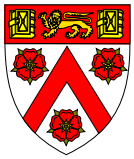
UNIVERSITY OF CAMBRIDGE

Exploring the Population Structure, Recombination Landscape, and Pan-Genome of the Global *Neisseria meningitidis* Population

Neil Z. MacAlasdair



Trinity College

This thesis is submitted on the 30th of June, 2021 for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Neil Z. MacAlasdair
June, 2021

Abstract

Exploring the Population Structure, Recombination Landscape, and Pan-Genome of the Global *Neisseria meningitidis* Population

Neil Z. MacAlasdair

Neisseria meningitidis is a gram-negative species of bacteria which causes meningitis, septicaemia, urethritis, and pneumonia worldwide. Infections are typically asymptomatic carriage, but those which cause disease are extremely difficult to treat, leading to a high case-fatality rate. As such, there is considerable interest in studying *N. meningitidis* to understand its spread, what causes development from carriage to invasive disease, and how its evolution impacts efforts to control the disease. The latter has been of particular concern in regions where there have been outbreaks, particularly the 'meningitis belt' that spans from West Africa to East Africa, where there is greater disease burden and periodic epidemics which can span the region. Due to difficulties in treatment, the primary method of controlling invasive meningococcal disease is vaccination. Currently, available vaccines target five of the extant serogroups of *N. meningitidis*, chosen through study of the serogroups most frequently found in disease. However, either the replacement of disease lineages with those of different serogroups or capsular switching within disease-associated lineages may undermine the success of mass vaccination efforts and create the need for additional campaigns. *N. meningitidis* specifically possesses characteristics which make vaccine escape likely and unpredictable. The most important are the adaptions which allow frequent homologous recombination with other *Neisseria*. The evolutionary consequences of this sporadic partial chromosomal recombination are

not well understood, but the transfer of alleles between distant lineages – including those associated with virulence – has been observed. Another gap in our understanding of bacterial evolution is in the evolutionary effect of population structure. Obilgately human-parasitic species such as *N. meningitids* have a global distribution and opportunities for rapid migration, and therefore may have a complex population structure. To study these problems, I have assembled a collection of over 15,000 whole-genome sequenced *N. meningitids* isolates from 70 distinct countries with isolation dates spanning over a hundred years. These data consist of a mixture of publicly published data, and three collections of newly sequenced isolates. Using these data, I determine the global population structure of *N. meningitids*. Subsequently, I infer phylogenetic trees for and find patterns of recombination within major lineages in the global population. Separately, I also infer and analyse the species-wide pan-genome. The results of these analyses indicate that *N. meningitidis* has a deep well of generally unsampled diversity in an extremely complex population structure which is primarily made up of a few globally distributed lineages. Within these lineages, population bottlenecks are a frequent occurrence. The 25 major lineages differ significantly in both their rates of recombination and the distribution of recombination across their genomes, but evidence suggest that most recombination occurs within *N. meningitidis*. In a local population, recombination generally acts to reduce the effect of deleterious mutations, although an example also exists of recombination acting in concert with positive selection. The pan-genome reveals the extent to which recombination can disrupt tree-like evolution, with most major lineages containing patterns of relatedness in their accessory gene content inconsistent with their whole-genome phylogenies. Trends in the pan-genome indicate that most gene gain is from other *N. meningitidis* isolates, but is governed primarily by evolutionary forces and not recombination rate. Together, these results demonstrate the profound complexity present in the population structure of *N. meningitidis*, and distinct evolutionary trends in individual lineages. This work also underscores the importance of carriage sampling and the value of a global perspective when

studying a globally-distributed species. Further sampling in regions which are under-sampled and ongoing carriage surveillance will be a crucial part of any long-term efforts to successfully control the disease through vaccination.

Acknowledgements

I must begin by acknowledging those with whom I have engaged in fruitful collaboration during the course of my doctoral studies. In particular, I would like to acknowledge those with whom I co-authored the manuscript entitled "*The effect of recombination on the evolution of a population of* Neisseria meningitidis" (*Genome Res*, 2021) results from which are included in Chapters 3 and 4. Their assistance with analysis methodology, insightful commentary on the results, and patience with my occasionally lackadaisical pace in making revisions was much appreciated during the entire process of drafting and revising the manuscript. Furthermore, although they have not contributed directly to this thesis, co-workers over the past four and half years have shaped my thinking enormously, be it in the office on the squash court, or at the pub. Aaron Weimann, Chris Ruis, Leonor Sanchez Buso, Stephanie Lo, Dorota Jamrozy, Christine Boinett, Chrispin Changuza, and Sophie Belnman have been a pleasure to work alongside. I must also acknowledge the direct funders of this research, the Wellcome Trust, who have underwritten my stipend and all of the novel sequencing associated with this thesis, and also the Meningitis Research Foundation, who contributed funding for the storage, culturing, and extraction of the MenAfriCar isolates.

Separately, I must extend my gratitude to the entire *Neisseira meningitidis* research community, who by and large, have fastidiously shared the raw sequence data and metadata generated in the course of their research. Without this open attitude towards data sharing, the analysis presented in this thesis would simply not be possible. In no particular order, designated thanks must go to the Menigicoccal Reference Unit at Public Health England, who do the routine sequencing of cases in the UK and

performed the DNA extractions for the MenAfriCar sequencing. From the MRU, I must specifically thank Jay Lucidarme and Aiswarya Lekshmi, who did the majority of the MenAfriCar DNA extractions and also spent two long days teaching me how to safely culture and extract DNA from *N. meningitidis* two years after I had last been in a laboratory. My thanks go to the rest of the MenAfriCar consortium as well, who spent years of considerable effort collecting an enormous dataset. I must also extend my gratitude to Dominique Caugant at the Norwegian Institute of Public Health, who provided the global archival dataset and led the collection of the Burkina Faso dataset, both studied in this thesis, as well as Ingerid Kirkeleitte who performed the DNA extractions for both these collections. My appreciation also goes out to the DNA sequencing pipeline team at the Sanger, who were responsible for doing all of the sequencing which underlies the research I have done. Similarly, I would like to thank the core and pathogen informatics teams at the Sanger, who have maintained the robust computational infrastructure integral to this research, and put up with my annoying requests and suggestions, respectively. Finally though they have not directly impacted this work, I owe my thanks to the maintainers of *Neisseria* pubMLST and the Meningitis and Vaccine Preventable Diseases Branch of the US's CDC, whose active sharing and maintenance of repositories to share *N. meningitidis* sequence data and metadata has much enriched this work.

I am personally extremely grateful to my supervisors, Stephen Bentley, Julian Parkhill, and Caroline Trotter. I am quite sure that I have not always been the easiest student to supervise, and their guidance, encouragement, and above all, patience has been fundamental to my scientific development during the course my doctoral studies. Special thanks must go to Ste, whose supervision has been exactly what I needed during my PhD. He also has recently had the unenviable task of closely supervising the preparation of this thesis during a global pandemic and while other personal circumstances have greatly disrupted the speed and quality of my writing. His practical guidance and comments on draft chapters under such challenging circumstances, coupled

with quiet encouragement and good humour have made what at times seemed like an impossible task a reality.

My deepest gratitude is also extended to my parents, Liling Zhang and Duncan MacAlasdair. They have given me a lifetime of love and support, even as they have become progressively more confused by what it is exactly that I am doing. Above all, however, their constant faith in my ability and desire to do what is in my best interests has, more than anything else, enabled me to even consider undertaking doctoral study, and for that, I am forever grateful. I also greatly appreciate how my girlfriend, Alexandra Forrester, has supported me throughout the process of drafting this thesis. She has been a steadfast source of love and support, sacrificing substantial amounts of her time and sleep – even as she has been grieving the death of her mother and working on her own doctoral thesis – to ensure that this thesis could be successfully completed. I would also like to thank some of my old friends, who have not only put up with my moaning about the thesis, but also supported me and each other during a difficult time. In *alphabetical order by last name*, thanks must go to Arjun Biyani, Kenneth Chow, Sebastian Thomas, and Jeffrey Weaver. I would like to also thank Gerry Tonkin-Hill, who has been a kind friend, enormously supportive both professionally and personally during the course of my doctoral studies, and Ian Forrester, who has not only graciously welcomed me into his home for the past 9 months, but also done his best to support my unusual work hours. Special thanks must also be made to the late Cassiopeia, Alex's cat, who was the best thesis-writing-in-times-of-difficulty companion any human could have wished for, and who died shortly following the submission of this thesis.

Finally, I come to the acknowledgement I have been dreading. Thanks are quite insufficient for my dear friend of twenty years, the late Matthew Brennan, who died on the 26$^{\text{th}}$ of January, 2021. Four days earlier, we had the last academic discussion we would ever have, where I double-checked my understanding of the statistical theory underlying some of the methods presented in detail in Chapter 2 of this thesis. Matt's presence is pervasive in all aspects of my life, but the preparation of this thesis has highlighted the absence of our late-night calls to have a break

and a laugh, ask him about math and LaTeX, and commiserate over the shared frustrations of being postgraduate students. Completing this thesis would not have been possible without my certainty that it is what he would have wanted. Even in death, his example has continued to inspire me throughout this difficult process. Dr. Matthew Brennan received his PhD in Electrical Engineering and Computer Science posthumously on the 4^{th} of June, 2021, the conferral of which included an animated wooden rotor-copter carrying a diploma followed a jolly dancing beaver and a powerpoint slide with an AI voice synthesiser reading his name. I think he would have been amused – such a ceremony would have been "Great…", and made it all worthwhile.

Contents

chapter 1

Introduction

Neisseria meningitidis – the meningicoccus – was first observed by human eyes in the year 1884, in a sample of cerebrospinal fluid from an unknown patient [1], and first isolated in 1887[2] less than thirty years after the theory of evolution was first published [3], yet more than ten years before the widespread acknowledgement of Mendelian genetics [4], and more than twenty years before the first papers were published beginning the work of the modern synthesis [5]. All of this, of course, happened at least half a century before DNA was identified as the molecule of heredity which, in turn, was twenty-five years before a method of genome sequencing was first developed in 1977. The first whole-genome sequenced *N. meningitids* isolates were published in the year 2000 [6, 7]. In the one hundred and sixteen years between the first observation of *N. meningitidis* and the reading of its genome sequence, the various fields of medicine, genetics, evolution, microbiology, and molecular biology made enormous progress. The diversity of approaches to biological research in the 20^{th} century, however, has often led to a fragmented understanding of some fundamental questions, including the complexity of bacterial evolution [8, 9]. This thesis is concerned with the genetics of *N. meningitids*, a gram-negative obligately human-commensal proteobacteria, which colonises and inhabits musosal surfaces in *Homo sapiens*, occasionally causing severe disease. It is especially interested in using genetic information to understand the evolution of *N. meningitidis*, in examining how knowledge of its evolution may impact the ability

to control the disease it causes through vaccination, and understanding how its evolutionary history and dynamics sit in the context of research into bacterial evolution, generally. Details of the methodology, results, and insights from this research will form the bulk of this dissertation, Chapters 2-6. In this chapter, however, I briefly examine the wider motivation for this research by discussing what is presently known regarding meningococcal disease, the genetics of *N. meningitidis*, and bacterial evolution.

1.1 Meningococcal disease

N. meningitidis is so called because on the rare occasions that it causes an invasive disease, a small fraction of all infections [10], it most frequently causes disease by moving from asymptomatic carriage in the nasopharynx to infecting the meninges – the protective membranes covering the brain and spinal cord – though it can also infrequently cause infections at other sites [11]. Meningitis caused by *N. meningitidis* is often life-threatening. Since the 1980s, when modern treatment using antimicrobials began, cases of the disease have a fatality rate of around 9-12% [12], though this rises to around 40% in cases where *N. meningitidis* also causes sepsis. Among the 80-90% of patients who are successfully treated and survive, up to 20% of cases meningococcal disease lead to lifelong sequelae including deafness, amputations, and mental impairment [13–15]. Even in cases without sequelae, survivors often experience an adverse effect on their health-related quality of life [14]. Unsurprisingly, there has therefore been significant interest in reducing the disease incidence with vaccination [16]. The first vaccines targeting *N. meningitidis* were developed in 1969, targetting two of the dozen serogroups [17] (serogroup A and serogroup C) of *N. meningitidis* [18]. Since then vaccines have been developed to target three additional serogroups. A vaccine targeting serogroups W and Y were developed in 1981 [19], and vaccines targetting serogroup B were first developed in 2011 [20]. Vaccines have been widely used in order to prevent disease [21–23], though routine vaccinations are relatively recent [16].

Despite the effectiveness of recent immunization programs

[24], the global disease burden of invasive disease caused by *N. menigitidis* remains high enough for the World Health Organisation to include it as part of it's 2020 roadmap *Defeating meningitis by 2030* [25]. This is in part due to the variable nature of the epidemiology of *N. meninigitis* around the world. In general, incidence is low as *N. meningitidis* is a sporadic, opportunistic pathogen [10]. However, incidence can range from 1 in 100,000 to 100 in 100,000 in different times and parts of the world [24, 26], particularly when there are epidemic outbreaks, a feature that is particularly associated with *N. meningitidis* among the meningitis-causing bacteria [10]. This is especially the case in a region of Africa commonly known as the 'meningitis belt' [27], where there is a generally higher incidence, but also frequent epidemics every two to five years [28]. A mass serogroup A vaccination campaign, MenAfriVac$^{\text{TM}}$, has taken place across the meningitis belt [29], and although it resulted in a substantial decrease in the incidence of *N. meningitidis* generally [23, 30, 31], outbreaks of disease due to non serogroup A have been observed [32, 33], leading to concerns that there may be a rise in disease caused by *N. meningitids* not covered by vaccines [34].

This concern is echoed globally due to the fact that of the twelve serogroups of *N. meningitidis*, [17], six of which (A, B, C, W, X, Y) cause the overwhelming majority of disease [10, 24, 35], only five (A, B, C, W, Y) are covered by currently available vaccines. The reported expansion of serogroup X disease [36], as well as cases of detected *N. meningitidis* invasive disease caused by serogroup E [37] and non-groupable/capsule null isolates [38, 39] suggests that a strategy of vaccination against currently prevalent disease-causing serogroups may be insufficient. Long-term control of the disease and further reduction in its incidence relies upon not only developing new vaccines – like the pentavalent vaccine effective against serogroup X currently under development [40] – but also taking into account the possibility of strain replacement or capsule switching, as has been observed in other species [41, 42]. In order to do so successfully, there must be a robust understanding of what the evolutionary consequences of a vaccine roll-out would look like.

This in turn depend upon two things. A thorough understanding of the genetics of *N. meningitidis*, and a general understanding of bacterial evolution.

1.2 The genetics and genomics of *N. meningitidis*

N. meningitidis has a single, circular chromosome of roughly two mega-base pairs and two thousand genes [6, 7, 43]. Like all bacteria, *N. meningitidis* reproduces by duplicating its chromosome and dividing into two daughter cells, each with a copy of its chromosome. Despite this theoretically entirely clonal descent, it was observed through laboratory study that bacteria of *N. meningitis* also routinely transport exogenous DNA inside their cells and incorporate it into their chromosome [44]. While this is not unknown in bacteria [45], it was soon discovered that not only does *N. meningitidis* engage in recombination with exogenous DNA, but its genome also contains specific mechanisms to promote such activity. A 10 base-pair DNA sequence was discovered to be particularly effective [46] at causing what is referred to as 'transformation', the genetic alteration of a cell through the introduction of external DNA. After this initial discovery, subsequent decades of research have revealed further mechanisms which have evolved to facilitate homologous recombination within *N. meningiditis* [47]. These include the genus-specific DNA uptake sequence which promotes the uptake of DNA fragments containing that sequence into *Neisseria* bacteria [48], a variety of membrane proteins which variously identify and bind exogenous DNA containing the uptake sequence (ComP) [49], and proceed to transport it across the cell membrane at a constant rate [50]. After uptake, a series of proteins process the imported DNA [51], and genes from the Rec pathway appear to then affect homologous recombination between those fragments and the bacterial chromosome [52]. Separately, it has also been shown that *N. meningitidis* maintains multiple copies of the *pilE* locus in the form of silent *pilS* cassettes, which recombine with one another to generate antigenic diversity in the type IV pilus [47], another adaptive mechanism to promote recombination,

although within a single chromosome. Other repeat mechanisms, such as the inverted dRS3 + RS repeat elements that make up neisserial intergenic mosaic elements, have been hypothesised to promote recombination at specific loci in the genome [7, 43], though no laboratory analyses supporting this hypothesis have been published. All together, these various adaptations to enable and promote recombination in the *N. meningitidis* genome have meant that, despite its nominally clonal reproduction, recombination plays an important role in its evolution. Estimates have often implicated recombination as the primary mechanism through which diversity is generated in *N. meningitidis* [53].

As a result of the extent to which *N. meningitidis* recombines, efforts to understand the genetic structure of its populations have occasionally described it as "panmictic" [54] or "freely recombining" [55, 56], and therefore lacking an identifiable population structure. Research using the Multi-Locus Sequence Typing (MLST) method for typing bacteria, initially developed for use in *N. meningitidis* [57], has demonstrated that bacteria with the same sequence type – or ST – can be grouped into lineages of isolates which have recently reproduced clonally [10]. MLST uses patterns of nucleotide variation (alleles) in seven different genes (six originally [58]), of the *N. meningitidis* genome to classify isolates into lineages, where ST-1 for instance, has alleles 1, 3, 1, 1, 1, 1, 3. Different patterns of different alleles at the seven loci make up the 8000 sequence types which have been identified as of June, 2021 [59]. Sequence types from MLST data, like the multilocus enzyme electrophoresis (MLEE) types, based on proteins which proceeded MLST [60], can be reduced into clonal complexes, which in *N. meninigitidis* group together sequence types which have up to two mismatches in their allele types [61] – for instance an isolate with allele types 1,1,1,1,1,1,1 would be of the same clonal complex as the previous example. The ease of identifying the sequence type of an isolate of interest – only seven genes need to be typed – has meant that even early in the genomic era, it was possible to quickly and cheaply type relatively large numbers of isolates [62].

The use of MLST to describe and categorise *N. meningitidis* prompted substantial research into the structure of the

N. meningitidis population, and the identification of virulent lineages which appear to be consistently identified in cases of disease and outbreaks, yet rarely found in carriage [10]. Of these lineages, the ST-11 clonal complex has been one of the most well studied. It was originally identified in outbreaks in the United States and then Europe [63], before causing a large outbreak associated with the hajj pilgrimage [64], and spreading first to the meningitis belt [65], and then eventually to three other continents [66–68]. This lineage has continued to be closely studied, and recent research has identified additionally invasive variants of the lineage which have spread in Europe [69], and even a variant of this lineage which has become associated with urogenital infections [70]. Another lineage identified through MLST which has been very well studied is the ST-5 clonal complex, which, prior to large-scale vaccination against serogroup A *N. meningitidis* caused several epidemics of invasive disease across the meningitis belt in the 2000's [71]. The comparative study of MLST lineages of *N. meningiditis* has led to some insights regarding its evolution. These include the identification of novel restriction-modification systems between different clonal complexes [72], the importance of selection in generating the clonal lineages of *N. meningitidis* [73], and the importance of recombination in the evolution of *N. meningitidis* [74].

1.3 *N. meningiditis* and bacterial evolution

The use of multi-locus sequence typing has consequently proven enormously powerful in studying the genomic epidemiology of *N. meningitidis* worldwide, allowing for the identification and tracking of virulent lineages, and helping to guide public health intervention. It has even allowed for some comparative research between species [75, 76]. Despite its successes in these regards, MLST-based study has been unable to discern any relationship between different outbreaks of the various clonal complexes, or explain how the clonal complexes manage to persist between outbreaks [77]. Using a whole-genome approach to studying the genetics and evolution of *N. meningitidis* has

several downsides, namely that computation takes much longer, and results are often far less readily interpretable than typing methods. However, the increased resolution such whole-genome methods offer [78] may shed some light on hitherto unresolvable problems.

In the two decades since the first whole-genome sequenced *N. meningitids* isolates were published [6, 7], the number of sequenced *N. meningitidis* has grown by a factor of approximately 10^4. This amount of whole-genome data is "big" in the biological context, as it currently remains beyond what typical software for the purposes of the analysis of bacterial data can handle (Section 2.2.2). This has led to the development of new methods which are designed to account for datasets of such size[79, 80], but these typically partition the data into biological subsets for further analysis instead of directly working on whole-genome data at scale [79]. The reasons for this are numerous, but it is partially a result of the fact that despite this enormous increase in the amount of sequencing data available not only in *N. meningitids*, but also in many other bacterial species[81, 82], many fundamental questions in the study of bacterial evolution remain unanswered [9]. In the century since the modern synthesis was first proposed, the overwhelming majority of evolutionary research has been focused on understanding evolution in organisms with a ploidy greater than two and which are randomly mating and sexually reproducing, and are therefore capable of being at equilibrium. Even research into populations which violate the assumptions allowing for Hardy-Weirnberg equilibrium for one reason or another – assortative mating, polyploidy, small population sizes, migration, introgression – often fails to consider haploid microorganisms, on the basis that their clonal reproduction makes their population structures and evolutionary dynamics relatively simple [8]. The significant increase in the number and diversity of genomes sequenced in many bacterial species has demonstrated that this is not the case, and has led to a number of alternatives being proposed [56, 83], one particularly with *Neisseria* in mind [55]. Despite the recent increase in studying evolution from the perspective of haploid microorganisms, many of the the questions which have been studied in

detail in eukaryotes, such as the reasons for the evolution and maintenance of recombination [84–86], remain poorly studied in bacteria [87]. Many of these questions are of substantial clinical relevance, however, in the public health management of diseases caused by bacteria [42]. This is particularly the case in species which maintain large populations in asymptomatic infections [88–90] around the world, and are primarily managed through vaccination [16]. The use of vaccines on a large scale creates a strong selective pressure in these species, and understanding how that shapes the evolution of these pathogens at the global level, as has been called for in *N. meningitidis* [91], is crucial for designing a successful vaccination strategy [92].

This is the context in which the research from this thesis is nested, with detailed discussion of the data and theoretical basis of the analysis methodology to follow in Chapter 2. *N. meningitidis* is a well-studied bacteria that is recombinant and possesses specific adaptations to facilitate unique levels of recombination, shared with other members of its genus [93]. It causes a serious, life-threatening illness which has been effectively managed with vaccination, but the long-term success of these efforts is not guaranteed as *N. meningitidis* may evolve to reduce the effectiveness of current vaccine strategies. To ensure the success of these efforts in the future, we must develop comprehensive understanding of how bacterial populations evolve. The population genetics of *N. meningitidis* have been well-studied at the regional level using MLST data, but a need for research into the population at higher resolution [78] and a global level has been identified [91]. To address this need, I have assembled a collection of over 15,000 whole-genome sequenced *N. meningitidis*, which I use to assess the population structure at a global scale, the extent of variation in recombination across the species, and its pan-genome evolution.

CHAPTER 2

Data and Methodology

2.1 The Global *Neisseria meningitidis* Dataset

This thesis aims to study the genetic diversity of *Neisseria meningitidis*, to understand how the population of organisms making up this species is structured at a global scale. Also, it aims to ascertain the extent of gene flow within the species, and to investigate how genetic variants within the species may be responsible for observed variation in bacterial phenotypes. We are able to study these phenomena thanks to significant advances in sequencing and population genomics analysis methods over the last two decades, allowing for large-scale sequencing of thousands of samples and their subsequent analysis. This is particularly true in the case of bacteria that infect human hosts, which are relatively straightforward to sequence due to their small genomes, the ability to culture samples for many species, and ease of sampling their host organism. This has allowed whole-genome population genetics at a scale unprecedented in non-prokaryotic species (with the exception of *Homo sapiens*), involving sequencing on the order of 10^4 individual organisms from the species' population. As such, there are now efforts to conduct novel surveys of the global populations of various prokaryotic species [82]. This remains, however, enormously expensive due to the costs and logistics involved in collecting samples around the world. However, for many bacterial species,

particularly those relevant to human health and disease, the past decade of genomic research has produced many smaller-scale population genomic surveys on the orders of 10^2 or 10^3 sampled organisms, all around the world. In order to investigate the global population of *Neisseria meningitidis* this study will amalgamate dozens of these datasets, some published, some newly sequenced, with a newly-sequenced archival collection of *Neisseria meningitidis* spanning almost 100 years. Gathering this dataset, as will be detailed in this chapter, produces a collection of 15,450 (post quality control) *Neisseria meningitidis* isolates, which captures enormous temporal and geographic diversity in its sampling, and allows us to perform in-depth investigations into the global population of *Neisseria meningitidis*. Compared to designing a global sample population, using an amalgamated dataset has some drawbacks, most notably the imbalanced nature of the sampling overall, and the unrepresentative nature of many of the smaller datasets. Despite this, however, the added depth of sampling provide by amalgamating all of the data is such a significant advantage that being careful with the interpretation of any results is preferable to down-sampling to generate a more representative dataset.

2.1.1 Newly sequenced data

Three separate new sequencing projects have been included in this study, with a combined total of 6,332 newly sequenced isolates (before quality control). These projects were collected long before my involvement, which has been limited to co-coordinating their sequencing and the subsequent data analysis. These three projects are: the MenAfriCar African meningitis belt carriage collection (1449 samples), the Burkina Faso carriage collection (2839 samples), and an archive of global samples, stored at the Norwegian Institute of Public Health (2046 samples). These three studies fill important gaps in the extant sequenced samples currently published and publicly available, and as such were essential in order to enable this effort to study the global population of *N. meningitidis*. As they are newly collected and sequenced, and, at the time of writing not yet published, we will hereafter describe their collection, preparation,

and sequencing in some detail.

2.1.1.1 The MenAfriCar dataset

The MenAfriCar isolate collection are the stored isolates from a large-scale series of carriage surveys in the meningitis belt, between August 2010 and October 2012, inclusive. The MenAfriCar study was initially established to monitor the prevalence of serogroup A *N. meningitidis* carriage in the populations of 7 countries of the meningitis belt – Chad, Ethiopia, Ghana, Mali, Niger, Nigeria, and Senegal – before and after the administration of a meningitis-belt wide vaccination campaign against serogroup A meningococcal meningitis (MenAfriVac^{TM}). Within each country, oropharyngeal swab samples were collected from two study sites, an urban site and a rural site, as shown in Table 2.1. Sampling was organised into two main studies, a cross-sectional survey of households within the study sites, and a follow-up longitudinal household cohort study. In the cross-sectional survey, randomly selected individuals were swabbed and samples plated out in the field, and returned to the collaborating centres for further microbiological analysis and storage. Upon successful identification of *N. meningitidis* in a sample, the entire household from the sampled individual were invited to join the longitudinal follow-up studies, where every member of the household was swabbed every two weeks for two months, and then monthly for a further four months. In total, 48,490 swab samples were collected as part of the cross-sectional sampling, from which 1687 laboratory-confirmed isolates of *N. meningitidis* were identified. An additional 9809 nasopharyngeal swab samples were collected as part of the follow-up longitudinal studies, resulting in a further 991 laboratory identified *N. meningitidis* isolates. Molecular and epidemiological analysis on these samples was conducted by the MenAfriCar consortium, and published in two publications [94, 95]. The boilates (boiled culture pellets) used for those analyses were not suitable for whole-genome sequencing, so to enable this project, additional DNA extractions needed to be performed. The Meningitis Research Foundation (MRF) awarded a grant for this purpose, and samples were shipped to collaborators at the Meningococcal

Country	Site	Number of Isolates
Mali	Bamako, Djicoroni-para (urban)	61
	Narena and Siby (rural)	61
Niger	Yantala (urban)	384
	Say (rural)	416
Ghana	Navrongo town (urban)	79
	Kassena-Nankana District (rural)	178
Nigeria	Konduga (rural)	1
Senegal	Fatick (urban)	30
	Niakkar (rural)	122
Chad	N'Djamena Sud (urban)	23
	Mandelia (rural)	89

Table 2.1: Table of sampling locations in the MenAfriCar carriage survey, and the number of isolates successfully sequenced from that location in the whole-genome dataset.

Reference Unit of Public Health England, where isolates were cultured and extracted according to their protocol. In brief, samples were cultured and extracted in the following way:

Frozen sample beads were plated out and cultured overnight on Columbia agar plus horse blood. Samples were inspected the next day to detect signs of contaminating bacteria, and plates with evidence of substantial contamination from other species were sub-cultured overnight onto a new plate, from at least five colony picks, in an effort to isolate pure samples of *N. meningitidis*. Plates of single cultures were then swabbed to collect material for DNA extraction, after which cells were lysed at 80 degrees. DNA extraction was then performed using a Promega Wizard© Genomic DNA purification kit, and stored between 2°C and 8°C. Original culture plates were kept overnight for at least one additional night, and those that developed signs of contamination had their samples discarded, and were then sub-cultured and extracted again. Extracted DNA was sent on dry ice to the Wellcome Sanger Institute (WSI) for sequencing, where it was sequenced using Illumina Hi Seq X sequencing technology, with 150 bp paired-end reads.

These DNA extractions led to 1449 successfully sequenced isolates. Substantial metadata is available for these samples, including not only the date of isolation, location, and the age and sex of the carrier, but also information about the household, such as the presence of risk factors (such as open stove, smoking)

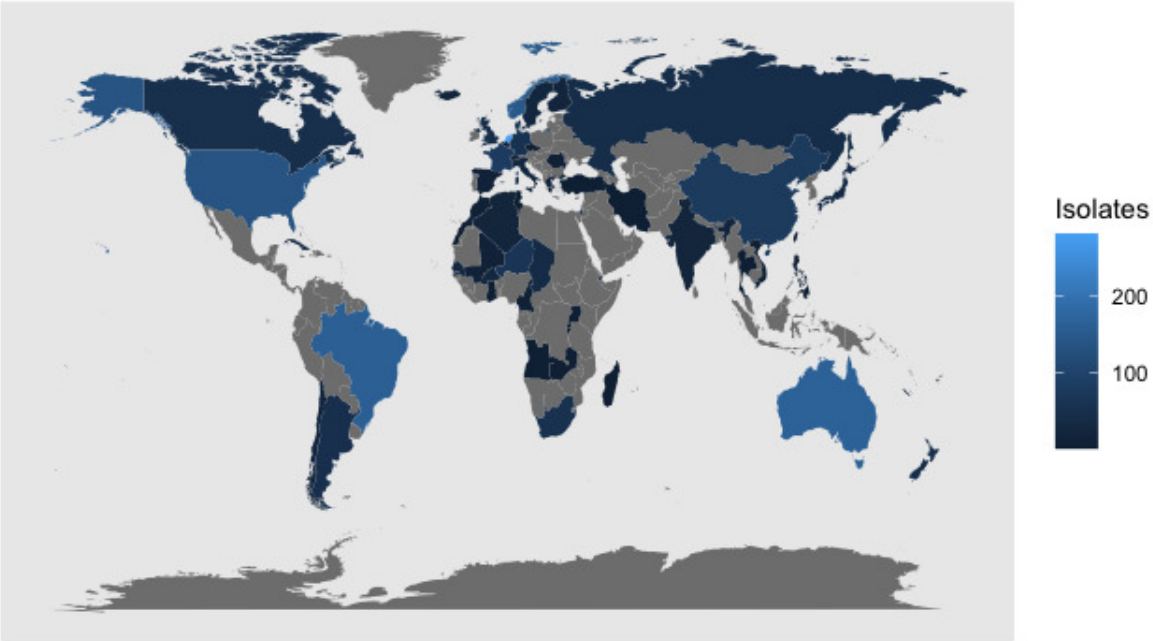


Figure 2.1: Map of the distribution of isolates which make up the global archival *Neisseria meningitidis* collection

or the number of inhabitants and their ages. The number of samples successfully sequence from each survey site is shown in table 2.1.

2.1.1.2 The Global Archival collection

The Global Archival collection was provided by Professor Dominique Caugant of the Norwegian Institute of Public Health (NIPH). It encompasses nearly a hundred years of sampling, between 1915 and 2008, six continents, and 58 countries, as shown in Figure 2.1. These samples were collected in various settings over the course of decades. These samples were stored at -80°C and were reawakened, grown, and extracted at the NIPH in Norway. Sample DNA was then shipped to the WSI on dry ice for sequencing, again using Illumina Hi Seq X sequencing technology, with 150 bp paired-end reads.

The DNA extractions and sequencing from this project led to 2046 successfully sequenced samples. As these samples form part of an archival collection, and not a novel sampling project, the available metadata is relatively limited. Year and country of isolation is available for every isolate, but a more specific geographic location is only available for 517 isolates, or 25.2%

of the isolates. Disease manifestation is known for a greater proportion of isolates, 1999 out of 2046, or 97.7% of isolates, but the anatomical sample site is known for a smaller number of isolates, 266, or 13% of the isolates. Again, due to the nature of this collection, beyond these limited metadata, there is no further information about the carriers or their household environment.

2.1.1.3 The Burkina Faso carriage collection

The Burkina Faso carriage collection is, similar to the MenAfriCar carriage collection, a large-scale sampling effort conducted in parallel with a mass serogroup A vaccination campaign. As Burkina Faso was the first country to implement the campaign, the sampling occurred earlier than MenAfriCar, over the course of four years, from 2009-2012, in 10 rounds of sampling. Sampling took place across three sites in Burkina Faso – one urban the Bogodogo arrondissment of the capital city, Ouagadougou; and two rural, 10 villages in the Kaya district, 100km north-east of Ouagadougou, and 10 villages in the Dandé district, 350km west of Ouagadougou, as shown in Figure 2.2. Oropharyngeal swabs were taken from healthy volunteers, and associated metadata were collected from the individuals. 50,811 samples were collected over the course of the sampling period, and a total of 2848 meningococcal isolates were recovered from the 10 rounds of sampling and confirmed as *N. meningitidis* at the Norwegian Institute of Public Health (NIPH), Oslo. The isolates were stored frozen in Greaves medium at -70°C. DNA was extracted from 2839 of the 2848 of these collected isolates at the NIPH, and these were sent on dry ice to be sequenced at the WSI, using Illumina HiSeq 2000 sequencing technology and 125bp paired-end reads.

DNA extractions from this study led to the successful sequencing of 2838 isolates. As with the MenAfriCar dataset, there are considerable metadata available for each isolate in the Burkina Faso carriage collection, including the date of sampling, the precise location of the sampling, and the age and sex of the carrier. Some information about the household of swabbed individuals was also collected, though no specific details about

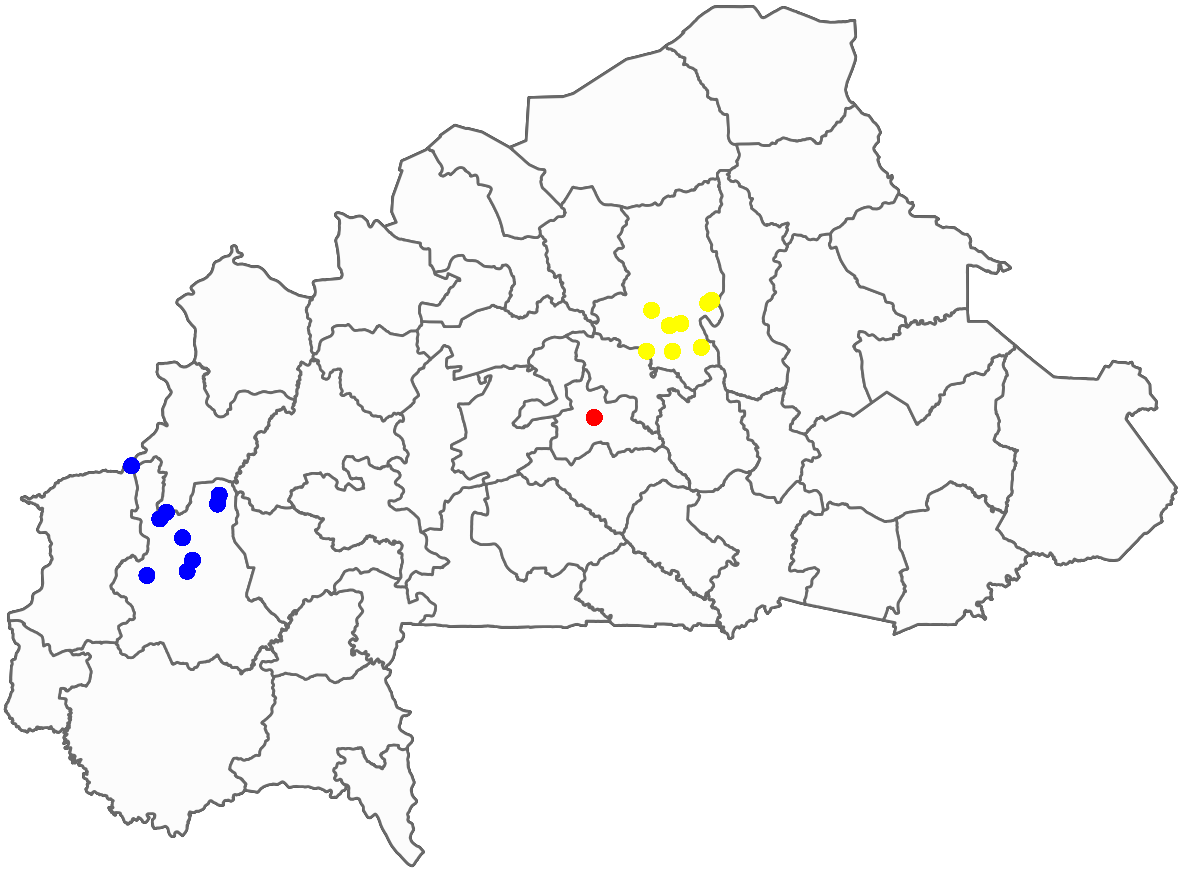


Figure 2.2: Map of the distribution of sampling sites from which isolates were sampled in the Burkina Faso carriage collection. Ougadougou sites are in red, Kaya sites in yellow, and Dandé sites in blue.

other individuals in the household were collected.

2.1.2 Publicly available data

Although the projects which make up this research have generated 6,333 new whole-genome sequences of *N. meningitidis* from around the world, the quantity and distribution of the isolates sequenced in those projects is not enough to, on their own, conduct an analysis of the global population of *N. meningitidis* as this project aims to, as the global archival collection is a very shallow sample from across the world (Figure 2.1). To increase the depth of coverage in regions other than the meningitis belt for these analyses, we will rely upon an additional 9,408 publicly available whole-genome sequenced *N. meningitidis*. These data have all had their raw Illumina short-read sequencing data uploaded to the mirrored European Nucleotide Archive (ENA)/Sequence Read Archive (SRA) and separately had metadata published or manually made available through upload into the *Neisseria* pubMLST BIGSdb database [59]. There are three primary sources for these data: 1) isolates available on *Neisseria* pubMLST BIGSdb and linked to raw whole-genome sequence through an ENA accession number, 2) a series of large-scale sequencing projects carried out by the United States of America Centres for Disease Control and Prevention (CDC), with sequencing data made publicly available on the SRA and basic metadata made available with the publications, and 3) a number of smaller-scale sequencing projects which had previously been conducted at the WSI and had already been published.

