2010. 19

[88] S. R. Santos and H. Ochman. **Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins**. *Environmental Microbiology*, **6**(7):754–759, 2004. 19

[89] T. Adekambi, M. Drancourt, and D. Raoult. **The rpoB gene as a tool for clinical microbiologists**. *Trends in Microbiology*, **17**(1):37–45, 2009. 19, 100

[90] K. J. Boor, M. L. Duncan, and C. W. Price. **Genetic and Transcriptional Organization of the Region Encoding the Subunit of Bacillus subtilis RNA Polymerase**. *Journal of Biological Chemistry*, **270**(35):20329–20336, 1995. 19

[91] T. Adekambi, P. Colson, and M. Drancourt. **rpoB-Based Identification of Nonpigmented and Late-Pigmenting Rapidly Growing Mycobacteria**. *Journal of Clinical Microbiology*, **41**(12):5699–5708, 2003. 20, 99

[92] B. L. Scola, L. T. M. Bui, G. Baranton, A. Khamis, and D. Raoult. **Partial rpoB gene sequencing for identification of Leptospira species**. *FEMS Microbiology Letters*, **263**(2):142–147, 2006. 20, 99

[93] D. E. Hunt, V. Klepac-Ceraj, S. G. Acinas, C. Gautier, , S. Bertilsson, and M. F. Polz. **Evaluation of 23S rRNA PCR Primers for Use in Phylogenetic Studies of Bacterial Diversity**. *Applied and Environmental Microbiology*, **72**(3):2221–2225, 2006. 20

[94] K. Trebesius, D. Harmsen, A. Rakin, J. Schmelz, and J. Heesemann. **Development of rRNA-Targeted PCR and In Situ Hybridization with Fluorescently Labelled Oligonucleotides for Detection of Yersinia Species**. *Journal of Clinical Microbiology*, **36**(9):2557–2564, 1998. 20

[95] J. Kluytmans, A. van Belkum, and H. Verbrugh. **Nasal carriage of Staphylococcus aureus: epidemiology, underlying mechanisms, and associated risks.** *Clinical Microbiology Reviews*, **10**(3):505–520, 07 1997. 20

[96] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman. **Diversity of the Human Intestinal Microbial Flora**. *Science*, **308**(5728):1635–1638, 06 2005. 20

[97] B. Gao and R. S. Gupta. **Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria**. *Microbiology and Molecular Biology Reviews*, **76**(1):66–112, 03 2012. 20

[98] H. M. Wexler. **Bacteroides: the Good, the Bad, and the Nitty-Gritty**. *Clinical Microbiology Reviews*, **20**(4):593–621, 10 2007. 20

[99] D. J. Brenner, D. G. Hollis, C. W. Moss, C. K. English, G. S. Hall, J. Vincent, J. Radosevic, K. A. Birkness, W. F. Bibb, and F. D. Quinn. **Proposal of Afipia gen. nov., with Afipia felis sp. nov. (formerly the cat scratch disease bacillus), Afipia clevelandensis sp. nov. (formerly the Cleveland Clinic Foundation strain), Afipia broomeae sp. nov., and three unnamed genospecies.** *Journal of Clinical Microbiology*, **29**(11):2450–2460, 11 1991. 20

[100] G. H. W. Bowden and I. R. Hamilton. **Survival of Oral Bacteria**. *Critical Reviews in Oral Biology and Medicine*, **9**(1):54–85, 1998. 20

[101] A. Molander, C. Reit, G. Dahlén, and T. Kvist. **Microbiological status of root-filled teeth with apical periodontitis**. *International Endodontic Journal*, **31**(1):1–7, 1998. 21

[102] R. Podschun and U. Ullmann. **Klebsiella spp. as Nosocomial Pathogens: Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors**. *Clinical Microbiology Reviews*, **11**(4):589–603, 10 1998. 21

[103] A. Hejazi and F. R. Falkiner. **Serratia marcescens**. *Journal of Medical Microbiology*, **46**(11):903–912, 11 1997. 21

[104] R. Keller, M. Z. Pedroso, R. Ritchmann, and R. M. Silva. **Occurrence of Virulence-Associated Properties in Enterobacter cloacae**. *Infection and Immunity*, **66**(2):645–649, 02 1998. 21

[105] T. Adekambi and M. Drancourt. **Dissection of phylogenetic relationships among 19 rapidly growing Mycobacterium species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing**. *International Journal of Systematic and Evolutionary Microbiology*, **54**(6):2095–2105, 2004. 21

[106] F. Corpet. **Multiple sequence alignment with hierarchical clustering**. *Nucleic Acids Research*, **16**(22):10881–10890, 1988. 21, 44, 45, 46, 47

[107] **http://www.promega.com/a/apps/biomath** [online, cited 27th January 2011]. 22

[108] W. A. Kibbe. **OligoCalc: an online oligonucleotide properties calculator**. *Nucleic Acids Research*, **35**(suppl 2):W43–W46, 07 2007. 22

[109] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. **Basic local alignment search tool**. *Journal of Molecular Biology*, **215**:403 – 410, 1990. 22, 35, 47

[110] G. Ficetola, E. Coissac, S. Zundel, T. Riaz, W. Shehzad, J. Bessiere, P. Taberlet, and F. Pompanon. **An In silico approach for the evaluation of DNA barcodes**. *Bmc Genomics*, **11**(1):434, 2010. 22, 46

[111] J. Handelsman. **Metagenomics: Application of Genomics to Uncultured Microorganisms**. *Microbiol. Mol. Biol. Rev.*, **68**(4):669–685, 2004. 25

[112] F. Sanger, S. Nicklen, and A. R. Coulson. **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12):5463–5467, 1977. 25

[113] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood. **Fluorescence detection in automated DNA-sequence analysis**. *Nature*, **321**(6071):674–679, 1986. 25

[114] J. M. Prober, G. L. Trainor, R. J. Dam, F. W. Hobbs, C.W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A Jensen, and K. Baumeister. **A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides**. *Science*, **238**(4825):336–341, 1987. 25

[115] R. S. Mandabhushi. **Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions**. *ELECTROPHORESIS*, **19**(2):224–230, 1998. 25

[116] National Human Genome Research Institute. **The Human Genome Project Completion: Frequently Asked Questions**. *http://www.genome.gov/11006943*, 2003. 26

[117] S. T. Bennett, C. Barnes, A. Cox, L. Davies, and Clive Brown. **Toward the $1000 human genome**. *Pharmacogenomics*, **6**(4):373–382, 2005. 26

[118] N. Hall. **Advanced sequencing technologies and their wider impact in microbiology**. *Journal of Experimental Biology*, **210**(9):1518–1525, 2007. 26

[119] D. MacLean, J.D.G. Jones, and D. J. Studholme. **Application of 'next-generation' sequencing technologies**

to microbial genetics. *Nature Reviews Microbiology*, **7**(4):287 – 296, 2009. 27, 29

[120] H. Peng and J. Zhang. **Commercial high-throughput sequencing and its applications in DNA analysis**. *Biologia*, **64**(1):20–26, 2009. 28

