UNIVERSITY COLLEGE LONDON

The microbial ecology of human-associated bacterial communities

Liam Shaw

January 2018

A thesis submitted to University College London for the degree of
DOCTOR OF PHILOSOPHY

Declaration

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Liam Shaw
London, April 2018

Abstract

The bacterial communities within the human body have important associations with health and disease. Understanding their complexity requires ecological approaches. In this thesis, I apply ecological techniques and models to explore the microbial ecology of human-associated bacterial communities at multiple scales. In the first half of this thesis, I explore the oral microbiome using 16S rRNA gene sequencing data to characterise the effect of various factors on its diversity. Multiple factors apart from disease can also affect the oral microbiome, but their relative importance remains a matter of debate. In Chapter 2, I use a dataset of saliva samples from a family of related Ashkenazi Jewish individuals to show that host genetics plays much less of a role than shared household in explaining bacterial community composition. In Chapter 3, I use a large dataset of plaque samples from women in Malawi to investigate associations between bacterial taxa and periodontal disease. I show that the signals from gingivitis and periodontitis can be distinguished, and use correlation networks to identify important taxa for the development of disease. The second half of this thesis deals with the effect of antibiotics on the human microbiome. I demonstrate new approaches at two extremes of scale: abstracting the gut microbiome to a single metric, and also investigating the worldwide distribution and diversity of a single resistance gene. In Chapter 4, I develop a new and simple mathematical model of the gut microbiome's response to antibiotic perturbation and fit it to empirical data, showing that in some individuals the gut microbiome appears to return to an alternative stable state, raising questions about the long-term impact of antibiotics on previously healthy bacterial communities. Antibiotic use also selects for resistance, which is a growing concern, particularly as resistance can be transmitted horizontally on mobile genetic elements. In Chapter 5, I describe a global dataset of isolates containing the mobilized colistin resistance gene *mcr-1* and use the diversity present within a composite transposon alignment to explore its distribution and spread across multiple bacterial communities.

Impact statement

The last decade has seen great interest in understanding the role of the bacterial communities within the human body for health and disease, but their complex ecology is still far from satisfactorily understood. In this thesis, I present a number of different approaches to the sequencing datasets that are now increasingly available about these communities. These approaches represent distinct contributions, but have in common an ecological perspective that attempts to infer ecological units and interactions from the data at different scales. My findings have several implications. My work on the oral microbiome provides important examples of how to integrate detailed host genetic and demographic data into a microbiome analysis to avoid confounding. The effects of antibiotics on our bacterial communities and therefore our health are a global concern, both in terms of the development and spread of resistance, but also possible long-term detrimental effects on healthy individuals. The new mathematical model I develop for the perturbation response of the gut microbiome could form a common basis for comparing different antibiotic treatments and assessing their effects, something that is urgently needed to rationally decide lengths of antibiotic courses. For the spread of resistance, I show how a phylogenetic analysis of a mobile genetic element carrying an antibiotic resistance gene is possible by using a combined approach to search all publicly available sequence data. Future work may routinely use this approach, including integrating the analysis for multiple resistance genes to understand how resistance spreads globally.

Acknowledgements

My first and principal thanks go to my supervisors. From the start, Nigel Klein successfully passed on to me his excitement and enthusiasm for bacterial communities. He has generously allowed my exploration of this developing field in multiple directions while remaining supportive. Robin Callard's attitude to mathematical modelling was a strong influence. His support and advice early on in my PhD was invaluable, and I am very glad to have had him as a supervisor. Sarah Walker never failed to make time to provide robust advice on statistics and to frame problems in a characteristically incisive way. I always looked forward to leaving our meetings with a new set of questions to address. Finally, Francois Balloux has been perhaps my closest scientific mentor throughout the three years, and I am extremely thankful for his encouragement and scepticism in roughly equal measures. His diverse scientific interests and democratic attitude to collaboration have provided me with an ideal environment in which to develop as a researcher.

The work I have completed required many collaborators. In rough chronological order and at the risk of omission: Ronan Doyle introduced me to the world of 16S and much more, and generously allowed me to analyse the data he had spent so much effort collecting together with Ulla Harjunmaa and others (Chapter 3). Andre Ribeiro, Adam Roberts, and Andrew Smith were a friendly and relaxed team to work with, and Adam Levine was extremely patient with my near-endless enquiries about the dataset (Chapter 2). Developing the antibiotic perturbation model was much easier with Chris Barnes on hand to offer a fresh perspective (Chapter 4). Working closely with Lucy van Dorp to understand the story of *mcr-1* was one of my most rewarding scientific experiences to date (Chapter 5). Our common CoMPLEX background made it a particular pleasure to investigate the complex world of bacterial genetics together. I would also like to thank the anonymous peer reviewers whose comments considerably improved all of the work in this thesis.

Thanks to all those who indirectly made this thesis a very enjoyable experience. Camilo Chacón-Duque, Florent Lassalle, Stephen Price, Matteo Fumagalli, Javier Mendoza, and the wider GEE department were a wonderful group of people to share office space and science with. Vania de Toledo was an indefatigable source of administrative support. My friends and family were uniformly great, but particular thanks to my parents and to all inhabitants of Senrab Street and Ebbisham Drive (permanent or otherwise). I would like to highlight for a special mention those who I've talked to about my thesis: my scientific peers Joel Hellewell, John Lees, Rox Middleton, and Tim Russell, and my non-scientific peer Claire Hall.

Finally, I am grateful to the Engineering and Physical Sciences Research Council for supporting my research. It is a privilege to have been provided with public money to work in such a fascinating field.

Contents

List of Figures

List of Tables

Acronyms

AIC Akaike information criterion.

AMR antimicrobial resistance.

BoP bleeding-on-probing.

EUCAST European Committee on Antimicrobial Susceptibility Testing.

FMT faecal microbiota transplant.

GWAS genome-wide association study.

HGT horizontal gene transfer.

HIV human immunodeficiency virus.

HMP Human Microbiome Project.

HOMD Human Oral Microbiome Database.

HOT human oral taxon.

HSCT haematopoetic stem cell transplant.

IFA iron folate.

iLiNS International Lipid-Based Nutrient Supplements Project.

IR inverted repeat.

IS insertion sequence.

LNS lipid-based nutritional supplement.

MDS metric multidimensional scaling.

MED minimum entropy decomposition.

MIC minimum inhibitory concentration.

MMN mixed micro-nutrients.

mRNA messenger RNA.

NMDS non-metric multidimensional scaling.

ORF open reading frame.

OTU operational taxonomic unit.

PCoA principal coordinates analysis.

PCR polymerase chain reaction.

RNA ribonucleic acid.

rRNA ribosomal RNA.

SNP single nucleotide polymorphism.

SRA Short Read Archive.

Bacterial species

The first mention of a species in a chapter is always written in full. Only those species mentioned more than once in this thesis have an abbreviated form listed here. Following the *Journal of Bacteriology* style guide I italicize all taxonomic levels (Journal of Bacteriology, 2018).

A. pleuropneumoniae *Actinobacillus pleuropneumoniae.*
C. difficile *Clostridium difficile.*
E. coli *Escherichia coli.*
E. nodatum *Eubacterium nodatum.*
F. alocis *Filifactor alocis.*
F. nucleatum *Fusobacterium nucleatum.*
K. pneumoniae *Klebsiella pneumoniae.*
P. stomatis *Peptostreptococcus stomatis.*
S. enterica *Salmonella enterica.*
T. denticola *Treponema denticola.*

Chapter 1

Introduction

From the moment we are born, our bodies are homes for a vast and diverse array of bacteria. They live within us in complex communities, interacting with each other and with our immune systems, collectively forming the human microbiome. Recent developments in sequencing technology mean that we can now survey the diversity of the communities within the human microbiome at high resolution. Understanding their diversity, the factors that shape their composition, and their importance for our health requires an ecological approach. In this thesis I apply mathematical models to human-associated bacterial communities to understand aspects of their microbial ecology. The first half of this thesis focuses on the oral microbiome: Chapter 2 tests the relative importance of environment and genetics for the composition of the salivary microbiome, and Chapter 3 looks at using severity scores to separate out different aspects of periodontal disease and their associations with bacteria in plaque. The second half of the thesis focuses on the effect of antibiotics: Chapter 4 outlines a model for the temporal response of the gut microbiome to perturbation by antibiotics, and Chapter 5 analyzes the global distribution of a specific resistance gene. These chapters are each relatively self-contained but are inter-related by the approaches and techniques used. In this introductory chapter I provide a brief background to the human microbiome and the techniques used for analysing it. I review the current understanding of the oral microbiome and the effects of antibiotics on the human microbiome. I also discuss some important ecological concepts applicable to bacterial communities, along with the challenges and opportunities these present.

1.1 The human microbiome

1.1.1 History

The human microbiome comprises all the microorganisms that live on and inside human bodies. The first recorded use of ‘microbiome’ in the *Oxford English Dictionary* is in the 1950s (OED Online, 2018)[1], but the term was re-invented and popularised by Joshua

[1] “The protozoan fauna (as a matter of fact, the whole microbiome) [of a harbour station] is poor in species and individuals, and those present are typically polysaprobes.” *Science Monthly*, January 7th 1952.

Lederberg in the early 2000s to refer to “the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease” (Lederberg and McCray, 2001). While in its original meaning the term includes bacteria, archaea, viruses, and fungi, it is now commonly used to refer to only to the bacterial component of these communities. This component is often studied on its own due to the ease of marker gene sequencing. In this thesis I follow common usage and use ‘microbiome’ to refer only to bacterial communities unless otherwise specified. While there is a growing literature on the human virome (Lecuit and Eloit, 2013), the human mycobiome (Cui et al., 2013), and most recently even the human archaeome (Koskinen et al., 2017), I do not investigate these in any of my analyses.

The human microbiome is not homogeneous. The body contains many different environmental niches, each with their own corresponding bacterial community. Estimates of the number of bacteria present in a human body vary, but the figure is likely to be $\sim 10^{13}$ bacterial cells, of the same order as the number of human cells (Sender et al., 2016). It is not a new notion that the trillions of non-human cells that inhabit our bodies might have important roles in health, but it is only in recent years that we have had the ability to use DNA sequencing to profile these communities at high resolution. Exponential reductions in the cost and efficiency of sequencing technology in recent decades have facilitated investigation of the role of human-associated bacterial communities in health and disease, as well as their complex structures and genetic diversity.

The Human Microbiome Project (HMP) was created in 2008 to perform the important work of characterising the core features of these bacterial communities using an initial cohort of healthy individuals (Turnbaugh et al., 2007). In 2012, multiple papers established a benchmark for the normal composition of oral, skin, gut, and vaginal communities (Gevers et al., 2012; Methé et al., 2012; Huttenhower et al., 2012; Faust et al., 2012; K. Li et al., 2012; Segata et al., 2012). One of the interesting findings of this initial work was that individuals appear to have a relatively stable taxonomic composition over time (Schloissnig et al., 2012), but that taxonomic composition is typically much more variable between individuals than metabolic pathways (Huttenhower et al., 2012). This suggests that many different possible combinations of bacteria can perform the same community functions, but once a particular combination is established in a niche it usually persists. Even since 2012, the datasets available on the human microbiome have increased dramatically in size and the work of cataloguing the normal composition of these communities is largely complete. Microbiome research is therefore now at a stage to address ecological questions about the structure and complexity of the bacterial communities we are home to and their role in our health.

1.1.2 Role in health and disease

At the moment of birth, previously sterile environments within a baby's body are rapidly colonized by environmental bacteria. The mode of delivery can affect this first colonization with variation between body sites immediately after birth, although by six weeks this variation has disappeared and the developing microbiome exhibits body-site specificity (Chu et al., 2017). Over time, these assemblages from environmental colonization form communities and stabilise, typically reaching adult-like levels of diversity within a few years (Koenig et al., 2011; Yassour et al., 2016). These interactions between the first colonizers and the nascent immune system are believed to be extremely important for the proper development of both (C. Petersen and Round, 2014). Immune maturation in mice is dependent on colonization by a host-specific microbiome (Chung et al., 2012). In humans, early-life antibiotic use is associated with increased risk of allergies and autoimmune conditions, suggesting the disruption of immune development (Vangay et al., 2015). Clearly, human-associated bacterial communities cannot be completely neutral assemblages because of the highly specific conditions of particular niches. For internal environments, homeostasis maintains a constant temperature around 37 °C, and oxygen and light levels are low, so e.g. bacteria within the gastrointestinal tract tend to be largely anaerobic (Huttenhower et al., 2012). There can also be more specialized adaptations based on interactions with the host immune system, even among these early colonizers. For example, specific and reproducible phenotypic adaptations consistently occur in the same genomic regions of *Escherichia coli* during colonization of the mouse gut (Barroso-Batista et al., 2014). Even if a given bacterial community appears to perform no useful function for the host, it still provides 'colonization resistance': by occupying available environmental niches, the resident microbiota prevent opportunistic pathogens from gaining a foothold.

The existence of specialized interactions with the immune system has led some to propose the human microbiome as a 'missing organ' with influence over many different aspects of human health (S. V. Lynch and Pedersen, 2016). However, a major issue in microbiome research is the risk of reporting spurious disease assocations due to confounding factors that are known to also be associated with the microbiome such as: diet (David et al., 2013), environment (Lax et al., 2014), lifestyle (J. Wu et al., 2016), and host genetics (Blekhman et al., 2015).[2] These factors can also interact in interesting ways. Bonder et al. (2016) performed a genome-wide association study for microbial abundances and found evidence of a gene-diet-microbiome interaction involving a genetic variant with a recessive effect on lactose intolerance, dietary milk intake, and abundance of *Bifidobacterium*.

There are many challenges in correctly identifying associations between the human microbiome and human health. These include:

[2] See Section 1.2.2 for a discussion of these in the context of the oral microbiome.

- **Inter-individual variation.** Longitudinal studies are not always possible, so cross-sectional datasets in microbiome research are common. Numerous factors affect the variation between bacterial communities. Quantifying their relative strength and importance is difficult due to the confounding of many of these factors.

- **Multi-symptom diseases.** Many diseases are broad groupings of multiple clinical features, and are in fact more appropriately termed as syndromes (e.g. obesity). Identifying associations with the microbiome using a case-control design cannot provide information about which of these features are actually involved.

- **The vagueness of dysbiosis.** A common feature of 'dysbiotic' microbiome states associated with disease is that inter-individual variation appears to be greater than in healthy individuals. Standards exist for identifying individual pathogens (e.g. Koch's postulates), but microbiome research lacks concrete definitions of dysbiosis.

- **Identifying the appropriate scale of association.** The human microbiome and its associated environmental meta-community contain astonishing genetic diversity, meaning that associations can be found at multiple levels in large sequencing datasets with high-resolution available (i.e. from phylum to species to gene). It is not always clear which of these levels is appropriate.

In this thesis, I engage with these four challenges in different contexts, corresponding to each of the four main chapters, and aim to show that it is possible to address them using statistical techniques adapted from ecology.

1.1.3 Sequencing

For many years the human microbiome remained poorly characterized because bacteria that live inside the body tend to be anaerobic and adapted to specialized conditions, making it difficult to find appropriate culture conditions to grow them *in vitro*. Here I describe two important technologies that allow the sequencing of DNA directly from samples without culturing. As this thesis is not experimental, this is only a very brief survey and readers interested in further details should consult the references cited.

16S rRNA marker gene sequencing

A marker gene constitutes a stable region of the genome across organisms that can be used to accurately determine phylogenetic relationships. To be a good marker gene, a gene should have a critical functional role – meaning that it is highly conserved – but also have variable regions – meaning it contains information for taxonomic identification. The 16S rRNA gene[3] is around 1,550 bp long and is universal among prokaryotes

[3] *16S* refers to the sedimentation rate of the translated gene in a centrifuge measured in Svedbergs, a unit equivalent to 10^{-13} seconds. *rRNA* refers to the product of the translation of the gene from DNA

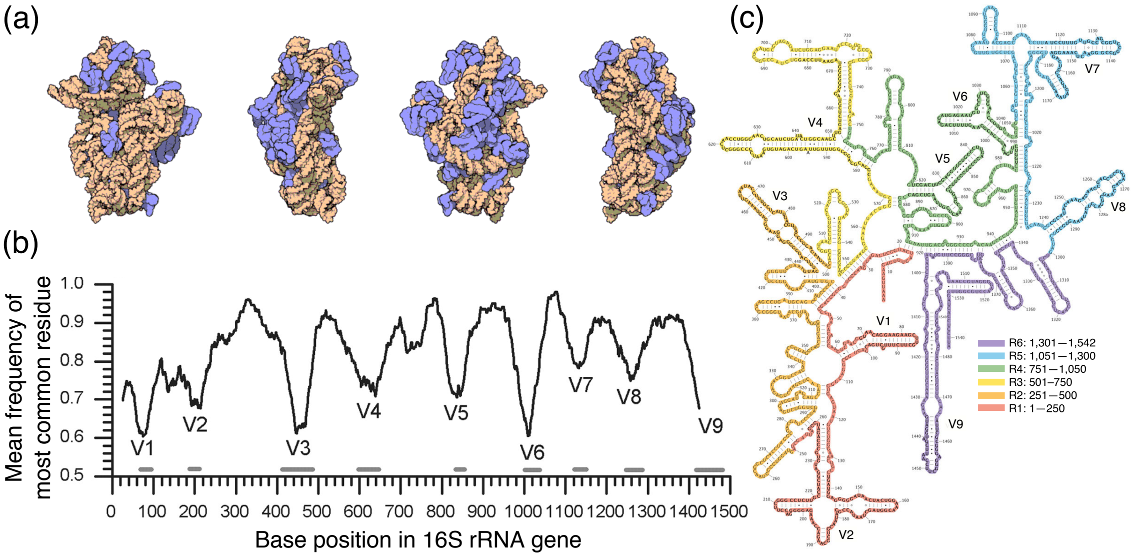


Figure 1.1: Structure and sequence features of 16S ribosomal RNA (rRNA) and the 16S rRNA gene. **(a)** Structure of the 30S subunit of *Thermus thermophilus*, shown from four different angles of rotation. The structure consists of 19 proteins and a small polypeptide (blue) bound to folded 16S rRNA (khaki) (Schluenzen et al., 2000). Adapted from an animation by David S. Goodsell (Goodsell, 2017). **(b)** Mean frequency of the most common residue at each base position within the 16S rRNA gene from 4,383 seqences. Data has been smoothed with a 50 base sliding window. Adapted from Ashelford et al. (2005). **(c)** Secondary structure of 16S rRNA. Different regions are shown in different colours, with variable regions in bold. Adapted from Yarza et al. (2014).

(Clarridge, 2004). It codes for an important part of the 30S subunit of the ribosome (Figure 1.1a), the molecular machine that translates and assembles proteins from messenger RNA (mRNA). Translating proteins is a fundamental cellular process, meaning that certain regions of the 16S rRNA gene are under high selection to preserve the function of the ribosome. These regions are highly conserved across the bacterial domain. However, the subunit is a large folded piece of RNA and there exist regions that loop around and do not directly contribute to function (Figure 1.1c). These regions are not under high selection and so tend to accumulate more mutations, making them more variable than other regions of the gene (Figure 1.1b). The sequence of these hypervariable regions is closely linked to the evolutionary history of bacteria, meaning that sequences can be associated with taxonomic groups to varying levels of resolution. Generally, the more evolutionarily divergent two bacterial species are the more divergent one would expect their hypervariable regions to be. Because of this, rRNA genes have been called “the ultimate molecular chronometers” (Woese, 1987). Woese and Fox (1977) used this insight for phylogenetic inference and analyzed association coefficients between 16S or 18S rRNA from thirteen species of eukaryotes and prokaryotes, showing three major groupings and providing the first evidence for at least three primary branches in the tree of life by resolving the prokaryotes into two domains: the bacteria and the archaea.

The 16S rRNA gene has a total of nine hypervariable regions interspersed with con-

into ribosomal ribonucleic acid (RNA). The *16S rRNA gene* is often also referred to as *16S rDNA*. Its homologue in eukaryotes is the 18S rRNA gene.

served regions (Figure 1.1b). The benefit of this combination is that universal DNA primers can be designed to target a conserved region that is highly similar in all bacteria. After the primer bonds to the conserved region, replication moves along the sequence of DNA and continues into a hypervariable region. Thus, the polymerase chain reaction (PCR) can be used to amplify a particular region of the 16S rRNA gene. When applied to a sample containing many different species of bacteria, the relative abundances of the taxa present can then be identified by partitioning sequences into units and counting the abundances of these units. It is common to use primers that amplify in two directions (forward and reverse) to cover a larger region of the gene using 'paired-end' sequencing (Caporaso et al., 2012).

Identifying ecological units

The first challenge of marker gene sequencing data is to cluster or partition sequences into units. To be practically useful, these sequence units should correspond in a meaningful way to taxonomic units and ultimately ecological units. Sequence-based units are often called phylotypes, and ecological units ecotypes. Achieving as high a correlation as possible between phylotype and ecotype is essential for microbial ecology.

The most widely-used methods for inferring ecological units from marker gene data use sequence similarity. Sequence reads that have a similarity above a certain cutoff are grouped together into an operational taxonomic unit (OTU). This reduces the complexity of datasets, saves computational time, and simplifies downstream analysis. A variety of increasingly sophisticated algorithms have been developed to perform this clustering into OTUs. The main methods can be thought of in three groupings:

- Closed – sequences are clustered using seed sequences from a reference database.
- *De novo* – sequences are clustered without a reference database.
- Open – a mixture; sequences are first clustered using a closed algorithm against a reference database, then any that do not cluster are clustered *de novo*.

In this thesis, I largely use *de novo* clustering methods, as these have been shown to be generally superior to closed or open clustering methods (Westcott and Schloss, 2015). All OTU clustering methods require a similarity cutoff below which sequences are considered to be different units. A convention in microbial ecology is to use an arbitrary similarity cutoff of 97% for largely historical reasons. 97% was given as the approximate 16S rRNA gene sequence similarity level that was equivalent to a 70% DNA-DNA hybridisation (which had previously been recommended as a heuristic for determining bacterial species) by the *Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics* (Moore et al., 1987). In other words, the 97% cutoff is a rough approximation built on a rough approximation, which was intended only as a proxy for species identity. Sequence-based similarity cutoffs will not be identical for all pairs of extant species and are strongly influenced by phenotypic consistency (Moore et al., 1987). Added to this, bacterial taxonomy is usually highly historically contingent, which has led some to suggest entirely divorcing strain classification from previous species names (Baltrus, 2016). Furthermore, species with virtually identical 16S rRNA genes can have very different habitats. For example, two strains of *Candidatus Pelagibacter ubique* with >98% similarity across the whole 16S rRNA gene have distinct geographical distributions: HTCC1062 dominates in polar regions and HTCC7211 in tropical locations (Brown et al., 2012). Remarkably, similar distinct distributions to those shown at a global scale in the world's oceans can exist even within the human mouth (see Section 1.2.1).

OTU clustering is a powerful approach, but because highly similar phylotypes with only a few base changes can still correspond to meaningful ecotypes, there exist newer methods to avoid losing this ecological information. These 'oligotyping' methods avoid clustering by overall similarity and use a substantially different methodology based on positional entropy, aiming to find the most informative bases in an alignment of sequences and iteratively partition sequences into clusters using these bases. Oligotyping (Eren et al., 2013) and its unsupervised counterpart minimum entropy decomposition (MED) (Eren, Morrison, et al., 2014) provide much higher resolution of taxonomic and therefore ecological units, even in the presence of sequencing errors. All marker gene datasets will contain sequences with bases in their alignment that are inaccurate; a sequencing error per base of 0.03 (the probability of a base being inaccurate) intuitively suggests that clustering will be limited to at most a 97% overall similarity cutoff.[4] However, the insight used in oligotyping methods is that as long as these sequencing errors are randomly distributed they can be distinguished from true sequence differences by the different signals of positional entropy: positions with true varying bases will show much higher positional entropy compared to those with no differences but random errors. These positions will appear as peaks of positional entropy. Phylotypes are generated by clustering sequences based on their bases at the position with the highest entropy, then repeating this process to generate sub-clusters from remaining positions until no position has entropy above a pre-specified limit. This can generate higher resolution phylotypes separated by just a single base, often referred to as oligotypes.

There are further subtleties to working with the 16S rRNA gene. Bacteria can carry multiple copies of the gene, with some species having as many as 15 different copies (Vtrovský and Baldrian, 2013; Stoddard et al., 2015). Furthermore, PCR primer bias can also result in preferential amplification of certain taxa (Edgar, 2017) as can the choice of amplification region, with different regions detecting phyla in different proportions and so potentially biasing diversity metrics (Cai et al., 2013). Therefore, relative abundances do not necessarily correspond to true numbers of bacterial cells. Note that differential

[4] In fact, even an intuitive 97% cutoff will lead to incorrect clusters in this scenario. For a given sequence the percentage of bases with errors will be binomially distributed ($n = 100$, $p = 0.03$). If we now sample from a large population of sequences, by the Central Limit Theorem their percentage error will be normally distributed with a mean of 3% and a standard deviation of $\sim 1.7\%$. Thus, a practical sequence similarity cutoff will be lower than 97% to ensure clusters can be trusted.

abundances across samples for a particular taxa can still be calculated.

Ultimately, phylotypes from 16S rRNA gene sequences are only a proxy for true ecotypes, so sometimes they simply do not resolve ecologically distinct populations. This fact has been known for many years: Jaspers and Overmann (2004) analyzed 11 strains of the planktonic bacterium *Brevundimonas alba* and showed that despite completely identical 16S rRNA genes they had non-overlapping niches in terms of their ability to synthesize carbon substrates. Bacterial genomes are often highly flexible and dynamic, and mobile genetic elements can result in very different phenotypes and therefore ecotypes for species with identical 16S rRNA genes. For example, antibiotic resistance genes can be transferred on integrons between species in the oral microbiome (Tansirichaiya et al., 2016). These studies illustrate dramatically that while 16S rRNA is extremely useful for profiling bacterial communities, it cannot directly give information on the other genomic content of the community. To a certain extent phylogeny is correlated with function. Methods exist that attempt to infer functional content from 16S rRNA data alone (Langille et al., 2013). However, these methods have serious shortcomings: they rely on the availability of complete genomes; they cannot capture frequent horizontal gene transfer (HGT); and they can only give information on broad KEGG pathways which are often highly uninformative. In summary, 16S rRNA contains only a small fraction of the total genetic diversity in a bacterial community. It is the most generally ecologically informative and practical fraction to sequence, but it is not always sufficient for answering ecological questions with a focus on functional capacity. Exploring this deeper functional diversity requires other sequencing approaches.

Shotgun sequencing

Unlike marker gene surveying, where PCR is used to target a specific region of genomic DNA, in shotgun sequencing all DNA in a sample is extracted and sequenced. The short reads generated can then be mapped to existing metagenomic databases to identify genes of interest (e.g. 16S rRNA) and also used to assemble genomes. When higher taxonomic resolution and information on functional capacity is needed, metagenomic sequencing is undeniably superior to 16S rRNA sequencing. It also presents a correspondingly more complex challenge for data analysis. Identifying ecological units within shotgun sequencing data is often even more challenging than for marker gene data, particularly for metagenomic datasets. Approaches exist to cluster metagenomic datasets of mixed communities into genomes based on *k*-mer abundances and coverage (Alneberg et al., 2014; Cleary et al., 2015) with recent methods even claiming strain-level resolution (Quince et al., 2017). Developments in long-read sequencing promise to address some of these difficulties, but it still remains important to consider the biology and ecology. The questions to be answered from whole genome sequencing are fundamentally the same as for marker gene sequencing: how to organise data to identify meaningful ecological units, how to quantify variation in these units, and how to associate this variation with other factors.

1.1.4 Summary

The advent of high-throughput sequencing technologies and the use of universal primers for the 16S rRNA gene has permitted the characterization and description of the many different communities within the human microbiome. There are many challenges to understanding how to correctly infer ecological units from this sequencing data and investigate their microbial ecology, but the past few decades have seen many advances in our understanding. In the next section, I go into more detail about a particular group of communities within the human microbiome: the oral microbiome.

1.2 The oral microbiome[5]

Research into the bacteria that live in our mouths has the longest history of any part of the human microbiome. In 1683, Antoni van Leeuwenhoek scraped plaque from his teeth, mixed it with rainwater, and examined it under a microscope. Despite what he thought of as a rigorous daily tooth-cleaning regime – involving rubbing his teeth with salt every morning and vigorously rubbing them with a cloth after eating – he was astonished to describe *dierken* ('little animals' or 'animalcules') "very prettily a-moving", representing the first recorded observations of oral bacteria (Leeuwenhoek, 1683). The oral microbiome is now perhaps the most well-characterized environmental niche in the human body due to two factors: not only is sampling from the mouth easier than from internal environments, but also the culture conditions for bacteria are more easily reproduced than for truly internal body niches. As of 2017 just 32% of taxa are estimated to remain uncultivated (www.homd.org). The oral microbiome typically has very high diversity, with samples from the mouth typically having higher alpha diversities than those from other body sites (Stearns et al., 2011; Huttenhower et al., 2012). In this section I summarise the current state of knowledge about the oral microbiome, the investigation of which is the topic of the first two chapters of this thesis.

1.2.1 Structure and characterization

The biogeography of the oral environment

Referring to 'the' oral microbiome might suggest a degree of homogeneity within the mouth, but it should be stressed that the biogeography of the oral cavity leads to highly structured and differentiated microenvironments with correspondingly different microbial populations. Examples of microenvironments within the oral cavity include the periodontal sulcus, tongue, hard palate, buccal mucosa, and saliva (Krishnan et al., 2016), although there is clearly overlap and some degree of mixing between these sites (Figure 1.2). More generally, a broad distinction can be made between hard tissue surfaces

[5] Parts of this section have been published as: L. P. Shaw, A. M. Smith, and A. P. Roberts (2017). The oral microbiome. *Emerging Topics in Life Sciences* **1**(4), 287296. doi: 10.1042/ETLS20170040.

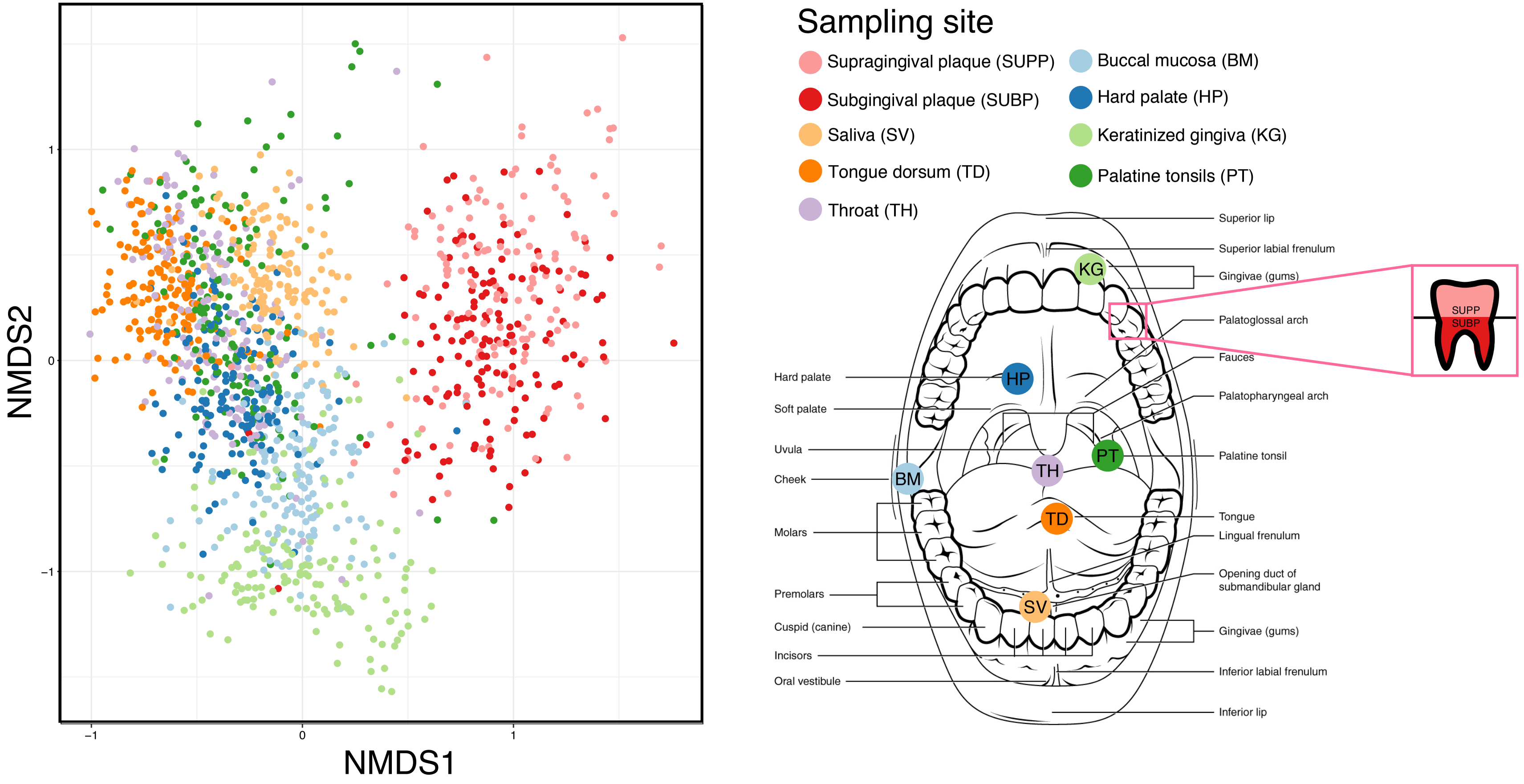


Figure 1.2: Sampling sites in the mouth. Left panel: non-metric multidimensional scaling (NMDS) plot of samples from various habitats in the oral environment. To generate this plot I used a table of 16S rRNA V3-V5 MED phylotypes from data described in Eren, Borisy, et al. (2014). Right panel: the location of these habitats on a diagram of the mouth. Diagram adapted from Humananatomyly.com (2017).

(dental plaque) and soft tissue surfaces (Warinner et al., 2015), which clearly separate in terms of microbial community composition.

The complex nature of oral biofilms is beginning to be explored with new techniques that promise to reveal a great deal about their development and progression. A recent pioneering study by Mark Welch et al. (2016) combined metagenomic sequencing with fluorescence *in situ* hybridization to reveal complex radial structures in supragingival plaque, with anaerobic taxa at the centre and aerobes at the edges. Co-localization of consumers and producers of metabolites within such structures supports the real functional importance of such spatial organization within the oral microbiome. Such biofilm structure can also be investigated with *in vitro* models that allow the culturing of previously unculturable oral microbes (Vartoukian et al., 2016). The associations between taxa that facilitate the buildup of biofilms are crucial in the etiology of oral disease and are still not well understood. In Chapter 3 I demonstrate that even without explicit information on structure, correlations in relative abundance between disease-associated taxa reveal patterns in a correlation network that can be analyzed to identify central taxa that are consistent with previous experimental work.

Development and resilience

The oral microbiome already exhibits body-site specificity six weeks after birth (Chu et al., 2017) and undergoes a substantial increase in diversity up until 3 years, especially after the eruption of teeth (Cephas et al., 2011), followed by a maturation process that continues into adulthood (Sampaio-Maia and Monteiro-Silva, 2014). Even once established, the oral microbiome is subject to continual perturbation. Unlike more internal environments within the body, the mouth experiences daily physicochemical fluctuations in temperature, oxygen content, acidity, and carbohydrate availability. Despite this, the oral microbiome exhibits marked stability over time (Utter et al., 2016). It has been suggested that this need to be robust to multivariate fluctuations may explain the salivary microbiome's greater resilience to antibiotic perturbation when compared to the more homogenous gut microbiome (Zaura et al., 2015).

The resilience of the oral microbiome once established contributes to colonization resistance, where established microorganisms confer protection from external pathogens by occupying available surfaces and environmental niches (Zaura et al., 2014). Many authors have observed that the normal 'commensal' microorganisms that confer protection from external pathogens are also responsible for a wide range of oral diseases (Ruby and Barbeau, 2002; Wade, 2013). This apparent paradox can be resolved by relaxing the strict distinction between the symbiotic and the pathogenic, which can be artificial and misleading in the context of human-associated microbiomes. Indeed, the etiology of oral microbial diseases such as caries and periodontitis has undergone several paradigm shifts over the twentieth century, as molecular techniques have expanded in scope from individual pathogens to the entire oral microbiome (Teles et al., 2013). There is a growing consensus that community-level dysbioses involving feedback loops between the oral environment and oral bacteria are important in oral disease.

Difficulties in characterizing the oral microbiome

Marker gene sequencing allows apparently easy characterisation of the oral microbiome, but this ease can be misleading. Importantly, it is well established that many oral microbes with highly similar 16S rRNA gene sequences can have different genomic content and correspondingly different ecological niches. For example, Eren, Morrison, et al. (2014) reanalyzed HMP data sampled from multiple oral locations within the same individual using oligotyping, and found that *Neisseria* oligotypes varied greatly in spatial distribution. An oligotype of *Neisseria flavescens/subflava* that was detected in high abundance in keratinized gingiva but rare at all other sites sampled had over 99% sequence similarity in the V3-V5 region of the 16S rRNA gene. Furthermore, different choices of primers can result in differential PCR amplification from different bacterial families because of primer mismatch (Morales and Holben, 2009) that can lead to biased diversity metrics (Cai et al., 2013; Kumar et al., 2011), and differences between variable regions can lead to reduced specificity depending on the bacterial genus (William Wade, personal communication).

While partitioning oral microbes into ecological units based on marker genes is a powerful technique, it is important to bear in mind that while positional entropy methods may offer higher resolution and specificity than OTU clustering, they still may not separate out true ecological differences (Section 1.1.3). Indeed, oral microbes with identical 16S rRNA can still possess dramatically different gene complements due to mobile DNA e.g. highly dynamic integron gene cassette arrays (Y.-W. Wu et al., 2012; Tansirichaiya et al., 2016).

1.2.2 Factors shaping the oral microbiome

The core oral microbiome

As the first identified human-associated microbiome, it is unsurprising that the oral microbiome has been extensively characterized compared to other microbiomes. The Human Oral Microbiome Database (HOMD) (www.homd.org, T. Chen et al. (2010)) provides a curated collection of full-length 16S rRNA gene sequences of common oral microbes, together with genome sequences where available. The characterized oral microbiome is dominated by six major phyla making up 96% of the taxa: *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, *Actinobacteria*, *Spirochaetes*, and *Fusobacteria* (Dewhirst et al., 2010). These major phyla define the core oral microbiome determined by the common nature of the oral cavity across individuals: microbes subsisting on endogenous nutrients from the human host with secondary differences in composition due to other factors (Wade, 2013).

Despite an overall similar core oral microbiome, individuals appear to have a stable oral microbiome ‘fingerprint’ over timescales of a few months (Utter et al., 2016) to a year (David et al., 2014), despite rapidly fluctuating relative abundances on a timescale of days (Mark Welch et al., 2014). Oral viruses have also been shown to be personalized and persistent over similar timescales (Abeles et al., 2014), consistent with known phage-bacteria interactions in the oral microbiome (K. Wang et al., 2016). The relative importance of all factors that could conceivably lead to individual-level differences is difficult to establish due to the complexity of performing a comprehensive controlled analysis, although studies of various combinations allow some conclusions to be drawn. These differences between individuals are typically at the sub-genus level, and do not appear to translate into larger-scale geographic differences across global scales (Nasidze et al., 2009).

Diet

The primary source of nutrients for oral microbes is saliva and gingival crevicular fluid rather than food ingested by the host (Wade, 2013), suggesting that diet may not be a key modulator of the oral microbiome in terms of its healthy composition. However, there have been many postulated associations between diet and oral disease, most notably dental caries (see Section 1.2.3). The higher prevalence of oral disease in industrialized countries may be linked to diet-associated dysbioses in the oral microbiome (Marsh, 2003). Chronic disorders like diabetes and inflammatory bowel disease have been linked to a ‘Western diet’ high in sugar and starch (Cordain et al., 2005), and the oral microbiome may play a role in this interaction. Despite a lack of global structuring in the oral microbiome, the oral microbiomes of specific populations with distinct diets that are perhaps more similar to historical human diets can show differences with Western oral microbiomes. Lassalle et al. (2017) investigated the salivary microbiome of hunter-gatherers and farmers in the Philippines, using a carefully controlled study design to identify shifts in composition that were likely to be due to diet alone.[6] Species regarded as oral pathogens linked to periodontal disease were more abundant in the hunter-gatherers’ mouths but anecdotally this did not seem to lead to poorer oral health, suggesting the possibility that the Western diet may have selected for more virulent strains. Further insight into the possible interaction of diet and the oral microbiome over evolutionary timescales may come from investigations of ancient dental calculus (Warinner et al., 2015). It has been claimed that there are major identifiable shifts in composition that correspond to the Neolithic and Industrial Revolutions (Adler et al., 2013).

Lifestyle

Different behaviours may affect the oral microbiome in different ways, either by changing the oral environment directly or through regular seeding by particular taxa due to repeated environmental exposure. Perhaps the most striking example is smoking, which measurably affects the oral microbiome. A study of 1204 American adults found that

[6] I was a co-author of this paper, but do not discuss it again in this thesis.

current smokers had distinct oral microbiome composition from those who had never smoked, with lower levels of *Proteobacteria* and an increased abundance of *Streptococcus* spp. (J. Wu et al., 2016). Interestingly smokers also experience higher susceptibility, severity, and faster progression of periodontal disease, although the mechanisms underlying this faster disease progression remain unclear (Nociti et al., 2015). Other lifestyle factors such as physical activity may also influence the oral microbiome through links to general health and immune status, although these appear to have less of an effect than smoking (Michaud et al., 2013).

Genetics and the environment

There are several conceivable ways that host genetics could affect the oral microbiome, including salivary composition, immune phenotype, or indirectly through gene-diet interactions as observed in the gut microbiome (Bonder et al., 2016). Typically genetics is confounded with multiple other factors, most notably environment. Understanding the role of the environment in determining the oral microbiome is of particular relevance for conditions that show familial aggregation that could be driven by either genetics or shared environment, such as inflammatory bowel disease (Nunes et al., 2011). While there is a generally observable correlation between human genetics and oral microbiome composition, a number of lines of evidence lead to the conclusion that environmental effects are dominant.

It is well established that cohabiting individuals share overlapping oral microbiomes (Lax et al., 2014; Abeles et al., 2016) including – in some cases – with their cohabiting dogs (Song et al., 2013). Stahringer et al. (2012) performed a longitudinal study of the salivary microbiome of twins over several years and concluded that 'nurture trumps nature', with the effect of shared upbringing larger than that of genetics. They observed that monozygotic and dizygotic twins did not have statistically more similar microbiomes – in agreement with observations on the gut microbiome (Turnbaugh et al., 2009) – and that oral microbiome similarity decreased over time once twins no longer co-habited, pointing to the dominant effect of environment.

It has been suggested that there may be ethnic differences in the oral microbiome, possibly linked to differing susceptibilities to periodontitis (Takeshita et al., 2014; Mason et al., 2013). Conclusions reached simply by comparing ethnic groups without any direct genetic evidence should be viewed with scepticism, because they rely on the assumption that other factors (such as lifestyle) are unimportant or controlled for, when these are often confounded with ethnicity. A more rigorous analysis by Blekhman et al. (2015) explicitly used human genetic information extracted from HMP samples from 93 individuals, and did find that host genetic variation correlated with the composition of the oral microbiome. Notably, the most significant association was between genes involved in the signalling pathway for leptin (Cava and Matarese, 2004) and taxa in keratinized gingiva and subgingival plaque, suggesting a link between immunity and the oral microbiome at these sites.

Twin studies are one way to investigate the effects of genetics and environment simultaneously, but the relationship between the oral microbiome in parents and children is less clear: could associations be due to host genetic similarity as well as shared household environment? If so, which component dominates? If the hypothesis is that host genetic similarity affects the microbiome via immune interactions which involve specific genes, genetic relatedness based on pedigree may not capture the relevant portion of true genetic similarity (Speed and Balding, 2014). In Chapter 2 I address this question using a cohort of Ashkenazi Jewish individuals with both host exome and salivary microbiome sequencing to investigate the simultaneous impact of genetics and the environment in a population that controls for other common confounders of microbiome studies.

1.2.3 Associations with disease

Oral biofilms build up progressively in our mouths, and their daily removal is necessary to prevent their establishment and progression. The oral microbiome is indicative of the general relationship between the host immune system and the human microbiome. Individuals with compromised immune systems – either through genetic mutations, chronic infection, immunomodulatory treatments, or pregnancy – have a greater risk of bacterial infection, and a high proportion of these infections occur in the oral cavity. While there is often debate about the direction of causation, it is clear that many oral diseases can be associated with specific bacterial populations. Here I discuss some notable examples of bacterial-associated disease linked to diet (dental caries), immune status (human immunodeficiency virus (HIV)), and multiple factors (periodontal disease).

Dental caries

Dental caries refers to tooth decay caused by acids produced by oral bacteria (Laudenbach and Simon, 2014). These acids are byproducts of the breakdown of oral carbohydrates (Takahashi and Nyvad, 2011). The association between dental caries and carbohydrates was first hypothesised by Miller (W. D. Miller, 1890), and is now supported by extensive evidence (Rugg-Gunn, 2013). Reduced sugar diets have been shown to be associated with fewer dental caries (Moynihan and Kelly, 2014), and it is known that cooked starches can act as a stimulus that produces elevated acidity and aciduric species at caries-prone sites (Bradshaw and R. J. M. Lynch, 2013). In response to this body of evidence, the World Health Organization has issued guidelines that free sugars in diet should provide <5% of total energy intake (Moynihan, 2016). Other important prevention strategies include oral hygiene (to prevent the buildup of aciduric biofilms) and dietary fluoride (to encourage the remineralisation of tooth enamel) (Rugg-Gunn, 2013).

HIV infection

The importance of the host immune system in maintaining the balance with the commensal oral microbiome is clearly indicated by the oral manifestations of HIV infection, with oral abnormalities related to HIV occurring in up to 80% of HIV-infected individuals (Reznik, 2005). HIV has been associated with an increased prevalence of oral mucosal infections and general oral dysregulation, including the overgrowth of the yeast *Candida albicans* and the development of candidiasis as in other immunosuppressed populations (Heron and Elahi, 2017). Candidiasis results from the loss in neutrophil recruitment to the oral tissue through a depletion in number of mucosal associated Th17 lymphocytes. Furthermore, impaired oral immunity in HIV-infected individuals may predispose them to periodontal diseases. The precise effects HIV infection has on the oral microbiome are complicated by potential effects of the anti-retroviral treatment. A study comparing HIV-positive individuals to controls found only minor differences in the composition of the salivary microbiome, although certain taxa including *Haemophilus parainfluenzae* were significantly associated with HIV-positive individuals (Kistler et al., 2015).

Periodontitis

Periodontal disease is inflammation of oral tissues in reaction to oral biofilms (Van Dyke, 2008). Gingival inflammation (gingivitis) is associated with bleeding of the gums, but if oral bacteria progress deeper into the gums this can lead to the formation of periodontal pockets (periodontitis). The etiology of periodontitis remains a matter of debate, although it is accepted that bacterial-derived factors can stimulate the inflammatory response in the gingivae (Cochran, 2008). In general, after an earlier focus on specific pathogens that were identifiable by culture techniques, newer paradigms take a more ecological view where microbial communities enter a disrupted alternative stable state. This is due to synergistic feedback between bacteria and their environment, shifting from homeostasis into destructive inflammation (Hajishengallis and Lamont, 2012). However, it is undoubtedly true that species such as *Porphyromonas gingivalis*, *Porphyromonas intermedia* and *Aggregatibacter actinomycetemcomitans*, which reside within plaque, are highly important in activating the host immune response and driving a chronic inflammatory reaction within the gingivae. Tissue inflammation or gingivitis can lead to a cascade of events, resulting in osteoclastogenesis and subsequent local bone loss via the receptor activator of nuclear factor-kappa B (RANK)-RANK ligand (RANKL). Activation of RANKL drives macrophage differentiation into osteoclasts and bone reabsorption, which results in the development of periodontitis (Belibasakis and Bostanci, 2012).

There is a detailed literature on the molecular mechanisms involved in periodontal etiology (Teles et al., 2013). For example, *P. gingivalis* is known to express a range of virulence factors which facilitate survival within the oral cavity and avoidance of the host immune system (Sheets et al., 2008). Rubrerythrin, a nonheme iron protein, protects the bacteria from neutrophil mediated oxidative killing and exacerbates the local and systemic inflammation within the gingivae (Mydel et al., 2006). The gingipains Kgp and RgpA are the major proteases involved in hemin acquisition, binding, and accumulation, and protect *P. gingivalis* from oxidative damage through the formation of an oxidative sink (Sheets et al., 2008).

A historical focus on individual taxa means that the formation, persistence, and development of bacterial biofilms that can eventually result in destruction of tissue is still not fully understood. Particular species seem to act as bridging bacteria that are required to facilitate the development of biofilms. As I have mentioned, the structure of these biofilms has recently been studied using pioneering fluorescence *in situ* hybridization (Mark Welch et al., 2016) but this approach is not practical across large cross-sectional datasets. Case/control study designs do not reflect the true progression of periodontal disease which is continuous rather than binary. While longitudinal studies of gingivitis exist (Huang et al., 2014) permitting the development of periodontal pockets is not possible for ethical reasons. In Chapter 3 I aim to address both these problems using a large cross-sectional cohort of homogenized supragingival plaque samples. I demonstrate that using two clinical features of periodontal disease allows distinguishing between the community features involved with gingivitis and periodontitis, and that bridging bacteria can be identified using correlation network analysis despite an absence of direct information on spatial organization.

1.2.4 Summary

Our understanding of the oral microbiome has improved significantly since Leeuwenhoek's observations in Delft over 350 years ago, with next-generation sequencing methods beginning to provide us with a much fuller picture of its true taxonomic diversity. However, despite great success in establishing its composition and variation across different sites in the mouth and associations with various external factors, we still have much to discover about the interactions within oral biofilms.

The work I present in this thesis contributes to our understanding of the oral microbiome in two ways. In Chapter 2 I investigate the effect of different factors on the composition of the salivary microbiome, contributing to the debate about whether genetics or the environment is more important in shaping the oral microbiome. Then in Chapter 3 I find associations between taxa in supragingival plaque and periodontal disease using two clinical variables rather than a case/control study design. I demonstrate how network analysis can be used to tentatively identify members of the community that seem to be central to the development of biofilms and the progression of disease.

1.3 Antibiotics and the microbiome

Our relationship with our resident bacterial communities is not always harmonious. Bacteria are responsible for many life-threatening infections, including species which under normal conditions are commensal members of the microbiome e.g. *Streptococcus pneumoniae* (O'Brien et al., 2009). Bacterial infection into regions of the body that are not normally colonized – such as the blood – can lead rapidly to an uncontrolled inflammatory response, sepsis, and death (Ramachandran, 2014).

The discovery of antibiotics led to a revolution in modern medicine, meaning that conditions that were previously life-threatening could be effectively cured. An antibiotic is any substance which acts against bacteria, either by preventing growth (bacteriostatic) or directly killing cells (bacteriocidal). Many antibiotics are in fact naturally occurring substances that are made by bacteria to act against other bacteria. After the discovery of the first antibiotics the 20th century saw a great boom in the isolation, identification, purification, and synthesis of these substances to allow mass production of antibiotics for use in healthcare. It is difficult to overstate the benefits that antibiotics have provided for healthcare as a crucial part of modern medicine, both as direct treatment for infection and as prophylaxis in situations where infection is likely to develop. High concentrations of antibiotics constitute a strong selective pressure on the microbiome that is probably unprecedented in human evolutionary history, and there is an increasing awareness that continuing to use antibiotics as we presently do constitutes a serious problem. In this section, I describe two related issues concerning antibiotics: the damaging long-term effects of antibiotics on the microbiome, and the development and spread of antibiotic resistance.