In addition to the short-read sequencing data, for methodological reasons to be discussed in section 2.2, it is important to have a collection of finished genomes, that is, sequenced isolates whose primary chromosome have been assembled into a single contiguous sequence. A collection of these was also assembled, and detailed in Section 2.1.2.4

2.1.2.1 Samples submitted to pubMLST

5,886 samples with metadata available on *Neisseria* pubMLST and whole-genome sequencing were downloaded and integrated

into the dataset used for the research conducted in this thesis, with the final sequences downloaded and added from a version of *Neisseria* pubMLST accessed on the 24th of February, 2020. *Neisseria* pubMLST contains, as of the 13th of January, 2021, 23,817 assembled genomes of *N. meningitidis* isolates. However, for many of these assembled genomes, there is either no link to raw sequence read in the ENA/SRA, so they could not be included, or no metadata with regard to the year or country of isolation, which were intentionally not included here due to the difficulty in interpreting genomic data with incomplete metadata.

There is one notable exception to the above principle for determining whether isolates available on pubMLST should have been included, and that is the study published under the title "*Whole genome sequencing reveals within-host genetic changes in paired meningococcal carriage isolates from Ethiopia*" [96]. The isolates from this study are available on pubMLST, but at the time of accessing the data in February 2020, did not have associated ENA accession numbers. These isolates represent the only large-scale sequencing project in Ethiopia, as for various reasons, the MenAfriCar samples from Ethiopia were unable to be extracted and sequenced. Ethiopia is the easternmost part of the meningitis belt, and thus important to include for geographic representation. As such, metadata for these samples was separately obtained through correspondence with the original authors, in addition to the raw sequences being downloaded from the ENA.

Finally, it should be noted that a significant proportion of the publicly available whole-genome sequenced *N. meningitidis* isolates available on *Neisseria* pubMLST are due to the funding and effort of the MRF Meningococcus Genome Library, a project to sequence all disease isolates in the UK between the years 2009 and 2013, which was then followed by routine sequencing of all disease cases within the UK in the years since. It was developed by Public Health England, the Wellcome Sanger Institute, and the University of Oxford, funded by the MRF.

Metadata for pubMLST isolates included in this study is all almost complete. In order to be included in the dataset studied

in this thesis, isolates had to have a defined year and country of isolation. In most cases, isolates also had a known site of isolation, disease state, and some basic information about the host, such as sex and age. For many isolates further information was also available, principally additional geographical information, the month and day of isolation, and the results of AMR screening.

2.1.2.2 USA CDC studies

1941 samples were downloaded and incorporated into the dataset used in this research from three distinct whole-genome sequencing studies carried out by the CDC. These data were uploaded to the SRA, and then published in the following publications:

1. Expansion of a urethritis-associated *Neisseria meningitidis* clade in the United States with concurrent acquisition of *N. gonorrhoeae* alleles (297 isolates) [70]

2. Whole genome sequencing for investigations of meningococcal outbreaks in the United States: a retrospective analysis (201 isolates) [97]

3. Insights on Population Structure and Within-Host Genetic Changes among Meningococcal Carriage Isolates from U.S. Universities (1519 isolates) [98].

The complete details of the sampling and related methodology is published in the respective publications of these studies, but for the convenience of the reader, I will briefly summarise the aforementioned below. Readers familiar with these studies may wish to skip to section 2.1.2.3.

In "Expansion of a urethritis-associated *Neisseria meningitidis* clade in the United States with concurrent acquisition of *N. gonorrhoeae* alleles" there were two primary sources of isolates. The first source of isolates was contemporary sampling made by state public health bodies after the CDC made a request for samples from urethritis cases which tested GNID positive and NAAT negative for *N. gonorrhoeae* on the 17$^{\text{th}}$ of February, 2016 through the Epidemic Information Exchange system. This led to the collection of 209 isolates between the 1$^{\text{st}}$ of January,

2015, and September 30th, 2016. The remaining 88 isolates were take from the archives of the CDC as they were judged to be part of the same clade, containing a mixture of urogenital and non-urogenitally sourced bacteria. DNA was extracted from isolates using ArchivePureTM DNA purification kits and sequenced using either Illumina HiSeq or MiSeq technology, with 250bp reads. For this dataset, key metadata is available, including: year of isolation, the state the isolate was collected in, the clinical site of isolate collection, and disease state.

"Whole genome sequencing for investigations of meningococcal outbreaks in the United States: a retrospective analysis" consists entirely of isolates collected as part of routine disease surveillance in the USA. In particular, the CDC requested cases they classified as "outbreaks" where there occurred 2 or more cases of IMD, identified as the same serogroup, in an "organization", in less than three months, or an "increase in disease rates in a community", between 2009 and 2015. These 'outbreak' cases represent 84 of the isolates published in this study, and the remaining 117 isolates are 'sporadic' disease cases, not identifiable as part of an outbreak, from the national surveillance network, chosen to be in the same 15 states where the outbreaks occurred. DNA was extracted using 5 Prime ArchivePure DNA Purification kits and sequenced using Illumina HiSeq2500 or MiSeq technology and 250bp reads. Available metadata is relatively limited for this dataset, it consists solely of the year the isolates were collected.

Finally, in "Insights on Population Structure and Within-Host Genetic Changes among Meningococcal Carriage Isolates from U.S. Universities", the authors performed 10 cross-sectional carriage surveys at three universities in the USA, two in Rhode Island, and one in Oregon. Four surveys were performed at the university in Oregon, and one of the universities in Rhode Island, following mass vaccination campaigns which had begun after disease outbreaks. Two surveys were performed at the second university in Rhode Island. In total, 8,905 swabs were collected from 7,001 unique individuals, which resulted in the collection of 1,514 *N. meningitidis* carriage isolates. DNA was again extracted from isolates using 5Prime ArchivePure

DNA Purification kits, and sequencing performed using Illumina HiSeq2500 or MiSeq short-read sequencing technology, with 250bp reads. Metadata beyond year and country of isolate are available for the isolates published in this study by virtue of our knowledge of the sampling technique (nasopharyngeal swab), the locations of the universities involved in the study, and the disease status of all the isolates (carriage).

2.1.2.3 Work previously done at Sanger

There have been a number of sequencing projects involving *N. meningitidis* at the WSI, however, for many the metadata is not stored or available within a central database locally. As such, we are limited to including samples sequenced at the WSI which have data stored in the isolate tracking database, or whose authors were able to provide metadata by correspondence. As such, this is limited to two studies. The first, with ENA project accession ERP004245, titled "Integration of the host and pathogen genetics in bacterial meningitis" [99], was a multi-species project with the aim of studying how host genetics and bacterial genetics contribute to the development of bacterial meningitis. Its *N. meningitidis* component consisted of 1086 samples, of which 788 will be used in this study. These samples were collected in the Netherlands as part of the MeninGene cohort, a study begun in 2006 using samples from bacterial meningitis surveillance network of the national reference laboratory, as well as samples stored in their archive. Metadata available for these samples, for the purposes of this dataset, are somewhat limited, including only the year of isolation, the site of isolation, the country, and the disease state, but all isolates meet this minimum criteria.

The other historical project sequenced at the WSI with available metadata is titled "Microevolution of *Neisseria meningitidis* during clonal waves of colonization and disease", with ENA project accession ERP002590 [100]. It consists of 185 whole-genome sequenced *N. meningitidis*, from Burkina Faso and Ghana, all of which are used in the dataset here. Metadata is relatively limited, consisting in most cases of only the country and date of isolation, though in some cases it also includes information on disease status and site of isolation.

2.1.2.4 Assembling a collection of finished reference genomes

With the advent of long-read sequencing, the ability to produce finished genomes has become considerably cheaper and possible to do at scale. As a result, there now thousands of finished genomes available in the RefSeq database. Many of these 'references' are in fact closely related isolates from different clades within an outbreak, and therefore it is not useful to gather all of these sequences to potentially use as references. Instead, a collection of references from sequencing projects with the explicit aim of creating reference genomes, long-read references generated from sequencing projects included in the dataset – in particular the Burkina Faso carriage collection and the Ethiopia carriage collection – and a small sample of references created for other sequencing projects, was created. to enable reference-based analysis methods. Sequences were collected from the following sources: published, finished genomes ([6, 7, 43, 101–104]; PacBio long read sequenced genomes from the National Culture Type Collecion NCTC 3000 project to sequence complete genomes from the NCTC; Oxford Nanopore long-read sequenced isolates from studies which are included in the amalgamated dataset [96, 105]; and long-read sequenced and manually quality-controlled isolates from the FDA-ARGOS database [106]. In total, 45 of these reference sequences were collected. In some cases, suitable publicly-available finished genomes could not be found for use as a reference sequence. In these cases, the short-read *de novo* assembly scaffolder MeDuSa [107] was used with many short-read *de novo* assemblies to produce a reference sequence.

2.1.3 The amalgamated global *Neisseria meningitidis* collection

After assembling all the datasets and removing contaminating isolates – methodology detailed below – the final dataset consisted of 15,332 isolates, collected between 1915 and 2020, from 70 different countries or territories. In Figure 2.3, the global distribution of these isolates is plotted country-by-country on a

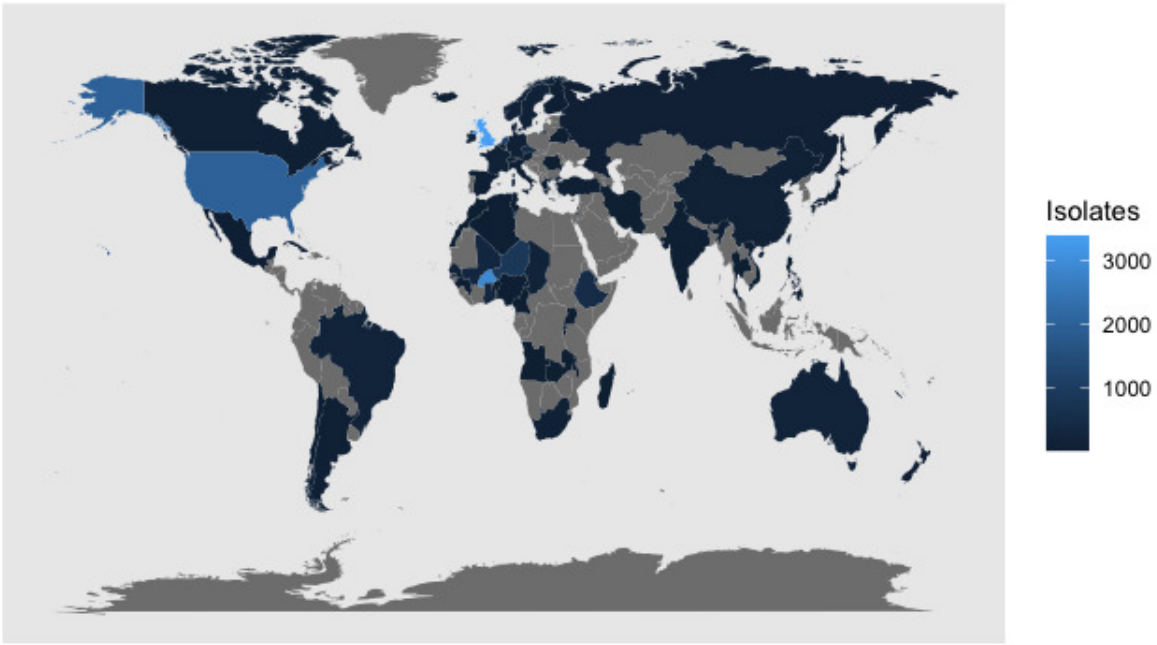


Figure 2.3: Map of the distribution of isolates which make up the final global *Neisseria meningitidis* collection

global map, which shows how the geographic sampling is uneven. Europe and North America are relatively well-covered, whereas Central and South America, and South East Asia are much less so. The global distribution of isolates also reveals some strong biases within our dataset – most countries have fewer than a thousand isolates, and indeed all but 5 have fewer than 500 isolates. However, three countries have more than 2000 isolates – the UK, USA, and Burkina Faso. Particularly in the UK and Burkina Faso, this represents a significantly greater sampling density than anywhere else in the world, and when interpreting the results of any further analyses, particularly how they pertain to geographical location, it will be important to consider whether the differential sampling density may be affecting the results. This is also true of the data we have from within the meningitis belt, where Burkina Faso has several times the number of samples compared to any other country in the region.

The number of samples in the collection per year is plotted in Figure 2.4, and from that figure we can see that the overwhelming majority of sequenced samples have been collected between 2010 and 2020. This is to be expected due to the widespread

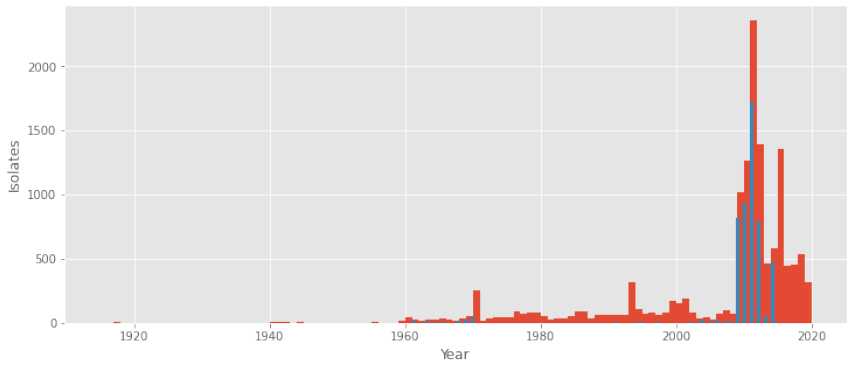


Figure 2.4: Bar chart of samples per year plotted for the entire global *Neisseria meningitidis* collection (red) and samples per year for isolates from the African meningitis belt (blue)

increase in the availability of whole-genome sequencing techniques allowing for many samples to be cultured, extracted, and sequenced routinely all around the world. Earlier samples, particularly those from the 20$^{\text{th}}$ century, would have had to survive storage in a freezer for decades in order to have been sequenced. As such, there are predictably many fewer samples from before the turn of the millennium. Despite the challenges in producing these samples, however, there is a consistent sampling all the way back to 1960, and sporadic sampling before that. As such, we can be relatively confident that any temporal inferences made from our analyses should be robust until at least 1960. Isolates sampled per year solely in the meningitis belt, unfortunately are only consistent back to 2004, before which there are sporadic samples back in time to 1961.

Finally, the percentage of isolates based on their carriage/disease status in every continent, and the global collection, is shown in Figure 2.5. Though it demonstrates that overall the collection is very well balanced between carriage and disease, there are significant differences between continents in terms of the carriage and disease state of the isolates sampled. Should there be evidence of significant differences in the population structures between continents, the differential sampling in terms of the disease state of sampled hosts could confound any potential interpretation of the results with regard to associations with carriage versus disease.

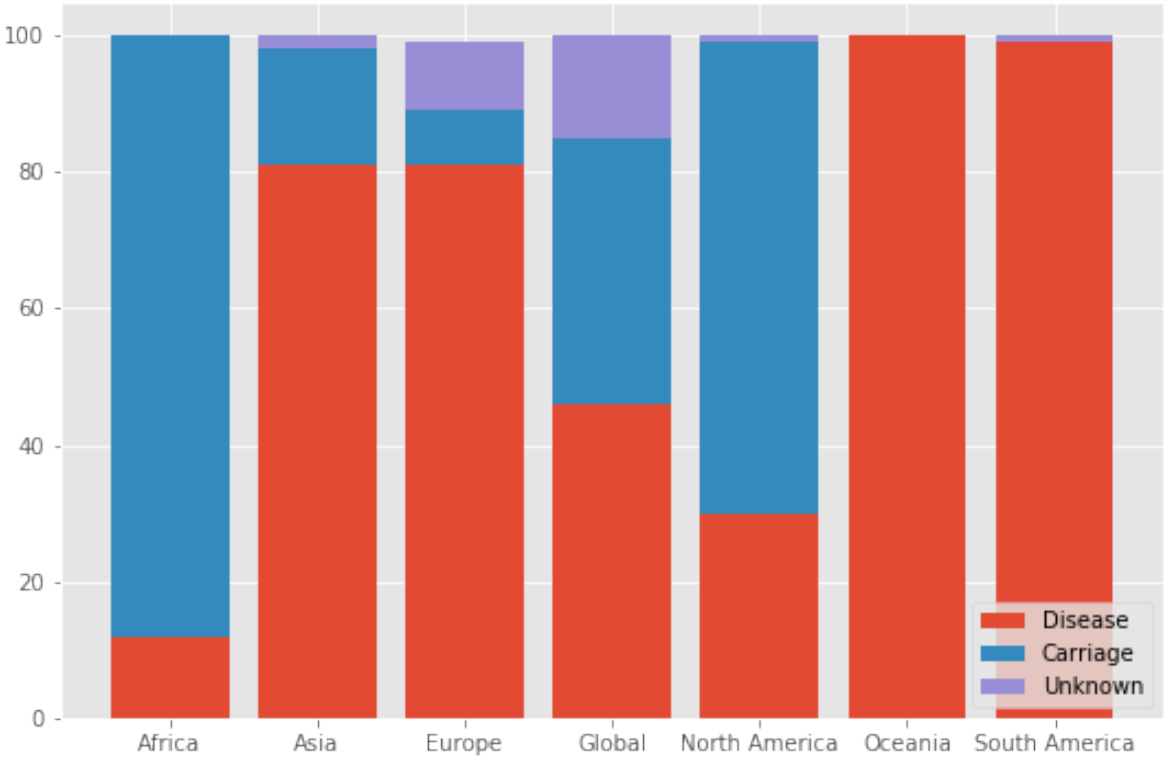


Figure 2.5: Stacked bat chart of the percentage of carriage/disease metadata available for isolates in the global *Neisseria meningitidis* collection, arranged by continent of origin

2.2 Bacterial Population Genomic Analysis Methods

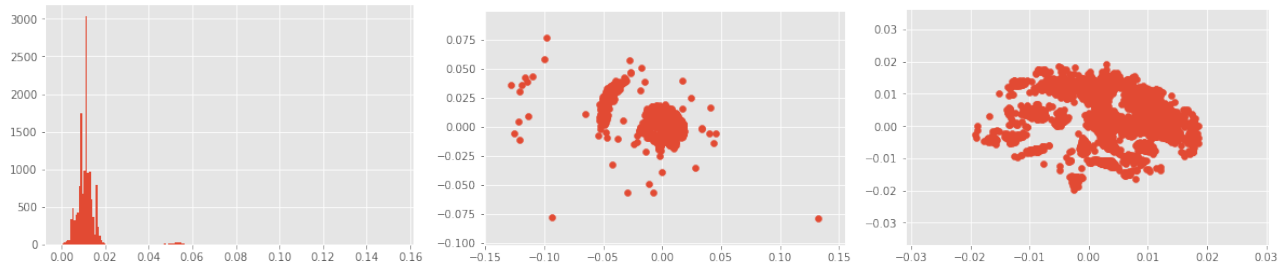
The process of analysing a dataset of over 15,000 whole-genome sequenced bacterial isolates presents challenges that require careful consideration. The central issue of ancestral descent and variation in gene content become substantially more challenging as the magnitude of the dataset increases, and even more so in species, such as those in the genus *Neisseria*, which are naturally competent and frequently recombine their genomes with exogenous DNA. In addition to issues caused by the biologically unique properties of bacteria, analysing a dataset consisting of over 15,000 isolates also inevitably runs into issues of errors introduced by the imperfect nature of short-read sequencing and the subsequent processing of short-read data. It is impossible at present, however, to avoid creation of these errors, so they must be taken into account when performing downstream analyses.

2.2.1 Basic genomic methods

2.2.1.1 Basic quality control, assembly, annotation, and *in silico* typing

All of the raw reads, whether sequenced at the WSI or downloaded from the ENA, were first run through a preliminary QC pipeline, to check that sequencing had been successful by making sure that FASTQ files contained sufficient numbers of reads with expected lengths and GC content. Isolates which passed this initial QC were then assembled using the prokaryotic genome assembly pipeline at the WSI, which primarily relies upon the short-read genome assembler Velvet [108]. The resulting genome assemblies were annotated using Prokka, the prokaryotic annotation pipeline [109].

Though sequence typing data exists for a plurality of isolates in the database, all isolates were also sequence typed *in silico* using the short-read sequence typing software SRST2 [110], the Neisseria pubMLST [59] sequence type database, and the raw reads for each isolate. Similarly, genogrouping, determining the presence of specific serogroups' capsular genes, was also



(a) Histogram of pairwise distances in the collection **(b)** MDS plot of pairwise distances after initial filtering **(c)** MDS plot of pairwise distances of isolates which passed QC

Figure 2.6: Plots of distribution of the entire collection's set of pairwise distances at various points in the QC process. (a) before any isolates were filtered, (b) after isolates were filtered based on discontinuity in the histogram of (a), and (c), the resulting dataset after filtering based on a threshold in MDS space, based on (b).

performed for each isolate *in silico* as lab-based serogroup data was available for some but not all isolates. The tool SeroBA [111] was used to perform this analysis, with a sequence database of published capsular reference sequences [17].

2.2.1.2 Further quality control

Though the sequencing pipelines perform some quality control for successful sequencing, with a dataset as large as this, some quality control for heavily contaminated sequencing runs needed to be performed as well. There are a variety of methods and no accepted standard for removing contaminated sequencing as it depends on the nature of the dataset and the aim of the research. For these studies, we have a substantial interest in understanding the extent of genetic diversity within *N. meningitidis*, and as such, we use a series of QC metrics (available in the panaroo-qc module [79]) which have no *a priori* assumptions regarding the expected genetic information sequenced, and as such avoid using any type of reference genomes.

We begin by using Mash [112] to calculate a pairwise distance matrix between all pairs of isolates in our collection, made feasible due to the fast MinHash algorithm used by Mash to estimate pairwise distances. From these pairwise distances, panaroo-qc uses the multidimensional scaling (MDS) function from the scikit-learn [113] python library to map these isolates into Cartesian 2D-space. Using these co-ordinates, we can then

calculate a norm for each isolate, representing its distance from the theoretical centre of the distribution, where the 'average' isolate in our collection lies. A histogram of these distances is shown in Figure 2.6 (a), and it is easily discernible that there is a clear break in the distribution around 0.04, where there is a distribution of isolates where the overwhelming majority of isolates are very close to the origin, with distances from the origin less than 0.1 in two small peaks, apart from a very small number of isolates which have distances close to 1. We can remove these isolates as they will represent completely contaminated isolates which are likely from a different genus. After removing these outlying sequences, we create an MDS plot to see what the distribution of isolates in terms of their pairwise distances looks like – shown here in Figure 2.6 (b). The plot looks fairly tight, but there are still a number of outlier isolates, and particularly a cluster on the upper left side of the origin. These likely represent isolates which are contaminated by the same, common, contaminant. After filtering with an even lower threshold to remove those isolates, we can see the centre of the distribution, and it is clear that there is some structure within the main oval, which should represent the population structure of *N. meningitidis*, in 2.6 (c). Finally, I checked the number of contigs and genes in the annotated assemblies to assess the effectiveness of the contamination quality control. In both cases, the MDS-based removal of outliers as a first step seems to have been a successful way of controlling the quality of the dataset, though one outlier remained in terms of contig number, which is also included among the seven outliers in terms of gene number. These were also removed from the dataset to produce the final dataset of *N. meningitidis* carried forward into further analyses.

2.2.2 Methods used in further analyses

The rest of this thesis relies upon a number of methods specifically for the analysis of bacterial whole-genome sequencing data, which may be unfamiliar to a wider biological audience. As such, the remainder of this chapter will present in-depth summaries of the rationale and theory behind the bacteria-specific analysis methods used in the research conducted in this thesis. Readers

who are familiar with some or all of these methods may wish to selectively read the remainder of this chapter. In general, however, many standard computational methods are used in the analysis of these data and the preparation of this thesis. Python [114] is used throughout in the processing of all kinds of data, including metadata, as well as for performing various mundane tasks relating to data analysis. Specifically, the IPython [115], NumPy [116], and Pandas [117] libraries were used for those purposes. Most figures have been made using matplotlib [118], though phylognies have been drawn using the ggtree [119] package, and maps in ggplot2 [120] and R [121]. Statistical tests were mostly performed using the SciPy package [122], and algorithmic techniques using the scikit-learn [113] package. Select figures were made using circos [123], cytoscape [124], and microreact [125] as identified in the text. Additional general methods for the analysis of biological data are occasionally used throughout the results chapters, and their use is detailed in the text.

2.2.2.1 Determining the population structure

Bacterial population genetics is very much complicated by the fact that the extent of homology between genomes can vary enormously within a species of interest. This introduces a number of practical complications, but it also introduces conceptual and practical difficulties in identifying the relatedness of any collection of bacteria. The conceptual difficulty tends to only arise in cases where the collection being analysed is particularly diverse, due to the fact that over short evolutionary timescales a subset of the genome, the 'core' genome, evolves in a sufficiently tree-like manner for standard phylogenetic methods to work. Using an aligned 'core' region of DNA to infer a phylogeny has been a standard methodology for inferring the relationships between isolates in smaller datasets for some time [126]. It leads to the practical difficulty, however, of how to choose the sequence data which is considered to be 'core', or consistently present and tree-like in its descent across the collection of interest from raw read data. Many methods exist for selecting this region as this has been an active area of research for at least a quarter of

a century. The oldest method for selecting a 'core' region still in use is the aforementioned Multi-locus Sequence Type (MLST) [58] method, developed at the end of the last century before the widespread availability of whole-genome sequencing, it selects some number of core genes, typically between five and seven, and uses the pattern of variation within these genes to identify the relatedness of some collection of bacteria. Though this technique has many practical benefits, the widespread availability of whole-genome sequencing data means that it is now possible to identify a much more expansive core region of the genome, allowing for much more accuracy and resolution in the identification of the relatedness of isolates within a population [78]. Two methods are generally accepted for identifying a maximal 'core' region, either mapping to a standard reference, or assembling raw reads into a draft genome assembly and using pan-genome inference methods to identify a core region of orthology, which will be discussed in detail in section 2.2.2.4.

In a dataset as large and diverse as the one considered in this thesis, however, it is clear that even if a 'core' region were to be accurately identified, standard phylogenetic methods would not be up to the task of accurately identifying the population structure, in addition to the obvious issues of runtime and interpretability that would arise. As such, research in recent decades has lead to the development of a variety of methods which allow a large collection of bacterial isolates to be partitioned into smaller groups [80, 127, 128], which could then in turn be subject to an in-depth phylogenetic analysis. These clustering methods have different underlying methodology for both determining the number of clusters present within the collection, and how to assign isolates to a specific cluster, but all of these methods rely upon being given some core variation data in order to partition the collection.

A recently developed whole-genome clustering method, PopPUNK [80], aims to resolve both of these key issues, as it partitions a population into clusters using both core and accessory genome content. To do this, it uses the MinHash similarity-estimation algorithm in order to estimate the pairwise core and accessory distances between every single pair of isolates in a

collection. This is done by generating k-mer sketches of every isolate in the collection, for variable k between an inferred minimum to avoid false positive matches, and a maximum set at 29bp, to allow for efficient computation. MinHash then allows for the rapid calculation of the Jacard distance, J , between all samples, pairwise. This distance, when not affected by mismatches caused by SNPs – practically, when k-mer length is as short as possible without causing false positive mismatches – is itself an effective measure of the pairwise accessory distance between two samples as the Jacard distance, J, between two samples $J(A, B)$ is defined as $\frac{|A \cap B|}{|A \cup B|}$, or in terms of k-mer sketches, the fraction of the number of k-mers present within both samples over the total number of k-mers present.

From this estimate of the pairwise accessory distance between samples at a variety of k-mer lengths, it is then possible to estimate the core distances between all samples, pairwise, as well. This is due to the fact that as k increases, the number of mismatched k-mers will increase with the length of the k-mers in the following relationship: $p_{corematch} = (1 - \pi)^k$, where k is the length of the k-mers and π is the SNP polymorphism between two species in the core. This is due to the fact that as the k-mer length increases, the likelihood of that k-mer not including a polymorphism is the inverse of the per-base pair rate of SNP polymorphism – the probability that any given position is not a polymorphism – multiplied by the number of base pairs, the length of the k-mer. This relationship, though complicated by needing to account for the probability of accessory genome mismatches with the above estimate of accessory mismatches, allows for the straightforward estimation of a core genome distance using linear regression in log space between the length of the k-mer and the overall probability of matching, the Jacard distance J.

The above method allows a rapid pairwise estimation of both core and accessory distances for each pair of isolates in the studied collection. Using these distances, it is then possible to determine a threshold in the two-dimensional space of core and accessory distances. This is done by using spatial clustering methods – either a Gaussian mixture model or the density-based

clustering method HDBSCAN – to find a cluster which extends to the origin. A linear threshold is then created in the core and accessory space to encompass the lowest upper bound of the spatial cluster which extends to the origin. This threshold is then taken to be the determinant of what is a lineage in PopPUNK's estimation, and what is not. A network of these lineage clusters is then created, where isolates whose pairwise core and accessory distances fall beneath the two dimensional threshold are joined into a network. The network properties of all clusters in this network are then used to refine the accessory-core cluster threshold within the neighbourhood of the initial boundary, and the optimal result is then output as the final clustering.

As a clustering method, PopPUNK has many advantages, both theoretical and practical. Theoretically, it resolves the key issue of bacterial clustering - how to take into account the variability in presence and absence of genetic material across a diverse collection, by considering both core and accessory distance in creating a discernment criteria for the clusters. Practically, the creation of a simple threshold as opposed to maximising some type of likelihood estimator means that the clustering is as extensible to new data like simpler methods such as MLST, without losing any information compared to other whole-genome methods. Finally, the use of the MinHash technique makes computation feasible on exceptionally large and diverse datasets, such as the one studied in this thesis.

2.2.2.2 Detecting recombination

After determining the population structure in species where that is complicated by non-tree-like patterns of descent, the next logical course of study is to attempt to determine where in the population, and in the genome, there are examples of alternative coalescent histories – in other words, recombination events. Detecting recombination events remains an open problem without a single methodological solution in all cases. The problem is complicated by the differing magnitude and genetic diversity of the various datasets which researchers would like to detect recombination events within, leading to different

methods being appropriate in different circumstances. This is primarily due to the nature of detecting exceptions to tree-like evolution necessarily requires determining what the 'correct' or tree-like pattern of descent is first. In this thesis, we use Gubbins [129], software which relies upon having a whole-genome multiple sequence pseudoalignment – created through mapping reads to a reference, and then instantiating variants and adding gaps for insertions – as it first generates a phylogeny for the sequences using RAxML [130], and looks for regions within the alignment which represent deviations from that phylogeny. This, and also the method for detecting recombinant regions explained below, means that Gubbins is only suitable for use on relatively non-diverse datasets. In our case, partitioning the dataset into lineages with PopPUNK provides sufficient genetic conservedness for Gubbins to run accurately.

To detect regions of the genome in a multiple sequence alignment that are potentially recombinant, Gubbins uses a sliding window statistic across SNPs in the genome to look for regions where there is a higher density of SNPs than would be expected by random chance. For each branch in the phylogeny generated by Gubbins, which is made up of the tip sequences supplied and inferred ancestral sequences, Gubbins slides a window of variable size (between 100bp and 10,000bp) along the genome and tests the density of single nucleotide variants in that window relative to a random distribution across the genome – modelled as a binomial distribution based on the size of the window and the average density of mutations in the entire sequence. Every set of contiguous windows where the p-value of a Bonferroni-corrected binomial test is greater that 0.05 is then considered a putative recombination. To identify accurate start and end points of a putative recombinant region, the distribution of SNPs in the putative recombination is also modelled as a binomial distribution, based on the length of the region and the overall density of SNPs within the region – strictly higher estimate of recombination than the binomial distribution which models SNPs across the entire genome. The boundaries of the putative recombination are then refined by reducing the boundaries of the putative recombination from the left to the next left out-

ermost SNP, then from the right to the next right outermost SNP iteratively, until further reduction does not increase the estimated likelihood of the recombination region under it's elevated binomial distribution of SNPs, compared with likelihood estimate under the binomial distribution derived from the whole-genome parameters. Finally, Gubbins tests the likelihood of the putative recombination with refined boundaries by comparing its Bonferroni-corrected p-value under the alternate 'recombination' binomial distribution of SNPs to the probability that a window of the refined size would contain as many SNPs as the putative recombination under the whole-genome distribution of SNPs, and rejecting the putative recombination unless the former probability is greater.

The initial recombination region is then masked in the focal branch and all its descendants, and the algorithm is repeated with the remaining SNP data in the focal branch until no windows significantly deviate from the density implied by the random whole-genome binomial distribution, or fewer than 3 recombinations remain unmasked in the branch. Finally, after this algorithm has run through every branch in the phylogeny, Gubbins masks all recombination regions which have passed the likelihood threshold, and reconstructs a phylogeny, free from the recombinant regions previously identified, and repeats the entire algorithm. This is repeated up to 5 times by default, or until the weighted Robinson-Foulds distance between the phylogenies generated from any two iterations is negligible.

Gubbins has a number of limitations, most of which arise from the fact that it relies upon a whole-genome pseudo-alignment [129], which can only be generated with the use of a reference genome. Many bacteria, including *N. meningitids*, have flexible genomes which accumulate diversity in the form of structural rearrangement and large insertions and deletions of genetic content, making reference-based approaches unreliable. As will be discussed in detail in Section 2.2.2.4, inferring the pan-genome of a collection of *de novo* assembled isolates can allow for a reference-free approach to the comparative study of homologous sequences. In this thesis, I use the method FastGEAR to detect evidence of recombination in the aligned gene sequences of a *N.*

meningitis pan-genome.

FastGEAR [131] takes individual gene alignments as input, and as a first step identifies lineages in the alignment using BAPS [127] and a hidden Markov model (HMM) to infer whether or not lineages are identical and should be collapsed. If the population structure is known from some other clustering – such as a PopPUNK run, this can be supplied, and this first step is skipped. Two steps of inference using hidden Markov models are then run to detect recombination events. In the first step, each cluster in the multiple sequence alignment is analysed separately using a hidden Markov model on each isolate in the lineage. These HMMs use the nucleotides at each position in the focal isolate as the observed states of the HMM, and aim to infer the hidden states, which are taken to be the origin of the nucleotide in question at each position – possible origins being its own lineage, another lineage present in the collection, and an unknown lineage not present in the collection. The HMM is iterated, updating its hypperparameters each time, until convergence. Positions whose hidden states are inferred to be from other lineages or unknown lineages at convergence are taken to be a recent recombination from that lineage. The second, simpler, HMM is then run on all the clusters, where the observed states are the nucleotide frequencies at each position within clusters, and the hidden states can take two values, identical and not identical. After the HMM is iterated to covergence, positions where hidden states were 'identical' are then taken to be ancestral recombinations between two clusters.

Finally, FastGEAR assess inferred recombinations for significance thorough calculating a bayes factor for each putative recombination. This is calculated using SNP densities, comparing the SNP density in recombinations compared to the SNP density in non-recombinant sequences at the same loci.

2.2.2.3 Detecting Selection

Detecting selection in bacteria cannot rely upon many of the sequence-statistics based methods which are used in eukaryotes, as they often rely upon assumptions – such as free recombination – which are not met in bacteria, particularly true for methods

which aim to detect selection at the level of the whole genome [9]. This has led to the development of methods for detecting selection specifically with bacteria in mind [132, 133], and to scan for evidence of historical selection in the whole genome of each cluster, I use one of these methods, spydrpick [134], on the whole-genome pseudo-alignments of each cluster.

To detect selection, spydrpick relies upon the genome-wide epistasis study, or GWES, approach. GWES scans the genome for highly epistatically linked single-nucleotide variants whose association is so significant that it cannot be explained by any other factor, such as linkage disequilibrium or genetic hitch-hiking, except for direct co-selection of a pair of variants. Spydrpick begins with an input whole-genome pseudo-alignment, or other sequence structures which contain information regarding the relative locations of variants from each other, and initially reduces it to a set of variant sites. It then calculates the primary statistic it uses to asses the relatedness between two variant sites in the population, mutual information. Mutual information, precisely formulated in the case of discrete distributions as $I(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{XY}(x,y) \log(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)})$, where x and y are some values from discrete random variables X and Y, p_{XY} is the joint probability mass function of X and Y, and p_X and p_Y are the marginal probability mass functions of X and Y. In effect, mutual information therefore measures the extent to which information regarding one of the values of (X, Y) is informative regarding the other. In a sense, it is a metric on the extent of probabilistic independence or dependence between two random variables. In the case of detecting selection in a genome, these variables are sequence positions, and mutual information measures the extent to which these bases in one variant position are informative as to the bases in another variant position.

Of course, linkage within a DNA molecule means that many variant positions will have relatively high mutual information, so spydrpick uses a technique called sequence reweighting to take this into account when estimating the joint probability in the calculation of each variant's mutual information. Sequence reweighting is a technique which uses the per-site pairwise single nucleotide distance between a collection of sequences to

assign weights to each sequence based on the reciprocal of the number other sequences in that collection fall beneath some threshold, which spydrpick sets at 0.1 [134]. For instance, if five sequences had a pairwise per-site nucleotide distance of less than 0.1 compared to the focal sequence, the focal sequence would be assigned a weight of 0.2. In this way, it accounts for combinations which are frequently or highly linked in the entire collection of sequences because they generally similar and have descended from a recent common ancestor.

Selection on one or more variants is also known to lead to a phenomenon referred to as 'genetic hitch-hiking' [135] where variation that is linked to a specific variant under selection also appears to be under selection due to the effects of selection acting on a linked variant. To account for these cases, spydrpick uses a network filter algorithm called ARCANE [136], which filters out edges in an interaction network, in this case made up of all variants with pairwise mutual information, $I \geq 0$, through a simple inequality. For each triplet of three linked nodes, x, y, x, in the interaction network, $\min(I(x, y), I(x, z), I(y, z))$ is deleted, and only the greatest two links, in terms of mutual information, are reported. When this is iteratively done throughout the network of interactions, only a linear group of the strongest of a convoluted network of links will be reported, both removing the weakly-linked variants and simplifying interpretation.

Finally, in order to test the significance of the discovered mutual information links, spydrpick performs a Tukey outlier test [137] on the distribution of mutual information values, and reports positive outliers and extreme outliers, highly linked variant pairs whose linkage can only be explained by co-selection.