[121] Life Technologies. **Scalable, simple, fast microbial sequencing solutions**, 2013. 28

[122] Life Technologies. **Sequencing for all.**, 2012. 28

[123] M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu. **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers**. *BMC Genomics*, **13**(1):341, 2012. 28

[124] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. **Real-Time DNA Sequencing from Single Polymerase Molecules**. *Science*, **323**(5910):133–138, 2009. 30

[125] P. Coupland, T. Chandra, M. Quail, W. Reik, and H. Swerdlow. **Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation**. *Biotechniques*, **53**(6):365–372, 2012. 30

[126] G. F. Schneider and C. Dekker. **DNA sequencing with nanopores**. *Nat Biotech*, **30**(4):326–328, 04 2012. 30

[127] Illumina, Inc. **TruSeq DNA Sample Prep Kits**. Technical report, Illumina, 2012. 30

[128] **Fastqc: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/** [cited November 1st 2013]. 31

[129] Illumina, Inc. **Genome analyzer data analysis software**, 2008. 32

[130] T. Magoç and S. L. Salzberg. **FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies**. *Bioinformatics*, 2011. 31

[131] **http://hannonlab.cshl.edu/fastx_toolkit/index.html**. 32

[132] V. K. Sharma, N. Kumar, T. Prakash, and T. D. Taylor. **Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin**. *PLoS ONE*, **7**(4):e34030, 04 2012. 33

[133] D. Wei, Q. Jiang, Y. Wei, and S. Wang. **A novel hierarchical clustering algorithm for gene sequences**. *BMC Bioinformatics*, **13**(1):174, 2012. 33, 58

[134] A. Khamis, D. Raoult, and B. La Scola. **Comparison between rpoB and 16S rRNA Gene Sequencing for Molecular Identification of 168 Clinical Isolates of Corynebacterium**. *Journal of Clinical Microbiology*, **43**(4):1934–1936, 04 2005. 33, 66

[135] W. Chen, C. K. Zhang, Y. Cheng, S. Zhang, and H. Zhao. **A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs**. *PLoS ONE*, **8**(8):e70837, 08 2013. 33, 105

[136] P. D. Schloss and S. L. Westcott. **Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis**. *Applied and Environmental Microbiology*, **77**(10):3219–3226, 05 2011. 34

[137] X. Hao, R. Jiang, and T. Chen. **Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering**. *Bioinformatics*, 2011. 34

[138] W. Li, L. Jaroszewski, and A. Godzik. **Clustering of highly homologous sequences to reduce the size of large protein databases**. *Bioinformatics*, **17**(3):282–283, 2001. 34

[139] W. Li and A. Godzik. **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics*, **22**(13):1658–1659, 2006. 34

[140] *The NCBI Handbook, 2nd edition*. National Center for Biotechnology Information, 2013. 36, 106

[141] S. Anders and W. Huber. **Differential expression analysis for sequence count data**. *Genome Biology*, **11**(10):R106, 2010. 37, 70

[142] J. C. Marioni, C. E. Mason, Shrikant M. Mane, M. Stephens, and Y. Gilad. **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Research*, **18**(9):1509–1517, 2008. 37

[143] L. Wang, Z. Feng, X. Wang, X. Wang, and X. Zhang. **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data**. *Bioinformatics*, **26**(1):136–138, 01 2010. 37

[144] **Deseq: http://www.bioconductor.org/packages/release/bioc/html/DESeq.html** [cited 12th July 2013]. 38

[145] J. Townend. *Practical statistics for environmental and biological scientists*. Wiley-Blackwell, 2002. 39

[146] F. Taroni, S. Bozza, A. Biedermann, C. Aitken, and P. Garbolino. *Data analysis in forensic science: A Bayesian decision perspective*. John Wiley and Sons, Chichester, 2010. 40

[147] **Hierarchical clustering: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html** [cited 12th July 2013]. 40

[148] **http://stat.ethz.ch/R-manual/R-patched/library/stats/html/dendrogram.html**. 40

[149] J. R. Cole, B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. **The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis**. *Nucleic Acids Research*, **33**:D294 – D296, 2005. 48

[150] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB**. *Nucleic Acids Research*, **35**(21):7188–7196, 2007. 48

[151] T. Chen, W-H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, , and F. E. Dewhirst. **The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information**. *Database*, **2010**. 48

[152] M. Vos, C. Quince, A. S. Pijl, M. de Hollander, and G. A. Kowalchuk. **A Comparison of rpoB and 16S rRNA as Markers in Pyrosequencing Studies of Bacterial Diversity**. *PLoS ONE*, **7**(2):e30600, 2012. 58, 100, 101

BIBLIOGRAPHY

[153] P. I. Diaz, N. I. Chalmers, A. H. Rickard, C. Kong, C. L. Milburn, R. J. Palmer, and P. E. Kolenbrander. **Molecular Characterization of Subject-Specific Oral Microflora during Initial Colonization of Enamel**. *Applied and Environmental Microbiology*, **72**(4):2837–2848, 2006. 66

[154] R. J. Case, Y. Boucher, I. Dahllof, C. Holmstrom, W. F. Doolittle, and S. Kjelleberg. **Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies**. *Applied and Environmental Microbiology*, **73**(1):278–288, 2007. 66

[155] D. Sweet, M. Lorente, J. A. Lorente, A. Valenzuela, and E. Villanueva. **An Improved Method to Recover Saliva from Human Skin: The Double Swab Technique**. *Journal of Forensic Sciences*, **42**(2):320–2, 1997. 98

[156] L. Abusleme, B-Y. Hong, A. Dupuy, L. Strausbaugh, and P. Diaz. **Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing**. *Journal of Oral Microbiology*, **6**(0), 2014. 98

[157] A. Edelmann, U. Eichenlaub, S. Lepek, D. H. Krüger, and J. Hofmann. **Performance of the MagNA Pure 96 System for Cytomegalovirus Nucleic Acid Amplification Testing in Clinical Samples**. *Journal of Clinical Microbiology*, **51**(5):1600–1601, 05 2013. 98

[158] J. Rajendhran and P. Gunasekaran. **Microbial phylogeny and diversity: Small subunit ribosomal RNA sequence analysis and beyond**. *Microbiological Research*, **166**(2):99 – 110, 2011. 99

[159] Y-H. Lin, B. C.H. Chang, P-W. Chiang, and S-L. Tang. **Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes**. *Gene*, **416**(1–2):44 – 47, 2008. 99

[160] Y. Wang, Z. Zhang, and N. Ramanan. **The actinomycete Thermobispora bispora contains two distinct types of transcriptionally active 16S rRNA genes.** *Journal of Bacteriology*, **179**(10):3270–3276, 05 1997. 99

[161] L. Kraal, S. Abubucker, K. Kota, M. A. Fischbach, and M. Mitreva. **The Prevalence of Species and Strains in the Human Microbiome: A Resource for Experimental Efforts**. *PLoS ONE*, **9**(5):e97279, 2014. 100