1.3.1 Antibiotics and the individual

At any one time, it is estimated that between 1-3% of people worldwide are currently taking antibiotics (Goossens et al., 2005). This does not include consumption of antibiotics through food or environmental contamination, which may provide a significant source of low concentrations of antibiotics (Martinez, 2009). The effect of even a short course of antibiotics can be extremely dramatic, both for the bacterial communities and for their human host, but due to the effectiveness of antibiotics there has been little investigation of these side-effects on the human microbiome. Here I outline some important points to consider when thinking about how bacterial communities are affected by antibiotics.

Inter-species interactions dictate the *in vivo* effects of antibiotics

There is surprisingly limited evidence into the actual effects of antibiotics *in vivo* on bacterial communities. Experiments with single species *in vitro* cannot account for potential interactions between different groups of bacteria, and these may be significant when considering the overall impact of antibiotics on a microbial community and its subsequent time-evolution. By way of example, consider the trivial case of a community with species that are either **resistant** or *susceptible* to an antibiotic. If **bacteria A** depends on *bacteria B* via some crucial metabolite M, then using an antibiotic will affect not only *bacteria B* but also **bacteria A**, due to the associated depletion of metabolite M. Investigating the actual impact of antibiotics on real bacterial communities is important to understand these interactions and the potential long-term impacts of antibiotic use.

Current prescribing practices rarely explicitly consider possible *in vivo* interactions and are only guidelines based on clinical heuristics (see e.g. Cosgrove et al. (2015)). The killing rates and tolerances of bacteria for specific antibiotics are based on single cultures; the European Committee on Antimicrobial Susceptibility Testing (EUCAST) provides clinical breakpoints and minimum inhibitory concentration (MIC) distributions for many widely-used antibiotics, as well as clinical guidelines for their usage based on such experiments from multiple laboratories (EUCAST, 2017). It seems that rates and tolerances for individual species cultured *in vitro* do not generally correlate with the *in vivo* activity of antibiotics (Parijs and Steenackers, 2017), suggesting there is little possibility for reliably extrapolating measured *in vitro* effects to *in vivo* effects. The collateral damage of antibiotics is therefore potentially very high and is only beginning to be seriously investigated.

Collateral damage to the microbiome

The mechanisms of action of many antibiotics target fundamental parts of cellular machinery which are shared by many different bacteria. In most cases, using an antibiotic to treat an infection caused by a single species inadvertently affects other bacteria as well and causes collateral damage. Such a depletion of normal bacterial communities constitutes a strong perturbation to the stability of the ecosystem. Gastrointestinal disturbances such as diarrhoea have always been recognized as a relatively common side-effect of antibiotics (Hempel et al., 2012). The risk of further unintended consequences for the gut microbiome has been known about for many decades: Bohnhoff and C. P. Miller (1962) reported that streptomycin increased the likelihood of *Salmonella* infection in mice. Recently, the advent of high-throughput sequencing has allowed quantification of this disturbance at the community level.

Several pioneering studies have established the dramatic impact of short courses of common antibiotics on the gut microbiome in healthy individuals (Modi et al. (2014) provides a useful review). Löfmark et al. (2006) looked at the effects of a week-long course of clindamycin in eight healthy volunteers, and found an enrichment of clindamycin-resistant strains of *Bacteroides* and long-term persistence of resistance genes two years afterwards (Löfmark et al., 2006; Jernberg et al., 2007). Dethlefsen et al. (2008) gave a week's course of ciprofloxacin to three individuals, expecting that effects would be minimal because ciprofloxacin has little *in vitro* activity against members of the gut microbiota. After four weeks, they observed that community composition had mostly returned to its pre-treatment state, but even six months afterwards the depletion of certain taxa remained. In a follow-up study, around fifty samples each were collected over ten months from three individuals, with two five-day courses of ciprofloxacin separated by six months (Dethlefsen and Relman, 2011). The effect of each of the courses of ciprofloxacin was different, suggesting the existence of hysteresis in the gut microbiome; elsewhere in ecology, it is well established that compounded perturbation can produce ecological surprises (Paine et al., 1998). The longevity of this altered microbiome state after antibiotics can be extraordinary when contrasted with the length of treatment: Jakobsson et al. (2010) found that three individuals who received combined clarithromycin and metronidazole for seven days had persistent effects up to four years afterwards.

Certain effects of antibiotics appear to be highly reproducible across individuals while other aspects remain highly individualized. Raymond, Ouameur, et al. (2016) reported that a week-long course of cefprozil caused an increase of *Lachnoclostridium bolteae* in sixteen out of eighteen individuals. However, in other ways there were strong associations between the intitial state of the gut microbiome and the response: increased levels of *Enterobacter cloacae* were only observed in individuals who had higher levels of *Bacteroides* before treatment. These and other studies typically use small numbers of individuals and approach the problem anecdotally. It is therefore striking that the effects of antibiotics are dramatic enough to be clearly observable with such study designs when associations between the microbiome and disease are usually quite subtle. A course of antibiotics is perhaps the most abrupt perturbation that a bacterial community can experience. What is lacking is systematic quantification of the effects of antibiotics in a consistent framework to allow associations with long-term impacts on health using large populations.

Long-term impacts on health

Depletion of the diversity of bacterial communities can have negative impacts on the health of the host. The main reason for this is that the colonization resistance provided by the stable normal microbiome is reduced, allowing pathogenic colonization or overgrowth. *Clostridium difficile* infection (CDI) is often used as a poster child for microbiome research. Individuals who suffer from CDI have typically recently undergone broad-spectrum antibiotic treatment, which reduces the diversity of the gut microbiome, and have markedly different gut microbiome composition to both healthy individuals and those with non-*C. difficile*-associated diarrhoea (Schubert et al., 2014). The opportunistic overgrowth by *C. difficile* is the manifestation of the deeper community-wide dysbiosis; it is the symptom, not the cause. Conventional treatments for *C. difficile* typically have a very low success rate and a high relapse rate. A simple but effective treatment is faecal microbiota transplant (FMT): stool from a healthy donor is transplanted into the sufferer with the aim of suppressing the overgrowth and restoring a healthy gut microbiome. Consistent success rates of above 90% have been reported in multiple trials (Aroniadis and Brandt, 2014). Although the mechanism by which FMT actually works is still not entirely understood, it somehow restores the gut microbiome to a community state that controls the growth of *C. difficile* (Stein et al., 2013). Indication that the general principle of dysbiosis is important – rather than specifically *C. difficile* – comes from the observation that intestinal colonization can also occur from other parts of the human microbiome during gut dysbiosis; it has recently been shown that oral *Klebsiella* species can ectopically colonize the intestine under certain conditions (Atarashi et al., 2017).

Colonization is not the only risk posed by a low diversity gut microbiome. Interactions between the gut microbiome and the immune system can lead to associations between abnormal states of the gut microbiome and the outcomes of other treatments. Individuals undergoing haematopoetic stem cell transplant (HSCT) for underlying immune conditions are typically immunocompromised or immunosuppressed before transplant, meaning they are placed on high levels of antibiotic prophylaxis. Taur et al. (2014) looked at survival after three years in 80 patients undergoing HSCT and found that low gut microbiome diversity at the time of engraftment was significantly associated with lower survival in a multivariate model taking other clinical variables into account. The mechanism for this may be connected to the risk of graft-versus-host-disease, which can be modulated by the normal gut microbiota (Simms-Waldrip et al., 2017). Similarly, in cancer treatment immune checkpoint inhibitors used to treat tumors fail to work in some individuals due to abnormal gut microbiome composition; the efficacy of these inhibitors can be restored by probiotics (Routy et al., 2017).

There is abundant evidence that perturbation by antibiotics can alter the long-term composition of the gut microbiome and have associated detrimental effects on human health. However, we lack a mathematical framework to systematically compare these perturbations. The development of a common framework for comparing the effect of antibiotics on the state of the gut microbiome would be an important step for population-level studies into their long-term impact. In Chapter 4 I develop a mathematical model to describe this perturbation derived from simple ideas in classical ecological theory.

1.3.2 Antimicrobial resistance

Any antibiotic exerts a selective pressure on bacteria to evolve or acquire resistance to that antibiotic, contributing to the development and spread of antimicrobial resistance (AMR), which has been described as one of the greatest threats to global health (O'Neill, 2016). AMR is an umbrella term for any form of increased tolerance or resistance of microorganisms to drugs used to treat infections involving them, but there is considerable diversity. Some forms of AMR occur due to chromosomal mutations in specific genes e.g. in in *Myobacterium tuberculosis*, rifampicin resistance repeatedly arises through predictable mutations in the *rpoB* gene (Ford et al., 2013). Other forms involve the evolution of novel genes that confer resistance to specific antibiotics e.g. the mobilized colistin resistance gene *mcr-1* which was recently described by Liu et al. (2016). In bacteria, such genes can be transferred on mobile genetic elements (Frost et al., 2005).

The fact that bacteria in the body exist in communities is relevant for AMR, because antimicrobial tolerance is known to be higher in multi-species biofilms due to a combination of multiple mechanisms including physical barriers, mutual cross-species protection, and the induction of tolerance phenotypes due to the presence of other species (Hathroubi et al., 2017). The persistence of plasmids carrying resistance elements has been shown to be greater in biofilms of single species compared to well-mixed liquid cultures (Ridenhour et al., 2017). AMR is fascinating because the genetic architecture of bacteria means that it can be approached at multiple levels. There is a surprisingly limited evidence base for antibiotics compared to other commonly-used drugs, and this is reflected in the variety of ways that antibiotics are used worldwide: antibiotic availability and usage vary dramatically between countries (Blommaert et al., 2014) and even within countries, with different recommendations for empirical treatment of common syndromes at different English hospital Trusts (Llewelyn et al., 2014). Such a complexity of differing, individualized responses might appear overwhelming, but this does not preclude the application of general principles about ecosystems.

1.3.3 Summary

I have argued that understanding both the effect of antibiotics on the human microbiome and the problem of antibiotic resistance requires thinking about the effect of antibiotics in ecological terms, rather than focusing on single species. In Chapter 4 I develop a model of perturbation to the gut microbiome by antibiotics and demonstrate that it describes the time-response behaviour and transition to a different long-term community state. In Chapter 5 I look at how the spread of resistance on mobile genetic elements can be tracked using whole genome sequencing, using a global dataset of isolates that carry a gene conferring resistance to the ‘last-resort’ antibiotic colistin. Both these chapters are framed as ecological problems. In the next section, I outline some important approaches for analyzing bacterial communities that I adopt in this thesis.

1.4 Ecological approaches

The sequencing data that can now be collected on bacterial communities require different approaches to those of culture-based microbiology – which by necessity usually focused on culturable single organisms – and can benefit from the application of techniques adopted from traditional macro-ecology to understand the communities as ecosystems. These approaches are collectively referred to as microbial ecology. In some instances, the ease of sequencing means that more is known about the genetics of bacterial communities than about their traditional ecology, leading some to suggest the advent of ‘reverse ecological’ approaches that attempt to infer ecological units from genomic data alone (Shapiro and Polz, 2014; Lassalle et al., 2015). I believe there is merit in this approach, and in this thesis I attempt to synthesize information from multiple sources to address ecological questions about bacterial communities. This section outlines some of the important concepts and modelling approaches I use to achieve this.

Metric	Definition	Reference
Richness	N	
Shannon index	$H = -\sum_{i=1}^{N} p_i \ln p_i$	Shannon (1948)
Simpson index	$\lambda = \sum_{i=1}^{N} p_i^2$	Simpson (1949)
Hill diversity	${}^{q}D = \left(\sum_{i=1}^{R} p_i^q\right)^{1/(1-q)}$	Hill (1973)
Phylogenetic diversity	Sum of branch lengths of minimum spanning path of tree	Faith (1992)

Table 1.1: Commonly used diversity metrics mentioned in this thesis. N is the total number of species observed in the community, and p_i is the proportional or relative abundance of species i. Most diversity indices can be expressed in terms of the Hill diversity for various values of q e.g. richness is ${}^{q=0}D$, but phylogenetic diversity cannot.

1.4.1 Diversity and composition

Diversity is variously defined as the number, evenness, and types of taxa within a community. Diversity metrics are summary statistics that measure different aspects of this definition. There is no absolute measure of diversity and a variety of metrics exist – Table 1.1 gives a list of some commonly-used metrics mentioned in this thesis. Readers interested in the ecological debate about the consistent framework for defining diversity should consult Tuomisto (2010). I take the stance of Ricotta (2005) that the substantial disagreements about what diversity actually is are best resolved by taking diversity to refer to “a set of multivariate summary statistics for quantifying different characteristics of community structure”.

Higher diversity is sometimes assumed to be a positive quality of an ecosystem, implicitly or otherwise. This is wrong: there is no such universal law of diversity. It is true that higher diversity is associated with health in the gut microbiome (Lloyd-Price et al., 2016) but a counterexample is that higher diversity correlates with disease in the oral microbiome (Wade, 2013). More generally, the diversity of human-associated bacterial communities varies substantially depending on the environmental niche (Huttenhower et al., 2012). Shade (2017) describes diversity as “the question, not the answer” – it should provide a starting point for enquiry into ecological mechanisms, rather than a conclusion in itself. I agree with this assessment. In this thesis, I use diversity as a way of summarising a community and as good first step for investigating communities within appropriate contexts. It is also a useful way to incorporate bacterial communities into other models using a single variable. Using diversity to compare samples can be misleading because very different communities can have identical diversities. Composition of samples can also be directly compared with pairwise dissimilarity metrics. In this thesis I use the Bray-Curtis dissimilarity metric (Bray and Curtis, 1957) defined as

$$BC_{ij} = \frac{\sum_k |x_{ik} - x_{jk}|}{\sum_k (x_{ik} + x_{jk})}, \tag{1.1}$$

where x_{ik} is the abundance of species k at site i. The metric is bounded by 0 (completely identical composition) and 1 (completely dissimilar composition). A matrix of pairwise dissimilarities can then be used to produce an ordination of samples using many different techniques. These ordinations are not a statistical test in themselves and the best technique depends heavily on the dataset and research question (McMurdie and Holmes, 2013) but they are often valuable for an exploratory analysis. Using ordinations can show clear general trends that can then be further investigated quantitatively (for example, Figure 1.2). Quantitatively, associations between environmental factors or metadata can be tested using a permutational analysis of variance (Anderson, 2001; Ramette, 2007), as used in Chapter 2.

1.4.2 Associations and interaction networks

The ability to test for associations with disease for many different species simultaneously rather than specifying in advance which species are to be tested allows a ‘hypothesis-free’ data-fishing approach to finding potential associations. The large number of simultaneous tests produces a multiple testing problem, which it is important to correct with multiple testing corrections such as the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). However, reporting associations with taxa as if they were independent tests neglects the fact that in reality taxa are interacting with each other, and this information is encoded in the correlations of their abundances. Homogenized samples of bacterial communities provide no direct information on spatial structure but they do contain information on the co-occurrence of taxa across samples, hinting at possible ecological interactions (whether direct or indirect). There are problems with calculating correlation coefficients from compositional data; the restriction on the total sum of abundances means that relative abundances are on a simplex, so coefficients are not independent. In Chapter 3, I use correlations of log-transformed relative abundances of disease-associated taxa (Friedman and Alm, 2012) combined with a measure originally developed for social network analysis (Freeman, 1977) to identify important members of this network.

1.4.3 Stability and variability

The temporal stability of human-associated bacterial communities varies depending on their location. The human gut microbiome has been suggested to be less resilient than the oral microbiome (Zaura et al., 2015) but appears to be stable over time in the absence of major perturbations (David et al., 2014). How a system responds to perturbation can give insight on the forces that underly its continued stability. In classical ecology such perturbation experiments are common. As I have discussed, the effect of antibiotics on the gut microbiome represents an ideal perturbation experiment, but this perturbation is typically reported fairly anecdotally (Section 1.3.1).

The idea that stable communities in a niche are usually similar and perturbation leads to increased dissimilarity has led to the adaptation of the 'Anna Karenina' principle for microbiome research: "dysbiotic individuals vary more in microbial community composition than healthy individuals" (Zaneveld et al., 2017). There have been calls for the application of ecological theory to the gut microbiome (Costello et al., 2012). The concept of a fitness landscape from evolution has been adapted for ecological communities as a stability landscape (Holling, 1973), and applied heuristically to the human microbiome (Relman, 2012). In Chapter 4 I argue that this concept can be used to derive a simple mathematical model that captures the time-response of the gut microbiome to a short but intense perturbation.

1.4.4 Nested genetic diversity and horizontal gene transfer

Bacterial communities can exhibit complex nested genetic diversity. The amount of data from a 16S rRNA analysis may seem large but it is only a tiny fraction of the true genetic diversity in any given community. Bacteria reproduce asexually by clonal cell division and inherit genetic material vertically. They can also undergo HGT i.e. genetic material can move *across* this vertical line of descent. HGT can prevent the accumulation of deleterious mutations and loss of fitness, rescuing populations from Muller's ratchet (Takeuchi et al., 2014), and provides a valuable source of genetic novelty for phenotypic innovation and niche adaptation (Ravenhall et al., 2015). Genetic adaptation to gut colonization in *E. coli* is conferred by gene inactivation or modulation by an insertion sequence (IS) (Barroso-Batista et al., 2014). Bacteria also have mechanisms in place to control or prevent HGT to prevent the accumulation of selfish genetic elements on their chromosomes. The mechanisms (and counter-mechanisms) of HGT are complex but are traditionally divided into three classes: transformation, transduction, and conjugation (Juhas, 2015). There is accumulating evidence that HGT is not just important for evolutionary novelty in stress situations (Aminov, 2011; Peterson et al., 2011) but is a normal feature of bacterial communities. HGT appears to be pervasive in the human microbiome, with 25-fold increased rates observed between human-associated bacteria compared to between non-human isolates (Smillie et al., 2011). These rates are likely much higher in stress situations: Stecher et al. (2012) demonstrated that gut inflammation can boost rates of conjugative HGT between pathogenic and commensal *Enterobacteriaceae*.

Considering these fuzzy ecosystems presents new challenges for microbial ecology. Unlike macroscopic ecology where species remain distinct, the boundaries for bacteria are not as clear-cut. Indeed, some have questioned whether there are in fact meaningful bacterial 'species', as bacteria do not fit standard definitions which stress reproductive isolation (Mayr, 1942). It seems a mistake to conclude that because microorganisms do not fit macroecological species definitions then they cannot exhibit *any* characteristics of species; more sensibly a spectrum of speciation can be defined based on the level of overlap of genomic and ecological units (Shapiro and Polz, 2014). The clonality of bacteria is determined by the balance of recombination and natural selection, and appears to be largely stable over time for given populations (Shapiro, 2016).

Although inferring the presence of HGT is in general difficult (Ravenhall et al., 2015), there are cases where the unit of selection is clear; one such case is resistance to antibiotics. Resistance genes are often carried on mobile genetic elements such as transposons: the smallest transposable elements are ISs, which contain transposases to catalyze their movement between genomic backgrounds. Tracking the spread of these genes can still be challenging due to the nested genetic mobility of bacteria; transposons can jump between plasmids, and plasmids can jump between species (Sheppard et al., 2016). Antibiotic resistance genes are extremely common in the environment and in the human microbiome. Bacterial communities in the human body are repeatedly exposed to a range of natural antimicrobials in our diet. These include plant-based essential oils and flavonoids (Donsì and Ferrari, 2016; D. Wu et al., 2008), as well as anthropogenically added antimicrobials used for personal hygiene such as chlorhexidine and triclosan (Brading and Marsh, 2003). In addition, exposure of the microbiome to antibiotics (whether clinical or environmental) is likely to select for HGT events which lead to the acquisition of genes that confer resistance or increased tolerance. The human microbiome can harbour multiple antibiotic resistance genes (Sommer et al., 2009; Seville et al., 2009), often in association with mobile genetic elements, and has previously been described as a resistance 'reservoir' (Roberts and Mullany, 2010).

Bacterial genomes are now usually thought of as consisting of a conserved core genome (found in all members of a species) and a variable mobile genome or mobilome (found in only some species). The total diversity of genes is referred to as the pan-genome (Tettelin et al., 2008). There are analogies here to the human microbiome, which also consists of core components found in all individuals and variable components (whether species or genes). The wider community of environmental bacteria represents the metacommunity of the microbiome (Adair and Douglas, 2017), and can provide a source of new species and new genes, including those for antibiotic resistance. Understanding the flexibility of bacterial genomes at this level is thus extremely important for efforts to control disease (Adair and Douglas, 2017).

Genes viewed as part of the core genome for some species can confer antimicrobial resistance when expressed in a different bacterial host. For example, resistance to triclosan can be conferred when *E. coli* expresses a house-keeping metabolic gene (Enoyl-[acyl-carrier-protein] reductase, fabI) derived from the oral microbiome (Tansirichaiya et al., 2017). The fabI gene is present on a transposon in *Staphylococcus* and can be selected for by triclosan (Furi et al., 2016). This is just one example to indicate how resistance

Chapter	Sample source	Sequencing approach
2	Saliva	16S rRNA (V5-V7)
3	Supragingival plaque	
4	Faeces	
5	Multiple (including faeces, blood, urine)	Whole genome sequencing

Table 1.2: Human-associated bacterial communities analyzed in this thesis. For more details on the datasets, see the associated chapter.

elements can be mobile across species and have differential effects depending on their host, making quantifying and tracking antimicrobial resistance extremely challenging. In Chapter 5 I describe work studying bacterial isolates from multiple species that contain a recently-described gene encoding resistance to colistin on a composite transposon (Liu et al., 2016). Using a global dataset allows the dating of the emergence of this composite transposon and provides insight into its origins.

1.5 Conclusions

As I have outlined above, the recent dramatic increase in availability of sequencing data about human-associated bacterial communities facilitates analysis of them in their entirety as ecosystems. In this thesis, I analyze a total of four datasets. The first three contain 16S rRNA marker gene sequences from various human niches. For these communities, I adapt existing modelling techniques and develop new models to extract as much information as possible from the data. This allows me to draw novel conclusions about their role in the etiology of disease, the impact of various factors in shaping their composition, and their temporal stability. The final dataset is a global whole genome sequencing dataset of *mcr-1* positive isolates, and contains samples from multiple human niches (including faeces, blood, urine, and sputum) as well as many agricultural and environmental samples that form part of the wider ecological metacommunity.

Broadly, this thesis is divided into two complementary halves on the oral microbiome (Chapters 2 and 3) and antibiotics (Chapters 4 and 5). Each chapter mainly addresses one of the four challenges for associating bacterial communities with health and disease that I outlined in Section 1.1.2, although components of each appear in all chapters. In Chapter 2, I demonstrate how to assess the role of different factors in determining *inter-individual variation* in salivary microbiome composition. In Chapter 3, I look at a *multi-symptom disease* (periodontitis), and show that a cross-sectional dataset of homogenized samples can be used to distinguish associations with different features of disease and even derive information on possible biofilm interactions. In Chapter 4, I go from a schematic picture showing the *vagueness of dysbiosis* to a simple but effective mathematical model that can be fitted to real data on antibiotic perturbation of the gut microbiome. Finally, in Chapter 5 I investigate a global dataset containing whole genome sequences of *mcr-1*-positive isolates and *identify the appropriate scale of assocation* as a composite transposon. I investigate this unit's distribution in several species across the multiple countries and sample sources that make up the meta-community of human-associated bacterial communities.

Chapter 2

Genetics and the environment in the salivary microbiome

Declaration of contributions

Andre Ribeiro performed the 16S rRNA library preparation and sequencing. Nikolas Pontikos calculated host kinships between participants. Adam Levine validated the inference of household information and other metadata. I performed all subsequent data analysis and wrote the associated paper, with feedback from all co-authors.

Publication

This work has been published in *mBio* as Shaw, Ribeiro, et al. (2017):
L. P. Shaw[†], A. L. R. Ribeiro[†], A. P. Levine, N. Pontikos, F. Balloux, A. W. Segal, A. P. Roberts, and A. M. Smith (2017). The human salivary microbiome is shaped by shared environment rather than genetics: evidence from a large family of closely related individuals. *mBio* **8**(5), e0123717. doi: 10.1128/mBio.01237-17.

† : equal contribution.

2.1 Introduction

As I have outlined in the introduction, the oral microbiome is one of the most diverse of any human-associated bacterial community (Huttenhower et al., 2012; Wade, 2013), and is a causative factor in conditions such as dental caries (F. Yang et al., 2012) and periodontal disease (Teles et al., 2013). It has also been implicated as a reservoir for infection at other body sites (Wade, 2013) and in the pathogenesis of non-oral diseases, such as inflammatory bowel disease (Lucas López et al., 2017). Strictly speaking there is no single 'oral microbiome' as its composition is highly heterogeneous across different sites in the mouth (Eren et al., 2013; Mark Welch et al., 2016), but the term is commonly used to encompass all of these. Site-specific microbiomes can be observed in the periodontal sulcus, dental plaque, tongue, buccal mucosa and saliva (Krishnan et al., 2016). The salivary microbiome exhibits long-term stability and can be considered as an important reservoir that contains microorganisms from all distinct ecological niches of the oral cavity. Characterizing and understanding the factors defining the composition of the salivary microbiome is thus crucial to understanding the oral microbiome (Takeshita et al., 2016; Belstrøm et al., 2016).

Some factors that are thought to influence the human microbiome include environment, diet, disease status and host genetics (Section 1.2.2). The relative importance of these factors for the oral microbiome is still under debate, with the majority of previous studies focusing on the gut microbiome, although it seems reasonable to assume some potential interaction between the salivary microbiome and microbial communities in other parts of the human body including the intestinal tract. The concept of a vertically transmissible microbiome that is directly encoded into the genome of individuals rather than being purely environmental is extremely provocative. At extreme scales, there are undoubtedly signals of long-term co-evolution between different primates and their gut microbiomes (Moeller et al., 2014) and signals of short-term strain evolution within individuals (Garud et al., 2017; S. Zhao et al., 2017). At intermediate scales, there is a question over to what extent differences in the composition of oral microbial communities can be linked to genetic differences between their human hosts.

There is evidence that genetically related individuals tend to share more gut microbes than unrelated individuals, whether or not they are living in the same house at the time of sampling (Turnbaugh et al., 2009; Yatsunenko et al., 2012). However, the level of covariation is similar in monozygotic and dizygotic twins, suggesting that a shared early environment may be a more important factor than genetics (Turnbaugh et al., 2009; Stahringer et al., 2012). The effect of co-habitation with direct and frequent contact is greatest when considering the skin microbiome, with a less-evident effect on the gut and salivary microbiomes (Song et al., 2013).

There is also evidence that genetic variation is linked to microbiome composition across other body sites, including the mouth (Blekhman et al., 2015), with a recent genome-wide association study (GWAS) identifying several human loci associated

($p < 5x10^{-8}$) with microbial taxonomies in the gut microbiome (Bonder et al., 2016). However, no study to date has incorporated both genetic relatedness as a continuous variable and shared environment into the same analysis of the salivary microbiome.

Despite high diversity between individuals, the salivary microbiome appears to have little geographic structure at genus level at the global scale (Nasidze et al., 2009). Nevertheless, at smaller geographical scales it appears that the environment plays a role in the oral microbiome. Song et al. (2013) studied 60 household units and found that the bacterial composition of dorsal tongue samples was more similar between cohabiting family members than for individuals from different households, with partners and mother-child pairs having significantly more similar communities. However, this did not include information on genetic relatedness in addition to family relationships. It appears that household-level differences in the salivary microbiome may also apply to genetically unrelated individuals and non-partners, with a similar pattern observed in analysis of 24 household pairs of genetically unrelated individuals, only half of whom were considered romantic couples at the time of sampling (Abeles et al., 2016).

The establishment of the oral microbiome appears to proceed rapidly in the first few years of life, with a notable increase in diversity up to three years (Song et al., 2013), especially after the eruption of teeth (Cephas et al., 2011). The plaque microbiome also appears stable within adult individuals over at least a period of three months, with a unique ‘fingerprint’ of oligotypes discernible even within a single bacterial genus (Utter et al., 2016). Another study indicates that the salivary microbiome is relatively stable over a year, despite measurable effects of interventions like flossing (David et al., 2014). Taken together, these findings suggest the intriguing hypothesis that once a particular oral microbiome is established earlier in life it can potentially persist over months and perhaps even over years, particularly if external factors such as diet remain fixed. If this were true, shared upbringing effects would continue to be detectable in the salivary microbiome even after individuals are no longer living in the same household (Stahringer et al., 2012).

A recently described large Ashkenazi Jewish family (Levine et al., 2016) offers an opportunity to investigate the effect of both environment and genetics in closely-related individuals. The availability of host genetic data for this cohort means it is possible calculate similarity between individuals based on single nucleotide polymorphisms (SNPs), rather than using measures of relatedness from pedigrees that do not precisely correspond to shared genetic content (Speed and Balding, 2014). I hypothesized that using this more accurate measure of host genetic similarity could lead to different conclusions about the proportion of shared microbiome composition attributable to genetics compared with previous studies. Furthermore, while like other studies this dataset lacks information on potential confounders such as diet and lifestyle (Nasidze et al., 2009), due to shared cultural practices between members of the ultra-orthodox Ashkenazi Jewish community (Levine, 2015) they are likely to be more controlled for in this cohort than in others. For this reason, this cohort represents a unique opportunity to compare the salivary microbiome within a large number of individuals living in separate locations but nevertheless sharing a similar diet, lifestyle, and genetic background, and to investigate the long-term effect of shared upbringing on salivary microbiome composition.

2.2 Materials and Methods

2.2.1 Cohort

Cohort

The cohort contained data from 133 individuals within the same extended family (Family A) living in four different cities (I, II, III, IV) across three continents (see Levine et al. (2016) for more information). There were also samples available from 18 individuals from a separate smaller family (Family B), and 27 unrelated Ashkenazi Jewish controls. All individuals studied were of genetically confirmed Ashkenazi Jewish ancestry (Levine, 2015; Levine et al., 2016). Shared household was not directly available but Adam Levine's doctoral work on this cohort had established that individuals within this community grow up in the shared household that their parents live in, then marry and subsequently leave the family home, having children at a median age of 21 (95% CI: 19-26) (Levine, 2015). Therefore, I inferred shared household according to age: I assumed that individuals aged 18 or younger at the time of sampling were living with their parents and individuals aged 25 or older were not. These tentative households were then verified by Adam Levine. Individuals aged between 19 and 24 were excluded from household analyses as household could not be reliably inferred.

For analysis of the effects of household, I included only households with two or more individuals so as to remove the possibility that I was only measuring inter-individual differences, which can be large in the salivary microbiome (Nasidze et al., 2009; Utter et al., 2016). 26 individuals were living with at least one other individual at the time of sampling in a total of nine households. An additional 35 individuals who had grown up in a shared household with at least one other individual in the cohort, but who were no longer living together were subsequently included in the analysis.

I took the possibility of identification of participants due to the combined inclusion of (inferred) household, age, sex, and city seriously, and anonymized the four cities to remove this risk. Ethical and research governance approval was provided by the National Research Ethics Service London Surrey Borders Committee and the UCL Research Ethics Committee. Written informed consent was provided by all participants.

2.2.2 Sampling and extraction

Sampling

Saliva samples were collected in sterile tubes containing saliva preservative buffer as per the method of Quinque et al. (2006). 2x saliva preservative buffer was prepared according to the following protocol: 50 mM Tris pH 8.0, 50 mM EDTA pH 8.0, 200 mM NaCl, 1% (w/v) SDS and 50mM sucrose dissolved in ddH_2O, followed by a filter sterilization through a 0.2 μm filter. 2ml of saliva was collected from each participant and 500ml of 2x saliva preservative buffer was added. After that, 1μl of proteinase K (Sigma-Aldrich Company Ltd, Dorset, UK), 75μl of 10% SDS and 2μl of 10% azide per ml was added to the samples and incubated overnight at 50°C and stored at -200°C. Bacterial DNA was extracted with the PurElute Bacterial Genomic Kit (Edge Biosystems, Gaithersburg, MD) according to the manufacturers instructions from 0.5 ml sample/buffer mix and the dried DNA pellet was re-suspended in 40μl of DNase RNase free water. After DNA extraction, three spikes were added to all samples for the purposes of quality control in a final concentration of 4pg/ml, 0.4pg/ml and 0.08pg/ml, respectively (see below).

PCR amplification, purification and sequencing

The Mastermix 16S Basic containing MolTaq 16S DNA polymerase (Molzym GmbH & Co.KG, Bremen, Germany) was used to generate PCR amplicons. PCR amplicons were purified in two rounds using the Agencourt AMPure system (Beckman Coulter, Beverly, Massachusetts) in an automated liquid handler Hamilton StarLet (Hamilton Company, Boston, Massachusetts). DNA quantitation and quality control was performed using the Agilent 2100 Bioanalyzer system (Agilent Technologies, Inc., Santa Clara, CA). Amplification was performed with 785F and 1175R 16S rRNA primers that amplified the V5-V7 region of the 16S rRNA gene. Sequencing was performed with Illumina MiSeq (Illumina, San Diego, CA).

Quality control

With the aim of assessing technical variation across runs, samples had been spiked during library preparation with a fixed amount of synthetic DNA. Three unique spike sequences (350 bases in length) were designed which could be easily identifiable for quality control purposes. I was not involved in the decision to use these spikes, which was made before my involvement in the project; for more information on them see the Supplementary Material of the associated publication (Shaw, Ribeiro, et al., 2017).

I found that the number of spike sequences and the number of putative 16S rRNA sequences (length between 350 and 380 bases) were negatively correlated with each other, which would be expected due to the limited total sequencing depth of the Illumina Miseq (Figure A.2a). The variation in reads corresponding to this spike across samples was independent of run. After initial concern about the possibility of spikes affecting the

Variable	OTU F-statistic (p-value)	MED F-statistic (p-value)
Sequencing plate	2.876 (0.001)	3.132 (0.001)
Family	2.134 (0.004)	2.319 (0.001)
Gender	0.963 (0.442)	1.135 (0.243)
Age	2.516 (0.009)	2.560 (0.001)

Table 2.1: Results from an example permutational analysis of variance for de novo OTUs at 98.5% sequence similarity and MED phylotypes. For every variable, MED phylotypes result in a greater explanation of variance.

downstream analysis, Andre Ribeiro resequenced a subset of samples without spikes. My analysis confirmed the same qualitative differences, although with a clear batch effect (Figure A.2b), implying that the addition of spikes would not have had a negative impact on downstream analysis.

Phylotyping and taxonomic classification

Paired-end reads were merged with `fastq-mergepairs` in VSEARCH v1.11.1 (Rognes et al., 2016), discarding reads with an expected error >1. As the expected length of the V5-V7 region was 369 bases, I discarded sequences with <350 or >380 bases. I then clustered sequences with MED (Eren, Morrison, et al., 2014). MED requires that the variation in read depth across samples does not differ by several orders of magnitude, so I discarded samples with fewer than 5,000 reads and subsampled to a maximum number of 20,000 sequences, resulting in 6,353,210 sequences. I ran MED v2.1 with default parameters: minimum substantive abundance of an oligotype, $M = 627$ (1/5000 total sequences); maximum nucleotide variation allowed within an oligotype $d = 4$. This analysis removed 1,044,114 sequences as outliers due to the minimum substantive abundance criterion (853,159) and the nucleotide variation criterion (190,955). I assigned taxonomy to MED phylotypes with RDP (Q. Wang et al., 2007) against HOMD (T. Chen et al., 2010).

As mentioned in Section 1.1.3, MED offers higher resolution compared to OTU picking methods, and has previously been shown to differentiate the composition of the oral microbiome of individuals over time even within the same genus in a study of plaque samples (Utter et al., 2016). I wanted to see if this higher resolution improved results in this dataset as well, so also clustered sequences into *de novo* OTUs at 98.5% sequence similarity with VSEARCH v1.11.1 (Rognes et al., 2016) following the pipeline outlined in Figure A.1. The compositional dissimilarities between samples from using MED and OTUs were highly correlated (Spearman's $\rho = 0.88$, $p < 0.001$; Figure A.3). In an example analysis of variance MED phylotypes allowed increased statistical power, with a greater F-statistic for every variable considered (Table 2.1) confirming that MED offers greater differentiating power between samples, consistent with the literature (Eren, Morrison, et al., 2014). All further analysis in this chapter uses MED phylotypes rather than

OTUs.

Comparison to HMP samples from various sites in the mouth sequenced with V3-V5 primers (Eren, Borisy, et al., 2014) also indicated that Ashkenazi Jewish individuals do not have a significantly different oral microbiome from other populations (Figure A.4). However, the use of different primers makes it difficult to reach a robust conclusion on this point.

2.2.3 Inclusion of host genetics

Genetic relatedness was clearly linked with salivary microbiome dissimilarity with a simple correlation (Figure 2.1a), but I wanted to test the effect of using different measures of relatedness, as unusually this dataset had a genuine genetic distance available rather than a proxy. I investigated the effect of relatedness between individuals on salivary microbiome composition using both genetic kinships k_g (based on genome-wide SNPs) and pedigree kinships k_p (based on the pedigree). Pedigree kinships were calculated with kinship2 (Sinnwell et al., 2014) and genetic kinships with LDAK v5.94 (Speed et al., 2012) using genome-wide SNP data from either the Illumina HumanCytoSNPv12 (Illumina, USA) or the Illumina HumanCoreExome-24, as described elsewhere (Levine et al., 2016).

These genetic kinships k_g are normalized to have a mean of zero, and correspond to genetic similarity between individuals. k_g correlates with the pedigree kinship k_p but there can be substantial spread around the expected values due to the random nature of genetic inheritance making k_g a more accurate measure of true genetic similarity between individuals (Speed and Balding, 2014). These kinships are differently defined so have different values and scalings, although they are correlated (Figure 2.1b). I converted kinships to dissimilarities scaled between 0 and 1 with:

$$d_g = 1 - \frac{k_g - \min(k_g)}{\max(k_g) - \min(k_g)} \tag{2.1}$$

$$d_p = 1 - 2k_p \tag{2.2}$$

I then converted these dissimilarities to Euclidean distances using `dist()` (Figure 2.1c) and then used `metaMDS()` to produce a MDS ordination using vegan v2.4.4 (Oksanen, 2016). For the ordination I followed Blekhman et al. (2015) who investigated host genetic variation and its association with microbiome composition. I similarly used $k = 5$ dimensions, and found that using more did not affect the conclusions. I normalised the MDS axes using a Box-Cox transformation, with parameter λ calculated from `BoxCox.lambda` using forecast v8.2 in R (Hyndman, 2017) with the formula

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda}. \tag{2.3}$$

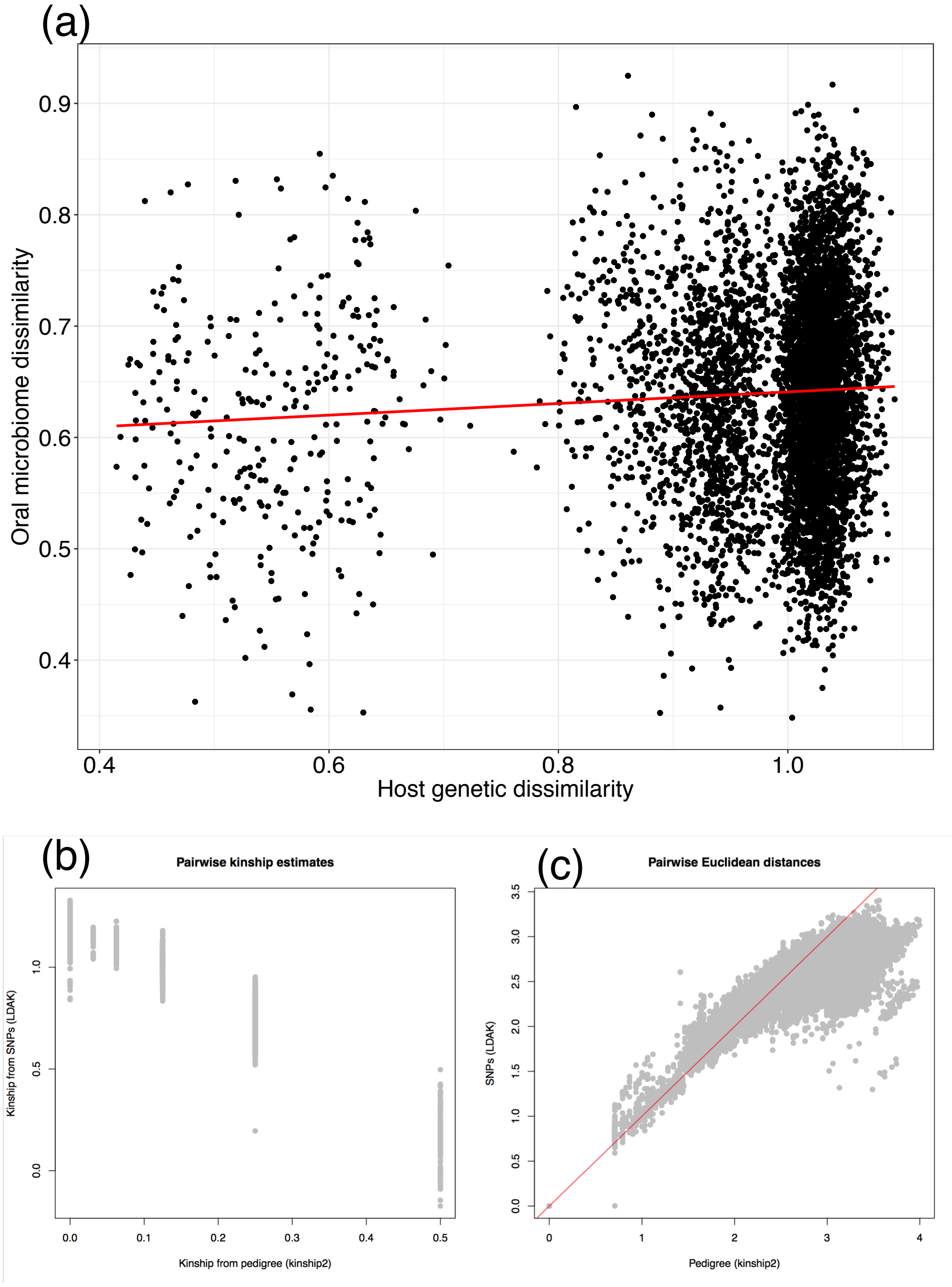


Figure 2.1: Genetic dissimilarities based on SNPs weakly correlate with salivary microbiome dissimilarities and are different from pedigree kinships. (a) Pairwise genetic dissimilarity appears weakly but significantly correlated with pairwise salivary microbiome dissimilarity (Mantel statistic $r = 0.065$, $p = 0.001$). Genetic dissimilarity was based on kinship calculated with LDAK, such that higher values indicate lower relatedness. Salivary microbiome dissimilarity was calculated with the Bray-Curtis metric. The rough clusters visible from left to right correspond to (i) siblings/parent-child pairs, (ii) first cousins, and (iii) less-related individuals. **(b)** Correlation of observed calculated kinship with LDAK against the expected kinship from the pedigree, showing pairwise kinship estimates before rescaling to dissimilarities (see above). **(c)** Pairwise Euclidean distances calculated from rescaled dissimilarities.

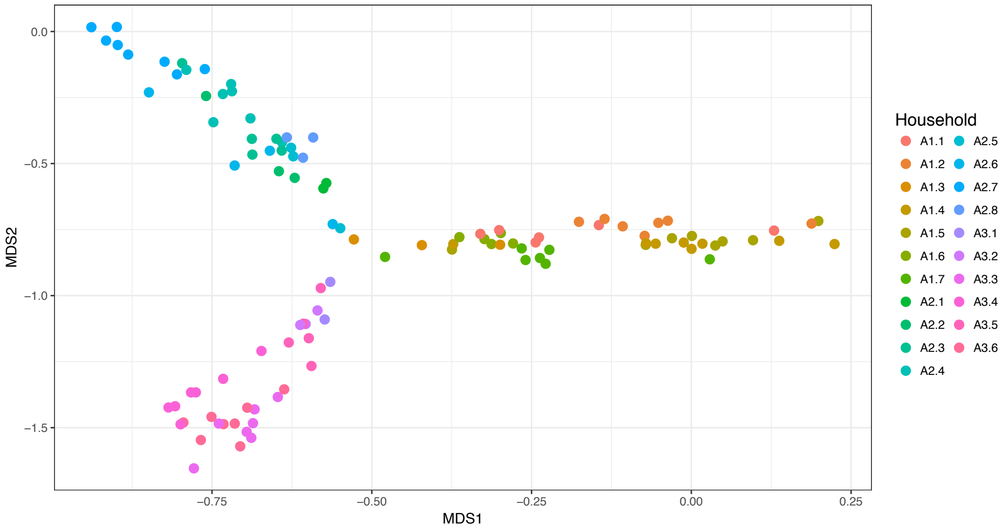


Figure 2.2: Multidimensional scaling of kinships shows the familial structure of the cohort. This example multidimensional scaling of samples based on kinships calculated using LDAK showing the first and second metric multidimensional scaling (MDS). The family structure is visible (colors) from the three arms of the pedigree (Figure A.5). The first five MDS axes were used to describe host genetic variation, following Blekhman et al. (2015).

2.2.4 Statistical analysis

I calculated Bray-Curtis dissimilarities between samples based on relative abundances of phylotypes, excluding samples with fewer than 1,000 reads. Variance explained in Bray-Curtis dissimilarities was calculated using the adonis function from the vegan v2.4.1 package in R (Oksanen, 2016), which performs a permutational analysis of variance of distance matrices (Anderson, 2001). I used $n = 9,999$ permutations with the following order of variables,

```
plate+Gender+samplingAge+MDS1+MDS2+MDS3+MDS4+MDS5+(environment)
```

where `(environment)` was either `city`, `household`, or `city+household`. When including both city and household, permutations were stratified by city to avoid permuting mixing the nested hierarchical levels by permuting households across cities.

2.3 Results

2.3.1 Description of cohort

The families analysed in this study have been already described in detail by Levine et al. (2016). All individuals sampled were from the ultra-orthodox Jewish community. Family A comprised over 800 individuals living in at least eight cities in four countries. Family B comprised over 200 individuals living in at least four cities in three countries. The unrelated controls were sampled from the same community as the two families. In total, data were generated from 133 individuals in Family A, 18 individuals in Family B, and 27 controls.

There were 271 phylotypes in the total dataset, all of which were present when considering just Family A. 49 of these phylotypes were present in >95% of individuals within Family A, with the Firmicutes the most abundant phyla (Figure 2.3a) as observed in previous oral microbiome studies (Stahringer et al., 2012; Dewhirst et al., 2010). The most abundant genera were *Streptococcus* (30.4%), *Rothia* (18.5%), *Neisseria* (17.1%), and *Prevotella* (17.1%). Composition of samples was similar between the two families (A and B) and the unrelated controls (Figure 2.3b). These groupings had a small but significant effect in an analysis of variance ($R^2 = 0.015$, $p < 0.01$) but this is typical of comparisons between such large groups that may differ in an unknown number of confounded variables (e.g. diet, genetics, lifestyle). I concluded that Family A was at the very least a representative sample capturing the majority of the variation present in the wider Ashkenazi Jewish population, if not also non-Ashkenazi-Jewish individuals. This cohort was originally collected for a study of the genetics of Crohn's disease (Levine et al., 2016), and 28 individuals had a diagnosis of the disease at the time of saliva sample acquisition. I found no significant effect of Crohn's disease on salivary microbiome composition with an exploratory analysis of variance ($R^2 = 0.009$, $p = 0.101$, $n = 148$) accounting for other variables. It was therefore not included as a covariate in further analysis.

2.3.2 Host genetic similarity and microbiome similarity

I performed an exploratory analysis on individuals in Family A with both genetic and microbiome data available ($n = 111$), and found that genetic kinship was weakly but significantly associated with salivary microbiome dissimilarity computed using Bray-Curtis dissimilarities (Figure 2.1a; Mantel test $r = 0.065$, $p = 0.001$). This analysis does not take into account confounding by shared environment, and therefore sets a probable upper bound on the variation that can be attributed to host genetics. An exploratory analysis of microbiome variation across a subfamily within Family A ($n = 44$) showed that individuals from the same household had a more similar microbiome composition as measured by Bray-Curtis dissimilarity (mean ± s.d, 0.623 ± 0.088) compared with individuals from different households (0.652 ± 0.084), and this difference was significant (two-sided t-test, $p < 0.001$). An exploratory visual representation of this variation showed some possible clustering by household, with large overlap between households in a two-dimensional NMDS plot (Figure 2.4). However, such an analysis is insufficient; household is obviously correlated with variation in host genetics (Figure 2.2) because parents tend to live with their children. This emphasizes the need for a quantitative approach looking at the effect of both household and genetics simultaneously as well as other potential confounders.

The approach I chose to use was adonis, which performs a permutational analysis of

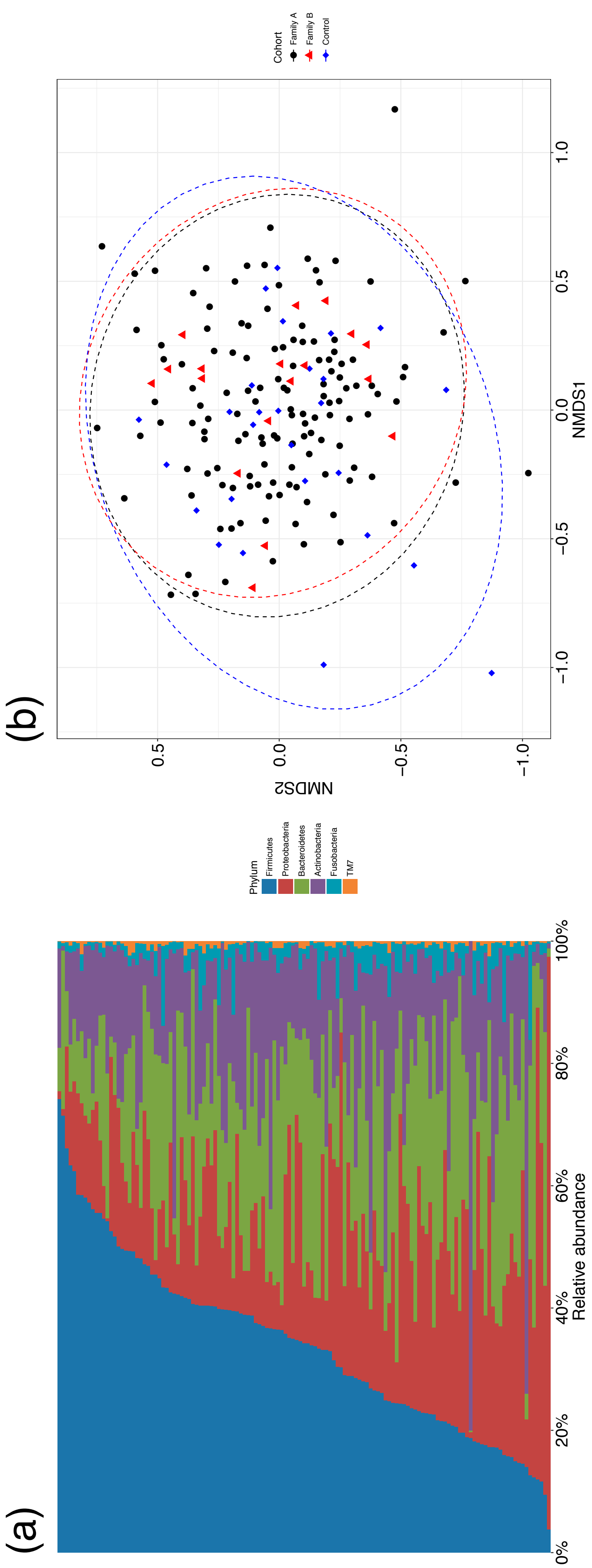


Figure 2.3: This cohort contains a representative sample of variation in oral microbiome composition. (a) Relative abundance of the six bacterial phyla found in saliva samples from Family A, sorted by decreasing Firmicutes content. Color scheme adapted from Stahringer et al. (2012). Taxonomy was assigned to 271 MED phylotypes using RDP based on the HOMD database. **(b)** Non-metric multidimensional scaling based on Bray-Curtis dissimilarities between samples shows high overlap between Family A (black circles), Family B (red triangles), and unrelated Ashkenazi Jewish controls (blue diamonds).