2.2.2.4 Inferring the pan-genome

An alternative reference-free approach to the comparative study of substantial numbers of bacterial genomes is to infer a pan-genome[138] of the collection of interest. This approach typically reduces the collection of isolates to their coding regions (though it is also possible to extend this approach to intergenic regions [139]), which are then manageable in an all-against-all sequence similarity comparison, in order to find homologous genes, which

can then be further processed. This approach, first suggested early this millennium, has been the foundation of all pan-genome methods which have been developed since, despite nearly two decades of research [79]. Substantial refinements have been made in the detection and processing of paralogous genes [140], for instance, but these advances unfortunately did not account for the central issue which would arise as a result of increasingly large datasets: the errors introduced from incomplete assembly and computational gene annotation [79]. A sequence-similarity comparison of all genes in every sample, the starting point for modern pan-genome inference algorithms, requires first identifying the gene regions in each sample. Short read sequencing has vastly expanded the number of isolates that can be sequenced but has also generally lead to a decrease in the completeness of draft genome assemblies, due to technological limitations. Furthermore, the time and effort involved in annotating a genome manually has meant that for any sequencing study which aims to study more than a handful of isolates, computational genome annotation is required in order to identify the locations of genes. These two processes, particularly when combined, introduce errors at very low rate, normally no more than a few erroneous annotations per isolate. However, when combining thousands, or even tens of thousands of isolates to infer a pan-genome for those isolates, erroneous annotations quickly add up and severely affect the accuracy and interpretability of the results [141]. Despite relying on the same basic principles as most pan-genome pipelines, Panaroo accounts for these errors by building a pan-genome network based on the positions of genes in draft assemblies across the input collection. It then uses the context of the entire pan-genome network, therefore including information from other isolates, in order to correct various types of error introduced in assembly and annotation, as well as accurately splitting paralogous genes.

To do this, Panaroo [79] first relies on cd-hit to cluster the set of genes from all isolates into clusters [142]. These clusters are then joined to other clusters in a network, with genes as nodes and edges created between genes where there is sequence evidence that two genes were on the same draft

assembly contig in at least one sample. This forms the initial inferred pan-genome, which is then run through a series of error-correction and further processing algorithms in order to produce an accurate polished pan-genome.

The first stage of post-processing is to handle paralogous genes – genes where two annotated genes from the same sample are present in the same cluster. They are initially split into separate clusters for each copy in the genome, and then merged based on their position in the pan-genome network. Nodes that either closest or have the most similar context in the graph are iteratively merged until the final number of merged clusters is equal to the greatest number of times the paralogous gene family appears in the same genome. With paralogous gene families grouped into their differing positions in the genome, Panaroo then runs an algorithm to correct for frame-shift insertion/deletion mutations, or annotation in the wrong frame by clustering all genes which are neighbours of genes which are connected to more than 2 other genes – sections of the pan-genome graph which are non-linear – at the DNA level. A similar algorithm is then used to reduce diverse gene families which occupy the same location across many genomes, but instead iteratively clusters the sequences in the neighbourhood of a node of high degree, down to 50% identity. After these corrections, the main error-correcting step of Panaroo is run, the removal of genes which are suspected to be annotation errors, or contamination. Depending on the user-specified stringency, Panaroo will iteratively delete genes whose degree in the graph is 1, and whose existence is therefore poorly supported in the graph, with the level of support adjustable to a user-specified stringency. In circular bacterial genomes, it is theoretically impossible to have genes, nodes in the graph, which are connected on a DNA molecule to less than 2 other genes. Though some regions of the genome and mobile elements are difficult to assemble, trailing 'ends' of the graph with low support, either from a single isolate or a very small fraction of isolates, almost definitely represent mis-annotations at the ends of draft assembly contigs, a known cause of error in computational annotation methods [143]. Finally, Panaroo goes on to its final error correction, finding genes which may have

been missed by computational annotation of genomes. This is done by extracting the assembled contig sequence for genes where isolate is present in a neighbouring gene, but not the focal gene. Regular expression pattern matching is then used to search the contig for the missing sequence. Found sequences above the percentage identity threshold are then added to the pan-genome network. A final round of collapsing gene families is then performed, and then the final graph is output.

Panaroo's ability to accurately infer the pan-genome allows it to be used on large datasets [79]. However, for datasets on the order of 10^4 isolates, the initial pan-genome graph can become so large, due to the number of annotation errors, that the algorithm becomes prohibitively slow to run. To account for these cases, Panaroo also includes the Panaroo-merge module, which allows the merging of the output of several Panaroo runs on partitions of a larger dataset, or different datasets of the same species. This is done by clustering a reference sequence from each gene in each pan-genome, and then merging the gene clusters from the various pangenome based on this clustering.

2.2.2.5 Bacterial genome-wide association studies

Genome-wide association studies (GWAS) are a common and well-established method for exploring the links between phenotypic data and their genetic determinants. Bacterial genome variation complicates the use of GWAS methodologies in much the same way as they complicate determination of population structure. Furthermore, their clonal population structure complicates the use of GWAS due to the significantly increased effect it has on background selection and genetic hitch-hiking. Various methods have been developed over years of research to account for these problems, and as with many issues caused by particular patterns of inheritance in bacteria, there is no one-size-fits-all solution – the best approach depends on the nature of the dataset and the aim of the GWAS.

Pyseer [144] is a bacterial GWAS pipeline which implements many different methodologies so that it may be used as a general toolkit for bacterial GWAS. In particular, it allows association studies to be performed as a variety of variant types including

SNPs, k-mers, unitigs, and presence/absence matrices, with a variety of methods – fixed effect generalised linear models and generalised linear mixed models – for performing the association while controlling for the population structure, and also allowing for additional covariates to be taken into account. This flexibility allows for easy computation of a variety of input data and association methodology on the same data, an important method of validating any potential significant associations.

CHAPTER 3

The Population Structure of *Neisseria meningitidis*

Historically, the effects of within-species population structure on our understanding of evolution has been substantially understudied. This is partially because of the dominance of diploid, sexually reproducing organisms in studies of evolution, but also largely a result of the presumed simplicity of population structure in clonally reproducing organisms. However, population structure is now understood to be an important evolutionary force in microorganisms, and hence a crucial part of understanding the epidemiology of bacterial disease. This has led to the widespread use of early genetic techniques (such as MLST) to categorise bacterial pathogens into 'strains' among researchers and clinicians. The early era of whole-genome sequencing led first to direct comparisons based on finished genomes [43, 101]. In recent years, as the capacity to create whole-genome bacterial sequence data has increased beyond the point where phylogenetic inference can meaningfully describe a population, there has been substantial interest in the development of methods to identity the population structure of large population datasets. An overview of the theory behind these methods can be found in the methods section of the previous chapter. In this chapter, I will apply those whole-genome clustering methods to identify the population structure of the global collection of *N. meningi-*

tidis. This is an essential first step in order to allow for further analysis of these data, but is also itself an interesting area of study. Priority questions include: What is the genetic population structure of *N. meningitidis*, compared to other species with similar life histories, and how is that related to their geographical distribution? How does the population structure change over time? How do the major lineages which make up this population differ in terms of their epidemiology?

3.1 Whole-genome clustering to determine major lineages

3.1.1 Validating the whole-genome clustering

To determine the population structure of the global collection, a species-level core/accessory distance threshold for lineage partitioning in *N. meningitidis* was determined by running PopPUNK's sketching and model fitting functions on 13,391 isolates in the collection without the 1941 CDC isolates from the USA, as they were latterly added to the collection. Those isolates were then assigned to the previously determined model, and from that, the entire collection of 15,332 isolates was grouped into 1262 clusters. Before discussing the clustering results in detail however, it is important to confirm that the clustering threshold determined by PopPUNK and then used to partition the collection into cluster lineages is accurate and biologically meaningful. There are two ways in which we can validate the species-level clustering of a dataset of this size, the first is to examine the additional outputs of PopPUNK, in particular checking structure of the network of isolates joined into clusters by their relative distance and the position of the clustering threshold in the two-dimensional space of core and accessory distance. The second way is to compare the whole-genome clustering of PopPUNK to other methods of categorising *N. meningitidis* into lineages such as MLST or core-genome clustering, though as discussed in sections 1.2 and 2.2.2.1, these methods use less information to draw their conclusions [78] and might therefore result in less accurate lineage determination, either by incorrectly grouping

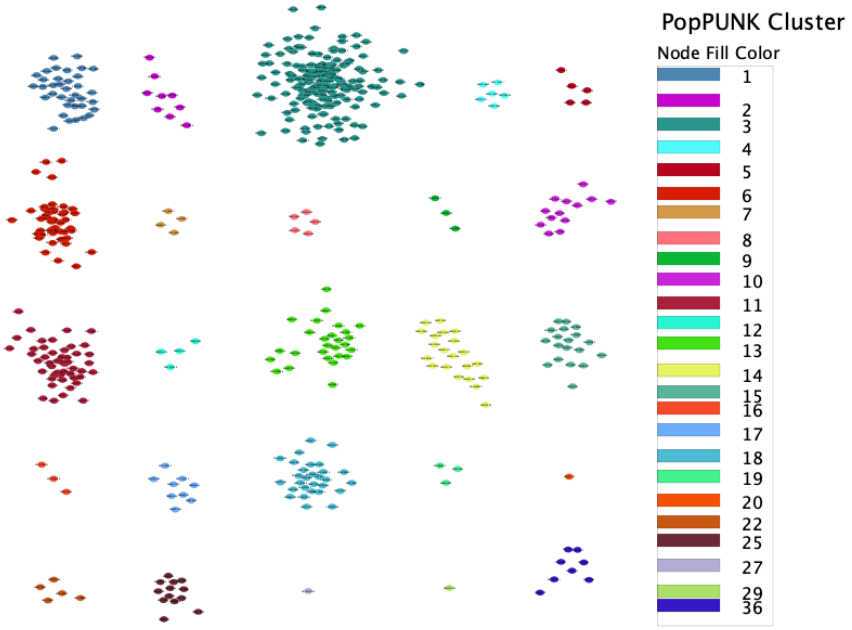


Figure 3.1: Reduced network diagram of PopPUNK clustering of the global *N. meningitidis* collection, consisting of 'reference' isolates selected by reducing the network to a single isolate per adjacency cliques in the full network. Clusters are arranged from left to right, then top to bottom.

isolates together due to a recombination event, or by splitting single lineages into multiple clusters.

Figure 3.1 shows the reduced network diagram of the 25 major lineages with over 100 isolates, where isolates within lineages are joined together to form a network where the magnitude of their pairwise core-accessory distance falls below the clustering threshold. The whole network was then drawn using a force-directed drawing algorithm, which lays out the nodes of the graph, in this case isolates, based on the distances between them, and due to the number of isolates, the network was reduced to a smaller representative set before plotting. Generally, it shows that PopPUNK has successfully found a clustering threshold for delineating lineages within *N. meningitidis* – most clusters are well formed and highly interconnected, some so much so that they can be represented by a single isolate. This means that the pairwise distances between most, and in some cases all of the isolates within a cluster are underneath the lineage delineation threshold, an important characteristic of clusters which are biologically accurate. There are some differences in

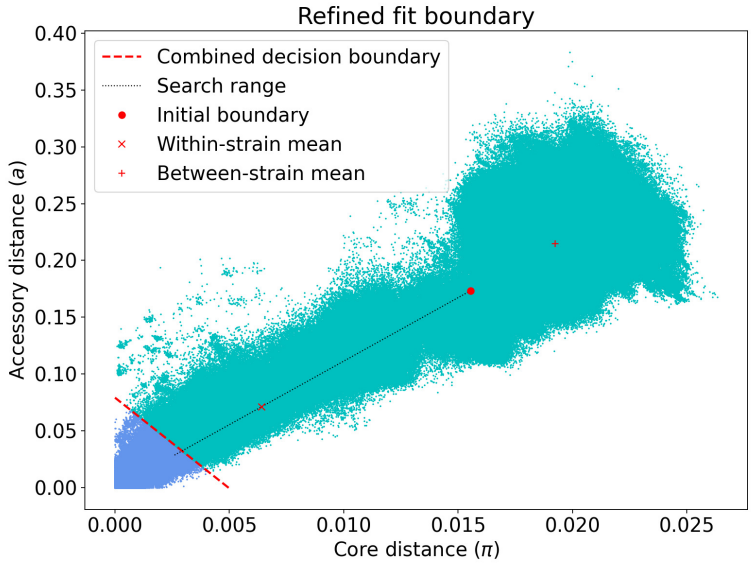


Figure 3.2: Plot of pairwise core and accessory distances for each pair of samples used in the initial model determination step of PopPUNK (13,391 isolates). Various regions of interest and the final clustering threshold are as indicated in the plot legend. The 'Combined decision boundary' refers to the final clustering threshold.

the shape of the cluster networks, however, with some clusters (3, 6, 11, 13) being more diffuse than others. This could be due to either the inherent differences in the diversity of clusters, or biases in how clusters were sampled, and it is impossible to determine which factor may be most important from this dataset alone. In either case, this is an expected outcome and does not reflect poorly on the accuracy of the clustering.

In Figure 3.2 we see the pairwise core and accessory distances for each pair of samples in our collection, plotted in the upper right quadrant of Cartesian space, where the pairwise core genome distance between samples is plotted on the x axis , and the accessory genome distance between samples is plotted along the y (see section 2.2.2.1 for further detail on how these distances are calculated). On this graph, the initial clustering boundary before network-property based refinement, the search range for network-property based refinement, and the final clustering threshold are also indicated as specified in the figure legend.

There are a few things which this figure tells us about both *N. meningitidis* and our efforts to discover its biological population structure. PopPUNK typically looks for a discrete cluster which

extends from the origin and is non-continuous with the remainder of pairwise core-accessory distances. Figure 3.2 clearly shows that in *N. meningitidis*, no such straightforward boundary exists – across the species, the pairwise difference between isolates are continuous from 0 to their maximum extent. This suggests that the population structure in *N. meningitidis* is not fixed and stringent, but rather exists on a fluid continuum. Given its natural competence and the knowledge that it is highly recombinant[55, 56], this is unsurprising, but Figure 3.2 provides direct evidence that the extent to which genetic information is shared between different isolates in *N. meningitidis* makes resolving the population structure with strict categorisation into lineages difficult to the extent that it should not be considered entirely possible. Instead, we might view multiple levels of classification as accurate, and ought to determine a population structure which is precise enough to capture relevant features of interest while remaining biologically accurate, and also allowing large-scale datasets to be reduced to lineages of sizes which are more susceptible to current methods of genomic data analysis.

The network-based model refinement step of PopPUNK does exactly this – by comparing the network properties of the resulting clusters at different thresholds, it is possible to assess how different clustering thresholds perform in terms of the various aforementioned factors. Figure 3.1 demonstrates that the final threshold succeeds in all of these areas – the networks are well-formed and highly interconnected, and, as will be discussed below, confirm the relatedness of common, well-studied lineages. A final feature of Figure 3.2 which suggests that the final clustering threshold, or 'Combined decision boundary' is accurate is its position away from the extreme minimum of the search range – generating increasingly smaller sub-clusters will always result in better network transitivity and connectedness, while significantly reducing how informative those clusters are in terms of the biological reality. The fact that the final threshold for splitting isolates into clusters, or not, is not at the origin suggests that it did not fall into this endless regression toward the origin.

Although the output of PopPUNK suggests that the cluster-

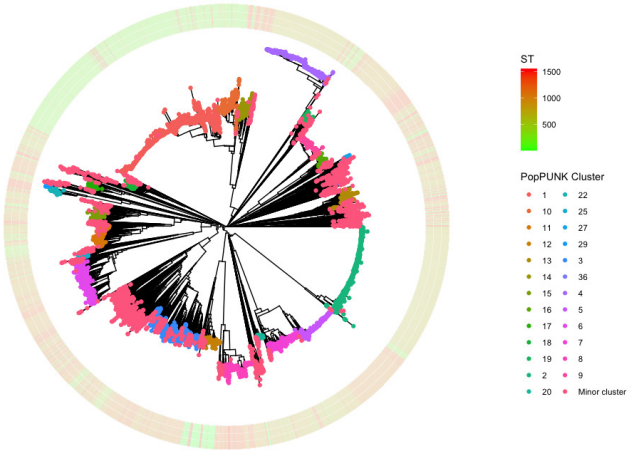


Figure 3.3: Core-genome distance (as calculated by PopPUNK, section 2.2.2.1) phylogeny inferred by Fasttree, annotated with PopPUNK clusters on the branch tips, and MLST Sequence Type on the outer ring.

ing threshold is accurate, and therefore the inferred population structure is biologically meaningful, in order to be confident that this is the case, the PopPUNK clustering can be compared to what is presently known about the population structure of *N. meningitidis*. As discussed in Chapter 1, the current knowledge of the global population structure of *N. meningitidis* is primarily based on the MLST method of identifying lineages based on allele patterns [10]. We expect whole-genome clustering to broadly recapitulate the population structure suggested by MLST, with a significant number of minor exceptions, due to recombination events which overwrite MLST genes and hence distort the lineages suggested by MLST. Figure 3.3 indicates how MLST compares to the PopPUNK clustering by mapping PopPUNK cluster and sequence type onto a pairwise core distance phylogeny – and it can be seen roughly that the expected pattern holds. The largest clusters on the phylogeny, particularly clusters 1-7, can be seen to predominantly correspond to a single ST. Most clusters are composed of a single dominant ST (Table 3.1, Figures 3.7-3.31) which in most cases represents more than 80% of the isolates within that PopPUNK cluster, though there are some notable exceptions where the dominant ST represents as little as 25% of the isolates in a cluster – in the 25 major clusters,

there are 4 such clusters: clusters 6, 11, 18, and 36. However, the pattern of a single dominant ST and other STs at very low frequency holds across the vast majority of the major clusters where $n > 100$. Among the four clusters with a lower frequency primary ST, clusters 6 and 11 are within the top 3 clusters when ordered by number of STs present within the clusters. This is expected if some clusters are more recombinant than others and therefore more rapidly switch STs. Variation in recombination rate, particularly between major lineages, is explored in chapter 4, but these initial results are enough to demonstrate that the PopPUNK clustering of this global collection is consistent with what is currently known about the population structure of *N. meningitidis* from MLST. Not only does this further confirm that the PopPUNK clustering threshold is indeed accurate and biologically meaningful, it also serves to demonstrate how whole-genome clustering offers enhanced precision and accuracy when compared to MLST methods, at the expense of a significant computational cost. 3 of the four clusters with a main ST which makes up less than 35% of the isolates are well-studied invasive lineages, which have been characterised primarily as clonal complexes. Categorisation with sequence types relies on similarities at specific allele sites, and relying on clonal complexes which groups together isolates despite mismatches in the sequence type can therefore miss isolates which are highly related overall but differ at specific sites, as well as erroneously linking isolates which are distantly related, but have the same alleles at 5 of the 7 MLST sites used to define clonal complexes. This is exemplified in Clusters 1 and 10, for example, would be grouped into the same clonal complex, despite falling above the same threshold of relatedness as any other two clusters, and both forming well-formed clusters in the distance network, Figure 3.1. Whole genome clustering, as we can see from Figure 3.3 broadly captures the same genuine population structure without these potential pitfalls, supporting claims which suggest it offers higher precision and accuracy.

The neighbour-joining core-genome phylogeny (Figure 3.3) inferred by using the core genome distances calculated by PopPUNK also provides good evidence that the PopPUNK cluster-

ing is accurate and biologically meaningful. Within the core genome, and on the long timescale considered in a species-wide phylogeny, the accumulation of single nucleotide mutation should, on average, accurately delineate the clonal pattern of evolution of *N. meningitidis*. As such, we expect the patterns of relatedness within the species, over long enough timescales, to be more or less accurately represented with a branching phylogeny, with long branches separating distinct clusters. This is precisely what is evident in Figure 3.3, where the main clusters map cleanly onto monophyletic clades of the phylogeny, and are separated by long branches which go deep into the phylogeny. Interestingly, there are a number of exceptions to this, where isolates within a monophyletic clade on the core-genome phylogeny – corresponding to a PopPUNK cluster – are clustered into a minor cluster, and not the cluster which generally corresponds to the monophyletic clade in question. These isolates, several of which are particularly evident in the clade of the phylogeny corresponding to PopPUNK Cluster 1, likely represent isolates with substantial accessory genome divergence, causing them to fall above the PopPUNK clustering threshold while still remaining well within the monophyletic clade corresponding to Cluster 1 in a phylogeny made with core-distances. This is most likely caused by substantial deletion or importation of novel gene content within these isolates' genomes, and possibly with missing intermediary samples which would have allowed them to fall under the clustering threshold. In any case, these isolates represent at most few interesting cases of significant gene transfer, while overall the pattern in the core genome phylogeny strongly supports the PoPUNK clustering as biologically accurate.

3.1.2 PopPUNK clustering results

As the PopPUNK clustering seems to accurately capture the population structure of our global collection of *N. meningitidis*, it is then possible to consider what the clustering actually reveals about the nature of that population structure. Of the 1262 clusters that PopPUNK partitioned this dataset into, 811 were singleton clusters composed of a single isolate, and only 142 clusters contained more than 5 isolates. 25 clusters contained more

than 100 isolates and the sizes of these clusters are shown in Figure 3.4, coloured by their continent of origin. This figure is one of several which suggests that *N. meningitidis* has a complex population structure at the global level, as many clusters are clearly primarily associated with a specific continent, despite sampling being relatively balanced between 3 different continents, as discussed in Chapter 2. This will be discussed in detail later on in this chapter, but it is worth noting that even this initial result confirms the existence of a complex global population structure with distinct patterns in different regions of the globe. Apart from their distribution, however, the whole-genome clustering also confirms that *N. meningitidis* has a complex global population structure, where many lineages within the global species population co-exist worldwide and are separated by substantial evolutionary divergence. Research using MLST data – particularly the fact that it is possible to reduce significant numbers of isolates and sequence types into relatively few clonal complexes – has also suggested that. Whole-genome data, however, provide some novel evidence for the longer-term trends governing the evolution of *N. meningitidis*. In particular, the number of singleton clusters, as well as the core distances phylogeny of all the isolates in the collection (Figure 3.3) demonstrates that there is an enormous well of deep, ancient diversity within the species, and that current clusters/clonal complexes of interest represent a substantial population increase in a small subset of the standing diversity present within the species. Though it is not yet possible to confirm what forces govern changes to the composition of the species population over the long-term, these whole-genome data demonstrate unequivocally that there remains significant ancient diversity within the species which has not been driven to extinction by the success of lineages which are currently of interest due to their effect on human health.

Despite the evidence of such diversity, Figure 3.3 also shows how our sample of the global population is dominated by 6 lineages which contain more than 500 isolates, and therefore make up approximately 44 % of the sample collection, shown coloured by continent of origin in Figure 3.4. As the collection is

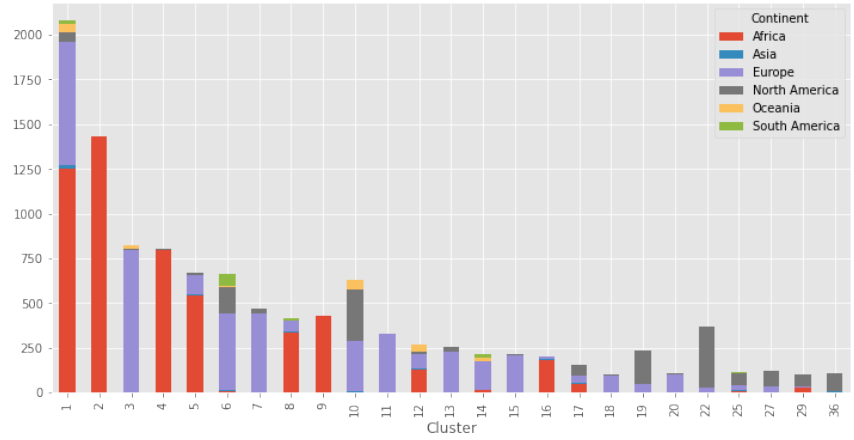


Figure 3.4: Bar chart of the sample sizes of the 25 major clusters with more than 100 isolates, coloured by continent of origin

not randomly sampled in cross-section, however, it is impossible to know how the relative sizes of these lineages, as observed in our sampling, correspond to their size in the current extant global population. Cluster 1, for instance, is the well-studied ST-11 lineage which has been responsible for many outbreaks of invasive disease during and immediately before the period from 2009 when most of the samples – 59.8%, Figure 2.4 – were collected. It has therefore been the subject of substantial interest, and as such has had the benefit of significant sample collection and sequencing efforts. Given that it was causing recorded outbreaks of disease, it is probable that this lineage also underwent an expansion in its population size during this period – however, it is, and will be, extremely difficult to differentiate these two effects. Similar circumstances apply for most of the major clusters – discussed throughout Section 3.2 – so in general, and unfortunately, it is not possible for this dataset to reliably inform us regarding the true frequencies of these various lineages within the global *N. meningitidis* population. Given the sharp decline in the sampled sizes of the minor clusters, however, we can say with relative confidence that the recent global population consists of these samples. Furthermore, while the biases in our sampling do complicate our ability to draw robust conclusions, PopPUNK clusters do tend to be primarily associated with a single continent, with the exception of clusters 1 and 10, which are likely well-sampled across multiple continents due to the intense public health interest in these lineages. The extremely

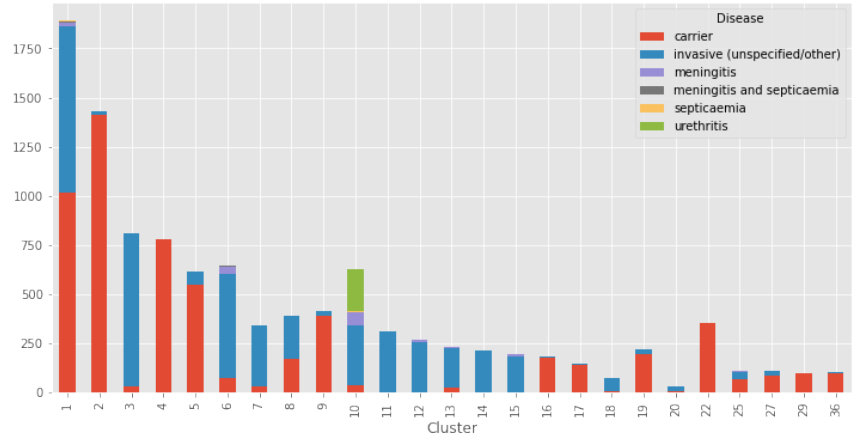


Figure 3.5: Bar chart of the sample sizes of the 25 major clusters with more than 100 isolates, coloured by the disease, or lack thereof, caused in the host at the time of isolation

strong association in the remaining lineages with one of the three main continents sampled in this dataset (Africa, Europe, and North America) seems to suggest that despite the highly non-random nature of the sampling, there still remain distinct population dynamics in different parts of world. That the continent with the most carriage isolates – Africa – has three large clusters which are located almost exclusively within it and barely detected in Europe or North America further supports the suggestion that geography is extremely important in determining the population structure and dynamics of *N. meningitidis*, as it is generally believed that most of the population does not cause disease. The accumulated carriage sampling here suggests that it may also evolve quickly and in ways which are not readily discernible from sampling disease-causing isolates.

The PopPUNK lineages in this collection do show a substantial tendency to be associated with either causing meningococcal disease, or being found in carriage (Figure 3.5). However, again due to the imbalanced sampling of the dataset, isolates collected in Europe, Asia, Oceania, and South America are predominantly disease isolates, and isolates collected in Africa and North America are predominantly carriage isolates (Figure 2.4). This makes it difficult to disentangle the effect of the geographical sampling with propensity for causing invasive disease. Despite this, apart from clusters 1 and 8, the remaining 25 major clusters are strongly associated with either carriage or invasive disease.

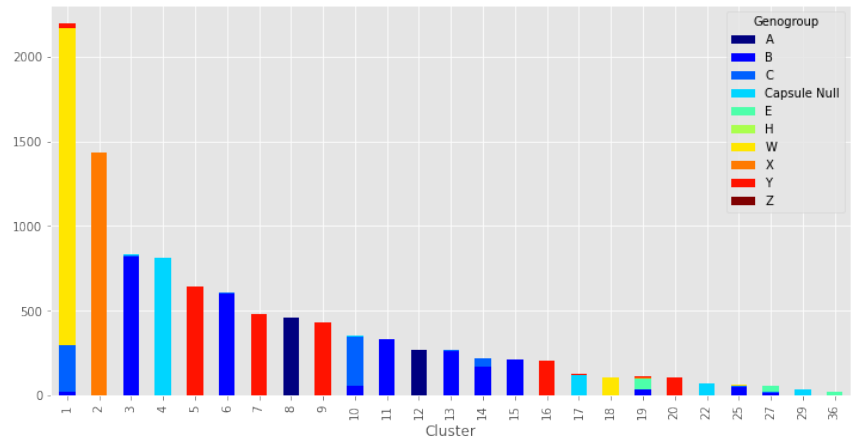


Figure 3.6: Bar chart of the sample sizes of the 25 major clusters with more than 100 isolates, coloured by their *in silico* genogroup

It is well known that certain lineages of *N. meningitidis* are associated with causing invasive disease, and the whole-genome data here similarly serve to illustrate this.

The samples in this collection date back to 1915, with 24 isolates having been collected before 1950. None of these isolates, however, are clustered among the 25 major clusters, and the earliest sampling from among the major clusters dates to 1959. Of the 748 isolates collected before 1975, fewer than half of the isolates (360) are clustered together with the major 25 lineages, and these 360 isolates are found in only 12 of the 25 major lineages – 1, 2, 5, 7, 8, 10, 12 ,14, 18, 27, 29, and 36. The earliest year of sample collection in the remaining major clusters ranges from 1975 to 2012. Though they are not as definitive as molecular evolution dating methods, phylogenetic estimates of the date of the most recent common ancestors of these lineages (Table 3.1) suggest that this is generally not unexpected as most lineages have recent common ancestors (discussed in depth in section 3.2). Regardless, the fact that most of the lineages which have been sampled and sequenced recently are not or barely represented in the collection prior to 1975, strongly suggest that changes to the population structure of *N. meningitidis* occur on a timescale which is observable within a human lifetime. This is a somewhat unexpected result, as MLST data for samples from that time period have suggested that many of the contemporary disease-causing lineages have existed for as long as we have had typeable isolates. In our collection, for instance, the ST-11, ST-

4, ST-32, ST-41/44, ST-1, and ST-4240/6688 clonal complexes are all represented by multiple isolates in the pre-1950 isolates. The whole genome data is very clear in this regard, however, and none of these isolates fall below the core-accessory threshold for clustering with any of the major clusters, and instead are grouped into 19 minor clusters, among which the most recent isolate was sampled in 1966. These facts, combined, do not point to an obvious causal origin for the shift in population structure and seeming replacement of lineages which has occurred over the time frame of our isolate collection, but it provides strong evidence that such a shift has indeed occurred.

Examining the population structure in the different regions provides further evidence that the populations are different in different parts of the globe. In the three main continents where we have sufficiently thorough sampling, Europe (n = 5862), Africa (n = 5476), and North America (n = 2115), we see populations which share some common lineages, but are generally quite distinct. 826 different clusters are found in Europe, whereas only 215 were found in North America, and 100 in Africa. Among these clusters, 696 – 84% of the total number found in Europe – were unique to Europe, 68 – 68% of the total number found in Africa – were unique to Africa, and 147 – 68% of the total found in North America – were unique to North America. Of the 25 major lineages, 23 were present in Europe, 20 in North America, and 18 in Africa. One major lineage, Cluster 11, was only found in Europe and another, Cluster 9, was only found in Africa. These results suggest two things: first, that the number of lineages detected in a given region is primarily driven by the breadth of sampling temporally. Europe, where the earliest sample in this collection dates to 1934, and North America, where the earliest sample in this collection dates to 1915, have a higher number of different lineages despite there similar or smaller sample size to Africa. Secondly, these results suggest that the current diversity is no more isolated or unique in any particular region. It has long been suspected that the population of *N. meningitidis* in the meningitis belt must have different evolutionary dynamics and consequently, population structure, to the population in

the rest of the world[10, 71] due to the substantially different epidemiology of meningococcal disease in the meningitis belt, but the population structure seen in this global collection does not support that view. Africa remains under-sampled, particularly due to the historical samples available in other regions which likely account for the higher numbers of unique lineages identified in Europe and North America, but the level of unique diversity in Africa remains comparable to elsewhere. Furthermore, each continent has a different set of major lineages, and in this regard the population sampled in Africa is similarly unexceptional.

One of the largest contiguously sampled datasets, in both space and time, within this global meta-collection is the newly sequenced collection of 2838 isolates from Burkina Faso, collected from 2009 to 2012 in three regions of the country. Section 2.1.1.3 fully describes the detail of how the isolates were collected, but briefly, it is a carriage collection of collected from healthy volunteers over a period of 4 years. We observe such a deeply divergent population structure over the whole collection but is this maintained within Burkina Faso as well? In short, yes. The collection of 2838 isolates is dominated by one large cluster making up 47.67% of isolates (1353), and generally is composed of 9 clusters which have more than 10 isolates, 8 of which are represented in the 25 major global lineages. In order of their size within the Burkina Faso dataset, these are the clusters 2, 5, 1, 9, 16, 8, 4, 29, and 34. Though no other constituent dataset is as large, regional subsamples echo the trend of having a subset of major global lineages in local populations, but at different frequencies from each other, and the frequencies within the global sample collection. Much like distribution of lineages across the different continents, this is further evidence to suggest that the complex population structure of *N. meningitidis* is extremely localised. Though different regions may contain isolates from some shared lineages of the species, their relative proportions and the presence of some unique lineages means that the interactions between lineages will not necessarily occur in the same way, leading to the emergence of differently structured regional populations.

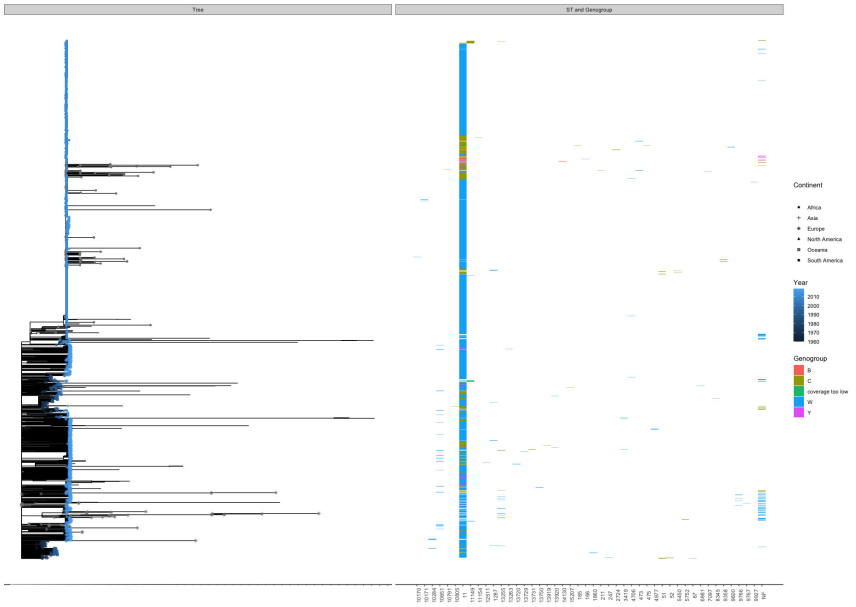


Figure 3.7: Whole-genome phylogeny of Cluster 1 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/cwHXXJCX9V8PvVKe9tq2sz`

3.2 Population structure within the major lineages of *Neisseria meningitidis*

Cluster 1, the largest cluster present in this collection, corresponds to the very well-studied serogroup W ST-11 clonal complex lineage which iss a major cause of invasive meningococcal disease around the world. It is possible that it is the largest or one of the largest lineages of *N. meningitidis* extant today, but its size without a doubt also reflects a bias in the samples collected and sequenced for study. It is impossible to know for certain the size of that effect, but we can speculate based on the structure of the tree (Figure 3.7) which suggests that approximately half of the 2161 isolates – 1267 – are part of a large, closely related outbreak as they are for the most part connected by very short branches. That said, however, their localisation to Europe and Africa means that these short branches may be accurately reflecting reflect a growth in their census population size after introduction into those two regions [65, 68].

Putting aside question of the relative size of this lineage within the greater *N. meningitidis* population, the tree also reflects the existence of more substantial evolutionary distances between clades within this lineage. The 'outbreak' clade of this phylogeny is restricted to Europe and Africa, and to sampling between the years 2000 and 2017. The lineage, however, contains isolates from 6 continents, and 48 different years between 1960 and 2019. The sampling is generally fairly well-distributed as well, with a significant peak around 2011 and 2012, where 660 and 328 isolates respectively were collected and therefore make up just under half of the cluster, but only two other years had more than 100 isolates collected (2017 and 2018), meaning the remaining 1000 isolates are split across 44 years in counts of less than 100 per year. Isolates collected in Africa also dominate the collection (1254) but a significant number also originate in Europe (690), with the remainder spread across North America, Oceania, Asia, and South America. The single largest country sampled within this lineage is the UK, but the countries of the meningitis belt are also generally well-sampled, particularly Burkina Faso, Niger, Mali, and Senegal. These 5 countries account for approximately 77% of all the isolates within this lineage, with the remaining isolates spread across 33 different countries. Interestingly, some isolates collected in the same country within 1-2 years from another are actually part of different and relatively distant clades on the phylogeny, suggesting the existence of population structure within this lineage. This is particularly evidence in recent (> 2015) European and particularly UK isolates, which can often be found contemporaneously in distant parts of the tree.

Unlike the distribution of isolates across different countries, the distribution of genogroups across the phylogeny follows the expected trend of relatively few switching events. Genogroup W is, as expected, the dominant genogroup in this lineage accounting for 84.87% of isolates, but there are significant numbers of genogroup C (12.91%) isolates as well, and notable numbers of genogroup Y and B isolates – 25 and 24 respectively, corresponding to 1.133% and .9968% of the isolates in the lineage. They are generally grouped into closely related lineages, however, in

line with the expectation that capsule switching is a relatively rare occurrence. The same can generally be said of switches in sequence type, though they do not always appear to be grouped as tightly as the switches in genogroup. There exist 46 different sequence types within this cluster, as well as 70 isolates with an allele pattern which did not match any of the currently defined sequence types.

Although Cluster 1 is generally associated with disease and is frequently referred to as a hyperinvasive lineage [10], it is actually primarily represented by carriage isolates in the amalgamated global collection, with 53.65% of isolates in this cluster have been collected from cases of healthy carriage. This likely reflects a sampling bias which saw the collection of many carriage samples particularly through the three large carriage collections in the meningitis belt when this lineage was present at a relatively high frequency in the local population of *N. meningitidis*. This does naturally lead to the question as to whether or not a lineages are inherently more likely to cause disease, as is often suggested [10], or if an observed bias in disease cases toward a certain lineage may also partially reflect the fluctuations in the population dynamics of *N. meningitidis*. Regardless, the shape of the whole-genome phylogeny clearly demonstrates that the isolates collected from Cluster 1 form what we might describe as an 'outbreak' cluster, with a population structure which generally is in line with what we would expect from a clonal organism. Older isolates tend to form out-groups of the phylogeny, and more inward isolates tend to be more recent isolates forming monophyletic clades in based on local geography. This is particularly true of the large monophyletic clade found only in the meningitis belt in the centre of this cluster, though it is true of many smaller clades in different regions as well. This is consistent with the most recent common ancestor of this cluster being dated to around 1934.47 (CI: 1930.6-1937.4). This is somewhat more recent than is expected based on epidemiological history and sequence type data, but those data may reflect sister lineages within this cluster which have now gone extinct, and the current extant diversity of Cluster 1 dates back to a single isolate sometime in the first half of the 20^{th} century.