[162] D. P. Chandler, F. J. Brockman, T. J. Bailey, and J. K. Fredrickson. **Phylogenetic Diversity of Archaea and Bacteria in a Deep Subsurface Paleosol**. *Microbial Ecology*, **36**(1):37–50, 1998. 100

[163] P. D. Schloss, D. Gevers, and S. L. Westcott. **Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies**. *PLoS ONE*, **6**(12):e27310, 2011. 103

[164] 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. Methé, T. Z. DeSantis, J. F. The Human Microbiome Consortium, Petrosino, R. Knight, and B. W. Birren. **Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons**. *Genome Research*, **21**(3):494–504, 2011. 103

[165] N. A. Hasan, B. A. Young, A. T. Minard-Smith, K. Saeed, H. Li, E. M. Heizer, N. J. McMillan, R. Isom, A. S. Abdullah, D. M. Bornman, S. A. Faith, S. Y. Choi, M. L. Dickens, T. A. Cebula, and R. R. Colwell. **Microbial Community Profiling of Human Saliva Using Shotgun Metagenomic Sequencing**. *PLoS ONE*, **9**(5):e97699, 2014. 103, 115

[166] **http://www.swiftbiosci.com/products/accel-ngs-2s-dna-library-kit-for-illumina-performance**. 103

[167] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis. **Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample**. *PLoS ONE*, **7**(2):e30087, 2012. 104

[168] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. **Comparison of Next-Generation Sequencing Systems**. *Journal of Biomedicine and Biotechnology*, 2012. 104

[169] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen. **Performance comparison of benchtop high-throughput sequencing platforms**. *Nat Biotech*, **30**(5):434–439, 05 2012. 104

[170] W. Li, L. Jaroszewski, and A. Godzik. **Tolerating some redundancy significantly speeds up clustering of large protein databases**. *Bioinformatics*, **18**(1):77–82, 2002. 105

[171] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. **NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy**. *Nucleic Acids Research*, **40**(D1):D130–D135, 2012. 106

[172] K. Li, M. Bihan, and B. A. Methe. **Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome**. *PLoS ONE*, **8**(5), 2013. 107, 108

[173] S. M. Huse, Y. Ye, Y. Zhou, and A. A. Fodor. **A Core Human Microbiome as Viewed through 16S rRNA Sequence Clusters**. *PLoS ONE*, **7**(6):e34242, 2012. 107, 108

[174] R. R. Sokal and P. H. A. Sneath. *Principles of Numerical Taxonomy*. Freeman, San Francisco, 1963. 109

[175] S. Hands and B. Everitt. **A Monte Carlo Study of the Recovery of Cluster Structure in Binary Data by Hierarchical Clustering Techniques**. *Multivariate Behavioral Research*, **22**(2):235–243, 1987. 109

[176] F-J. Lapointe and P. Legendre. **Comparison tests for dendrograms: A comparative evaluation**. *Journal of Classification*, **12**(2):265–282, 1995. 109

[177] S. Saracli, N. Dogan, and I. Dogan. **Comparison of hierarchical cluster analysis methods by cophenetic correlation**. *Journal of Inequalities and Applications*, **2013**(1):203, 2013. 109

[178] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. Freeman, San Francisco, 1973. 109

[179] I. Nasidze, D. Quinque, J. Li, M. K. Li, K. Tang, and M. Stoneking. **Comparative analysis of human saliva microbiome diversity by barcoded pyrosequencing and cloning approaches**. *Analytical Biochemistry*, **391**(1):64–68, 2009. 110

[180] S. L. Leake. **Is human DNA enough? - Potential for bacterial DNA**. *Frontiers in Genetics*, **4**(238), 2013. 111

[181] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. **A bioinformatician's guide to metagenomics**. *Microbiology Molecular Biology Reviews*, **72**(4):557 – 578, 2008. 111

[182] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight. **Microbial community resemblance methods differ in their ability to detect biologically relevant patterns**. *Nat Meth*, **7**(10):813–819, 10 2010. 111

[183] V. Lazarevic, K. Whiteson, P. François, and J. Schrenzel. **The salivary microbiome, assessed by a high-throughput and culture-independent approach**. *Journal of Integrated OMICS*, **1**, 2010. 112, 115

[184] B. Budowle, A. J. Onorato, T. F. Callaghan, A. D. Manna, A. M. Gross, R. A. Guerrieri, J. C. Luttman, and D. L. McClure. **Mixture Interpretation: Defining the Relevant Features for Guidelines for the Assessment of Mixed DNA Profiles in Forensic Casework***. *Journal of Forensic Sciences*, **54**(4):810–821, 2009. 113

[185] P. M. A. Corby, W. A. Bretz, T. C. Hart, M. Melo Filho, B. Oliveira, and M. Vanyukov. **Mutans streptococci in preschool twins**. *Archives of Oral Biology*, **50**(3):347–351, 3 2005. 114

[186] P. M. Corby, W. A. Bretz, T. C. Hart, N. J. Schork, J. Wessel, J. Lyons-Weiler, and B. J. Paster. **Heritability of Oral Microbial Species in Caries-Active and Caries-Free Twins**. *Twin Research and Human Genetics*, **10**:821–828, 2007. 114

[187] L. Dethlefsen, S. Huse, M. L Sogin, and D. A Relman. **The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing**. *PLoS Biol*, **6**(11):e280, 2008. 114

[188] L. Dethlefsen and D. A. Relman. **Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation**. *Proceedings of the National Academy of Sciences*, 2010. 114

[189] H. E. Jakobsson, C. Jernberg, A. F. Andersson, M. Sjölund-Karlsson, J. K. Jansson, and L. Engstrand. **Short-Term Antibiotic Treatment Has Differing Long-Term Impacts on the Human Throat and Gut Microbiome**. *PLoS ONE*, **5**(3):e9836, 2010. 114

[190] V. Lazarevic, S. Manzano, N. Gaïa, M. Girard, K. Whiteson, J. Hibbs, P. François, A. Gervaix, and J. Schrenzel. **Effects of amoxicillin treatment on the salivary microbiota in children with acute otitis media**. *Clinical Microbiology and Infection*, **19**(8):E335–E342, 2013. 114

[191] T. Ding and P. D. Schloss. **Dynamics and associations of microbial community types across the human body**. *Nature*, **509**:357–360, 2014. 115

[192] D. Belstrøm, P. Holmstrup, C. Nielsen, N. Kirkby, S. Twetman, B. Heitmann, V. Klepac-Ceraj, B. Paster, and N-E. Fiehn. **Bacterial profiles of saliva in relation to diet, lifestyle factors, and socioeconomic status**. *Journal of Oral Microbiology*, **6**(0), 2014. 115

[193] A. Benitez-Paez, P. Belda-Ferre, A. Simon-Soro, and A. Mira. **Microbiota diversity and gene expression dynamics in human oral biofilms**. *BMC Genomics*, **15**(1):1471–2164, 2014. 115