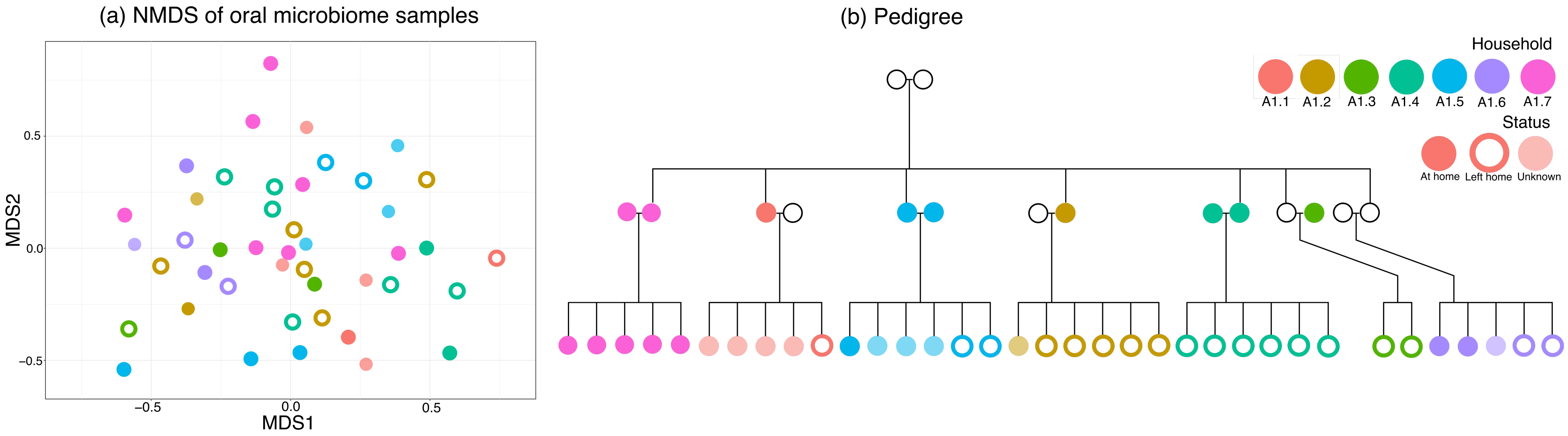


Figure 2.4: Oral microbiome composition is possibly associated with household. Oral microbiome samples weakly cluster by household (colours), shown by **(a)** a non-metric multidimensional scaling based on Bray-Curtis dissimilarities between samples from **(b)** $n = 44$ individuals in a particular subfamily within Family A. This figure includes individuals who are currently living together (filled circles), those who had moved out of their childhood home (empty circles), and those for whom data was missing (faint circles). This weak qualitative clustering could be due to shared environment or also due to shared genetics, suggesting a quantitative analysis is required.

variance in community composition using a sequential sum-of-squares approach (Anderson, 2001). I used Bray-Curtis dissimilarities to quantify differences in salivary microbiome composition between individuals. The following sections present a combination of analyses attempting to quantify the effects of shared environment and genetics. The analysis groups were as follows: A_{26} ($n = 26$ individuals co-habiting with at least one other; Table 2.2a), A_{61} ($n = 61$ individuals who had co-habited with at least one other, either at time of sampling or beforehand; Table 2.2b), A_{82} ($n = 82$ individuals across four different cities who were not necessarily co-habiting with another; Table 2.3), and A_{111} ($n = 111$ individuals with host genetic information available; Table 2.4).

These groups are nested within each other i.e. $A_{26} \in A_{61} \in A_{82} \in A_{111}$. The magnitude of the effect of a variable is given by the amount of variance explained (R^2 in tables).

2.3.3 Shared household affects salivary microbiome composition

I performed a permutational analysis of variance on the salivary microbiome dissimilarities for A_{26} (26 individuals within Family A, each of whom lived in a household with at least one other individual in the cohort). At the time of sampling, these co-habiting individuals lived across a total of 16 households in four cities (I, II, III, IV). Host genetics was controlled for by using $k = 5$ axes from a MDS of pairwise genetic distances between individuals (see Section 2.2.3).

There was no significant effect of any of the MDS axes of host genetics, suggesting that host genetics in closely-related individuals does not significantly affect microbiome composition. I investigated the effect of environment using two levels of geography: city and household (Table 2.2). A city-only model showed no significant effect of environment ($R^2 = 0.08$, $p = 0.4$), whereas a household-only model showed a significant effect ($R^2 = 0.30$, $p = 0.001$). This was reproduced in a model containing both geographic variables, with permutations stratified by city, where household was still a significant effect ($R^2 = 0.22$, $p = 0.001$), suggesting that differences at the level of household are more important than at larger geographical scales. I confirmed that city-level effects were small by extending the sample to A_{82} ($n = 82$ individuals across the four cities who were not necessarily cohabiting with others; I: 48, II: 13, III: 12, IV: 9), and found that city still had a small effect, although it was significant ($R^2 = 0.053$, $p < 0.01$). In this analysis I also found no significant effect of genetics, but age was significant ($R^2 = 0.028$, $p = 0.01$) (Table 2.3).

2.3.4 Spouses share taxa at the sub-genus level

Restricting the analysis to only married couples within Family A ($n = 16$, eight couples), shared household explained even more of the variance ($R^2 = 0.591, p = 0.001$). Subtle variations in the relative abundance of phylotypes within the same genus between households were observable, even within the same city location. For example, *Leptotrichia* phylotypes qualitatively varied consistently between spouse pairs and these patterns were

(a) A_{26}	City only		Household only		City and household†	
	R^2	p	R^2	p	R^2	p
Sequencing plate	0.048	0.19	0.048	0.075	0.048	0.458
Gender	0.032	0.724	0.032	0.4	0.032	0.467
Age	0.069	0.017	0.069	0.004	0.069	0.013
MDS1	0.031	0.757	0.031	0.537	0.031	0.727
MDS2	0.05	0.142	0.05	0.052	0.05	0.099
MDS3	0.03	0.807	0.03	0.585	0.03	0.862
MDS4	0.049	0.162	0.049	0.054	0.049	0.097
MDS5	0.029	0.824	0.029	0.614	0.029	0.791
City	0.08	0.4			0.08	0.178
Household			0.3	0.001	0.22	0.001
Residuals	0.582		0.362		0.362	
Total	1		1		1	

(b) A_{61}	City only		Household only		City and household†	
	R^2	p	R^2	p	R^2	p
Sequencing plate	0.029	0.018	0.029	0.012	0.029	0.013
Gender	0.018	0.258	0.018	0.219	0.018	0.257
Age	0.038	0.002	0.038	0.001	0.038	0.002
MDS1	0.014	0.668	0.014	0.607	0.014	0.74
MDS2	0.017	0.362	0.017	0.305	0.017	0.44
MDS3	0.02	0.173	0.02	0.141	0.02	0.263
MDS4	0.02	0.15	0.02	0.118	0.02	0.147
MDS5	0.012	0.783	0.012	0.744	0.012	0.943
City	0.056	0.149			0.056	0.934
Household			0.239	0.021	0.183	0.044
Residuals	0.777		0.594		0.594	
Total	1		1		1	

Table 2.2: Permutational analysis of variance (adonis) results for co-habiting individuals. (a) A_{26}: 26 individuals who lived in the same household as at least one other individual. **(b)** A_{61}: 61 individuals who had at least co-habited at some point. Household is always significant and explains the most variance of any variable (>18%) even in a model that nests permutations within cities. The order of variables in the model is given by their order in the table. † Permutations stratified by city in this analysis

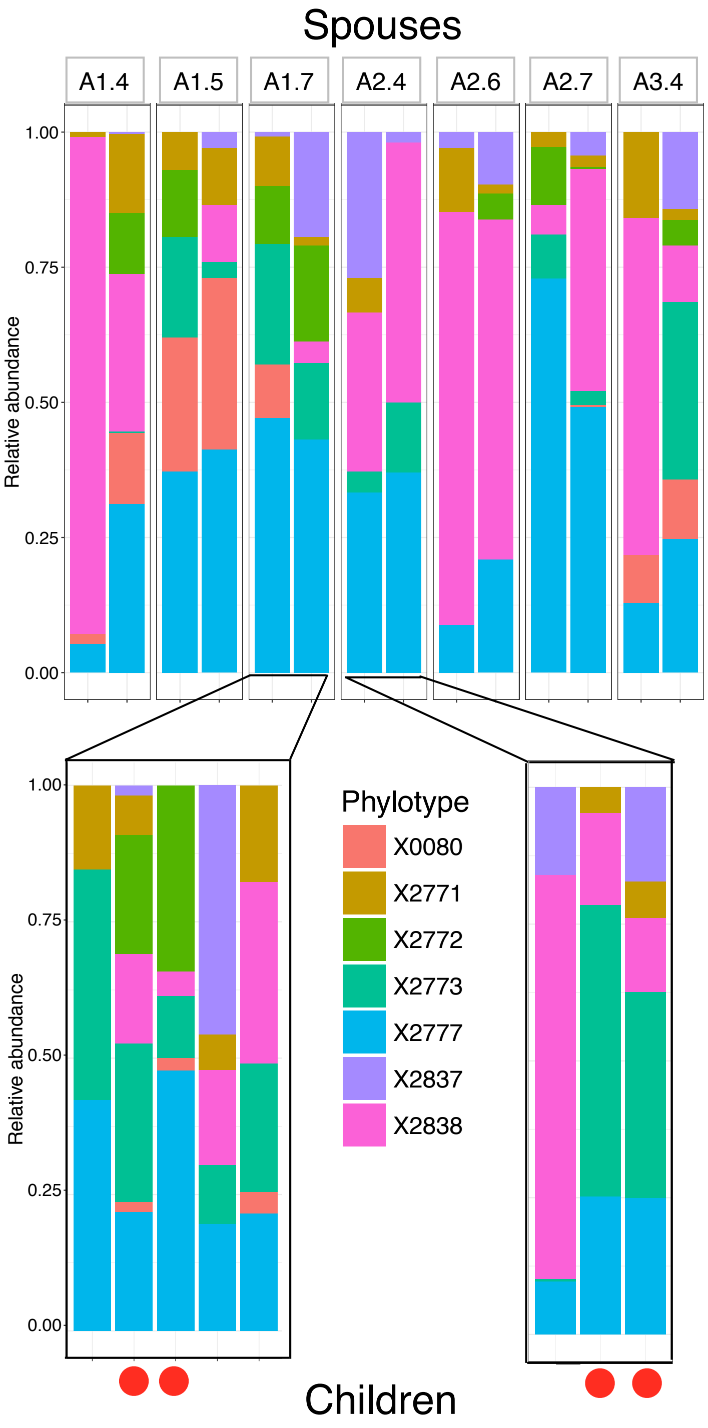


Figure 2.5: Household-level variation within a genus, shown here with the relative abundance of phylotypes within *Leptotrichia.* The relative abundance of phylotypes within seven pairs of spouses shows clear associations with household. These patterns are to some extent recapitulated in their children. Looking at children still living at home, MED phylotype X2772 is not observed in any individual from household A2.4, but is found in both spouses and two children living in household A1.7. Red dots indicate children aged 10 or under at time of sampling, who appear more similar to each other than other pairs of children. For an indication of variation in other genera between spouses, see Figure 2.6.

	R^2	p
Sequencing plate	0.028	0.001
Gender	0.015	0.136
Age	0.028	0.001
MDS1	0.009	0.775
MDS2	0.013	0.311
MDS3	0.015	0.141
MDS4	0.017	0.05
MDS5	0.013	0.306
City	0.053	0.005
Residuals	0.809	
Total	1	

Table 2.3: Permutational analysis of variance for individuals living across four cities worldwide. This group A_{82} contained 82 individuals who were not necessarily co-habiting with others. In this analysis, city has no significant effect on salivary microbiome composition.

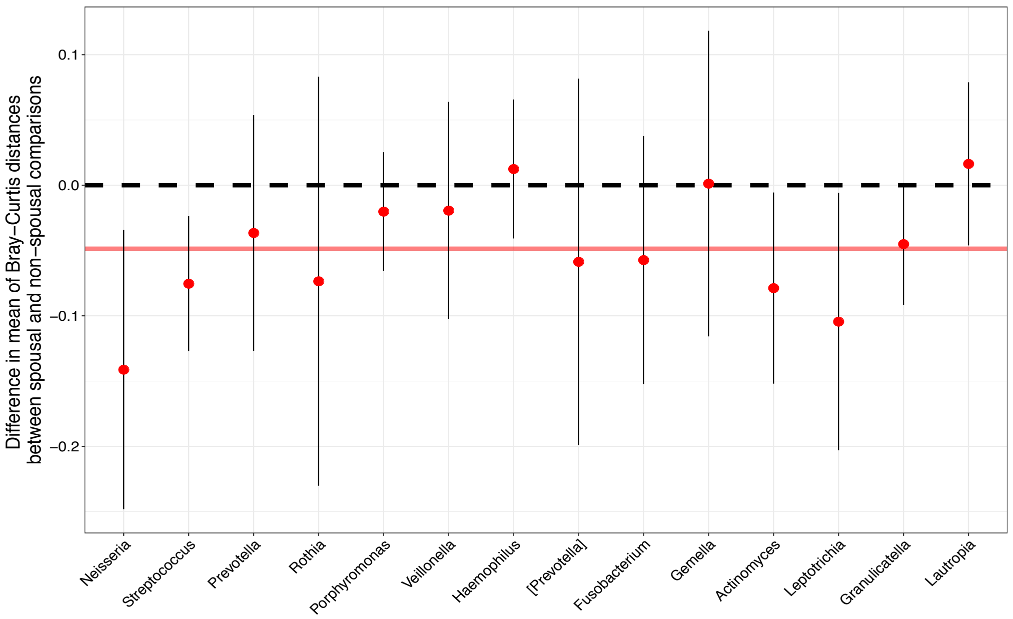


Figure 2.6: Spouse pairs significantly share taxa at the sub-genus level. The difference in mean Bray-Curtis dissimilarities (red, with 95% confidence intervals in black) for spouses vs. non-spouses at the sub-genus level i.e. calculated using only the phylotypes within that genus. The similarity at sub-genus level is significantly lower for spouses on average across these genera, as indicated by the red line.

also seen in children living at home (Figure 2.5). MED phylotype X2772 was present in both spouses in household A1.7, and was also present in the two youngest children within that household (aged 10 or under). Similarly, within household A2.4 the two children aged 10 or under were more similar in *Leptotrichia* phylotypes than an older child.

Quantitatively, repeating a permutational analysis of variance based only on the composition of phylotypes within *Leptotrichia* showed that spousal pair explained 68.4% of variance, although this was not significant ($R^2 = 0.684$, $p = 0.068$). Extending to the top 12 abundant genera, similar patterns were also visible (not shown). Spouses on average had a significantly more similar sub-genus phylotype composition than non-spouses (mean ± s.e. difference in Bray-Curtis dissimilarities for each genus: -0.048 ± 0.013; Figure 2.6).

2.3.5 Persistence of household effects after co-habiting

There were an additional 35 individuals who had grown up in a household with at least one other individual present, but who no longer lived together at time of sampling. To see if the effects of household persisted, I repeated analysis of variance with these individuals included along with the cohabiting-individuals (A_{61}, Table 2.2). The effect of household remained significant ($R^2 = 0.183$, $p = 0.044$), and no axes of human genetic variation were significant ($p > 0.05$). Age had a significant effect ($R^2 = 0.038$, $p < 0.01$).

Other variables such as age and sequencing plate had smaller effects than household in all the analyses of variance.

Order of variables does not change conclusions

I reasoned that the order of variables presented in this chapter for the adonis analyses was the appropriate one for the intended purpose of testing for effects of household after controlling for other variables (Section 2.2.4). However, it should be noted that adonis explains variance by a sequential sum-of-squares approach, also known as a Type I sum-of-squares (Oksanen, 2016). This means that the ordering of variables can have an effect with an unbalanced design. To check this was not biasing the results and therefore the conclusions about the important factors for salivary microbiome composition, I also investigated the effect of randomly permuting the order of variables in the model formula. I ran adonis ($n = 999$ permutations) on 1000 permutations of the variables in the model formula (2.2.4) for the A_{61} dataset i.e. individuals who had cohabited at some point with at least one another. .

Some permutations ($n = 498$) resulted in variables dropping out of the model due to the unbalanced design i.e. the variable added no additional information, and therefore I did not include results from these for fairness of comparison. The remaining permutations ($n = 502$) gave a full model. I corrected for multiple testing using the Benjamini-Hochberg correction (Benjamini and Hochberg, 1995). Age and household were always significant ($q < 0.05$) in all models. Sequencing run was significant in 213 out of 502 models. Gender was never significant. Crucially, no MDS axis of genetic variation was significant in any of the 502 models. Therefore, I concluded that household was the dominant factor, and not host genetics, as found with the original variable order.

2.3.6 Relying on pedigree kinships produces a genetic signal

	Pedigree (kinship2)		SNPs (LDAK)	
	R^2	p	R^2	p
Sequencing plate	0.028	<0.001	0.028	<0.001
Gender	0.011	0.094	0.011	0.096
Age	0.023	<0.001	0.023	<0.001
MDS1	0.01	0.174	0.011	0.119
MDS2	0.007	0.706	0.01	0.231
MDS3	0.012	0.063	0.011	0.131
MDS4	0.016	0.009	0.011	0.111
MDS5	0.009	0.325	0.007	0.617
Parental household	0.215	<0.001	0.217	<0.001
Residuals	0.67		0.671	
Total	1		1	

Table 2.4: Comparison of pedigree-based and genome-wide measures of kinship to take host genetics into account. Shown here are results from adonis on salivary microbiome dissimilarities of $n = 111$ individuals. Using pedigree information to produce kinship results in a significant association with human genetics via the fourth MDS axis ($p = 0.011$), which is not present using kinships calculated with LDAK based on genome-wide SNPs.

To test whether the conclusions required using kinships estimated from genome-wide SNP data for individuals, or whether pedigree information was sufficient, I also repeated the analyses using pedigree kinships for A_{111} (Section 2.2.3). Using pedigree kinships resulted in a small but significant amount of variation in microbiome composition being attributable to host genetics via the fourth MDS axis ($R^2 = 0.016$, $p < 0.01$, Table 2.4).

2.4 Conclusions

2.4.1 Discussion

In this chapter I have conducted the first simultaneous investigation of the role of environment and host genetics in shaping the human salivary microbiome, using a cohort of closely-related individuals within a large Ashkenazi Jewish family. I found a weak correlation between host kinship and salivary microbiome similarity before taking shared household into account, and an apparent small but significant effect of genetics when using kinships based on the family pedigree as proxies for genetic similarity. However, when using kinship estimates based on genome-wide SNPs between individuals and simultaneously controlling for shared household with a permutational analysis of variance, I found no support for any clear effect of human genetics, suggesting that shared environment has a much larger effect than genetics and is the dominant factor affecting the salivary microbiome. Typically shared household had an order of magnitude greater effect compared with other significant variables. For example, in the analysis where city was also used as an environmental variable, the variance explained was as follows: household (18.3%), age (3.8%), sequencing plate (2.9%) (Table 2.2b).

I also found that younger children living in the same household shared subtle variations in phylotype abundance within genera with their parents (Figure 2.5). However, despite a persistence of household effects it would be wrong to conclude that the salivary microbiome is completely fixed once established, as it clearly has aspects that can change over time. For example, shared household explained more variation for spousal pairs (likely due to frequent contact between them) and phylotypes observed in younger children and their parents were not seen in older children (likely due to less frequent contact between them). Taken together, these observations support the view that human genetics does not play a major role in shaping the salivary microbiome, at least not in individuals of the same ethnicity, compared to the environment and contact with other individuals.

These results confirm the seemingly paradoxical situation that the salivary microbiome is largely consistent across global geographical scales, but can show large variation between households in the same city. Previous studies have also found evidence of small variations in salivary microbiome composition comparing samples across a global scale (Nasidze et al., 2009). As noted previously, this variation could be influenced by differences in environmental or cultural factors, in which case controlling for these differences would decrease the amount of geographical variation. All individuals in this study followed a traditional Ashkenazi Jewish lifestyle and subsequently are thought to share a similar diet and lifestyle regardless of geographic location (Levine, 2015) which may reduce the variation attributable to city-level differences.

The establishment of the oral microbiome early in life may lead to the persistence of a similar composition over several years. The microbial composition of sites within the mouth has been previously observed to be persistent within individuals over periods of months (Utter et al., 2016) to a year (David et al., 2014) and I found similar strain-level variation between spouses and their young children as observed between individuals by Utter et al. (2016) (Figure 3). These findings indicate the persistence of household effects in individuals no longer co-habiting, suggesting that the salivary microbiome composition established early in life via shared upbringing is able to persist for at least several years. It has been observed that monozygotic twins do not have significantly more similar gut microbiomes than dizygotic twins (Turnbaugh et al., 2009). Stahringer et al. (2012) observed the same effect in the salivary microbiome, and also found that twins' salivary microbiomes became less similar as they grew older and ceased cohabiting, concluding that 'nurture trumps nature' in the salivary microbiome. My results from a large number of related individuals (rather than twins) support this view including the persistence of shared upbringing effects. Shared upbringing appears to be the dominant factor affecting microbiome composition in both the gut and the mouth, rather than genetic similarity. This may have implications for understanding the familial aggregation of diseases such as inflammatory bowel disease, which has been suggested to have an environmental component (Nunes et al., 2011).

The salivary microbiome appears far more resilient to perturbation compared to the gut microbiome, with a rapid return to baseline composition after a short course of antibiotics (Zaura et al., 2015). While this could be because of the pharmacokinetics of the antibiotics involved, Zaura et al. speculate that this difference may be due to the salivary microbial ecosystem's higher intrinsic resilience to stress, as the mouth is subject to more frequent perturbation (Marsh et al., 2015). This chapter supports the dominant role of the environment in affecting salivary microbiome composition and suggests that another important factor in long-term persistence may be the regular reseeding of the ecosystem with bacteria from the external environment.

The fact that the conclusion on the lack of effect of genetics required kinship based on genome-wide single nucleotide polymorphism (SNP) markers rather than pedigree (Table 2.4) casts doubt on the reliability of pedigrees for calculating relatedness. There are several possible reasons for a discrepancy between kinship estimates from pedigrees and allele sharing (Speed and Balding, 2014). One possibility is errors in the pedigree, most likely due to extra-pair paternities, although this explanation can be ruled out in this dataset. More importantly, inherent stochasticity in the Mendelian process of inheritance means that although parents always pass on 50% of their genes to their offspring, SNPs are inherited together in blocks (i.e. haplotypes), meaning that the relatedness between two offspring in a family can be substantially different from 50%. Finally, and most importantly for this closely-related population, shallow pedigrees cannot fully capture complex inbreeding patterns. Thus, while pedigrees are a good model for host relatedness in microbiome studies of large randomly mating populations, I recommend that they be used with caution in closely-related families like this one.

2.4.2 Limitations

Because all individuals in the main cohort were members of the same extended Ashkenazi Jewish family, the genetic variation in this dataset is therefore much lower than between individuals from a wider population. It is conceivable that host genetics between more distantly-related individuals may play a significant role in affecting salivary microbiome composition. However, a recent study of the nasopharyngeal microbiome among Hutterite individuals (a founder population in North America) found detectable associations between host variation and microbial composition with a similar cohort size (Igartua et al., 2017), demonstrating that limited genetic variation can be associated with the composition of other microbiomes. It may simply be that the salivary microbiome is relatively unaffected by such variation, perhaps because of reduced interaction with mucosal surfaces.

Furthermore, I only looked at overall genetic similarity, assessed using community comparison metrics based on taxa abundances. The results presented therefore do not preclude the existence of fine-scale links between particular microbial taxa and individual genetic loci, particularly in immune-sensing genes such as those identified in the gut microbiome by Bonder et al. (2016) using a much larger cohort of $1,514$ individuals; my more opportunistic investigation was not designed or powered to detect such associations.

Additionally, the dataset lacked detailed information on the diet and lifestyle factors of individuals. However, the shared cultural practices within this ultra-orthodox Ashkenazi Jewish family mean that it is not unreasonable to assume they share a similar lifestyle and diet despite living in different locations around the world (Levine, 2015).

The continued effect of shared upbringing after leaving a household could be confounded by the fact that individuals may continue living near to the household where they grew up and interacting with the same individuals. If this were the case, then the apparent persistence could instead be due to the persistence of a shared environment beyond the household at a level intermediate between household and city, rather than the persistence of a stable salivary microbiome following environmental change. Finally, these samples represent only a single cross-sectional snapshot in time. More long-term longitudinal studies like the work of Stahringer et al. (2012) on twins are necessary to investigate the persistence of the salivary microbiome after its establishment early in life in a variety of relatedness settings.

2.4.3 Summary

By incorporating a measure of genetic relatedness using SNPs I have demonstrated that the overall composition of the human salivary microbiome in a large Ashkenazi Jewish family is largely influenced by shared environment rather than host genetics. An apparent significant effect of host genetics using pedigree-based estimates disappears when using genetic markers instead, which recommends caution in future microbiome research using pedigree relatedness as a proxy for host genetic similarity. Geographic structuring occurs to a greater extent at household level within cities than between cities on different continents. Living in the same household is associated with a more similar salivary microbiome, and this effect persists after individuals have left the household. This is consistent with the long-term persistence of the salivary microbiome composition established earlier in life due to shared upbringing, although longitudinal studies with more detailed metadata would be required to satisfactorily establish this link.

Chapter 3

Periodontal disease and the plaque microbiome

Declaration of contributions
This work was based on a dataset collected by the International Lipid-Based Nutrient Supplements Project (iLiNS) collaboration. I am grateful to the iLiNS committee for granting me permission to use the data for this work and also for the demographic metadata on participants. Ulla Harjunmaa performed dental assessments and provided the dental data. Ronan Doyle performed the 16S rRNA library preparation and sequencing. I performed all subsequent data analysis and wrote the associated paper, with feedback from all co-authors.

Publication
This work has been published in *Applied and Environmental Microbiology* as Shaw et al. (2016):
L.P. Shaw, U. Harjunmaa, R. Doyle, S. Mulewa, D. Charlie, K. Maleta, R. Callard, A. S. Walker, F. Balloux, P. Ashorn, and N. Klein (2016). Distinguishing the signals of gingivitis and periodontitis in supragingival plaque: a cross-sectional cohort study in Malawi. *Applied and Environmental Microbiology* **82**(19), 60576067. doi: 10.1128/AEM.01756-16.

3.1 Introduction

Periodontal disease is a major public health problem, particularly in low-income settings like those found in sub-Saharan Africa (P. E. Petersen et al., 2005). In periodontal disease, the immune system responds with inflammation to oral biofilms (Van Dyke, 2008) which can eventually result in the formation of periodontal pockets and loss of teeth (Section 1.2.3). Aside from irreversible tooth loss, chronic periodontitis may also increase the risk of adverse systemic conditions (X. Li et al., 2000) such as cardiovascular disease (Y.-H. Yu et al., 2015) and preterm birth, although for the latter, different studies have reported conflicting results (Ide and Papapanou, 2013). The association between periodontitis and systemic disease may be due both to increased systemic inflammation and to translocation of bacteria into the blood stream (Hajishengallis, 2014). Despite its importance, the microbial ecology of periodontal disease in different oral habitats remains incompletely understood. Studies of the oral microbiome in periodontal disease typically focus on small populations in developed countries with advanced dental healthcare systems, which may not be representative of the natural history of periodontal disease in the absence of treatment (Baelum and Scheutz, 2002).

After an initial focus on identifying particular periodontal ‘pathogens’ (Socransky et al., 1998) it is now widely accepted that oral bacterial communities undergo a shift into dysbiosis (Jiao et al., 2014) and that the presence of particular disease-associated species may exacerbate the inflammatory reaction to commensal bacteria (Wade, 2013). In periodontitis, bacteria progress from supragingival plaque (on the gums) to subgingival plaque (below the gums) in periodontal pockets; studies typically focus on the subgingival communities. The two main features of periodontal disease are gingival inflammation (gingivitis) and the formation of periodontal pockets (periodontitis). While it is clear that gingivitis always precedes periodontitis (Schätzle et al., 2003), gingivitis does not always progress to periodontitis (Batchelor, 2014) suggesting that these may not simply represent different stages of a continuous spectrum of disease. While there is some evidence that a steady continuous progression can be expected (Jeffcoat and Reddy, 1991) most models involve acute bursts of exacerbation and longer periods of remission (Batchelor, 2014; Mdala et al., 2014).

Despite this knowledge, studies of oral bacteria in periodontal disease often fail to capture the full range of periodontal conditions: from health through gingivitis to periodontitis. Considering supragingival plaque in particular, comparing only healthy subjects with subjects suffering from periodontitis may lead to associations being attributed to periodontitis alone, despite the fact that they might also be present in subjects with gingivitis. To explain the progression of disease and identify factors uniquely attributable to periodontitis it is necessary to compare subjects across the full range of periodontal severities. In itself this is not a novel concept, with many previous studies investigating bacterial associations with disease using checkerboard DNA-DNA hybridization (Ximénez-Fyvie, Haffajee, Som, et al., 2000; Ximénez-Fyvie, Haffajee, and Socransky,

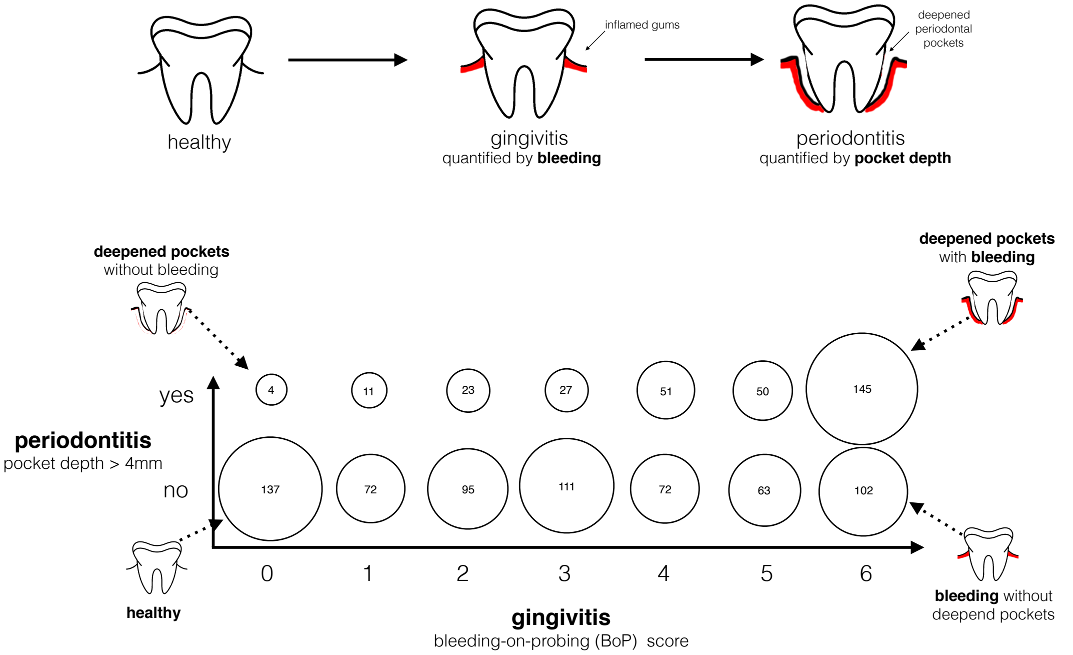


Figure 3.1: The clinical manifestations of periodontal disease include both periodontitis and gingivitis. Top. Considering periodontal disease as a continuous progression from health through gingivitis to periodontitis ignores the fact that clinical diagnosis usually rests on two separate criteria for each form of the disease. In this chapter, I look at associations with bleeding (used to diagnose gingivitis) and pocket depth (used to diagnose periodontitis), in order to identify signal that differentiates periodontitis from gingivitis. **Bottom.** The two-dimensional landscape of periodontal disease I use to model assocations in this chapter, with fourteen possible combinations of severities. The numbers in circles (and their size) indicate the number of women with these features in the cohort.

2000; Haffajee et al., 2009). These and other earlier studies were targeted at a small number of bacterial species, typically around 40, due to the limitations of the technology. The advent of high-throughput 16S rRNA gene amplicon sequencing has facilitated improved analysis of the total bacterial diversity in the oral cavity (Griffen et al., 2011; T. Chen et al., 2010) identifying around 1,000 species that may be present (Wade, 2013). Recent studies have used such amplicon sequencing to characterize subgingival plaque across a range of periodontal conditions, finding differences between subjects with gingivitis and periodontitis (Park et al., 2015; Camelo-Castillo et al., 2015). Work on supragingival plaque has been less common due to the fact that it does not have a direct link to inflammation and subsequent loss of attachment in periodontitis. It therefore remains ambiguous whether, for supragingival plaque, periodontitis can be simply considered as an advanced stage of gingivitis, or if there are detectable differences in bacterial composition.

Traditional case-control studies of the microbiome and disease that use a binary distinction and fail to take into account the continuous nature of disease are losing information that can be used to understand this etiology (Section 1.1.2). This point is of particular relevance for periodontal disease, where the milder form (gingivitis) and the more severe form (periodontitis) have distinct clinical manifestations defined in terms of different variables. To address this issue, in this chapter I investigate bacterial abundances in supragingival plaque using quantitative modelling that takes into account both gingivitis and periodontitis in a cross-sectional cohort of 962 Malawian women who had recently given birth (Ashorn et al., 2015). I quantify gingivitis using bleeding-on-probing (BoP) and periodontitis via a binary assessment (see Section 3.2.1) giving a two-dimensional 'landscape' of periodontal disease (Figure 3.1). I model the effect of gingivitis and periodontitis both separately and together in order to understand the features of the plaque community that are linked to more severe disease, rather than merely being associated with increased bleeding.

I use negative binomial models, originally developed for RNA-seq experiments (Love et al., 2014), making use of absolute (i.e. un-normalized) read counts to avoid losing information – a downside of other statistical approaches applied to marker gene data like rarefying (McMurdie and Holmes, 2014). After fitting a negative binomial distribution to count data for a given species, the mean of this distribution is then used as the output of a generalized linear model with a logarithmic link using experimental variables (e.g. disease severity) as inputs, allowing the identification of differentially abundant species. This approach considers bacterial species as independent, but in reality oral bacteria exist in complex polymicrobial biofilms (Teles et al., 2013; Mark Welch et al., 2016). Therefore, I also apply co-occurrence analysis to periodontitis-associated taxa to identify important members of the periodontal biofilm community.

In summary, here I aim to identify the effects of periodontitis on supragingival plaque after controlling for gingivitis severity, separating and distinguishing the signals of these two features of periodontal disease.

3.2 Materials and Methods

3.2.1 Data collection

Study population

Women analyzed in this study were participants in the iLiNS-DYAD-M trial (registration ID: NCT01239693) (Ashorn et al., 2015). This was a randomized controlled trial into the effects of three nutritional supplements on birth outcomes: lipid-based nutritional supplement (LNS), mixed micro-nutrients (MMN) or iron folate (IFA). Women were eligible for enrolment in the trial if they were <20 weeks pregnant, >14 years old, had no chronic illnesses requiring frequent medical care, no allergies, no evident pregnancy complications (edema, blood hemoglobin < 50 g/l, systolic blood pressure > 160 mmHg or diastolic > 100 mmHg), no earlier participation in the same trial and no concurrent participation in any other.

1,391 pregnant women were enrolled between February 2011 and August 2012 at antenatal clinics at two hospitals (Mangochi and Malindi) and two health centers (Lungwena and Namwera) in Mangochi district, Malawi. All women were self-reported non-

BoP	Periodontitis	N	Age (yrs)	Positive HIV test	Malaria[a]	BMI	Education (yrs)
0	No	140	23.4 (5.8)	27 (19.3%)	37 (26.6%)	22.7 (3.2)	5.6 (3.6)
1	No	72	23.9 (5.9)	7 (9.7%)	16 (22.2%)	22.6 (3.4)	5.1 (3.8)
	Yes	11	31.6 (6.1)	1 (9.1%)	1 (9.1%)	22.7 (2.4)	4.4 (3.3)
2	No	95	24.7 (6.2)	11 (11.6%)	22 (23.2%)	22.1 (2.6)	4.4 (3.6)
	Yes	23	27.5 (6.2)	5 (21.7%)	5 (21.7%)	21.7 (2.0)	2.7 (3.3)
3	No	111	24.4 (5.4)	11 (9.9%)	32 (28.8%)	21.7 (2.3)	4.3 (3.3)
	Yes	27	26.5 (5.7)	4 (14.8%)	3 (11.1%)	22.2 (2.7)	3.6 (3.0)
4	No	72	25.0 (6.4)	9 (12.5%)	16 (22.2%)	21.7 (2.2)	3.4 (3.0)
	Yes	51	26.9 (5.4)	8 (15.7%)	11 (21.6%)	21.8 (2.7)	3.3 (3.1)
5	No	63	24.9 (5.2)	7 (11.1%)	12 (19.0%)	21.6 (2.4)	4.0 (3.6)
	Yes	50	26.6 (5.9)	5 (10.0%)	7 (14.0%)	21.8 (3.1)	2.4 (2.8)
6	No	102	24.5 (5.5)	10 (9.8%)	18 (17.0%)	21.9 (2.3)	3.5 (3.0)
	Yes	145	28.3 (7.0)	30 (20.7%)	28 (19.3%)	22.1 (2.5)	2.9 (3.0)

BoP	Periodontitis	Anemia[b]	socioeconomic status[c]	Site[d]	Nutritional intervention[e]	Sequencing run[f]
0	No	36 (25.7%)	0.38 (1.22)	36 / 37 / 18 / 49	43 / 53 / 44	47 / 49 / 41 / 3
1	No	12 (16.7%)	0.19 (1.11)	25 / 9 / 17 / 21	32 / 19 / 21	34 / 26 / 12 / 0
	Yes	3 (27.3%)	-0.35 (0.62)	6 / 2 / 1 / 2	8 / 0 / 3	3 / 5 / 3 / 0
2	No	19 (20.0%)	0.10 (1.10)	39 / 19 / 13 / 24	38 / 34 / 23	31 / 41 / 23 / 0
	Yes	4 (17.4%)	-0.16 (0.91)	13 / 1 / 4 / 5	5 / 11 / 7	9 / 7 / 7 / 0
3	No	21 (18.9%)	-0.12 (0.84)	41 / 22 / 22 / 26	40 / 34 / 37	36 / 34 / 39 / 2
	Yes	6 (22.2%)	-0.20 (0.91)	11 / 6 / 3 / 7	11 / 4 / 12	11 / 6 / 10 / 0
4	No	11 (15.3%)	-0.16 (0.80)	28 / 16 / 10 / 18	16 / 26 / 30	26 / 28 / 18 / 0
	Yes	7 (13.7%)	-0.17 (0.81)	27 / 3 / 7 / 14	14 / 19 / 18	23 / 7 / 21 / 0
5	No	15 (23.8%)	-0.16 (0.81)	22 / 11 / 9 / 21	22 / 23 / 18	26 / 13 / 24 / 0
	Yes	5 (10.0%)	-0.36 (0.61)	18 / 11 / 7 / 14	16 / 15 / 19	21 / 12 / 17 / 0
6	No	26 (25.7%)	-0.20 (0.81)	36 / 24 / 16 / 26	33 / 41 / 28	18 / 46 / 36 / 2
	Yes	32 (22.1%)	-0.27 (0.74)	66 / 28 / 17 / 34	45 / 48 / 52	59 / 43 / 41 / 2

Table 3.1: Demographic characteristics broken down by severity of periodontal disease.
a. Malaria was diagnosed with a rapid diagnostic test obtained from a finger prick. *b.* Anemia was defined as a haemoglobin count Hb < 110 g/l. *c.* A proxy for socioeconomic status was created from principal components analysis by combining information on the building material of the house, main source of water and electricity, sanitary facilities, and main type of cooking fuel used. *d.* Women were enrolled at four sites: Lungwena / Malindi / Namwera / Mangochi. *e.* Women received one of three nutritional interventions: IFA / MMN / LNS. *f.* Supragingival samples were run on one of four sequencing runs on Illumina MiSeq.

smokers, and were given two courses of preventive malaria treatment with sulfadoxine-pyrimethamine (three tablets of 500 mg sulfadoxine and 25 mg pyrimethamine orally): one at enrolment and one between the 28th and 34th gestational week. After giving birth, 1229 women completed an oral health examination, consisting of a clinical examination and a panoramic X-ray of the jaws conducted by Ulla Harjunmaa or colleagues under her supervision. 1024 women had this examination within six weeks of delivery of a single infant (mothers of twins were excluded) and were included in further analysis. After excluding women without a supragingival sample ($n = 59$) and those with unknown HIV status ($n = 3$) our cross-sectional dataset included 962 women. Demographic characteristics are given in Table 3.1 – for more information on how these were collected, see Ashorn et al. (2015).

Classification of periodontal disease

Gingivitis was measured by the number of dental arch sextants with BoP out of six, with three sextants on each jaw (left, middle, and right). For periodontitis classification, each tooth was examined for evidence of deepened dental pockets both clinically and radiologically. A tooth was defined as having periodontitis if either a ≥ 4 mm pocket was measured in clinical examination or a vertical bony pocket was identified at least at the cervical root level radiologically. A woman was defined as having periodontitis

Total $n = 962$	Gingivitis (BoP score)						
	0	1	2	3	4	5	6
No periodontitis	137	72	95	111	72	63	102
Periodontitis	4	11	23	27	51	50	145

Table 3.2: Breakdown of all women by severity of periodontal disease.

if she had at least three teeth with periodontitis or at least one dental arch sextant with horizontal bone loss (at least at cervical level). The measurements forming the basis for periodontitis were on a per tooth basis, and involved converting a continuous variable into a binary one for each tooth, then converting a continuous count into a binary variable. The examination and classification methods are explained in more detail in Harjunmaa et al. (2015). The number of women with each possible combination of gingivitis and periodontitis defined in this way is given in Table 3.2 (for graphical representation see Figure 3.1).

While initially I wanted to use a continuous scale for both gingivitis and periodontitis, it was not clear what this scale should be for periodontitis. It was clear that a simple linear scale would be inappropriate for two reasons: firstly, the number of teeth with periodontitis had a long tail (Figure A.6), and secondly, there was no way of controlling for the total number of teeth already lost to periodontitis or otherwise. Therefore, a binary classification based on clinical standards used by dentists seemed most appropriate in order to investigate the features of the bacterial community that were associated with periodontitis at multiple levels of gingivitis.

Sample collection

Supragingival dental plaque samples were collected by swabbing the gingival margin of each tooth with a sterile plastic swab stick with a nylon fiber tip (microRheologics no. 552, Coban, Brescia, Italy). Only one swab was used per woman meaning that each sample represents a homogenized mixture of multiple locations in the mouth. After transfer in a cold box with ice packs to a laboratory, swabs were stored in cryovials at -20°C before being transferred to -80°C.

3.2.2 Sample extraction

DNA extraction and sequencing

The V5-V7 region of the 16S rRNA gene was amplified with 785F/1175R Illumina-compatible primers (see Section 3.2.3) (Bonder et al., 2012) to generate a sequencing library (R. M. Doyle et al., 2014). Each sample was amplified with dual indexes on the forward and reverse primer. All barcodes and adapter sequences used have been previously published (Caporaso et al., 2012). Each reaction was set up with 1 X MolzymPCRBuffer (Molzym), 200 μM of dNTPs (Bioline), 0.4 μM of forward and reverse primer with bar-

Criteria	Reads remaining
Maximum expected errors < 1	14,466,591
Minimum length (350 bases)	14,466,222
Maximum length (380 bases)	14,458,493
Samples <1,000 reads discarded	14,449,794

Table 3.3: Filtering of reads prior to MED analysis. The number of reads remaining at each point in the filtering pipeline.

code attached, 0.025 μM of Moltaq (Molzym), 5 μl of template DNA and PCR grade water (Bioline) to make a final reaction volume of 25 μl. Cycling parameters were as follows: 94°C x 3 min, 30 cycles of $-$94°C x 30 sec, $-$60°C x 40 sec, $-$72°C x 90 sec and one final extension at $-$72°C x 10 min.

Samples were purified and pooled into an equimolar solution using SequalPrep Normalization Plate Kit (Life Technologies) and further cleaned using AMPure XP beads (Beckman Coulter), both as per manufacturer's instructions. After quantification using the Qubit 2.0 (Life Technologies), the library was diluted and loaded into the MiSeq reagent cartridge at 10 pM. MiSeq runs were set to generated 250bp paired-end reads and two 12bp index reads for each sample. Reads were deposited in the European Nucleotide Archive under project accession PRJEB15035.

3.2.3 Analysis methods

Taxonomic classification

Sequenced reads were merged, demultiplexed, and quality filtered (minimum average Phred score > 25) using QIIME v1.8.0 (Caporaso et al., 2010). Closed-reference OTUs were picked at 98.5% similarity against HOMD v13.2 (T. Chen et al., 2010) using USEARCH v6.1.544 (Edgar, 2010) in QIIME v1.8.0 (Caporaso et al., 2010) with `parallel_pick_OTUs_usearch_61.py`. I used 98.5% sequence similarity because this is the threshold used to define taxa in HOMD, as it approximately corresponds to species level clusters for most oral bacteria (T. Chen et al., 2010). This approach identified 664 bacterial OTUs corresponding to 13,049,932 reads. The mean number of reads per sample was $13,565 \pm 6,833$.

Closed-reference OTU picking suffers from a number of issues, including sensitivity to the order of reference sequences when sequences are identical over the region considered (Westcott and Schloss, 2015). This is a particular problem when sequences are similar; as discussed earlier there exist oral bacteria that have >99% sequence similarity in given regions of the 16S rRNA gene but occupy separate oral habitats (Section 1.2.1). For this reason, I also performed MED on reads to define ecological units at higher resolution (Section 1.1.3).

After the merging of overlapping reads, the average sequence length was 369 bases.

I filtered sequences with an expected error greater than 1 using `fastq_filter` in VSEARCH v1.11.1 (Rognes et al., 2016). I then discarded all sequences shorter than 350 or longer than 380 bases, but performed no other quality filtering (e.g. length truncation) because MED assumes that length variation is biologically meaningful. Table 3.3 gives the sequences remaining at each stage of the filtering process. I ran MED v2.1 on 14,449,794 sequences. Because I wanted to be able to detect rare sequences, I set the minimum substantive abundance parameter (M) to 1444 (0.1% of the total number of reads) and the maximum variation allowed within a node (V) to 3. All other parameters were set to their default values. I assigned taxonomy to MED phylotypes using GAST (Huse et al., 2008) with VSEARCH v1.11.1 replacing the closed-source software USEARCH (Edgar, 2010) which has shown to be a satisfactory replacment (Westcott and Schloss, 2015).

Primer mismatch and its effect on phylotype detection

The protocol for amplification of the V5-V7 region of the 16S rRNA gene was chosen by Ronan Doyle from a comparison between primer sets conducted as part of his doctoral thesis (R. Doyle, 2016). However, the protocol was optimized for classifying larger numbers of OTUs across multiple body sites in order that samples from iLiNS-DYAD could be analyzed consistently, and was not designed specifically for oral periodontal pathogens. It is well established that different primer pairs can differentially amplify DNA from different taxa, biasing detection and subsequent analysis (Morales and Holben, 2009; Kumar et al., 2011; Cai et al., 2013). Therefore, one must always be cautious in interpreting marker gene data obtained using this approach: most importantly, absence of evidence is not the same as evidence of absence.

The standard 785F/1175R primer pair that was used for amplification has several degenerate positions, indicated here in **bold**:

785F: GGATTAGATACCC**BR**GTAGTC

1175R: ACGTC**R**TCCCCD**D**CCTTCCTC

The degeneracy symbols have the following meanings:

R → A or G

B → C, G, or T

D → A, G, or T

To identify phylotypes that one would *a priori* expect to be less efficiently amplified by the primers, I screened all possible versions of the primers (2 'R' options × 3 'B or D' options = 6 possibilities for each primer) against the HOMD v13.2 database (T. Chen et al., 2010) with blastn v2.2.31 (Camacho et al., 2009) to identify HOMD sequences that had mismatches with the primers. While most sequences had full-length matches at 100% similarity with possible primers for 785F and 1175R, there were eight and 51 HOMD sequences respectively that did not (Table A.1).

Of course, this *in silico* approach does not rule out differential amplification even for those sequences which have a full-length match to a primer pair, as other reasons for differential amplification exist (Edgar, 2017). But it conversely does provide a powerful prior that those phylotypes *without* a perfect primer pair match may well be absent (or detected at misleadlingly low levels) using this protocol, even if they were present in the original sample. In particular, the list of phylotypes with a mismatch to 1175R includes the well-established periodontal pathogens *Porphyromonas gingivalis* (human oral taxon (HOT) 619), *Tannerella forsythia* (HOT 601), and *Treponema denticola* (HOT 584) (Socransky et al., 1998). The absence of any of the taxa in this list from the dataset should not be interpreted as proof that they are not associated with periodontal disease in Malawian women.

3.2.4 Statistical analyses

(i) Diversity

I fitted a multivariate linear regression model to predict two measures of diversity (Table 1.1) – species richness (observed number of species) and Shannon index (a measure of richness and evenness) – using gingivitis, periodontitis, and the variables listed in Table 3.1 for 811 out of 962 samples with complete data and >5,000 reads. Richness and Shannon index were averaged over 100 iterations of rarefying to 5,000 reads per sample. I used backwards stepwise reduction using Akaike information criterion (AIC) (Akaike, 1974) to select the final model.

(ii) Differential abundances

I used negative binomial models to identify taxa that were significantly associated with disease using the DESeq2 package v1.6.3 (Love et al., 2014) as recommended by McMurdie and Holmes (2014) to avoid losing statistical information. DESeq2 fits generalized linear models to the count K_{ij} of taxa i in sample j using a negative binomial distribution NB with mean μ_{ij} and dispersion parameter α_i :

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i) \tag{3.1}$$

$$\log_2(q_{ij}) = x_j \beta_i \tag{3.2}$$

The coefficients β_i give the log2-fold changes for each taxa i for each column of the specified model matrix. The reason for the choice of a negative binomial model is that in ecology it is common for organisms to aggregate together, meaning that count data follows an overdispersed Poisson distribution instead of a standard Poisson distribution where the variance is equal to the mean. The parameter α_i governs the strength of this aggregation. Gingivitis was included as a continuous variable (BoP ranging from 0 to 6) and periodontitis as a binary factor. The model also contained terms controlling for potential confounders (study site, nutritional intervention, HIV status, and sequencing run). I corrected for multiple testing using the Benjamini-Hochberg procedure to control the false discovery rate at $q = 0.05$ (Benjamini and Hochberg, 1995). Full DESeq2 results for gingivitis and periodontitis are available for download from the associated publication as Data Set S7 (Shaw et al., 2016).