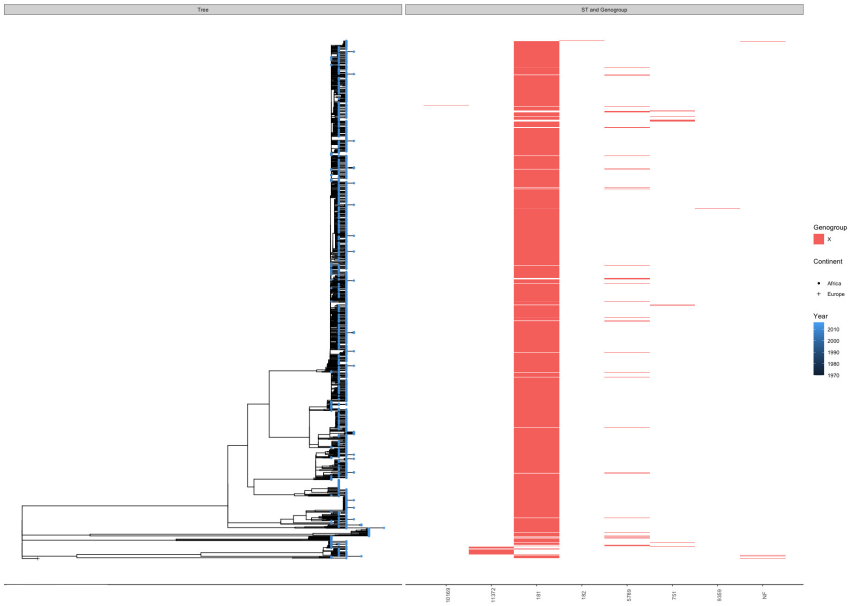


Figure 3.8: Dated whole-genome phylogeny of Cluster 2 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/4gd5k8ndjM5WRnhSwwaRFz`

Cluster 2 makes a sharp contrast from Cluster 1, as it is very much more localised in space, being found almost entirely within Africa, with only a single isolate being found elsewhere, in the Czech Republic. Even within Africa, the vast majority of isolates (1370, or 95.67%) were collected in Burkina Faso. The lack of diversity in genogroup is therefore unsuprising; every isolate was found to be genogroup X after *in silico* genogrouping. The pattern among STs is also low diversity; there are only 7 different STs found in the collection, along with 8 isolates with allele profiles not found in the current database of *N. meningitidis* sequence types. Interestingly, and in contrast to the pattern in Cluster 1, switches in ST are generally distributed across the phylogeny (Figure 3.8), particularly the switch between ST-181 and ST-5789, seen many times in Figure 3.8. Given the collections from which these isolates originate, the overwhelming majority of isolates within this lineage are carriage isolates, 98.53% of all isolates for which there is disease metadata. The disease isolates are from Chad, Togo, and the Czech Republic, and some from Burkina Faso contemporaneous with the carriage

collection.

The fact that this lineage is so dominated by Burkina Faso makes interpretation much more straightforward, as there are relatively few isolates which are not part of that collection. The next most frequent country of origin in this dataset is Ethiopia, with 59 isolates (21 of which were high-quality enough to be in the whole-genome phylogeny) then Chad, Togo, and the Czech Republic with a single isolate each. Most of the isolates in Cluster 2 were sampled between 2009 and 2012, as expected given the dominance of the Burkina Faso dataset within this lineage. The remaining isolates are sampled in 2014 (the Ethiopian isolates), 2016 (the isolate from Togo) 2013, (the isolate from Chad), and 1970 (the isolate from the Czech Republic). Reassuringly, the isolate from the Czech Republic in 1970 is an outgroup to the rest of the lineage, and close to the root of the phylogeny. The phylogeny then has a major split into the large clade which contains most of the Burkina Faso isolates and the isolate from Togo, and a number of smaller clades. One of the smaller clades consists entirely of the isolates from Ethiopia, while another is a mixture of 46 isolates from Burkina Faso and the isolate from Chad. Based on the distribution of isolates across their countries of origin in this dataset, it seems very likely that Cluster 2 is relatively rare globally, and highly over-represented in our dataset due to the scale of the Burkina Faso collection, which was sampled at a very specific point in time when there was a significant outbreak of this lineage in the local population. Still, the existence of long branches between different clades in this lineage, one of which is present in only Burkina Faso, another in both Burkina Faso and Chad, and another only in Ethiopia, combined with the branch near the root dating to 1970 and found in the Czech Republic, lends enormous support to the population structure suggested by the overall clustering of the entire global collection, that many minor lineages persist at low frequency in different populations, and occasionally expand significantly. The lack of any samples in this lineage from beyond 2016 is surprising given its size in the collection, and suggests that it has once again declined in relative frequency within the population after

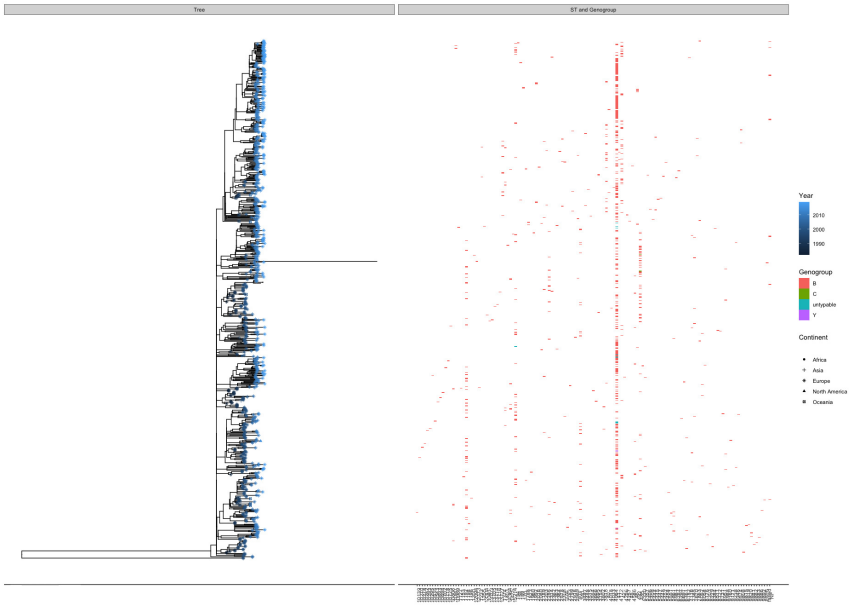


Figure 3.9: Whole-genome phylogeny of Cluster 3 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/7bBGQ44sqvUVCpxsFyNjwC`

the outbreak, in keeping with the current thinking regarding this lineage of *N. meningitidis*. Within the cluster itself, the structure of the phylogeny again supports the suggestion that this cluster has undergone a large clonal population expansion, with the outgroup of the entire cluster being geographically and temporally distant, and recent clades within the phylogeny showing geographical structure. The estimated date of the most recent common ancestor supports all of the above as well, with the time estimated around 1967.91 (CI: 1966.85-1968.81).

In terms of sequence types, Cluster 3 is enormously different from Cluster 2, having the most sequence types present of any cluster, 138 different STs. The primary ST, ST-41 makes up 38.48% of isolates, and the next most prevalent ST, ST-485, makes up only 6.129% of isolates. There is frequent switching between STs, and it is generally not localised to monophyletic clades, instead ST switches are frequently spread out over the entire phylogeny. This echoes what is known about the ST-41 clonal complex, which is known to be especially diverse in terms of ST, but seeing their distribution on a phylogeny underscores the extent to which that diversity is driven by consistent

switching between alleles at MLST sites. In terms of serogroup diversity, Cluster 3 is more normal, with one genogroup-level switch to genogroup C in a UK-sampled lineage from 2018 and 2019, and two separate switches to genogroup Y in distantly related UK isolates sampled from 2010 and 2012. There are also two clades which were non-groupable *in silico*, having apparently lost the genes required to produce capsule. Cluster 3 is so dominated by isolates sampled from cases of invasive disease it is difficult to draw any conclusions from their distribution across the phylogeny, with only 3.941% of isolates having been collected from healthy carriage infections, primarily from the UK but also including 1 of the 2 US isolates. This likely primarily reflects a lack of concurrent carriage sampling in these regions as opposed to any biological phenomena.

From the perspective of the geographical origin of the isolates in Cluster 3, it is perhaps unsurprising that so much diversity is observed within the UK. The vast majority of the isolates in this lineage were collected in Europe (793 isolates making 96.35% of the lineage), within which the UK was the single largest source of isolates (479) followed by the Netherlands (261) and Ireland (41). In addition to the UK isolates, 23 isolates were collected in Oceania (22 in New Zealand, 1 in Australia) 2 in North America (USA), 2 in Asia (Israel), and 1 in Africa (Réunion). There is some temporal diversity in the non-European isolates, they range from 1991-2016, but that is dwarfed by the temporal diversity found the European isolates present in the collection, which were sampled in 36 years between 1982 and 2019, with the only year in which no samples from this lineage were collected being 2003. Most of the samples were collected between 2010 and 2012, with 68, 114, and 88 samples having been collected in those years, respectively. That said, the spread of sampling is generally very good, with substantial numbers being collected between 1998 and 2001, as well as 2017 and 2019. The imbalanced sampling between different regions does reveal an interesting pattern, however, where contemporaneously sampled isolates from the same country are quite separate in the phylogeny, and nested within clades separated by long branches, particularly evident in the isolates from New Zealand, at both

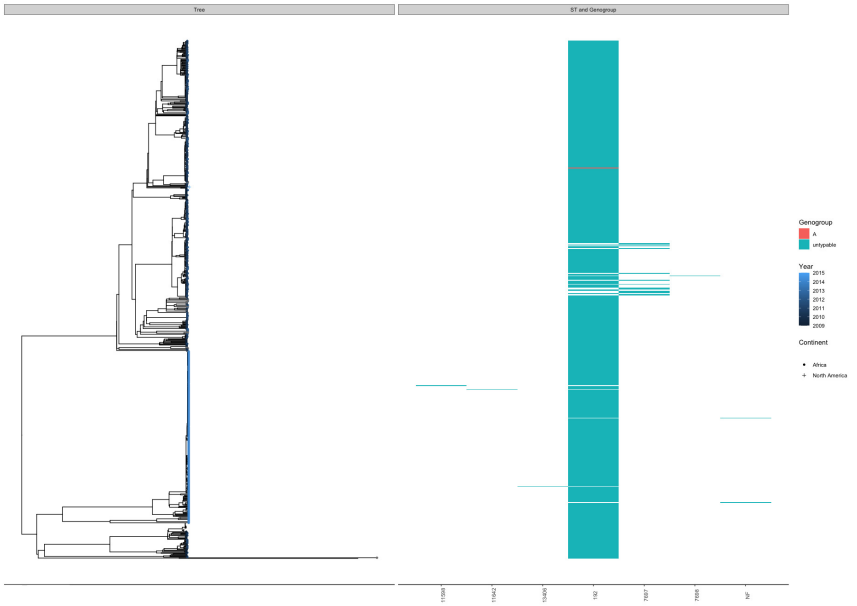


Figure 3.10: Whole-genome phylogeny of Cluster 4 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/4xEvHVpyjYMp5T4gmt6tB8`

points of sampling, as well as the couple of isolates from the Czech Republic and the United States. This is a further illustration of the existence of population structure within the lineages of an already deeply structured population at the species level. Perhaps the most interesting aspect of the Cluster 3 phylogeny, however, is its outgroup, two isolates from the UK, sampled in 1998, which exist isolated on a long branch away from the rest of the phylogeny. Unlike the distant outgroup in cluster 2, the outgroup isolates here are contemporaneous and co-localised with a great deal of other isolates, and were sampled at the start of a period of relatively intense sampling (1998-2001). Despite this, the monophyletic lineage containing these two isolates has apparently gone extinct. The most recent common ancestor of all the isolates is unreasonably distant in the past, with the estimate dating to 1783.1, with a confidence interval of over a hundred years between 1702.93 and 1826.79, additional evidence that within Cluster 3 there are a number of deeply divergent lineages, and a complex population structure.

Unlike Cluster 3, Cluster 4 is composed almost entirely of carriage isolates, at around 99.87% carriage, with only a

single disease isolate. This primarily reflects the geographical origin of the isolates, which almost entirely come from the meningitis belt, apart from two exceptions, one isolate collected in the United States, and one isolate collected in Malawi. Niger and Ethiopia are the origin of most of the isolates, 375 and 269, respectively, with smaller numbers from other meningitis belt countries: 90 from Burkina Faso, 41 from Mali, 15 from Ghana, 7 from Chad, and 1 from Senegal. As these isolates were all primarily collected in large-scale carriage surveys of the meningitis belt, their temporal spread is also relatively minimal, ranging from 2009-2015, with samples collected in every year apart from 2013.

Possibly in part due to the relatively narrow sampling window compared to Clusters 1-3, Cluster 4 is much less diverse in terms of sequence types and genogroups. The majority of isolates (97.54%) in Cluster 4 have sequence type ST-192, with a few isolates with different ST, and 2 isolates with allele profiles not found in the pubMLST database. All of the Cluster 4 isolates apart from one were non-serogroupable with SeroBA, apart from a single serogroup A isolate. The distribution of STs on the phylogeny is largely as expected, with switches to different STs from ST-192 being largely restricted to specific areas in the phylogeny. The shape of the phylogeny in general, however, is somewhat unexpected given the phylogenies examined so far. Two large monophyletic clades make up the phylogeny, with one clade containing the all the isolates from Ethiopia, 40 isolates from Burkina Faso, an isolate from Niger, and the isolate from Malawi, whereas the other isolate consists of the remainder of the isolates from the meningitis belt, and the single disease isolate from the United States. In the other phylogenies examined thus far, the phylogeny has not been so clearly divided into two monophyletic lineages, and despite the evidence of population strucutre in Cluster 3, Cluster 4 seems to be an even more profound illustration of the extent to which a single cluster can have internal population structure. The date of the most recent common ancestor in Cluster 4 is similarly unresolvable, with an estimate in the 17^{th} century and again a confidence interval of over a hundred years. Our understanding of the population

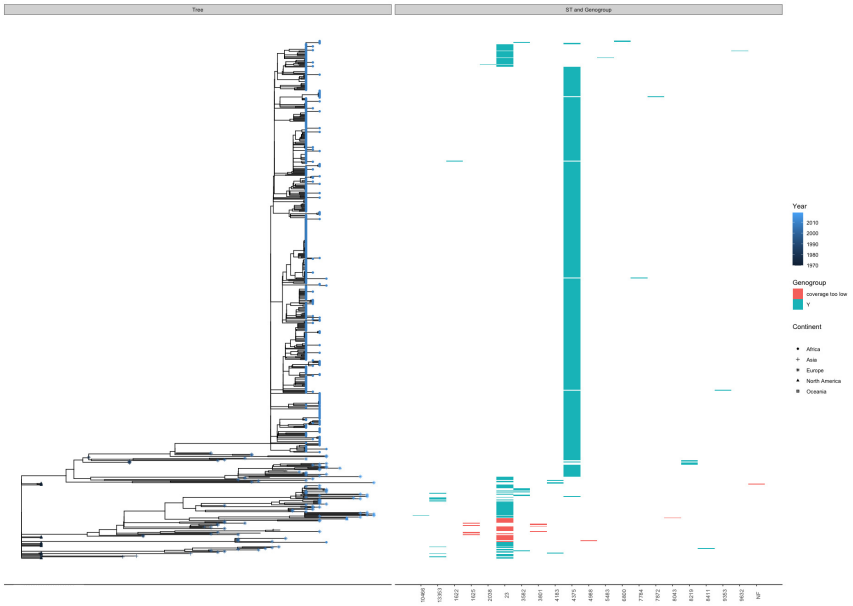


Figure 3.11: Whole-genome phylogeny of Cluster 5 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/fJMqp8Jxwawhb2NH5iQMBd`

structure in Cluster 4 might also be confounded the narrower sampling window, though apart from a single isolate collected in 1970, Cluster 2 has a very similar temporal spread, suggesting that this cluster might simply consist of two primary lineages. Finally, a minor aside, the existence of two geographically distant isolates on long, single branches of phylogeny as outgroups of an internal monophyletic clade – in this case isolates located in Malawi and the US – underscores the importance of sampling as much as possible from different geographical regions. Though this cluster seems to be primarily endemic to only the meningitis belt based on our collection, it is possible that this is driven entirely by insufficient sampling in other regions.

Compared with Cluster 4, Cluster 5 seems much more globally distributed, though in this collection the majority of its isolates were still collected in Africa (540). Nonetheless, 102 isolates were collected in Europe, 17 in North America, 11 in Asia, and 1 in Australia. Most African isolates are from Burkina Faso (538, two from Ghana), whereas the European isolates are mostly split between the UK (57) and Sweden (34). The isolates in this cluster have been sampled from 31 different years

between 1970 and 2019, though most of the sampling is from 2009 and 2011. The spread of geographical origin and year on the phylogeny serves to underscore its most notable feature – in contrast to Cluster 4, the phylogeny of Cluster 5 is made up of one large monophyletic clade, and several smaller outgroup clades separate from the monophyletic clade which includes most isolates (around 85%), and more similar to the shape of the phylogenies of Clusters 1, 2, and 3. The large monophyletic clade and its immediate sister clade contain isolates from most of the sampling locations, apart from five isolates sampled from the United States in 1970, an isolate collected in 1970 in Canada, two Japanese isolates from 1982 and 1984, an isolate from Israel collected in 1992, and 8 isolates collected between 1995 and 2005 in Sweden. These disparate isolates do not form a single monophyletic cluster, instead they form several monophyletic clusters generally linked by geography – the Swedish isolates, for instance, form a single monophyletic cluster – but they, like minor lineages in the other clusters, point toward a feature not generally noticeable when observing the entirely global collection as a whole, that there appears to be monophyletic lineages within the global clusters which have gone extinct. Much like in Cluster 3, where one of the apparently extinct outgroup lineages overlapped in space and time with the main lineage which continued to be sampled, in this cluster the sampling in Sweden continues to 2010, isolates from both the main clade and the outgroup clade are sampled between 1995 and 2005, after which no more outgroup clade isolates are detected, in any part of the global collection. Despite this apparent complex population structure within the lineage, Cluster 5 has an estimaed most recent common ancestor in 1967, precisely 1967.12 (CI: 1966.75-1967.41), supporting the suggestion that the population structure observed is likely relatively recently arisen and does not reflect a divergent and complex population structure within the lineage.

The ST diversity in Cluster 5 consists of a number of switches of ST generally in line with expectations. The main lineage is mostly made up of ST-4375, which makes up 78.43% of the Cluster as a whole, while its sister lineage and the outgroup

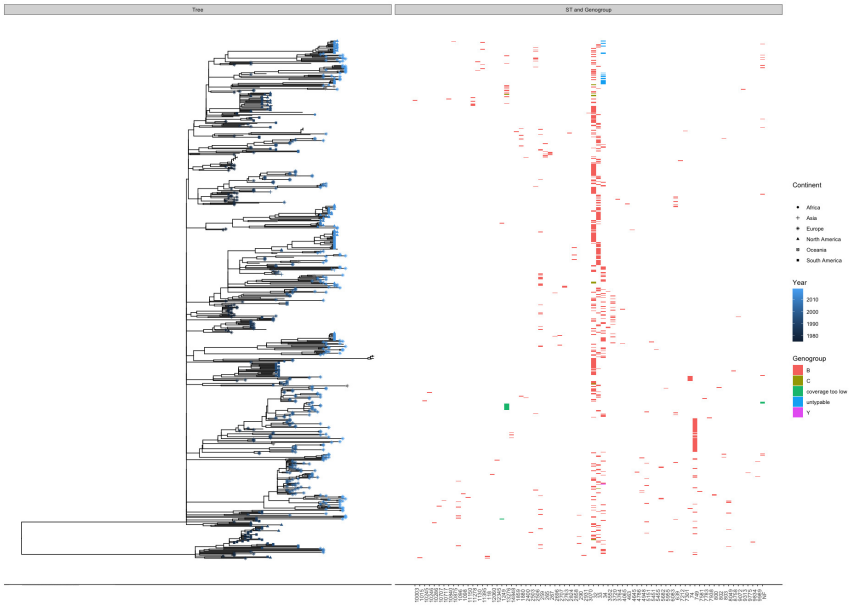


Figure 3.12: Whole-genome phylogeny of Cluster 6 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/qArNBsKamTLbYiagt2fSZL`

lineages are primarily ST-23. There is absolutely no serogroup diversity with this cluster, with all isolates being identified as genogroup Y or not possessing sufficient read depth to be genogrouped *in silico*. Finally, Cluster 5 is primarily made up of isolates sampled from healthy carriage, which make up 88.47% of sampled isolates. There are 71 isolates sampled from disease, fairly well spread across the phylogeny apart from the portion of the main clade which is made up of carriage isolates from Burkina Faso.

Cluster 6 is another extremely diverse cluster in terms of the sequence types present within, with 76 different sequence types present and 17 isolates with allele profiles not found in the database. The most common sequence type is ST-32, which includes only 31.70% of isolates, and the next two most common STs, ST-33 and ST-34 make up 17.6% and 11.45% of the cluster's isolates, respectively. Their distribution across the phylogeny does show distinct phylogenetic grouping, that is to say that there are multiple monophyletic clades of several isolates of a given sequence type, their distribution is highly non-random across the tips of the phylogeny. The same is true of the

serogroup diversity, though there is generally much less diversity in the *in silico* genogroups than in sequence types. 95.83% of isolates are genogroup B, with only 17 being non-groupable, 12 genogroup C, and 1 genogroup Y. The genogroups are generally distributed in monophyletic clades, with the non-groupables in particular forming a single clade of US isolates. Serogroup B isolates of the major sequence types identified within Cluster 6 have previously been identified as sequence types likely to cause disease, and perhaps as a result of both the propensity to cause disease and the increased interest in studying disease cases, a substantial majority of the isolates – 88.70% – in Cluster 6 have been collected from cases of invasive disease.

Perhaps related to their serogroup diversity, Cluster 6 isolates also span six continents, though the isolates were predominantly collected from Europe – 137 from the Netherlands, 127 from the UK – and North America, with 113 isolates having been collected in the United States, four from Canada, and 31 from Cuba. Notably, Cluster 6 has the largest number of South American isolates of the 25 main clusters, with 66 isolates from the continent, 35 from Brazil, 16 from Chile and 15 from Argentina. Smaller numbers of isolates in this cluster were collected from Africa, Asia, and Oceania. None of the four Africa isolates are from the meningitis belt, however, they were collected in South Africa and Réunion. With so many isolates being from the Global Archival collection, Cluster 6 also has a substantial sampling window, with isolates having been collected every year except 2013 from 1975 to 2019, inclusive. Regardless of the number of samples obtained from a country or their temporal spread, isolates from almost every country are well distributed across the phylogeny, apart from the Maltese isolates, which form a single monophyletic clade. Again, a striking feature of the Cluster 6 phylogeny similar to Cluster 4 is a deep split between two different two sister clades which make up the entirety of the phylogeny. The two clades in Cluster 6 are not of roughly equal size, as they are in Cluster 4, but both are spread over several continents, and both contain contain sequence type and serogroup diversity. Unlike Cluster 4, however, the date of the most recent common ancestor is not inestimable from

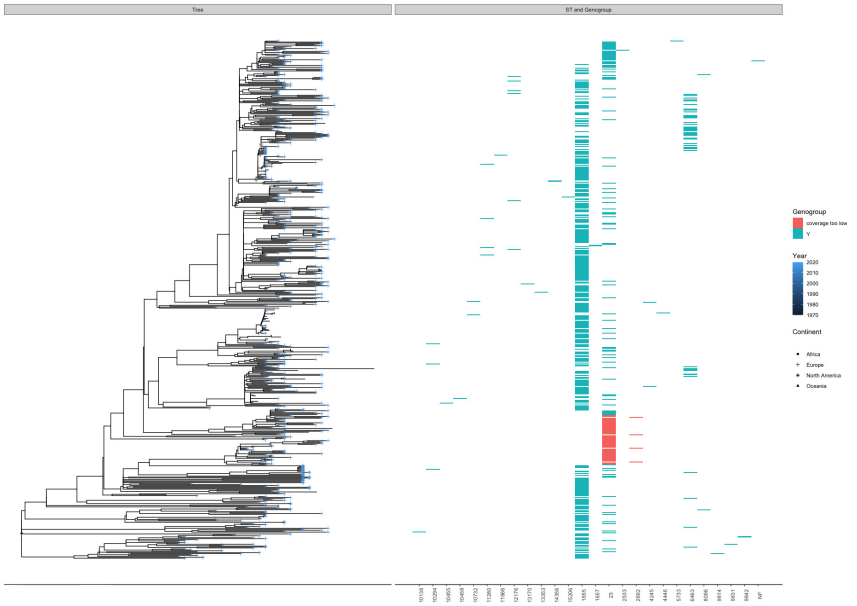


Figure 3.13: Whole-genome phylogeny of Cluster 7 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/dZYaayKgh21RzPeALYE1Y7`

the phylogeny, though it is on the borderline of what can be predicted accurately, with a confidence interval of around 45 years, the tMRCA of Cluster 6 is expected to be around 1902.73 (1880.65-1915.71). Though it is difficult to be confident of that date itself, It does provide good evidence that the population structure observed with the two sister lineages which make up this cluster is likely to have arisen in the past century, as opposed to the more deeply divergent population structure found in Clusters 3 and 4.

Cluster 7 is much less diverse in terms of sequence types and genogroups, with only 26 different STs present, and no genogroup diversity, with 100% of isolates which had enough sequencing coverage to genogroup being identified as genogroup Y. 60.11% isolates in Cluster 7 are identified as the sequence type ST- 1655, with the next most prevalent sequence type being ST-23, accounting for around 25.81% of isolates in the cluster. Similar to Cluster 6, Cluster 7 is primarily made up of disease isolates, with only 9.3% of isolates having been collected from carriage infections. The geographical and temporal distributions of the origins of the isolates in Cluster 7 are also similar, with

isolates in Cluster 7 having been collected primarily in Europe (91.15%), with 359 of those isolates originating in the UK, 58 in Sweden, and 13 in Ireland. A smaller but still notable number of isolates, 22, were collected in the United States, two in Australia, and one in South Africa. The temporal overlap is similar as well, with isolates ranging from 1970-2020. This similarity is only superficial, however, as unlike Cluster 6, there is a substantial gap in the sampling between 1971 and 1992, which was not present in Cluster 6. Most of the sampling is again between 2009 and 2012, and 2017 and 2019.

The distribution of isolates from different times and locations across the phylogeny recapitulates many of the patterns seen in Clusters 1-6. Isolates from the same geographical region are often dispersed across the tips of the phylogeny, suggesting the existence of either a complex population structure, or rapid migration similar to Cluster 1. There is also the existence of a sister, outgroup clade made up of 26 isolates, sampled in the UK and Sweden between 1995 and 2009, further indicating that there may be some deeply diverged population structure within this cluster as well. Despite the existence of such an outgroup clade, however, the date of the most recent common ancestor of this lineage, 1970 (CI: 1969.26-1970) suggests that what structure exists is likely to have arisen very recently, indeed from an most recent common ancestor similar to the isolate collected in 1970. This gives us some idea of how long it takes for population structure to develop in *N. meningitidis*, and it seems that even under conditions of sympatry, distinct lineages can emerge within 20-30 years.

Cluster 8 is one of the few clusters which is relatively balanced in its composition between carriage and disease isolates, with 42.93% isolates having been collected from healthy carriage. Isolates from carriage and disease are not particularly well mixed, but this is likely to be due to their different geographic regions and dates of isolation. Carriage isolates were mostly collected from the meningitis belt of Africa, in particular from Burkina Faso (121 isolates), Chad (96), and Ghana (65), with smaller numbers in Mali (26), Guinea (4), and Niger (1). There are also sporadic samples from parts of Africa other than the

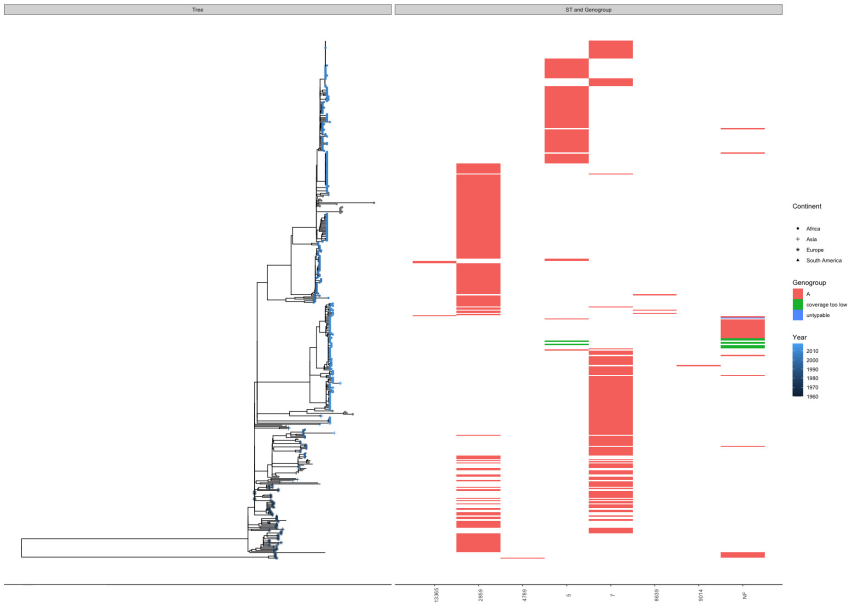


Figure 3.14: Whole-genome phylogeny of Cluster 8 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/5EDxMq6KjkUkBiPHeK497f`

meningitis belt, with 8 isolates from 3 counties in Southern Africa, and 4 isolates from 3 countries in central Africa. Other than the African isolates, Cluster 6 includes 62 isolates collected in Europe, primarily from Norway (28) and Sweden (22), but also from Denmark, the Netherlands, France, Russia, and the UK. Cluster 8 also includes 8 isolates from Asia (China and the Philippines) and 7 isolates from Brazil. 35 years between 1960 and 2018 are represented in the sampling, and as with the collection as a whole, the majority of the sampling is around the late 2000's and early 2010's with 2009 and 2011 having the most isolates per year (83, 67 respectively). The isolates are generally well spread between 1970 and 2014, with the biggest gap in sampling between 1990 and 1995.

Cluster 8 predominantly consists of genogroup A isolates (99.78%) though there is also one non-groupable isolate. Despite the lack of genogroup diversity, there is some diversity in sequence types, where Cluster 8 is made up of 7 different sequence types. The most common is ST-2859, making up 37.39% of Cluster 8, followed by ST-7, making up 19.66%, and ST-5, at 7.69%. The sequence types are generally distributed in line with

common ancestry in the phylogeny, with different STs making up generally monophyletic clades of many isolates. Unlike the distribution of countries across the phylogeny of Cluster 7, in Cluster 8 countries also tend to be distributed this way, with the exception of some of the meningitis belt countries, which are more intermixed with one another, in a pattern which suggests that there is not the same deep diverge between different lineages within this cluster. Despite this, there is also an outgroup clade in Cluster 8, consisting of 20 isolates sampled in the Nordic countries between 1970 and 1978, contemporaneous with other isolates in the main clade, suggesting that even in clusters without strong evidence of an internal globally-distributed population structure, there seems to be evidence that there are sub-lineages sampled in historical sampling which have either gone extinct or are generally undetectable in contemporary sampling. There is an additional isolate, unfortunately without geographical or temporal data, which is on a long branch very distant from the rest of the cluster this isolate likely represents a minor isolate sampled from the otherwise unsampled diversity of Cluster 8, hinting that the population structure may be more complex than our current sampling reveals. Unfortunately, it is also likely this sample on a distant branch renders the accurate dating of the most recent comon ancestor of the lineage impossible, giving a confidence interval of nearly 100 years, with the estimate being 1802.3 (CI: 1745.45-1841.55).

Cluster 9 is unique among the 25 clusters in that all 427 of its isolates were collected in meningitis belt, predominantly in Burkina Faso and Ghana – from which 193 and 188 isolates were collected, respectively, but also from Niger, Mali and Chad, where 24, 21, and one isolate were collected. Despite the limited geographical distribution of the isolates in this cluster, their temporal distribution is more diverse than expected, ranging from 1978 to 2012. Though most of the sampling, as with the entire collection, is between 2008 and 2012, small numbers of isolates were collected in 2006, 1997, and a single isolate in 1978. Unsurprisingly given the lack of diversity in the isolates' geography, there is also very little diversity in the sequence types and genogroups which make up Cluster 9, where 99.77%

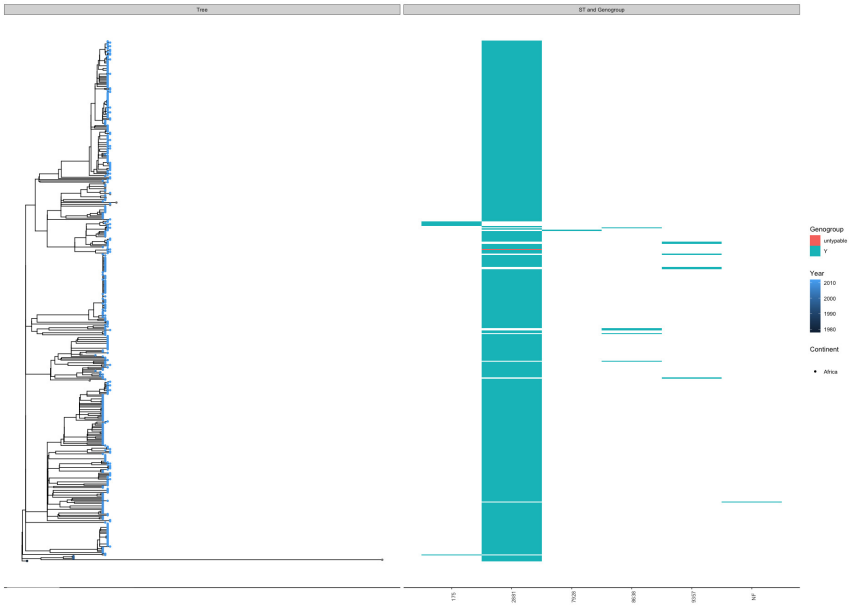


Figure 3.15: Whole-genome phylogeny of Cluster 9 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/n1JzowU9pQ9SCuLQgDV5fe`

of the isolates in the cluster are genogroup Y isolates, with a single non-groupable isolate, and 95.84% of the isolates are typed as ST-2881, with 3 other sequence types making up the remaining 4% of isolates, and one isolate in the cluster which had allele patterns which did not match any of the sequence types currently in the database. Also given the sampling, it is unsurprising that the majority of isolates in Cluster 9 are carriage isolates, 92.81%. Despite this collection's bias toward carriage samples in the meningitis belt there are 30 isolates in this cluster collected from cases of invasive disease.

The distribution of sampling dates and locations across the phylogeny of Cluster 9 is exactly as one would expect for a clonal lineage of bacteria with the straightforward population structure that implies. The oldest isolate, collected from a case of invasive disease in Burkina Faso in 1978, is the outgroup to the entire collection, which we would expect if there were no deep divergences or other complex population structure. Not only is the oldest isolate the outgroup to the entire lineage, the outermost group of the sister clade to the outgroup contains the three second-oldest isolates, collected in 1997. Further, while there

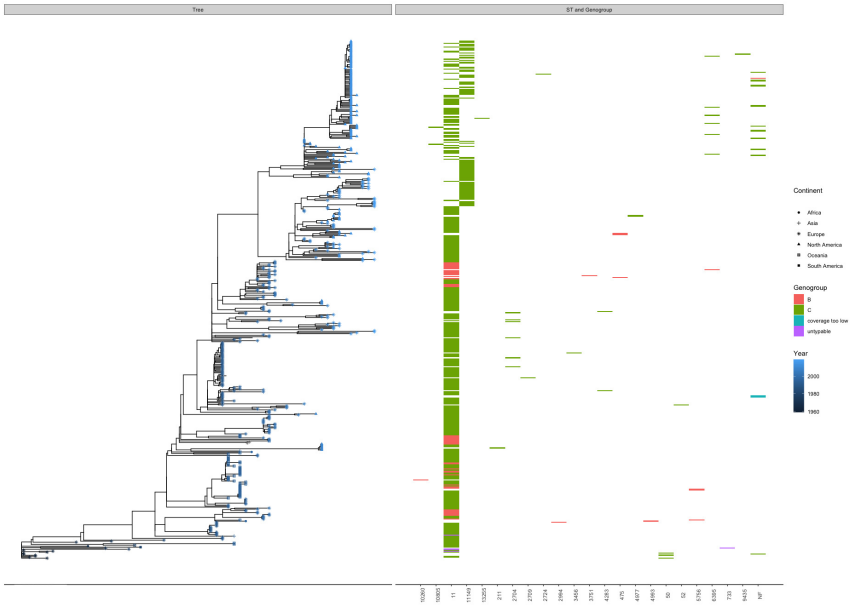


Figure 3.16: Whole-genome phylogeny of Cluster 10 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/umuo5d3uxK3vmJYF5Mu7hn`

are some isolates which have inconsistent geography given their location on the phylogeny, these are likely to simply represent migration events, as in general isolates from each country form one or more monophyletic clades. The most recent common ancestor of the lineage is also estimated with high confidence to have been in 1975.94, (CI: 1975.04-1976.61), suggesting a recent population expansion. All of this is consistent with Cluster 9, or at least the isolates we have sampled of it, having a simple clonal population structure, and combined with its highly localised geography, it seems that it potentially represents a lineage which formerly existed at very low frequency in the meningitis belt, much like many of the hundreds of minor clusters in the global collection, before expanding substantially before or around the time period where much of the sampling in this collection took place.

Cluster 10 is primarily split between North America and Europe, with 287 of its isolates having been collected in North America, and 286 in Europe. Smaller numbers were collected in Oceania – 51 in Australia and one in New Caledonia, Africa – one in Niger, Asia – two in Israel, and one isolate in South Amer-

ica, in Brazil. Within North America, the sampling is almost entirely from the United States, with three isolates from Mexico and the remaining 284 from the US. 10 countries are represented in the samples from Europe, but the sampling is dominated by the Netherlands, with 121 samples, the Czech Republic, with 76, and the UK, with 73. Alongside their geographical distribution, there is a significant spread in the temporal distribution of Cluster 10 isolates as well, which have been collected in 42 years between 1959 and 2019, with consistent sampling from 1991 onward, 2015 being the peak of the sampling with 137 isolates collected within that year. Perhaps unsurprisingly given that the sampling in Cluster 10 is focused on regions where the amalgamated collection consists mostly of disease samples, relatively few isolates from Cluster 10 were collected from carriage infections, only 5.911%. Cluster 10 does contain just under half of the urethritis isolates studied in *Expansion of a urethritis-associated Neisseria meningitidis clade in the United States with concurrent acquisition of N. gonorrhoeae alleles* [70], meaning that 34.19% of the isolates in Cluster 10 were collected from cases of urethritis.

In keeping with its geographic and temporal diversity, Cluster 10 is also fairly diverse in terms of the sequence types and genogroups present within. 25 different known sequence types were detected in Cluster 10, of which the dominant sequence type was ST-11, making up 72.53% of the lineage. Most of the switches to other sequence types are sporadic single isolate switches, apart from two slightly larger clades which have switched to ST-11149. In terms of genogroup, Cluster 10 is predominantly composed of genogroup C isolates, making up 90.39% of the cluster's isolates, but also contains many clades which have switched to genogroup B, making up 8.661% of the collection, and an additional two non-groupable isolates.