[194] R. Francavilla, D. Ercolini, M. Piccolo, L. Vannini, S. Siragusa, F. De Filippis, I. De Pasquale, R. Di Cagno, M. Di Toma, G. Gozzi, D. I. Serrazanetti, M. De Angelis, and M. Gobbetti. **Salivary microbiota and metabolome associated with celiac disease**. *Applied and Environmental Microbiology*, 2014. 116

[195] S. J. Song, C. Lauber, E. K. Costello, C. A. Lozupone, G. Humphrey, D. Berg-Lyons, J. G. Caporaso, D. Knights, J. C. Clemente, S. Nakielny, J. I. Gordon, N. Fierer, and R. Knight. **Cohabiting family members share microbiota with one another and with their dogs**. *eLife*, **2**, 2013. 116

[196] J. G. Kang, S. H. Kim, and T. Y. Ahn. **Bacterial diversity in the human saliva from different ages**. *Journal of Microbiology*, **44**(5):572–576, 2006. 116

[197] J. V. Pereira, L. Leomil, F. Rodrigues-Albuquerque, J. O. Pereira, and S. Astolfi-Filho. **Bacterial diversity in the saliva of patients with different oral hygiene indexes**. *Brazilian Dental Journal*, **23**:409 – 416, 2012. 117

[198] M. Matsui, N. Chosa, Y. Shimoyama, K. Minami, S. Kimura, and M. Kishi. **Effects of tongue cleaning on bacterial flora in tongue coating and dental plaque: a crossover study**. *BMC Oral Health*, **14**(1):4, 2014. 117

[199] R. S. Murch, E. L. Bahr, B. Budowle, S. E. Schutzer, R. G. Breeze, P. S. Keim, and S. A. Morse. *Chapter 38 - Validation of Microbial Forensics in Scientific, Legal, and Policy Contexts*, pages 649–663. Academic Press, San Diego, 2011. 117

[200] N. Gligorov, J. Azzouni, D. P. Lackey, and A Zweig. *The Human Microbiome: Ethical, Legal and Social Concerns*, chapter 2 - Personal Identity, pages 55–70. Oxford University Press, 2013. 117, 118

[201] L. Zhao. **Genomics: The tale of our other genome**. *Nature*, **465**:879–880, 2010. 118

[202] N. Gligorov, L. E. Frank, A. P. Schwab, and B. Trusko. *The Human Microbiome: Ethical, Legal and Social Concerns*, chapter 4 - Privacy, Confidentiality, and New Ways of Knowing More, pages 107–127. Oxford University Press, 2013. 118

[203] M. W. Eshoo, J. Picuri, D. D. Duncan, D. J. Ecker, B. Budowle, S. E. Schutzer, R. G. Breeze, P. S. Keim, and S. A. Morse. *Chapter 10 - Microbial Forensic Analysis of Trace and Unculturable Specimens*, pages 155–171. Academic Press, San Diego, 2011. 120

[204] N. J. Parkinson, S. Maslau, B. Ferneyhough, G. Zhang, L. Gregory, D. Buck, J. Ragoussis, C. P. Ponting, and M. D. Fischer. **Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA**. *Genome Research*, **22**(1):125–133, 2012. 120

[205] S. L. Fordyce, M. C. Ávila-Arcos, E. Rockenbauer, C. Børsting, R. Frank-Hansen, F. T. Petersen, E. Willerslev, A. J. Hanse, N. Morling, and Gilbert. M. T. P. **High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform**. *Biotechniques*, **51**:127–133, 2011. 122

[206] C. Van Neste, F. Van Nieuwerburgh, D. Van Hoofstat, and D. Deforce. **Forensic STR analysis using massive parallel sequencing**. *Forensic Science International: Genetics*, (0), 2012. 122

[207] D. H. Warshauer, D. Lin, K. Hari, R. Jain, C. Davis, B. LaRue, J. L. King, and B. Budowle. **STRait Razor: A length-based forensic STR allele-calling tool for use with second generation sequencing data**. *Forensic Science International: Genetics*, **7**(4):409–417, 2013. 122

[208] Illumina. **Targeted Next-Generation Sequencing for Forensic Genomics**. PDF on website (www.illumina.com), 2013. 122

Appendix A

8.1 *Streptococci* species/strains found by *rpoB*1 and 16S rRNA

Table 8.1: *Streptococci* species/strains found by *rpoB*1 and 16S rRNA

RpoB1	16S rRNA
Streptococcus agalactiae GD201008-001	*Streptococcus constellatus clone*
Streptococcus anginosus isolate VS113A	*Streptococcus constellatus gene*
Streptococcus anginosus	*Streptococcus mutans clone WWC_C1MKM077*
Streptococcus anginosus strain CIP	*Streptococcus salivarius strain HNL13*
Streptococcus constellatus	*Streptococcus sp. 2944*
Streptococcus cristatus	*Streptococcus sp. LMG 27206*
Streptococcus dysgalactiae subsp.	*Streptococcus sp. LVRI_101*
Streptococcus equi subsp. zooepidemicus	*Streptococcus sp. oral taxon 071*
Streptococcus gordonii str. Challis	*Streptococcus thoraltensis strain*
Streptococcus infantarius subsp.	
Streptococcus intermedius JTH08	
Streptococcus macedonicus ACA-DC	
Streptococcus mitis B6 complete	
Streptococcus mitis isolate VS779	
Streptococcus mitis RNA polymerase	
Streptococcus mutans UA159	
Streptococcus oralis isolate VS113B	
Streptococcus oralis isolate VS2971R	
Streptococcus oralis isolate VS2971S	
Streptococcus oralis isolate VS745	
Streptococcus oralis isolate VS79	
Streptococcus oralis strain ATCC 10557	
Streptococcus oralis Uo5 complete	
Streptococcus parasanguinis ATCC	
Streptococcus parasanguinis FW213	
Streptococcus parasanguinis isolate	
Streptococcus pneumoniae 670-6B	
Streptococcus pneumoniae gamPNI0373	
Streptococcus pneumoniae R6, complete	

Streptococcus pneumoniae rpoB gene	
Streptococcus pneumoniae SPNA45	
Streptococcus pneumoniae ST556	
Streptococcus pneumoniae strain NCTC	
Streptococcus pneumoniae strain RifR-13	
Streptococcus pneumoniae strain RifR-16	
Streptococcus pneumoniae strain RifR-24	
Streptococcus pneumoniae strain RifR-25	
Streptococcus pneumoniae strain RifR-31	
Streptococcus pneumoniae strain RifR-56	
Streptococcus pneumoniae strain RifR-65	
Streptococcus pseudopneumoniae IS7493	
Streptococcus pyogenes A20	
Streptococcus salivarius 57.I	
Streptococcus salivarius CCHSS3	
Streptococcus salivarius JIM8777	
Streptococcus sanguinis isolate VS395	
Streptococcus sanguinis	
Streptococcus sanguinis SK36	
Streptococcus sp. CSL 7508	
Streptococcus suis S735	
Streptococcus suis ST1	
Streptococcus thermophilus MN-ZLW-002	
Streptococcus uberis 0140J	

8.2 Filtering at 10 sequences

In order to justify the choice of only keeping clusters containing 20 sequences or more, the data was re-analysed keeping all clusters containing 10 sequences or more, to see if any differences were observed. Table 8.2 compares the relative distance between individuals for each experiment separately and both experiments combined. Only species with a p-value <0.1 from a t-test between the samples from each individual or a BF <1 were used. The differences between the relative distances from samples filtered at 20 sequences or 10 sequences are neglible indicating that the conservative approach of filtering at 20 sequences is justified. Lowering the filtering threshold to 10 sequences risks adding error for no reason. Table 8.3 shows the same comparison but only for both experiments combined with species filtered at a p-value <0.01 and for each target separately and pairwise combinations of targets. This corroborates the results presented above along with the choice of p-value presented in section 6.2.1.