(iii) Co-occurrence networks

Beyond the identification of individual taxa associated with disease, co-occurrence analysis can allow the identification of important members of bacterial communities (Faust and Raes, 2012). To facilitate a higher resolution analysis of the network of periodontitis-associated bacteria, I selected all MED phylotypes that had representative sequences with >98.5% sequence similarity to periodontitis-associated HOMD OTUs. Calculating correlations from compositional data has long been known to be problematic because it can result in spurious correlations (Pearson, 1897). Fortunately, methods to deal with these problems have been developed; one such method specifically for marker gene data is SparCC, which uses log-ratio transformed abundances and pseudocounts to deal with zero values (Friedman and Alm, 2012). I calculated pairwise Spearman correlation coefficients between these MED phylotypes across samples using the SparCC procedure with default parameters (20 inference iterations and a correlation strength exclusion threshold of 0.1). To calculate pseudo-p values (two-sided t test), I shuffled the data sets for each group 100 times and repeated the procedure, removing correlations that were not significant ($p < 0.05$, no multiple testing correction). Networks of strong correlations, defined as being outside of the 95% confidence interval (CI) for the mean correlation between nodes (mean + 1.96 × standard deviation) were visualized as networks with qgraph v1.3.1 (Epskamp et al., 2012) using the Fruchterman-Reingold algorithm for node placement (Fruchterman and Reingold, 1991).

The betweenness centrality was originally defined for social networks, and gives a measure of how central a point is in terms of the flow of information across a graph (Freeman, 1977). For a given point p_k in a graph with n points, this can be defined by a sum over all other pairs of points ($i \neq j \neq k$). Let g_{ij} be the total number of geodesics (shortest paths) between i and j, and $g_{ij}(p_k)$ be the number of geodesics between i and j that pass through p_k. Then the normalized betweenness centrality $C_B^{'}(p_k)$ is defined by:

$$C_B^{'}(p_k) = \frac{2\sum_{i<j}^{n}\sum^{n}\frac{g_{ij}(p_k)}{g_{ij}}}{n^2 - 3n + 2} \tag{3.3}$$

(a) Gingivitis	Estimate (standard error)	Pr(>$\|t\|$)
(Intercept)	3.428 (0.616)	<0.001
Age (yrs)	0.056 (0.011)	<0.001
Malaria	-0.32 (0.163)	0.05
BMI	-0.046 (0.026)	0.077
Education	-0.097 (0.022)	<0.001
Socioeconomic status	-0.337 (0.081)	<0.001
(b) Periodontitis	Estimate (standard error)	Pr(>$\|z\|$)
(Intercept)	-4.714 (0.454)	<0.001
Age (yrs)	0.083 (0.014)	<0.001
Education (yrs)	-0.056 (0.028)	0.043
Socioeconomic status	-0.163 (0.11)	0.136
BoP	0.516 (0.047)	<0.001

Table 3.4: Final regression models to predict (a) gingivitis and (b) periodontitis. Full models started with all demographic variables in Table 3.1 and were then subject to stepwise reduction by AIC.

To understand this measure intuitively, consider the following example from Freeman (1977): a 'wheel' graph with a central point and spokes radiating out with points on the end. The central point will have $C_B'(p_k) = 1$ and all other points will have $C_B'(p_k) = 0$. Thus, the measure gives us a way of ranking points in a graph by how central they are. In the context of a correlation network of bacterial phylotypes believed to have some association (with disease) it identifies candidates for those taxa that are important members of that community.

3.3 Results

3.3.1 Demographics and initial modelling

Description of cohort

The cohort included 962 Malawian women with a mean age of 25.4 ± 6.2 years, of whom 140 (14.6%) had no periodontal disease, 822 (85.4%) had gingivitis (BoP $\geq$ 1), and 307 (32.0%) had periodontitis (Table 3.2). Gingivitis and periodontitis were significantly correlated (Spearman's $\rho = 0.44$) with the majority of women with periodontitis having high levels of gingivitis.

Demographic characteristics are predictive of periodontal disease

I fitted a linear regression model to predict gingivitis severity using selected demographic variables (Table 3.1) for 946 out of 962 women without any missing data. After backwards stepwise elimination of variables using AIC as a criterion for model selection (Akaike, 1974), the final model indicated that gingivitis was more severe in older women (OR 1.06 per year; 95% CI 1.03-1.08) with lower BMI (0.96; 0.91-1.00), fewer years

(a) Gingivitis	Estimate (standard error)	Pr(>$\|t\|$)
(Intercept)	2.885 (0.674)	<0.001
Age (yrs)	0.053 (0.011)	<0.001
BMI	-0.045 (0.027)	0.098
Education (yrs)	-0.093 (0.024)	<0.001
Socioeconomic status	-0.313 (0.088)	<0.001
Richness	0.011 (0.002)	<0.001
(b) Periodontitis	Estimate (standard error)	Pr(>$\|z\|$)
(Intercept)	-4.798 (0.494)	<0.001
BoP	0.495 (0.05)	<0.001
Age (yrs)	0.075 (0.015)	<0.001
Education (yrs)	-0.097 (0.028)	<0.001

Table 3.5: Final regression models to predict (a) gingivitis and (b) periodontitis. Full models started with all demographic variables in Table 3.1 and richness, and were then subject to stepwise reduction by AIC.

of education (0.91 per year; 0.87-0.95), a lower socioeconomic status (0.71; 0.61-0.84), and no malaria (0.73; 0.53-1.00) (Table 3.4a). HIV was not included in the best model, in agreement with previous research that found no association with periodontal disease (John et al., 2013; Khammissa et al., 2012). I also applied the same procedure to predict (binary) periodontitis using a logistic regression model that included gingivitis severity. The final model showed that periodontitis was more likely in women with more severe gingivitis (OR 1.68 per BoP; 1.53-1.84) who were older (1.09 per year; 1.06-1.12), had fewer years of education (0.95 per year; 0.90-1.00) and a lower socioeconomic status (0.85; 0.68-1.05) (Table 3.4b).

I wanted to see if adding information on the diversity of supragingival plaque bacterial communities improved the models, so I also added in the calculated richness to the full model to predict gingivitis and periodontitis for 811/962 women with >5,000 reads and no missing data, then again performed stepwise reduction according to AIC. Evenness of microbial communities was not included in the model due to high collinearity with richness (Spearman's $\rho = 0.88$). Richness was retained in the final model for gingivitis (Table 3.5a) but not periodontitis (Table 3.5b).

In summary, periodontal disease was more common in women who were older, had lower socioeconomic status, and fewer years of education, in line with previous research highlighting it as a sociopolitical public health problem (Batchelor, 2014). My preliminary analysis confirmed that the supragingival plaque community contained some information on periodontal disease after taking these factors into account, with richness retained in a final model for predicting gingivitis severity but not periodontitis. This suggested that a more sophisticated analysis than just considering community richness was required to identify signals unique to periodontitis.

Plaque richness and diversity are higher in more severe gingivitis and periodontitis

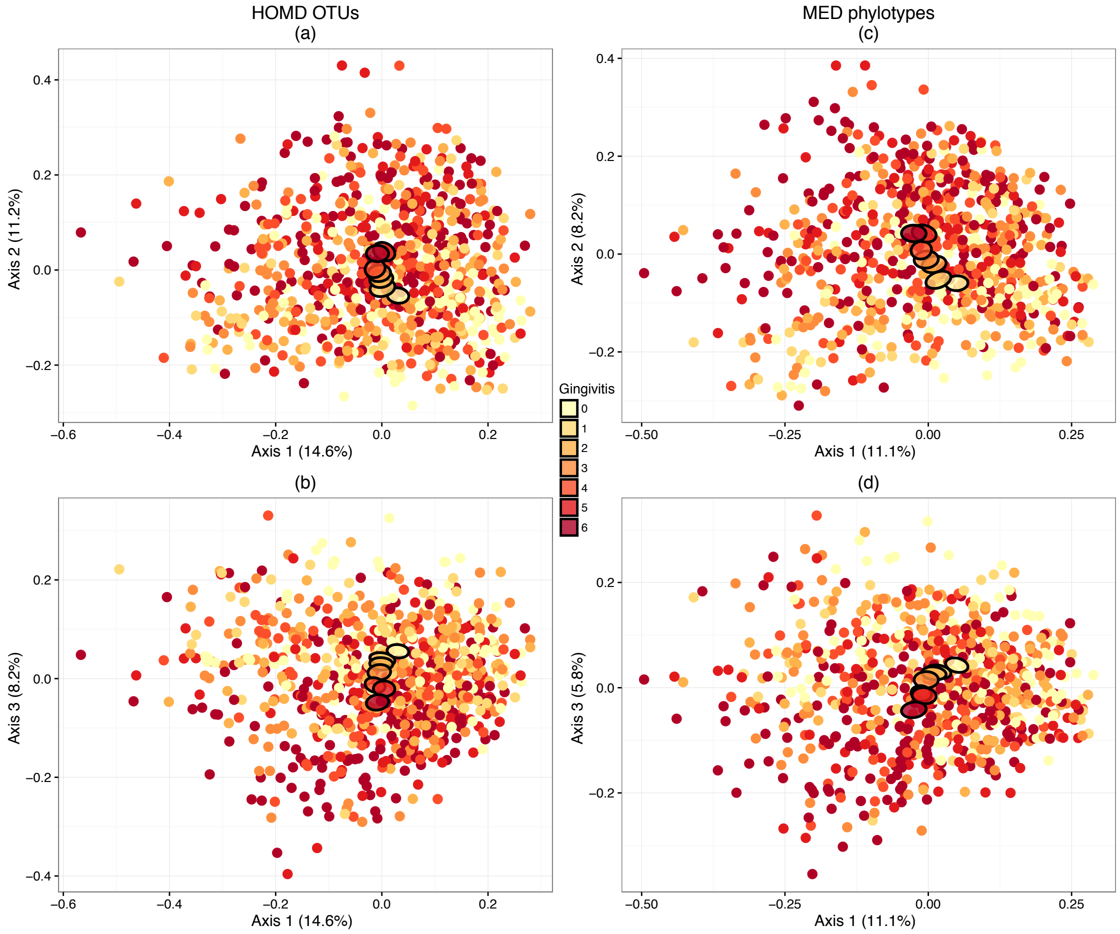


Figure 3.2: The principal coordinates analysis (PCoA) ordination of supragingival plaque samples shows an approximate trend with gingivitis severity that is robust to analysis methods. PCoA ordinations based on Bray-Curtis dissimilarities between samples for 626 HOMD OTUs (a, b) and 502 MED phylotypes (c, d). Filled ellipses show mean values for each gingivitis severity, ranging from 0 (yellow) to 6 (dark red). In both cases, an approximate trend is visible in the mean ellipses for each group, despite the noisiness of the data set. Before plotting, samples were rarefied to 5,000 reads to minimize the impact of sequencing depth.

Initial exploratory analysis with PCoA ordinations showed that although there was large variability in community composition across supragingival plaque samples, there was also a clear trend related to gingivitis severity that was robust to the analysis method used: HOMD OTUs or MED phylotypes (Figure 3.2). Stratifying by periodontitis in the same way did not indicate visually clear differences.

Quantitative analysis of diversity reflected this trend. Gingivitis was associated with higher microbial community richness (Figure 3.3a) and Shannon index (Figure 3.3b). Bacterial communities did not markedly differ between healthy women and those with low levels of gingivitis. Both gingivitis and periodontitis were associated with higher supragingival plaque richness in a linear regression controlling for demographic variables (Table A.2a). Interestingly, the site at which samples were collected was also associated with richness, perhaps because of local environmental effects similar to those I found

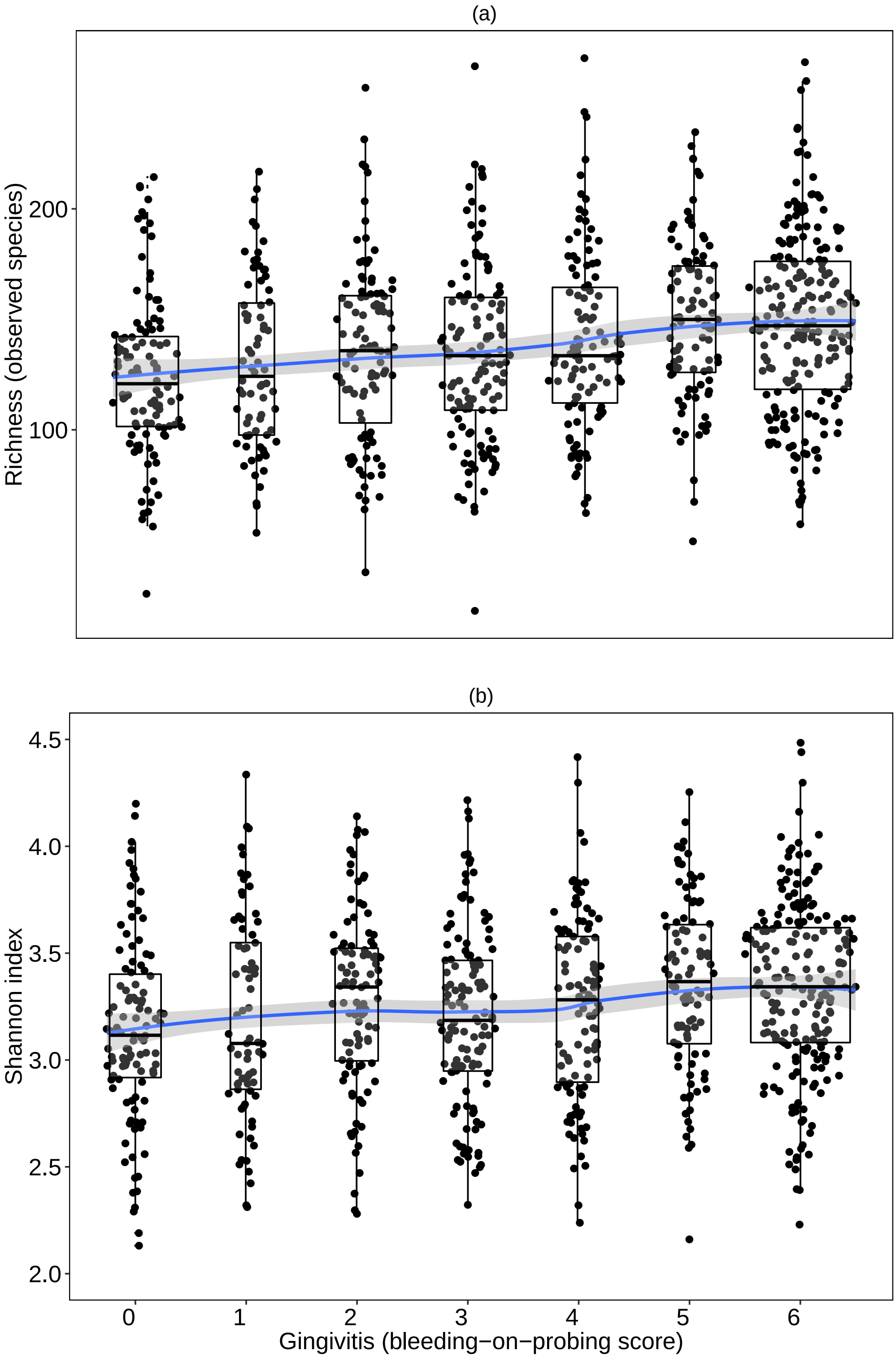


Figure 3.3: Both (a) richness and (b) Shannon index of the supragingival microbiome increase with gingivitis severity. Estimates for each sample were calculated by sampling with replacement at a rarefaction depth of 5,000 sequences per sample and averaging over 100 iterations. The fitted line shows a local polynomial regression fit calculated using `loess` in R, with the grey region indicating the 95% CI. 138 out of 962 samples were excluded due to having fewer than 5,000 sequences. Changing the rarefaction depth did not affect the conclusion that gingivitis severity was associated with an increase in both measures of diversity.

in the Ashkenazi cohort (Chapter 2). In the final model predicting Shannon index, periodontitis was not retained although gingivitis was (Table A.2b), in line with the earlier reversed analysis using demographic characteristics to predict disease, where richness was retained in the final model for gingivitis but not for periodontitis (Table 3.5).

3.3.2 Bacterial taxa associated with periodontal disease

Differences in bacterial abundances with gingivitis

Differential abundance analysis with DESeq2 (Section 3.2.4) identified 118 OTUs that were significantly ($q < 0.05$) associated with greater severity of gingivitis, making up 16.6% of the dataset in terms of reads. Conversely, 47 OTUs were associated with lower severity (18.7% of the dataset). Figures 3.4a and 3.4b show the cumulative abundances of health- and gingivitis-associated OTUs respectively, showing the progressive nature of changes with the amount of bleeding. Most of the pairwise comparisons of summed abundances of health- and gingivitis-associated OTUs were not significantly different between women with and without periodontitis (Kruskal-Wallis test, $p > 0.05$). However, for women with periodontitis, severity of gingivitis was important, as there were microbial differences between women with and without periodontitis for both moderate gingivitis (BoP of 3; $p = 0.014$) and severe gingivitis (BoP of 6; $p = 0.011$).The most significantly gingivitis-associated OTU was *Peptostreptococcus stomatis*, which was present in over 75% of samples across severity categories and was an average of 1.45-fold more abundant (95% CI 1.37-1.54) with a unit increase in BoP.

Differences in bacterial abundances with periodontitis

While gingivitis had a stronger association with supragingival microbiota, there were also differences in microbial community composition with periodontitis (Figures 3.4c and 3.4d), with 71 OTUs significantly ($q < 0.05$) more abundant in women with periodontitis, making up 4.4% of the dataset in terms of reads. A smaller number of OTUs were significantly more abundant in the absence of periodontitis (13 OTUs), making up 3.6% of the dataset by reads. These health-associated OTUs were *Lautropia mirabilis*, *Rothia aeria*, *Streptococcus pyogenes*, *Streptococcus mutans* and seven members of *Actinomyces*.

At the genus level for periodontitis-associated OTUs, *Prevotella* (14 OTUs) and *Treponema* (10 OTUs) were the most represented. Only one member of the pathogenic red complex (Socransky et al., 1998) was significantly associated with periodontitis: *T. denticola*. The other two members (*P. gingivalis* and *T. forsythia*) were additionally not identified as MED phylotypes in the dataset, probably due to primer mismatch (see Section 3.2.3). *Eubacterium nodatum*, previously identified as clustering with the red complex in supragingival plaque (Haffajee et al., 2008), was significantly associated with periodontitis.

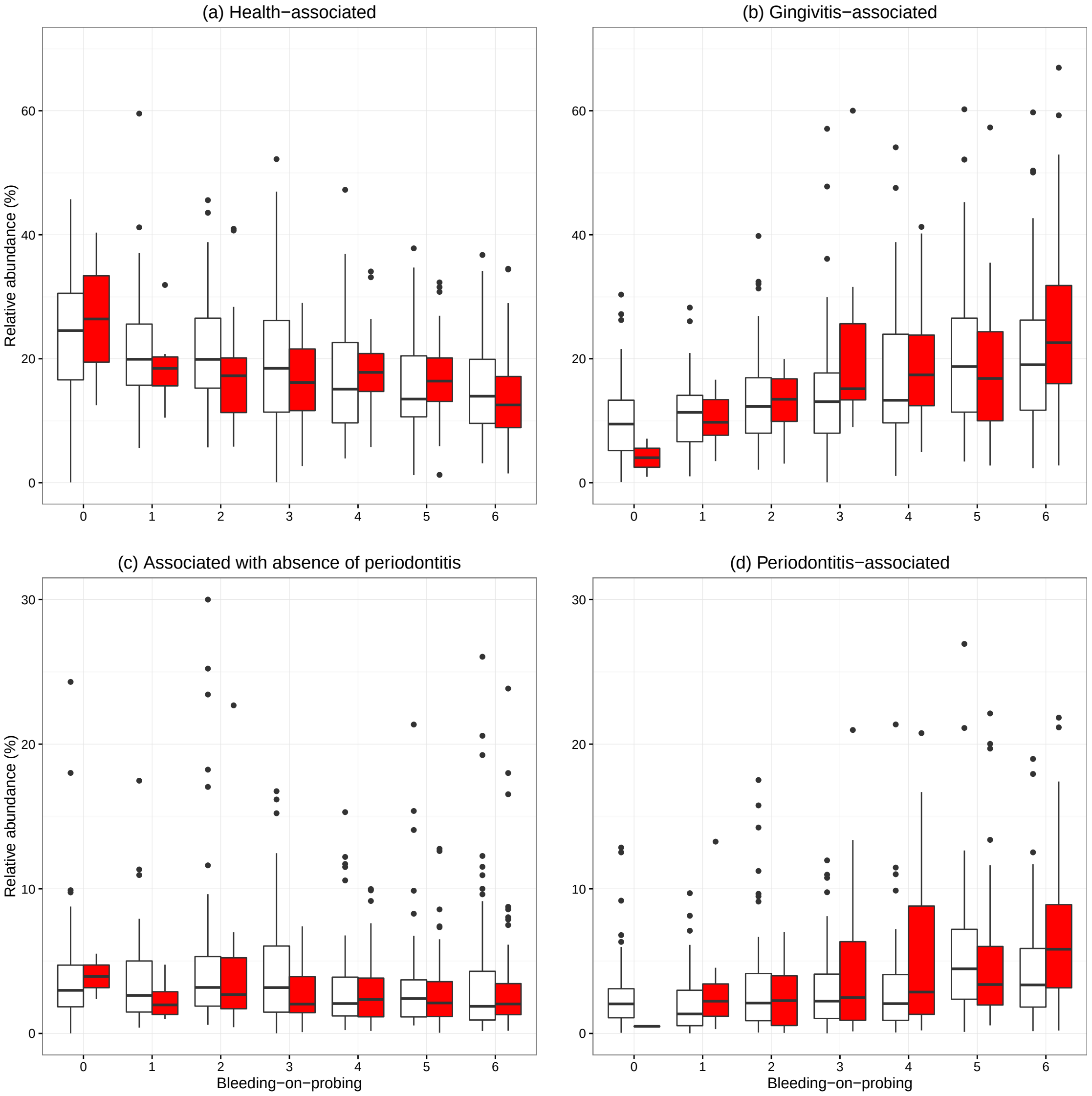


Figure 3.4: Summed percentage abundances of OTUs associated with (a) decreased gingivitis, (b) increased gingivitis, (c) absence of periodontitis, and (d) presence of periodontitis for each periodontal disease category. Colours indicate groups without (white) and with (red) periodontitis. For plotting purposes, samples were rarefied to 10,000 reads per sample, resulting in the removal of 269 out of 962 samples; this rarefaction was not used in the selection of the OTUs, which was performed using DESeq2 on the whole dataset (Section 3.2.4). One outlier and two outliers in (c) and (d) respectively are not shown due to trimming the y-axis at a relative abundance of 30%.

Differences in bacterial abundances unique to periodontitis

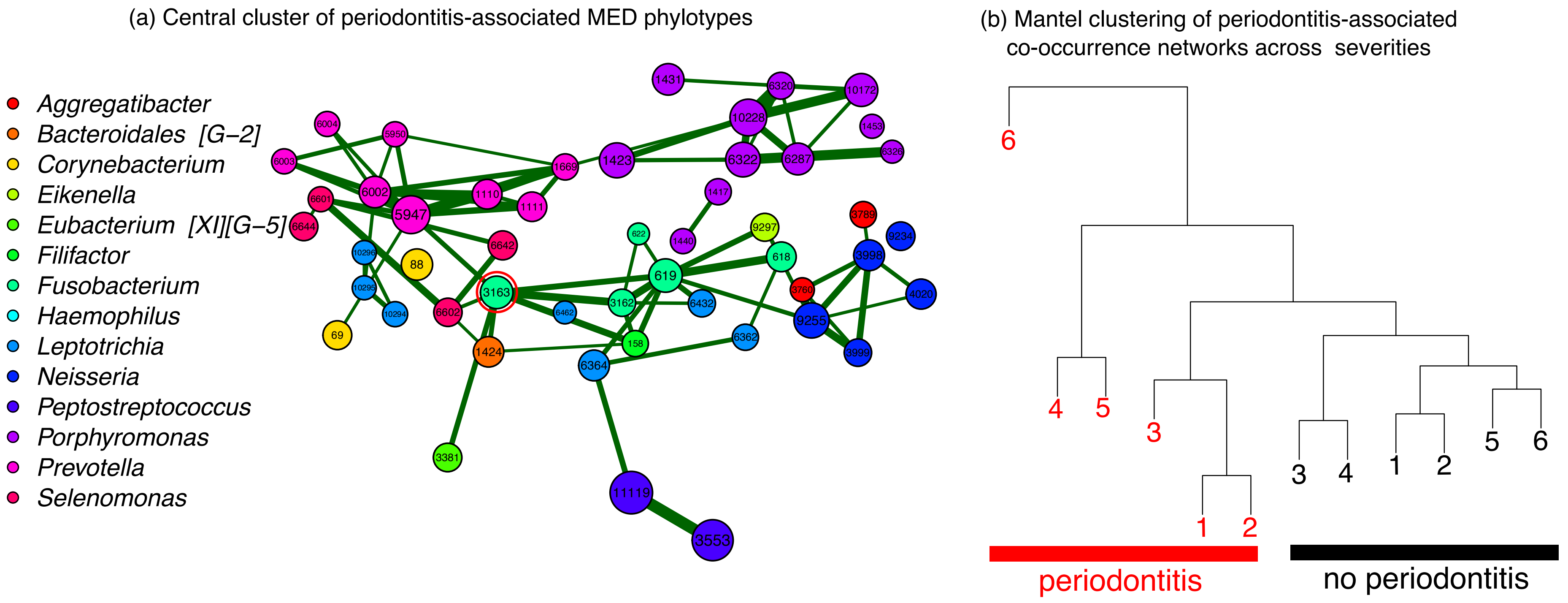
Forty out of seventy-one periodontitis-associated OTUs (56%) were not associated with gingivitis. These taxa were rare: their mean cumulative abundance was 2.2%, with only six OTUs having mean relative abundances $> 0.1\%$. The most represented genera were *Prevotella* (nine OTUs), *Treponema* (five OTUs) and *Selenomonas* (four OTUs). The presence or absence of periodontitis was not a significant determinant of cumulative abundances of these OTUs for women with the same levels of gingivitis (Kruskal-Wallis test, $p > 0.05$), except for women with a BoP of 4 ($p = 0.026$).

3.3.3 The co-occurrence network of periodontitis-associated taxa

The above analysis treats each OTU as independent but in reality oral bacteria exist in complex multi-species biofilms where interactions are extremely important (Teles et al., 2013). I therefore analyzed the co-occurence networks of periodontitis-associated bacteria across all periodontal severities.

In a preliminary analysis, co-occurrence network analysis using HOMD OTUs associated with periodontitis showed more connections in the network in women with periodontitis across gingivitis severities (Figure A.7). However, I wanted to verify this result with MED analysis to ensure co-occurrence patterns were not due to the limited resolution of the OTU picking process, which used a closed algorithm against HOMD with a 98.5% sequence similarity cutoff. Therefore, I selected all 81 MED phylotypes with >98.5% sequence similarity to a periodontitis-associated HOMD OTU. This selection included 19 members of *Streptococcus*, despite the fact that only *Streptococcus oligofermentans* (HOT 886) was associated with periodontitis, due to the high sequence similarity of this genus in the V5-V7 region. When plotted as a co-occurrence network, these phylotypes clearly clustered away from the periodontitis-associated phylotypes and had negative correlations with the rest of the network. I therefore removed them when preparing Figure 3.5.

The strongly-connected co-occurrence network in women with severe gingivitis (BoP of 6) and periodontitis showed several genus-level clusters, including *Selenomonas*, *Peptostreptococcus*, and *Prevotella* (Figure 3.5a). Notably, these clusters were connected by a small group of central bacteria including *Filifactor alocis* (phylotype 158) and several members of *Fusobacterium nucleatum* with phylotypes classified taxonomically as subspecies *vincentii* (phylotypes 3163 and 622) and *polymorphum* (phylotypes 618 and 619), suggesting their roles in co-aggregation of periodontal biofilms. Ranking phylotypes in the strongly-connected network according to their betweenness centrality, which measures potential for influence on information transfer in a network (Freeman, 1977), the most connected phylotype was *F. nucleatum* subsp. *vincentii* (phylotype 3163) (Table 3.6). *T. denticola* was not present in this network, but when MED analysis was repeated with the minimum substantive abundance parameter reduced by a factor of 10 to 0.01% it was placed in the network in a central position.



(a) Central cluster of periodontitis-associated MED phylotypes

(b) Mantel clustering of periodontitis-associated co-occurrence networks across severities

Figure 3.5: The co-occurrence network of periodontitis-associated bacteria shows a distinct community structure with the presence of periodontitis across gingivitis severities. **(a)** The strongly connected central co-occurrence network of periodontitis-associated bacteria across supragingival plaque samples from n=110 women with severe gingivitis (BoP=6) and periodontitis. Shown here are significant strong pairwise Spearman correlation coefficients ($p < 0.01$, $\rho > 0.405$) calculated with SparCC between MED phylotypes with >98.5% similarity to periodontitis-associated HOMD OTUs. Node color indicates taxonomic genus, size is proportional to log-transformed mean relative abundance, and edge weight indicates the strength of the correlation. The red circle indicates the node with the highest betweenness centrality, classified taxonomically as *F. nucleatum ss. vincentii* (Table 3.6). Node layout was determined using the Fruchterman-Reingold algorithm in qgraph v1.3.1. 22 nodes without any strong correlations connecting them to the rest of the network (i.e. no edges with $\rho > 0.405$) were removed during figure preparation. **(b)** Clustering using `hclust` of the correlation matrices calculated in this way for all severities of periodontal disease. The periodontitis-associated co-occurrence network is more similar between women with periodontitis regardless of gingivitis severity. Correlation matrices were not adjusted for significance due to the different numbers of women between groups.

Phylotype	Taxonomic classification	Centrality
3163	*Fusobacterium nucleatum ss. vincentii* (HOT 200)	0.519
5947	*Prevotella melaninogenica* (HOT 469)	0.501
619	*Fusobacterium nucleatum ss. polymorphum* (HOT 202)	0.445
1669	*Prevotella melaninogenica* (HOT 469)	0.288
10228	*Porphyromonas sp.* (HOT 284)	0.259
9255	*Neisseria flava* (HOT 609)	0.196
6002	*Prevotella veroralis* (HOT 572)	0.159
6364	*Leptotrichia sp.* (HOT 215)	0.125
10295	*Leptotrichia wadei* (HOT 222)	0.085
6601	*Selenomonas sputigena* (HOT 151)	0.063
3998	*Neisseria subflava* (HOT 476)	0.045
6320	*Porphyromonas sp.* (HOT 284)	0.045
6322	*Porphyromonas sp.* (HOT 275)	0.043
11119	*Peptostreptococcus stomatis* (HOT 112)	0.043
3162	*Fusobacterium sp.* (HOT 205)	0.031
6602	*Selenomonas sputigena* (HOT 151)	0.021
6287	*Porphyromonas sp.* (HOT 275)	0.003
6362	*Leptotrichia sp.* (HOT 215)	0.003

Table 3.6: Ranking of MED phylotypes in strongly-connected co-occurrence network by their normalized betweenness centrality score. Betweenness centrality scores for the network of significant strong correlations (Figure 3.5a) were calculated using the betweenness function in igraph v1.1.2 (Csardi and Nepusz, 2006). Only non-zero scores are shown. Phylotype ID is assigned randomly by MED.

To confirm that this altered community structure was a distinguishing feature of supragingival plaque between women with and without periodontitis, I clustered the matrices of SparCC correlations based on Mantel distances for each category of periodontal disease (Figure 3.5b). Networks clustered by the periodontitis status of the women in the group, confirming that the altered community structure with periodontitis was detectable even in women with low levels of gingivitis. Within the periodontitis groupings, matrices clustered by gingivitis severity.

3.4 Discussion

3.4.1 Conclusions

In this chapter I have investigated changes in the supragingival microbiome associated with periodontal disease severity in a large cross-sectional cohort in Malawi, which represents the largest study of its kind to date. My main finding was that even though the composition of supragingival plaque is primarily associated with gingivitis (as quantified by bleeding-on-probing) rather than the presence or absence of periodontitis, the presence of periodontitis does have detectable associations with the supragingival microbiota that distinguish it from gingivitis. In particular the differences in co-occurrence patterns of taxa between women with and without periodontitis support a more complex etiology of disease than a simple progression from health through gingivitis to periodontitis.

Gingivitis and periodontitis were both associated with higher bacterial community richness and Shannon index, and this association remained after adjustment for demographic factors including age, BMI, socioeconomic status, and local environment. This finding is consistent with previous research (Kistler et al., 2013; H. Chen et al., 2015), with higher diversity meaning that in periodontal disease the oral microbiota expands in membership rather than existing taxa undergoing replacement. This could correspond to primary ecological succession in a new environmental niche, as suggested by Abusleme et al. (2013).

Gingivitis and periodontitis are clearly linked; I found that many taxa were associated with both. The abundance of the majority of these taxa increased with gingivitis severity, and this pattern was not influenced by the presence of periodontitis. Furthermore, some women with no signs of gingivitis had similar summed percentage abundances of disease-associated taxa to women with severe gingivitis. It would appear that relative bacterial abundances alone are insufficient to explain the presence of disease, consistent with a requirement for other factors such as the host inflammatory response to cause disease.

Periodontitis-associated OTUs were also identified including known periodontal pathogens like *F. alocis* , *T. denticola*, *F. nucleatum*, and *P. stomatis*, consistent with findings from other populations (Teles et al., 2013). OTUs including members of *Prevotella*, *Treponema*, and *Selenomonas* were not significantly associated with gingivitis severity, supporting the idea that periodontitis is not just an advanced phase of gingivitis and involves additional bacteria. However, cumulative abundances of periodontitis-associated OTUs did not differ significantly between women with and without periodontitis who had the same levels of gingivitis, suggesting that abundances do not fully explain the disease.

The co-occurrence analysis of periodontitis-associated taxa found different co-occurrence patterns across disease categories, indicating the presence of a consistent community structure in women with periodontitis across all gingivitis severities. Central nodes in this periodontitis-associated network included *F. alocis* and several subspecies of *F. nucleatum*, which acted as hubs connecting different clusters (Figure 3.5a). Network analysis using betweenness centrality ranked *F. nucleatum* subsp. *vincentii* (phylotype 3163) as the most central phylotype in the strongly-connected co-occurrence network in women with severe gingivitis and periodontitis (Table 3.6). These findings are consistent with its proposed role as one of the ‘bridging bacteria’ which contribute to the co-aggregation of periodontal biofilms (Aruni et al., 2015). *F. nucleatum* has been shown experimentally to “facilitate the survival of obligate anaerobes in aerated environments” (Bradshaw et al., 1998), and has been identified as one of the important precursors for the growth of biofilms *in vitro* (Foster and Kolenbrander, 2004). Of relevance for this cohort, strains of *F. nucleatum* present in periodontal pockets have also been identified in isolates from amniotic fluid swallowed by babies born preterm, suggesting it may also be involved with pregnancy complications (Gonzales-Marin et al., 2013). However, a further study on this cohort that I was involved in found no direct association between oral bacteria and preterm birth, suggesting that oral infection may primarily affect birth outcomes via systemic inflammation (Harjunmaa et al., 2018). The other central species *F. alocis* has also been experimentally linked to the co-aggregation of periodontal biofilms (Schlafer et al., 2010; Fine et al., 2013) and correlates with greater inflammation in periodontitis (Camelo-Castillo et al., 2015). H. Chen et al. (2015) also identified a similar *F. alocis*-centered co-occurrence group of taxa that was enriched in multiple oral habitats during periodontitis compared with healthy controls.

3.4.2 Limitations

A great strength of this study was the inclusion of large numbers of women with different severities and combinations of periodontal disease, thanks to the access to the iLiNS-DYAD-M cohort. However, due to the difficulty of collecting such a large number of samples from a cohort in a resource-limited setting only supragingival plaque was sampled, meaning that my findings about periodontitis only apply to supragingival plaque. Previous work has shown that sampling supragingival plaque still allows the detection of bacteria associated with periodontitis while being minimally invasive and simple to perform (Galimanas et al., 2014). Similarly, I found changes in abundances of rare taxa known to be associated with the subgingival plaque of periodontitis. For example, *Fretibacterium fastidiosum* (HOMD ID: 360BH017) which accounted for a mean of just 0.009% of reads was still significantly more abundant (2.5-fold) in women with periodontitis, consistent with a recent finding of a higher abundance in subgingival plaque when periodontitis was compared to gingivitis (Park et al., 2015).

Another limitation was that samples were collected from across the mouth instead of localizing sampling to sites of specific interest. The distribution of bacterial species across the mouth is known to be heterogeneous, with supragingival plaque at sites adjacent to deepened periodontal pockets showing significantly higher counts of periodontitis-associated species (Haffajee et al., 2008). Again, due to the size of the cohort sampling used a single swab, which was probably at least partially responsible for the large amount of variability in our dataset when visualized in ordinations (Figure 3.2), and effectively pooled all supragingival sites. This precluded any investigation of heterogeneity between sites, but detectable associations with both gingivitis and periodontitis were still present even with this approach.

I treated gingivitis as a continuous variable but periodontitis as binary. In reality periodontitis is a complex disease with a problematic classification (Mdala et al., 2014), and it is likely that my simple treatment of periodontitis obscures this complexity. This could cause bacterial co-occurrence patterns in women with periodontitis to appear stronger, as women with more severe disease may have greater abundances of associated bacterial species. The effect of host genetics was also not investigated. It is known that a hyper-responsive immune phenotype can affect the risk of certain forms of periodontitis

(Shaddox et al., 2010) and this has been linked to the oral microbiome (Fine et al., 2013). A simultaneous analysis of genetic differences and microbiome differences would be interesting to unpick possible causation.

This study is the largest to be conducted so far in a sub-Saharan population and the results appear consistent for the most part with previous work on bacterial associations with periodontal disease (Ximénez-Fyvie, Haffajee, and Socransky, 2000; Teles et al., 2013; Haffajee et al., 2008; H. Chen et al., 2015; Haffajee and Socransky, 1994). However, it should be pointed out that the population under study was additionally notable in two respects. Firstly, all participants were women who had recently given birth. Pregnancy, particularly in its early to mid stages, is known to be linked to periodontal disease and potential changes in the oral microbiome (Fujiwara et al., 2015). Pregnant women have an increased susceptibility to gingivitis (Gürsoy et al., 2010) although subgingival levels of known periodontal pathogens may remain unchanged (Adriaens et al., 2009). Qualitative differences between periodontal pathogens found during pregnancy and postpartum have also been observed (Carrillo-de-Albornoz et al., 2010). It is not clear for how long after pregnancy the oral microbiome remains altered, but evidence that significant changes are mainly detectable in early pregnancy (Fujiwara et al., 2015) and the consistency of my results with other studies suggests that effects remaining after six weeks postpartum are small. Secondly, all women in the study were intermittently given sulfadoxine-pyrimethamine (SP) at enrolment and between the 28th and 34th gestational week for malaria prevention. Since systemic antibiotics can be given as a treatment for aggressive periodontitis (Rabelo et al., 2015), patients who have received antibiotic treatment in the previous 6 months are often excluded from studies of periodontitis. However, the salivary microbiome has been shown to be robust to disturbance by a week-long course of antibiotics (Zaura et al., 2015). Given that SP treatment was intermittent, involved antibiotics not targeted at periodontal bacteria, and took place around two months before the oral sampling, I believe that it is unlikely to have played an important role, but have no direct evidence to support this claim.

3.4.3 Summary

I have conducted the largest study to date investigating associations between supragingival plaque composition and varying severities of periodontal disease, in a low-income sub-Saharan population with limited oral hygiene. I modelled periodontal disease using two variables, allowing the identification of distinct signals associated with gingivitis and periodontitis in supragingival plaque. Network analysis of observed co-occurrence patterns and network analysis was consistent with the role of bridging bacteria like *F. nucleatum* and *F. alocis* in the co-aggregation of periodontal biofilms prior to penetrance into subgingival regions. Although some periodontitis-associated bacteria were also associated with gingivitis, the major change with periodontitis is in the network of co-occurrences. Viewed this way, gingivitis could set the stage for periodontitis to develop by providing an environment where periodontitis-associated taxa can increase in abundance and co-aggregate into pathogenic biofilms that may then penetrate to subgingival regions. More quantitative modelling of associations between oral bacteria and various clinical features of disease will be necessary to understand these complex relationships and explore the microbial ecology of periodontitis.

Chapter 4

A perturbation model of the gut microbiome's response to antibiotics

Declaration of contributions

I conceived the model, performed all analysis, and wrote the associated paper with feedback from all co-authors.

Publication

This work is available on *bioRxiv* as Shaw, Barnes, et al. (2017):

L.P. Shaw, C. P. Barnes, A. S. Walker, N. Klein, and F. Balloux (2017). A perturbation model of the gut microbiomes response to antibiotics. *bioRxiv*, 222398. doi: 10.1101/222398.

4.1 Introduction

The human gut microbiome is a complex ecosystem and, as such, can be thought of in ecological terms. The relative stability of the gut microbiome in the absence of large perturbations has been suggested to indicate the presence of restoring forces within a dynamical system (Relman, 2012). While stability appears to be the norm, disturbances to this ecosystem are also important when considering the impact of the gut microbiome on human health. One example of a major perturbation is a course of antibiotics, which typically leads to a marked reduction in species diversity before subsequent recovery (Modi et al., 2014). Even a brief course of antibiotics can result in long-term effects on microbial community composition, with species diversity remaining lower than its baseline value up to a year afterwards (Zaura et al., 2015). However, the nature of the reconstitution of the gut microbiome remains an active area of research.

Artificial perturbation experiments are widely used to explore the underlying dynamics of macro-ecological systems (Wootton, 2010). In the context of the gut microbiome, the response after antibiotics has been extensively investigated (Section 1.3.1). However, despite interest in the application of ecological theory to the gut microbiome (Pepper and Rosenfeld, 2012) it is still challenging to develop and fit quantitative models of this time response for the whole community due to the large number of species involved. Furthermore, while responses can appear individualized (Dethlefsen and Relman, 2011), this does not preclude the possibility of generalized models that are applicable at the population level. Additionally, recent work suggests that alterations due to specific antibiotics are predictable and reproducible (Raymond, Ouameur, et al., 2016).

Applying mathematical models to other ecological systems subject to perturbation has a long tradition of giving useful insight into their behaviour (May, 1973; Scheffer et al., 2001; Skellam, 1951). Crucially, it allows the comparison of different models based on different hypotheses about the subsequent behaviour of the system. Developing a consistent mathematical framework for quantifying the long-term effects of antibiotic use would facilitate comparisons between different antibiotics and different regimens, with the potential to inform antibiotic stewardship (Doron and Davidson, 2011). Some previous work has attempted to model species interactions in the context of antibiotics using Lotka-Volterra models (Stein et al., 2013), but such models require dense temporal sampling and restriction to a small number of species to make meaningful inference, limiting their applicability to broader ecological questions. Furthermore, it has recently been shown that pairwise microbial interactions in different scenarios cannot be captured by a single equation, suggesting that pairwise modelling will often fail to predict microbial dynamics (Momeni et al., 2017).

Holling (1973) introduced a now widely-used conceptual framework for thinking about ecosystem resilience, where ecosystems exist within a stability landscape analogous to a fitness landscape in evolutionary biology. This classical ecology framework has been used in several articles to visualise the state of the gut microbiome as a ball sit-

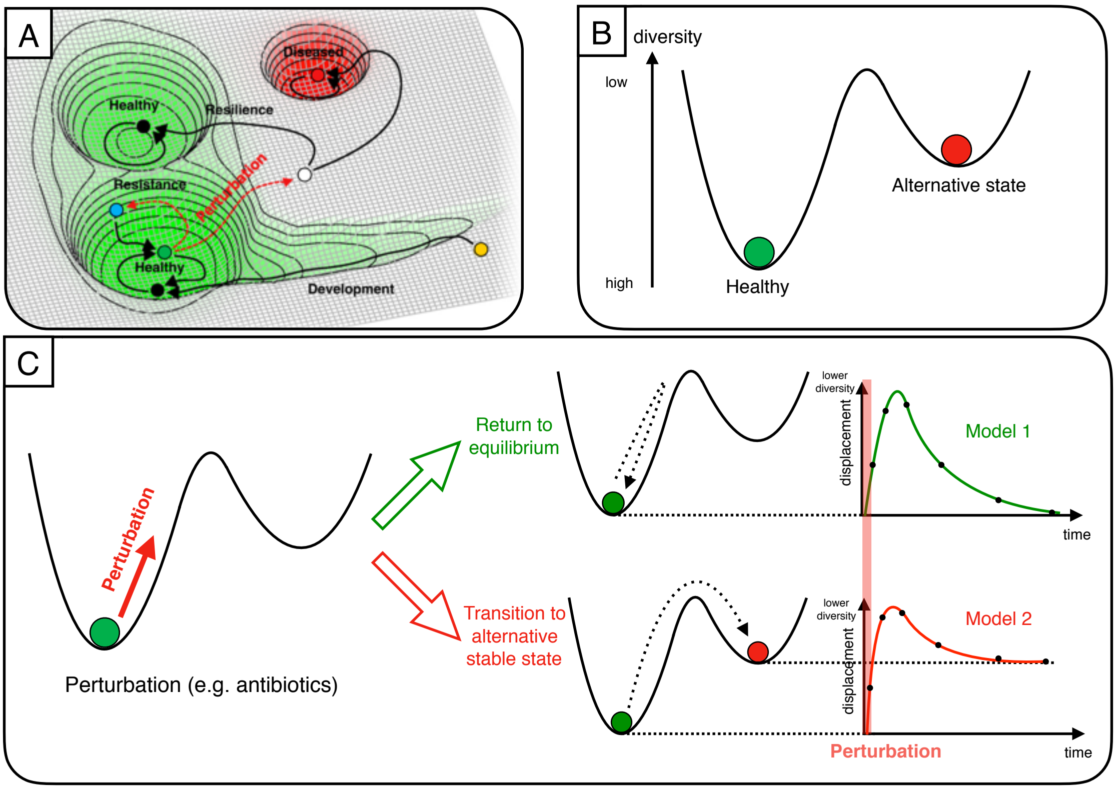


Figure 4.1: An impulse response model of antibiotic perturbation to the gut microbiome. The gut microbiome is represented as a unit mass on a stability landscape, where height corresponds to phylogenetic diversity. (A) The healthy human microbiome can be conceptualized as resting in the equilibrium of a stability landscape of all possible states of the microbiome. Perturbations can displace it from this equilibrium value into alternative states (adapted from Lloyd-Price et al. (2016)). (B) Choosing to parameterize this stability landscape using diversity, I assume that there are just two states: the healthy baseline state and an alternative stable state. (C) Perturbation to the microbiome (e.g. by antibiotics) is then modelled as an impulse, which assumes the duration of the perturbation is short relative to the overall timescale of the experiment. I consider the form of the diversity time-response under two scenarios: a return to the baseline diversity; and a transition to a different value of a diversity (i.e. an alternative stable state).

ting in a simple landscape (Lemon et al., 2012; Relman, 2012; Lloyd-Price et al., 2016). Perturbations can be thought of either as forces acting on the ball to displace it from its equilibrium position (Lloyd-Price et al., 2016), or alterations of the stability landscape (Costello et al., 2012). While this image is usually provided only as a conceptual model to aid thinking about the complexity of the ecosystem, I use it to derive a mathematical model to investigate whether it could provide mechanistic insight.

The model I outline here, based on simple ecological concepts, allows quantitative hypotheses about the effect of antibiotics on the gut microbiome to be tested. I model the effect of a brief course of antibiotics on the microbial community's phylogenetic diversity as the impulse response of an overdamped harmonic oscillator (Figure 4.1), and compare parameters for two widely-used antibiotics by fitting to empirical data previously published by Zaura et al. (2015). I find that a variant of the model with an extra parameter accounting for the possibility of an altered equilibrium value of diversity is better supported, providing evidence from a sparse dataset that antibiotics can produce transitions to alternative stable states.

4.2 Materials and methods

4.2.1 Ecological assumptions

I represent the state of the gut microbiome as a unit mass resting in a stability landscape (Figure 4.1A). Choosing to mathematically model the state of the gut microbiome in this way also requires choosing a mathematical representation with reference to an equilibrium value. While earlier studies sought to identify a core set of 'healthy' microbes, the disturbance of which would indicate displacement from equilibrium, it has become apparent that this is not a practical definition due to high inter-individual variability in taxonomic composition (Lloyd-Price et al., 2016). More recent concepts of a healthy 'functional core' appear more promising, but characterization is challenging, particularly as many gut microbiome studies use 16S rRNA marker gene sequencing rather than whole-genome shotgun sequencing.

Therefore, I choose to use a metric that offers a proxy for the general functional potential of the gut microbiome: phylogenetic diversity (Lloyd-Price et al., 2016). Higher diversity has previously been associated with health (Turnbaugh et al., 2007) and temporal stability (Flores et al., 2014). For these reasons, I assume the equilibrium position to have higher diversity than the points immediately surrounding it, forming a potential well (Figure 1B). However, there may be alternative stable states that represent possible dysbiotic' states (Figure 4.1B), which are of interest when considering the effect of perturbations (Figure 4.1C).

4.2.2 The model

I treat the local stability landscape as a harmonic potential, with a 'restoring' force proportional to the displacement x from the equilibrium position ($-kx$). I also assume the presence of a 'frictional' force acting against the direction of motion ($-b\dot{x}$). This system is equivalent to a damped harmonic oscillator (Riley et al., 1997) with the following equation of motion:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\frac{\mathrm{d}x}{\mathrm{d}t}+kx=0 \tag{4.1}$$

Additional forces acting on the system now appear on the right-hand side of this equation as perturbations. Consider a course of antibiotics of duration τ. If we are interested in the behaviour of the system at timescales $T \gg \tau$, we can assume for simplicity that this perturbation is of infinitesimal duration and model it as an impulse of magnitude D acting at time $t=0$:

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}+b\frac{\mathrm{d}x}{\mathrm{d}t}+kx=D\delta(t) \tag{4.2}$$

To solve this second order differential equation, let us assume that $b^2 > 4k$ (the overdamped' case) based on the lack of any oscillatory behaviour previously observed in the gut microbiome, which would imply $b^2 < 4k$ (underdamping). Then, subject to the initial conditions $x(0^+) = 0$ (system at equilibrium at time $t = 0$) and $\dot{x}(0^+) = D$ (a gain in momentum given by the magnitude of the impulse at time $t = 0$) we obtain the following equation describing the system's trajectory:

$$x(t) = \frac{D}{2\sqrt{\left(\frac{b}{2}\right)^2 - k}} \left(e^{-\left(\frac{b}{2} - \sqrt{\left(\frac{b}{2}\right)^2 - k}\right)t} - e^{-\left(\frac{b}{2} + \sqrt{\left(\frac{b}{2}\right)^2 - k}\right)t} \right) \tag{4.3}$$

Fitting the model therefore requires fitting three parameters: b (the damping on the system), k (the strength of the restoring force), and D (how strong the perturbation is). For the purposes of fitting the model, I choose to reparameterise the model using the following definitions:

$$b = e^{\phi_1} + e^{\phi_2} \tag{4.4}$$

$$k = e^{\phi_1 + \phi_2} \tag{4.5}$$

Resulting in the following model M_1 (Figure 4.1C):

$$x_1(t) = \frac{De^{\phi_1} e^{\phi_2}}{e^{\phi_2} - e^{\phi_1}} \left(e^{-e^{\phi_1} t} - e^{-e^{\phi_2} t} \right) \tag{4.6}$$

Antibiotics may lead not just to displacement from equilibrium, but also state transitions to new equilibria (Modi et al., 2014). To investigate this possibility, I also consider a model M_2 where the value of equilibrium diversity asymptotically tends to a new value A (Figure 4.1C).

$$x_2(t) = x_1(t) + A\left(1 - e^{-e^{\phi_1} t}\right) \tag{4.7}$$

4.2.3 Empirical dataset

To validate the model and test whether antibiotic perturbation caused a state transition I fitted both models to an empirical dataset and compared the results. Zaura et al. (2015) conducted a study on the long-term effect of antibiotics on the gut microbiome which provides an ideal test dataset. As part of this study, 30 Swedish individuals (15 males and 15 females, average age 26 years, range 18-45 years) were randomly assigned to either ciprofloxacin, clindamycin, or a placebo. The antibiotics (150 mg clindamycin four times a day, 500 mg ciprofloxacin twice a day) and placebo were administered for $\tau = 10$ days and longitudinal faecal samples collected until 1 year afterwards (i.e. $\frac{\tau}{T} \approx 0.027 \ll 1$) at baseline, after treatment, one month, two months, four months, and one year. Samples underwent 16S rRNA gene amplicon sequencing, targeting the V5-V7 region (SRA: SRP057504). I reanalysed this data, doing *de novo* clustering into OTUs at 97% similarity with VSEARCH v1.1.1 (Rognes et al., 2016) as described previously (Section 2.2.2). Taxonomy was assigned with RDP (Q. Wang et al., 2007).