The distribution of sampling dates and locations across the phylogeny of Cluster 10, similar to the phylogeny of Cluster 9, does not indicate the existence of deep divergence and complex structure within Cluster 10, instead, the outgroups primarily consist of pre-1991 isolates. Monophyletic clades away from the root of the phylogeny, consisting of samples from 1991 onward,

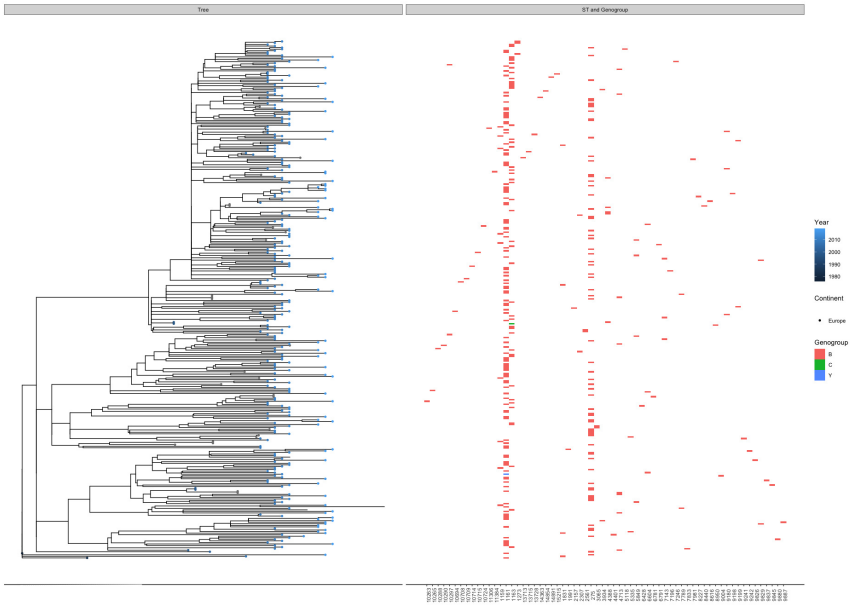


Figure 3.17: Whole-genome phylogeny of Cluster 11 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/2hCKFkfdLGdL1X1EyEUu7l`

do show some evidence of migration between regions, particularly between the US and Europe. Combined with a relatively recent estimate of the date of the most recent common ancestor of the lineage, 1958.66 (CI: 1958.41; 1958.87), this suggests relatively straightforward population structure within this cluster, albeit combined with high rates of movement across what we imagine are likely geographical boundaries. While this migration has been evident in the phylogenies of some clusters previously, the extent of migration between different continent within a relatively short time frame (2007-2017) in Cluster 10 suggests that this is definitely an active, ongoing process in Cluster 10, unlike the relatively or entirely regional dynamics observed in some other clusters, such as Cluster 9.

Cluster 11 is unique in that it is the only cluster among the 25 major lineages in the collection to be located entirely in Europe. In fact, even within Europe, most isolates (300) were collected in just one country, the UK. Other than isolates collected in the UK, Cluster 11 includes 17 isolates collected in Ireland, 11 collected in the Netherlands, and a three single isolates collected in the Czech Republic, Denmark, and Malta.

Despite its narrow geographic range, Cluster 11 does have a relatively wide temporal spread, containing isolates sampled between 1976 and 2019, though the sampling is very shallow, apart from the years between and including 2009-2013 and 2017-2019, no more than two samples were collected in any other year. Like Cluster 10, the fact that this cluster is geographically restricted to regions where the samples are predominantly collected from invasive disease unsurprisingly means that the isolates are overwhelmingly collected from cases of invasive disease. Only 2 isolates (0.639%) were collected from cases of healthy carriage in Cluster 11.

Despite the relatively little geographic and temporal diversity in this cluster, there remains significant diversity in the sequence types of Cluster 11 isolates. 64 different sequence types were identified in this cluster, with three sequence types in particular making up most of the cluster. The largest being ST-1161, making up 34.43% of isolates, followed by ST-275, at 20.66% of isolates, and finally ST-1163, which includes 14.07% of the isolates in Cluster 11. Despite the relatively substantial diversity in sequence types, there is substantially less diversity in the genogroups of Cluster 11 isolates, 99.40% of isolates are genogroup B, with the remaining 0.6% of isolates consisting of two single genogroup Y and genogroup C isolates.

Unlike Clusters 9 and 10, Cluster 11 has more evidence of some population structure within the cluster, as apart from four outgroup isolates which were collected in 1976, 1985, 2007, and 2018, there are two main monophyletic clades which make up the rest of the phylogeny. Both of these lineages span the late 1990's to 2019, and have overlapping country distributions, between the UK, the Netherlands and Ireland. Despite this, the most recent common ancestor of Cluster 11 is estimated to be very recent with quite high confidence, in 1976 (CI: 1975.04-1976), suggesting that the evidence of within-cluster structure in the phylogeny has arisen very recently.

Cluster 12 is the first of the smaller major clusters made up of only a few hundred isolates to have more of a truly global distribution. The continent most represented in terms of isolate origins is Africa, where 130 of the cluster's isolates were sampled

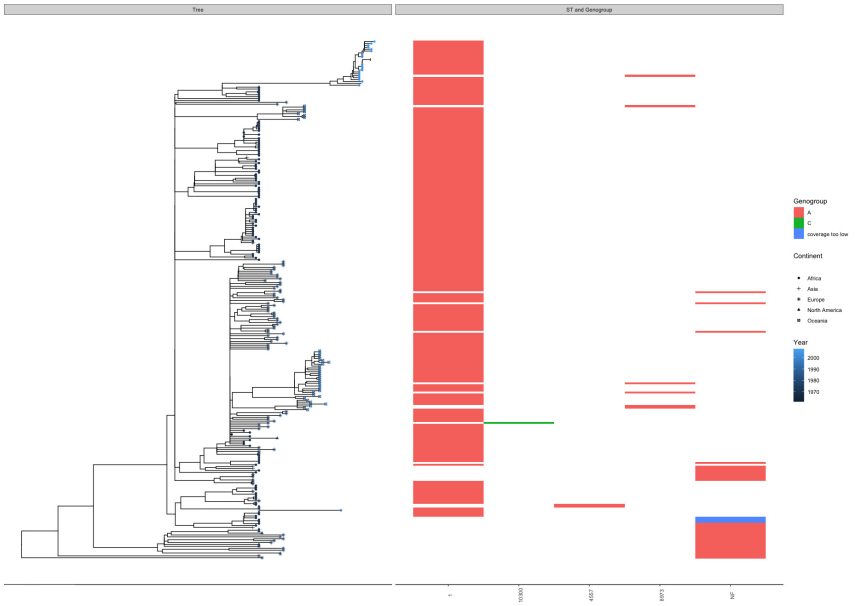


Figure 3.18: Whole-genome phylogeny of Cluster 12 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/jrrcWdcMtfKGHL3fAMWWU7`

but it is closely followed by Europe, where 80 isolates in Cluster 12 originate, and then Oceania with 40 isolates, eleven in North America, and four in Asia. Its temporal spread is similarly wide, with isolates sampled from 31 years between 1961 and 2008, with most of the isolates having been collected between 1968 and 1970, though a substantial number of samples were also collected in 1990. Unlike other clusters predominantly located in Africa which were primarily sampled in carriage surveys, Cluster 12 mostly consists of isolates collected from cases of invasive disease, with only one carriage isolate collected in the UK in 1978.

Despite the relatively diverse geographical and temporal origins of the isolates in Cluster 12, they possess very little diversity in terms of ST and genogroup. All but one of the isolates are genogroup A, with a single genogroup C isolate. In terms of sequence type, only four defined sequence types are present, predominantly ST-1 (88.03%). The second largest sequence type present is actually not a sequence type at all, but instead the absence of an identified sequence type in the database, which included 13.36% of isolates.

Cluster 12 does not have much evidence of especially deep diversity in the shape of and distribution of metadata across its phylogeny, it general follows the expected pattern of having older and geographically distant isolates in outgroup branches, and more recently sampled isolates in larger numbers form the innermost clades. There are some exceptions to this – the outermost outgroup is contemporaneous with the bulk of the sampling despite those isolates forming an internal monophyletic clade, and it is likely that this outgroup is driving the extreme inferred age of the most recent common ancestor, which is estimated to be in 1891.87 with some degree of confidence (CI: 1876.92-1904.56). The extent of distance between this outgroup and the rest of the phylogeny, despite originating in the same place (France) and within 7 years earlier and one year later than a total of 52 other isolates within the cluster demonstrates the extent to which present sampling cannot necessarily capture historical diversity, and the extent to which bottlenecks have recently affected the evolution of *N. meningitidis*. This is echoed in a unique property of Cluster 12 among the major clusters, that none of its isolates were sampled from this cluster after 2008, when the most intensive sampling in entire amalgamated collection is in the decade following 2010. Time will tell if this lineage has truly gone extinct, but given the diversity which has been sampled in regions where the sampling has been particularly intense, it does not seem unreasonable that this lineage may persist in parts of the world which are not very well represented within this amalgamated global collection. This is the first major lineage which has not persisted up until the end of the entire sampling window despite the regions in which it is prevalent being well-covered after the last sample in this cluster was collected, and regardless of whether or not it is truly extinct globally, it does again suggest that significant changes in the population structure of *N. meningitidis* can occur within relatively short periods of time – of around 20 years.

Cluster 13 is made up of predominantly isolates collected in Europe, 225, combined with an additional 27 isolates collected in North America, two in Asia, and one in Oceania. Within Europe, most of the isolates were collected in the UK (192) though a

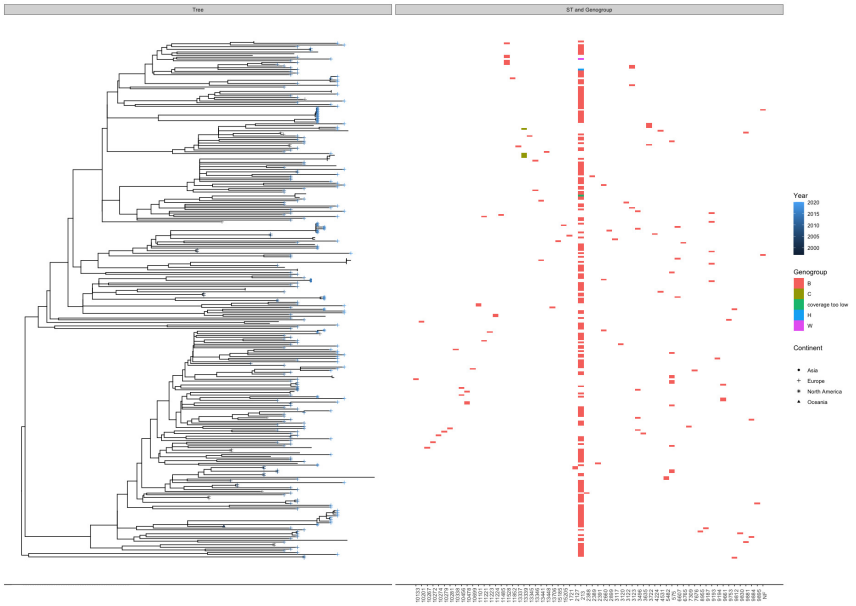


Figure 3.19: Whole-genome phylogeny of Cluster 13 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/tV1NkodWeCFcJDRR76u7QL`

notable number (> 10) were also collected in Netherlands and Ireland. The temporal spread of the isolates in Cluster 13 is similarly limited, being restricted to 20 years between 1997 and 2020. Despite the geography being largely limited to regions with primarily disease isolates, the cluster contains a relatively high number of carriage isolates compared to other lineages largely localised to Europe and Oceania – with 13.37% of isolates in the cluster having been collected from carriage infections.

Despite how concentrated the sampling distribution of Cluster 13 is around isolates which were collected in the UK between 2010 and 2020, there is substantial diversity present within the cluster in terms of the sequence types and genogroups present. Isolates from Cluster 13 were identified as 62 different sequence types, though ST-213 is by far the most prevalent, with 61.28% of isolates in Cluster 13 having been typed ST-213. Apart from ST-213, no other ST is present in any substantial numbers, reflecting what is known about the ST-213 clonal complex. A similar pattern exists in terms of the genogroups identified within Cluster 13. The vast majority of isolates (97.97%) have been identified *in silico* as possessing the genes required to produce a

serogroup B capsule, but much diversity exists in low frequency. Three other geogroups are identified within Cluster 13 – four genogroup C isolates, clustered together on the phylogeny, one genogroup H isolate, and one genogroup W isolate.

The phylogeny of Cluster 13, like those of many other clusters which have been examined, shows some evidence of a structured population, even within the cluster. The outgroup clade of the cluster's phylogeny consists of four isolates, collected 2012, 2013, and 2018 from the UK, despite the existence of older isolates within the cluster. The main clade of phylogeny is then split into two sister clades of roughly the same size, both overlapping in time and space, and with the same dominant sequence type. As with many other clusters with a similarly shaped phylogeny, this provides good evidence that there are several lineages of this cluster circulating within the global population, two of which have been sampled quite heavily in this collection, but at least three of which exist. Despite this evidence of deep population structure, the most recent common ancestor of the entire lineage is estimated to be quite recent, in 1970.78 (CI: 1963.55-1975.73) suggesting, like some other clusters, that this structure has arisen recently.

Again found primarily in Europe, with 158 of its isolates having been collected there, is Cluster 14. Despite its similarity to many other clusters in this regard, it is unusually also found in relatively substantial numbers in South America, with 22 of its isolates having been collected there. An additional 19 of its isolates were also collected in Oceania, 13 in Africa, four in North America, and one in Asia. Within Europe, the Netherlands is unusually the primary source of the collected isolates, with a handful of isolates from Iceland, Denmark, France, the UK, and Ireland, and single isolates from Malta, Norway, and Sweden. The South American isolates are predominantly from Argentina (21) with the remaining one from Chile. The isolates are fairly well spread temporally, having been collected in 39 different years between 1961 and 2012, with continuous sampling (at least one isolate per year) between 1965 and 1988, and missing only one year (1997) between 1990 and 2002.

For its size, relatively few different sequence types are present

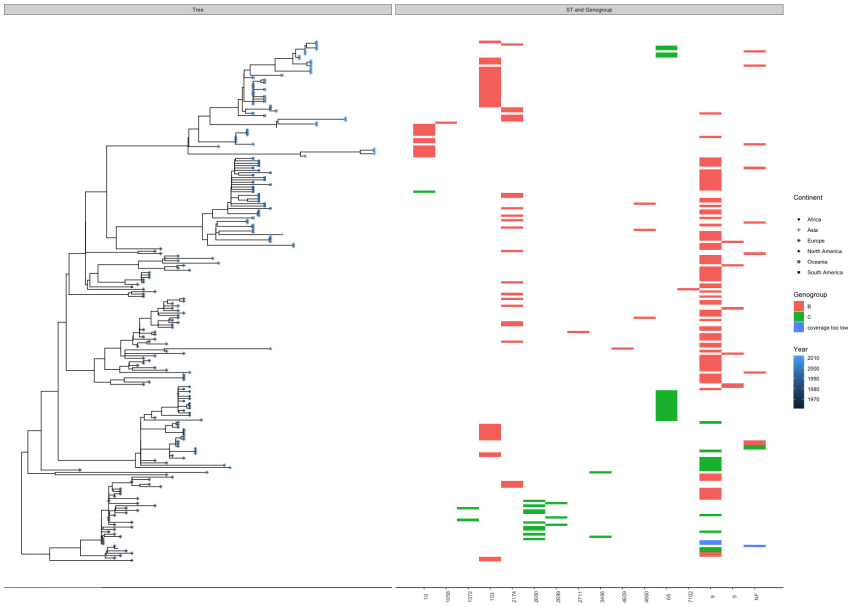


Figure 3.20: Whole-genome phylogeny of Cluster 14 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/6LQuGB2EHGdac2ZyAcfVxa`

within Cluster 14, only 15 unique sequence types found in the database. Cluster 14 is not, however, dominated by a single sequence type, the most prevalent is ST-8, which includes 41.70% of isolates, but other sequence types, particularly ST-153 and ST-2174, are also present and include more than 10% of isolates (14.35% and 10.31%, respectively). Similarly, though Cluster 14 does not contain many different genogroups, it contains several different clades which are partially or mostly made up of genogroup C isolates, making up 22.22% of all the isolates successfully genogrouped *in sillico*, despite the predominant genogroup in Cluster 14 being genogroup B, consisting of the remaining 77.78% of isolates.

The phylogeny of Cluster 14 is one which is generally in keeping with a simple population structure of a clonally expanding organism, though with some exceptions. More ancestral clades, including the outgroup, tend to also consist of older isolates, but the oldest isolate is actually located within the three nodes from the root. The two most derived clades consist of the most recent isolates, though only one of them extends until the present day, whereas the other was last sampled in 1998. The most recent

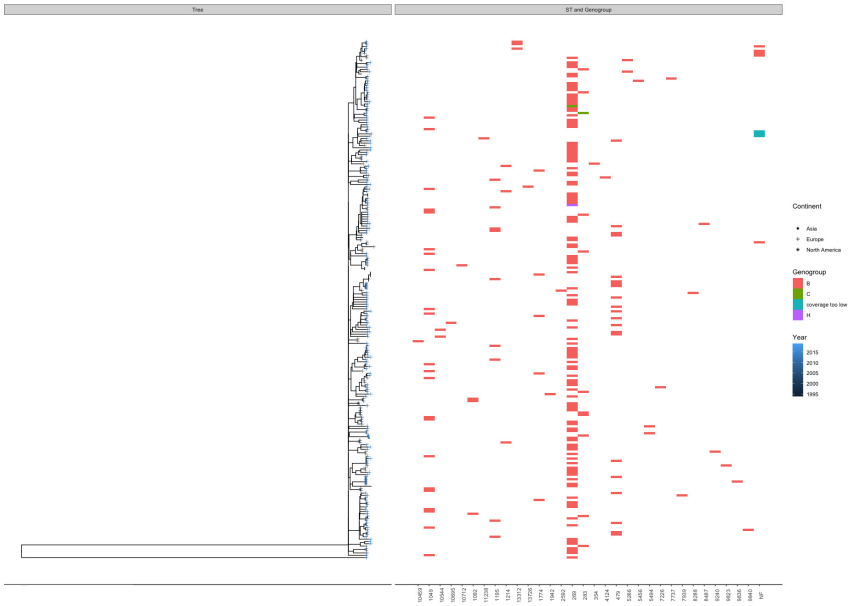


Figure 3.21: Whole-genome phylogeny of Cluster 15 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/oam5pGj29nRAUvhRukiAVp`

common ancestor is estimated to be around 20 years earlier than the earliest isolate collected in this cluster, in 1950.92 (CI: 1947.66-1953.15). This pattern fits a population undergoing waves of clonal expansion from which some isolates persist and go on to expand into a subsequent daughter lineage, but most go extinct. Indeed other than the isolates in the most derived lineage in the phylogeny of Cluster 14 – furthest from the root – no isolate is sampled more recently than 1998.

Cluster 15 is one of the more geographically restricted clusters, being found almost entirely in Europe (203 isolates), with only 10 isolates collected outside Europe, eight in North America (7 US, 1 Canada) and two in Asia (Turkey). The isolates are similarly restricted in their temporal spread, having been collected in 21 years between 1994 and 2019, with more than half of all the cluster's samples having been collected between 2009 and 2013. Though it would be unsurprising, given the distribution of the entire collection, for a cluster with isolates collected in these regions to be predominantly composed of disease isolates, Cluster 15 exceeds this expectation, being the only cluster among the 25 major clusters to contain no isolates

collected from healthy carriage infections.

The diversity in sequence types identified within Cluster 15 is around the mean, with 31 different sequence types present in the set of isolates making up Cluster 15, with a single predominant sequence type, ST-269, including 50.44% of isolates. Despite this being a relatively low proportion for a dominant sequence type, no other sequence type present in Cluster 15 makes up of more than 10% of the total number of isolates. Genogroup diversity is present but not substantial in Cluster 15, with 98.65% of isolates typed as genogroup B isolates, with the remaining 1.35 % consisting of two genogroup C isolates and one genogroup H isolate.

One outgroup isolate on an extremely long branch renders the date of the most recent common ancestor of Cluster 15 far older than can be inferred with confidence. The date of the most recent common ancestor is estimated to be in 1320.47, but with a confidence interval of nearly 500 years between 1096.37 and 1500.72. Apart from this long branch, however, the phylogeny is made up of very short branches which form a substantial number of monophyletic clades of relatively small degree. These small monophyletic clades tend to be made up of isolates from the same country, with the exception being the UK and Ireland, which tend to form mixed clades with one another. From this, Cluster 15 again appears to be a cluster in which almost all of the isolates collected, apart from the single outgroup isolate, are from a recent clonal expansion. This is despite being the outgroup isolate being contemporaneous (2010) and in the same country (UK) as many other isolates in the collection, which when combined with the other characteristics of the phylogeny seem to suggest a recent population expansion from low frequency but high standing diversity, which favoured one of the lineages present within this cluster expanded, while the others have not, based on the isolates collected here.

Cluster 16 is another cluster whose isolates were mostly collected as part of the large-scale carriage surveys in the meningitis belt, and as such, is predominantly sampled in Africa, with only 18 of its isolates having been collected elsewhere, 16 in Europe and 2 in Asia. Even within Africa, and within the

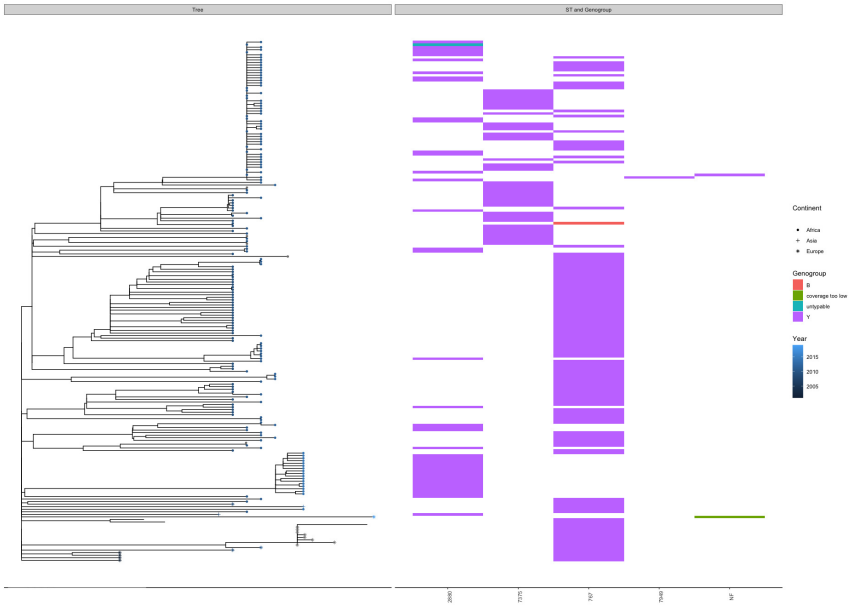


Figure 3.22: Whole-genome phylogeny of Cluster 16 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/8nop2CjXvBomum8ai2iAop`

meningitis belt, most of the samples were collected in Burkina Faso (135) with smaller numbers of isolates collected in Ethiopia (18), Mali (12), Senegal (7), Niger (7), and Ghana (4). Outside the meningitis belt, most isolates were collected in the UK (15) with single isolates collected in Ireland, Turkey, and India. The sampling window of Cluster 16 is, in keeping with its relatively narrow geographic range, one of the more narrow sampling windows, with isolates having been collected from 8 different years between 2001 and 2019, though almost all of the sampling occurred between 2009 and 2012. Cluster 16, having been primarily sampled in carriage surveys from the meningitis belt, is almost entirely compose of carriage isolates, apart from 3 single disease isolates collected in the UK, Ireland, and Turkey.

Perhaps unsurprisingly given its narrow sampling range, in terms of both geography and time, Cluster 16 has very little sequence type diversity present, with only four different sequence types identified within the cluster. The most prevalent is ST-767, with 57.35% of isolates in the cluster being typed as this sequence type. Two of the remaining three sequence types are present at a similar frequency, ST-2889 at 21.08%, and ST-7375

at 20.10% of isolates. The final sequence type present within Cluster 16, ST-7949, consists of only a single isolate. Unlike the level of sequence type diversity, genogroup diversity within Cluster 16 is at a similar level with other clusters. It is 99.01% made up of isolates that are genogroup Y, but does contain a genogroup B isolate and a non-groupable isolate, likely capsule null.

The phylogeny of Cluster 16 is unusual and in some ways unlike any of the other clusters' examined so far. The two of the most derived – furthest from the root – clades are the largest monophyletic clades in the phylogney, and include more than half the isolates in the Cluster. Isolates in these two clades originate from Mali and Burkina Faso, and were collected between 2009 and 2012, roughly in the middle of the cluster's sampling window. The immediate sister clade to the two most derived clades, consisting of less than a quarter of the total number in the cluster, follows a similar pattern, as it is detected between 2009 and 2012 in the western portion of the meningitis belt, but then is also found in 2014 in Ethiopia, in the east. The more ancestral lineages making up the rest of the cluster form outgroups to these these two main lineages – some of which consist of a single isolate – which include isolates collected before, in the same time period, and after the isolates in the two main lineages. Despite this unusual shape of the phylogeny, the date of the most recent common ancestor is estimated to be shortly before the earliest isolates in this cluster were collected, 1994.05, with relatively high confidence (CI: 1992.19-1995.31). Combined, these facts seem to suggest that what we see in the population structure of Cluster 16 is a lineage of bacteria which existed at low frequency in the meningitis belt around the time the first isolates from this cluster were collected, then underwent a population expansion, possibly due to the population perturbation caused by the mass vaccination campaign in the meningitis belt around that time. This was detected in the large number of samples collected in the middle of the total time period this cluster was sampled, but then not in the same region toward the end of the sampling, where instead an outgroup isolate, presumably from the extant global diversity within this cluster, was detected in Ireland, a

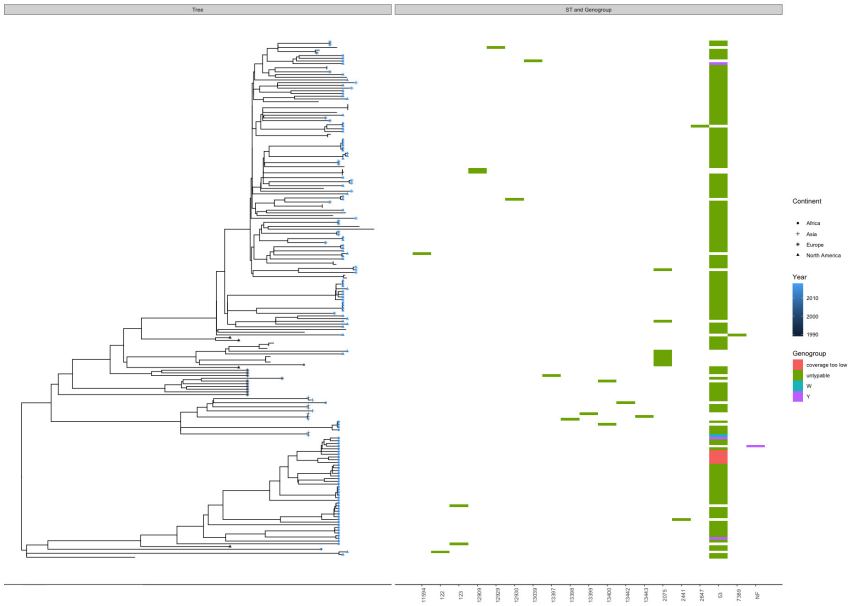


Figure 3.23: Whole-genome phylogeny of Cluster 17 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/qdj3w5JXLiCABVEGwvDpvb`

surprising example of the extent global spread of lineages of this cluster which have only barely been sampled into this collection.

Unlike the previous four clusters of a similar size, Cluster 17 is much more globally distributed. Most of the isolates were collected in North America, (64) and given the national origins of North American isolates within the global collection, it is unsurprising that the US – being the largest source of isolates within North America – is the single country from which the most isolates were sampled, with 60 isolates in the cluster having been collected there. Outside of North America, substantial numbers of samples were also collected in Africa (47), mostly from Ethiopia (44), Europe (34) mostly from the UK (10) and the Czech Republic (10), and Asia, where 10 isolates in this cluster were collected in India. The temporal spread of the isolates is also relatively substantial with isolates having been collected in 17 different years between 1989 and 2018. There is a significant gap in the temporal continuity of isolates in this cluster, however, with only a single isolate having been collected between 1993 and 2006, in 2001. Despite its global spread in regions which are predominantly composed of disease isolates,

Cluster 17 is predominantly made up of carriage isolates, with only 7 disease cases spread between the US, UK, Cuba, and South Africa.

Cluster 17 does not have a particularly notable amount of either sequence type diversity or genogroup diversity. 18 sequence types are present in the cluster, though a single sequence type, ST-53, dominates, with 85.43% of the cluster's isolates being identified as this type. The overwhelming majority of isolates in the cluster are non-genogroupable, strongly suggesting that they are capsule null, though there are four isolates identified as genogroup Y and a single isolate identified as genogroup W spread out across the phylogeny.

Its phylogeny, however, is a familiar shape, like the phylogeny of Cluster 4, it bisects into two main lineages. Both of these lineages contain isolates which span the sampling interval and contain some geographical diversity, although the larger is spread much more evenly between 1989 and 2018, whereas the smaller lineage is restricted to the meningitis belt and North America, and contains only isolates from 1989 and 2010-2016. Unlike Cluster 4, despite the bisecting phylogeny, the date of the most recent common ancestor of Cluster 17 is still estimated with a relatively high degree of confidence as being in 1941.01 (CI: 1922.83-1949.99) suggesting that this divergence is relatively recently evolved.

Cluster 18 is another cluster which is almost entirely restricted to Europe, with only seven of its isolates having been collected elsewhere, four in the US, and single isolates in South Africa, Turkey, and Australia. Within Europe, most of its isolates – 68 –were collected in the UK, though a substantial number of isolates were also sampled in the Netherlands (11) and the Czech Republic (4), and a handful across Ireland, France, and Malta. Despite their relative lack of geographic spread, the isolates were collected across 22 years between 1970 and 2018, though the relatively few isolates in this cluster does mean that the sampling, which is concentrated in time between 2010 and 2013, is spread quite thinly across the rest of the sampling interval. Like other clusters which are predominantly sampled from Europe, Cluster 18 consists mostly of isolates sampled

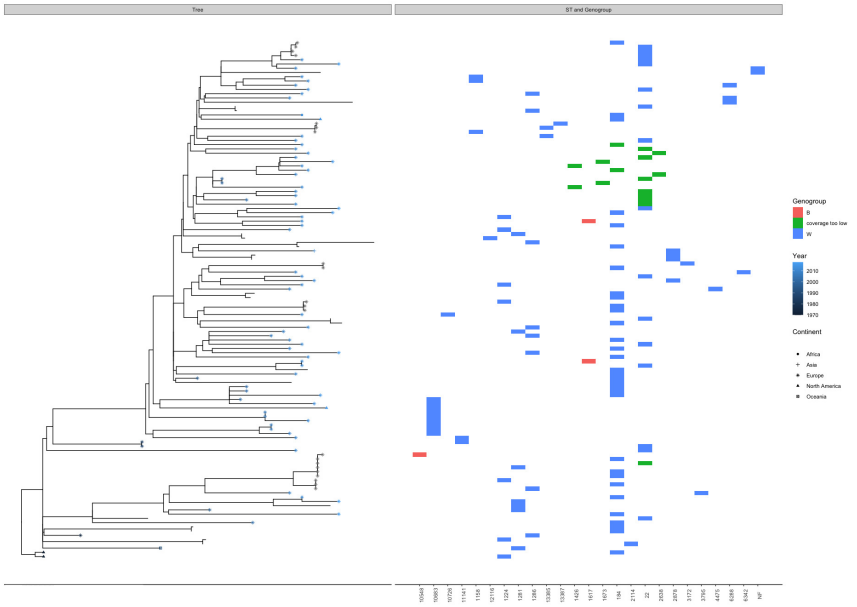


Figure 3.24: Whole-genome phylogeny of Cluster 18 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/af169PMU9gf5fnBVCZedGJ`

from disease, but still contains a decent proportion of carriage isolates: 10.53% of the isolates in the cluster were collected from healthy carriage infections.

Like a few other clusters, Cluster 18 is another cluster which does not have a single dominant sequence type. ST-184 is the most prevalent sequence type, at 27.86% of the cluster's isolates, and ST-22 is the next most prevalent, at 19.67% of the cluster's isolates. No other ST comprises more than 10% of the total number of isolates in Cluster 18. Interestingly, neither of these two sequence types form a large monophyletic clade, and are instead interspersed among isolates of different sequence type across both major clades. As for genogroups, Cluster 18 is overwhelmingly typed as genogroup W (97.17%), with 3 exceptions spread out through the phylogeny, all of which are genogroup B.

The phylogeny of Cluster 18 has, as an outgroup, two isolates collected in 1970 from the US. The remaining isolates in the phylogeny then form two sister clades, one around four times the size of the other, though both display some temporal and geographical diversity, overlapping with one another in Europe.

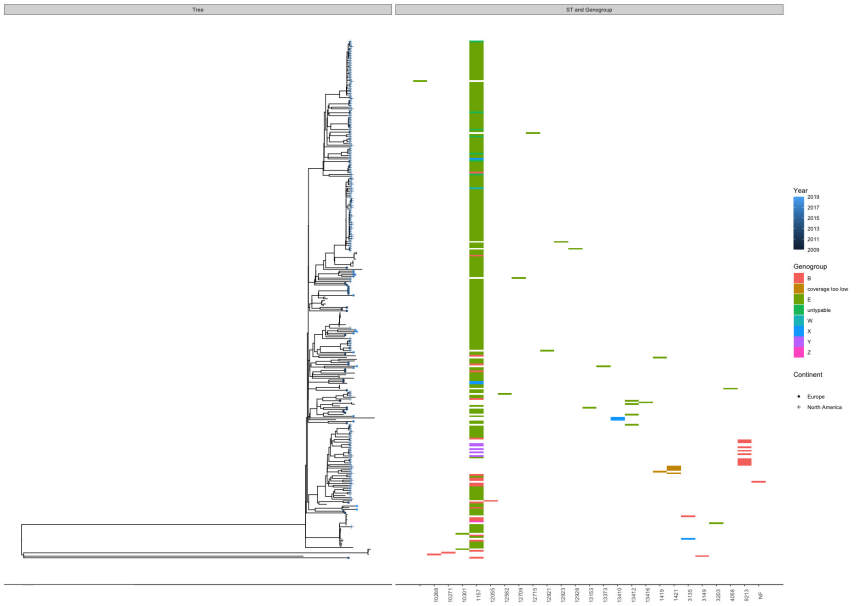


Figure 3.25: Whole-genome phylogeny of Cluster 19 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/iwyfcw9zzMatahAANU3SAe`

The date of the most recent common ancestor of the entire cluster, 1966.43 (CI: 1964.84-1967.53) confirms that the significant split between the two lineages is somewhat recently developed possibly due to independent founder events from the diversity present around 1970.

Cluster 19 is the first cluster to have been predominantly collected in North America, predominantly in the United States. 182 of its isolates were collected there, with an additional 50 collected in Europe. Most of the European isolates were collected in the UK (39), with the remaining eleven split between Ireland, the Netherlands, the Czech Republic, and Malta. The temporal spread is fairly narrow, from 2009 to 2019, but with consistent sampling, with samples having been collected during every one of those eleven years. Despite this narrow temporal spread, more than half of the samples were collected in 2015, with 2016 being the only other year to have had more than 10 samples collected. The samples were predominantly collected from healthy carriage infections, which made up 88.58% of the total number of isolates in the cluster for which disease metadata was available, and there were an additional 25 isolates collected from cases of invasive

disease, in the US, Malta, and the UK.

The diversity in sequence types in Cluster 19 is relatively minimal compared to its size, with 24 different known sequence types present. The cluster is dominated by isolates identified as ST-1157, however, which make up 86.29% of the cluster's isolates. Other than this main sequence type, no other sequence type accounts for more than 10% of the isolates in the cluster. The genogroup diversity in this cluster, however, is exceptional in terms of the number of genogroups present within. Like most clusters, Cluster 19 is mostly made up of a single dominant genogroup, in this case genogroup E, but it also contains five other genogroups, as well as non-groupable isolates, which are presumably cluster null. The second most prevalent genogroup in Cluster 19 is genogroup B, with 35 of the cluster's isolates being typed as genogroup B *in silico*. This is followed by seven non-groupable isolates, six genogroup X isolates, five genogroup Y isolates, one genogroup Z isolate, and one genogroup W isolate. This genogroup diversity is found throughout the phylogeny of Cluster 19, but is particularly prevalent in one clade of isolates found in the US and Europe which includes many of the serogroup B and serogroup Y isolates.

The shape of the Cluster 19 phylogeny, with a long branch separating 6 isolates from the rest of the phylogeny serves to emphasise the extent to which it is possible for low-level diversity from a former cluster expansion to survive to the present day and lead to a complex within-cluster population structure. The date of the most recent common ancestor of the entire phylogeny confirms this, with the unreliable estimate being 1818.82, with a confidence interval of over a hundred years (1756.07-1889.79). Even excluding this outgroup lineage, however, the remainder of the phylogeny contains evidence of a deep population structure. Several sister lineages containing between 50 and 150 isolates overlap in their spatial and temporal distribution, despite being separated by deep branches in the phylogeny.

The isolates of Cluster 20 were primarily collected in Europe, though there are three isolates interspersed throughout the phylogeny collected in North America, in the US. Within Europe, the collection is similarly dominated by isolates collected in

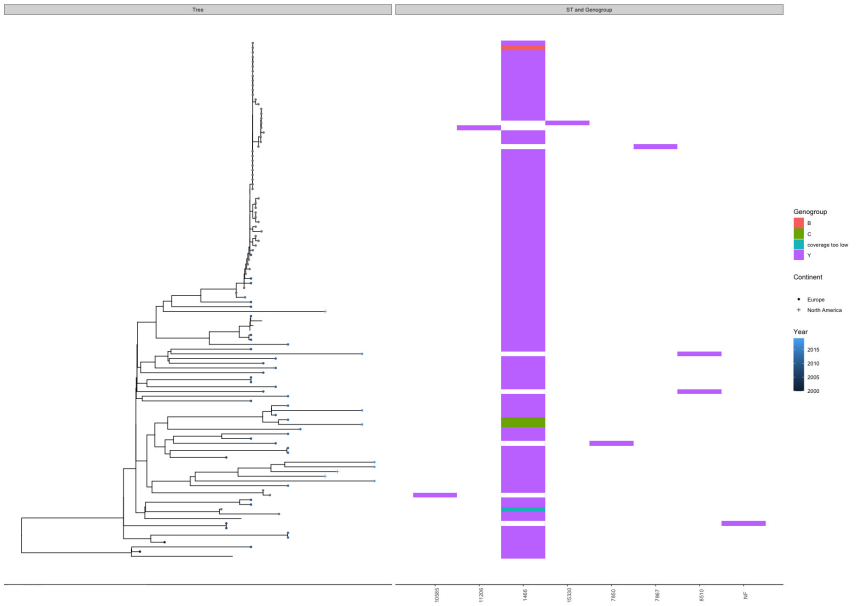


Figure 3.26: Whole-genome phylogeny of Cluster 20 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/tQyJzaBxKMMqEub8zcSeRE`

the UK, with only 9 isolates having been collected outside the UK, six in the Netherlands, two in Ireland, and one in Sweden. Isolates are still fairly restricted temporally, having been collected between 2000 and 2019, but given the relatively small size of this cluster, they are spread very thinly across even a relatively narrow sampling window, particularly because the distribution of year of collection is biased heavily toward the years 2009 and 2012, the only two years to have had more than 10 isolates collected. Across the cluster, samples were primarily collected from infections causing invasive disease, making up 82.14% of isolates for which there is disease metadata, though there are 6 carriage isolates, spread between the US and the UK.