Target	Filtered 20	Filtered 10
Experiment 1		
***RpoB*1**	21.53	21.71
***RpoB*2**	15.26	16.41
16S rRNA	44.06	48.36
Experiment 2		
***RpoB*1**	25.36	25.36
***RpoB*2**	22.59	23.19
16S rRNA	51.25	54.66
Experiments comb.		
***RpoB*1**	22.44	22.85
***RpoB*2**	15.13	15.79
16S rRNA	54.22	57.69

Table 8.2: Comparison of relative distance between individuals, per target, for each experiment separately and both experiments combined. The relative distance was calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised (only for both experiments combined) and logged species abundance. Only species with a p-value <0.1 from a t-test between the samples from each individual or a BF <1 were used.

Target	Filtered 20	Filtered 10
***RpoB*1**	25.38	25.35
***RpoB*2**	13.81	15.30
16S rRNA	57.36	60.91
***RpoB*1+*rpoB*2**	32.02	29.70
***RpoB*2+16S rRNA**	59.81	62.84
16S rRNA+*rpoB*1	60.90	66.24

Table 8.3: Comparison of relative distance between individuals, per target and combined targets, for both experiments combined. The relative distance was calculated using the Euclidean distance and the Ward method of hierarchical clustering, on the normalised and logged species abundance. Only species with a p-value <0.01 from a t-test between the samples from each individual or a BF <1 were used.

8.3 Cophenetic distance

The cophenetic distance can be used to test how accurately a dendrogram represents the data. The closer the value is to one the better the representation. The cophenetic distance was calculated for all dendrograms and the results are presented in Table 8.4. Nearly all the values are over 0.99 indicating the dendrograms accurately represent the data. The lowest value is seen for *rpoB*2, for both experiments combined, indicating that the separation provided by *rpoB*2 is not as reliable as that of *rpoB*1 and 16S rRNA. This further confirms the proposition made in section 6.5.2 that the best combination of targets

is *rpoB*1 and 16S rRNA.

Target	Cophenetic distance
Experiment 1	
***RpoB*1**	0.9937
***RpoB*2**	0.9938
16S rRNA	0.9933
Experiment 2	
***RpoB*1**	0.9917
***RpoB*2**	0.9956
16S rRNA	0.9984
Experiments comb.	
***RpoB*1**	0.9716
***RpoB*2**	0.8028
16S rRNA	0.9948
16S rRNA+*rpoB*1	0.9933

Table 8.4: Cophenetic distance for dendrograms from the hierarchical clustering of individual experiments and both experiments combined, per target.

Appendix B

8.4 Protocols

8.4.1 PCR protocol

Shown in Table 8.5 are the quantites used for each component of the PCR mix, along with their final concentrations for one reaction. Subsequently, shown in Table 8.6 are the cycling parameters which were used for all PCR amplifications. For the annealing step, the melting temperatures of the primer pairs were used, hence this value changes depending on the primer pair used, (the melting temperatures for each primer pair can be found in table 3.2).

Mix	**1x**	**Final conc.**
Water	22 μl	-
dNTP (2mM)	5 μl	0.2 mM
HF buffer (5x)	10 μl	1x
$MgCl_2$ (50mM)	1 μl	1 mM
Forward primer	2.5 μl	0.5 μM
Reverse primer	2.5 μl	0.5 μM
DNA	5 μl	-
DMSO	1.5 μl	3 %
Phusion polymerase	0.5 μl	0.02 U/μl

Table 8.5: PCR mix components for one reaction - where applicable the final concentration of each component is included.

8.4.2 Acrylamide gel

The following protocol was used to analyse the PCR products on an acrylamide gel:

1. Make sure all glassware is clean (no dried polyacrylamide)
2. Take 1x plate with spacer and 1x glass plate and place together with bottoms of both plates flat and level.

Temperature (°C)	Time (seconds)
98	30
98	5
Tm	15
72	10
72	300
4	∞

Table 8.6: PCR cycling parameters - Tm = melting temperature for the primer pair. The middle 3 parameters were repeated for 35 cycles.

3. Place plates in plastic stand, ensuring plates are level then place the stand in the stand grip making sure the plates are sealed.
4. Mix together all components (as shown in Table 8.7) adding the TEMED last, once the TEMED has been added need to work fast.
5. Pipette mix into plate gap, fill to top, then add comb and wipe off any excess mix that spills over.
6. Once gel has polymerised can then place it in the gel block (if only running one gel need a plastic plate to finish the connection)
7. Add 1x TBE buffer making sure to fill up inside the gel holder as well as outside.
8. Remove comb/s.
9. Prepare samples for gel: 2 μl loading buffer and 5 μl sample, then load samples into gel wells, including the ladder.
10. Attach power supply and run gel. In this case gels were run at 90V for 4.5 hours.
11. To stain gels add 2 drops of gel red and 1x TBE buffer and leave for at least 10 minutes.
12. Drain off buffer, rinse with fresh buffer then the gel is ready to be photographed using UV light.

Gel mix	1.5x
Acrylamide mix (30%)	10.4 ml
Water	0.85 ml
TBE (10x)	1.25 ml
APS (10%)	81.25 μl
TEMED	16.25 μl

Table 8.7: Acrylamide gel mix components - 1.5x is used to ensure there is enough mix for one complete gel.