4.2.4 Phylogenetic diversity

There are many possible diversity metrics that could be used to compute the displacement from equilibrium (see Table 1.1). Because of the assumption that phylogenetic diversity approximates functional potential, which is itself a proxy for ecosystem 'health' (see Section 4.2.1), I chose to use Faith's phylogenetic diversity (Faith, 1992) calculated with the `pd()` function in the picante R package v1.6-2 (Kembel et al., 2010). Calculating Faith's phylogenetic diversity requires a phylogeny, which I produced with RaxML v8.1.15 (Stamatakis, 2014) after aligning 16S rRNA V5-V7 OTU sequences with Clustal Omega v1.2.1 (Sievers et al., 2011). To obtain values for fitting the model, I used mean bootstrapped values ($n = 100$, sampling depth $r = 2000$) of the phylogenetic diversity at timepoint i for individual j: d_i^j. These were scaled relative to the baseline phylogenetic diversity d_0^j for that individual, representing the displacement from equilibrium in the model:

$$\tilde{d}_i^j = d_i^j - d_0^j \tag{4.8}$$

4.2.5 Model fitting

I used a Bayesian framework to fit models 1 and 2 (equations 4.6 and 4.7) using Stan (Carpenter et al., 2017) and RStan (Stan Development Team, 2017) to the three separate groups: placebo, ciprofloxacin, and clindamycin. In brief, I used 4 chains with a burn-in period of 10,000 iterations and 100,000 subsequent iterations, verifying that all chains converged ($\hat{R}$ = 1) and the effective sample size for each parameter was sufficiently large ($n_{\text{eff}} > 10{,}000$).

I used uninformative priors for the three parameters in the original model M_1 without a state transition (equation 4.6). For ciprofloxacin and clindamycin I used the same uniformly distributed prior for D, and uniform priors for ϕ_1, ϕ_2. For the model M_2 with a state transition (equation 4.7) I used the same priors, with a normal prior centred at zero for the new equilibrium value A with a standard deviation given by the standard deviation of the displacement of placebo samples from baseline after a year ($\sigma = 1.263$), with bounds between -2 and 2. The priors are as follows:

$$D \sim \text{uniform}(0, 15) \tag{4.9}$$

$$\phi_1 \sim \text{uniform}(-1.99, 1.99) \tag{4.10}$$

$$\phi_2 \sim \text{uniform}(-2, 2) \tag{4.11}$$

$$A \sim \text{normal}(\mu = 0, \sigma = 1.263) \tag{4.12}$$

Group	n	% males	% Caucasian	Age, yrs	Weight, kg	Height, cm
Placebo	10	50	100	26 (4)	74 (9)	179 (10)
Ciprofloxacin	10[a]	50	80	26 (3)	69 (13)	176 (10)
Clindamycin	9[b]	56	100	24 (5)	67 (11)	175 (9)

Table 4.1: Summary demographic characteristics of participants in each treatment group. Age, weight, and height are given as mean value ± standard deviation. Adapted from Zaura et al. (2015).

a. On reanalysis after downloading data from SRA Run Selector, I found that participant KI17 was missing 2/6 faecal samples, so they were excluded from analysis i.e. leaving $n = 9$ for the reanalysis of ciprofloxacin as well as clindamycin. However, these summary statistics apply before the exclusion of KI17.

b. One female participant who was initially recruited dropped out of the study after enrolment.

The marginally different priors for ϕ_1 and ϕ_2 are because of the way Stan chooses initial values. For the placebo group, I expected no perturbation response so used a uniform prior for D centred at zero:

$$D \sim \text{uniform}(-5, 5) \tag{4.13}$$

I compared models M_1 and M_2 for each treatment group using the Bayes factor (Aitkin, 1991; Kass and Raftery, 1995) after extracting the model fits using bridge sampling with the bridgesampling R package v0.2-2 (Gronau et al., 2017). A prior sensitivity analysis showed that choice of priors did not affect the conclusion that M_2 outperformed M_1 for the two antibiotics, although the strength of the Bayes factor varied.

Full code for fitting the models is available as Supplementary Material on *biorxiv* (Shaw, Barnes, et al., 2017).

4.3 Results

4.3.1 Dataset

I fitted the model to published data from Zaura et al. (2015) where 30 individuals received a ten-day course of either a placebo, ciprofloxacin, or clindamycin (Table 4.1). Clindamycin is a lincosamide with a broad spectrum of activity against Gram-positive aerobes and anaerobes Gram-negative anaerobes (Guay, 2007). Ciprofloxacin is a quinolone which targets bacterial DNA topoisomerase and DNA gyrase, making it active against a range of Gram-positive and Gram-negative bacteria (Mustaev et al., 2014). Faecal samples were taken at baseline (i.e. before treatment), then subsequently at ten days, one month, two months, four months, and one year after treatment.

4.3.2 An impulse response model for the effect of antibiotics

The model (Figure 4.1) assumes that a short course of antibiotics can be modelled as an impulse on the gut microbiome. With some additional simplifying assumptions about

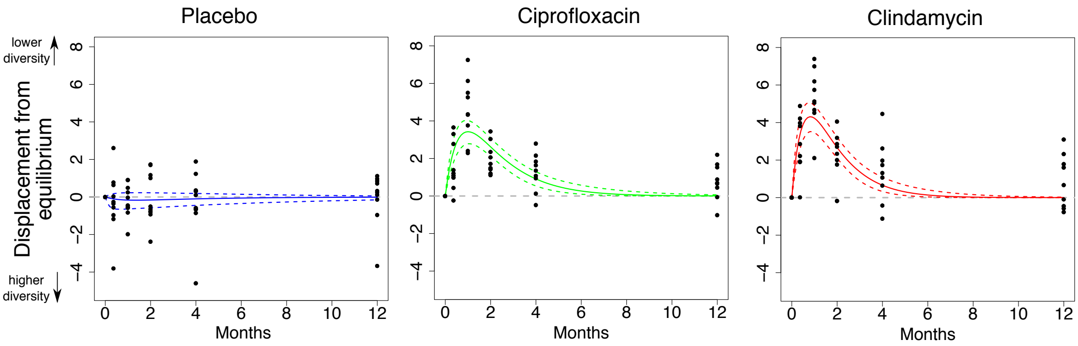


Figure 4.2: An impulse response model captures the dynamics of the effect of antibiotics on the gut microbiome. Bayesian fits with Stan for participants taking either a placebo ($n = 10$), ciprofloxacin ($n = 9$), or clindamycin ($n = 9$). The mean phylogenetic diversity from 100 bootstraps for each sample (black points) and median and 95% credible interval from the posterior distribution (bold and dashed coloured lines, respectively). The grey line indicates the equilibrium diversity value, defined on a per-individual basis relative to the mean baseline diversity. The biased positive skew of residuals after a year suggests the possibility of a transition to an alternative stable state with persistently reduced diversity.

the form of the stability landscape (see Section 4.2.1), I derive an analytical form for this overdamped impulse response in terms of the phylogenetic diversity of the gut microbiome (M_1; equation 4.6).

The model M_1 appeared to adequately describe the initial response to antibiotics (Figure 4.2), where diversity decreases (i.e. displacement from equilibrium increases) before returning gradually towards equilibrium. Despite large variability between samples from the same treatment group, reassuringly the placebo group clearly did not warrant an impulse response model whereas data from individuals receiving ciprofloxacin and clindamycin was qualitatively in agreement with the model.

However, the residuals suggested that diversity after a year was not well-captured by the model, with a substantially positive skew of samples in Figure 4.2. In their analysis, Zaura et al. (2015) noted significantly ($p < 0.05$) reduced Shannon diversity when comparing samples a year after receiving ten days' ciprofloxacin to baseline, but this could have in principle merely been due to slow reconstitution and a return to the original equilibrium under the dynamics I have described but with greater damping.

Fitting the impulse model to the data and taking into account the whole temporal response suggests that the lack of return to the initial equilibrium state is not due to slow reconstitution of the initial microbiome species community. Instead, the distribution of residuals indicates that, while the initial response fits a standard impulse response model well, the longer-term dynamics of the system did not, as might be expected under a scenario involving a long-term transition to an alternative community state (Figure 4.1). I therefore developed a variant of the model (equation 4.7) to take into account potential shifts to alternative stable states.

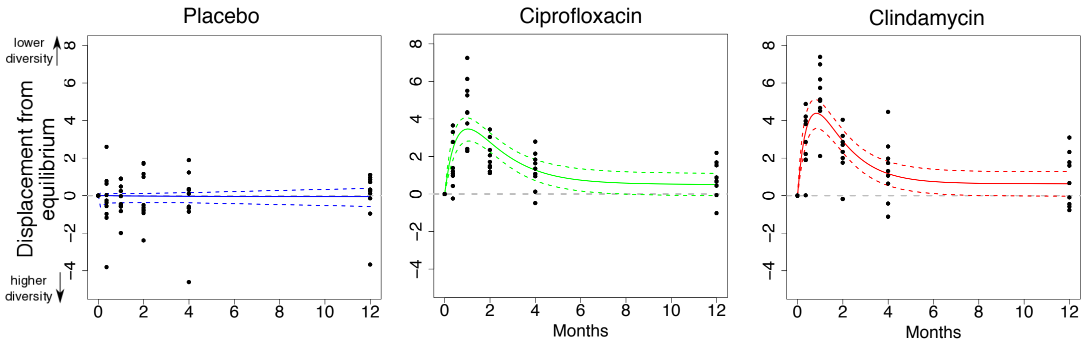


Figure 4.3: A model with a possible state transition improves the fit to empirical data. Bayesian fits with Stan for participants taking either a placebo ($n = 10$), ciprofloxacin ($n = 9$), or clindamycin ($n = 9$). The mean phylogenetic diversity from 100 bootstraps for each sample (black points) and median and 95% credible interval from the posterior distribution (bold and dashed coloured lines, respectively). The grey line indicates the equilibrium diversity value, defined on a per-individual basis relative to the mean baseline diversity. The biased positive skew of residuals after a year suggests the possibility of a transition to an alternative stable state with persistently reduced diversity. ăThe non-zero-centred asymptote indicates support for a state transition.

4.3.3 Support for an antibiotic-induced state transition

To test the hypothesis that the course of antibiotics could have moved individuals' gut microbiomes into alternative states, I fitted an extended version of the model that allowed a potential non-zero asymptotic value (M_2; equation 4.7), representing a new long-term value of diversity. I assumed a normally distributed prior for the asymptote parameter A centred at zero (i.e. return to original equilibrium) with a variance given by the variance of the displacement of placebo samples from baseline after a year.

Qualitatively, this slightly more complex model gave a similar fit (Figure 4.3) but with a positive displacement from equilibrium, corresponding to an alternative equilibrium state with lower diversity. I compared models with the Bayes factor K, where $K > 1$ indicates support for one model over another. There was no support for M_2 over M_1 for the placebo ($K = 0.96$) but support for ciprofloxacin ($K = 3.36$) and clindamycin ($K = 3.99$). The posterior estimates for the asymptote parameter for ciprofloxacin and clindamycin were substantially positively skewed (Figure 4.4), providing evidence of a transition to a state with lower phylogenetic diversity than the baseline.

4.3.4 Comparison of parameters between antibiotics

Comparing the posterior distribution of parameters for fits of M_2 between treatment groups (Figure 4.4), the strength of the perturbation parameter D was not substantially different between antibiotics. The asymptotic equilibrium parameter A was positively skewed for both antibiotics (median (95% CI): $A_{\text{clinda}} = 0.66$ (-0.13 – 1.41); $A_{\text{cipro}} = 0.58$ (-0.14 – 1.27), strongly suggesting persistent detrimental effects on microbiome diversity and a transition to an alternative stable state.

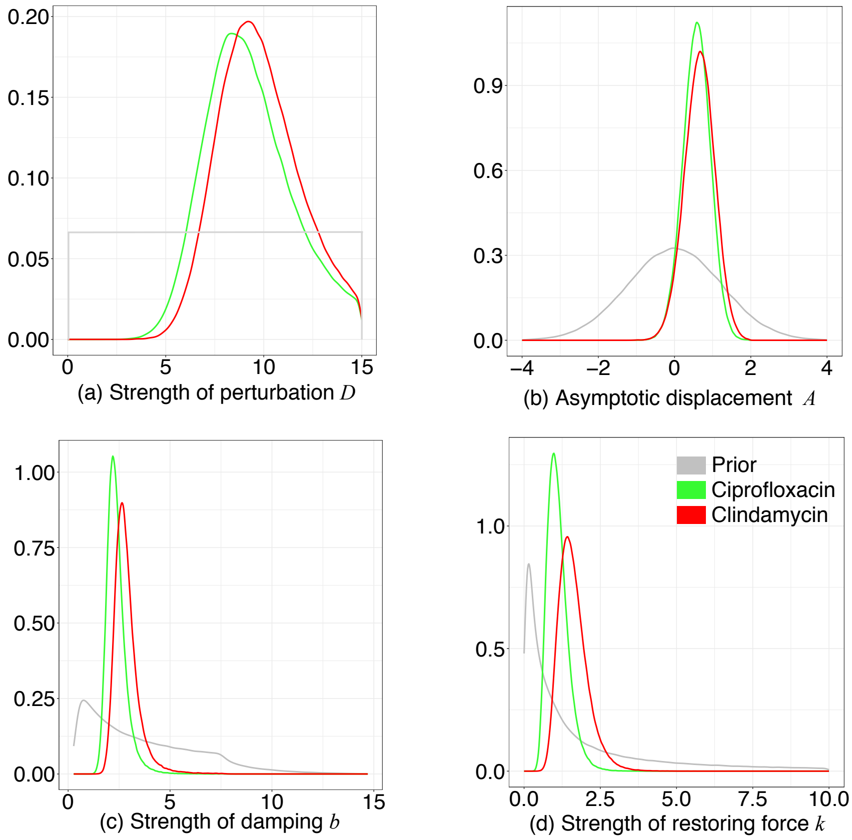


Figure 4.4: Posterior parameter estimates for model with a possible transition to an alternative stable state. The posterior distributions from Bayesian fits of M_2 (equation 4.7) to empirical data for ciprofloxacin (green) and clindamycin (red). Each posterior distribution represents 400,000 iterations in total.

The parameters b and k were both greater in clindamycin compared to ciprofloxacin. The damping ratio $\zeta = \frac{b}{2\sqrt{k}}$ summarises how perturbations decay over time, and is an inherent property of the system independent of the perturbation itself. Therefore, if the modelling framework and ecological assumptions were valid we would expect a consistent damping ratio across both the clindamycin and ciprofloxacin groups. This is indeed what is observed, with median (95% CI) damping ratios of $\zeta_{\text{clinda}} = 1.07\ (1.00 - 1.65)$ and $\zeta_{\text{cipro}} = 1.07\ (1.00 - 1.66)$, substantially different from both the prior and the posterior distribution in the placebo group of $\zeta_{\text{placebo}} = 1.21\ (1.00 - 3.00)$, supporting the view of the gut microbiome as a damped harmonic oscillator.

4.3.5 True complexity of response does not prevent modelling

While it is not my intention to repeat a comprehensive description of the precise nature of the response for the different antibiotics, I note here some interesting qualitative observations from my reanalysis that highlight the complexity of the antibiotic response. While modelling these interactions is far beyond the scope of this model, I wish to make the point that the approach is unaffected by this underlying complexity. I discuss here observations at the level of taxonomic family (Figure 4.5).

Despite their different mechanisms of action, both clindamycin and ciprofloxacin

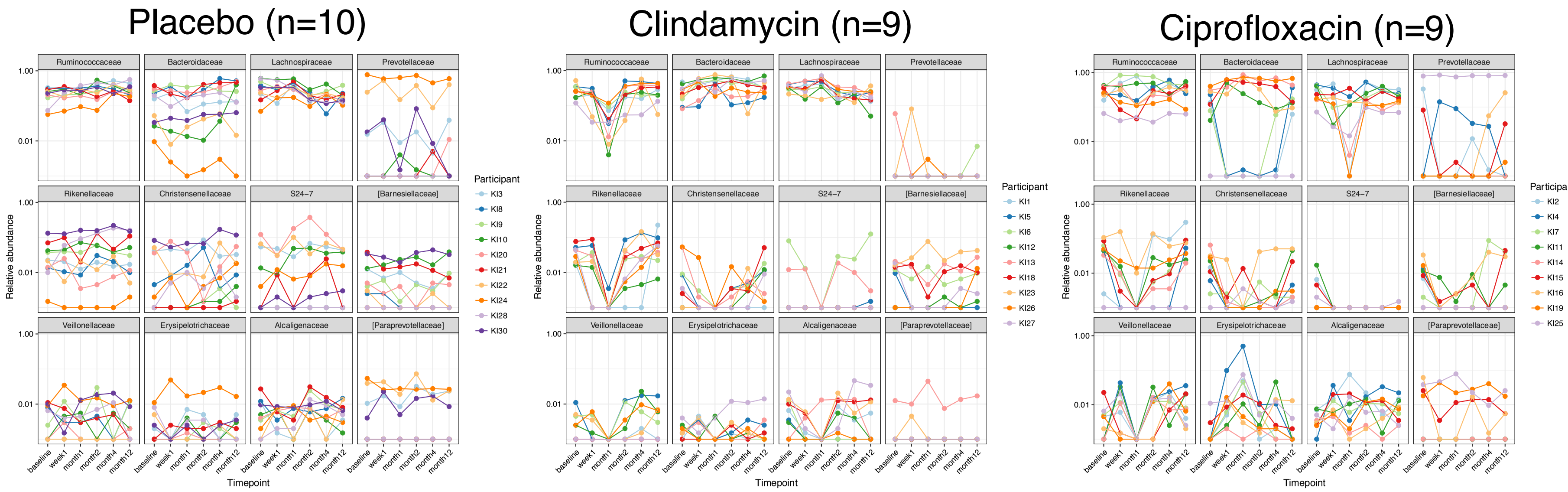


Figure 4.5: Differences in individual response over time for the top twelve most abundant taxonomic families for each treatment group. Relative abundances (log-scale) of the top twelve most abundant bacterial families plotted at each sampled timepoint. Observations are linked by coloured lines for each individual. Despite some consistency in changes between antibiotics across individuals, there is inter-individual variability and evidence of possible interactions between bacterial families.

caused a dramatic decrease in the Gram-negative anaerobes *Rikenellaceae*, which was most marked a month after the end of the course. However, for ciprofloxacin this decrease had already started immediately after treatment, whereas for clindamycin the abundance after treatment was unchanged in most participants. The different temporal nature of this response perhaps reflects the bacteriocidal nature of ciprofloxacin (Mustaev et al., 2014) compared to the bacteriostatic effect of clindamycin, although concentrations in vivo can produce bacteriocidal effects (Spizek and Rezanka, 2004).

There were clear differences in response between antibiotics. For example, clindamycin caused a decrease in the anaerobic Gram-positives *Ruminococcaceae* after a month, whereas ciprofloxacin had no effect. Conversely ciprofloxacin caused lower levels of *Barnesiellaceae* which was largely unaffected by clindamycin.

Some families appeared unaffected by antibiotics: the *Bacteroidaceae* were largely unaffected in most individuals. Furthermore, while overall diversity decreased, this can still be consistent with increases in the relative abundance of certain taxa. For example, ciprofloxacin led to increases in *Erysipelotrichaceae*, which were dramatic in some individuals. Interestingly, for these individuals these increases coincided with marked decreases in *Bacteroidaceae*, suggesting the relevance of inter-family microbial interactions (Figure 4.5). The individualized nature of the ciprofloxacin response was also noticeable in *Lachnospiraceae* – which was largely unaffected by clindamycin – as its abundance dropped below detectable levels in some individuals after a month but remained unchanged in other individuals.

Comparing relative abundances at the family level, there were few differences between community states of different treatment groups after a year. Equal phylogenetic diversity can be produced by different community composition, and this suggests against consistent trends in the long-term dysbiosis associated with each antibiotic. However, I did find that *Peptostreptococcaceae*, a member of the order *Clostridiales*, was significantly more abundant in the clindamycin group when compared to both the ciprofloxacin group and the placebo group separately ($p < 0.05$, Wilcoxon rank sum test). In a clinical setting, clindamycin is well-established to lead to an increased risk of a life-threatening infection caused by another member of the same order: *Clostridium difficile* (Thomas et al., 2003). The long-term reduction in diversity may well similarly increase the risk of colonization and overgrowth of pathogenic species.

4.4 Discussion

4.4.1 Conclusions

Starting from a common qualitative conceptual picture of the gut microbiome as resting within a stability landscape, I have developed a simple mathematical model of its response to perturbation. With a few simplifying ecological assumptions, most notably that the phylogenetic diversity of the gut microbiome relative to its baseline value can parameterise this stability landscape, I have demonstrated that the response of the gut microbiome to a short course of antibiotics can be modelled as an impulse acting on a damped harmonic oscillator. Crucially, the simplifications involved appear to be justified at some fundamental level, as this model proves to successfully capture empirical dynamics from a previous study (Zaura et al., 2015). From this, I suggest that the restoring forces that contribute to the gut microbiome's resilience to perturbation are proportional to displacement from equilibrium and that the system is overdamped.

This approach uses a simple conceptual model to give mechanistic insight. Zaura et al. (2015) made the observation from their dataset that the lowest diversity was observed after a month rather than immediately after treatment stopped. This cannot be due to a persistence of the antibiotic effect, as clindamycin and ciprofloxacin only have short half-lives of the order of hours (Bergan et al., 1987; Leigh, 1981). Furthermore, measured concentrations of ciprofloxacin and clindamycin in faeces were higher than the MICs of most members of the gut microbiome, with mean concentrations of 168.5 ± 41.4 mg/kg and 147.4 ± 126.9 mg/kg respectively (Rashid et al., 2015). The model gives us a mechanistic framework for thinking about this temporal delay: the full effects of the transient impulse take time to be realized due to the overdamped nature of the system, and I found a consistent damping ratio for both antibiotics analyzed.

I have also demonstrated how this modelling framework could be used to compare different hypotheses about the long-term effect of antibiotic perturbation on the gut microbiome by fitting different models and using Bayesian model selection. This modelling work provides an additional line of evidence that while short-term restoration obeys a simple impulse response model, the underlying long-term community state can be fundamentally altered by a brief course of antibiotics, as suggested previously by others (Dethlefsen and Relman, 2011), raising concerns about the long-term impact of antibiotic use on the gut microbiome. Despite the noisiness of the dataset and use of uninformative priors, I found better support for a model with a state transition, which was not observed in individuals taking a placebo. The transition to a new state with reduced diversity may increase the risk of colonization and overgrowth of pathogenic species. Even if only marginal, when considered at a population level this may mean that antibiotics have substantial negative health consequences that could support reductions in the length of antibiotic courses, in addition to concerns about antibiotic resistance (Llewelyn et al., 2017). Modelling the long-term impact on the microbiome of different doses and courses could help to influence the use of antibiotics in routine clinical care.

While the evidence for a long-term state transition is weak at present, I argue that we can at the very least conclude that the restoration of diversity after a year does not seem to obey the same underlying dynamics that govern the initial response – even if we should perhaps remain agnostic about the most appropriate model refinement. Implicit in some definitions of ecological resilience is the assumption that the fundamental shape of the stability landscape remains unaltered (Gunderson, 2000), which I have assumed when drawing the visual aids to accompany the model. However, an alternative schematic picture could be drawn where a harmonic potential with a single equilibrium value gradually shifts over time to a new equilibrium value. In this interpretation the landscape has been fundamentally altered, representing an irreversible change.

4.4.2 Limitations

The sample size is small so the precise posterior estimates for parameters that I obtained should not be over-interpreted, but comparing antibiotics using these estimates represents another practical application of such simple models. However, these posterior estimates for the model parameters were fairly wide, which is to be expected with a sparse and small dataset. Hierarchical mixed effects models may offer an improved fit, particularly if they take into account other covariates; however, here I lacked any metadata on the participants from the original study, with only summary statistics available for each treatment group (Table 4.1).

A single metric clearly fails to capture all the complexity of the microbial community and its interactions. Nevertheless, the observation that treating phylogenetic diversity as the variable underlying the stability landscape leads to a reasonable fit of a simple model is interesting, as it supports observations of functional redundancy in the gut microbiome (Turnbaugh et al., 2007). An interesting extension of this work would be to systematically fit the model to a variety of diversity metrics and assess the model fit to see which metric (or combination of metrics) is most appropriately interpreted as the state variable parameterizing the stability landscape. Such an analysis could use the Hill diversity ${}^{q}D$ to assess model fit as q was varied (Table 1.1). A possible complementary approach could consider the diversity of the gut resistome, which is the collection of antibiotic resistance genes harboured in the gut microbiome (Schaik, 2015).

I would not expect the behavior with longer or repeated courses of antibiotics to be well-described by an impulse response model, but it would be possible to use the mathematical framework given here to obtain an analytic form for the possible system response by convolving any given perturbation function with the impulse response. It remains to be seen whether this simple model would break down in such circumstances.

The detailed nature of the gut microbiome's response to clindamycin and ciprofloxacin was individualized in the dataset, as others have also observed with shotgun sequencing of samples from healthy participants given a second-generation cephalosporin (Raymond, Déraspe, et al., 2016). I believe it would be a mistake to react to this complexity by assuming that no simplified model can capture general details of the ecosystem. At this stage of our understanding, creating a comprehensive inter-species model of the hundreds of members of the gut microbiome appears intractable. My recommendation is that microbiome research instead starts with ecologically-informed simple models and believe there is a place for both 'bottom-up' models using pairwise interactions for systems of reduced complexity (like bioreactors) and 'top-down' models using general ecological principles, as I have attempted to demonstrate here.

4.4.3 Summary

I have shown that comparing different hypotheses about the response of the gut microbiome to antibiotics is possible by using a simple model derived from minimal assumptions about the nature of its equilibrium diversity and response to perturbation. Future mathematical models of the gut microbiome, in conjunction with carefully designed longitudinal studies, will offer many more opportunities to rigorously test ecological hypotheses.

Chapter 5

The global distribution and spread of an antibiotic resistance gene

Declaration of contributions

This work was the product of a collaboration involving many individuals: Ruobing Wang, Qi Wang, Xiaojuan Wang, Longyang Jin, Qing Zhang, Yuqing Liu and Hui Wang collected samples. Ruobing Wang, Qi Wang, Xiaojuan Wang, Longyang Jin, Qing Zhang and Yuqing Liu performed microbial identification, antimicrobial susceptibility testing, screening for *mcr-1*, and DNA extraction for whole-genome sequencing. Ruobing Wang, Lucy van Dorp and I assembled the new sequence data. Lucy van Dorp and I jointly curated the global dataset and performed all the computational analyses. Thamarai Dorai-Schneiders advised on functional aspects of colistin resistance. Adrien Rieux, Lucy Anne Weinert, and Xavier Didelot helped with the phylogenetic reconstructions. Phelim Bradley and Zamin Iqbal performed the search for *mcr-1*-positive samples on the Short Read Archive. I jointly wrote the associated paper with Lucy van Dorp and Francois Balloux, with feedback from all co-authors. For the avoidance of doubt, in this chapter I use 'we' when referring to analysis that was a joint collaboration between myself and others and 'I' when referring to analysis exclusively performed by me.

Publication

This work has been published at *Nature Communications* and is available as R. Wang et al. (2018):

R. Wang†, L. van Dorp†, L. P. Shaw†, P. Bradley, Q. Wang, X. Wang, L. Jin, Q. Zhang, Y. Liu, A. Rieux, T. Dorai-Schneiders, L. A. Weinert, Z. Iqbal, X. Didelot, H. Wang and F. Balloux (2018). The global distribution and spread of the mobilized colistin resistance gene *mcr-1*. *Nature Communications* **9**(1):1179. doi: 10.1038/s41467-018-03205-z.

† : equal contribution

5.1 Introduction

This chapter delves in far greater detail into the bacterial genetics of antibiotic resistance, focusing on a single gene that confers resistance to colistin: *mcr-1*. I demonstrate how using a combination of publicly available data and novel data allows the identification of a consistent unit across hundreds of sequences from within the human microbiome and beyond. Colistin resistance is particularly interesting because it is emblematic of the growing problems of antimicrobial resistance worldwide, which represent a major concern for future human health (Section 1.3.2). Colistin was largely abandoned as a treatment for bacterial infections in the 1970s due to its high toxicity and low renal clearance, but has been reintroduced in recent years as an antibiotic of 'last resort' against multi-drug-resistant (MDR) infections (Grégoire et al., 2017). It is therefore alarming that resistance to colistin may be becoming more widespread, following the identification of plasmid-mediated colistin resistance in late 2015 (Liu et al., 2016).

Up until 2015, resistance to colistin had only been linked to mutational and regulatory changes mediated by chromosomal genes (Olaitan et al., 2014; Lee et al., 2016). The mobilized colistin gene *mcr-1* was first described in a plasmid from *Enterobacteriaceae* isolated in China in April 2011 (Liu et al., 2016). The presence of colistin resistance on mobile genetic elements poses a significant public health risk, as these can spread rapidly by horizontal transfer, and may entail a lower fitness cost (Carattoli, 2013). At the time of writing, *mcr-1* has been identified in numerous countries across five continents. Significantly, *mcr-1* has also been observed on plasmids containing other antimicrobial resistance genes such as carbapenemases (Poirel et al., 2016; Du et al., 2016; Yao et al., 2016) and extended-spectrum β-lactamases (ESBL) (X.-F. Zhang et al., 2016; Falgenhauer et al., 2016; Haenni et al., 2016).

The *mcr-1* element has been characterized in a variety of genomic backgrounds (Y. Wang, R. Zhang, et al., 2017; R. Li et al., 2017; Matamoros et al., 2017; Zhou et al., 2017), consistent with the gene being mobilized by a transposon. To date, *mcr-1* has been observed on a wide variety of plasmid types, including IncI2, IncHI2, and IncX4 (Matamoros et al., 2017). Intensive screening efforts for *mcr-1* have found it to be highly prevalent in a number of environmental settings, including the Haihe River in China (D. Yang et al., 2017), recreational water at public urban beaches in Brazil (Fernandes et al., 2017), faecal samples from otherwise healthy individuals in China (Y. Wang, Tian, et al., 2017), and Dutch travellers who had recently visited Southern Asia (Wintersdorff et al., 2016). While both Brazil and China have now banned the use of colistin in agriculture, this evidence that *mcr-1* can spread within hospital environments – even in the absence of colistin use (Y. Wang, Tian, et al., 2017) – as well as in the community (Wintersdorff et al., 2016) raises the possibility that the spread of *mcr-1* will not be contained by these bans. The spread of *mcr-1* across multiple bacterial communities worldwide, including into the human microbiome, demonstrate the reality of the wider environmental meta-community of human-associated bacterial communities. The human microbiome can both acquire antibiotic resistance genes from the wider community and serve as a reservoir of those genes (Section 1.3.2).

The global distribution of *mcr-1* over at least five continents is well documented, and an evolutionary model for its mobilization has been proposed. Snesrud et al. (2016) analyzed a collection of 77 *mcr-1*-containing sequences and identified a common 2,607-bp sequence flanked at one or both ends by the insertion sequence IS*Apl1*. They proposed that this composite IS*Apl1*-*mcr-1*-PAP2- IS*Apl1* transposon has mobilized the *mcr-1* gene (Figure 5.2). However, little is known about the origin, acquisition, emergence, and spread of *mcr-1*; in principle, these issues can be addressed by identifying the composite transposon (the ecological unit) and conducting phylogenetic analysis to understand its evolutionary past.

In this chapter, I extend the work of Snesrud et al. (2016) and report investigations into these fundamental issues. I use a combination of sources to build a global dataset: whole genome sequencing data from 110 novel *mcr-1*-positive isolates from China, and an extensive collection of publicly available sequence data sourced from the NCBI RefSeq database and Short Read Archive (SRA). This dataset and our analyses support an initial single mobilization event of *mcr-1* by an IS*Apl1*-*mcr-1*-PAP2- IS*Apl1* transposon around 2006. The transposon was immobilized on several plasmid backgrounds following the loss of the flanking IS*Apl1* elements, and spread through plasmid transfer. The current distribution of *mcr-1* points to a possible origin in Chinese livestock. These results illustrate the complex dynamics of antibiotic resistance genes across multiple embedded genetic levels (transposons, plasmids, bacterial lineages and bacterial species), previously described for another resistance gene as a nested ‘Russian doll’ model of genetic mobility (Sheppard et al., 2016).

5.2 Materials and methods

5.2.1 Compilation of genomic dataset

I blasted for *mcr-1* in all NCBI GenBank assemblies (as of 16th March 2017, $n = 90,759$) using a 98% identity cut off. 195 records (0.21%; 121 assemblies, 73 complete plasmids, 1 complete chromosome) contained at least one contig with a full-length hit to *mcr-1* (1,626 bases). I only included samples with a single copy of *mcr-1*. The only isolate with multiple copies was a previously published isolate with three chromosomal copies of *mcr-1* and seven copies of IS*Apl1* (C. Y. Yu et al., 2016).

Our collaborators Phelim Bradley and Zam Iqbal also searched a snapshot of all whole-genome sequenced bacterial raw read datasets in the NCBI SRA (December 2016), looking for samples containing *mcr-1* by using a k-mer index ($k = 31$) which they had previously constructed (Bradley et al. (2017), https://github.com/phelimb/cbg). This snapshot consisted of 455,632 samples, of which 7,799 were excluded as they exceeded an arbitrary threshold of 10 million kmers after error-cleaning with McCortex (Turner et al., 2017), and identified 184 datasets that contained at least 70% of the 31-mers in *mcr-1*. After removing duplicates (i.e. those with a draft assembly available) we could assemble contigs with *mcr-1* for 153 of these.

The final combined dataset comprised 457 isolates from six genera across 31 different countries, ranging in date from 2008 to 2017. Where only a year was provided as the date of isolate collection the date was set to the midpoint of that year.

Whenever identified isolates did not comprise previously assembled genomes or complete plasmids, Lucy van Dorp built assemblies using a pipeline I had originally written and she had further adapted. In brief, raw fastq files were first inspected using FastQC and trimmed and filtered on a case by case basis. *De novo* assembly was then conducted using Plasmid SPADES 3.10.0 using the `-careful` switch and otherwise default parameters (Antipov et al., 2016). For those isolates sequenced using PacBio a different pipeline was employed, which Lucy van Dorp wrote. Correction, trimming and assembly of raw reads was performed using Canu (Koren et al., 2017) and assembled reads were corrected and trimmed using the tool Circlator (Hunt et al., 2015). The quality of resultant assemblies was assessed using infoseq. In both cases I identified *mcr-1* carrying contigs from these assemblies using blastn v2.2.31 (Camacho et al., 2009).

I also investigated the wider genomic context of *mcr-1* beyond the transposon. Lucy van Dorp ran Plasmid Finder 1.348 (Carattoli et al., 2014) with 95% identity to identify plasmid replicons on the *mcr-1*-carrying contigs. 182 unique contigs could be assigned a plasmid type using this method.

Assembling this large dataset was a collaborative effort primarily between Lucy van Dorp and me. We attempted wherever possible to combine information from multiple sources (across NCBI databases) to add metadata to the isolates. In several instances this identified duplicate records that were not apparent from checking accessions. We still lacked information on some isolates. At the time we finished assembling it this dataset represented all publicly available sequences with (to the best of our knowledge) all publicly available metadata. I took the responsibility for curating the final dataset and identifying duplicate records.

5.2.2 Novel samples from China

Hui Wang and her group selected 110 *mcr-1*-positive isolates from China for whole genome sequencing from a larger survey effort of both clinical and livestock isolates. Non-repetitive clinical isolates, including 1,637 *Escherichia coli* and 1,187 *Klebsiella pneumoniae*, were collected from 15 provinces of mainland China from 2011 to 2016. 72 isolates were resistant to polymyxin B, comprising 40 *E. coli* and four *K. pneumoniae* carrying *mcr-1*. Livestock samples were collected from four provinces of China in 2013 and 2016. One broiler farm of the Shandong province provided chicken anal swabs, liver, heart and wastewater isolated in 2013. In 2016, samples including faeces, wastewater, anal swabs, and internal organs of sick livestock were collected from swine farms, cattle farms and broiler farms in four provinces (Jilin, Shandong, Henan and Guangzhou). A total of 601 *E. coli* and 126 *K. pneumoniae* were isolated, of which 167 (137 *E. coli* and 30 *K. pneumoniae*) were resistant to polymyxin B. They detected *mcr-1* in 135 *E. coli* and two *K. pneumoniae*, as well as in eight *E. coli* isolated from environmental samples, which were collected from influents and effluents of four tertiary care teaching hospitals.

All of the isolates were sent to the microbiology laboratory of Peking University People's Hospital and were confirmed with routine biochemical tests, the Vitek system (bioMérieux, Hazelwood, MO) and/or MALDI-TOF (Bruker Daltonics, Bremen, Germany). The minimal inhibitory concentrations (MICs) of polymyxin B were determined using the broth dilution method. The breakpoints of polymyxin B for *Enterobacteriaceae* were interpreted with the EUCAST guidelines (EUCAST, 2017). Colistin-resistant isolates (defined as having an MIC of ≥ 2 μg/ml) were screened for *mcr-1* by PCR and sequencing as described previously (X. Wang et al., 2017).

5.2.3 Identification and alignment of *mcr-1* transposon

I searched for the *mcr-1* carrying transposon across isolates by blasting for its major components: IS*Apl1* (*Actinobacillus pleuropneumoniae* reference sequence: EF407820), *mcr-1* (from *E. coli* plasmid pHNSHP45: KP347127.1), and short sequences representing the sequences immediately upstream and downstream of *mcr-1* (from KP347127.1) using `blastn-short`. I aligned contiguous sequences containing *mcr-1* with Clustal Omega (Sievers et al., 2011) and then manually curated and amended this alignment to correct misaligned sequences using jalview v2.10.3 (Waterhouse et al., 2009), resulting in a 3,679bp alignment containing the common 2,600bp identified by Snesrud et al. (2016). The downstream copy of IS*Apl1* was more often fragmented or inverted. 28 isolates which were all assemblies from the same study in Vietnam had a 1.7kb insertion downstream of *mcr-1* (Figure 5.3e) before the downstream IS*Apl1* element.

5.2.4 Phylogenetic analyses

For constructing the transposon phylogeny, I excluded the downstream IS*Apl1* and the insertion sequence observed in a small number of samples, as well as regions identified as having signals of recombination with ClonalFrameML (Didelot and Wilson, 2015), resulting in a 3,522bp alignment. I removed two homoplastic sites (requiring >1 change on the phylogeny), before constructing a maximum parsimony neighbor-joining tree based on the Hamming distance between sequences. I calculated branch lengths using non-negative least squares with nnls.phylo in phangorn v2.2.0 (Schliep, 2011) and visualized phylogenies with ggtree v1.8.1 (G. Yu et al., 2017).

5.2.5 Phylogenetic dating

Recombination can conceal clonal phylogenetic signal. Therefore, Lucy van Dorp also applied ClonalFrameML (Didelot and Wilson, 2015) to identify regions of high recombination in a subset of IncI2 and IncX4 plasmid background alignments that I had selected. Where recombination hotspots were identified, they were removed from the alignment. In the IncI2 alignment this resulted in removing 1,281 positions. No regions of high recombination were detected in the IncX4 alignment. We applied root-to-tip correlations to test for a temporal signal in the data using TempEST (Rambaut et al., 2016) and found a significantly positive slope for all three alignments. Lucy van Dorp applied BEAUTi and BEAST v2.4.7 (Drummond et al., 2012; Bouckaert et al., 2014) to estimate a timed phylogeny from an alignment of IncI2 plasmids (7,161 sites, 110 isolates) and IncX4 plasmids (34,761 sites, 8 isolates). Sequences were annotated using their known sampling times expressed in years. For both plasmid alignments, the HKY substitution model was selected based on evaluation of all possible substitution models in bModelTest. Beast analyses were then applied under both a coalescent population model (the coalescent Bayesian skyline implementation) and an exponential growth model (Coalescent Exponential population implementation). Additionally, a strict clock, with a lognormal prior, and a relaxed clock (both lognormal and exponential) were tested. MCMC was run for 50,000,000 iterations sampling every 2,000 steps and convergence was checked by inspecting the effective sample sizes (ESS) and parameter value traces in the software Tracer v1.6.0. Analyses were repeated three times to ensure consistency between the obtained posterior distributions. Posterior trees for the best-fitting model were combined in TreeAnnotator after a 10% burn-in to provide an annotated Maximum Clade Credibility (MCC) tree. MCC trees were plotted using ggtree v1.8.2 (G. Yu et al., 2017) for both backgrounds: IncI2 (Figure A.10a) and IncX4 (Figure A.10b). The model fit across analyses was compared using the Akaikes information criteria model (AICM) through 100 boot-strap resamples as described in Baele et al. (2012) and implemented in Tracer v1.6.

Phylogenetic dating on the transposon was performed using an alignment of 364 isolates, which included only those with information on isolation date, across 3,522 sites. As before BEAST analyses were applied under both a coalescent population model (coalescent Bayesian skyline implementation) and an exponential growth model (coalescent exponential population implementation). Additionally, a strict clock, with a lognormal prior, and a relaxed clock (both lognormal and exponential) were tested. Analyses were run under a HKY substitution model for 600 million iterations sampling every 5,000 steps. Only analyses using a strict clock model reached convergence after 600 million iterations. The resultant set of trees were thinned by sampling every 10 trees and excluding a 10% burn-in and combined using TreeAnnotator to produce a MCC tree. MCC trees were plotted using ggTree (G. Yu et al., 2017). As before the model fit was evaluated using AICMs implemented in Tracer v1.6.

5.2.6 Environmental distribution

For the purpose of testing the distribution of sequences containing some trace of IS*Apl1*, I classed isolates into broad categories as either environmental ($n = 39$; bird, cat, dog, fly, food, penguin, reptile, vegetables), agricultural ($n = 213$; chicken, cow, pig, poultry feed, sheep, turkey), or human ($n = 108$). I did not correct for study site with subsampling as we found great diversity within sites, consistent with a recent study showing multiple diverse *mcr-1*-positive strains within a single hospital sewage sample (F. Zhao et al., 2017).

The human-associated samples demonstrated the presence of *mcr-1*-positive bacteria in multiple environmental niches within the human body. The largest group was blood cultures ($n = 30$), followed by faeces or rectal swabs ($n = 12$), urine ($n = 10$), wounds ($n = 5$), abdominal fluid ($n = 5$), and sputum ($n = 3$). The dataset also contained samples from wastewater influent and effluent ($n = 8$), with most other human-associated samples having missing information.

5.3 Results

5.3.1 Dataset

I compiled a global dataset of 457 *mcr-1*-positive isolates (Figure 5.1A), including 110 new whole genome sequences from China, of which 105 were sequenced with Illumina short reads and five with PacBio long read technology. 195 isolates were sourced from publicly available assemblies in the NCBI sequence repository (73 completed plasmids, 1 complete chromosome, 121 assemblies). A further 153 sequences were sourced from the Short Read Archive (NCBI-SRA), after being identified as *mcr-1*-positive using a k-mer index of a snapshot of the SRA as of December 2016 (see Methods). The whole dataset consists of 256 short-read datasets, 6 long-read PacBio WGS, 121 draft assemblies, and 74 completed assemblies.

Isolates carrying *mcr-1* were identified from 31 countries (Figure 5.1A). The countries with the largest numbers of *mcr-1*-positive samples are China (212), Vietnam (58) and Germany (25). Within China, nearly half (45%) of positive isolates stem from Shandong province (Figure 5.1B). The vast majority of *mcr-1*-positive isolates belong to *E. coli* ($n = 411$), but the dataset also comprises *mcr-1*-positive isolates from another seven bacterial species: *Salmonella enterica* ($n = 29$), *K. pneumoniae* ($n = 8$), *Escherichia fergusonii* ($n = 2$), *Kluyvera ascorbata* ($n = 2$), *Citrobacter braakii* ($n = 2$), *Cronobacter sakazakii* ($n = 1$) and *Klebsiella aerogenes* ($n = 1$) (Figure 5.1A). The majority of isolates for which sampling dates were available (80%), were collected between 2012 and 2016, with the oldest available isolates dating back to 2008 (Figure 5.1C).

The large number of *mcr-1*-positive isolates from China, and the high incidence in Shandong province can be largely ascribed to the inclusion of our 110 newly sequenced

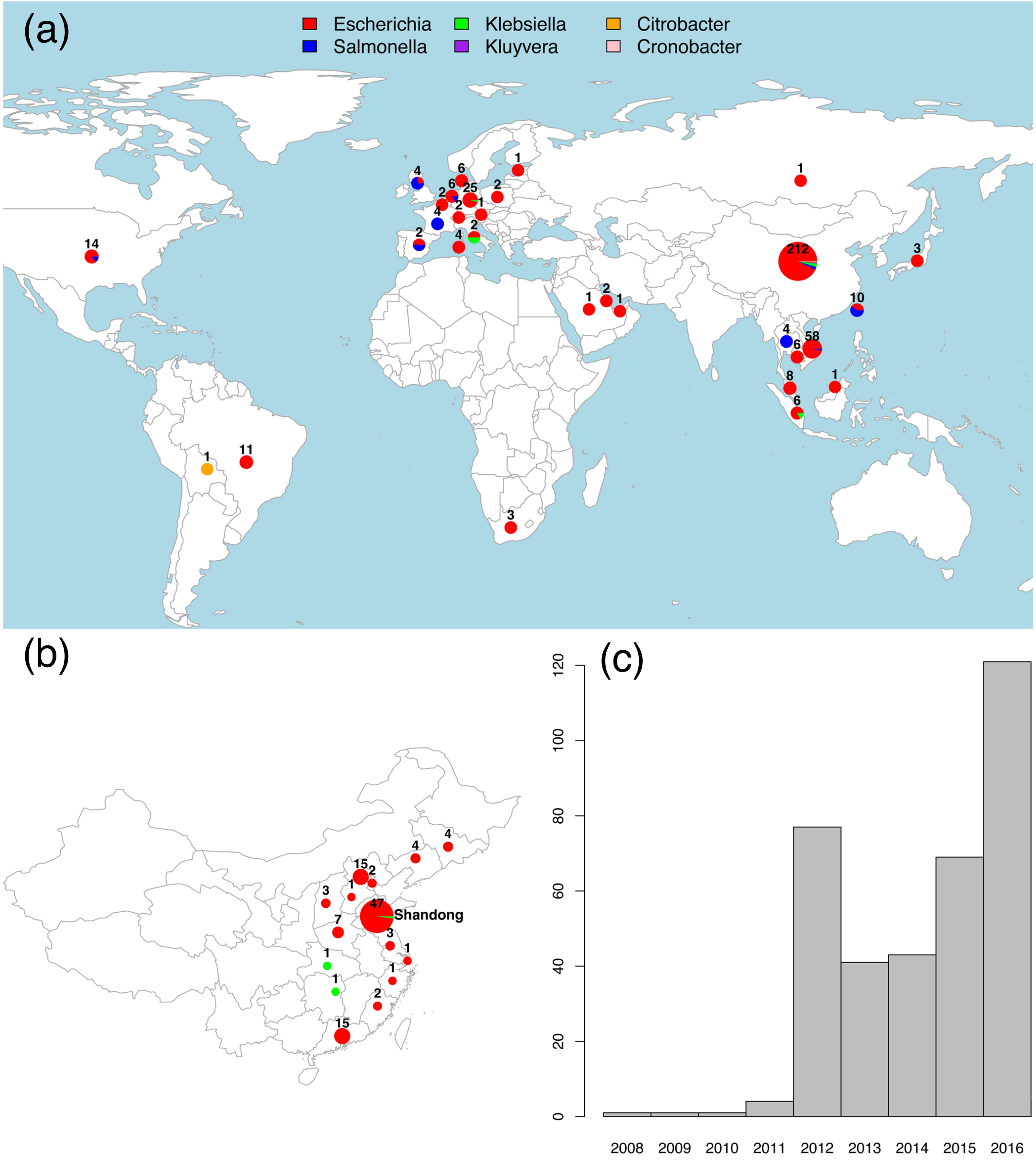


Figure 5.1: Overview of the *mcr-1*-positive isolates in the global dataset compiled for this work. (a) Isolates displayed on a per-country basis, with pie charts showing the proportion of isolates from six genera. **(b)** Map of novel Chinese isolates sequenced for this study. **(c)** Histogram of sampling dates of the isolates. Lucy van Dorp prepared this figure and I am grateful for her permission to include it.

isolates including 49 from Shandong and to another 37 isolates from a previous large sequencing effort (Y. Wang, R. Zhang, et al., 2017). However, even after discounting the isolates from these two sources, China remains one of the two countries with the highest number of sequenced *mcr-1*-positive isolates, the other being Vietnam.