Cluster 20 is one of the less diverse clusters in terms of sequence type, with only seven sequence types present, and despite the relatively few sequence types present, a single sequence type, ST-1466, still accounts for 91.89% of isolates, with no other sequence type accounting for more than 10% of isolates, indeed, no other sequence type has more than a single isolate within Cluster 20 being identified as being of that type. Genogroup di-

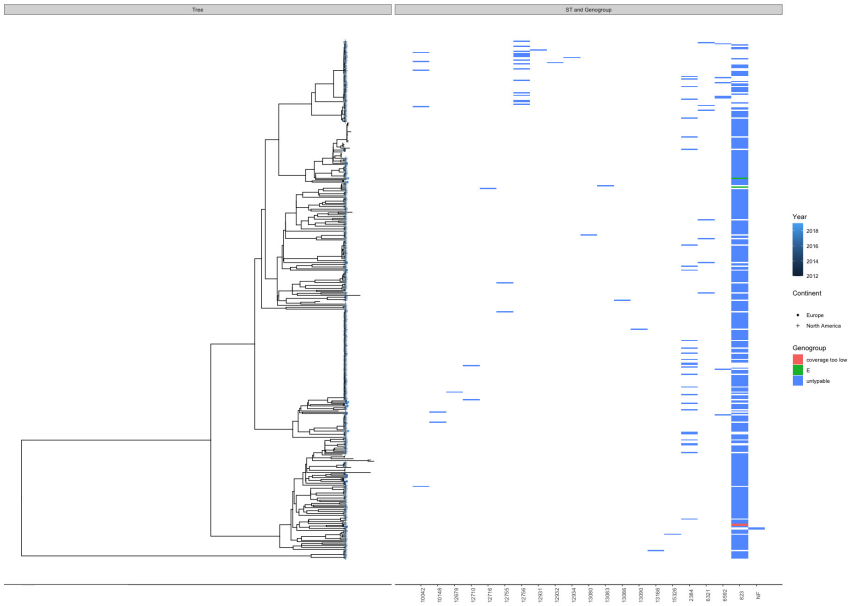


Figure 3.27: Whole-genome phylogeny of Cluster 22 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/iH7RmJ2QXRw5dT81QLEFvj`

versity remains low as well, but in line with other clusters. Most of the isolates –96.36%– in Cluster 20 were typed as genogroup Y, but there are also two genogroup C isolates and a single genogroup B isolate.

The phylogeny of Cluster 20 is much more consistent with a simple clonal outbreak, with older isolates forming outgroups, and isolates tending to be more recent as they move further away from the root. The date of the most recent common ancestor, one of the more recent of the 25 major lineages, at 1990.42 (CI: 1987.14-1993.18), supports the view that what has been collected of this cluster within the collection of isolates studied here, if not the cluster as a whole, is recently expanded and therefore is consistent with a relative lack of complexity in the population structure.

Chronologically, the second cluster to consist primarily of isolates collected in North America, and in particular the United States, is Cluster 22. Although over 90% of its isolates were collected in the US, there are 28 isolates belonging to this cluster which were collected in Europe, 15 in the UK, nine in Ireland, and one each in Austria and Malta. The time period in which

Cluster 22's isolates were collected is the narrowest of any of the 25 major clusters, though they consequently are well spread and cover all eight years of the sampling window between 2012 and 2019, with most isolates having been collected as part of the large US carriage survey, in 2015 and 2016. Consequently, Cluster 22 is composed almost entirely of isolates collected from healthy carriage, with only 4 disease isolates – 2 in the UK, and two single isolates in Austria and the US.

The pattern of sequence types across the phylogeny in Cluster 22 matches many other clusters, in that single sequence type is most prevalent, and many other sequence types are present but in very low numbers. In Cluster 22's case, the dominant ST is ST-823, which describes 80.10% of the cluster's isolates. The pattern in genogroups is similar, with almost all isolates (99.51%) being non-groupable, and therefore likely capsule null. The exception to the non-groupables in this cluster are two isolates, separate on the phylogeny, and genogrouped as genogroup E *in silico*.

Based on the shape of its phylogeny, Cluster 22 appears to have a deeply divergent and complex population structure. The estimated date of the most recent common ancestor also suggests this, as it is estimated to be beyond what can be accurately estimated using current methods, at 1564.32, with a wide confidence interval of nearly 400 years (CI:1282.97-1663.86). This is likely principally driven by the outgroup of seven isolates collected in the US in 2015, separated from the rest of the cluster on a long branch. Even excluding this outgroup cluster, however, the rest of the phylogeny also shows signs of population structure with deeply separate clades overlapping in space and time. This has all been observed in previous clusters, but the most interesting thing about Cluster 22 in particular is how the seven isolates collected in the US in 2015 are part of a large carriage survey, and detected at such low frequency. This underscores the importance of intensive carriage sampling in allowing a full understanding of the population of *N. meningitis* which may be present at any given point in time.

Although Cluster 25 is again primarily composed of isolates which were collected in the United States (65), it has a more global distribution than Cluster 22, with 30 isolates that are part

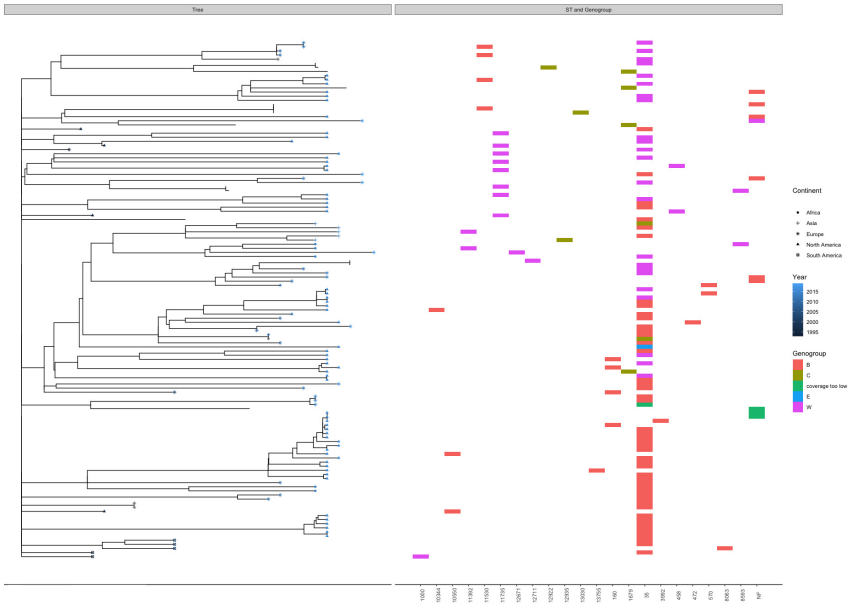


Figure 3.28: Whole-genome phylogeny of Cluster 25 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/fjdaGXQYUiQp3NpHuxB8eh`

of the cluster having been collected in Europe, seven in Africa, five in Asia, and five in South America. The European isolates are primarily from the UK (22), with smaller numbers of isolates collected in Ireland, the Czech Republic, the Netherlands, and Greece. The African isolates are split between Mali and Ethiopia, the South American isolates between Argentina and Chile, and the Asian isolates were all collected in Turkey. The time period from which the isolates were sampled is longer than Cluster 22, with the isolates having been collected between 1993 and 2019, inclusive, though like Cluster 22, the sampling is remains dominated by the isolates collected in 2015 as part of the carriage survey in the US. Primarily as a result of those isolates, Cluster 25 is mostly made up of carriage isolates, though generally fairly evenly split, with 62.96% of isolates having been collected from asymptomatic carriage infections.

21 different known sequence types are found in Cluster 25, in a pattern similar to other other clusters where a single sequence type predominates and others are found at low frequency. In Cluster 25, ST-35 is the most prevalent ST, though it only accounts for 59.25% of isolates. No other sequence type makes

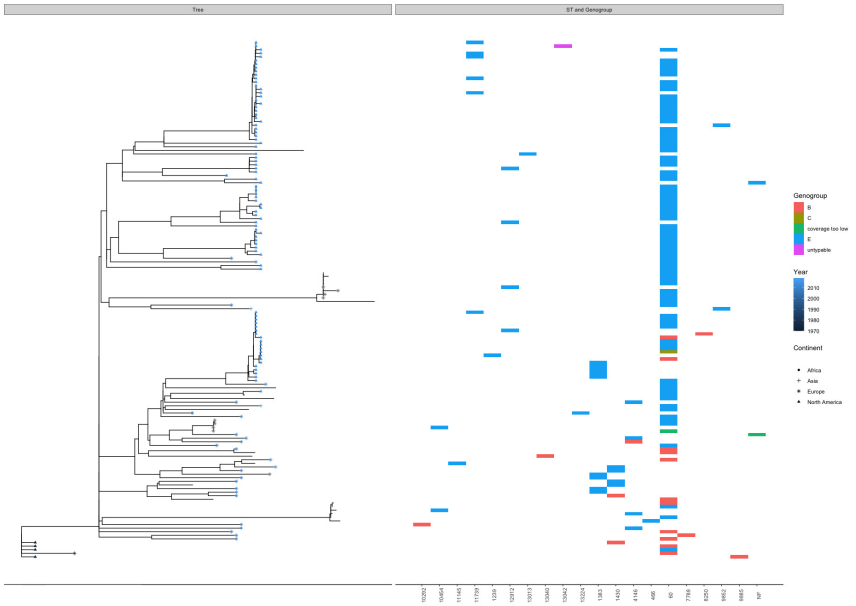


Figure 3.29: Whole-genome phylogeny of Cluster 27 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/gWr4kQVg4mzequT7dAT49B`

up for more than 10% of isolates, however, though there is some association between specific clades and certain sequence types. Cluster 25 does contain an unusual amount of diversity in terms of the genogroups present within the cluster, however, both in terms of the number of genogroups present and their proportions within the cluster. Most clusters are dominated by a single genogroup, and while that it true in the case of Cluster 25, where genogroup B predominates at 58.20% of isolates, the minor genogroups in Cluster 25 are at much high frequencies than is usual, with 33.61% of its isolates having been typed as genogroup W isolates, 7.377% of its isolates having been typed as genogroup C, and a single genogroup E isolate.

The relatively small number of isolates in the phylogeny of Cluster 25 and the multiple branch shape of the Cluster 25 phylogeny make it difficult to infer anything from the shape or distribution of attributes across the tips. The date of the most recent common ancestor, however, 1988.95 (CI:1987.74-1989.93) suggests that this is a recent, clonally expanded lineage without a complex population structure, which is not inconsistent with the general shape of the phylogeny.

Cluster 27 is another cluster which contains predominantly isolates collected in the United States (84), though like Cluster 26 it contains some isolates from elsewhere – 34 from Europe, again mostly from the UK (27), one from South Africa, and two from Turkey. Unlike some of the other smaller clusters which mostly contain isolates collected in the US, Cluster 27 contains four isolates dating back to 1970, making the sampling window range from 1970-2019, though with a significant gap between 1978 and 2002. Most isolates, like other clusters primarily containing isolates collected in the US, were collected in 2015 and 2016, and as a result Cluster 27 is similarly primarily made up of isolates collected from healthy carriage infections (78.36%) though there are 25 isolates collected from invasive disease, mostly collected in Europe, though some were also collected in the US.

The diversity of sequence types found in cluster 27 roughly matches what has typically been found in other clusters. 19 different sequences types are present, of which ST-60 is the most common, with 67.37% of the isolates in Cluster 27 being identified as ST-60 isolates, and the remaining 18 sequence types being present at low frequency. The genogroup diversity is more substantial, but similar to many of the other clusters which are on the more diverse side, there is again a single predominant genogroup, in this case genogroup E, which in this case accounts for 85.21% of the cluster's isolates. Other than this main genogroup, 13.88% of Cluster 27's isolates are genogroup B, and it also contains a single genogroup C isolate and a single non-groupable isolate.

The phylogeny of Cluster 27 again does not lend itself to clear-cut interpretation. Four isolates from 1970 and one from 1978 form a series of outgroup clades to the rest of the cluster, which is made up of one main clade consisting of two sister clades, and then a series of three smaller clades. The date of the most recent common ancestor of the entire cluster is estimated to be in 1967.23 (CI: 1966.55-1967.95) suggesting that the outgroups close to the root of the phylogeny represent historic diversity which has now likely gone extinct, as has been the case with other clusters. Perhaps the phylogeny of Cluster 27 reflects decades of relatively consistent demographics,

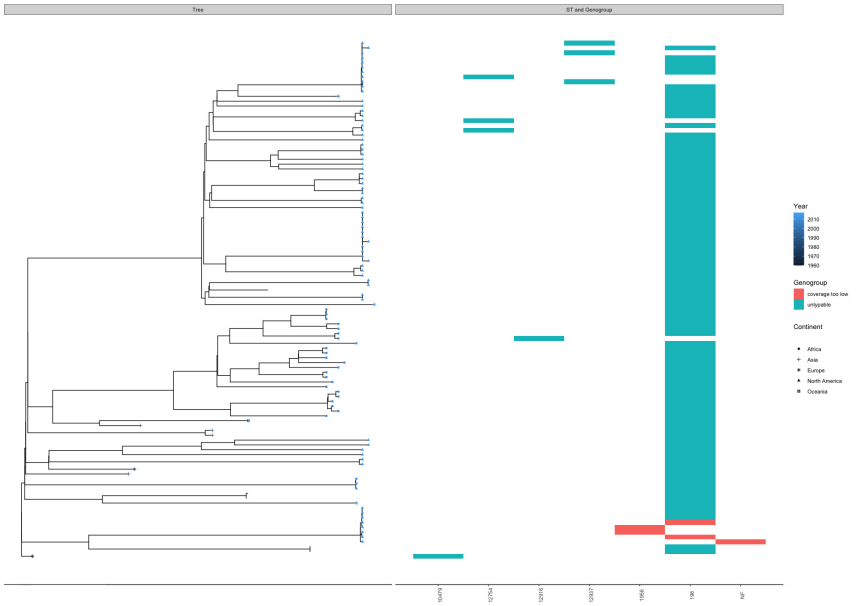


Figure 3.30: Whole-genome phylogeny of Cluster 29 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/9Kvyc65PooZdKXdwQhU16Z`

without any major clonal expansions or bottlenecks since the node separating the outgroups from the rest of the phylogeny. Given its relatively small sample size – even within the individual datasets before they were amalgamated – lends some credence to this suggestion.

Although Cluster 29 is similar to Clusters 22, 25, and 27 in that it is primarily made up of isolates from North America, and specifically the US (66 isolates), Cluster 29 is unlike them in that it also contains a substantial number of isolates which were collected in Africa (26 isolates), and relatively fewer isolates from Europe (5), as well as four isolates collected in Asia, and one collected in Oceania, in New Caledonia. The African isolates originate primarily from Burkina Faso, where 22 isolates belonging to this cluster were collected, though four isolates belonging to this cluster were also collected in Ethiopia. The five isolates collected in Europe are split between the UK and Norway, and the four Asian isolates were collected in Japan (2), China, and Singapore. Cluster 29 is also unique in terms of the range of its isolates' year of isolation, which fall (inclusively) between 1960 and 2017. The relatively smaller size of the cluster means

that the sampling within that interval is relatively sporadic, though it is relatively continuous from 2009 to 2017. Like the other clusters primarily collected in North America, Cluster 29 is predominantly composed of carriage isolates, though there are four isolates collected from cases of invasive disease belonging to the cluster, two collected in Norway, one in China, and one in New Caledonia.

Cluster 29 is one of the less diverse clusters both in terms of sequence type and serogroup diversity. Only six different sequence types have been identified among Cluster 29's isolates, and almost ninety percent (89.72%) are a single sequence type, ST-198, with the remaining sequence five types being present at low frequency. This is a surprising result given what is known about the ST-198 clonal complex, which in publicly available MLST data has ST-823 as its most common sequence type, which within this collection is all found in Cluster 22. 100% of the isolates in this cluster were non-groupable, very strong evidence that the entire cluster is capsule null. Cluster 29 is interestingly the only capsule null cluster to not include a single isolates with a capsule identified *in silico*, though this may simply reflect its relatively smaller sampled size compared to the other 3 non-groupable clusters.

The date of the most recent common ancestor of Cluster 29 is estimated to be 1958.21 (CI: 1957.02 -1959.12). The immediate outgroup of its phylogeny dates to 1960, Norway, which when combined with the date of the most recent common ancestor suggests that what we have sampled from the cluster may be a simple example of a clonal expansion. The clades between one and three nodes from the root, however, paint a very different picture. These clades are separated relatively deeply, and are overlapping in their temporal and geographical distributions, confirming that these lineages were present in the same place at the same time, some still to the present day. The most parsimonious interpretation is perhaps that this lineage diversified in the course of its global spread, without ever having a significant increase in its population size. This has led to a population which consists of several lineages, none of which are deeply divergent, but are nonetheless in sister clades to other

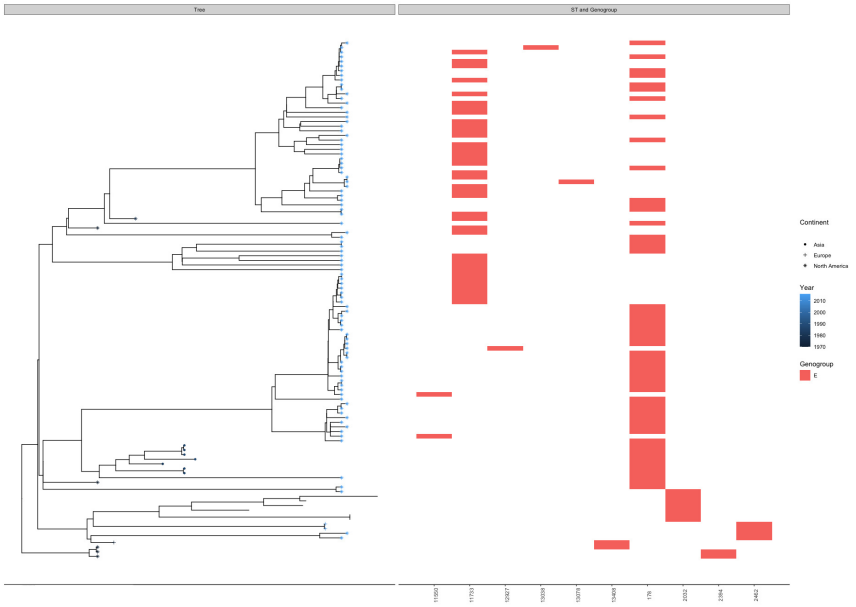


Figure 3.31: Whole-genome phylogeny of Cluster 36 annotated with continent and year of collection, plotted alongside the Sequence Type and Genogroup of each sample. An interactive and extended version of this figure is available at the following uniform resource locator: `https://microreact.org/project/w8SpTS7YpBZSvMvkyAm7cT`

isolates from this cluster which may be in the same place at the same time.

Like Clusters 22-29, Cluster 36 is one of the clusters whose isolates were predominantly collected in North America. 96 isolates belonging to Cluster 36 were collected from North America, 95 from the US, and one from Canada. Only ten isolates belonging to Cluster 36 were collected from outside of North America, seven in Asia, collected in Japan, and three in Europe, two in the UK and one in Germany. Despite this relatively focused sampling in terms of geography, the isolates from Cluster 36 have a fair spread in their year of collection, ranging from 1970 to 2016. As with all the clusters whose isolates were predominantly collected in the United States, most of the isolates were collected in 2015 and 2016, the years with the most sampling in North America. Apart from those two years, there is no other continuous group of years over which isolates belonging to Cluster 36 were collected. Finally, and similarly to Cluster 22-29, the isolates from Cluster 36 are again mostly collected from healthy carriage infections, though like them, Cluster 36 does also contain a number of isolates collected from cases of

invasive disease, eight in total, six collected in Japan, one in Germany, and one in the US.

The distribution of sequence types identified in Cluster 36 is unusual as it does not have a single sequence type which includes most of the isolates present in the cluster and several other sequence types at low frequency. Instead, the most prevalent sequence type, ST-178, accounts for just under half, 49.17%, of isolates, and another sequence type, ST-11733 accounts for most of the remainder, 33.04% of the total. There are also an additional eight sequence types present, though these are at very low frequency. Genogroup diversity is non-existent, Cluster 36 is one of six clusters with a single genogroup identified among its isolates, which have all been typed genogroup E.

The phylogeny of Cluster 36 is made up of an outgroup lineage of three isolates collected in the United States in 1970, and then it bisects into two main lineages, unequal in size, but both spanning a significant time interval. The larger, which itself bisects, is found in the US and Japan, and contains isolates from 1970-2016, while the smaller, found in the US and Europe, contains isolates from between 1973 and 2016. The most recent common ancestor of the entire lineage is estimated to be in 1956.01 (CI: 1951.18-1959.06), fairly shortly before the oldest isolates in the cluster. It seems that in Cluster 36, two different lineages from around the time of the most recent common ancestor have persisted, while the third has gone extinct, and uniquely, there are samples dating back to relatively close to the common ancestor in both of the lineages which have persisted.

3.3 Concluding remarks

From this in-depth analysis of the internal population structure of the 25 major lineages – summarised in Table 3.1, and Figures 3.32 and 3.33, a few clear trends emerge regarding the patterns of descent in *N. meningitidis* generally. First, with regard to changes in genogroup, and hence capsule switching, we see that this is generally a very rare event. Most (19 of 25) clusters contain some evidence of a change in genogroup occurring at least once somewhere in the phylogeny, but its occurrence in the 10-50

Cluster	Dominant Genogroup	Dominant ST	Primary Continent	Disease	Est. Date of MRCA
1	W	ST-11	Multiple	Mixed	1934.47 (1930.6–1937.4)
2	X	ST-181	Africa	Carriage	1967.91 (1966.85–1968.81)
3	B	ST-41	Europe	Disease	1783.1 (1702.93–1826.79)
4	NG	ST-192	Africa	Carriage	1666.84 (1592.69–1708.85)
5	Y	ST-4375	Africa	Carriage	1967.12 (1966.75–1967.41)
6	B	ST-32, 33, 34	Europe	Disease	1902.63 (1880.65–1915.71)
7	Y	ST-1655	Europe	Disease	1970 (1969.26–1970)
8	A	ST-2859	Africa	Mixed	1802.33 (1745.45–1841.55)
9	Y	ST-2881	Africa	Carriage	1975.94 (1975.04–1976.61)
10	C	ST-11	North America, Europe	Disease	1958.66 (1958.41–1958.87)
11	B	ST-1161, ST-275, ST-1163	Europe	Disease	1976 (1975.04–1976)
12	A	ST-1	Africa	Disease	1891.87 (1876.92–1904.56)
13	B	ST-213	Europe	Disease	1970.78 (1963.56–1975.73)
14	B	ST-8, ST-153, ST-2174	Europe	Disease	1950.92 (1947.66–1953.15)
15	B	ST-269	Europe	Disease	1320.47 (1096.37–1500.72)
16	Y	ST-767	Africa	Carriage	1994.05 (1992.19–1995.31)
17	NG	ST-53	North America	Carriage	1941.01 (1922.83–1949.99)
18	W	ST-184	Europe	Disease	1966.43 (1964.84–1967.53)
19	E	ST-1157	North America	Carriage	1818.82 (1756.07–1889.79)
20	Y	ST-1466	Europe	Disease	1990.42 (1987.14–1993.18)
22	NG	ST-823	North America	Carriage	1564.32 (1282.97–1663.86)
25	B	ST-35	North America	Carriage	1988.95 (1987.74–1989.93)
27	E	ST-60	North America	Carriage	1967.23 (1966.55–1967.95)
29	NG	ST-198	North America	Carriage	1958.21 (1957.02–1959.12)
36	E	ST-178	North America	Carriage	1956.01 (1951.18–1959.06)

Table 3.1: Summary of characteristics of the 25 major lineages present in this collection. Dominant Genogroup and ST were chosen based on the largest genogroup and ST present, and multiple STs are reported for lineages with several STs at high frequency ($> 10\%$). Primary continent refers to the most frequent continent present in that collection, whereas Disease is reported as "Carriage" or "Disease" if the proportion of isolates collected from one or the other disease state exceeded 60% of all the isolates in that lineage.

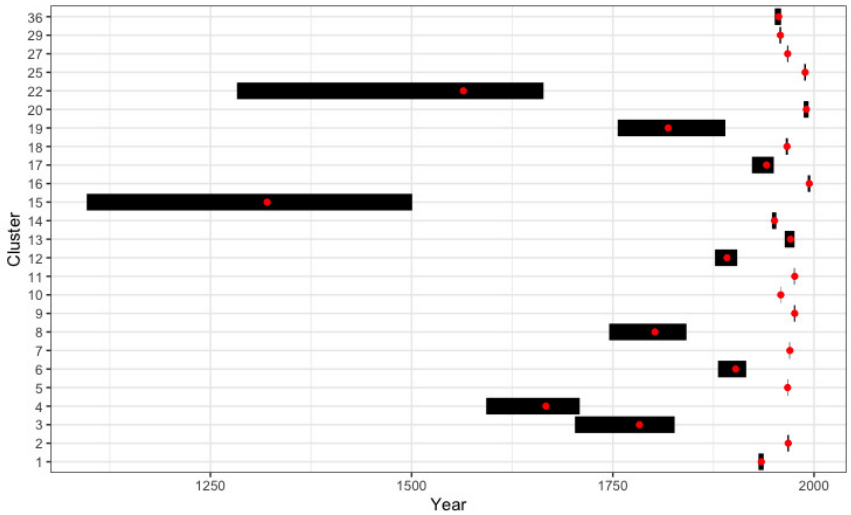


Figure 3.32: Dates of the most recent common ancestors for all 25 major lineages. Date estimates are indicated in red, with black bars indicating the 95% confidence interval of that estimate.

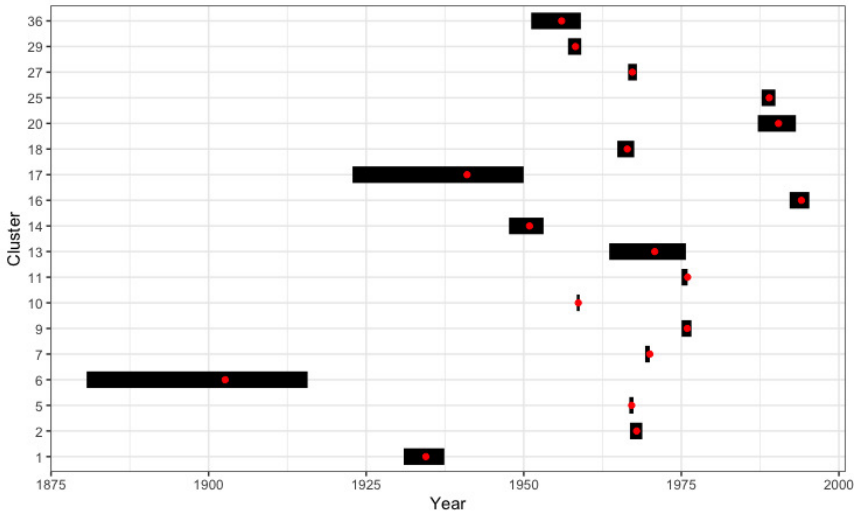


Figure 3.33: Dates of the most recent common ancestors for the lineages among the 25 major lineages with a most recent common ancestor within the 20^{th} century. Date estimates are indicated in red, with black bars indicating the 95% confidence interval of that estimate

year sampling span of most clusters did not exceed the second (10^1) order of magnitude, and typically only occurs a handful of times per cluster, with clusters containing relatively high proportions of different serogroups being primarily driven by the successful growth of a lineage after a single capsule-switching event. In contrast, sequence type switching is extremely common within lineages, with a common set of switches in sequence type occurring in both directions up to dozens of times. For sequence types, the frequency of switches and the amount of diversity within a given cluster also differ radically between clusters, with Cluster 3, for instance, having over a hundred different sequence types while Cluster 5 has only 18. This is often resolved by collapsing sequence types into clonal complexes, but that approach is fraught with further complications, as it can potential over-collapse isolates which are relatively divergent, as seen with Clusters 22 and 29, among others. These results clearly show that while sequence typing remains an unparalleled technique in terms of its speed and simplicity in understanding the local or regional population of *N. meningitidis*, like in the Burkina Faso dataset where there are often only a small number of deeply divergent lineages present and are therefore well-resolved with sequence types and clonal complexes, on a global scale using

sequence type to identify and classify lineages of *N. meningitidis* is likely to lead to the under-clustering of fairly divergent lineages (like Cluster 1 and Cluster 10) while also potentially causing over-clustering of relatively closely related isolates on different continents. In particular, understanding the biological properties and evolutionary history of *N. meningitidis* would be impossible using only clustering based on sequence types. This is a particularly salient point as the within-cluster population structures also strongly point toward distinct populations in different regions of the globe, where although the same cluster is present in many different regions, its prevalence and propensity for causing disease in different regions may differ strongly. The very existence of major serogroup E lineages within this collection (Clusters 19, 27, and 36) comes as a surprise, but this may be primarily be due to the extent to which the study of *N. meningitidis* has been focused on regions of the globe where prominent lineages of serogroup E *N. meningitidis* occur only at very low frequency. Clusters which are not even considered here due to their size may be dominant in regions not yet sampled. In the South American and Asian isolates from this collection, for instance, the five most frequent cluster are 6, 14, 41, 40 and 1, and 1, 69, 5, 54 and 17. The sample sizes from South American and Asia are much too small to conclusively determine if these lineages truly are the most prevalent in those regions, but given the differences between Europe, Africa, and North America, in terms of dominant lineages, it would not be surprising to find lineages which are present at a very low frequency elsewhere in the world could be present at much higher frequency in those areas of the world in which there is currently very little sampling of whole-genome sequenced isolates available.

Perhaps the most surprising and consistent pattern from the examination of the 'within-cluster' population structure through their phylogenies is the strong evidence of the existence of complex population structures within these major lineages, despite most models of bacterial population structure [55, 56] and the term 'clonal complexes' suggesting that they should all be clonal. Though not every cluster showed evidence of deeply divergent population structure in the shape of its phylogeny,

deeply branching sister lineages spanning multiple decades and different continents appeared in the majority of the clusters' phylogenies. While this is in some ways perhaps unsurprising given the pattern of major lineages coupled with variation at low frequency at the whole collection level, this parallelism does suggest that the properties which govern the emergence of new lineages are in general the same as those which govern the population structure that develops within lineages, which does not seem to be necessarily true *a priori*. The deep divergences within some of these clusters goes as far as to suggest how new lineages might indeed be formed; this is particularly true in the case of Cluster 4. It is more or less perfectly split into two halves of roughly equal size, both of which span the sampling interval and have at least a somewhat diffuse geographical distribution. The most recent ancestor is estimated to be in distant past, far beyond what we could reasonably expect to infer with any confidence, and indeed the enormous 95% confidence interval reflects this. By all appearances, this seems to be a cluster where the extant lineages within the cluster can only get more divergent, eventually presumably surpassing the threshold of divergence in both their core and accessory genomes used to distinguish and separate clusters, if neither lineage goes extinct. Extinction, of course, is the other surprising result from the within-cluster phylogenies. It can actually be seen as surprising, in some sense, that 18 of the 25 major clusters have a most recent common ancestor in the 20^{th} century. If it is possible for some lineages to persist for at least over a hundred years, if not much longer, why are most of the 25 major clusters showing signs of a relatively recent common ancestor? The phylogenies of many of these lineages with a recent common ancestor provides some clue, in the form of short branches near the root which are outgroups to the entire collection. These branches must reflect a sample from the former diversity present in the population which has since gone extinct. Based on the date of the most recent common ancestor of the major lineages in our collection of *N. meningitidis*, these population bottlenecks have been enormously important in the evolution of the current global population of *N. meningitidis*. This does not preclude

long-term stability in the global population of *N. meningitidis*, as the 20th century is in many ways a unique in the extent of global movement, but if stability in the global population does exist, it seem that it must exist at the level of major clusters and be principally based on whether or not clusters are able to survive severe population bottlenecks, which seems likely to be primarily driven by conditions in the various regions to which a Cluster has dispersed. The evolutionary effects of population bottlenecks in bacteria – and in particular how they might affect a structured population – are still very much being explored, but at the very least, their prevalence in this collection of *N. meningitidis* suggests a much more important role for both competition leading to selection between lineages and genetic drift in the evolution of the species than is currently generally presumed [55, 73].

That there should be significant population structure in a clonally reproducing organism in particularity *N. meningitidis*, is not a novel or groundbreaking observation. That said, however, such a profoundly deeply separated population structure is not necessarily our expectation in the case of *N. meningitidis* due to its relative propensity for homologous recombination between different isolates of not only its own species, but also other *Neisseria*. How is the population structure related to recombination in *N. meningitidis*?

CHAPTER 4

Recombination and Selection in *Neisseria meningitidis*

This chapter aims to explore the following questions regarding the global population of *N. meningitidis* as represented in the collection studied in this thesis:

1. Which loci are recombining and under selection in the main lineages of *N. meningitidis*?
2. How do these compare between different lineages, in terms of specific loci as well as general trends across all loci?
3. Are there any associations between recombination, selection, and genetics which provide further insight into how these two forces shape the evolution of the species?

To answer these questions I primarily use two well-established methods (outlined in sections 2.2.2.2 and 2.2.2.3) for detecting single nucleotide variants which show evidence of having recently been under selection or recombinant.

The population structure of *N. meningitidis* at a global scale sketches a portrait of a bacterial species which, even over long timescales, does have a core component which is primarily tree-like in its evolution. It is possible to distinguish different monophyletic lineages of the bacteria, though there is good evidence for major changes in the relative frequencies of these

lineages within the global species' population. However, our knowledge from laboratory studies [47] and smaller scale genomic surveys [74, 145] informs us as to the numerous adaptations in *N. meningitidis* which allow it to be naturally competent is the laboratory and routinely recombine in the wild. Though this recombination is not frequent enough or significant enough in scale to cause large-scale disruption to the overall population structure, given how frequently it occurs and the existence of genetic mechanisms to facilitate it, recombination must play a significant role in the evolution of *N. meningitidis*. This collection of whole-genome sequenced *N. meningitidis* will allow us to explore how recombination proceeds across the entire species in unprecedented depth, allowing us to develop our understanding of the importance of recombination to the evolution of the species.

4.1 Differences in recombination between the main lineages

The first stage of studying recombination in our collection of ËĘ*N. meningitidis* is to determine the locations of recombinant loci in each lineage. Gubbins was used to perform this analysis; the initial output provides the locations of each stretch of recombinant DNA within each lineage's whole-genome pseudoalignment and the set of nodes in its phylogeny where that recombination is found. Gubbins then processes these results, along with the initial input alignment, to provide various summary statistics for each node in the phylogeny. One of these summary statistics is ρ/θ, or the relative rate of recombination scaled to the mutation rate, θ. Though this is simple to estimate for each node – by dividing the number of recombination events, ρ, on the branch leading to the node by the number of mutation events on that branch, θ – it provides a good estimate of the recombination rate due to the fact that it is scaled to the mutation rate. Although mutation rate very likely does not remain constant across the diversity of *N. meningitidis*, the true intrinsic rate of mutation across the species should be invariant enough, on average and relative to recombination rate, to pro-

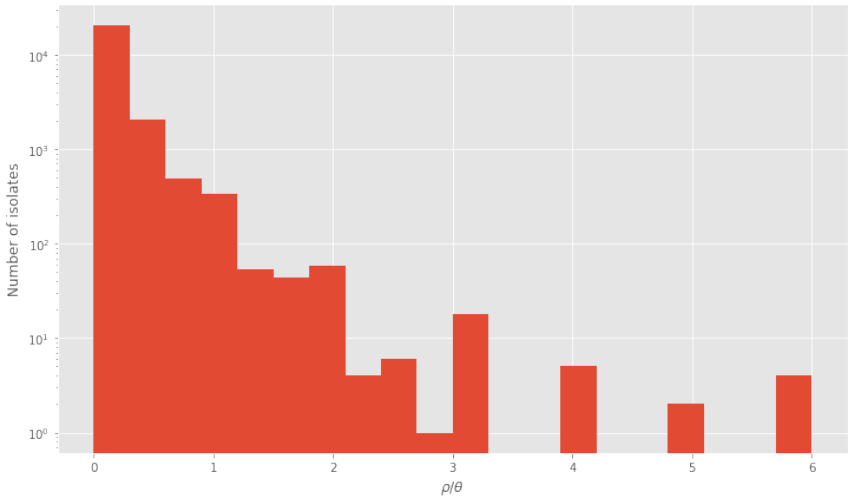


Figure 4.1: Log-scaled histograph of the recombination rates of all samples in the 25 major lineages, as well as ancestral nodes within those lineages.

vide an accurate enough scale for estimating the recombination rate.