8.5 Scripts

8.5.1 filter_cluster.py

filter_cluster.py

```
# import libraries

import numpy as np  # numerical python
import sys # needed for passing arguments
import re # needed for regular expressions
import os # for creating files, making directories etc

################################################
# USAGE:  python example.py <in_fn1> <in_fn2> <out_fn2> <out_3>
#fn1 - corresponds to file with multiple sequences >97%
#fn2 - corresponds to * sequence from each cluster

f=open(sys.argv[1],'r')

# create a dictionary lookup to reduce memory load
f1=open(sys.argv[2],'r')

toggle_rd_ln = True
seq_ref_dict = {}

while toggle_rd_ln:
    tmp_ln = f1.readline()
    #print tmp_ln
    if tmp_ln.find('>')!=-1: # we have found a seq ID line
      # generate dict and save any existing data
      try:
        seq_ref_dict[seq[:19]] = seq_GTAC
        print 'new key added to dict: ',seq[:19]
      except:
        print 'initialise sequence'
      seq = tmp_ln.split('>')[1].split(' ')[0]
      seq_GTAC = tmp_ln.split('>')[1].split(' ')[1]#.split('\n')
          [0]
    else:
      seq_GTAC += tmp_ln.split('\n')[0] # explicit could also just
          .split()
    # catch an empty redline command and exit
    if tmp_ln=='':
      toggle_rd_ln = False

f1_out=open(sys.argv[3],'w')
f2_out=open(sys.argv[4],'w')
f2_out.write('Cluster\t')
samples_nm = ['A1','A2','B1','B2']
for item1 in samples_nm:
   f2_out.write(item1+'\t')
f2_out.write('Species\t\n')
```

```
46
47 toggle_rd_ln = True
48 toggle_rd_clstr = True
49 iter=0
50 iter1=0
51
52 while toggle_rd_ln:
53     while toggle_rd_clstr:
54       tmp_ln = f.readline()
55       if tmp_ln.find('Cluster')!=-1: # we have found a cluster ID
             line
56         # start a new cluster
57         # write any data that exists
58         try:
59           #print iter
60           if iter>=19: #edit this number to change the sequence
                filtering threshold
61             print 'saved:','>'+rep_seq2find
62             #print clstr_no, samples_dict
63             #print tmp_ln
64             try:
65               # gets a bit silly because we don't have whole
                    sequence
66               # I had to manually edit the key length :19! beware
67               str1 = '>'+rep_seq2find+' '+seq_ref_dict[
                    rep_seq2find]+'\n'
68               f1_out.write(str1)
69               #print 'f1out',str1
70             except:
71               print 'check your dictionary'
72             f2_out.write('%i\t'%clstr_no)
73             #print samples_dict.keys()
74             #for key in samples_dict:
75             for nm in samples_nm:
76               f2_out.write('%i\t'%samples_dict[nm])
77             f2_out.write('species')
78             f2_out.write('\n')
79         except:
80           # no cluster exists, i.e 1st iteration
81           print "Initialise cluster "
82         # find cluster ID
83         clstr_no = int(re.findall(r'\d+',tmp_ln)[0])
84         print 'Cluster: ',clstr_no
85         # clear dictionaries and counters
86         iter=0
87         toggle_rd_clster = True
88         samples_dict={}
89         for key in samples_nm:
90           samples_dict[key]=0
91       else:
92         iter+=1
93         #print tier
94         # search for representative sequence
```

```
95          rep_seq = re.findall(r'.\D{1}\d{1}_\D+_\d+.\d+\.\.\.\s\*',
               tmp_ln) #gramn and strep
96          #rep_seq = re.findall(r'.\D{1}\d{1}_\d+\D+_\d+.\d+\.\.\.\s
               \*',tmp_ln) # 16S
97          if len(rep_seq)!=0:
98            # extract representative sequence to reference
99            rep_seq2find = re.findall(r'\D{1}\d{1}_\D+_\d+.\d+',
                 rep_seq[0])[0]
100           #rep_seq2find = re.findall(r'\D{1}\d{1}_\d+\D+_\d+.\d+',
                 rep_seq[0])[0]
101           #print rep_seq
102           #print 'rep_seq',rep_seq2find
103         # for access to all the sequences
104         try:
105           seq_curr_iter = re.findall(r'\D{1}\d{1}_\D+_\d+.\d+',
                 tmp_ln)[0]
106           #seq_curr_iter = re.findall(r'\D{1}\d{1}_\d+\D+_\d+.\d
                 +',tmp_ln)[0]
107           #print seq_curr_iter
108         except:
109           print 'failed to find sequence for this cluster'
110           iter1+=1
111         # search for samples present in cluster
112         try:
113           seq_sample= re.findall(r'.\D{1}\d{1}_',tmp_ln)
114         except:
115           print 'reg. exp. error: the sample name is not
                 compatible'
116         # update the dictionary
117         try:
118           samples_dict[seq_sample[0].split('>')[1].split('_')
                 [0]]+=1
119         except:
120           toggle_rd_clstr = False
121       # catch an empty readline command and exit
122       if tmp_ln=='':
123           # the file has been read
124           toggle_rd_ln = False
125
126 f1_out.close()
127 f2_out.close()
128
129 print 'failed to find sequences: ', iter1
130 print 'script complete'
```