5.3.2 Evolutionary model

It has been proposed that *mcr-1* is mobilized by a composite transposon formed of a 2,600bp region containing *mcr-1* (1,626bp) and a putative open reading frame encoding a PAP2 superfamily protein (765bp), flanked by two IS*Apl1* insertion sequences (Snesrud et al., 2016). IS*Apl1* is a member of the IS30 family of insertion sequences, which utilize

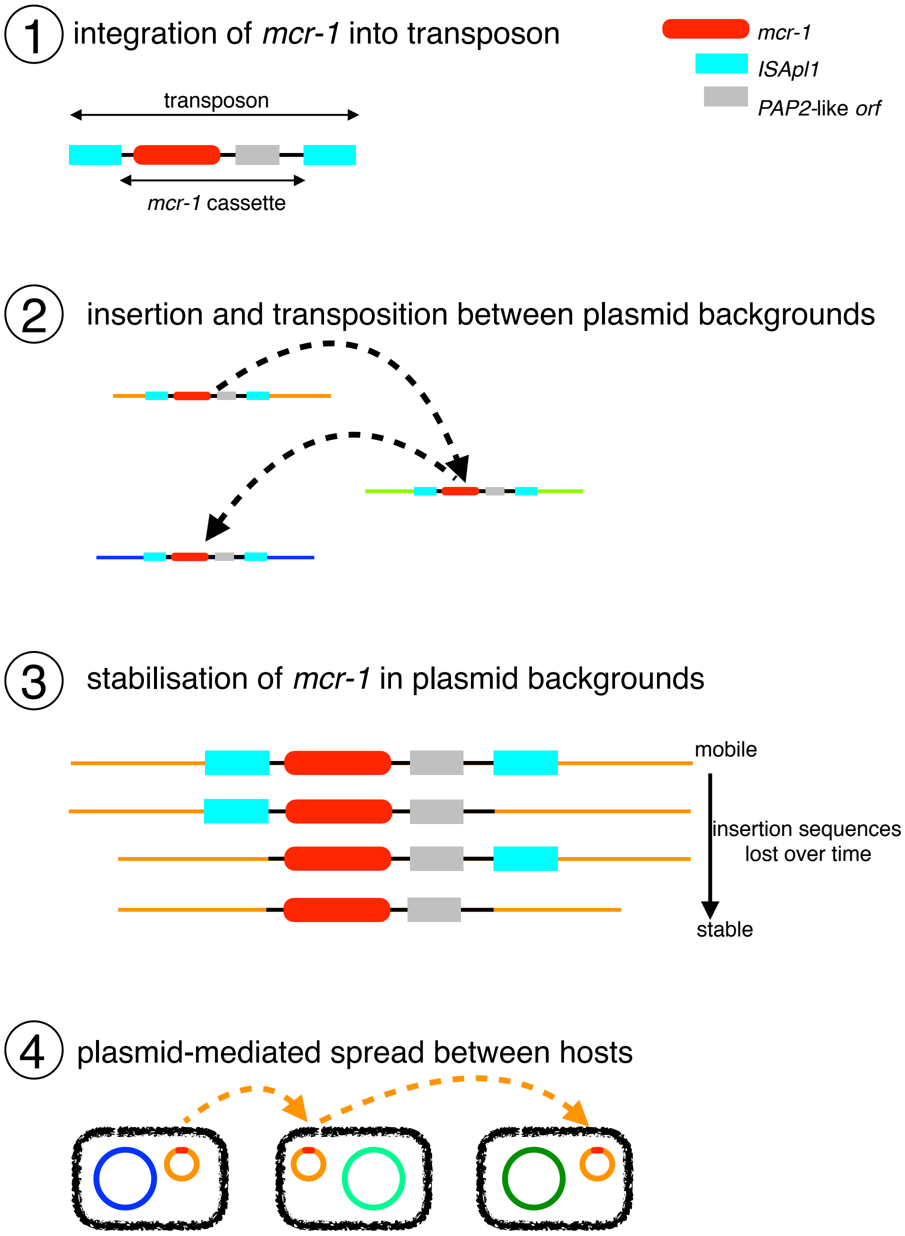

1 integration of *mcr-1* into transposon
mcr-1
ISApl1
PAP2-like *orf*
transposon
mcr-1 cassette
2 insertion and transposition between plasmid backgrounds
3 stabilisation of *mcr-1* in plasmid backgrounds
mobile
insertion sequences lost over time
stable
4 plasmid-mediated spread between hosts

Figure 5.2: Schematic representation of the evolutionary model for the steps in the spread of the *mcr-1* gene. (1) The formation of the original composite transposon, followed by (2) transposition between plasmid backgrounds and (3) stabilisation via loss of IS*Apl1* elements before (4) plasmid-mediated spread.

a ‘copy-out, paste-in’ mechanism with a targeted transposition pathway requiring the formation of a synaptic complex between an inverted repeat (IR) in the transposon circle and an IR-like sequence in the target. Snesrud et al. (2016) hypothesized that after the initial formation of such a composite transposon, these insertion sequences would have been lost over time, leading to the stabilization of *mcr-1* in a diverse range of plasmid backgrounds (Figure 5.2). With this dataset, we sought to test this model by performing an explicit phylogenetic analysis of the region surrounding *mcr-1* using our comprehensive global dataset.

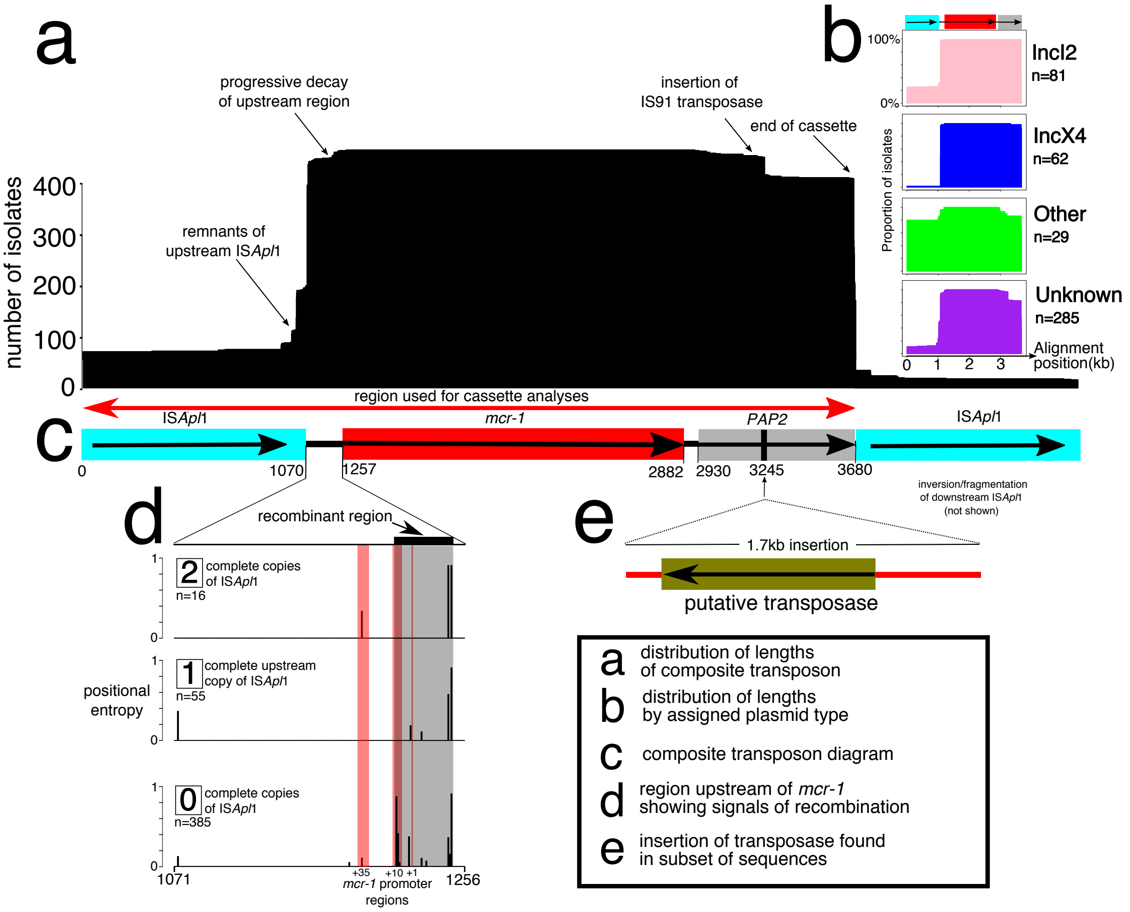


Figure 5.3: The genetic element carrying mcr-1 is a composite transposon and is alignable across all available sequences. **(a)** Length distribution of the alignment across sequences. **(b)** Length distribution subset by plasmid type. **(c)** The composite transposon, consisting of IS*Apl1*, *mcr-1*, a PAP2 , and IS*Apl1*. The region indicated by the red arrow was used in phylogenetic analyses, after the removal of recombination. Numbers underneath represent position in the alignment in bases. **(d)** The 186bp region upstream of *mcr-1* showed strong signals of recombination (grey box) that coincided with the promotor regions of *mcr-1* (red box), and this diverse region was removed from the subsequent alignment. **(e)** Some sequences from Vietnam ($n = 28$) had a 1.7kb insertion containing a region with a putative transpose, suggesting subsequent rearrangement after initial mobilization.

5.3.3 Immediate genomic background of *mcr-1*

If there had been a unique formation event for the composite transposon, followed by progressive transposition and loss of insertion sequences, one would expect to be able to identify a common immediate background region for *mcr-1* in all samples. Indeed, I was able to identify and align a shared region or remnants of it in all 457 sequences surrounding *mcr-1* (Section 5.2.3), supporting a single common origin for all *mcr-1* elements sequenced to date (Figure 5.3a). The majority of the sequences contained no trace of IS*Apl1* ($n = 260$) indicating that the *mcr-1* transposon had been completely stabilized in their genomic background. 42 sequences contained indication of the presence of IS*Apl1* both upstream and downstream, either in full copies ($n = 16$), a full copy upstream and a partial copy downstream ($n = 7$), a partial copy upstream and a full copy downstream (n=1), or partial copies upstream and downstream ($n = 18$). Some sequences only had IS*Apl1* present upstream as a complete ($n = 55$) or partial ($n = 99$) sequence, and one sequence had only a partial downstream IS*Apl1* element. The downstream copy of IS*Apl1* was inverted in some sequences ($n = 3$) and some sequences had full copies of IS*Apl1* present elsewhere on the same contig ($n = 7$), consistent with its high observed activity in transposition (Snesrud et al., 2017).

Further inspection of the transposon alignment revealed that the 186bp region between the 3' end of the upstream IS*Apl1* and *mcr-1* contained IR-like sequences similar to the IRR and IRL of IS*Apl1* (respectively: 93-142bp, 23/50 identity; and 125-175bp, 21/50 identity). The most variable positions in this 186bp region were at 177bp and 142bp, approximately coinciding with the end of the alignment with the IRs and were more variable in sequences lacking IS*Apl1*, suggesting possible loss of function of the transposition pathway associated with IS*Apl1* (Figure 5.3d). Some of these SNPs occurred in a stretch previously identified as the promoter region for *mcr-1* (Poirel et al., 2016), and this region showed strong signals of recombination. A small number of sequences (3%) had SNPs present in *mcr-1* itself. These tended to be at the upstream or 5' end of the sequence, particularly in the first three positions. A subset of the sequences from Vietnam ($n = 28$) include a secondary 1.7kb insertion downstream of *mcr-1* containing a putative transposase, indicating subsequent rearrangements involving this region after initial mobilization of the transposon (Figure 5.3e).

To reconstruct the phylogenetic history of the composite *mcr-1* transposon, I created a sequence alignment for 457 sequences (Figure 5.3c) with recombinant regions identified with ClonalFrameML removed by Lucy van Dorp, including the region immediately upstream of *mcr-1* between positions 1,212-1,247 (Figure 5.3d). The midpoint-rooted maximum parsimony phylogeny I constructed showed that there was a dominant sequence type with subsequent diversification, likely indicating the ancestral form of the composite transposon (Figure 5.4). There was no discernible clustering of isolates by sample source (Figure 5.5a) or bacterial species (Figure 5.5b), suggesting the composite transposon does not evolve differently in these different backgrounds.

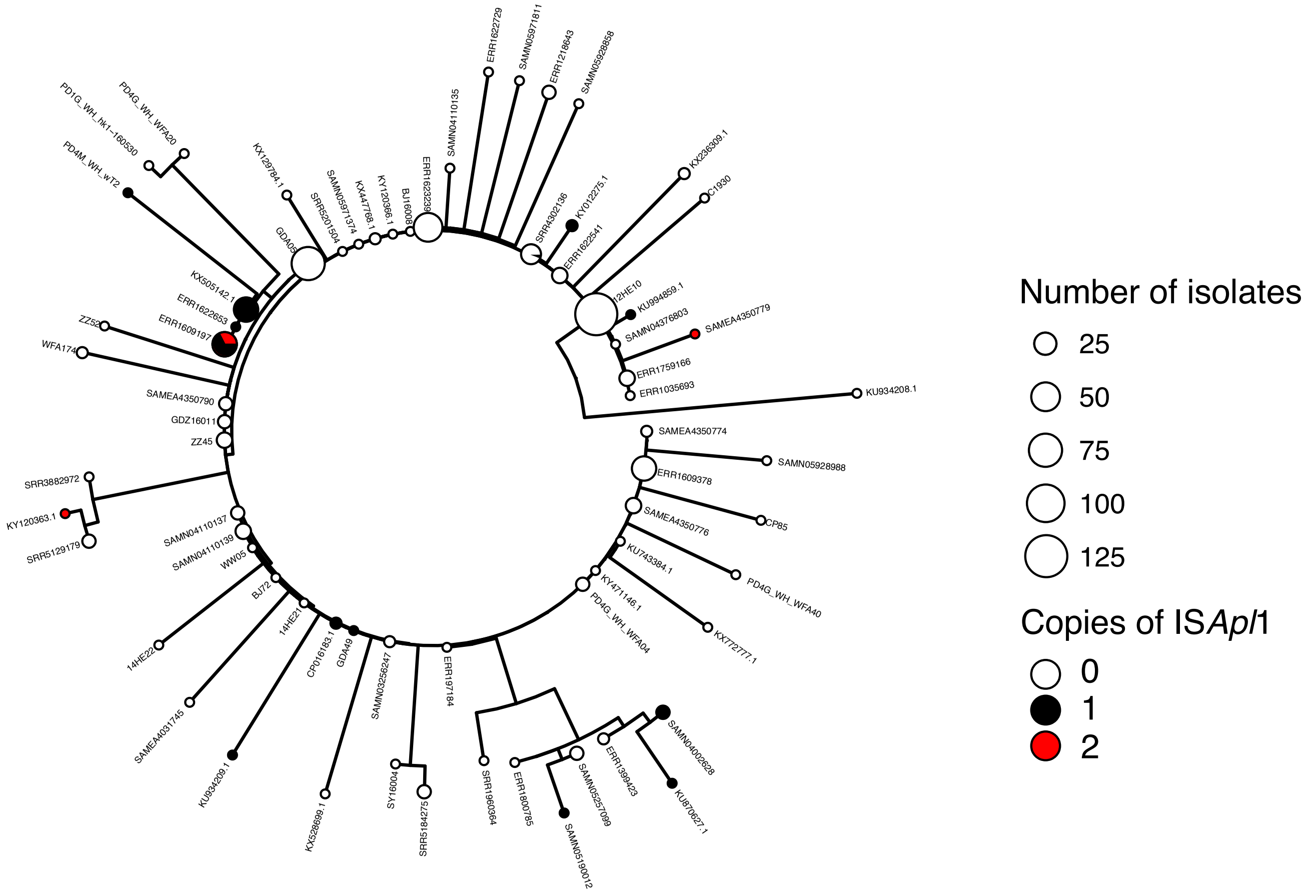


Figure 5.4: Phylogeny of the *mcr-1* composite transposon indicates a dominant sequence type with subsequent diversification. Midpoint-rooted maximum parsimony phylogeny based on the 3,522bp alignment of 457 sequences (recombinant regions removed). Size of points indicates the number of identical sequences, with a representative sequence for each shown next to each tip.

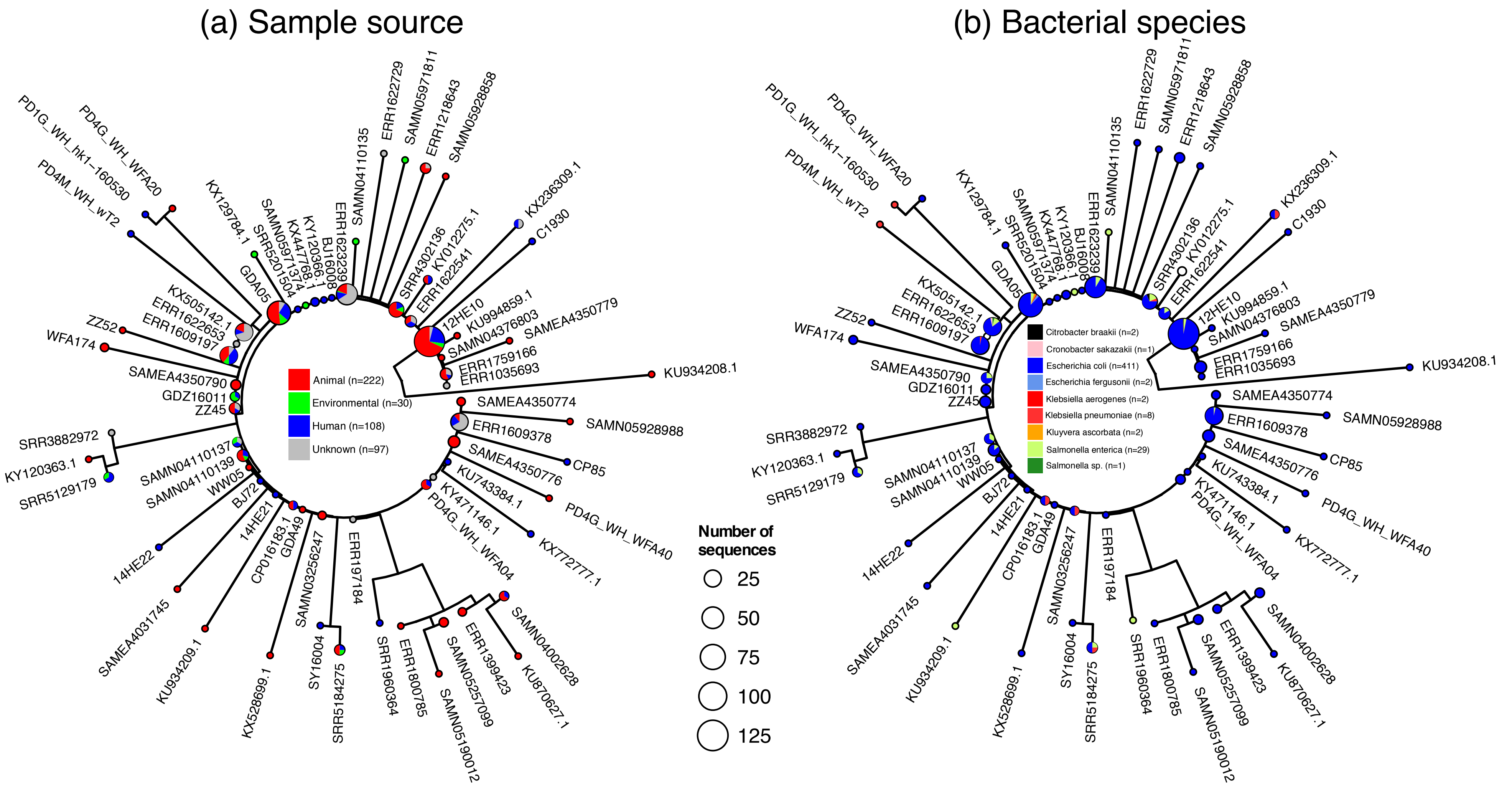


Figure 5.5: There is no clustering of (a) sample source or (b) bacterial species on the composite transposon phylogeny. Maximum parsimony tree (homoplastic sites removed, mid-point rooted, as in Figure 5.4). Size of points indicates the number of identical sequences, with a representative sequence for each shown next to each tip.

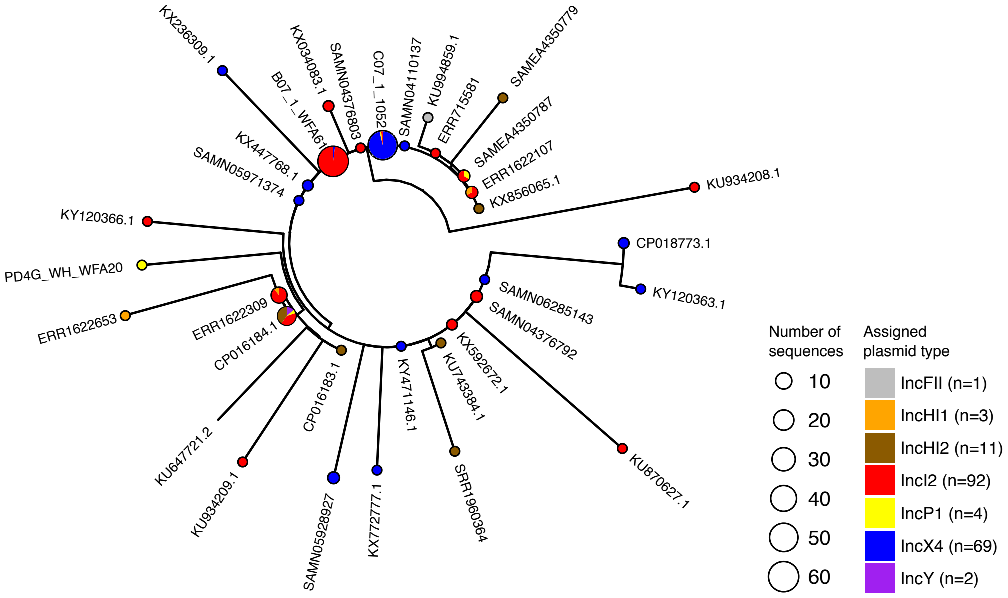


Figure 5.6: The distribution of plasmid types shown on the transposon phylogeny. Maximum parsimony tree (homoplastic sites removed, mid-point rooted, as in Figure 5.4) based on the composite transposon alignment for 172 sequences containing a plasmid replicon on the same contig i.e. those with an assigned plasmid type (color). IncI2 and IncX4 are the most common plasmid types. An example sequence ID is shown for each unique sequence.

Lucy van Dorp applied a Bayesian dating approach (BEAST) to infer a timed phylogeny of the maximal alignable region of the *mcr-1* carrying transposon (Section 5.2.5). Based on this 3,522 site alignment we inferred a common ancestor for 364 dated isolates in 2006 (Figure A.8; 2002-2008 95% HPD strict clock, coalescent model) with a mutation rate around 7.51×10^{-5} substitutions per site per year. There was no clear overall geographic clustering in the maximum clade credibility tree (Figure A.9).

5.3.4 Wider genomic background of *mcr-1*

The dataset also allowed the exploration of the wider genomic background upstream and downstream of the conserved transposon alignment. There were sufficiently long assembled contigs for 182 isolates to identify plasmid types based on co-occurrence with plasmid replicons (Section 5.2.1) and identified *mcr-1* in 13 different plasmid backgrounds. IncI2 and IncX4 were the dominant plasmid types, accounting for 51% and 38% of the isolates, respectively (Figure 5.6) similar to the proportions observed by Matamoros et al. (2017). One isolate in the dataset was definitively located on a complete chromosome, although we cannot rule out the presence of a few other chromosomal copies of *mcr-1* located on short contigs.

The distribution of transposons carrying one or two copies of IS*Apl1* was highly heterogeneous across these plasmid types. For example, sequences with one or two copies of IS*Apl1* were found on six and four types, respectively, which supports their mobility compared to those without ISApl1, which were found in five plasmid types. Of the contigs carrying one copy of IS*Apl1*, 61% were found in IncI2 plasmids, and 50% of contigs carrying two copies of IS*Apl1* belonged to IncHI2 plasmids. Conversely, the common IncX4 plasmids carried only two transposons with two copies of IS*Apl1* and none with a single copy of the element.

I identified two extended plasmid backbone sequences that could be aligned. The first such alignment encompassed a shared sequence of 7,161bp between 108 plasmid backgrounds and has been previously referred to as ‘Type A’ (Y. Wang, R. Zhang, et al., 2017). These sequences contain 54 sequences co-occurring with an IncI2 replicon, with 54 of unknown plasmid type, and encompass a large fraction of the genetic diversity found in the *mcr-1* transposon, although a large proportion (nine out of 108) belonged to the dominant sequence type. The second alignment was 34,761 bp long and was common to nine IncX4 plasmids and partly overlaps with a background previously defined as ‘Type D’ (Y. Wang, R. Zhang, et al., 2017).

Lucy van Dorp then applied BEAST to infer a timed phylogeny for each of these alignable regions after removal of SNPs showing evidence of recombination, and we jointly analyzed the results. For the IncI2 background we could infer that a common ancestor to all 108 isolates existed in 2006 (1998-2010 95% CI relaxed exponential clock model) assuming a constant population size model. For the IncX4 backgrounds we dated the common ancestor of the eight isolates to 2011 (2010-2013 95% CI relaxed exponential clock model) assuming a constant population size model. Posterior density distributions of root dating for these two alignments under different population and clock models are shown in Figure A.10. The difference in dating inferred for these two plasmid backgrounds and the recent date obtained for IncX4 highlight the dynamic nature of the integration of the *mcr-1* carrying transposon, even if in the IncX4 phylogeny isolates from East Asia and Europe and the Americas cluster together. The inferred mutation rates obtained for the IncI2 and IncX4 backgrounds consistently lie around $5-10\times10^{-5}$ substitutions per site per year, as did the rate for the composite transposon (Table A.3).

5.3.5 Environmental distribution of the composite transposon

It has been suggested that agricultural use of colistin – widespread in China since the early 1980s – caused the initial emergence and spread of *mcr-1* (Poirel et al., 2016; Schwarz and Johnson, 2016). According to the evolutionary model in Figure 5.2, the ancestral mobilizable state is represented by the transposon carrying both its IS*Apl1* elements. The transposon is thought to lose its capability for mobilization after the loss of both IS*Apl1* elements (Snesrud et al., 2016), although a single copy is reportedly sufficient to keep some ability to mobilize, with the upstream copy being functionally more important. I compared human ($n = 108$) and non-human ($n = 252$) isolates and found significantly more sequences with some trace of the insertion sequence IS*Apl1* both upstream and downstream in non-human isolates (32 out of 220 vs. 5 out of 108, χ^2 test, $p = 0.033$). This comparison held when only comparing agricultural isolates to human isolates ($n = 213$) (28 out of 213 vs. 5 out of 108, χ^2 test, $p = 0.029$). Furthermore, of the 42 isolates that had IS*Apl1* fragments both upstream and downstream, the majority were from Asia ($n = 30$) with only a quarter from Europe ($n = 10$) (χ^2 test, $p = 0.12$). This result was not driven by an over-representation of agricultural isolates from Asia in the dataset (χ^2 test, $p = 0.38$)

5.4 Discussion

5.4.1 Conclusions

In this chapter, I have described how I assembled a global dataset of 457 *mcr-1*-positive sequenced isolates and used this as the basis for a set of analyses to gain insight about the origins of *mcr-1*. This collaborative effort shows that there was a single integration event of *mcr-1* into an IS*Apl1* composite transposon, followed by its subsequent spread between multiple genomic backgrounds. Our phylogenetic analyses suggest an age of insertion of *mcr-1* into the gene transposon shared across our isolates in the mid 2000s (2002-2008 95% HPD). We could identify the likely sequence of the ancestral transposon type and show the pattern of diversity supports a single mobilization with subsequent diversification during global spread.

Despite the limited number of whole genome sequences for samples before 2012, with the oldest sequence available from 2008 (Figure 5.1C), our estimate is consistent with the majority of available evidence from retrospective surveillance data (Poirel et al., 2016) which has found the presence of *mcr-1* in samples dating back to 2005 in Europe (Haenni et al., 2016). One retrospective study of Chinese isolates from 1970-2014 reported three *mcr-1*-positive *E. coli* dating from the 1980s (Shen et al., 2016), although *mcr-1* then did not reappear until 2004. This observation seems surprising in light of our results, which clearly exclude such an early spread of *mcr-1*, at least on this IS*Apl1* transposon background. A constant population size model gave a better fit than an exponential model, suggesting the dramatic increase in reports of the presence of *mcr-1* across the past two years may not reflect a sudden global spread after its initial discovery (Liu et al., 2016) and highlighting the difficulty of interpreting novel surveillance data from previously unknown resistance elements.

Our estimates of the age of spread of the representative IncI2 and IncX4 plasmid backgrounds are more recent, dating to around 2008 and 2013, respectively, but are both consistent with the age of the transposon mobilization event. We did not constrain the evolutionary rates in any of our phylogenetic analyses. It is thus encouraging that the different rates are highly consistent between the *mcr-1* transposon and the two plasmid backgrounds. While this points to high internal consistency between our estimates, I was unable to find any previously published estimates for the evolutionary rate of bacterial plasmids to compare them to.

The current distribution and observed genetic patterns are in line with a centre of origin in China. This is the place where we observe the highest proportion of isolates carrying intact or partial copies of the IS*Apl1* flanking elements. Transposon sequences carrying IS*Apl1* elements were also overrepresented in environmental and agriculture isolates, relative to those collected form humans. This pattern is in line with agricultural settings acting as the source of *mcr-1* within bacteria isolated from humans (Y. Wang, Tian, et al., 2017). The current global distribution has been achieved through multiple translocations, and is illustrated by the interspersed geographic origins in our phylogenetic reconstructions. A likely driver for the global spread is trade, in particular food animals (Grami et al., 2016) and meat, although direct global movement by colonized or infected humans (Wintersdorff et al., 2016) is also likely to have played a role in the current distribution.

The origin of *mcr-1* prior to its mobilization remains elusive. Despite an exhaustive search of sequence repositories, including the SRA, I did not find a single *mcr-1* sequence outside the IS*Apl1* transposon background. IS*Apl1* was first identified in the pig pathogen *Actinobacillus pleuropneumoniae* (Tegetmeyer et al., 2008) suggesting that it may also have been an ancestral host for *mcr-1*, although to our knowledge no *mcr-1*-positive *A. pleuropneumoniae* isolates have been described. The phosphoethanolamine transferase (EptA) from *Paenibacillus sophorae* has also been proposed as a possible candidate (Gao et al., 2016). However, this seems most unlikely as *Paenibacili* are Gram positive and are thus intrinsically resistant to polymixins (Di Conza et al., 2017). Moreover, while the two sequences share functional similarities, this should be interpreted as a case of possible parallel evolution rather than direct filiation (Di Conza et al., 2017). *Moraxella* has been suggested as being the source of *mcr-1* (Kieffer et al., 2017), following the identification of genes in *Moraxella* with limited homology to *mcr-1* (60% nucleotide sequence identity). However, this sequence identity seems too low for *Moraxella* to be considered as true candidates for the origin of *mcr-1*. Until a sequence with high homology to *mcr-1* is identified outside of the IS*Apl1* sequence background, the search for its initial source remains open.[1]

5.4.2 Limitations

I aimed to assemble a comprehensive dataset for this work. However, since this work was completed many more sequences have been deposited in public databases. It would be a simple extension to use these sequences in a further analysis. A more difficult issue to circumvent is that the dataset is likely affected by complex sampling biases, with an overrepresentation of samples from places with active surveillance and well-funded research communities.

I investigated a single mobilized colistin resistance gene, the eponymous *mcr-1* that

[1] Since the original paper was published, a recent paper identified the likely origin of *mcr-1* as a novel species of *Moraxella* (Snesrud et al., 2018), based on the publication of an isolate containing a chromosomal region sharing >96% identity with the canonical cassette sequence (AbuOun et al., 2017). Thus, the event that is dated in this chapter is likely the original copy-out of this region and the integration into the composite transposon.

was first observed only two years ago. However, there are in fact several mobilized genes that have now been confirmed to confer colistin resistance, with *mcr-2* reported less than a year after *mcr-1* was initially described (Xavier et al., 2016) and more recently the phylogenetically distant *mcr-3* (Yin et al., 2017), *mcr-4* (Carattoli et al., 2017), and *mcr-5* (Borowiak et al., 2017) have also been described. There appear to be commonalities between the mechanisms of the *mcr* genes, despite their different sequences and location near to different insertion sequences. For example, *mcr-2* has 76.7% nucleotide identity to *mcr-1* and was found in colistin-resistant isolates that did not contain *mcr-1*, and appeared to be mobilized on a IS1595 transposon (Xavier et al., 2016). Despite the different insertion sequences, intriguingly, this mobile element also contained a similar protein downstream of the *mcr* gene. Indeed, in *mcr-1*, *-2* and *-3*, the *mcr* gene has a downstream open reading frame (ORF) encoding, respectively, a putative PAP2 protein (Snesrud et al., 2016), a PAP2 membrane-associated lipid phosphatase (Xavier et al., 2016), and a diacylglycerol kinase (Yin et al., 2017), all of which have transmembrane domains and are involved in the phosphatidic acid pathway (Athenstaedt and Daum, 1999; Epand et al., 2016). While the PAP2-like ORF in *mcr-1* has been shown to not be required for colistin resistance (Zurfluh et al., 2016), the presence of similar sequences downstream of other mcr genes implies some functional role, either in the formation of the mobile element and/or in its continued mobilization.

Finally, I did not investigate co-occurrence of other resistance genes with mcr-1, but many isolates in the dataset were resistant to several other antibiotics (particularly those from our Chinese collaborators). There is considerable evidence that epistasis of mutations is widespread for mutations (Durão et al., 2015) and this seems likely to also be true for resistance genes, both environmentally (B. Li et al., 2015) and also specifically on mobile genetic elements. Future work investigating these co-occurrences could give an interesting perspective on their concurrent spread.

5.4.3 Summary

I took responsibility for compiling the largest dataset to date of sequenced *mcr-1*-positive isolates, using a combination of collaborative sequencing efforts and an exhaustive search of sequences deposited on publicly available databases, including unassembled datasets from the SRA. This allowed me to obtain a truly global dataset of 457 *mcr-1*-positive isolates covering 31 countries and five continents, which formed the basis of a set of analyses investigating the spread of *mcr-1* on a composite transposon and plasmids. While the complex Russian doll dynamics of the transposon, plasmids, and bacterial host made it challenging to reach strong conclusions on some important aspects of the spread of *mcr-1*, these results nevertheless demonstrate the potential for phylogenetic reconstruction of antimicrobial resistance elements at a global scale, and highlight the relevance of a ‘one health’ perspective that makes use of all available isolates from multiple sources. The wider meta-community of the human microbiome is crucial for human health, with identical sequences found in Chinese agricultural isolates and the gut microbiome of individuals in Europe. Future efforts relying on more sophisticated computational tools and even more extensive genetic sequence data are likely to become part of the routine toolbox in infectious disease surveillance, improving our understanding of how ecological units can move between multiple nested genetic levels at a global scale.

Chapter 6

Conclusion

The bacterial communities associated with the human body are ecosystems that can be considered at multiple levels. In the introduction to this thesis I identified four challenges in understanding the role of bacterial communities in health and disease (Section 1.1.2) and explained how each of the chapters addressed these challenges directly using real datasets (Section 1.5). In this concluding section, I briefly summarise the contributions this thesis makes to our knowledge of human-associated bacterial communities before surveying the opportunities for future research. For a more detailed discussion of the findings and limitations of each piece of work, see the conclusions of each individual chapter.

6.1 Summary of findings

In Chapter 2 I performed the first simultaneous analysis of the effects of host genetics and shared environment on the salivary microbiome using whole-genome based measures of host genetic distance rather than pedigrees. Over 100 closely-related Ashkenazi individuals in this cohort lived across four global cities but shared a common lifestyle and cultural practices, meaning that confounding by other factors was presumed to be lower. My analysis showed conclusively that the dominant effect was from shared household rather than overall host genetic similarity. I found no geographical structuring at the level of cities, supporting previous claims that the core oral microbiome is conserved at a global scale. There was a persistent effect of parental household in individuals who had moved out of the familial home, suggesting that shared upbringing has a long-term impact on the oral microbiome over a period of years. Fine-scale differences that were observable at the sub-genus level between spouses suggest that regular environmental contact is important for maintaining a similar composition. This was supported by the observation that children under the age of ten shared phylotypes with their parents that were not seen in older children, who presumably interact less with their parents or leave the familial household more. Intriguingly, I found that using measures of relatedness based on the known pedigree gave a spurious signal from genetics, raising concerns about future microbiome studies investigating the effect of genetics in this way.

In Chapter 3 I used a large cross-sectional dataset of supragingival plaque samples from Malawian women to investigate associations between bacterial communities in supragingival plaque and periodontal disease, while controlling for demographic factors. Unlike other studies of periodontitis, I wanted to take advantage of the cross-sectional dataset that captured the landscape of periodontal disease, so I used a two-factor approach to investigate associations between gingivitis (bleeding), periodontitis (deepened pockets), and the relative abundances of bacterial taxa. I showed that the signals from these two related aspects of periodontal disease could be distinguished. Bacteria in the mouth exist in oral biofilms. Despite a lack of explicit data on this spatial structure, I was able to extract it in a hypothesis-free manner using correlations in the relative abundances of disease-associated taxa. I used a simple measure of centrality from social network analysis to rank taxa in this periodontitis-associated network and showed that the results were consistent with experimental investigation of periodontal biofilms, identifying known bridging bacteria. The structure of this periodontitis-associated network was different between women with and without periodontitis across the range of gingivitis severities, showing that community structure associated with a subgingival condition is detectable using supragingival samples.

In Chapter 4 I developed a new model of antibiotic perturbation for the gut microbiome. In contrast to complex models built up mechanistically using pairwise interactions between species, I adopted a highly simplified top-down approach. I made minimal assumptions based on a popular heuristic of the gut microbiome resting in a stability landscape. I proved that this model could describe the time-response of the gut microbiome to a short course of antibiotics by reanalyzing data from Zaura et al. (2015) where individuals took widely-used antibiotics, suggesting that viewing the gut microbiome as a damped harmonic oscillator is a valid model with predictive power. I also introduced a variant of the model that allowed for a transition to a different equilibrium within the stability landscape. This model was better supported than the model without a transition, suggesting that indeed the microbiome had been altered long-term by antibiotics.

Finally, in Chapter 5 I reconstructed the evolutionary history of a small transposon carrying a resistance gene (*mcr-1*) using phylogenetic approaches on a global whole genome sequencing dataset. This work represents perhaps the first such use of phylogenetic reconstruction tools for a mobile genetic element in this way. I combined whole genome sequences from a range of different sources and was able to identify a consistent small genomic region of 3,500bp within every bacterial genome sequenced. This allowed an estimate to be made for the origin of the *mcr-1* composite IS*Apl1* transposon, which we could date to the early 2000s.

6.2 Future approaches

The human microbiome is of great interest in its own right as an ecosystem, but the majority of research is implicitly aimed at improving health. Understanding the role that human-associated bacterial communities really play in health and disease requires putting them in context, understanding the factors that govern their variation between individuals and their stability over time before identifying features that are associated with disease (see the list of challenges I gave in Section 1.1.2). One of the biggest obstacles faced by current microbiome research is the lack of a well-defined understanding of what is clinically important or useful. Ultimately, the microbiome is another factor to be incorporated into clinical and epidemiological models that contain as many other factors as possible, as I have attempted in this thesis (e.g. in Chapter 3). In some cases it will be important, and in others of little consequence. Integrating microbiome information into mathematical models containing other clinical data is required to begin to understand the role that the microbiome plays in health. The stability landscape approach I adopted in Chapter 4 is inherently flexible and could be applied to other bacterial communities. It could also be extended in other directions. For example, fitting a perturbation model to multiple bacterial families would allow the construction of a multidimensional vector of parameters, and this vector could then be used to parameterise the stability landscape in some multidimensional space.[1]

Rhetoric around the problem of AMR mainly focuses on antimicrobial stewardship: coordinated approaches to reduce the prescription and use of antimicrobials. There is little focus on public perception of antibiotics as having a possible detrimental impact on the microbiome. Perhaps this is due to the rhetoric around antibiotics as ‘miracle drugs’, making us reluctant to accept the possibility that antibiotics can both be life-saving in certain situations and harmful in others. In a real sense, antibiotics do not treat the host; they treat their associated bacterial communities, making them substantially unlike other classes of drugs. Harm – both individual and collective – is a real possibility where antibiotics are prescribed unnecessarily. Patients are familiar with the need to balance necessity-concern relationships in the context of other medications, with a prevailing belief that we should take as little medication as possible (Horne et al., 2013). The lack of investigation and communication of the personal costs of antibiotics has likely played a role in the public perception of them as only having societal costs. One powerful inducement to future behavioural change would be to challenge the narrative around antibiotics as drugs that cannot confer harm to the patient. The presentation of particular data about antibiotics can dramatically change public perceptions and bring about behavioural change, which brings with it a large ethical responsibility to summarise data accurately and correctly.

The *mcr-1* transposon is not a species, but it is a genomic unit that corresponds to the

[1] I am grateful to Sarah Walker for this suggestion.

phenotype of colistin resistance. It is therefore in some sense the fundamental ecotype to deal with when addressing the problem of mobilized colistin resistance. I believe it is coherent to present phylogenetic and other analyses of small genetic elements, even if such analsyes are more familiar when applied to whole genomes. As datasets become more comprehensive, and searching public databases with *k*-mer methods becomes easier, I believe the type of analysis presented in Chapter 5 will become much more common and an important part of the toolbox for understanding the spread of antimicrobial resistance. Also, an ecological view of these units encourages the application of existing methods; the wheel does not have to be re-invented. For example, correlation network approaches like the ones I used in Chapter 3 could also be used to look at the co-occurrence of *mcr-1* with other resistance genes on plasmids, which could identify potential epistatic interactions and central genes.

The portability of analysis techniques touches on a recurring theme of this thesis: the similarity between the different levels of nested genetic diversity in bacterial communities. I wish to finish with a few observations on this diversity, which is one of the remarkable features of microbial ecology. The diversity of communities is a widely-used summary property in ecology, with many different metrics devoted to capturing it. However, any definition of diversity depends on a definition of an ecological unit. In classical ecology, it is usually fairly simple to determine (a) how many species are present in a sample and (b) how many of each sort of species, because larger organisms have strict species definitions that are observable 'at a glance'. However, when considering bacterial communities there is great debate about what constitutes a reliable ecological unit, particularly when using sequencing approaches.

Different choices for the level of sequencing will lead to different appreciation of the diversity: using a small region of the 16S rRNA gene to cluster bacterial species will give a reduced estimate of the total diversity compared to sequencing the whole gene, which will in turn give a reduced estimate compared to sequencing all genomes present with a shotgun metagenomic approach. Thus, to speak of 'the' diversity of a microbial community is usually meaningless in itself, because our choice of measuring stick has such a large impact on the diversity we calculate. The fractality of bacterial genomes has been noted by others (Koonin, 2012). Consider the apparently trivial question raised by Mandelbrot (1967): "How long is the coast of Britain?" Satisfactorily answering that question required developing a definition for the fractal dimension of a curve, allowing meaningful discussion of its length. An interesting line of future research would be to develop a mathematical definition of an analogous quantity to fractal dimension, in order to talk meaningfully about the diversity of bacterial communities. If we treat them as hierarchies of fractal structures, the analogy demonstrates the extraordinary richness that can be progressively revealed by deeper sequencing technologies.

The future offers many exciting possibilities for investigating the remarkable nested genetic diversity of bacterial communities, including those associated with the human body. In this thesis I have outlined how considering the communities involved at multiple ecological levels can give different insights. Carl Woese, who pioneered the use of 16S rRNA sequencing for microbial ecology (Section 1.1.3), wrote later in his career that even the genome itself is “a set of one-dimensional ecosystems” (Goldenfeld and Woese, 2011). To conclude, I wish to point out that all that is needed for an ecological approach is:

- A definition of a unit
- A method to discriminate these units into different varieties
- A method to count the varieties of units

These units are themselves made up of other smaller units, producing a self-similarity across levels and a certain degree of self-reference (Shaw, 2018). Even clonal populations of *Escherichia coli* subject to fixed environmental conditions can still give rise to complex ecological interactions that are detectable only with whole genome sequencing (Good et al., 2017). In other words, long-term temporal stability can be a complex and dynamic process. The appropriate resolution to view the system at depends on the question being asked. One might ask: where is evolution in this ecological picture? Goldenfeld and Woese (2011) write about the central aspect of evolution:

> It is a process that continually expands the space in which it operates through a dynamic that is essentially self-referential. . . self-reference arises because the biological components of interest are emergent, and we are seeking a description of biological phenomena in terms of these biological components only.

In this thesis I have demonstrated that ecological approaches can be applied to human-associated bacterial communities at multiple scales. Statistical methods that operate on ecological units inferred from sequencing data are widely applicable, whether we choose to treat a given bacterial community as: an interaction network of marker gene phylotypes, a single metric representing displacement from equilibrium, or a collection of genomic backgrounds for a mobile genetic element. The fascinating nested genetic complexity of bacterial communities means that within the fundamental units of a particular analysis, there is undoubtedly still more diversity to discover.

Bibliography

Abeles, S. R., M. B. Jones, T. M. Santiago-Rodriguez, M. Ly, N. Klitgord, S. Yooseph, K. E. Nelson, and D. T. Pride (2016). Microbial diversity in individuals and their household contacts following typical antibiotic courses. *Microbiome* **4**(1), 39. doi: 10.1186/s40168-016-0187-9.

Abeles, S. R., R. Robles-Sikisaka, M. Ly, A. G. Lum, J. Salzman, T. K. Boehm, and D. T. Pride (2014). Human oral viruses are personal, persistent and gender-consistent. *The ISME Journal* **8**(9), 1753–1767. doi: 10.1038/ismej.2014.31.

AbuOun, M. et al. (2017). *mcr-1* and *mcr-2* variant genes identified in *Moraxella* species isolated from pigs in Great Britain from 2014 to 2015. *Journal of Antimicrobial Chemotherapy* **72**(10), 2745–2749. doi: 10.1093/jac/dkx286.

Abusleme, L., A. K. Dupuy, N. Dutzan, N. Silva, J. A. Burleson, L. D. Strausbaugh, J. Gamonal, and P. I. Diaz (2013). The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *The ISME Journal* **7**(5), 1016–1025. doi: 10.1038/ismej.2012.174.

Adair, K. L. and A. E. Douglas (2017). Making a microbiome: the many determinants of host-associated microbial community composition. *Current Opinion in Microbiology* **35**, 23–29. doi: 10.1016/J.MIB.2016.11.002.

Adler, C. J. et al. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics* **45**(4), 450–455. doi: 10.1038/ng.2536.

Adriaens, L. M., R. Alessandri, S. Spörri, N. P. Lang, and G. R. Persson (2009). Does pregnancy have an impact on the subgingival microbiota? *Journal of Periodontology* **80**(1), 72–81. doi: 10.1902/jop.2009.080012.

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 111–142. doi: 10.2307/2345730.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**(6), 716–723. doi: 10.1109/TAC.1974.1100705.

Alneberg, J., B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**(11), 1144–1146. doi: 10.1038/nmeth.3103.

Aminov, R. I. (2011). Horizontal gene exchange in environmental microbiota. *Frontiers in Microbiology* **2**, 158. doi: 10.3389/fmicb.2011.00158.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**(1), 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x.

Antipov, D., N. Hartwick, M. Shen, M. Raiko, A. Lapidus, and P. A. Pevzner (2016). plasmidSPAdes: assembling plasmids from whole genome sequencing data. *Bioinformatics* **32**(22), 3380–3387. doi: 10.1093/bioinformatics/btw493.

Aroniadis, O. C. and L. J. Brandt (2014). Intestinal microbiota and the efficacy of fecal microbiota transplantation in gastrointestinal disease. *Gastroenterology & Hepatology* **10**(4), 230–237. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4073534/.

Aruni, A. W., Y. Dou, A. Mishra, and H. M. Fletcher (2015). The biofilm community rebels with a cause. *Current Oral Health Reports* **2**(1), 48–56. doi: 10.1007/s40496-014-0044-5.

Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman (2005). At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology* **71**(12), 7724–7736. doi: 10.1128/AEM.71.12.7724-7736.2005.

Ashorn, P. et al. (2015). The impact of lipid-based nutrient supplement provision to pregnant women on newborn size in rural Malawi: a randomized controlled trial. *American Journal of Clinical Nutrition* **101**(2), 387–397. doi: 10.3945/ajcn.114.088617.

Atarashi, K. et al. (2017). Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. *Science* **358**(6361), 359–365. doi: 10.1126/science.aan4526.

Athenstaedt, K. and G. Daum (1999). Phosphatidic acid, a key intermediate in lipid metabolism. *European Journal of Biochemistry* **266**(1), 1–16. doi: 10.1046/j.1432-1327.1999.00822.x.

Baele, G., P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* **29**(9), 2157–2167. doi: 10.1093/molbev/mss084.

Baelum, V. and F. Scheutz (2002). Periodontal diseases in Africa. *Periodontology 2000* **29**(1), 79–103. doi: 10.1034/j.1600-0757.2002.290105.x.

Baltrus, D. A. (2016). Divorcing strain classification from species names. *Trends in Microbiology* **24**(6), 431–439. doi: 10.1101/037325.

Barroso-Batista, J., A. Sousa, M. Lourenço, M.-L. Bergman, D. Sobral, J. Demengeot, K. B. Xavier, and I. Gordo (2014). The first steps of adaptation of Escherichia coli to the gut are dominated by soft sweeps. *PLOS Genetics* **10**(3), e1004182. doi: 10.1371/journal.pgen.1004182.

Batchelor, P. (2014). Is periodontal disease a public health problem? *British Dental Journal* **217**(8), 405–409. doi: 10.1038/sj.bdj.2014.912.

Belibasakis, G. N. and N. Bostanci (2012). The RANKL-OPG system in clinical periodontology. *Journal of Clinical Periodontology* **39**(3), 239–48. doi: 10.1111/j.1600-051X.2011.01810.x.

Belstrøm, D., P. Holmstrup, A. Bardow, A. Kokaras, N.-E. Fiehn, and B. J. Paster (2016). Temporal stability of the salivary microbiota in oral health. *PLOS ONE* **11**(1), e0147472. doi: 10.1371/journal.pone.0147472.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300. doi: 10.2307/2346101.

Bergan, T., S. B. Thorsteinsson, R. Solberg, L. Bjornskau, I. M. Kolstad, and S. Johnsen (1987). Pharmacokinetics of ciprofloxacin: intravenous and increasing oral doses. *The American Journal of Medicine* **82**(4A), 97–102. url: http://www.ncbi.nlm.nih.gov/pubmed/3578334.

Blekhman, R. et al. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biology* **16**(1), 191. doi: 10.1186/s13059-015-0759-1.

Blommaert, A., C. Marais, N. Hens, S. Coenen, A. Muller, H. Goossens, and P. Beutels (2014). Determinants of between-country differences in ambulatory antibiotic use and antibiotic resistance in Europe: a longitudinal observational study. *Journal of Antimicrobial Chemotherapy* **69**(2), 535–547. doi: 10.1093/jac/dkt377.

Bohnhoff, M. and C. P. Miller (1962). Enhanced susceptibility to Salmonella infection in streptomycin-treated mice. *The Journal of Infectious Diseases* **111**, 117–127. url: http://www.ncbi.nlm.nih.gov/pubmed/13968487.

Bonder, M. J., S. Abeln, E. Zaura, and B. W. Brandt (2012). Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* **28**(22), 2891–2897. doi: 10.1093/bioinformatics/bts552.

Bonder, M. J. et al. (2016). The effect of host genetics on the gut microbiome. *Nature Genetics* **48**(11), 1407–1412. doi: 10.1038/ng.3663.

Borowiak, M., J. Fischer, J. A. Hammerl, R. S. Hendriksen, I. Szabo, and B. Malorny (2017). Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in d-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B. *Journal of Antimicrobial Chemotherapy* **72**(12), 3317–3324. doi: 10.1093/jac/dkx327.

Bouckaert, R., J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* **10**(4), e1003537. doi: 10.1371/journal.pcbi.1003537.

Brading, M. G. and P. D. Marsh (2003). The oral environment: the challenge for antimicrobials in oral care products. *International Dental Journal* **53**(S6P1), 353–62. url: http://www.ncbi.nlm.nih.gov/pubmed/14725379.

Bradley, P., H. den Bakker, E. Rocha, G. McVean, and Z. Iqbal (2017). Real-time search of all bacterial and viral genomic data. *bioRxiv*, 234955. doi: 10.1101/234955.

Bradshaw, D. J., P. D. Marsh, G. K. Watson, and C. Allison (1998). Role of Fusobacterium nucleatum and coaggregation in anaerobe survival in planktonic and biofilm oral microbial communities during aeration. *Infection and Immunity* **66**(10), 4729–4732. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC108582/.