When we collate the results for all the clusters, we find that there is an enormous diversity of relative recombination rates within the species. Their distribution is shown in Figure 4.1, a log-scaled histogram of the distribution of recombination rates for all nodes in the phylogenies of all 25 main lineages. This clearly shows that in general, recombination rates follow an exponential distribution, where the vast majority of branches across all the phylogenies have little to no recombination relative to mutation. However, there are a small number of extremely recombinant branches, with recombination rates that are several times that of the mutation rate. This is consistent with the view that in general, the principal determinant of recombination rate in *N. meningitidis* is stochastic opportunity, as we would expect to see more structure in the distribution if this were not the case. This is perhaps unsurprising given its life history means that chance of any given *N. meningitidis* bacterium encountering another bacterium with a similar enough genome to allow for homologous recombination – but sufficiently divergent to make that recombination events detectable in our analysis – is generally quite low. Our *a priori* view may therefore be that this more or less restricts potential donor bacterium to those in other major lineages of *N. meningitidis*, and possibly from

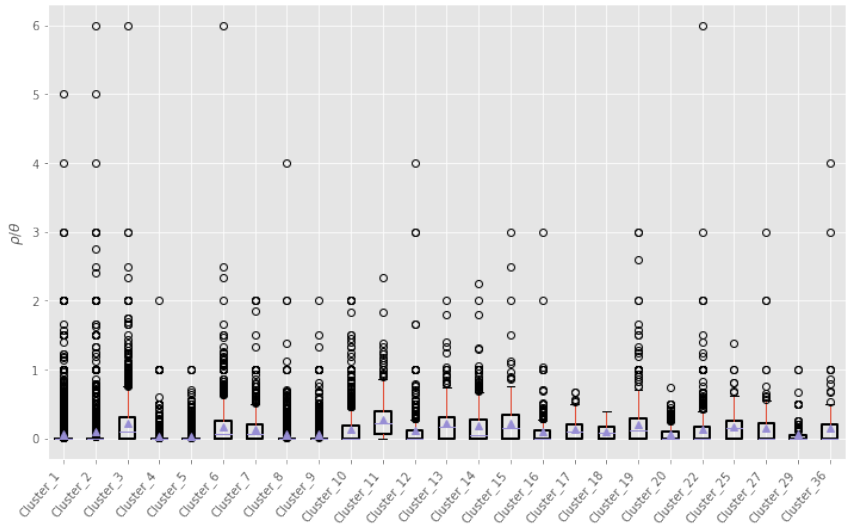


Figure 4.2: Boxplots of the distribution of the recombination rates in the 25 major clusters. The boxplot boxes indicate indicate the first quartlie, median, and third quartile of the set of recombination rates in each cluster. The mean recombination rates for each lineages are indicated with the purple triangles. The whiskers of the boxplots indicate 1.5× the inter-quartile range above and below the third and first quartiles, respectively. Outlier points beyond those whiskers are plotted with circles.

other *Neisseria* species, though laboratory evidence suggests mechanisms exists which limit the gene flow between different species of *Neisseria* [146]. The population structure observed in Chapter 3, where very few lineages dominate the population and minor lineages exist at low frequency, is also consistent with spasticity being the primary determinant of recombination rate.

Though the primary determinant of recombination in *N. meningitidis* seems to be stochastic opportunity, there are many reasons to still think that genetic factors may play a significant role in determining the likelihood of homologous recombination events occurring for any specific bacterium. This is in particular due to the known mechanisms of DNA uptake and recombination in *N. meningitidis*, where several molecular structures are involved in the detection, uptake, and integration of exogenous DNA into *N. meningitidis* chromosomes from their extracellular environment [47]. If genetic factors do affect differences in the rate of recombination between lineages of *N. meningitidis*, we would expect to be able to detect differences in different lineages of the bacteria. Figure 4.2 shows the distributions of ρ/θ recom-

bination rates for the 25 major lineages in the collection. It is not immediately obvious if the lineages are significantly different from one another due to the fact that, like the distribution of ρ/θ for all of the lineages combined, the distributions within each cluster are exponentially decreasing away from the origin – again consistent with the view that recombination in *N. meningitidis* in general is primarily determined by stochastic opportunity, and that this is true across lineages as they all have the same primary distribution, but making Figure 4.2 uninformative on its own with regard to the existence of differences in the parameters of the distribution between lineages.

Applying some statistical testing, in particular the non-parametric Kruskal–Wallis H test[147], allows us to determine the likelihood of whether the estimated recombination rates within each cluster are drawn from populations of the same distribution or not, in short, whether or not the clusters are inherently different in terms of their recombination rates. Computing the test statistic for the recombination rates of the 25 main lineages using the stats module of SciPy[122] returned a value of 4129.647. The p value associated with this test statistic is 0.0, which due to hardware limitations functionally means that $p < 2.2 \times 10^{-16}$. In short, the probability that the recombination rates for the 25 main lineages are drawn from the same distribution is extremely close to 0, we can conclude that at least one of the main lineages has a recombination rate distribution which is significantly different from the others.

Determining which lineages have significantly different recombination rates from one another is not an altogether straightforward process, complicated by the fact that pairwise statistical differences are not necessarily transitive, and sample x may be significantly different from sample y, which is significantly different from sample z, even though samples x and z cannot be shown to be arising from different distributions at high (> 0.95) probability. Nonetheless, performing these comparisons – in this case Dunn's *post hoc* pairwise comparisons[148] with Holm-Bonferroni correction – can still be informative as to the general patters of differences between the groups being compared. This is shown plotted in Figure 4.3, and to aid with its interpretation,

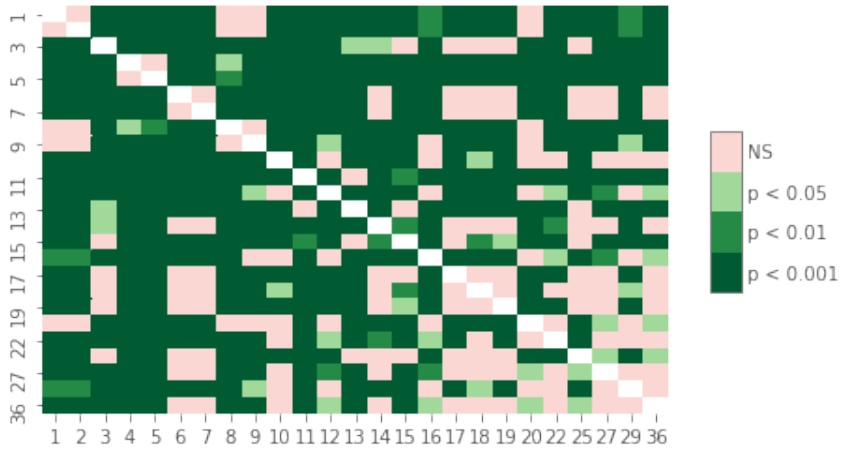


Figure 4.3: Heatmap of the significance results, with p values as indicated in the legend, of *post hoc* Dunn's pairwise comparisons tests for between-groups between the recombination rates of the 25 major lineages present in this collection.

the mean ρ/θ for each isolate is presented in Figure 4.4.

The figures show that generally, recombination rates are significantly different between lineages. Particularly for clusters 1-15, almost all lineages are significantly different from one another. Clusters beyond 15 tend to have smaller sample sizes, and therefore may be falling below the multiple-testing corrected threshold for rejecting the null hypothesis due to a lack of statistical power. The average ρ/θ values of Figure 4.4 do give some credence to this notion, as there are some clusters, for instance 27 and 29, and 29 and 36, which are as substantially different from one another as many of the clusters which are significantly different, but do not meet the lowest threshold for statistical significance ($p < 0.05$) after correcting for multiple testing. Fewer than half – eleven – of the clusters are below the average ρ/θ for the entire collection, which is a ρ/θ value of 0.117, or approximately 8.55 mutation events for every recombination event. This is likely due to some of the clusters with the lowest recombination rates, specifically Clusters 4, 5, 8, and 9 being some of the larger clusters and Clusters 4 and 5 having very low recombination rates. The average rate itself is a somewhat surprising estimate, given that the only published estimates of recombination rate relative to mutation rate in *N. meningitidis* estimated that it was on average around 0.8, with lineages ranging between 0.6357 and 0.9371, much higher than the estimates

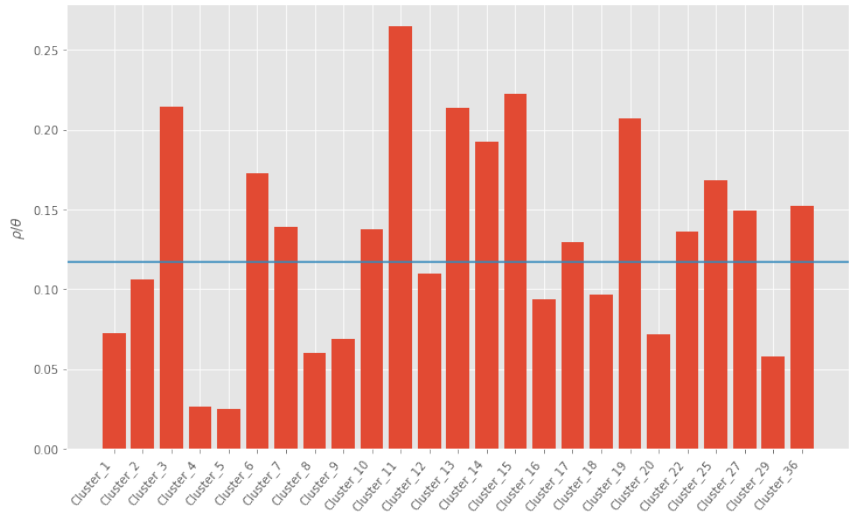


Figure 4.4: Bar chart of the average ρ/θ recombination rates for the 25 major lineages in this collection. The mean for the overall collection is represented by the blue line

here [74]. Those relative rates were estimated without the use of whole-genome data, however, and was therefore restricted to considering several small fragments of the *N. meningitidis* genome. Given the stochastic nature of the physical events which must take place for two *N. meningitidis* bacteria to exchange DNA, an estimate of average recombination rate around $^1/_{10}{}^{\text{th}}$ of the mutation rate seems more on the order of magnitude at which recombination is likely to occur.

As further interpretation of the pairwise significance difference matrix (Figure 4.3) by eye is largely impossible, it is useful to use algorithmic methods to see if representing that matrix in a different way might allow for further insight. In Figure 4.5, the difference matrix used to generate the heatmap in Figure 4.5 is projected onto 2-dimensional space using multidimensional scaling, where clusters that are found to have a p value of less than 0.05 are classed as having a distance of 1 from another, and clusters where the null hypothesis could not be rejected had a distance value of 0, as they are coloured in Figure 4.3. This allows us to see what the spread of the relative pairwise distances is, and it confirms the relative sense from looking at a heatmap of the matrix that the clusters are in general quite well spread out. The 25 clusters are split into 18 distinct groups in the MDS plot, and therefore the overall patterns of statistically

significant differences between samples represent a fairly differentiated set of samples. Though there are 8 isolates which were overlapping and therefore statistically indistinguishable from at least one other cluster, this does suggest that many different phenotypes of recombination rate are possible, and indeed, it is likely not a discrete phenotype. Figure 4.4 suggests this as well, as though some clusters have similar average rates, there are not an obvious set of levels which the genetic factors which control the rate of recombination can switch between. This is despite the differences in the inherent recombination rates of the clusters, which are difficult to explain outside of genetic factors. On a local scale the relative population size of a lineage may play a substantial role in the likelihood of recombination events occurring, but we see no effect of lineage size relative to the global population. Therefore, it seems that the variation in the inherent rate of recombination is itself continuous. This is perhaps unsurprising given the number of mechanisms in *N. meningitidis* which are believed to play a role in recombination, but it suggests that elasticity in the recombination rate might be even greater than we might believe. Two questions remain, however: What genetic factors control the recombination rate and cause there to be differences between lineages? And more generally, what evolutionary pressures drive shifts in recombination rate?

These two questions are not straightforward to fully answer, and will rely on more data and analysis than has currently been gathered and performed. However, much like with the population structure, the question of how applicable all of the above is in more localised geographic contexts: Are the significant differences in inherent recombination rate observed between lineages at a global level, also present within localised populations? This can be seen to be the case if we simply restrict our analysis to the Burkina Faso dataset. As discussed in Chapters 2 and 3, sections 2.1.1.3 and 3.1, the Burkina Faso carriage collection is one of the large constituent datasets which makes up this collection. It represents a period of extremely intense sampling over a relatively short timeframe, 2839 isolates in 3 locations within Burkina Faso from 2009-2012. The collection is dominated by 9

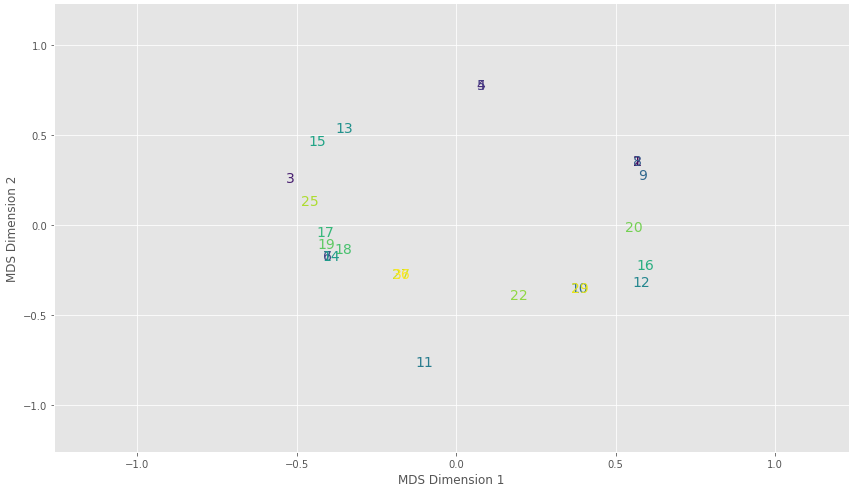


Figure 4.5: Multidimensional scaling plot of the significant differences detected in *post-hoc* Dunn's pairwise comparisons tests – where a distance of 1 represents significantly different, and 0 represents no significant difference – between the recombination rates of the 25 major lineages *N. meningitidis* present in this collection

of the global lineages: Clusters 2, 5, 1, 9, 16, 4, 8, 29, and 34, in order from largest to smallest. Figure 4.6 shows the boxplot of the distribution of recombination rates for only isolates from these lineages within Burkina Faso carriage collection, similar to Figure 4.2 for the entire Global Collection.

With fewer clusters, it is much easier to see that the distributions of recombination rate in the major lineages in 4.7 are substantially different. Statistical testing with a Kruskal–Wallis H test as per the entire global collection confirms that this is significant ($H = 440.977$,$p = 3.17 \times 10^{-90}$). Even with much a restricted sampling temporally and geographically, we are able to detect significant differences in the inherent recombination rate between lineages – further evidence that these differences in recombination rate are genuine and related to genetic factors. Working on a small localised dataset also allows us to explore a further unanswered question regarding recombination in *N. meningitidis*, namely, between what which isolates does the recombination occur? The entire global collections makes this analysis difficult to perform, as its substantial geographical and temporal spread mean that while it might be possible technically to infer recombinations between isolates, those results could not possibly be the result of a recent direct transfer of

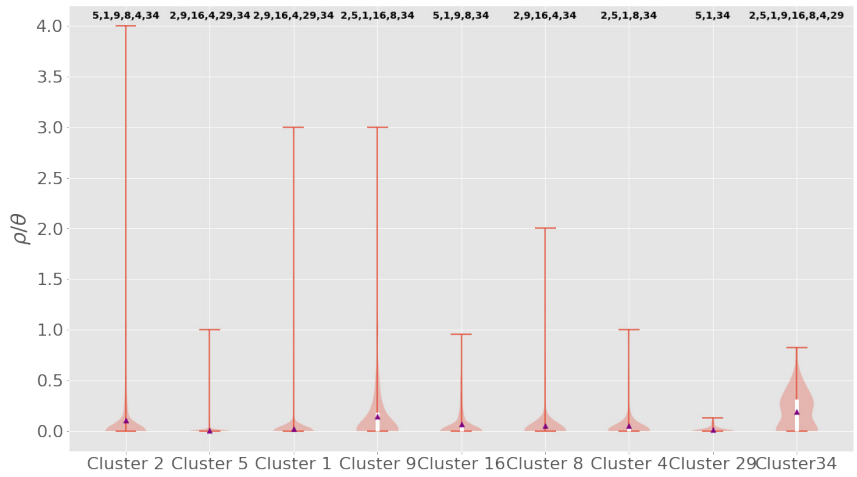


Figure 4.6: Violin plot of the distribution of per-isolate ρ/θ, recombination events per mutation event. The average ρ/θ per cluster is indicated by the purple triangles. The top and bottom of the white boxes indicate the third and first quartiles respectively, and the whiskers of the plot represent the maximum and minimum values. The orange background shading represents the distribution of inferred recombination rates within each cluster. Significant differences between clusters, as determined by a Kruskal-Wallis non-parametric analysis of variance on all the per-branch rates for each cluster followed by *post hoc* statistical testing for differences between groups using DunnâĂŹs test and the conservative Holm-Bonferroni correction for multiple testing, are indicated by cluster numbers above each clustersâĂŹ violin plot.

DNA and would be extremely difficult to interpret. Within the Burkina Faso collection, however, this is not a problem. Therefore, we can conduct an additional analysis on only the isolates from the Burkina Faso carriage collection to attempt to ascertain the sources of DNA in recombination events in *N. meningitidis*. By leveraging pan-genomic methods as described in Section 2.2.2.5, it is possible to align the coding regions of the core and accessory genomes for the entire *N. meningitidis* sample from the Burkina Faso collection. This allows us to then use the fastGEAR [131] (Section 2.2.2.3) method for detecting recombinations within short alignments of coding regions, which as part of its detection of recombinant regions, explicitly infers a source for the recombinant DNA, which could be any other lineage in the collection, or potentially an unknown source. The results from this analysis are shown in Figure 4.7, and the network of the recombination events clearly shows that most recombination events have donor DNA from another isolate of *N. meningitidis*, and our bias for detecting only sufficiently diverged homologous DNA means that, from what we can detect, this DNA particularly originates from isolates within other major lineages. Some recombinant DNA also appears to originate from unknown sources, potentially from other *Neisseria* which were not sampled in the collection.

The chord diagram of the network of recombination events inferred from the Burkina Faso collection also further demonstrates how stochastic encounters between lineages plays a significant role in recombination in *N. meningitidis*, as discussed in relation to Figures 4.1 and 4.2. The lineages present in the Burkina Faso collection with the greatest population size at the time of sampling, Clusters 2, 5, and 1, are involved in the greatest number of recombination events. Figure 4.7 also suggests that the recombination rate of clusters may have a significant effect, however, with Cluster 34 having one of the highest recombination rates in the Burkina Faso collection and being involved in 10 times the number of recombination events as Cluster 29, despite both having a similarly small number of isolates present in the population.

Finally, the metadata present for this collection, although

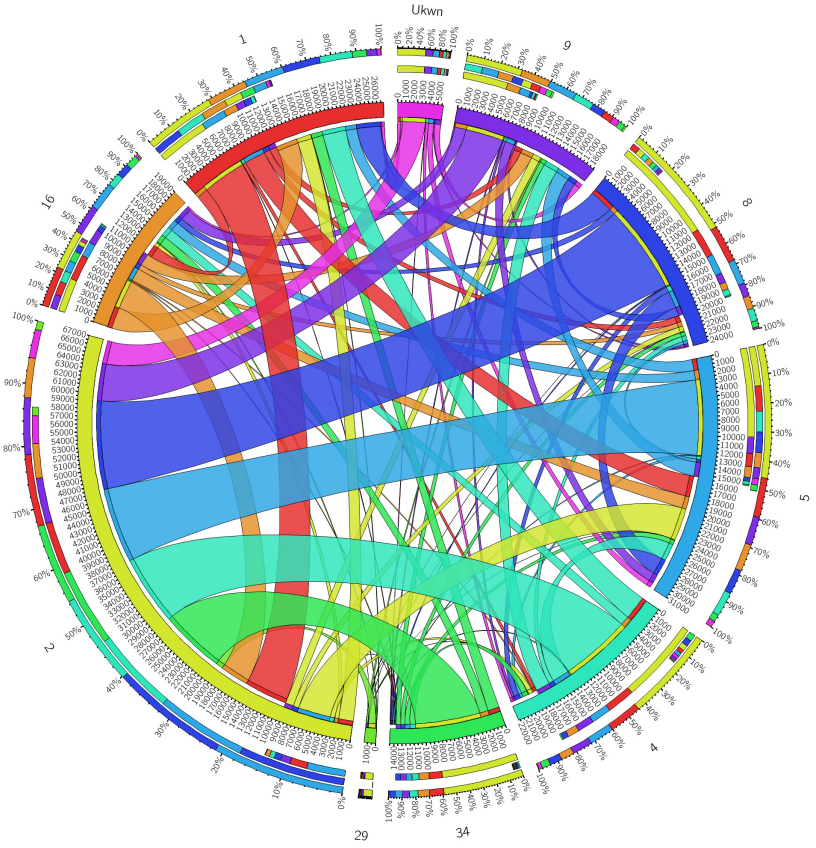


Figure 4.7: Chord diagram of the count of recombinant regions identified by fastGEAR between the clusters in the Burkina Faso *N. meningitidis* carriage collection. Clusters are positioned on the main circle of the diagram, with the arc length of the cluster indicating the number of recombination events. Linkages between clusters represent the number of recombination events occurring between those clusters with their width representing the number of events and the colour indicating the donating lineage. The three stacked bars outside the main diagram indicate, from outermost to innermost, the proportion of the total number of recombinant regions in each cluster coloured by the other cluster involved, those same proportions only for the count of recombinant regions received, and finally those proportions for regions donated, in the focal cluster.

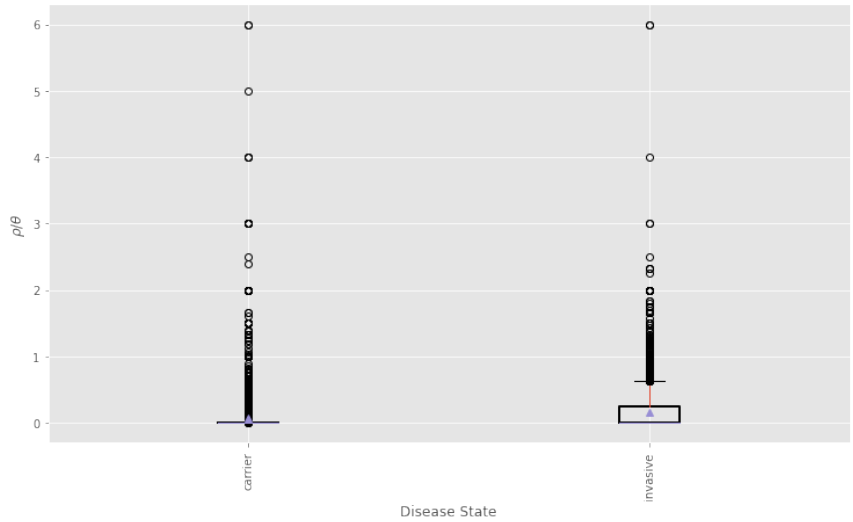


Figure 4.8: Boxplot of ρ/θ recombination rate for all isolates in the 25 major lineages of the global collection, grouped by whether or not they were isolated in carriage or cases of invasive disease

not particularly rich, do allow us to use these results to assess the association between different recombination phenotypes and the likelihood of causing disease. The first question is, is there a significant difference in the measured per-isolate ρ/θ between carriage and disease isolates? Figure 4.8 shows a boxplot of the distributions of recombination rates for carriage isolates versus isolates which were sampled from cases of invasive disease. There is quite clearly a difference in the distributions between the two groups, though interestingly, their ranges are very similar. A Mann–Whitney U test [149] for differences in the distributions of two populations confirms that they are significantly different, ($U = 8494146.5$, $p = 5.675 \times 10^{-136}$). This is not altogether expected, as it is not obvious why disease-causing isolates should have significantly elevated recombination rate. Despite the association of some lineages with being sampled primarily in cases of invasive disease, it is generally believed that most infections do not cause invasive disease, and an isolate which causes an infection resulting in invasive disease is most likely to have been transmitted from a carriage infection. One possible explanation for the significant difference between carriage and disease isolates is that it might be the case that by chance, some clusters with a higher proportion of disease causing isolates also have higher ρ/θ, and therefore are skewing the distribution of

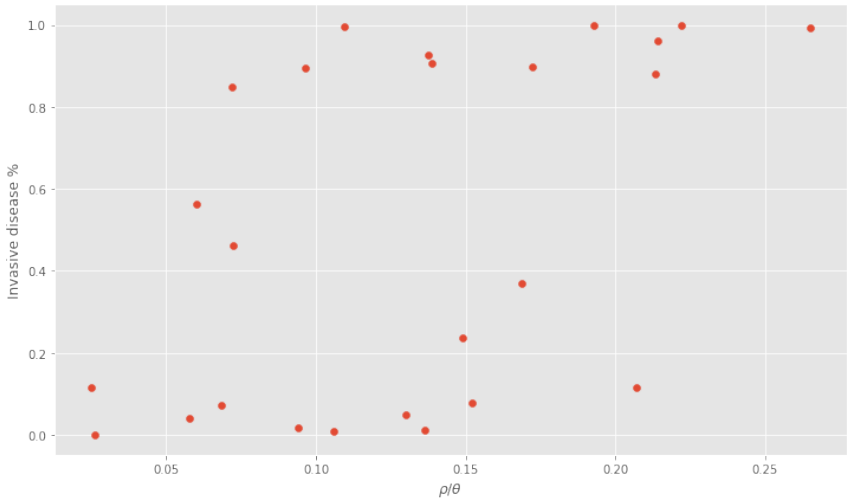


Figure 4.9: Scatter plot of the mean ρ/θ against the proportion of invasive disease isolates within the 25 major lineages of the global collection

disease causing isolates to have distribution of recombination rates which is significantly higher than carriage isolates.

If this were the case, we would expect there to be a very loose correlation between average recombination rate and the proportion of invasive disease isolates within a cluster, and as shown in Figure 4.9, this is to some extent true. The isolates are loosely correlated, (Spearman's [150] $r = 0.562$, $p = 0.003$), though significantly, and as most lineages are heavily biased toward either disease isolates or carriage isolates (with $> 80\%$ of one or the other) it is difficult to accurately ascertain how strongly being sampled from a disease causing infection is actually associated with higher levels of recombination. The ranges of recombination rates for lineages with over 80% recombinant isolates and under 20% recombinant isolates are not completely overlapping, however, suggesting that there may be some association which is not caused by lineage biases. An informative test of this would be to see whether the significant difference in ρ/θ persists within clusters as well as overall. Unfortunately, as most lineages are heavily biased towards either invasive disease isolates or carriage isolates, the relative sample sizes of the two classes of isolates within the lineages would severely hamper the power of any statistical tests to reject the null hypothesis of being sampled from distributions with similar distributions. Four lineages, however, Clusters 1, 8, 25, and 27 have a more inter-

mediate number of carriage and disease isolates, and might have enough statistical power to discern if there is a true difference between carriage and disease isolates. Performing Mann–Whitney U tests on the recombination rates of the carriage and disease isolates within these lineages returns test statistic, U values of 340752.5, 16328.5, 1018.0, and 490.0 respectively, corresponding to p values of 2.318×10^{-36}, 0.00138, 0.0212, and 1.757×10^{-6}. Using the Holm-Bonferroni multiple testing correction [151], we find that all these p values are sufficiently small to reject the null hypotheses that the distributions of recombination rate from which the carriage and disease isolates are drawn are the same.

If disease isolates therefore truly have higher recombination rates, measured as ρ/θ and independent from the cluster of origin, what could be causing this? There are a number of possibilities. First, it could be that an important precursor to causing disease that the recombination rate be elevated. Though it is not entirely clear why this would need to be the case, it is conceivable that an elevated recombination rate could assist the invasive strain in avoiding the immune system during the course of an infection. Another possibility is that disease isolates do not actually have an inherently elevated rate of recombination, but instead are much more likely to contain detected recombination events. It is again not entirely clear why this would be the case, but for a number of reasons, it might be the case that *N. meningitidis* bacteria are more likely to be invasive following a mixed infection, where they encounter other *N. meningitidis* bacteria and give rise to the possibility of recombination that is sufficiently diverse in order to be detected. It is unclear why or how this diversity would result in infections causing invasive disease, but there are a number of possible causes. The uptake and integration of sufficiently diverse DNA itself could lead to an increase in the likelihood of an invasive disease-causing infection, possibly due to content of the newly recombined DNA itself, or some other internal cellular process which is triggered by such a diverse recombination. The final possibility is that it isn't the recombination itself is not what leads to an increased likelihood of causing invasive disease, but the presence two lineages in a mixed infection may lead to competition between

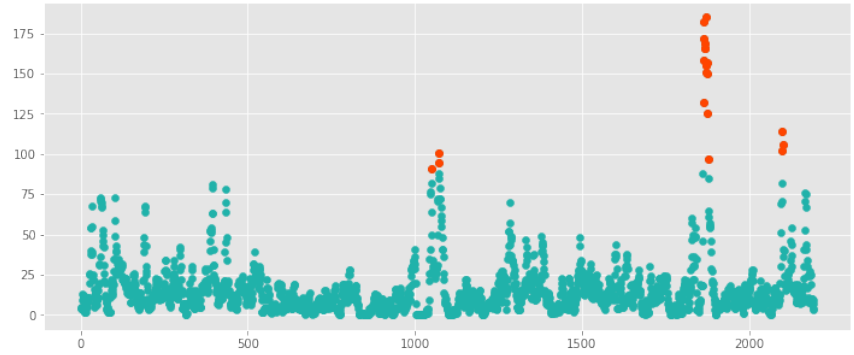


Figure 4.10: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 1. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

the bacteria, or other interaction which then causes an increase in the likelihood of invasive disease in the descendants of those lineages. Unfortunately, these data are not able to test these hypotheses, and indeed genomic data alone is unlikely to be capable of doing so. However, we can conclude that there is an association between isolates sampled from cases of invasive disease and higher recombination rate – an evolutionary genetic characteristic of *N. meningitidis* bacteria, and further research aimed at understanding why this is the case may lead to cracking open our understanding of what causes some *N. meningitidis* infections to result in invasive disease.

4.2 Differences in recombination within the main lineages

Thus far, this chapter has been examining recombination primarily in terms of difference in its rate between different lineages. However, recombination rate does not capture the extent of the diversity in how recombination proceeds as it also differs significantly in frequency between different loci in the genome. Hereafter we will examine how recombination is distributed across the genome in the 25 major lineages in this collection, and also compare how this distribution differs between the genomes of the 25 major lineages.

Cluster 1 (Manhattan plot of recombination events across

the genome shown in Figure 4.10) has, in general, a fairly high level of background recombination across the genome relative to other lineages, despite its relatively low recombination rate. This is likely due to its general diversity, meaning that relative to the mutation rate, the rate of recombination remains lower than the average. Despite its low rate, recombinations are distributed across every portion of the genome, with almost no regions having no recombination – only 93 1000 base-pair windows, or 4.24% of the genome, contain no recombination events. Despite how widespread recombination is across the genome in this lineage, there are still well-formed, identifiable 'peak' regions where recombination is significantly higher than the background. This is likely at least partially caused by the the size of the cluster (2218 isolates) as a proportion of the dataset, which will mean that more events will be picked up by virtue of the greater sample size. Within windows of the genome above the 99$^{\text{th}}$ percentile of recombinations per window, indicated in Figure 4.10 in orange, there were the following named genes: *dnaB*, the DNA helicase responsible for opening the replication fork during chromosomal replication, *yccS*, a transmembrane protein believed to be an efflux pump, *ftsY*, a signal receptor protein which is believed to be a homolgue of *pilA* in *N. gonorrhoeae*, *msrAB*, a surface-exposed methionine sulfoxide reductase, *pncC*, a nicotinamide-nucleotide amidohydrolase involved in pyiridine synthesis, *mscS*, a mechanosensitive cell membrane channel, *thiB*, a periplasmic thiamine-binding protein, *gph*, a phosphoglycolate phosphatase involved in glycolate biosynthesis and therefore DNA repair in response to oxidative stress, *mnmE*, a tRNA-modification GTPase, and *hap*, a secreted adhesion and penetration protein. There were also 7 additional hypothetical proteins with no known annotation or function.

Cluster 2 (Manhattan plot of recombination events across the genome shown in Figure 4.11) has a lower level of background recombination – there are long stretches of the genome which seldom recombine between peaks where significantly more recombination takes place. 37.1% of the 1000 base-pair windows across the reference genome did not contain any recombination events. Peaks are still easily distinguished from the background

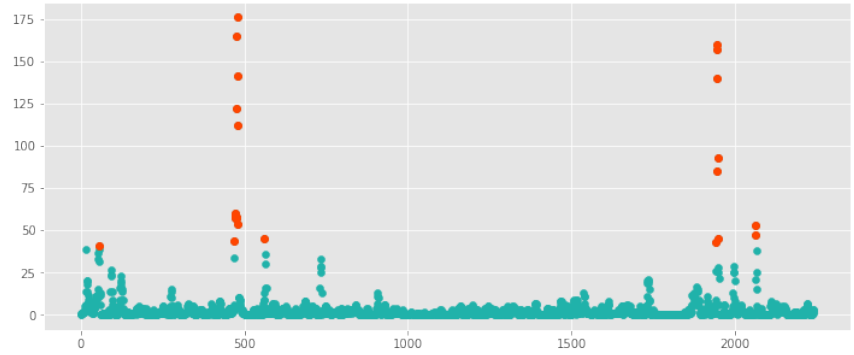


Figure 4.11: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 2. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

level, however, and those windows above the 99th percentile, shown in Figure 4.11, contained the following genes: *yccS*, a transmembrane protein believed to be an efflux pump, *amiC*, a N-acetylmuramyl-l-alanine amidase involved in cell separation and peptidoglycan release, *tsaE*, a tRNA threonylcarbamoyladenosine biosynthesis protein, *racE*, a cell wall biosynthesis protein, *tbpB*, the outer-membrane exposed transferrin-binding protein involved in iron uptake, *folP*, a cytoplasmic protein involved in tetrahydrofolate biosynthesis, *glmM*, a cytoplasmic phosphoglucosamine mutase, *dxs*, a cytoplasmic 1-deoxy-D-xylulose-5-phosphate synthase, *xerC*, a site-specific recombinase, *cbbA*, a fructose-1,6-bisphosphate aldolase, *rimI*, a ribosomal protein, *pyrE*, a phosphoribosyltransferase involved in pyrimidine synthesis, *argA*, an arginine biosynthesis protein, *hemR* a surface-exposed hemin receptor involved in iron uptake, and *yclQ*, an outer membrane transporter believed to be responsible for iron uptake. There were also an additional 5 hypothetical proteins without any known annotation present in 'hotspot' regions above the 99th percentile of recombinations per window.

Cluster 3 (Manhattan plot of recombination events across the genome shown in Figure 4.12) has one of the higher average recombination rates in this collection (Figure 4.4), behind only clusters 11 and 15, and it shows in the distribution of recombination events across its genome. The background level of recombination, though it does occasionally reach 0 recombination events per 1000 base-pair window, tends to hover around

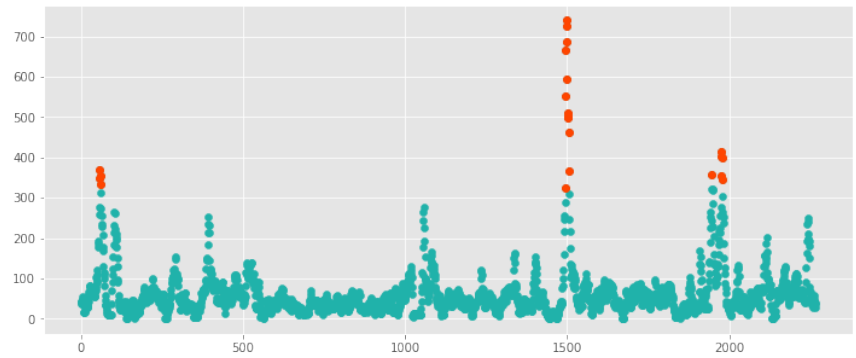


Figure 4.12: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 3. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

50 recombination events per 1000 base-pair window. It also has the most recombinations detected in any 1000 base-pair window across all 25 lineages, at 740 recombinations detected within a 1000 base-pair window. As possibly expected given the extent of recombination across the genome, there are only 7 windows across the entire reference genome without any overlapping recombination events, or 0.031% of all the windows. In the most recombinant windows in this cluster, there are the genes *apbC*, an iron-cluster carrier protein, the outer membrane transferrin-binding protein *tbpB*, *racE*, a protein involved in cell wall biosynthesis, *tsaE*, a tRNA threonylcarbamoyladenosine biosynthesis protein, *amiC*, a N-acetylmuramyl-l-alanine amidase involved which plays a role in cell separation and peptidoglycan release, *rlmL*, a ribosomal RNA large subunit methyltransferase, *acpS*, an acyl-carrier protein involved in fatty acid synthesis, *mscS*, a mechanosensitive cell membrane channel, *thiB*, a thiamine-binding periplasmic protein, *alx*, a membrane-bound redox modulator believed to be a transporter, *murA*, a UDP-N-acetylglucosamine 1-carboxyvinyltransferase involved in peptidogylcan synthesis and hence cell wall biogenesis, *pgk*, a phosphoglycerate kinase involved in glycolysis, *ibaG*, an iron-sulfate metabolism protein, and *ftsX*, a transmembrane protein involved in cell division. There were also an additional 7 hypothetical proteins with no known annotation.

Cluster 4 (Figure 4.13), along with Cluster 5, have the lowest two recombination rates of all the clusters. This is particularly

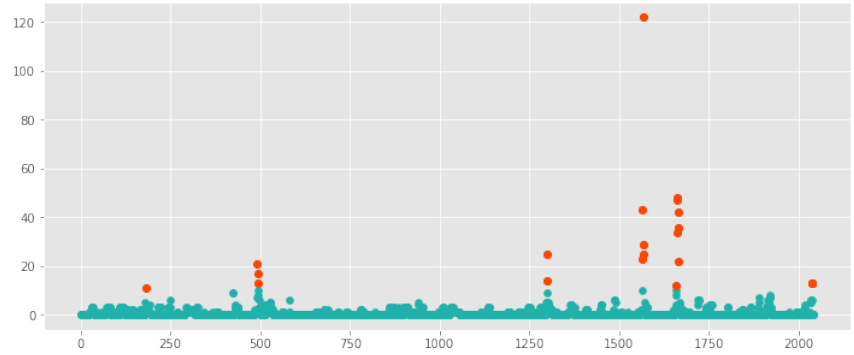


Figure 4.13: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 4. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

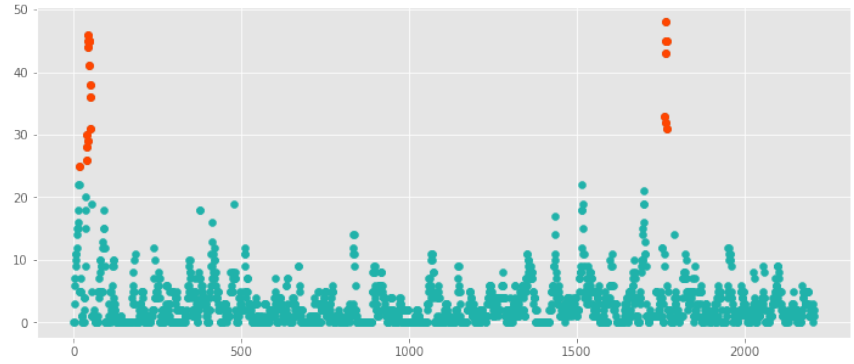


Figure 4.14: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 5. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

evident in the Manhattan plot of recombination events across the genome of Cluster 4, with most of the genome having no recombination or very low levels recombination, apart from a few specific regions with higher levels of recombination. 1506 of all 2043 of the windows across the genome, or 73.7% of all windows have 0 recombination events. Within the peak 'hotspot' regions above the 99$^{\text{th}}$ percentile of recombination, there are 11 genes, including *folP*, *glmM*, *ansA*, *lex1*, *pilE*, *lpxC*, *anmK*, *tbpB*, *tbp1*, *spuE*, *rpsT*, *lbpA*, and *groL*. There were 3 additional hypothetical genes of unknown function.