8.5.2 sort_cluster.py

sort_cluster.py

```
1 # import libraries
2
3 import numpy as np  # numerical python
4 import sys # needed for passing arguments
```

```
5  import re # needed for regular expressions
6  import os # for creating files, making directories etc
7  import time # used for timing
8
9  # Functions
10
11 def gen_dict_species_sequence(blast_hit_fn, dict_temp = {},
       debug_toggle = False):
12   f=open(blast_hit_fn,'r')
13   toggle_rd_ln = True
14   seq_spec_dict = {}
15   query_cntr = 0
16   nohits_cntr=0
17   query_str = ''
18   query_str_meta = ''
19   while toggle_rd_ln:
20     tmp_ln = f.readline()
21     #print tmp_ln
22     if tmp_ln.find('Query=')!=-1: # we have found a seq Query line
23       # complex data structure generate a string with all Query
             data
24       try:
25         # first save all existing data
26         # search the Query line to extract the info
27         # regular expression for 16S
28         query_seq = re.findall(r'\D{1}\d{1}_\d+\D+_\d+.\d+',
             query_str)
29         # extra info in case of duplicate sequences
30     query_meta = re.findall(r'\d{1}[_.:]\D{1}[_.:]\d{1}[_.:]\D+',
         query_str)
31         try:
32           species = query_str_meta.split('|')[8].split('\n')[0].
               strip(',')
33         except:
34           species = re.findall(r"\*\*\*\*\*\s\D{2}\s\D{4}\s\D{5}\s
               \*\*\*\*\*",query_str_meta)[0]
35           query_meta = ['']
36           nohits_cntr+=1
37         seq_spec_dict[query_seq[0][:19].lower()] = {'species':
             species,'meta':query_meta[0]}# [:19]
38         if debug_toggle == True:
39          try:
40           print 'new dictionary entry: ',query_seq[0][:19],
               species
41          except:
42           pass
43          if raw_input('test') == 'q':
44            sys.exit()
45         query_cntr+=1
46       except:
47         if query_cntr!=0:
48           print "Something failed with the Query: ", query_str
49
50       # once all pre existing data has been saved
```

```
51        # empty the query_str_meta for the next sequence
52        query_str_meta = ''
53        query_str = tmp_ln
54      else:
55        # if "Query=" is not in the line build the query meta str
56        try:
57          query_str_meta= query_str_meta+tmp_ln
58        except:
59          query_str_meta =''
60          print 'waiting for a Query_str_meta init'
61      if tmp_ln=='':
62        toggle_rd_ln = False
63
64    print 'total queries: ',query_cntr
65    print ' of which were no hits: ', nohits_cntr
66    return seq_spec_dict
67
68  #list_seqs=dict((x, list.count(x)) for x in list)
69  #for key in list_seqs.keys():
70  #    a+=list_seqs[key]
71  #f1.close()
72
73  ################################################
74  # USAGE:  python example.py <in_fn1> <in_fn2> <in_fn3> <out_fn2> <
      out_3>
75  #fn1 - corresponds to file with multiple sequences >97%
76  #fn2 - corresponds to * sequence from each cluster
77  #fn3 - blast hit filename
78
79  # number of species identified in a cluster i.e 20
80  spec_per_cluster = 20
81
82  dict_species_sequence = gen_dict_species_sequence(blast_hit_fn=sys
      .argv[3],dict_temp={})
83
84  f=open(sys.argv[1],'r')
85  f1=open(sys.argv[2],'r')
86
87  print "Searched filename: ", sys.argv[2]
88  toggle_rd_ln = True
89  seq_ref_dict = {}
90  iter=0
91
92  while toggle_rd_ln:
93      tmp_ln = f1.readline()
94      if tmp_ln.find('>')!=-1: # we have found a seq ID line
95        seq=tmp_ln.split('>')[1].split(' ')[0].lower()
96        seq_ref_dict[seq[:19]] = tmp_ln.split('>')[1].split(' ')[1].
            split('\n')[0]
97        iter+=1
98      else:
99        seq_ref_dict[seq[:19]]+= tmp_ln.split('\n')[0]
100     if tmp_ln=='':
101       toggle_rd_ln = False
```

```
102
103 print 'Found: ',iter, 'sequences'
104
105 f1_out=open(sys.argv[4],'w')
106 f2_out=open(sys.argv[5],'w')
107 f2_out.write('#')
108 #f2_out.write('Cluster\t')
109 samples_nm = ['A1','A2','B1','B2']
110 for nm in samples_nm:
111    f2_out.write(nm+'\t')
112 f2_out.write('Species\t')
113 f2_out.write('Rep. Seq.\t\n')
114
115 toggle_rd_ln = True
116 toggle_rd_clstr = True
117 iter=0
118 iter1=0
119
120 print 'Searched filename: ',sys.argv[1]
121
122 while toggle_rd_ln:
123     while toggle_rd_clstr:
124       tmp_ln = f.readline()
125       if tmp_ln.find('Cluster')!=-1: # we have found a cluster ID
              line
126         # start a new cluster
127         # write any data that exists
128         try:
129           print 'Total seqs found:', iter
130           if iter>=spec_per_cluster:
131             print 'saved:','>'+rep_seq2find
132             str1 = '>'+rep_seq2find+' '+seq_ref_dict[rep_seq2find.
                    lower()]+'\n'
133             f1_out.write(str1)
134             for nm in samples_nm:
135               f2_out.write('%i\t'%samples_dict[nm])
136             try:
137               print rep_seq2find, ' : ',
138                     dict_species_sequence_schnell[rep_seq2find.
                          lower()]['species']
139               f2_out.write('%s\t'%dict_species_sequence_schnell[
                       rep_seq2find.lower()]['species'].strip())
140             except:
141               f2_out.write('No species in blast file\t')
142             f2_out.write('%s\n'%rep_seq2find.lower())
143
144           print '------------------------'
145         except:
146           # no cluster exists, i.e 1st iteration
147           print "Initialise cluster "
148
149         # find cluster ID
150         clstr_no = int(re.findall(r'\d+',tmp_ln)[0])
151         print 'Cluster number:  ',clstr_no
```

```
152           # clear dictionaries and counters
153           iter=0
154           toggle_rd_clster = True
155           samples_dict={}
156           for key in samples_nm:
157             samples_dict[key]=0
158
159         else:
160           iter+=1
161           # search for representative sequence
162           rep_seq = re.findall(r'.\D{1}\d{1}_\d+\D+_\d+.\d+\.\.\.\s
                \*',tmp_ln) # 16S
163           if len(rep_seq)!=0:
164             # extract representative sequence to reference
165             rep_seq2find = re.findall(r'\D{1}\d{1}_\d+\D+_\d+.\d+',
                  rep_seq[0])[0]
166           # for access to all the sequences
167           try:
168             seq_curr_iter = re.findall(r'\D{1}\d{1}_\d+\D+_\d+.\d+',
                  tmp_ln)[0]
169             #print seq_curr_iter
170           except:
171             print 'failed to find sequence for this cluster'
172             iter1+=1
173
174           # search for samples present in cluster
175           try:
176             seq_sample= re.findall(r'.\D{1}\d{1}_',tmp_ln)
177           except:
178             print 'reg. exp. error: the sample name is not
                  compatible'
179
180           # update the dictionary
181           try:
182             samples_dict[seq_sample[0].split('>')[1].split('_')
                  [0]]+=1
183           except:
184             toggle_rd_clstr = False
185
186         # catch an empty readline command and exit
187         if tmp_ln=='':
188             # the file has been read
189             toggle_rd_ln = False
190
191 f1_out.close()
192 f2_out.close()
193
194 print 'failed to find sequences: ', iter1
195 print 'script complete'
```

8.5.3 adjust_table.py

adjust_table.py

Appendix B

```
# import libraries

import numpy as np  # numerical python
import sys # needed for passing arguments
import re # needed for regular expressions

########################################
# USAGE: python adjust_table.py <data_file_name_input> <
    file_name_output>

toggle_first_word_species = False
toggle_float = False
# load a data file with tab delimited columns
if toggle_float:
    file = np.loadtxt(sys.argv[1],delimiter = '\t',comments = '#'
        ,\
        dtype = {'names':['A1','A2','B1','B2','Species'],\
        'formats':['f4','f4','f4','f4','a100']})

else:
    file = np.loadtxt(sys.argv[1],delimiter = '\t',comments = '#'
        ,\
        dtype = {'names':['A1','A2','B1','B2','Species'],\
        'formats':['i4','i4','i4','i4','a100']})

# Generate output file
f1_out=open(sys.argv[2],'w')
f1_out.write('#A1\tA2\tB1\tB2\tSpecies\n')  # generate the file
    header
print 'total clusters in file: ',file.shape[0]
file.sort(order='Species')

spec_list=file['Species'].tolist()
spec_list_tmp=[]

if toggle_first_word_species:
    tmp_species=file["Species"].copy()

    for i in np.arange(tmp_species.shape[0]):
        tmp_species[i]=tmp_species[i].strip('"').split()[0]

    file['Species'] = tmp_species
    spec_list=tmp_species.tolist()


i=0
for item in spec_list: # cycle through all entries input file
    if spec_list_tmp.count(item)==0: # if no entry exists for the
        species create one
        tmp=np.compress(file['Species']==item,file) # filter all
            entries based on species name
        # require statistics for each species
        A1=tmp['A1'].sum()
        A2=tmp['A2'].sum()
```

```
49          B1=tmp['B1'].sum()
50          B2=tmp['B2'].sum()
51          if toggle_float:
52              s='%.4f\t%.4f\t%.4f\t%.4f\t%s\n'%(A1,A2,B1,B2,item) #
                    string for output file
53          else:
54              s='%i\t%i\t%i\t%i\t%s\n'%(A1,A2,B1,B2,item) # string
                    for output file
55          f1_out.write(s)
56          i+=1 # iterator, count number of different species
57
58      spec_list_tmp.append(item) # append this species to the
            tmp_list to avoid repeat
59
60  print 'total different Species identified: ',i
```