Bradshaw, D. J. and R. J. M. Lynch (2013). Diet and the microbial aetiology of dental caries: new paradigms. *International Dental Journal* **63**(S2), 64–72. doi: 10.1111/idj.12082.

Bray, J. R. and J. T. Curtis (1957). An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs* **27**(4), 325–349. doi: 10.2307/1942268.

Brown, M. V. et al. (2012). Global biogeography of SAR11 marine bacteria. *Molecular Systems Biology* **8**(1), 595. doi: 10.1038/msb.2012.28.

Cai, L., L. Ye, A. H. Y. Tong, S. Lok, and T. Zhang (2013). Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. *PLOS ONE* **8**(1), e53649. doi: 10.1371/journal.pone.0053649.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421. doi: 10.1186/1471-2105-10-421.

Camelo-Castillo, A., L. Novoa, C. Balsa-Castro, J. Blanco, A. Mira, and I. Tomás (2015). Relationship between periodontitis-associated subgingival microbiota and clinical inflammation by 16S pyrosequencing. *Journal of Clinical Periodontology* **42**(12), 1074–1082. doi: 10.1111/jcpe.12470.

Caporaso, J. G. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**(5), 335–336. doi: 10.1038/nmeth.f.303.

Caporaso, J. G. et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**(8), 1621–1624. doi: 10.1038/ismej.2012.8.

Carattoli, A. (2013). Plasmids and the spread of resistance. *International Journal of Medical Microbiology* **303**(6-7), 298–304. doi: 10.1016/J.IJMM.2013.02.001.

Carattoli, A., L. Villa, C. Feudi, L. Curcio, S. Orsini, A. Luppi, G. Pezzotti, and C. F. Magistrali (2017). Novel plasmid-mediated colistin resistance *mcr-4* gene in Salmonella and Escherichia coli, Italy 2013, Spain and Belgium, 2015 to 2016. *Eurosurveillance* **22**(31), 30589. doi: 10.2807/1560-7917.ES.2017.22.31.30589.

Carattoli, A., E. Zankari, A. García-Fernández, M. Voldby Larsen, O. Lund, L. Villa, F. Møller Aarestrup, and H. Hasman (2014). *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy* **58**(7), 3895–3903. doi: 10.1128/AAC.02412-14.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: a probabilistic programming language. *Journal of Statistical Software* **76**(1), 1–32. doi: 10.18637/jss.v076.i01.

Carrillo-de-Albornoz, A., E. Figuero, D. Herrera, and A. Bascones-Martínez (2010). Gingival changes during pregnancy: II. Influence of hormonal variations on the subgingival biofilm. *Journal of Clinical Periodontology* **37**(3), 230–240. doi: 10.1111/j.1600-051X.2009.01514.x.

Cava, A. L. and G. Matarese (2004). The weight of leptin in immunity. *Nature Reviews Immunology* **4**(5), 371–379. doi: 10.1038/nri1350.

Cephas, K. D., J. Kim, R. A. Mathai, K. A. Barry, S. E. Dowd, B. S. Meline, and K. S. Swanson (2011). Comparative analysis of salivary bacterial microbiome diversity in edentulous infants and their mothers or primary care givers using pyrosequencing. *PLOS ONE* **6**(8), e23503. doi: 10.1371/journal.pone.0023503.

Chen, H., Y. Liu, M. Zhang, G. Wang, Z. Qi, L. Bridgewater, L. Zhao, Z. Tang, and X. Pang (2015). A *Filifactor alocis*-centered co-occurrence group associates with periodontitis across different oral habitats. *Scientific Reports* **5**, 9053. doi: 10.1038/srep09053.

Chen, T., W.-H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, and F. E. Dewhirst (2010). The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* **2010**(0), baq013. doi: 10.1093/database/baq013.

Chu, D. M., J. Ma, A. L. Prince, K. M. Antony, M. D. Seferovic, and K. M. Aagaard (2017). Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nature Medicine* **23**(3), 314–326. doi: 10.1038/nm.4272.

Chung, H. et al. (2012). Gut immune maturation depends on colonization with a host-specific microbiota. *Cell* **149**(7), 1578–1593. doi: 10.1016/j.cell.2012.04.037.

Clarridge, J. E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews* **17**(4), 840–862. doi: 10.1128/CMR.17.4.840-862.2004.

Cleary, B., I. L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, and E. J. Alm (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology* **33**(10), 1053–1060. doi: 10.1038/nbt.3329.

Cochran, D. L. (2008). Inflammation and bone loss in periodontal disease. *Journal of Periodontology* **79**(8s), 1569–1576. doi: 10.1902/jop.2008.080233.

Cordain, L., S. B. Eaton, A. Sebastian, N. Mann, S. Lindeberg, B. A. Watkins, J. H. O'Keefe, and J. Brand-Miller (2005). Origins and evolution of the Western diet: health implications for the 21st century. *The American Journal of Clinical Nutrition* **81**(2), 341–354. url: http://www.ncbi.nlm.nih.gov/pubmed/15699220.

Cosgrove, S. E., E. Avdic, K. Dzintars, and J. Smith (2015). *Antibiotic guidelines 2015-2016: treatment recommendations for adult inpatients*. url: https://www.hopkinsmedicine.org/amp/guidelines/antibiotic_guidelines.pdf.

Costello, E. K., K. Stagaman, L. Dethlefsen, B. J. M. Bohannan, and D. A. Relman (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* **336**(6086), 1255–1262. doi: 10.1126/science.1224203.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal (Complex Systems)*, 1695. url: http://igraph.org.

Cui, L., A. Morris, and E. Ghedin (2013). The human mycobiome in health and disease. *Genome Medicine* **5**(7), 63. doi: 10.1186/gm467.

David, L. A., A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biology* **15**(7), R89. doi: 10.1186/gb-2014-15-7-r89.

David, L. A. et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**(7484), 559–563. doi: 10.1038/nature12820.

Dethlefsen, L. and D. A. Relman (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *PNAS* **108**(Supplement 1), 4554–4561. doi: 10.1073/pnas.1000087107.

Dethlefsen, L., S. Huse, M. L. Sogin, and D. A. Relman (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLOS Biology* **6**(11), e280. doi: 10.1371/journal.pbio.0060280.

Dewhirst, F. E., T. Chen, J. Izard, B. J. Paster, A. C. R. Tanner, W.-H. Yu, A. Lakshmanan, and W. G. Wade (2010). The human oral microbiome. *Journal of Bacteriology* **192**(19), 5002–5017. doi: 10.1128/JB.00542-10.

Di Conza, J. A., M. A. Radice, and G. O. Gutkind (2017). mcr-1: rethinking the origin. *International Journal of Antimicrobial Agents* **50**(6), 737. doi: 10.1016/j.ijantimicag.2017.06.003.

Didelot, X. and D. J. Wilson (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLOS Computational Biology* **11**(2), e1004041. doi: 10.1371/journal.pcbi.1004041.

Donsì, F. and G. Ferrari (2016). Essential oil nanoemulsions as antimicrobial agents in food. *Journal of Biotechnology* **233**, 106–120. doi: 10.1016/j.jbiotec.2016.07.005.

Doron, S. and L. E. Davidson (2011). Antimicrobial stewardship. *Mayo Clinic Proceedings* **86**(11), 1113–1123. doi: 10.4065/mcp.2011.0358.

Doyle, R. M., D. G. Alber, H. E. Jones, K. Harris, F. Fitzgerald, D. Peebles, and N. Klein (2014). Term and preterm labour are associated with distinct microbial community structures in placental membranes which are independent of mode of delivery. *Placenta* **35**(12), 1099–1101. doi: 10.1016/j.placenta.2014.10.007.

Doyle, R. (2016). Placental, oral and vaginal microbiomes and birth outcomes in rural Malawi. PhD thesis. UCL. url: http://discovery.ucl.ac.uk/1475188/1/Thesis_final.pdf.

Drummond, A. J., M. A. Suchard, D. Xie, and A. Rambaut (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**(8), 1969–1973. doi: 10.1093/molbev/mss075.

Du, H., L. Chen, Y.-W. Tang, and B. N. Kreiswirth (2016). Emergence of the *mcr-1* colistin resistance gene in carbapenem-resistant *Enterobacteriaceae*. *The Lancet Infectious Diseases* **16**(3), 287–288. doi: 10.1016/S1473-3099(16)00056-6.

Durão, P., S. Trindade, A. Sousa, and I. Gordo (2015). Multiple resistance at no cost: rifampicin and streptomycin a dangerous liaison in the spread of antibiotic resistance. *Molecular Biology and Evolution* **32**(10), 2675–2680. doi: 10.1093/molbev/msv143.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19), 2460–2461. doi: 10.1093/bioinformatics/btq461.

— (2017). UNBIAS: An attempt to correct abundance bias in 16S sequencing, with limited success. *bioRxiv*, 124149. doi: 10.1101/124149.

Epand, R. M., C. Walker, R. F. Epand, and N. A. Magarvey (2016). Molecular mechanisms of membrane targeting antibiotics. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1858**(5), 980–987. doi: 10.1016/j.bbamem.2015.10.018.

Epskamp, S., A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom (2012). qgraph: network visualizations of relationships in psychometric data. *Journal of Statistical Software* **48**(4), 1–18. doi: 10.18637/jss.v048.i04.

Eren, A. M., G. G. Borisy, S. M. Huse, and J. L. Mark Welch (2014). Oligotyping analysis of the human oral microbiome. *PNAS* **111**(28), E2875–84. doi: 10.1073/pnas.1409644111.

Eren, A. M., L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin (2013). Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution* **4**(12). doi: 10.1111/2041-210X.12114.

Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin (2014). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal* **9**(4), 968–979. doi: 10.1038/ismej.2014.195.

EUCAST (2017). *Clinical breakpoints*. url: http://www.eucast.org/clinical_breakpoints (visited on 12/16/2017).

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**(1), 1–10. doi: 10.1016/0006-3207(92)91201-3.

Falgenhauer, L. et al. (2016). Colistin resistance gene mcr-1 in extended-spectrum β-lactamase-producing and carbapenemase-producing Gram-negative bacteria in Germany. *The Lancet Infectious Diseases* **16**(3), 282–283. doi: 10.1016/S1473-3099(16)00009-8.

Faust, K. and J. Raes (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology* **10**(8), 538–550. doi: 10.1038/nrmicro2832.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower (2012). Microbial co-occurrence relationships in the human microbiome. *PLOS Computational Biology* **8**(7), e1002606. doi: 10.1371/journal.pcbi.1002606.

Fernandes, M. R., F. P. Sellera, F. Esposito, C. P. Sabino, L. Cerdeira, and N. Lincopan (2017). Colistin-resistant mcr-1-positive *Escherichia coli* on public beaches, an infectious threat emerging in recreational waters. *Antimicrobial Agents and Chemotherapy* **61**(7), e00234–17. doi: 10.1128/AAC.00234-17.

Fine, D. H., K. Markowitz, K. Fairlie, D. Tischio-Bereski, J. Ferrendiz, D. Furgang, B. J. Paster, and F. E. Dewhirst (2013). A consortium of Aggregatibacter actinomycetemcomitans, Streptococcus parasanguinis, and Filifactor alocis is present in sites prior to bone loss in a longitudinal study of localized aggressive periodontitis. *Journal of Clinical Microbiology* **51**(9), 2850–2861. doi: 10.1128/JCM.00729-13.

Flores, G. E. et al. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biology* **15**(12), 531. doi: 10.1186/s13059-014-0531-y.

Ford, C. B., R. R. Shah, M. K. Maeda, S. Gagneux, M. B. Murray, T. Cohen, J. C. Johnston, J. Gardy, M. Lipsitch, and S. M. Fortune (2013). Mycobacterium tuberculosis mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature Genetics* **45**(7), 784–790. doi: 10.1038/ng.2656.

Foster, J. S. and P. E. Kolenbrander (2004). Development of a multispecie oral bacterial community in a saliva-conditioned flow cell. *Applied and Environmental Microbiology* **70**(7), 4340–4348. doi: 10.1128/AEM.70.7.4340-4348.2004.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry* **40**(1), 35. doi: 10.2307/3033543.

Friedman, J. and E. J. Alm (2012). Inferring correlation networks from genomic survey data. *PLOS Computational Biology* **8**(9), e1002687. doi: 10.1371/journal.pcbi.1002687.

Frost, L. S., R. Leplae, A. O. Summers, and A. Toussaint (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* **3**(9), 722–732. doi: 10.1038/nrmicro1235.

Fruchterman, T. M. J. and E. M. Reingold (1991). Graph drawing by force-directed placement. *Software: Practice and Experience* **21**(11), 1129–1164. doi: 10.1002/spe.4380211102.

Fujiwara, N. et al. (2015). Significant increase of oral bacteria in the early pregnancy period in Japanese women. *Journal of Investigative and Clinical Dentistry*. doi: 10.1111/jicd.12189.

Furi, L., R. Haigh, Z. J. H. Al Jabri, I. Morrissey, H.-Y. Ou, R. León-Sampedro, J. L. Martinez, T. M. Coque, and M. R. Oggioni (2016). Dissemination of novel antimicrobial resistance mechanisms through the insertion sequence mediated spread of metabolic genes. *Frontiers in Microbiology* **7**, 1008. doi: 10.3389/fmicb.2016.01008.

Galimanas, V., M. W. Hall, N. Singh, M. D. J. Lynch, M. Goldberg, H. Tenenbaum, D. G. Cvitkovitch, J. D. Neufeld, and D. B. Senadheera (2014). Bacterial community composition of chronic periodontitis and novel oral sampling sites for detecting disease indicators. *Microbiome* **2**(1), 32. doi: 10.1186/2049-2618-2-32.

Gao, R. et al. (2016). Dissemination and mechanism for the mcr-1 colistin Rresistance. *PLOS Pathogens* **12**(11), e1005957. doi: 10.1371/journal.ppat.1005957.

Garud, N. R., B. H. Good, O. Hallatschek, and K. S. Pollard (2017). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *bioRxiv*, 210955. doi: 10.1101/210955.

Gevers, D., R. Knight, J. F. Petrosino, K. Huang, A. L. McGuire, B. W. Birren, K. E. Nelson, O. White, B. A. Methé, and C. Huttenhower (2012). The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biology* **10**(8), e1001377. doi: 10.1371/journal.pbio.1001377.

Goldenfeld, N. and C. R. Woese (2011). Life is physics: evolution as a collective phenomenon far from equilibrium. *Annual Review of Condensed Matter Physics* **2**, 375–399. doi: 10.1146/annurev-conmatphys-062910-140509.

Gonzales-Marin, C., D. A. Spratt, and R. P. Allaker (2013). Maternal oral origin of *Fusobacterium nucleatum* in adverse pregnancy outcomes as determined using the 16S-23S rRNA gene intergenic transcribed spacer region. *Journal of Medical Microbiology* **62**(Pt$_1$), 133–144. doi: 10.1099/jmm.0.049452-0.

Good, B. H., M. J. McDonald, J. E. Barrick, R. E. Lenski, and M. M. Desai (2017). The dynamics of molecular evolution over 60,000 generations. *Nature* **551**(7678), 45–50. doi: 10.1038/nature24287.

Goodsell, D. S. (2017). *Animation of the small subunit of the Thermus thermophilus ribosome*. url: https://commons.wikimedia.org/w/index.php?curid=2839678 (visited on 11/25/2017).

Goossens, H., M. Ferech, R. Vander Stichele, M. Elseviers, and ESAC Project Group (2005). Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *The Lancet* **365**(9459), 579–587. doi: 10.1016/S0140-6736(05)17907-0.

Grami, R., W. Mansour, W. Mehri, O. Bouallègue, N. Boujaâfar, J.-Y. Madec, and M. Haenni (2016). Impact of food animal trade on the spread of *mcr-1* -mediated colistin resistance, Tunisia, July 2015. *Eurosurveillance* **21**(8), 30144. doi: 10.2807/1560-7917.ES.2016.21.8.30144.

Grégoire, N., V. Aranzana-Climent, S. Magréault, S. Marchand, and W. Couet (2017). Clinical pharmacokinetics and pharmacodynamics of colistin. *Clinical Pharmacokinetics*. doi: 10.1007/s40262-017-0561-1.

Griffen, A. L., C. J. Beall, N. D. Firestone, E. L. Gross, J. M. Difranco, J. H. Hardman, B. Vriesendorp, R. A. Faust, D. A. Janies, and E. J. Leys (2011). CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. *PLOS ONE* **6**(4), e19051. doi: 10.1371/journal.pone.0019051.

Gronau, Q. F., H. Singmann, and E.-J. Wagenmakers (2017). bridgesampling: an R package for estimating normalizing constants. *arXiv*, 1710.08162. url: http://arxiv.org/abs/1710.08162.

Guay, D. (2007). Update on clindamycin in the management of bacterial, fungal and protozoal infections. *Expert Opinion on Pharmacotherapy* **8**(14), 2401–2444. doi: 10.1517/14656566.8.14.2401.

Gunderson, L. H. (2000). Ecological resilience in theory and application. *Annual Review of Ecology and Systematics* **31**(1), 425–439. doi: 10.1146/annurev.ecolsys.31.1.425.

Gürsoy, M., E. Könönen, U. K. Gürsoy, T. Tervahartiala, R. Pajukanta, and T. Sorsa (2010). Periodontal status and neutrophilic enzyme levels in gingival crevicular fluid during pregnancy and postpartum. *Journal of Periodontology* **81**(12), 1790–1796. doi: 10.1902/jop.2010.100147.

Haenni, M., L. Poirel, N. Kieffer, P. Châtre, E. Saras, V. Métayer, R. Dumoulin, P. Nordmann, and J.-Y. Madec (2016). Co-occurrence of extended spectrum β-lactamase and mcr-1 encoding genes on plasmids. *The Lancet Infectious Diseases* **16**(3), 281–282. doi: 10.1016/S1473-3099(16)00007-4.

Haffajee, A. D. and S. S. Socransky (1994). Microbial etiological agents of destructive periodontal diseases. *Periodontology 2000* **5**, 78–111. url: http://www.ncbi.nlm.nih.gov/pubmed/9673164.

Haffajee, A. D., S. S. Socransky, M. R. Patel, and X. Song (2008). Microbial complexes in supragingival plaque. *Oral Microbiology and Immunology* **23**(3), 196–205. doi: 10.1111/j.1399-302X.2007.00411.x.

Haffajee, A. D., R. P. Teles, M. R. Patel, X. Song, N. Veiga, and S. S. Socransky (2009). Factors affecting human supragingival biofilm composition. I. Plaque mass. *Journal of Periodontal Research* **44**(4), 511–519. doi: 10.1111/j.1600-0765.2008.01154.x.

Hajishengallis, G. and R. Lamont (2012). Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. *Molecular Oral Microbiology* **27**(6), 409–419. doi: 10.1111/j.2041-1014.2012.00663.x.

Hajishengallis, G. (2014). Periodontitis: from microbial immune subversion to systemic inflammation. *Nature Reviews Immunology* **15**(1), 30–44. doi: 10.1038/nri3785.

Harjunmaa, U., J. Järnstedt, L. Alho, K. G. Dewey, Y. B. Cheung, M. Deitchler, U. Ashorn, K. Maleta, N. J. Klein, and P. Ashorn (2015). Association between maternal dental periapical infections and pregnancy outcomes: results from a cross-sectional study in Malawi. *Tropical Medicine & International Health* **20**(11), 1549–1558. doi: 10.1111/tmi.12579.

Harjunmaa, U. et al. (2018). Periapical infection may affect birth outcomes via systemic inflammation. *Oral Diseases (in press)*.

Hathroubi, S., M. A. Mekni, P. Domenico, D. Nguyen, and M. Jacques (2017). Biofilms: microbial shelters against antibiotics. *Microbial Drug Resistance* **23**(2), 147–156. doi: 10.1089/mdr.2016.0087.

Hempel, S., S. J. Newberry, A. R. Maher, Z. Wang, J. N. V. Miles, R. Shanman, B. Johnsen, and P. G. Shekelle (2012). Probiotics for the prevention and treatment of antibiotic-associated diarrhea. *JAMA* **307**(18), 1959. doi: 10.1001/jama.2012.3507.

Heron, S. E. and S. Elahi (2017). HIV infection and compromised mucosal immunity: oral manifestations and systemic inflammation. *Frontiers in Immunology* **8**, 241. doi: 10.3389/fimmu.2017.00241.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**(2), 427–432. doi: 10.2307/1934352.

Holling, C. S. (1973). Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics* **4**, 1–23. url: http://www.annualreviews.org/doi/abs/10.1146/annurev.es.04.110173.000245.

Horne, R., S. C. E. Chapman, R. Parham, N. Freemantle, A. Forbes, and V. Cooper (2013). Understanding patients' adherence-related beliefs about medicines prescribed for long-term conditions: a meta-analytic review of the necessity-concerns framework. *PLOS ONE* **8**(12), e80633. doi: 10.1371/journal.pone.0080633.

Huang, S. et al. (2014). Predictive modeling of gingivitis severity and susceptibility via oral microbiota. *The ISME Journal* **8**(9), 1768–1780. doi: 10.1038/ismej.2014.32.

Humananatomyly.com (2017). *Labeled diagram of human mouth*. url: https://humananatomyly.com/wp-content/uploads/2017/07/diagram-of-human-mouth-labeled-the-mouth-pharynx-and-esophagus-c2b7-anatomy-and-physiology.jpg (visited on 12/02/2017).

Hunt, M., N. D. Silva, T. D. Otto, J. Parkhill, J. A. Keane, and S. R. Harris (2015). Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biology* **16**(1), 294. doi: 10.1186/s13059-015-0849-0.

Huse, S. M., L. Dethlefsen, J. A. Huber, D. M. Welch, D. A. Relman, and M. L. Sogin (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLOS Genetics* **4**(11), e1000255. doi: 10.1371/journal.pgen.1000255.

Huttenhower, C., D. Gevers, and R. Knight (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**(7402), 207–214. doi: 10.1038/nature11234.

Hyndman, R. (2017). *forecast: forecasting functions for time series and linear models*. url: http://pkg.robjhyndman.com/forecast.

Ide, M. and P. N. Papapanou (2013). Epidemiology of association between maternal periodontal disease and adverse pregnancy outcomes – systematic review. *Journal of Periodontology* **84**(4 Suppl), S181–S194. doi: 10.1902/jop.2013.134009.

Igartua, C., E. R. Davenport, Y. Gilad, D. L. Nicolae, J. Pinto, and C. Ober (2017). Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome* **5**(1), 16. doi: 10.1186/s40168-016-0227-5.

Jakobsson, H. E., C. Jernberg, A. F. Andersson, M. Sjölund-Karlsson, J. K. Jansson, and L. Engstrand (2010). Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLOS ONE* **5**(3), e9836. doi: 10.1371/journal.pone.0009836.

Jaspers, E. and J. Overmann (2004). Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. *Applied and Environmental Microbiology* **70**(8), 4831–4839. doi: 10.1128/AEM.70.8.4831-4839.2004.

Jeffcoat, M. K. and M. S. Reddy (1991). Progression of probing attachment loss in adult periodontitis. *Journal of Periodontology* **62**(3), 185–189. doi: 10.1902/jop.1991.62.3.185.

Jernberg, C., S. Löfmark, C. Edlund, and J. K. Jansson (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME Journal* **1**(1), 56–66. doi: 10.1038/ismej.2007.3.

Jiao, Y., M. Hasegawa, and N. Inohara (2014). The role of oral pathobionts in dysbiosis during periodontitis development. *Journal of Dental Research* **93**(6), 539–546. doi: 10.1177/0022034514528212.

John, C. N., L. X. Stephen, and C. W. Joyce Africa (2013). Is human immunodeficiency virus (HIV) stage an independent risk factor for altering the periodontal status of HIV-positive patients? A South African study. *BMC Oral Health* **13**, 69. doi: 10.1186/1472-6831-13-69.

Journal of Bacteriology (2018). *Nomenclature of Microorganisms*. url: http://jb.asm.org/site/misc/journal-ita_nom.xhtml.

Juhas, M. (2015). Horizontal gene transfer in human pathogens. *Critical Reviews in Microbiology* **41**(1), 101–108. doi: 10.3109/1040841X.2013.804031.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* **90**(430), 773–795. doi: 10.1080/01621459.1995.10476572.

Kembel, S. W., P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, and C. O. Webb (2010). picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**(11), 1463–1464. doi: 10.1093/bioinformatics/btq166.

Khammissa, R., L. Feller, M. Altini, P. Fatti, and J. Lemmer (2012). A comparison of chronic periodontitis in HIV-seropositive subjects and the general population in the Ga-Rankuwa area, South Africa. *AIDS Research and Treatment* **2012**, 620962. doi: 10.1155/2012/620962.

Kieffer, N., P. Nordmann, and L. Poirel (2017). Moraxella species as potential sources of mcr-like polymyxin resistance determinants. *Antimicrobial Agents and Chemotherapy* **61**(6), e00129–17. doi: 10.1128/AAC.00129-17.

Kistler, J. O., P. Arirachakaran, Y. Poovorawan, G. Dahlén, and W. G. Wade (2015). The oral microbiome in human immunodeficiency virus (HIV)-positive individuals. *Journal of Medical Microbiology* **64**(9), 1094–1101. doi: 10.1099/jmm.0.000128.

Kistler, J. O., V. Booth, D. J. Bradshaw, and W. G. Wade (2013). Bacterial community development in experimental gingivitis. *PLOS ONE* **8**(8), e71227. doi: 10.1371/journal.pone.0071227.

Koenig, J. E., A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, and R. E. Ley (2011). Succession of microbial consortia in the developing infant gut microbiome. *PNAS* **108**(Supplement 1), 4578–4585. doi: 10.1073/pnas.1000081107.

Koonin, E. V. (2012). *The Logic of Chance: the nature and origin of biological evolution*. Pearson Education, p. 516. url: https://www.pearson.com/us/higher-education/program/Koonin-Logic-of-Chance-The-The-Nature-and-Origin-of-Biological-Evolution-paperback/PGM280424.html.

Koren, S., B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**(5), 722–736. doi: 10.1101/gr.215087.116.

Koskinen, K., M. R. Pausan, A. K. Perras, M. Beck, C. Bang, M. Mora, A. Schilhabel, R. Schmitz, and C. Moissl-Eichinger (2017). First insights into the diverse human archaeome: specific detection of Archaea in the gastrointestinal tract, lung, and nose and on skin. *mBio* **8**(6), e00824–17. doi: 10.1128/mBio.00824-17.

Krishnan, K., T. Chen, and B. Paster (2016). A practical guide to the oral microbiome and its relation to health and disease. *Oral Diseases* **23**(3), 276–286. doi: 10.1111/odi.12509.

Kumar, P. S., M. R. Brooker, S. E. Dowd, and T. Camerlengo (2011). Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLOS ONE* **6**(6), e20956. doi: 10.1371/journal.pone.0020956.

Langille, M. G. I. et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**(9), 814–821. doi: 10.1038/nbt.2676.

Lassalle, F., D. Muller, and X. Nesme (2015). Ecological speciation in bacteria: reverse ecology approaches reveal the adaptive part of bacterial cladogenesis. *Research in Microbiology* **166**(10), 729–741. doi: 10.1016/j.resmic.2015.06.008.

Lassalle, F., M. Spagnoletti, M. Fumagalli, L. P. Shaw, M. Dyble, C. Walker, M. G. Thomas, A. Bamberg Migliano, and F. Balloux (2017). Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular Ecology*(In press). doi: 10.1111/mec.14435.

Laudenbach, J. M. and Z. Simon (2014). Common dental and periodontal diseases. *Medical Clinics of North America* **98**(6), 1239–1260. doi: 10.1016/j.mcna.2014.08.002.

Lax, S. et al. (2014). Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**(6200), 1048–1052. doi: 10.1126/science.1254529.

Lecuit, M. and M. Eloit (2013). The human virome: new tools and concepts. *Trends in Microbiology* **21**(10), 510–515. doi: 10.1016/j.tim.2013.07.001.

Lederberg, J. and A. T. McCray (2001). 'Ome Sweet 'Omics – a genealogical treasury of words. *The Scientist*, 15:8. url: https://www.the-scientist.com/?articles.view/articleNo/13313/.

Lee, J.-Y., M.-J. Choi, H. J. Choi, and K. S. Ko (2016). Preservation of acquired colistin resistance in Gram-negative bacteria. *Antimicrobial Agents and Chemotherapy* **60**(1), 609–612. doi: 10.1128/AAC.01574-15.

Leeuwenhoek, A. van (1683). Letter of 17 September 1683 to the Royal Society, London. *Royal Society*, MS. 1898. L 1. 69. url: http://www.dbnl.org/tekst/leeu027alle04_01/leeu027alle04_01_0008.php.

Leigh, D. A. (1981). Antibacterial activity and pharmacokinetics of clindamycin. *Journal of Antimicrobial Chemotherapy* **7**(Supplement A), 3–9. doi: 10.1093/jac/7.suppl_A.3.

Lemon, K. P., G. C. Armitage, D. A. Relman, and M. A. Fischbach (2012). Microbiota-targeted therapies: an ecological perspective. *Science Translational Medicine* **4**(137), 137rv5. doi: 10.1126/scitranslmed.3004183.

Levine, A. P. (2015). The genetics of inflammatory bowel disease in extended multiplex Ashkenazi Jewish kindreds. PhD thesis. UCL. url: http://discovery.ucl.ac.uk/1461013/.

Levine, A. P., N. Pontikos, E. R. Schiff, L. Jostins, D. Speed, L. B. Lovat, J. C. Barrett, H. Grasberger, V. Plagnol, and A. W. Segal (2016). Genetic complexity of Crohn's disease in two large Ashkenazi Jewish families. *Gastroenterology* **151**(4), 698–709. doi: 10.1053/j.gastro.2016.06.040.

Li, B., Y. Yang, L. Ma, F. Ju, F. Guo, J. M. Tiedje, and T. Zhang (2015). Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *The ISME Journal* **9**(11), 2490–2502. doi: 10.1038/ismej.2015.59.

Li, K., M. Bihan, S. Yooseph, and B. A. Methé (2012). Analyses of the microbial diversity across the human microbiome. *PLoS ONE* **7**(6), e32118. doi: 10.1371/journal.pone.0032118.

Li, R., M. Xie, J. Zhang, Z. Yang, L. Liu, X. Liu, Z. Zheng, E. W.-C. Chan, and S. Chen (2017). Genetic characterization of *mcr-1*-bearing plasmids to depict molecular mechanisms underlying dissemination of the colistin resistance determinant. *The Journal of Antimicrobial Chemotherapy* **72**(2), 393–401. doi: 10.1093/jac/dkw411.

Li, X., K. M. Kolltveit, L. Tronstad, and I. Olsen (2000). Systemic diseases caused by oral infection. *Clinical Microbiology Reviews* **13**(4), 547–558. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC88948/.

Liu, Y.-Y. et al. (2016). Emergence of plasmid-mediated colistin resistance mechanism *mcr-1* in animals and human beings in China: a microbiological and molecular biological study. *The Lancet Infectious Diseases* **16**(2), 161–168. doi: 10.1016/S1473-3099(15)00424-7.

Llewelyn, M. J., K. Hand, S. Hopkins, and A. S. Walker (2014). Antibiotic policies in acute English NHS trusts: implementation of 'Start Smart – Then Focus' and relationship with Clostridium difficile infection rates. *Journal of Antimicrobial Chemotherapy* **70**(4), 1230–1235. doi: 10.1093/jac/dku515.

Llewelyn, M. J., J. M. Fitzpatrick, E. Darwin, SarahTonkin-Crine, C. Gorton, J. Paul, T. E. A. Peto, L. Yardley, S. Hopkins, and A. S. Walker (2017). The antibiotic course has had its day. *The BMJ* **358**, j3418. doi: 10.1136/BMJ.J3418.

Lloyd-Price, J., G. Abu-Ali, and C. Huttenhower (2016). The healthy human microbiome. *Genome Medicine* **8**(1), 51. doi: 10.1186/s13073-016-0307-y.

Löfmark, S., C. Jernberg, J. K. Jansson, and C. Edlund (2006). Clindamycin-induced enrichment and long-term persistence of resistant Bacteroides spp. and resistance genes. *The Journal of Antimicrobial Chemotherapy* **58**(6), 1160–1167. doi: 10.1093/jac/dkl420.

Love, M. I., W. Huber, and S. Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**(12), 550. doi: 10.1186/s13059-014-0550-8.

Lucas López, R., M. J. Grande Burgos, A. Gálvez, and R. Pérez Pulido (2017). The human gastrointestinal tract and oral microbiota in inflammatory bowel disease: a state of the science review. *APMIS* **125**(1), 3–10. doi: 10.1111/apm.12609.

Lynch, S. V. and O. Pedersen (2016). The human intestinal microbiome in health and disease. *New England Journal of Medicine* **375**(24), 2369–2379. doi: 10.1056/NEJMra1600266.

Mandelbrot, B. B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* **156**, 636–638. url: http://www.jstor.org/stable/1721427.

Mark Welch, J. L., B. J. Rossetti, C. W. Rieken, F. E. Dewhirst, and G. G. Borisy (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences* **113**(6), E791–E800. doi: 10.1073/pnas.1522149113.

Mark Welch, J. L., D. R. Utter, B. J. Rossetti, D. B. Mark Welch, A. M. Eren, and G. G. Borisy (2014). Dynamics of tongue microbial communities with single-nucleotide resolution using oligotyping. *Frontiers in Microbiology* **5**, 568. doi: 10.3389/fmicb.2014.00568.

Marsh, P. D. (2003). Are dental diseases examples of ecological catastrophes? *Microbiology* **149**(2), 279–294. doi: 10.1099/mic.0.26082-0.

Marsh, P. D., D. A. Head, and D. A. Devine (2015). Dental plaque as a biofilm and a microbial community – implications for treatment. *Journal of Oral Biosciences* **57**(4), 185–191. doi: 10.1016/j.job.2015.08.002.

Martinez, J. L. (2009). Environmental pollution by antibiotics and by antibiotic resistance determinants. *Environmental Pollution* **157**(11), 2893–2902. doi: 10.1016/j.envpol.2009.05.051.

Mason, M. R., H. N. Nagaraja, T. Camerlengo, V. Joshi, and P. S. Kumar (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PLOS ONE* **8**(10), e77287. doi: 10.1371/journal.pone.0077287.

Matamoros, S. et al. (2017). Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the *mcr-1* gene indicates bacterial diversity but plasmid restriction. *Scientific Reports* **7**(1), 15364. doi: 10.1038/s41598-017-15539-7.

May, R. M. (M. (1973). *Stability and complexity in model ecosystems*. Princeton: Princeton University Press, p. 265.

Mayr, E. (1942). *Systematics and the Origin of Species*. New York: Columbia University Press.

McMurdie, P. J. and S. Holmes (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE* **8**(4), e61217. doi: 10.1371/journal.pone.0061217.

— (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLOS Computational Biology* **10**(4), e1003531. doi: 10.1371/journal.pcbi.1003531.

Mdala, I., I. Olsen, A. D. Haffajee, S. S. Socransky, M. Thoresen, and B. F. de Blasio (2014). Comparing clinical attachment level and pocket depth for predicting periodontal disease progression in healthy sites of patients with chronic periodontitis using multi-state Markov models. *Journal of Clinical Periodontology* **41**(9), 837–845. doi: 10.1111/jcpe.12278.

Methé, B. A. et al. (2012). A framework for human microbiome research. *Nature* **486**(7402), 215–221. doi: 10.1038/nature11209.

Michaud, D. S. et al. (2013). Lifestyle, dietary factors, and antibody levels to oral bacteria in cancer-free participants of a European cohort study. *Cancer Causes & Control* **24**(11), 1901–1909. doi: 10.1007/s10552-013-0265-2.

Miller, W. D. (1890). *The Micro-Organisms of the Human Mouth: the local and general diseases which are caused by them.* Reprinted. Philadelphia: The S.S. White Dental MFG. Co. url: https://archive.org/stream/microorganismsof00mill.

Modi, S. R., J. J. Collins, and D. A. Relman (2014). Antibiotics and the gut microbiota. *The Journal of Clinical Investigation* **124**(10), 4212–4218. doi: 10.1172/JCI72333.

Moeller, A. H., Y. Li, E. Mpoudi Ngole, S. Ahuka-Mundeke, E. V. Lonsdorf, A. E. Pusey, M. Peeters, B. H. Hahn, and H. Ochman (2014). Rapid changes in the gut microbiome during human evolution. *PNAS* **111**(46), 16431–16435. doi: 10.1073/pnas.1419136111.

Momeni, B., L. Xie, and W. Shou (2017). Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. *eLife* **6**, e25051. doi: 10.7554/eLife.25051.

Moore, W. E. C. et al. (1987). Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology* **37**(4), 463–464. doi: 10.1099/00207713-37-4-463.

Morales, S. E. and W. E. Holben (2009). Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by in silico analysis. *Applied and Environmental Microbiology* **75**(9), 2677–2683. doi: 10.1128/AEM.02166-08.

Moynihan, P. (2016). Sugars and dental caries: evidence for setting a recommended threshold for intake. *Advances in Nutrition* **7**(1), 149–156. doi: 10.3945/an.115.009365.

Moynihan, P. and S. A. M. Kelly (2014). Effect on caries of restricting sugars intake: systematic review to inform WHO guidelines. *Journal of Dental Research* **93**(1), 8–18. doi: 10.1177/0022034513508954.

Mustaev, A. et al. (2014). Fluoroquinolone-gyrase-DNA complexes. *Journal of Biological Chemistry* **289**(18), 12300–12312. doi: 10.1074/jbc.M113.529164.

Mydel, P. et al. (2006). Roles of the host oxidative immune response and bacterial antioxidant rubrerythrin during *Porphyromonas gingivalis* infection. *PLOS Pathogens* **2**(7), e76. doi: 10.1371/journal.ppat.0020076.

Nasidze, I., J. Li, D. Quinque, K. Tang, and M. Stoneking (2009). Global diversity in the human salivary microbiome. *Genome Research* **19**(4), 636–643. doi: 10.1101/gr.084616.108.

Nociti, F. H., M. Z. Casati, and P. M. Duarte (2015). Current perspective of the impact of smoking on the progression and treatment of periodontitis. *Periodontology 2000* **67**(1), 187–210. doi: 10.1111/prd.12063.

Nunes, T., G. Fiorino, S. Danese, and M. Sans (2011). Familial aggregation in inflammatory bowel disease: is it genes or environment? *World Journal of Gastroenterology* **17**(22), 2715. doi: 10.3748/wjg.v17.i22.2715.

O'Brien, K. L., L. J. Wolfson, J. P. Watt, E. Henkle, M. Deloria-Knoll, N. McCall, E. Lee, K. Mulholland, O. S. Levine, and T. Cherian (2009). Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *The Lancet* **374**(9693), 893–902. doi: 10.1016/S0140-6736(09)61204-6.

OED Online (2018). *microbiome, n.* url: http://www.oed.com/view/Entry/365863 (visited on 03/29/2018).

Oksanen, J. (2016). *vegan: community ecology package*. url: https://cran.r-project.org/package=vegan.

Olaitan, A. O., S. Morand, and J.-M. Rolain (2014). Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Frontiers in Microbiology* **5**, 643. doi: 10.3389/fmicb.2014.00643.

O'Neill, J. (2016). *Tackling drug-resistant infections globally: final report and recommendations*. London: The Review on Antimicrobial Resistance. url: https://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf.

Paine, R. T., M. J. Tegner, and E. A. Johnson (1998). Compounded perturbations yield ecological surprises. *Ecosystems* **1**(6), 535–545. doi: 10.1007/s100219900049.

Parijs, I. and H. P. Steenackers (2017). Competitive inter-species interactions underlie the increased antimicrobial tolerance in multispecies brewery biofilms. *bioRxiv*, 204628. doi: 10.1101/204628.

Park, O.-J., H. Yi, J. H. Jeon, S.-S. Kang, K.-T. Koo, K.-Y. Kum, J. Chun, C.-H. Yun, and S. H. Han (2015). Pyrosequencing analysis of subgingival microbiota in distinct periodontal conditions. *Journal of Dental Research* **94**(7), 921–927. doi: 10.1177/0022034515583531.

Pearson, K. (1897). Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London (1854-1905)* **60**(1), 489–498. doi: 10.1098/rspl.1896.0076.

Pepper, J. W. and S. Rosenfeld (2012). The emerging medical ecology of the human gut microbiome. *Trends in Ecology & Evolution* **27**(7), 381–384. doi: 10.1016/j.tree.2012.03.002.

Petersen, C. and J. L. Round (2014). Defining dysbiosis and its influence on host immunity and disease. *Cellular Microbiology* **16**(7), 1024–1033. doi: 10.1111/cmi.12308.

Petersen, P. E., D. Bourgeois, H. Ogawa, S. Estupinan-Day, and C. Ndiaye (2005). The global burden of oral diseases and risks to oral health. *Bulletin of the World Health Organization* **83**(9), 661–669. doi: /S0042-96862005000900011.

Peterson, G., A. Kumar, E. Gart, and S. Narayanan (2011). Catecholamines increase conjugative gene transfer between enteric bacteria. *Microbial Pathogenesis* **51**(1-2), 1–8. doi: 10.1016/j.micpath.2011.03.002.

Poirel, L., N. Kieffer, A. Brink, J. Coetze, A. Jayol, and P. Nordmann (2016). Genetic features of mcr-1-producing colistin-resistant Escherichia coli isolates in South Africa. *Antimicrobial Agents and Chemotherapy* **60**(7), 4394–4397. doi: 10.1128/AAC.00444-16.

Quince, C., T. O. Delmont, S. Raguideau, J. Alneberg, A. E. Darling, G. Collins, and A. M. Eren (2017). DESMAN: a new tool for *de novo* extraction of strains from metagenomes. *Genome Biology* **18**(1), 181. doi: 10.1186/s13059-017-1309-9.

Quinque, D., R. Kittler, M. Kayser, M. Stoneking, and I. Nasidze (2006). Evaluation of saliva as a source of human DNA for population and association studies. *Analytical Biochemistry* **353**(2), 272–277. doi: 10.1016/j.ab.2006.03.021.

Rabelo, C. C., M. Feres, C. Gonçalves, L. C. Figueiredo, M. Faveri, Y.-K. Tu, and L. Chambrone (2015). Systemic antibiotics in the treatment of aggressive periodontitis. A systematic review and a Bayesian network meta-analysis. *Journal of Clinical Periodontology* **42**(7), 647–657. doi: 10.1111/jcpe.12427.

Ramachandran, G. (2014). Gram-positive and Gram-negative bacterial toxins in sepsis. *Virulence* **5**(1), 213–218. doi: 10.4161/viru.27024.

Rambaut, A., T. T. Lam, L. Max Carvalho, and O. G. Pybus (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2**(1), vew007. doi: 10.1093/ve/vew007.

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* **62**(2), 142–160. doi: 10.1111/j.1574-6941.2007.00375.x.

Rashid, M.-U., A. Weintraub, and C. E. Nord (2015). Development of antimicrobial resistance in the normal anaerobic microbiota during one year after administration of clindamycin or ciprofloxacin. *Anaerobe* **31**, 72–77. doi: 10.1016/j.anaerobe.2014.10.004.

Ravenhall, M., N. Skunca, F. Lassalle, and C. Dessimoz (2015). Inferring horizontal gene transfer. *PLOS Computational Biology* **11**(5), e1004095. doi: 10.1371/journal.pcbi.1004095.

Raymond, F., M. Déraspe, M. Boissinot, M. G. Bergeron, and J. Corbeil (2016). Partial recovery of microbiomes after antibiotic treatment. *Gut Microbes* **7**(5), 428–434. doi: 10.1080/19490976.2016.1216747.

Raymond, F., A. A. Ouameur, et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME Journal* **10**(3), 707–720. doi: 10.1038/ismej.2015.148.

Relman, D. A. (2012). The human microbiome: ecosystem resilience and health. *Nutrition Reviews* **70**(Supplement 1), S2–S9. doi: 10.1111/j.1753-4887.2012.00489.x.

Reznik, D. A. (2005). Oral manifestations of HIV disease. *Topics in HIV Medicine* **13**(5), 143–148. url: http://www.ncbi.nlm.nih.gov/pubmed/16377852.

Ricotta, C. (2005). Through the jungle of biological diversity. *Acta Biotheoretica* **53**(1), 29–38. doi: 10.1007/s10441-005-7001-6.

Ridenhour, B. J., G. A. Metzger, M. France, K. Gliniewicz, J. Millstein, L. J. Forney, and E. M. Top (2017). Persistence of antibiotic resistance plasmids in bacterial biofilms. *Evolutionary Applications* **10**(6), 640–647. doi: 10.1111/eva.12480.

Riley, K. F., M. P. Hobson, and S. J. Bence (1997). *Mathematical methods for physics and engineering: a comprehensive guide*. Cambridge: Cambridge University Press, p. 1008.

Roberts, A. P. and P. Mullany (2010). Oral biofilms: a reservoir of transferable, bacterial, antimicrobial resistance. *Expert Review of Anti-infective Therapy* **8**(12), 1441–1450. doi: 10.1586/eri.10.106.

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

Routy, B. et al. (2017). Gut microbiome influences efficacy of PD-1based immunotherapy against epithelial tumors. *Science (in press)*, eaan3706. doi: 10.1126/science.aan3706.

Ruby, J. and J. Barbeau (2002). The buccale puzzle: the symbiotic nature of endogenous infections of the oral cavity. *The Canadian Journal of Infectious Diseases* **13**(1), 34–41. url: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/18159372/.

Rugg-Gunn, A. (2013). Dental caries: strategies to control this preventable disease. *Acta Medica Academica* **42**(2), 117–130. doi: 10.5644/ama2006-124.80.

Sampaio-Maia, B. and F. Monteiro-Silva (2014). Acquisition and maturation of oral microbiome throughout childhood: An update. *Dental Research Journal* **11**(3), 291–301. url: http://www.ncbi.nlm.nih.gov/pubmed/25097637.

Schaik, W. van (2015). The human gut resistome. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**(1670), 20140087. doi: 10.1098/rstb.2014.0087.

Schätzle, M., H. Löe, W. Bürgin, A. Anerud, H. Boysen, and N. P. Lang (2003). Clinical course of chronic periodontitis. I. Role of gingivitis. *Journal of Clinical Periodontology* **30**(10), 887–901. url: http://www.ncbi.nlm.nih.gov/pubmed/14710769.

Scheffer, M., S. Carpenter, J. A. Foley, C. Folke, and B. Walker (2001). Catastrophic shifts in ecosystems. *Nature* **413**(6856), 591–596. doi: 10.1038/35098000.

Schlafer, S., B. Riep, A. L. Griffen, A. Petrich, J. Hübner, M. Berning, A. Friedmann, U. B. Göbel, and A. Moter (2010). *Filifactor alocis* – involvement in periodontal biofilms. *BMC Microbiology* **10**(1), 66. doi: 10.1186/1471-2180-10-66.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics* **27**(4), 592–593. doi: 10.1093/bioinformatics/btq706.

Schloissnig, S. et al. (2012). Genomic variation landscape of the human gut microbiome. *Nature* **493**(7430), 45–50. doi: 10.1038/nature11711.

Schluenzen, F. et al. (2000). Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell* **102**(5), 615–623. doi: 10.1016/S0092-8674(00)00084-2.

Schubert, A. M., M. A. M. Rogers, C. Ring, J. Mogle, J. P. Petrosino, V. B. Young, D. M. Aronoff, and P. D. Schloss (2014). Microbiome data distinguish patients with *Clostridium difficile* infection and non-*C. difficile*-associated diarrhea from healthy controls. *mBio* **5**(3), e01021–14. doi: 10.1128/mBio.01021-14.

Schwarz, S. and A. P. Johnson (2016). Transferable resistance to colistin: a new but old threat. *The Journal of Antimicrobial Chemotherapy* **71**(8), 2066–2070. doi: 10.1093/jac/dkw274.

Segata, N., S. Haake, P. Mannon, K. P. Lemon, L. Waldron, D. Gevers, C. Huttenhower, and J. Izard (2012). Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biology* **13**(6), R42. doi: 10.1186/gb-2012-13-6-r42.

Sender, R., S. Fuchs, and R. Milo (2016). Revised estimates for the number of human and bacteria cells in the body. *PLOS Biology* **14**(8), e1002533. doi: 10.1371/journal.pbio.1002533.

Seville, L. A., A. J. Patterson, K. P. Scott, P. Mullany, M. A. Quail, J. Parkhill, D. Ready, M. Wilson, D. Spratt, and A. P. Roberts (2009). Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic DNA. *Microbial Drug Resistance* **15**(3), 159–166. doi: 10.1089/mdr.2009.0916.

Shaddox, L., J. Wiedey, E. Bimstein, I. Magnuson, M. Clare-Salzler, I. Aukhil, and S. Wallet (2010). Hyper-responsive phenotype in localized aggressive periodontitis. *Journal of Dental Research* **89**(2), 143–148. doi: 10.1177/0022034509353397.

Shade, A. (2017). Diversity is the question, not the answer. *The ISME Journal* **11**(1), 1–6. doi: 10.1038/ismej.2016.118.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27**(3), 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shapiro, B. J. (2016). How clonal are bacteria over time? *Current Opinion in Microbiology* **31**, 116–123. doi: 10.1016/j.mib.2016.03.013.

Shapiro, B. J. and M. F. Polz (2014). Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology* **22**(5), 235–247. doi: 10.1016/j.tim.2014.02.006.

Shaw, L. P. (2018). The microbial ecology of human-associated bacterial communities. PhD thesis. UCL, p. 158.

Shaw, L. P., C. P. Barnes, A. S. Walker, N. Klein, and F. Balloux (2017). A perturbation model of the gut microbiome's response to antibiotics. *bioRxiv*, 222398. doi: 10.1101/222398.

Shaw, L. P., A. L. R. Ribeiro, A. P. Levine, N. Pontikos, F. Balloux, A. W. Segal, A. P. Roberts, and A. M. Smith (2017). The human salivary microbiome is shaped by shared environment rather than genetics: evidence from a large family of closely related individuals. *mBio* **8**(5), e01237–17. doi: 10.1128/mBio.01237-17.

Shaw, L. P. et al. (2016). Distinguishing the signals of gingivitis and periodontitis in supragingival plaque: a cross-sectional cohort study in Malawi. *Applied and Environmental Microbiology* **82**(19), 6057–6067. doi: 10.1128/AEM.01756-16.

Sheets, S. M., A. G. Robles-Price, R. M. E. McKenzie, C. A. Casiano, and H. M. Fletcher (2008). Gingipain-dependent interactions with the host are important for survival of Porphyromonas gingivalis. *Frontiers in Bioscience* **13**, 3215–3238. url: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/18508429/.

Shen, Z., Y. Wang, Y. Shen, J. Shen, and C. Wu (2016). Early emergence of *mcr-1* in *Escherichia coli* from food-producing animals. *The Lancet Infectious Diseases* **16**(3), 293. doi: 10.1016/S1473-3099(16)00061-X.

Sheppard, A. E. et al. (2016). Nested Russian doll-like genetic mobility drives rapid dissemination of the carbapenem resistance gene blaKPC. *Antimicrobial Agents and Chemotherapy* **60**(6), 3767–3778. doi: 10.1128/AAC.00464-16.

Sievers, F. et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**(1), 539. doi: 10.1038/msb.2011.75.

Simms-Waldrip, T. R. et al. (2017). Antibiotic-induced depletion of anti-inflammatory Clostridia is associated with the development of graft-versus-host disease in pediatric stem cell transplantation patients. *Biology of Blood and Marrow Transplantation* **23**(5), 820–829. doi: 10.1016/j.bbmt.2017.02.004.

Simpson, E. H. (1949). Measurement of diversity. *Nature* **163**(4148), 688–688. doi: 10.1038/163688a0.