8.5.4 concatenate.py

concatenate.py

```
1  # import libraries
2
3  import numpy as np  # numerical python
4  import sys # needed for passing arguments
5  import re # needed for regular expressions
6
7  ##################################################
8  # USAGE: python concatenate.py <gramn_table_sum.txt> <
       gramn2_table_sum.txt> <test_debug.txt> <test.txt>
9  # argv1 = table
10 # argv2 = table
11 # argv3 = generated file A1 -> B4 SPECIES species_f1 species_f2
12 # argv4 = generated file A1 -> B4 SPECIES
13 # The reason for two outputs is for partial matches so you can
       check them
14
15 # load a data file with tab delimited columns
16 file1 = np.loadtxt(sys.argv[1],delimiter = '\t',comments = '#',\
17         dtype = {'names':['A1','A2','B1','B2','Species'],\
18         'formats':['i4','i4','i4','i4','a100']})
19
20 # load a data file with tab delimited columns
21 file2 = np.loadtxt(sys.argv[2],delimiter = '\t',comments = '#',\
22         dtype = {'names':['A3','A4','B3','B4','Species'],\
23         'formats':['i4','i4','i4','i4','a100']})
24
25 # Generate output file
26 out_file=open(sys.argv[3],'w')
27 out_file.write('#A1\tA2\tB1\tB2\tA3\tA4\tB3\tB4\tSpeciesMatch\t%s\
       t%s\n'%(sys.argv[1],sys.argv[2]))  # generate the file header
28 out_file1=open(sys.argv[4],'w')
29 out_file1.write('# Two files concatenated: %s , %s \n'%(sys.argv
       [1],sys.argv[2]))   # log the filenames of those combined
```

```
30 out_file1.write('#A1\tA2\tB1\tB2\tA3\tA4\tB3\tB4\tSpecies\n')  # generate the file header
31
32 file1.sort(order='Species')
33 file2.sort(order='Species')
34
35 # make a copy of file1 & file2 this identifies unmatched species
36 file1copy= file1.copy() # not needed but useful for debugging
37 file2copy= file2.copy()
38
39 # iterate through all species in file1 and test file2 for matches
40 for iter1 in file1['Species']:
41    match = False # switch to identify if a match has been located for this species
42    for iter2 in file2['Species']:
43       # search for the string in file2 species
44       try:
45         # intelligent string search length limit
46         # search for a matching species name, match all characters in file1
47         # file2 species can have more characters but will still produce a match
48         temp = re.findall(iter2.strip()[:len(iter1)],iter1)
49       except:
50         try:
51           temp = re.findall(iter2,iter1)
52         except:
53           temp = ''
54       # check for a match if so save it
55       if len(temp) > 0: # if a match has been found i.e has characters
56          match = True # switch to identify a match has been found
57          print iter1,iter2
58          # locate statistics related to species
59          tmp_stats_f1=file1.compress(file1["Species"]==iter1)
60          tmp_stats_f2=file2.compress(file2["Species"]==iter2)
61          # save to the output files
62          s = '%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%s\t%s\t%s\n'%(
                tmp_stats_f1['A1'],tmp_stats_f1['A2'],tmp_stats_f1['
                B1'],tmp_stats_f1['B2'],tmp_stats_f2['A3'],
                tmp_stats_f2['A4'],tmp_stats_f2['B3'],tmp_stats_f2['
                B4'],temp[0],iter1,iter2)
63          s1 = '%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%s\n'%(tmp_stats_f1
                ['A1'],tmp_stats_f1['A2'],tmp_stats_f1['B1'],
                tmp_stats_f1['B2'],tmp_stats_f2['A3'],tmp_stats_f2['
                A4'],tmp_stats_f2['B3'],tmp_stats_f2['B4'],temp[0])
64          out_file.write(s)
65          out_file1.write(s1)
66          # keep a log of unmatched iter2 species
67          file2copy = file2copy.compress(file2copy["Species"]!=
                iter2)
68    if match:
69       print 'match found' # some feedback
70    else:
```

```
71        # no match found so save file1 species with no match in the
              same format
72        tmp_stats_f1=file1.compress(file1["Species"]==iter1)
73        s = '%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%s\t%s\t%s\n'%(
              tmp_stats_f1['A1'],tmp_stats_f1['A2'],tmp_stats_f1['B1'
              ],tmp_stats_f1['B2'],0,0,0,0,iter1,iter1,'')
74        s1 = '%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%s\n'%(tmp_stats_f1['
              A1'],tmp_stats_f1['A2'],tmp_stats_f1['B1'],tmp_stats_f1[
              'B2'],0,0,0,0,iter1)
75        out_file.write(s)
76        out_file1.write(s1)
77        # keep a log of unmatched file1 species
78        #file1copy = file1copy.compress(file1copy["Species"]!=iter1)
              # used for debugging
79
80  # include file2 unmatched species in the output files
81  for iter in file2copy:
82     s = '%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%s\t%s\t%s\n'%(0,0,0,0,
           iter['A3'],iter['A4'],iter['B3'],iter['B4'],iter['Species'
           ],'',iter['Species'])
83     s1 = '%i\t%i\t%i\t%i\t%i\t%i\t%i\t%i\t%s\n'%(0,0,0,0,iter['A3'
           ],iter['A4'],iter['B3'],iter['B4'],iter['Species'])
84     out_file.write(s)
85     out_file1.write(s1)
86
87  # close all open files
88  out_file.close()
89  out_file1.close()
```

8.5.5 rand.py

rand.py

```
1  #import libraries
2  import numpy as np  # numerical python
3  import sys  # needed for passing arguments
4  import re   # needed for regular expressions
5
6  ################################################
7  # USAGE:  python example.py <in_fn1> <number_of_seq2select>
8  # fn1 - corresponds to * sequence from each cluster
9  # number_of_seq2select - 10/100/10000
10
11
12 f=open(sys.argv[1],'r') # f is a readable object defined by the
       first argument in this case
13 data=f.read() # data is a new object containing the content of f
14 #(in this case the first input file (fn1)) as one string
15
16 data_sp=data.split('>')[1:]
17 np.random.shuffle(data_sp)
18
19 f_out=open('rand_'+sys.argv[1].split('.')[0]+'.txt','w')
```

```
20 #try:
21 for item in data_sp[:int(sys.argv[2])]:
22     f_out.write('>'+item)
23 #except:
24 #    print 'total datapoints: ',len(data_sp)
25 #    print 'requested random sample: ',int(sys.argv[2])
26 f.close()
27 f_out.close()
```

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any others Swiss or foreign examination board.

The thesis work was conducted from December 2009 to October 2014 under the supervision of Professor Franco Taroni (School of Criminal Justice) and Professor Gilbert Greub (Institute of Microbiology, CHUV) at the School of Criminal Justice at the University of Lausanne.

Lausanne, Switzerland, 27th October, 2014

Sarah Leake