Sinnwell, J. P., T. M. Therneau, and D. J. Schaid (2014). The kinship2 R package for pedigree data. *Human Heredity* **78**(2), 91–93. doi: 10.1159/000363105.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika* **38**(1-2), 196–218. doi: 10.1093/biomet/38.1-2.196.

Smillie, C. S., M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, and E. J. Alm (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**(7376), 241–244. doi: 10.1038/nature10571.

Snesrud, E., S. He, M. Chandler, J. P. Dekker, A. B. Hickman, P. McGann, and F. Dyda (2016). A model for transposition of the colistin resistance gene *mcr-1* by IS*Apl1*. *Antimicrobial Agents and Chemotherapy* **60**(11), 6973–6976. doi: 10.1128/AAC.01457-16.

Snesrud, E., P. McGann, and M. Chandler (2018). The birth and demise of the IS*Apl1-mcr-1*-IS*Apl1* composite transposon: the vehicle for transferable colistin resistance. *mBio* **9**(1). Ed. by R. P. Novick, e02381–17. doi: 10.1128/mBio.02381-17.

Snesrud, E. et al. (2017). Analysis of serial isolates of*mcr-1*-positive Escherichia coli reveals a highly active IS*Apl1* transposon. *Antimicrobial Agents and Chemotherapy* **61**(5), e00056–17. doi: 10.1128/AAC.00056-17.

Socransky, S. S., A. D. Haffajee, M. A. Cugini, C. Smith, and R. L. Kent (1998). Microbial complexes in subgingival plaque. *Journal of Clinical Periodontology* **25**(2), 134–144. url: http://www.ncbi.nlm.nih.gov/pubmed/9495612.

Sommer, M. O. A., G. Dantas, and G. M. Church (2009). Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**(5944), 1128–1131. doi: 10.1126/science.1176950.

Song, S. J. et al. (2013). Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, e00458. doi: 10.7554/eLife.00458.

Speed, D. and D. J. Balding (2014). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* **16**(1), 33–44. doi: 10.1038/nrg3821.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**(6), 1011–1021. doi: 10.1016/j.ajhg.2012.10.010.

Spizek, J. and T. Rezanka (2004). Lincomycin, clindamycin and their applications. *Applied Microbiology and Biotechnology* **64**(4), 455–464. doi: 10.1007/s00253-003-1545-7.

Stahringer, S. S., J. C. Clemente, R. P. Corley, J. Hewitt, D. Knights, W. A. Walters, R. Knight, and K. S. Krauter (2012). 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. doi: 10.1101/gr.140608.112.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9), 1312–1313. doi: 10.1093/bioinformatics/btu033.

Stan Development Team (2017). *RStan: the R interface to Stan.* url: http://mc-stan.org/rstan/.

Stearns, J. C., M. D. J. Lynch, D. B. Senadheera, H. C. Tenenbaum, M. B. Goldberg, D. G. Cvitkovitch, K. Croitoru, G. Moreno-Hagelsieb, and J. D. Neufeld (2011). Bacterial biogeography of the human digestive tract. *Scientific Reports* **1**, 170. doi: 10.1038/srep00170.

Stecher, B. et al. (2012). Gut inflammation can boost horizontal gene transfer between pathogenic and commensal *Enterobacteriaceae*. *PNAS* **109**(4), 1269–1274. doi: 10.1073/pnas.1113246109.

Stein, R. R., V. Bucci, N. C. Toussaint, C. G. Buffie, G. Rätsch, E. G. Pamer, C. Sander, and J. B. Xavier (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLOS Computational Biology* **9**(12), e1003388. doi: 10.1371/journal.pcbi.1003388.

Stoddard, S. F., B. J. Smith, R. Hein, B. R. Roller, and T. M. Schmidt (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* **43**(D1), D593–D598. doi: 10.1093/nar/gku1201.

Takahashi, N. and B. Nyvad (2011). The role of bacteria in the caries process. *Journal of Dental Research* **90**(3), 294–303. doi: 10.1177/0022034510379602.

Takeshita, T. et al. (2014). Distinct composition of the oral indigenous microbiota in South Korean and Japanese adults. *Scientific Reports* **4**(1), 6990. doi: 10.1038/srep06990.

Takeshita, T. et al. (2016). Bacterial diversity in saliva and oral health-related conditions: the Hisayama study. *Scientific Reports* **6**, 22164. doi: 10.1038/srep22164.

Takeuchi, N., K. Kaneko, and E. V. Koonin (2014). Horizontal gene transfer can rescue prokaryotes from Muller's ratchet: benefit of DNA from dead cells and population subdivision. *G3* **4**(2), 325–339. doi: 10.1534/g3.113.009845.

Tansirichaiya, S., M. A. Rahman, A. Antepowicz, P. Mullany, and A. P. Roberts (2016). Detection of novel integrons in the metagenome of human saliva. *PLOS ONE* **11**(6), e0157605. doi: 10.1371/journal.pone.0157605.

Tansirichaiya, S., L. J. Reynolds, G. Cristarella, L. C. Wong, K. Rosendahl, and A. P. Roberts (2017). Reduced susceptibility to antiseptics is conferred by heterologous housekeeping genes. *Microbial Drug Resistance*, mdr.2017.0105. doi: 10.1089/mdr.2017.0105.

Taur, Y. et al. (2014). The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* **124**(7), 1174–1182. doi: 10.1182/blood-2014-02-554725.

Tegetmeyer, H. E., S. C. Jones, P. R. Langford, and N. Baltes (2008). IS*Apl1*, a novel insertion element of *Actinobacillus pleuropneumoniae*, prevents ApxIV-based serological detection of serotype 7 strain AP76. *Veterinary Microbiology* **128**(3-4), 342–353. doi: 10.1016/j.vetmic.2007.10.025.

Teles, R., F. Teles, J. Frias-Lopez, B. Paster, and A. Haffajee (2013). Lessons learned and unlearned in periodontal microbiology. *Periodontology 2000* **62**(1), 95–162. doi: 10.1111/prd.12010.

Tettelin, H., D. Riley, C. Cattuto, and D. Medini (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* **11**(5), 472–477. doi: 10.1016/j.mib.2008.09.006.

Thomas, C., M. Stevenson, and T. V. Riley (2003). Antibiotics and hospital-acquired *Clostridium difficile*-associated diarrhoea: a systematic review. *Journal of Antimicrobial Chemotherapy* **51**(6), 1339–1350. doi: 10.1093/jac/dkg254.

Tuomisto, H. (2010). A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* **164**(4), 853–860. doi: 10.1007/s00442-010-1812-0.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon (2007). The Human Microbiome Project. *Nature* **449**(7164), 804–810. doi: 10.1038/nature06244.

Turnbaugh, P. J. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* **457**(7228), 480–484. doi: 10.1038/nature07540.

Turner, I., K. V. Garimella, Z. Iqbal, and G. McVean (2017). Integrating long-range connectivity information into de Bruijn graphs. *bioRxiv*, 147777. doi: 10.1101/147777.

Utter, D. R., J. L. Mark Welch, and G. G. Borisy (2016). Individuality, stability, and variability of the plaque microbiome. *Frontiers in Microbiology* **7**, 564. doi: 10.3389/fmicb.2016.00564.

Van Dyke, T. E. (2008). The management of inflammation in periodontal disease. *Journal of Periodontology* **79**(8s), 1601–1608. doi: 10.1902/jop.2008.080173.

Vangay, P., T. Ward, J. S. Gerber, and D. Knights (2015). Antibiotics, pediatric dysbiosis, and disease. English. *Cell Host & Microbe* **17**(5), 553–564. doi: 10.1016/j.chom.2015.04.006.

Vartoukian, S. R., A. Adamowska, M. Lawlor, R. Moazzez, F. E. Dewhirst, and W. G. Wade (2016). In vitro cultivation of 'unculturable' oral bacteria, facilitated by community culture and media supplementation with siderophores. *PLOS ONE* **11**(1), e0146926. doi: 10.1371/journal.pone.0146926.

Vtrovský, T. and P. Baldrian (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLOS ONE* **8**(2), e57923. doi: 10.1371/journal.pone.0057923.

Wade, W. G. (2013). The oral microbiome in health and disease. *Pharmacological Research* **69**(1), 137–143. doi: 10.1016/j.phrs.2012.11.006.

Wang, K. et al. (2016). Preliminary analysis of salivary microbiome and their potential roles in oral lichen planus. *Scientific Reports* **6**(1), 22943. doi: 10.1038/srep22943.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**(16), 5261–7. doi: 10.1128/AEM.00062-07.

Wang, R. et al. (2018). The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nature Communications* **9**(1), 1179. doi: 10.1038/s41467-018-03205-z.

Wang, X., Y. Liu, X. Qi, R. Wang, L. Jin, M. Zhao, Y. Zhang, Q. Wang, H. Chen, and H. Wang (2017). Molecular epidemiology of colistin-resistant *Enterobacteriaceae* in inpatient and avian isolates from China: high prevalence of *mcr*-negative *Klebsiella pneumoniae*. *International Journal of Antimicrobial Agents* **50**(4), 536–541. doi: 10.1016/j.ijantimicag.2017.05.009.

Wang, Y., G.-B. Tian, et al. (2017). Prevalence, risk factors, outcomes, and molecular epidemiology of *mcr-1*-positive *Enterobacteriaceae* in patients and healthy adults from China: an epidemiological and clinical study. *The Lancet Infectious Diseases* **17**(4), 390–399. doi: 10.1016/S1473-3099(16)30527-8.

Wang, Y., R. Zhang, et al. (2017). Comprehensive resistome analysis reveals the prevalence of *NDM* and *mcr-1* in Chinese poultry production. *Nature Microbiology* **2**, 16260. doi: 10.1038/nmicrobiol.2016.260.

Warinner, C., C. Speller, and M. J. Collins (2015). A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**(1660), 20130376. doi: 10.1098/rstb.2013.0376.

Waterhouse, A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton (2009). Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9), 1189–1191. doi: 10.1093/bioinformatics/btp033.

Westcott, S. L. and P. D. Schloss (2015). *De novo* clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**, e1487. doi: 10.7717/peerj.1487.

Wintersdorff, C. J. H. von, P. F. G. Wolffs, J. M. van Niekerk, E. Beuken, L. B. van Alphen, E. E. Stobberingh, A. M. L. Oude Lashof, C. J. P. A. Hoebe, P. H. M. Savelkoul, and J. Penders (2016). Detection of the plasmid-mediated colistin-resistance gene *mcr-1* in faecal metagenomes of Dutch travellers. *The Journal of Antimicrobial Chemotherapy* **71**(12), 3416–3419. doi: 10.1093/jac/dkw328.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews* **51**(2), 221–271. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC373105/.

Woese, C. R. and G. E. Fox (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *PNAS* **74**(11), 5088–5090. doi: 10.1073/pnas.74.11.5088.

Wootton, J. T. (2010). Experimental species removal alters ecological dynamics in a natural ecosystem. *Ecology* **91**(1), 42–48. url: http://www.ncbi.nlm.nih.gov/pubmed/20380194.

Wu, D., Y. Kong, C. Han, J. Chen, L. Hu, H. Jiang, and X. Shen (2008). d-Alanine:d-alanine ligase as a new target for the flavonoids quercetin and apigenin. *International Journal of Antimicrobial Agents* **32**(5), 421–426. doi: 10.1016/j.ijantimicag.2008.06.010.

Wu, J. et al. (2016). Cigarette smoking and the oral microbiome in a large study of American adults. *The ISME Journal* **10**(10), 2435–2446. doi: 10.1038/ismej.2016.37.

Wu, Y.-W., M. Rho, T. G. Doak, and Y. Ye (2012). Oral spirochetes implicated in dental diseases are widespread in normal human subjects and carry extremely diverse integron gene cassettes. *Applied and Environmental Microbiology* **78**(15), 5288–5296. doi: 10.1128/AEM.00564-12.

Xavier, B. B., C. Lammens, R. Ruhal, S. Kumar-Singh, P. Butaye, H. Goossens, and S. Malhotra-Kumar (2016). Identification of a novel plasmid-mediated colistin-resistance gene, *mcr-2*, in *Escherichia coli*, Belgium, June 2016. *Eurosurveillance* **21**(27), 30280. doi: 10.2807/1560-7917.ES.2016.21.27.30280.

Ximénez-Fyvie, L. A., A. D. Haffajee, and S. S. Socransky (2000). Comparison of the microbiota of supra- and subgingival plaque in health and periodontitis. *Journal of clinical periodontology* **27**(9), 648–657. url: http://www.ncbi.nlm.nih.gov/pubmed/10983598.

Ximénez-Fyvie, L. A., A. D. Haffajee, S. Som, M. Thompson, G. Torresyap, and S. S. Socransky (2000). The effect of repeated professional supragingival plaque removal on the composition of the supra- and subgingival microbiota. *Journal of Clinical Periodontology* **27**(9), 637–647. url: http://www.ncbi.nlm.nih.gov/pubmed/10983597.

Yang, D., Z. Qiu, Z. Shen, H. Zhao, M. Jin, H. Li, W. Liu, and J.-W. Li (2017). The occurrence of the colistin resistance gene *mcr-1* in the Haihe River (China). *International journal of Environmental Research and Public Health* **14**(6), 576. doi: 10.3390/ijerph14060576.

Yang, F. et al. (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *The ISME Journal* **6**(1), 1–10. doi: 10.1038/ismej.2011.71.

Yao, X., Y. Doi, L. Zeng, L. Lv, and J.-H. Liu (2016). Carbapenem-resistant and colistin-resistant Escherichia coli co-producing NDM-9 and mcr-1. *The Lancet Infectious Diseases* **16**(3), 288–289. doi: 10.1016/S1473-3099(16)00057-8.

Yarza, P., P. Yilmaz, E. Pruesse, F. O. Glöckner, W. Ludwig, K.-H. Schleifer, W. B. Whitman, J. Euzéby, R. Amann, and R. Rosselló-Móra (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* **12**(9), 635–645. doi: 10.1038/nrmicro3330.

Yassour, M. et al. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science Translational Medicine* **8**(343), 343ra81. doi: 10.1126/scitranslmed.aad0917.

Yatsunenko, T. et al. (2012). Human gut microbiome viewed across age and geography. *Nature* **486**(7402), 222. doi: 10.1038/nature11053.

Yin, W., H. Li, Y. Shen, Z. Liu, S. Wang, Z. Shen, R. Zhang, T. R. Walsh, J. Shen, and Y. Wang (2017). Novel plasmid-mediated colistin resistance gene mcr-3 in Escherichia coli. *mBio* **8**(3), e00543–17. doi: 10.1128/mBio.00543-17.

Yu, C. Y., G. Y. Ang, T.-M. Chong, P. S. Chin, Y. F. Ngeow, W.-F. Yin, and K.-G. Chan (2016). Complete genome sequencing revealed novel genetic contexts of the mcr-1 gene in Escherichia coli strains. *Journal of Antimicrobial Chemotherapy* **72**(4), 1253–1255. doi: 10.1093/jac/dkw541.

Yu, G., D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**(1), 28–36. doi: 10.1111/2041-210X.12628.

Yu, Y.-H., D. I. Chasman, J. E. Buring, L. Rose, and P. M. Ridker (2015). Cardiovascular risks associated with incident and prevalent periodontal disease. *Journal of Clinical Periodontology* **42**(1), 21–28. doi: 10.1111/jcpe.12335.

Zaneveld, J. R., R. McMinds, and R. Vega Thurber (2017). Stress and stability: applying the Anna Karenina principle to animal microbiomes. *Nature Microbiology* **2**(9), 17121. doi: 10.1038/nmicrobiol.2017.121.

Zaura, E., E. A. Nicu, B. P. Krom, and B. J. F. Keijser (2014). Acquiring and maintaining a normal oral microbiome: current perspective. *Frontiers in Cellular and Infection Microbiology* **4**, 85. doi: 10.3389/fcimb.2014.00085.

Zaura, E. et al. (2015). Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. *mBio* **6**(6), e01693–15. doi: 10.1128/mBio.01693-15.

Zhang, X.-F., Y. Doi, X. Huang, H.-Y. Li, L.-L. Zhong, K.-J. Zeng, Y.-F. Zhang, S. Patil, and G.-B. Tian (2016). Possible transmission of mcr-1-harboring Escherichia coli between companion animals and human. *Emerging Infectious Diseases* **22**(9), 1679–1681. doi: 10.3201/eid2209.160464.

Zhao, F., Y. Feng, X. Lü, A. McNally, and Z. Zong (2017). Remarkable diversity of *Escherichia coli* carrying *mcr-1* from hospital sewage with the identification of two new *mcr-1* variants. *Frontiers in Microbiology* **8**, 2094. doi: 10.3389/fmicb.2017.02094.

Zhao, S., T. D. Lieberman, M. Poyet, M. Groussin, S. M. Gibbons, R. J. Xavier, and E. J. Alm (2017). Adaptive evolution within the gut microbiome of individual people. *bioRxiv*, 208009. doi: 10.1101/208009.

Zhou, H.-W., T. Zhang, J.-H. Ma, Y. Fang, H.-Y. Wang, Z.-X. Huang, Y. Wang, C. Wu, and G.-X. Chen (2017). Occurrence of plasmid- and chromosome-carried *mcr-1* in waterborne *Enterobacteriaceae* in China. *Antimicrobial Agents and Chemotherapy* **61**(8), e00017–17. doi: 10.1128/AAC.00017-17.

Zurfluh, K., N. Kieffer, L. Poirel, P. Nordmann, and R. Stephan (2016). Features of the *mcr-1* cassette related to colistin resistance. *Antimicrobial Agents and Chemotherapy* **60**(10), 6438–9. doi: 10.1128/AAC.01519-16.

Appendix A

Additional figures and tables

% identity with 1175R primer	Taxonomic classification	HOMD ID
90	*Rhodobacter capsulatus*	HOT_857_7798
90	*Desulfobulbus sp.*	HOT_041_R004
90	*Campylobacter rectus*	HOT_748_4317
90	*Campylobacter rectus*	HOT_748_6973
90	*Campylobacter showae*	HOT_763_6974
90	*Campylobacter concisus*	HOT_575_6977
94.74	*Campylobacter curvus*	HOT_580_4313
95	*Campylobacter gracilis*	HOT_623_4320
95	*Campylobacter sp.*	HOT_044BB120
95	*Campylobacter sputorum*	HOT_776_2768
95	*Bacteroides ureolyticus*	HOT_842_4321
95	*Prevotella micans*	HOT_378_1228
95	*Bacteroides heparinolyticus*	HOT_630_6487
95	*Bacteroides heparinolyticus*	HOT_630F0110
95	*Bacteroides zoogleoformans*	HOT_465_6488
95	*Porphyromonas endodontalis*	HOT_273AJ002
95	*Porphyromonas endodontalis*	HOT_273_7054
95	*Porphyromonas endodontalis*	HOT_273BB134
95	*Porphyromonas endodontalis*	HOT_273_6491
95	*Porphyromonas sp.*	HOT_395_7057
95	*Porphyromonas sp.*	HOT_285_F016
95	*Porphyromonas uenonis*	HOT_785F0120
95	*Porphyromonas asaccharolytica*	HOT_547_6490
95	***Porphyromonas gingivalis***	**HOT_619_3964**
95	*Tannerella sp.*	HOT_808BU045
95	*Tannerella sp.*	HOT_916-Wade
95	*Tannerella sp.*	HOT_286BU063
95	*Tannerella forsythia*	HOT_613_6495
95	*Bacteroidetes_[G-3] sp.*	HOT_281DA065
95	*Bacteroidetes_[G-3] sp.*	HOT_365_1206
95	*Bacteroidetes_[G-3] sp.*	HOT_365_3626
95	*Bacteroidetes_[G-3] sp.*	HOT_899-Wade
95	*Bacteroidetes_[G-3] sp.*	HOT_280DA064
95	*Bacteroidetes_[G-3] sp.*	HOT_436_1819
95	*Bacteroidetes_[G-3] sp.*	HOT_503_3613
95	***Treponema denticola***	**HOT_584_D011**
95	*Fusobacterium nucleatum ss polymorphum*	HOT_202BS019
95	*Lachnospiraceae_[G-3] sp.*	HOT_100EI074
95	*Oribacterium sp.*	HOT_102_1218
95	*Lachnospiraceae_[G-5] sp.*	HOT_080BB124
95	*Peptostreptococcaceae_[XIII][G-1] sp.*	HOT_113DA014
95	*Mycoplasma pneumoniae*	HOT_732_9061
95	*Mycoplasma genitalium*	HOT_616_7334
95	*Actinomyces sp.*	HOT_897-Wade
95	*Mobiluncus mulieris*	HOT_830_7625
95	*GN02_[G-1] sp.*	HOT_871_4L02
95	*GN02_[G-1] sp.*	HOT_872_CN02
95	*GN02_[G-2] sp.*	HOT_873_4Q04
95	*SR1_[G-1] sp.*	HOT_345_X112
95	*SR1_[G-1] sp.*	HOT_874_4Y03
95	*SR1_[G-1] sp.*	HOT_875_CN01
% identity with 785F primer		
95.24	*Leptothrix sp.*	HOT_025AV011
95.24	*Prevotella sp.*	HOT_296AU069
95.24	*Treponema sp.*	HOT_258_C009
95.24	*Treponema sp.*	HOT_270DD012
95.24	*Treponema sp.*	HOT_262AT040
95.24	*Selenomonas sputigena*	HOT_151_K168
95.24	*Solobacterium moorei*	HOT_678_1058
95.24	*Chloroflexi_[G-1] sp.*	HOT_439_1414

Table A.1: HOMD OTUs with mismatches to the 785F and 1175R primers used to amplify the V5-V7 region of the 16S rRNA gene. Members of the red complex (Socransky et al., 1998) are shown in ***red***.

1. `vsearch:fastq_filter`
   Filter sequences based on maximum errors, minimum/maximum length.
2. `vsearch:derep_fulllength`
   Dereplicate sequences.
3. `vsearch:sort_by_size`
   Sort by abundance and discard singletons (sequences that only appear once).
4. `vsearch:cluster_fast`
   Cluster into OTUs at defined sequence similarity e.g. 97%.
5. `vsearch:uchime_denovo`
   Filter chimeras *de novo*.
6. `vsearch:uchime_ref`
   Filter chimeras using ‘gold’ reference database at http://drive5.come/uchime/gold.fa
7. `vsearch:usearch_global`
   Map original reads back to OTUs.
8. `parallel_assign_taxonomy_rdp.py`
   Assign taxonomy to OTUs with RDP.

Figure A.1: Steps in the standard pipeline used in this thesis for OTU picking. This pipeline uses VSEARCH v1.1.1 (Rognes et al., 2016). See each chapter for more details and parameters used.

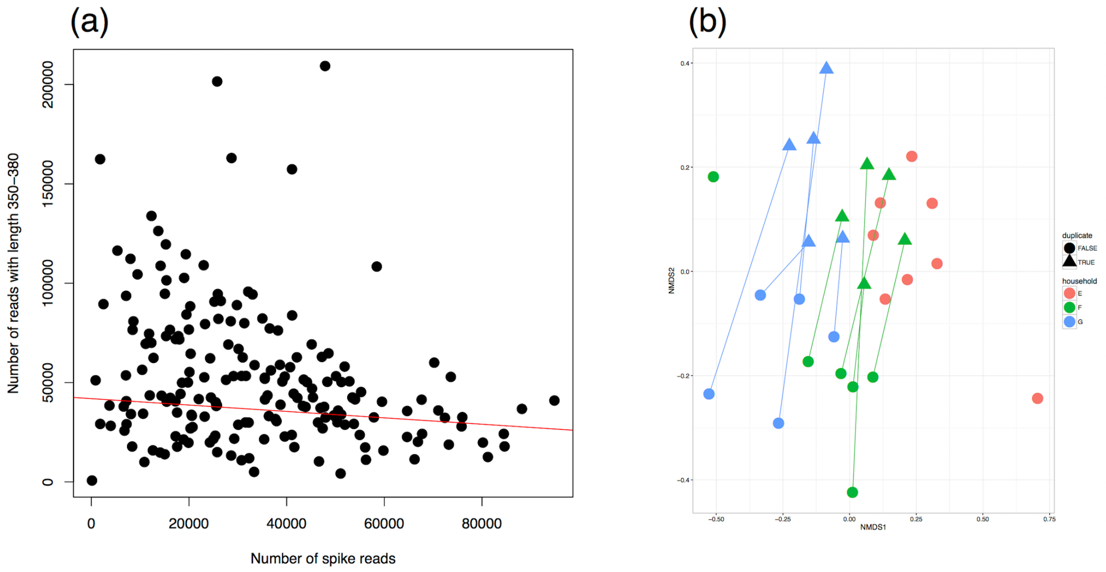


Figure A.2: Spikes added during library preparation do not have an important effect on analysis. (a) Number of reads from spikes of fixed concentrations added to samples during library preparation was weakly negatively correlated with the number of reads corresponding to true 16S reads. **(b)** Duplicated samples with and without spikes added during library preparation showed the same qualitative differences between households, indicating that addition of spikes did not negatively affect downstream analysis.

(a) Richness		Estimate (standard error)	Pr(>$\|t\|$)
Intercept		105.58 (4.23)	<0.001
Gingivitis		4.09 (0.71)	<0.001
Periodontitis		4.92 (3.18)	0.122
Anemia		-6.33 (3.37)	0.061
Site (vs. Lungwena)	Malindi	15.17 (3.75)	<0.001
	Namwera	16.25 (4.07)	<0.001
	Mangochi	18.89 (3.41)	<0.001
Intervention (vs. IFA)	MMN	8.52 (3.25)	0.009
	LNS	4.26 (3.32)	0.200
(b) Shannon index		Estimate (standard error)	Pr(>$\|t\|$)
Intercept		3.00 (0.05)	<0.001
Gingivitis		0.03 (0.01)	<0.001
Education (yrs)		-0.01 (0.01)	0.016
Anemia		-0.07 (0.04)	0.098
Site (vs. Lungwena)	Malindi	0.17 (0.05)	<0.001
	Namwera	0.20 (0.05)	<0.001
	Mangochi	0.23 (0.04)	<0.001
Intervention (vs. IFA)	MMN	0.08 (0.04)	0.029
	LNS	0.01 (0.04)	0.808

Table A.2: Final models predicting (a) richness and (b) Shannon index of supragingival plaque communities. Final models after backwards stepwise elimination according to AIC from a full model including gingivitis, periodontitis, and demographic variables from Table 3.1. Richness (observed number of species) and Shannon index (diversity measure) were averaged over 100 iterations of rarefying to 5,000 reads per sample, resulting in the removal of 138 out of 962 samples. A further 13 samples were removed due to missing data.

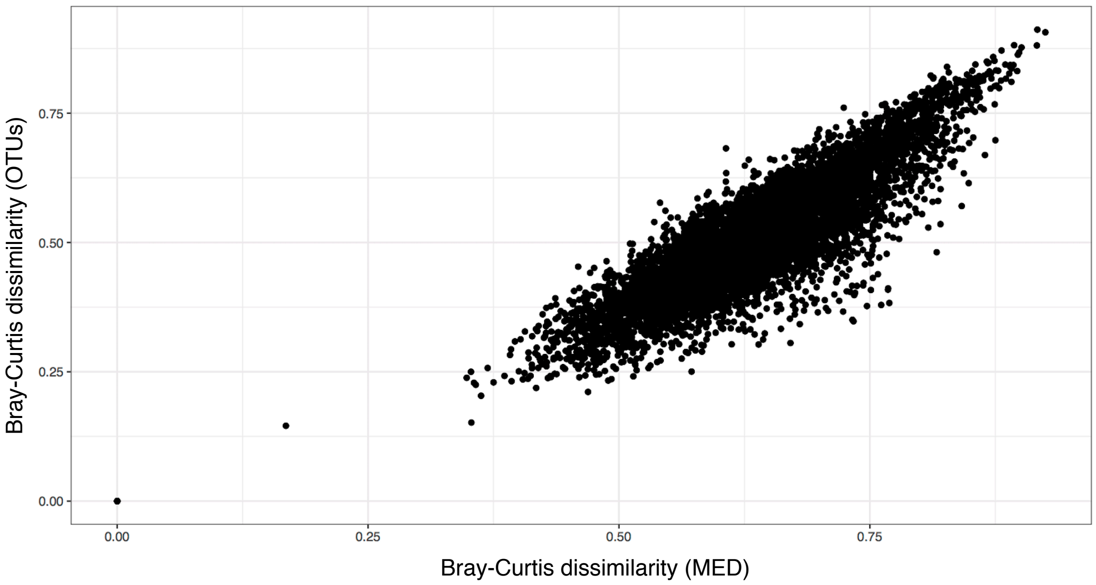


Figure A.3: MED and OTU picking give strongly correlated dissimilarities. Comparison of Bray-Curtis dissimilarities between samples calculated using compositions from MED and OTUs shows a high correlation between methods (Spearman's $\rho = 0.88$, $p < 0.001$). This correlation is expected, as both methods should identify sequences similarly to the genus level.

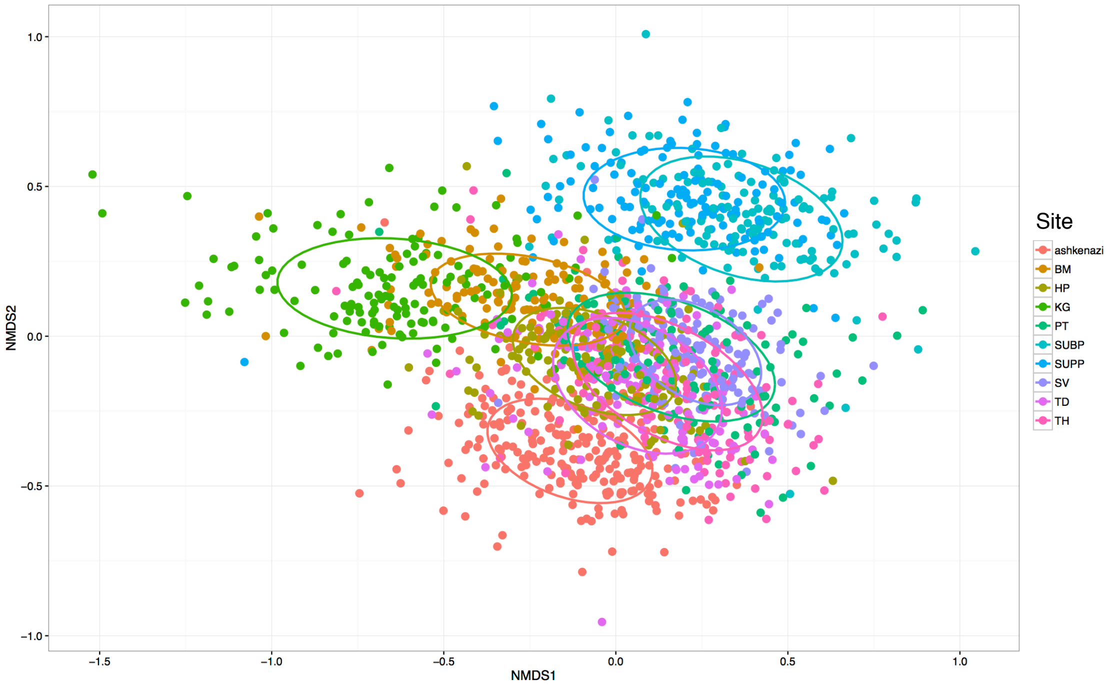


Figure A.4: NMDS ordination of samples from the Ashkenazi cohort ("ashkenazi") compared to samples from the Human Microbiome Project. Ashkenazi salivary microbiome samples were more similar to saliva (SV) and other non-plaque sites in the human mouth. Ashkenazi samples group near but separately from the HMP saliva samples, but we cannot distinguish whether this is due to a batch effect or a real biological difference. Site key: buccal mucosa (BM), hard palate (HP), keratinized gingivae (KG), palatine tonsils (PT), subgingival plaque (SUBP), supragingival plaque (SUPP), saliva (SV), tongue dorsum (TD), throat (TH).

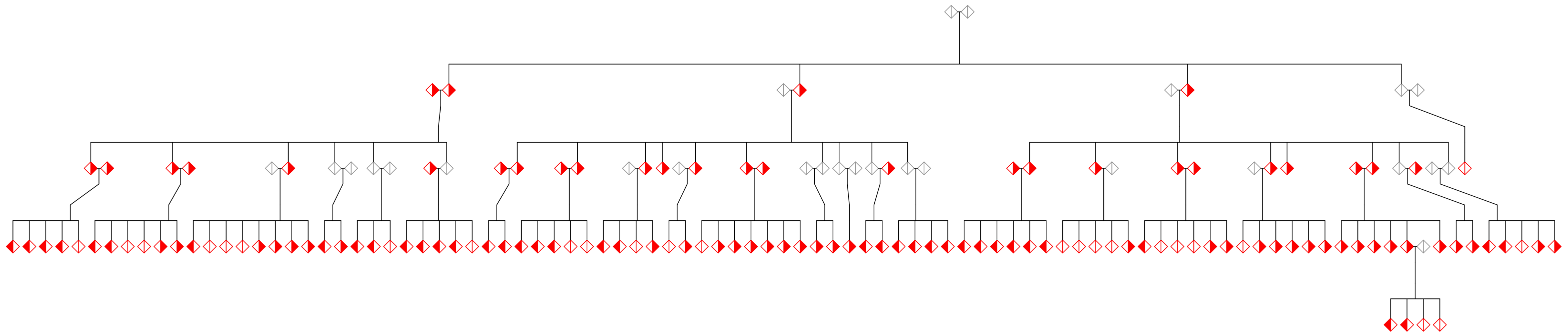


Figure A.5: Pedigree for Ashkenazi individuals within Family A. Diamonds are filled left for $\leq$ 18 years old, filled right for $\geq$ 25 years, grey for unsampled. There are three main branches of the family, which can also be seen in multidimensional scaling of the genetic distances between individuals.

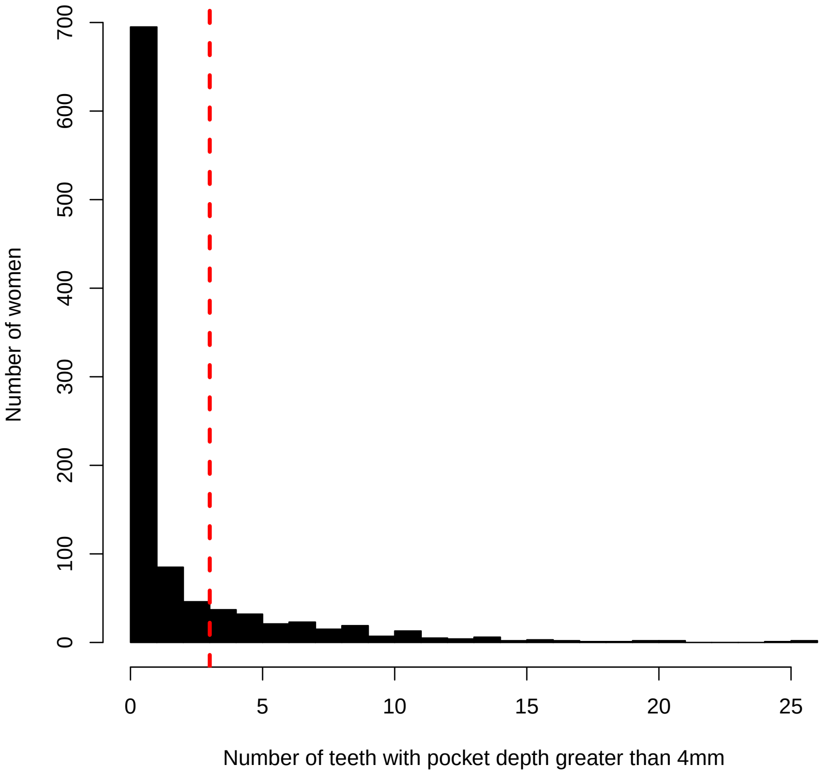
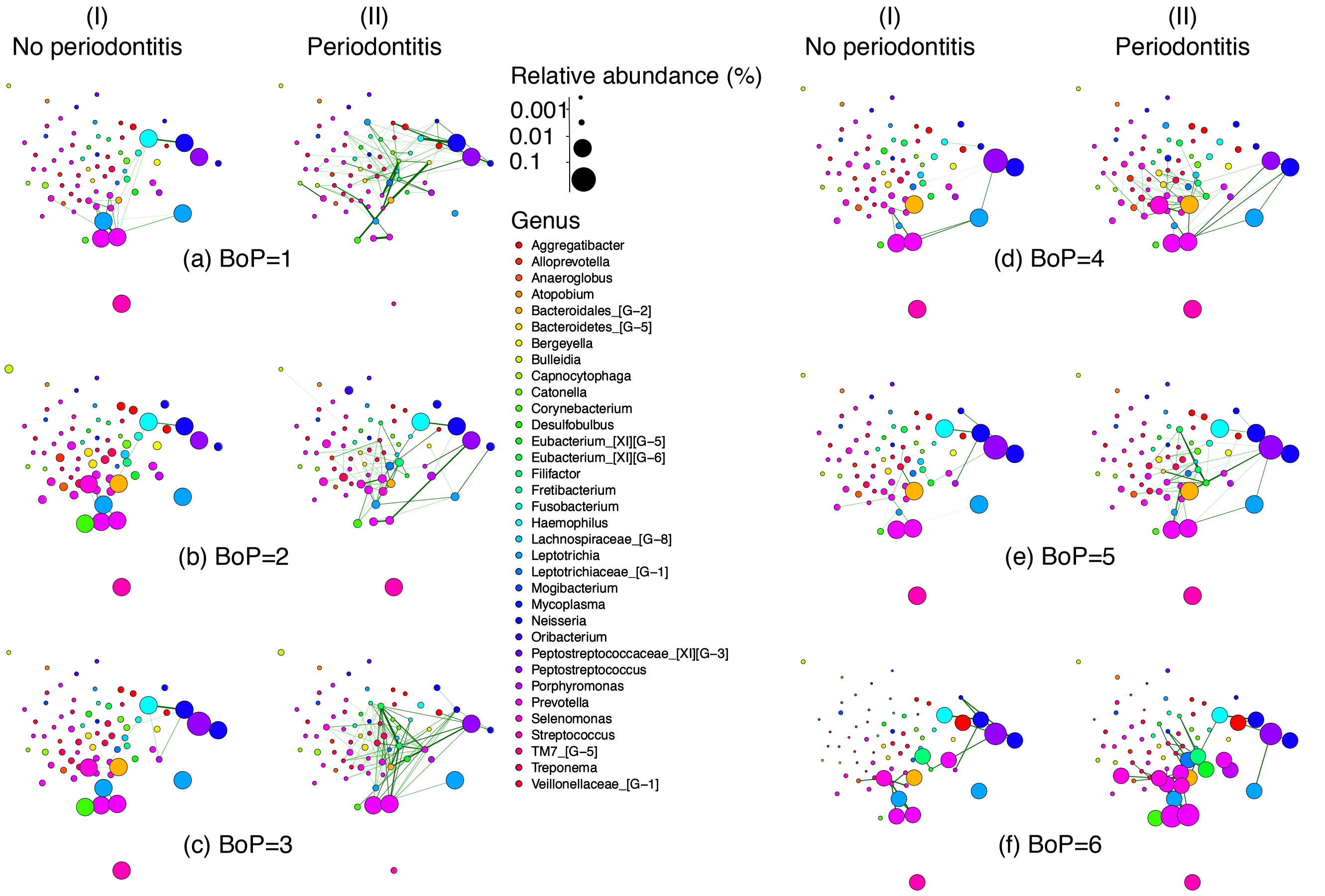


Figure A.6: Histogram of the numbers of teeth with pocket depth greater than 4mm. The long tail of the distribution and lack of normalization by total number of teeth means a simple linear scale for total periodontitis is not appropriate. The red dashed line indicates the cutoff used to define binary periodontitis.

	Constant population			**Exponential growth**		
Alignment	Mean	Median	95% HPD	Mean	Median	95% HPD
IncI2	5.8	5.7	3.3 – 8.9	6.6	6.5	3.9 – 10.2
IncX4	9.5	9.2	4.9 – 15.6	9.8	9.5	5.2 – 15.9
Transposon	7.6	7.4	4.3 – 11.9	7.5	7.3	4.3 – 11.8

Table A.3: Inferred clock rates expressed as substitutions per base pair per year for alignments under different population models. Shown here are relative rates (all $\times 10^{-5}$ for true rate) for the IncI2 and IncX4 plasmid backgrounds, as well as for the *mcr-1*-carrying composite transposon alignment under both constant population size and exponential growth models.



(a) BoP=1

(b) BoP=2

(c) BoP=3

(d) BoP=4

(e) BoP=5

(f) BoP=6

Figure A.7: Co-occurrence networks of periodontitis-associated taxa are more connected in women with periodontitis, independent of gingivitis severity. The co-occurrence network becomes more connected in women with periodontitis after controlling for gingivitis across the spectrum of gingivitis severity. Shown here are significant pairwise Spearman correlation coefficients ($p < 0.01$, $\rho > 0.4$) between periodontitis-associated OTUs in women both without periodontitis (I; left-hand columns) and with periodontitis (II; right-hand columns) at all severities of gingivitis **(a)-(f)** (BoP of 1-6 respectively). Node color indicates taxonomic genus and edge weight indicates the strength of the (positive) correlation between OTUs. Node layout was determined using the Fruchterman-Reingold algorithm in qgraph v1.3.1 on the correlations in (II.f).

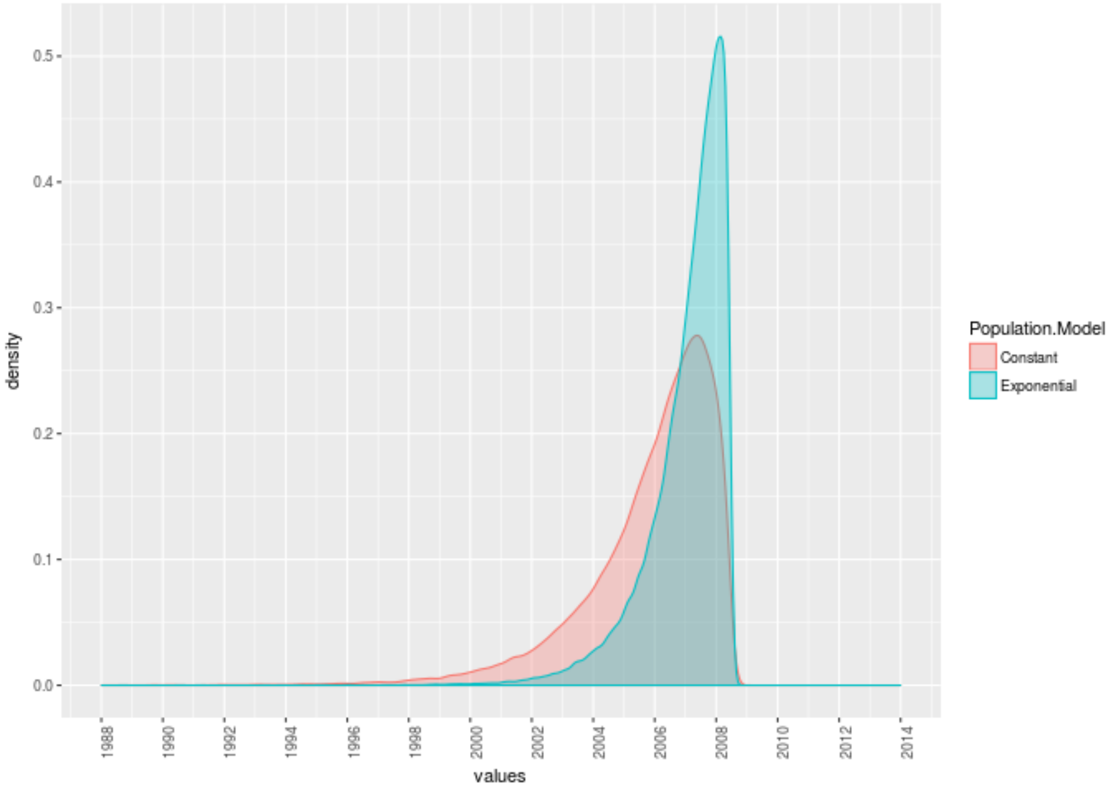


Figure A.8: Posterior density distributions of root height for the composite transposon alignment. Distributions under a constant population size model (pink) and a model of exponential population growth (green), both estimated under a strict clock model. I am grateful to Lucy van Dorp for permission to include this figure.

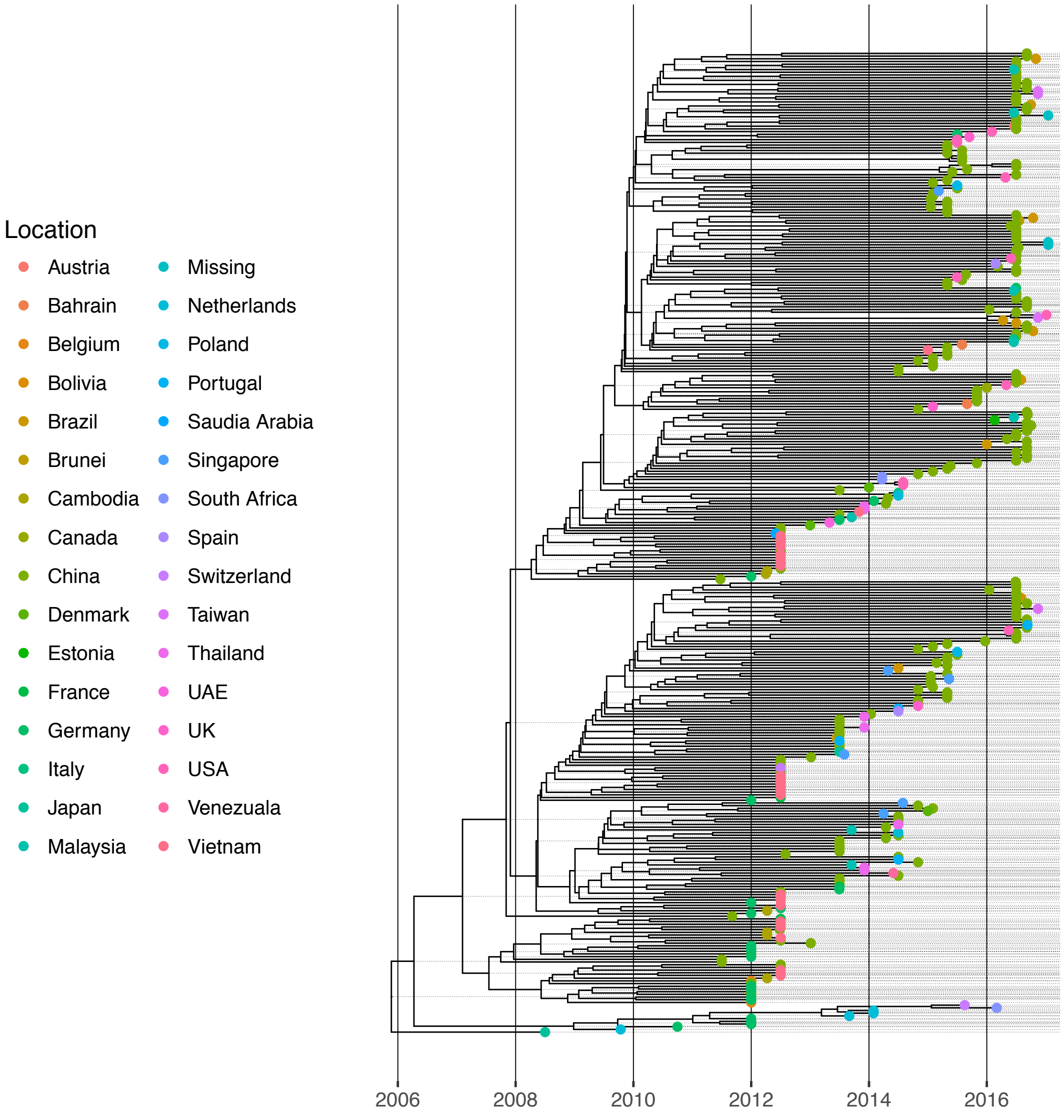
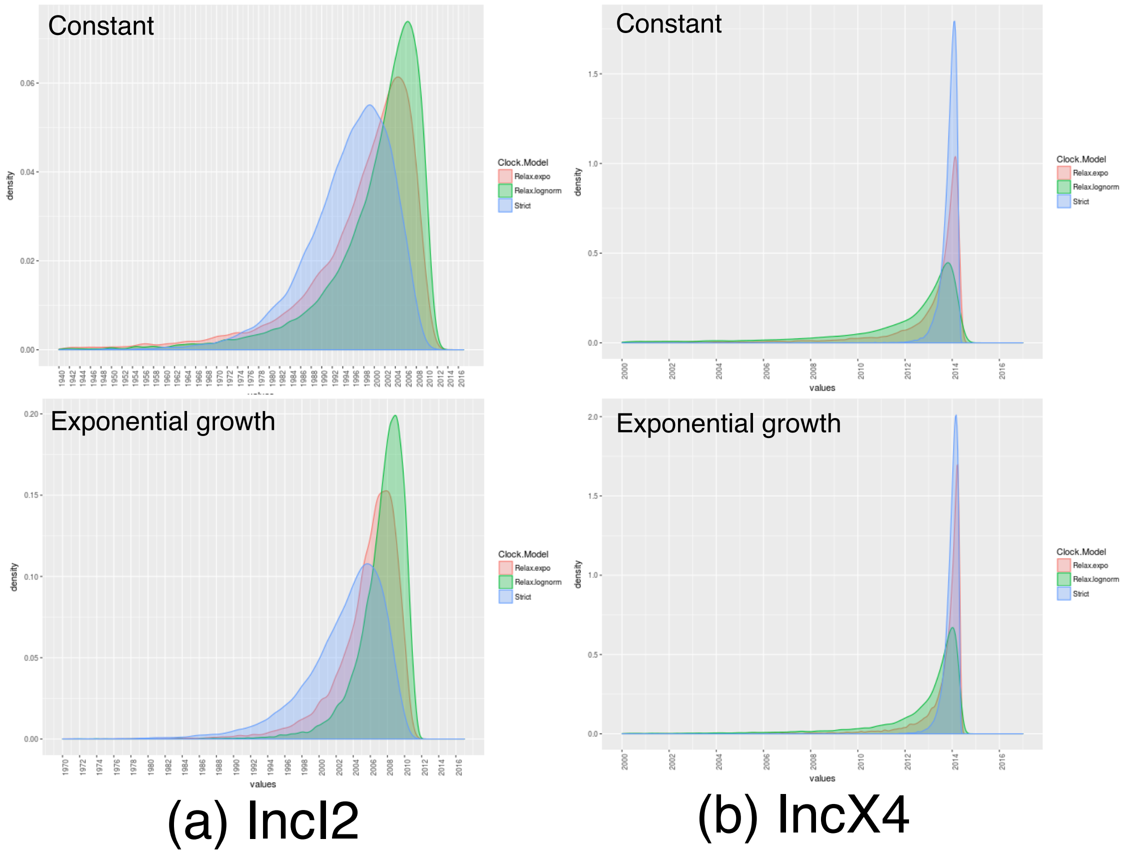


Figure A.9: Beast inferred maximum clade credibility tree for the composite transposon alignment. The timed phylogeny is based on a strict clock model under the coalescent, with tips coloured according to the country of sampling. I am grateful to Lucy van Dorp for permission to include this figure.



(a) IncI2

(b) IncX4

Figure A.10: Posterior density distributions of root heights for the (a) IncI2 and (b) IncX4 plasmid background alignments, assuming either a constant population or exponential population growth. Distributions clock models (colours) are shown under the Coalescent Bayesian skyline implementation. I am grateful to Lucy van Dorp for permission to include this figure.