Like Cluster 4, Cluster 5 (Figure 4.14) has a very low recombination rate, in fact the lowest of the 25 major lineages. Despite this, however, the Manhattan plot of recombinations looks quite different, due to recombination events occurring

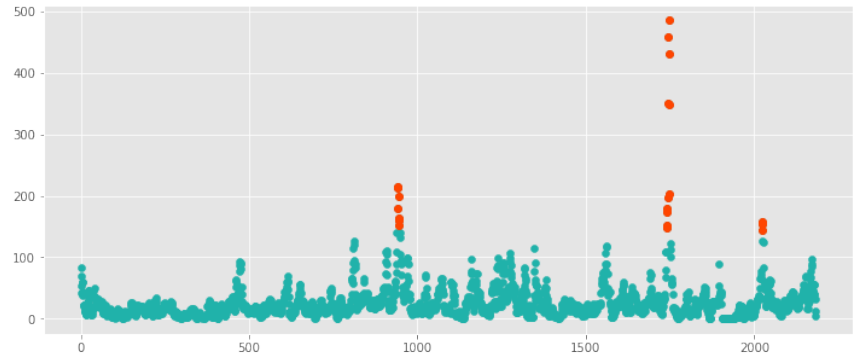


Figure 4.15: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 6. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

consistently across the genome. Unlike cluster 4, where 73.7% of the genome did not have any recombination events, only 653 windows of 2209, or 30% of windows had 0 recombination events. Despite the more consistent spread of recombination events across the genome, there are still two identifiable peaks above the 99$^{\text{th}}$ percentile of the number of recombination events per recombination window. These two peaks contain 16 genes, including *pilE*, the major pilin protein, *ftsY*, a signal recognition peptide involved in targeting membrane proteins to the cell membrane, *yccS*, a membrane protein believed to be an efflux pump, *pilT*, a pillus-associated protein believed to be involved in motility, *yggS*, a protein involved in pyridoxal 5'-phosphate homeostasis, *proC*, a pyrroline-5-carboxylate reductase believed to be involved in amino acid synthesis, *tbp1*, a transferrin-biding protein, *tbpB*, another transferrin-binding protein both of which are involved in iron uptake, *racE*, a protein involved in peptidoglycan and thus cell wall biosynthesis, *tsaE*, a tRNA synthesis protein, and *amiC*, a protein responsible for cell wall hydrolyse and is therefore involved in cell division. There were also an additional 4 hypothetical proteins of unknown function.

Cluster 6 (Figure 4.15) is a substantially more recombinant cluster than either cluster 4 or 5, with a higher rate of recombination than the all-cluster average (Figure 4.4), though it is not one of the most recombinant clusters in the collection. This is reflected in the distribution of recombinations across its genome, which contains only 57 our of 2183 windows that have

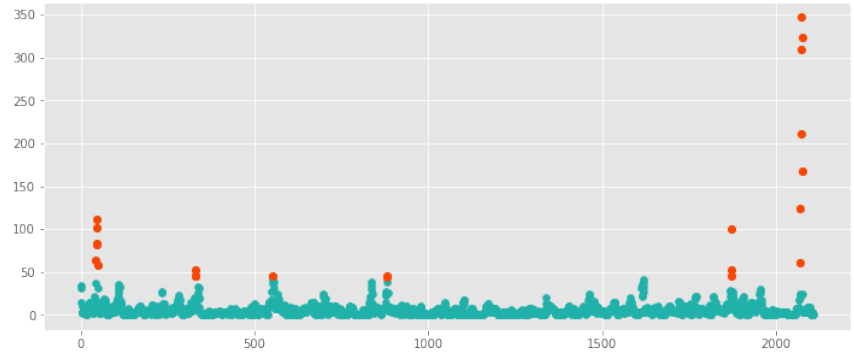


Figure 4.16: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 7. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

no recombination events (2.61%) and many minor peaks where windows have between 50 and 100 recombination events. In the major peaks above the 99th percentile there were 14 genes, which include *apbC*, an iron-cluster carrier protein, *carA*, a protien involved in carbamoyl-phosphate synthase and therefore arginine synthesis, *amiC*, a N-acetylmuramyl-l-alanine amidase involved which plays a role cell separation and peptidoglycan release, *tsaE*, a tRNA threonylcarbamoyladenosine biosynthesis protein, *tbpB*, a transferrin-binding protein *tbp1*, another transferrin-binding protein both of which are involved in iron uptake, and finally *lbpA*, a lactoferrin-binding protein, which is also involved in iron uptake. In addition, there were 4 hypothetical genes with unknown function.

Cluster 7 (Figure 4.16) is close to the average level of recombination and also has recombination events fairly well distributed across the genome, with 9.1% of windows, or 192 out of 2109 containing no recombination events. There are relatively few recombination peaks of significant scale, though the maximum value for a recombination window is 347. The 6 peaks which surpass the 99th percentile of recombinations per window contained the following genes: *yccS*, an inner membrane protein, *pilT*, a protein assosciated with pilin motility, *yggS*, a protein involved in pyridoxal phosphate homeostasis, *proC*, an protein involved in L-proline biosynthesis, *amiC*, a surface-exposed cell wall cleavage protein, *tsaE*, involved in tRNA synthesis, *racE*, responsible for a key step of the peptidoglycan synthesis path-

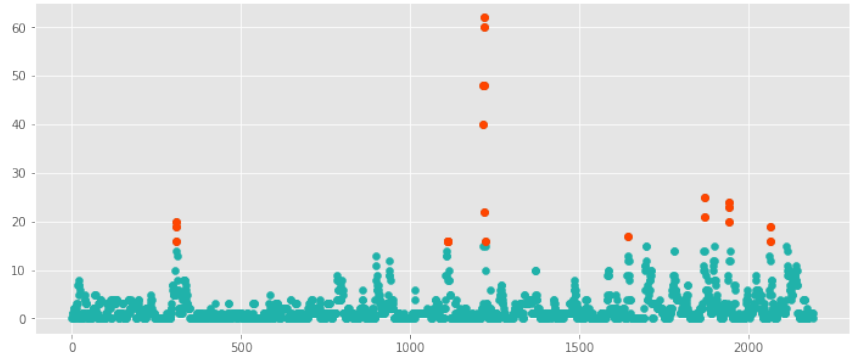


Figure 4.17: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 8. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

way, *lst*, part of the lipooligosaccharide biosynthesis pathway, *lbpA*, a lactoferrin-binding protein involved in iron update, *pilE*, the principal component of the type IV pillus, *lpxC*, an enzyme involved in lipid biosynthesis, and *tbp1* and *tpbB*, two transferrin-binding proteins involved in iron uptake.

Cluster 8 (Figure 4.17) has a recombination rate well below the average recombination rate for all the clusters. Unsurprisingly, it also has a higher proportion of the genome which has no recombination, at 39.4% of all the 1000 base-pair windows across the genome, or 864 out of 2192. Despite the relatively smaller amount of recombination distributed across the genome, there were still several specific loci where recombinations are significantly elevated, and those regions above the 99$^{\text{th}}$ percentile included: *dnaB*, a DNA helicase believed to be involved in chromosomal replication, *btuB*, a active vitamin B12 transporter, *pigA*, a protein involved in iron stress response, *pqiA*, a membrane protein, *tbp1*, a transferrin-binding protein involved in iron uptake, *tpbB*, another transferrin binding protein, *racE*, a protein involved in cell wall synthesis, *tsaE*, a protein which is responsible for a certain step in tRNA synthesis, *yggS*, a pyridoxal phosphate homeostasis *pilT*, a protein believed to be involved in pilin motility, *hemR*, the heme receptor involved in iron update, *glyS*, a tRNA synthase, *lex1*, a lipooligosaccharide biosynthesis protein believed to be involved in membrane synthesis, *arnB*, another protein believed to be involved in lipooligosaccharide biosynthesis, and *dapH*, a protein believed to be involved in

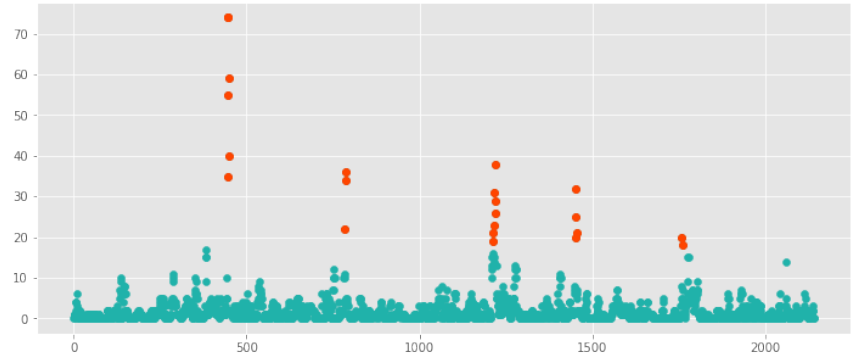


Figure 4.18: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 9. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

lysine biosynthesis.

Cluster 9 (Figure 4.18) is another cluster which has a recombination rate that falls below the average for all clusters. Like Cluster 8, a more substantial proportion of the reference genome for this cluster does not have any overlapping recombinations, at 42.3% or 906 out of 2141 1000 base pair windows. There are 5 well-defined peaks in the Manhattan plot which fall above the 99$^{\text{th}}$ percentile, however, and the following genes are contained within the windows which make up these hotspots: *tbp1*, a transferrin binding protein *tbpB*, another transferrin binding protein which is involved in iron uptake, *thiC*, involved in thiamine synthesis, *dksA*, an RNA polymerase-binding transcription factor, *proC*, an L-proline biosynthesis protein, *pilT*, a protein involved in pilin motility,*yccS*, a transmembrane transporter, *lbpA*, a lactoferrin-binding iron-uptake protein, *groL*, a chaperone protein, *tsaB*, a tRNA synthase, and *rimI*, an ribosomal acetyltransferase.

Cluster 10 (Figure 4.19) is close to the average for recombination rate, but has recombination events well-distributed across the genome. Despite this, 10.9% of the genome, or 246 out of 2252 windows, does and do not have any recombination events contained within. There are several peaks in the Manhattan plot which pass the 99$^{\text{th}}$ percentile of recombinations per 1000 base pair window, and these 'hotspot' regions contained *yhgF*, an mRNA binding protein, *kpsT*, a polysialic acid transporter, *kpsM*, a polysialic acid permease, *siaA*, a UDP-N-

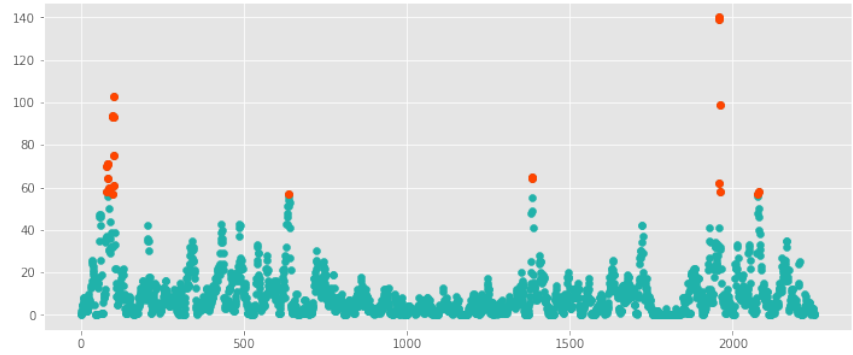


Figure 4.19: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 10. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

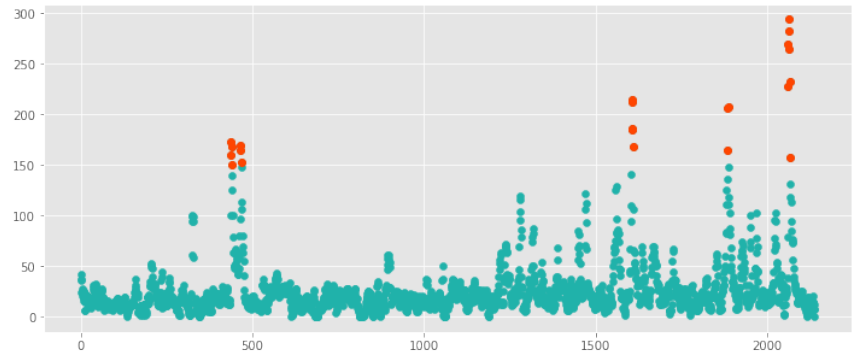


Figure 4.20: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 11. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

acetylglucosamine 2-epimerase, *rfbC*, dTDP-4-dehydrorhamnose 3,5-epimerase. Both epimerases are part of carbohydrate biosynthesis. Other genes include *gltS*, a sodium-dependant glutamate transporter, *ydjA*, an NADH nitroreductase, *piiC*, an outer membrane porin, *cbbA*, an aldolase involved in carbohydrate metabolism, *rimI*, ribosomal protein acetyltransferase, *pyrE*, a gene involved in UMP synthesis, *argA*, an acetyltransferase involved in argenine synthesis, *yclQ*, an iron transporter, and *yclN* a putative ABC transporter with unknown solute.

Cluster 11 (Figure 4.20) has the highest ρ/θ recombination rate out of all 25 major lineages, and the Manhattan plot of recombinations per 1000 base-pair window clearly reflects this, with a sustained high background level of recombination. This is also evident in the number of windows that did not overlap

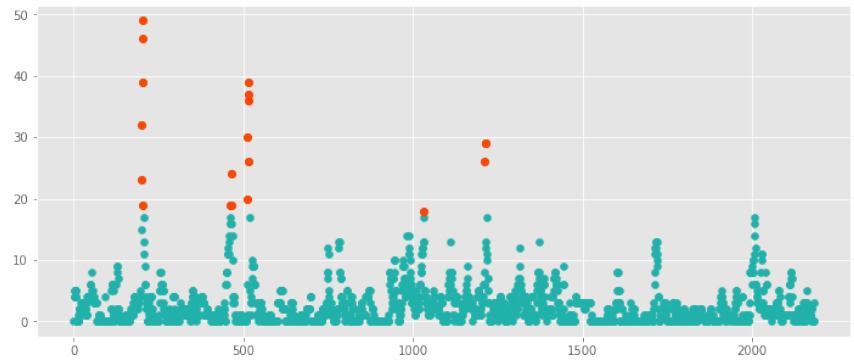


Figure 4.21: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 12. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

with any recombination events, only 0.327% of windows, or 7 out of 2139. The peaks are still well-defined despite the overall high level of recombination in this cluster, and the regions with recombinations above the 99$^{\text{th}}$ percentile included the following genes: *ftsY*, a membrane targeting protein, *msrAB*, an oxidation repair protein, *pilE*, the type IV pilus, *lpxC*, a protein involved in lipid biosynthesis, *tbp1*, a transferring binding protein, *tbpB*, another transferrin-binding protein both of which are involved in iron uptake, and *dnaB*, a DNA helicase believed to be involved in chromosome replication.

Unlike Cluster 11, the recombination rate of Cluster 12 (Figure 4.21) is close to but slightly below average. As we might have expected given that, there are substantially more regions which do not overlap any recombination events, with 712 of 2183 windows, or 32.6% of the windows did not contain any recombination events. This makes peaks in the Manhattan plot easier to distinguish, and 5 of the these peaks were above the 99$^{\text{th}}$ percentile. Within these peaks were the following genes: *tbp1*, a transferrin-binding protein, *tbpB*, another transferrin-binding protein responsible for iron update, *glmS*, an enzyme responsible for catalysing the first step of hexsoamine metabolism, *metG*, a tRNA ligase required for translation, *yccS*, a putative efflux pump, *pilT*, a pilin protein believed to be responsible for motility, *yggS*, a protein responsible for pyridoxal phosphate homeostasis, *proC*, a part of the pathway for proline biosynthesis, *glmM*, a phosphoglucosamine mutase, *folP*, involved in tetrahydrofolate

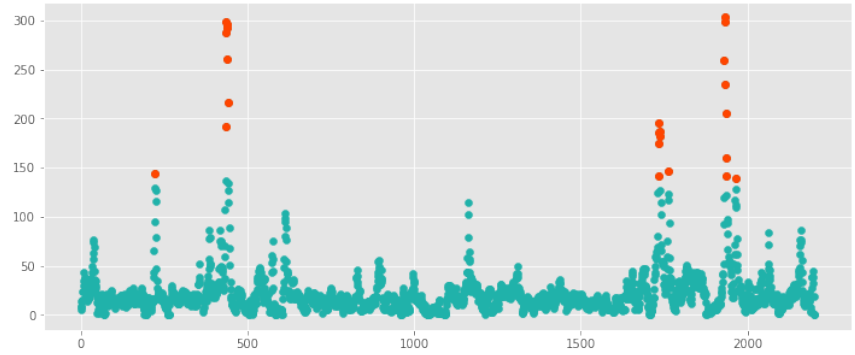


Figure 4.22: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 13. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

biosynthesis, and *aroF*, a protein involved in chorismate and thus amino acid synthesis.

Cluster 13 (Figure 4.22) has one of the higher recombination rates in the collection, which is very much reflected in the Manhattan plot of the recombinations per 1000 base-pair window, where it is clear that recombinations are well-distributed across the entire genome. The percetage of windows which do not overlap any recombination events is similarly very low, at only 0.64% of the the 2202 windows, or 14 windows in total. Despite the overall high level of recombinations across the whole genome, there are still several well defined peaks, 4 of which have windows above the 99$^{\text{th}}$ percentile. These peaks included the following genes: *lbpA*, a lactoferrin-binding protein, *tbpB*, a transferrin-binding protein, *tbp1*, another transferirn-binding protein, all three of which are responsible for iron uptake, *spuE*, a spermidine uptake protein, *carA*, the small chain of a carbamoyl-phosphate synthase, *carB*, the large chain of a carbamoyl-phosphate synthase involved in argenine synthesis, *msrAB*, a protein involved in repairing oxidative damage to other proteins, *pncC*, an enzyme involved in the pyridine cycle, *mscS*, a small mechanically sensitive transmembrane protein, *thiB*, a thiamine uptake protein, *gph*, a DNA repair protein, *glmU*, involved in UDP-N-acetylglucosamine synthesis, *efeB*, an iron uptake protein, and *alx*, a putative transmembrane transporter.

Cluster 14 (Figure 4.23) also has a recombination rate which is substantially above the average across all 25 major lineages,

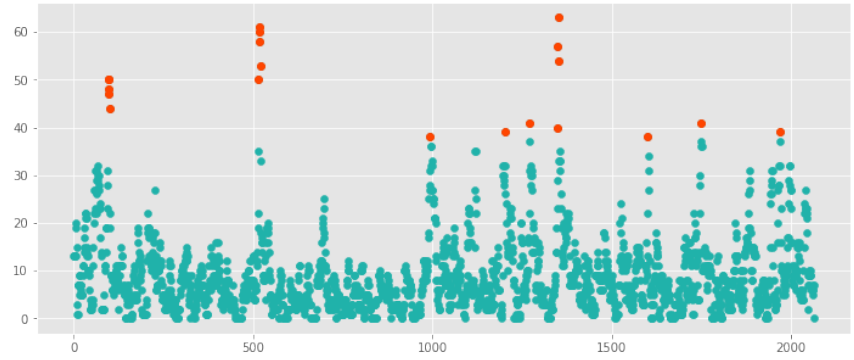


Figure 4.23: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 14. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

however the distribution of recombinations across the genome is a little unusual. Though no particular region has a very number number of recombinations per 1000 base-pair window, the background recombinations per window seems to hover around 10. Only 7% (144/2065) of the windows had zero overlaps with any recombination events. Several peaks had windows which surpassed the 99$^{\text{th}}$ percentile of recombinations per base pair window. The following genes were present within these 'hotspot' peaks: *dapH*, part of the lysine biosynthesis pathway, *mshA*, a protein from the Glycosyltransferase family, *ribD*, a protein involved in riboflavin biosynthesis, *dnaB*, a DNA helicase believed to be involved in chromosomal replication, *porA*, the outer membrane porin, *piiC*, another outer membrane porin, *ydjA*, a putative NAD(P)H nitroreductase, *folP*, a synthase involved in in the tetrahydrofolate biosynthesis pathway, *aroF*, a protein involved in chorismate biosynthesis, *tbp1*, a transferrin-binding protein, *tbpB*, another transferrin-binding protein, *pglA*, a protein involved in carbohydrate synthesis, *rpoN*, a transcription factor, *yccS*, a putative efflux pump, *yclO*, an iron transporter, and *yclN*, part of the same iron transporter complex.

The recombination rate in Cluster 15 (Figure 4.24) is the second-highest out of all 25 major lineages, and as such we see both a high level of recombinations across the genome and high peaks of recombinations per window in the Manhattan plot. Only 73 of 2225 1000 base pair windows in the genome – or 3.3% of windows – did not overlap with a single recombina-

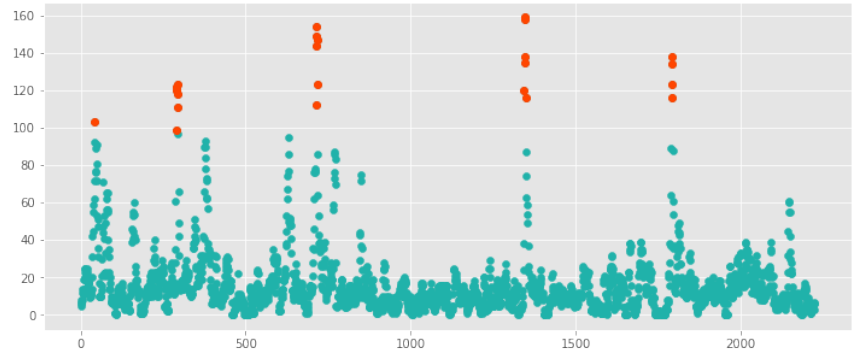


Figure 4.24: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 15. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

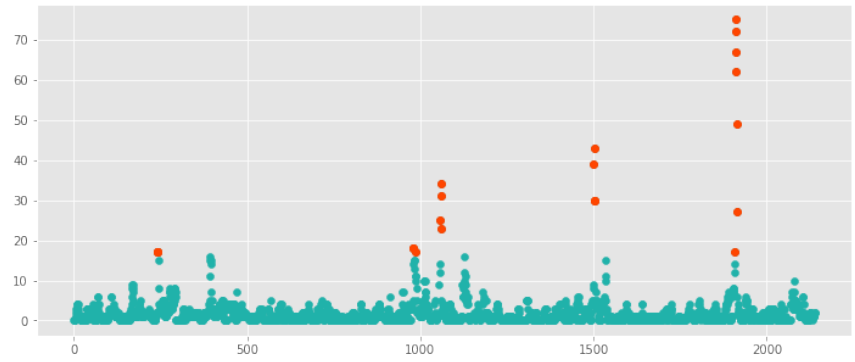


Figure 4.25: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 16. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

tion event. Due to the height of the peaks, 5 were still clearly distinguishable from the background levels and had windows greater than the 99^{th} percentile of recombinations per base pair window. These peaks contained the following genes: *glyQ*, one subunit of a glycine-tRNA ligase, *glyS*, another subunit of the same glycine-tRNA ligase, *gspA*, a membrane-bound protein believed to be involved in secretion, *lex1*, a protein involved in lipooligosaccharide biosynthesis and therefore cell membrane biogenesis, *lbpA*, a lactoferrin-binding protein involved in iron uptake, *dnaB*, a DNA helicase believed to be involved in chromosome replication, *tbp1*, a transferrin-binding protein, and *tbpB*, the other transferrin-binding protein.

Cluster 16 (Figure 4.25) has a ρ/θ recombination rate which is close to but below average, though the Manhattan plot of

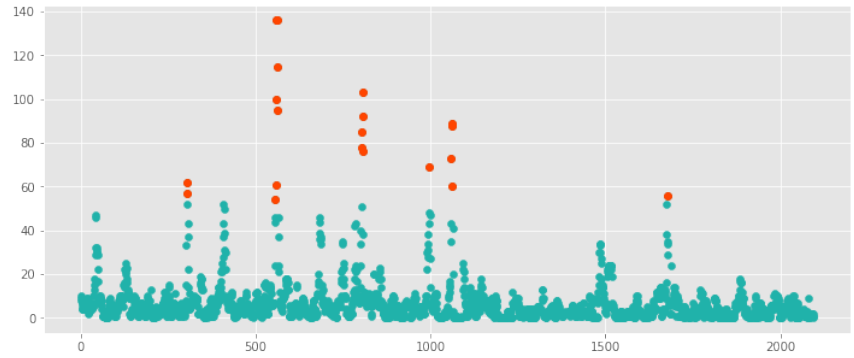


Figure 4.26: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 17. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

recombinations across the genome, and the fact that 37.4% of the 2140 1000 base-pair windows which cover the genome do not overlap with any recombinations shows that high levels of recombination tend to be localised to specific regions. Four regions have windows which fall above the 99$^{\text{th}}$ percentile of recombinations per window, and these regions contain: *pigA*, a protein involved in responding to iron starvation, *pqiA*, a transmembrane protein believed to be a transporter, *lbpA*, the lactoferrin-binding protein, two copies of the type IV pillus *pilE*, and the transferrin-binding proteins *tbpB* and *tbp1*.

The recombination rate in Cluster 17 (Figure 4.26) is close to the average recombination rate across all 25 major lineages, and the Manhattan plot reflects this. 268 of the 2096 windows covering the genome, or 12.8% of windows, have no recombination events, and the rest reflect a relatively low but consistent background rate of recombination. Peaks of recombinations in the Manhattan plot are fairly evenly distributed across the genome, and 6 peaks have windows which fall above the 99$^{\text{th}}$ percentile. Windows which make up these peaks contain the following genes: *hemR*, a heme receptor involved in iron uptake, two copies of the lipooligosaccharide biosynthesis protein *lex1*, *glyS*, a subunit of a glycine–tRNA ligase, *tbpB*, a transferrin-binding protein, *tbp1*, another transferrin-binding protein, the type IV pillus *pilE*, *ydfG*, a 3-hydroxy acid dehydrogenase, *lst*, a protein in the lipooligosaccharide biosynthesis pipeline, *icd*, a protein putatively involved in carbohydrate metabolism, and

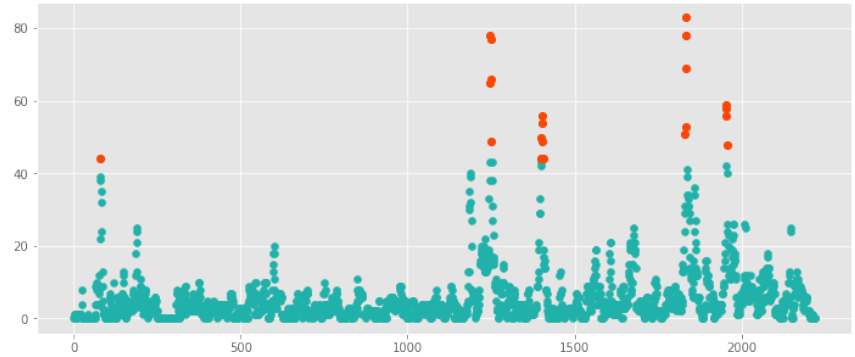


Figure 4.27: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 18. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

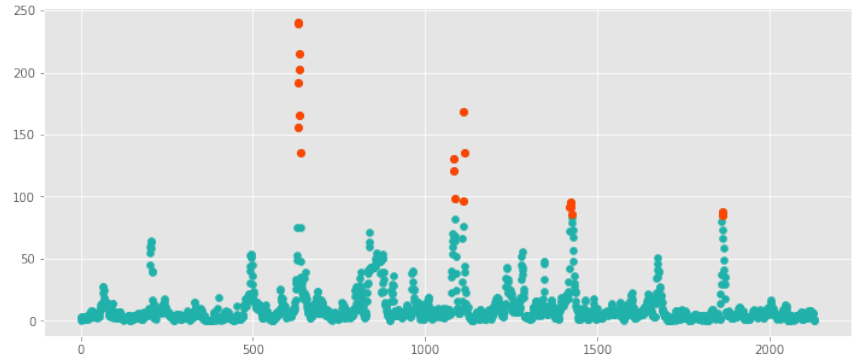


Figure 4.28: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 19. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

the lactoferrin-binding protein *lbpA*.

Cluster 18 (Figure 4.27) has a recombination rate which is similar to that of Cluster 16, slightly below the average across all clusters. It also has a similar Manhattan plot, though fewer windows are completely without recombination, at only 358 out of 2221, or 16.1% of all windows. Peaks are not quite as distinct, but there are a few definite regions which are highlighted by selecting points which fall above the 99$^{\text{th}}$ percentile. These 'hotspot' regions contain the following genes: *hxuC*, an outer-membrane protein involved in iron uptake, *apbC*, an iron-sulphur cluster carrier protein, *lbpA*, a lactoferrin-binding protein, two copies of the transferrin-binding protein *tbpB*, and *tbp1* another transferrin-binding protein.

Unlike the previous three clusters, Cluster 19 (Figure 4.28)

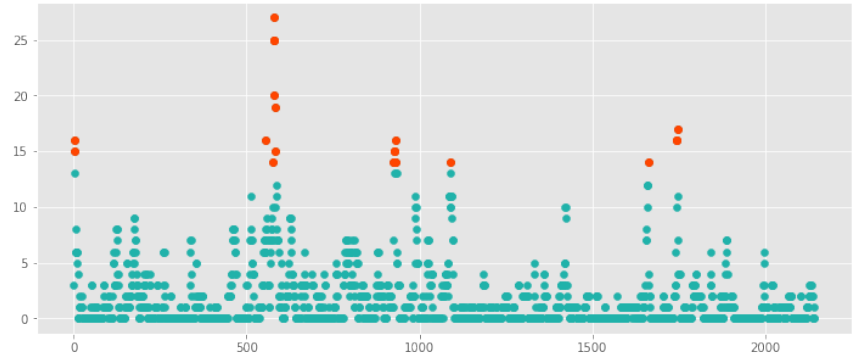


Figure 4.29: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 20. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

has a recombination rate which is within the top 5 recombination rates among the 25 major lineages. This is reflected in the Manhattan plot of recombinations across its genome, which has much clearer peaks, and fewer regions with no recombination, only 4.3% of windows – 92 out of 2132 – had no overlapping recombinations. In the four clearly defined peaks picked out by filtering for windows above the 99$^{\text{th}}$ percentile of recombination the following genes were present: *tbp1*, a transferrin-binding protein, *tbpB*, another transferrin-binding protein, *racE*, part of the peptidoglycan synthesis pathway, *tsaE*, involved in tRNA synthesis, *amiC*, a protein involved in cell wall cleavage, two copies of the type IV pillus *pilE*, *lpxC*, a part of the lipid (IV) biosynthesis pathway, *apbC*, an iron-sulphur carrier protein, *dsbE*, a putative oxioreductase, and *farR*, a transcriptional regulator.

In sharp contrast to Cluster 19, Cluster 20 (Figure 4.29)has a recombination rate that is well below average, and this is abundantly clear in the Manhattan plot of recombinations across its genome. The maximum number of recombinations in a window in this cluster is only 27 recombinations. This lower maximum value also allows the distinguishing of the differences between the counts of recombination in each window, giving a staggered view of the counts of recombination across the genome. There is also a substantial proportion of the genome – 51.6% of 2143 windows – which has no detected recombination events. Within the peak regions in the Manhattan plot above the 99$^{\text{th}}$ percentile of recombinations per window, the following

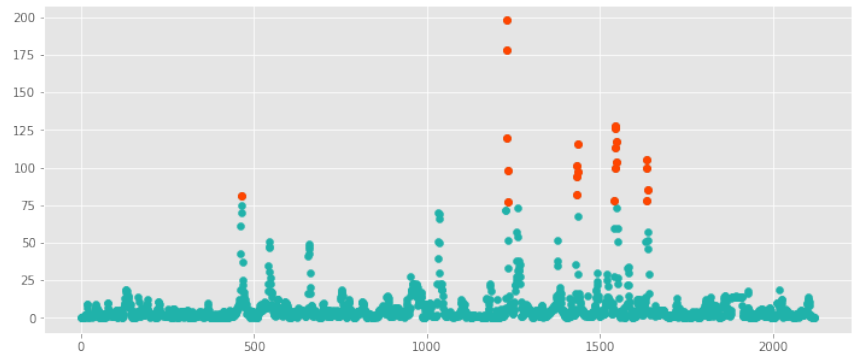


Figure 4.30: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 22. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

genes were present: *tbp1*, a transferrin-binding protein, *tbpB*, another transferrin-binding protein, *piiC*, an outer membrane porin, *apbC*, an iron-sulphur complex carrier, *dsbE*, a putative oxioreductase, *farR*, a transcriptional regulator, *dnaB*, a DNA helicase believed to be involved in chromosome duplication, *cysW*, part of an active sulfate transporter complex, *porA*, an outer membrane porin, *siaA*, a hydrolase of unknown general function, *neuA*, a synthase of unknown general function, *rpoN*, an RNA polymerase sigma-54 factor, and *pglA*, a protein involved in carbohydrate synthesis.

Cluster 22 (Figure 4.30) has a recombination rate that is slightly above average, and the Manhattan plot looks similar to other clusters with similar recombination rates. 13.3% of windows – 277 out of 2122 – did not overlap with any recombination events, but windows above the 99$^{\text{th}}$ percentile picked out 5 peaks in the Manhattan plot. Within these peaks, there were the following genes: *lex1*, a lipooligosaccharide biosynthesis protein, 5 copies of the type IV pillus, *pilE*, *tbpB*, a transferrin-binding protein, *tbp1*, another transferrin-binding protein, *pqiB*, a transmembrane protein, *pqiA*, another transmembrane protein, which like *pqiB* is believed to be a transporter, *pigA*, an oxygen starvation protein, and *lbpA*, a lactoferrin-binding protein.

The ρ/θ recombination rate in Cluster 25 (Figure 4.31) is also slightly above average, similar to Cluster 22. The maximum number of recombinations in a window is only around half of what it is in Cluster 22, however, with the maximum recombi-

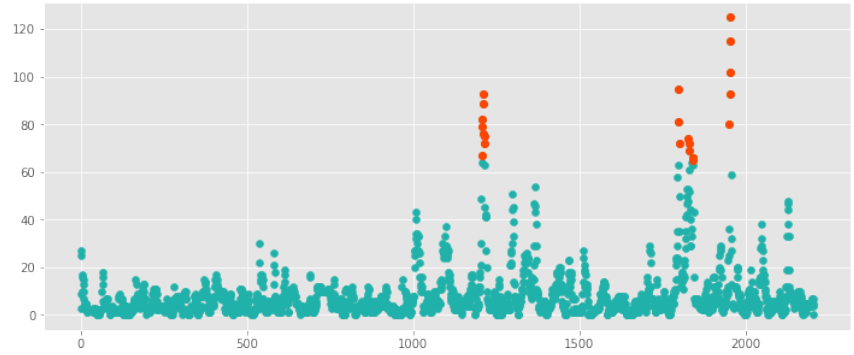


Figure 4.31: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 25. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

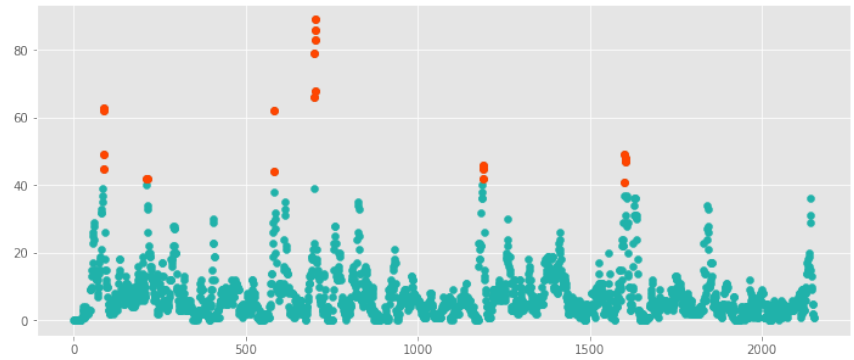


Figure 4.32: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 27. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

nations per window being 125. Recombinations are more evenly distributed across the genome, however, with only 140 of 2206 windows, or 6.34% of windows not having any recombination events within. Peaks seem to be primarily in the latter half of the genome, and three of these peaks fall above the 99$^{\text{th}}$ percentile for recombinations per window. The following genes were contained within these peak regions: *apbC*, an iron carrier protein, *yccS*, an putative efflux pump, 4 copies of *pilE*, the type IV pillus, and *tbp1* and *tbpB*, transferrin-binding proteins.

Cluster 27 (Figure 4.32) has a recombination rate that is comparable to the recombination rates in Clusters 22 and 25, slightly above average. Its Manhattan plot is similar to both of the other clusters' as well, though recombinations are even more well-distributed across the genome, with only 125 of 2152

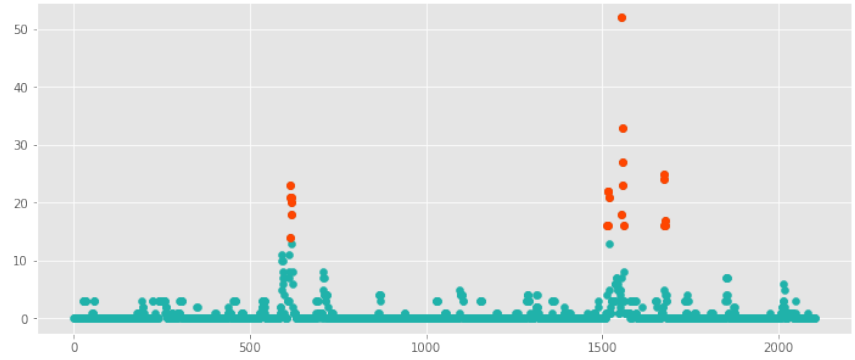


Figure 4.33: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 29. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

windows, or 5.8% having no recombination within them. Peaks in recombinations per window are dispersed across the genome, and 6 of these had windows which fell above the 99$^{\text{th}}$ percentile of recombinations per window. These peaks included the following genes: *menF*, an isochorismate synthase of unknown general function, *lpxC*, part of the lipid biosynthesis pathway, *pilE*, the type IV pillus, *tbpB*, a transferrin-binding protein, *tbp1*, another transferrin-binding protein, *piiC*, an outer membrane porin, and *kpsT* and *kpsM*, two components of a transmembrane transport protein.

The recombination rate in Cluster 29 (Figure 4.33) is a significant departure from the previous three clusters, being the third lowest recombination rate our of all 25 major lineages. The low recombination rate is reflected in the very few recombination events present in its genome, with 1672 out of 2104 windows – 79.5% – not overlapping with any recombination events, by far the highest proportion of any of the 25 major lineages. The are still 3 distinct peaks in recombination events, however, and all 3 are picked up above the 99$^{\text{th}}$ percentile of recombinations per window. These three peaks contained: *tbpB*, a transferrin binding protein, *tbp1*, an additional transferrin binding protein, *lpxC*, a protein involved in lipid biosynthesis, 8 copies of *pilE*, the type IV pilus, two copies of *lex1*, a protein involved in lipooligosaccharide biosynthesis, and *glyS* and *glyQ*, two components of a glycine-tRNA ligase.

Cluster 36 (Figure 4.34) is more in keeping with what has

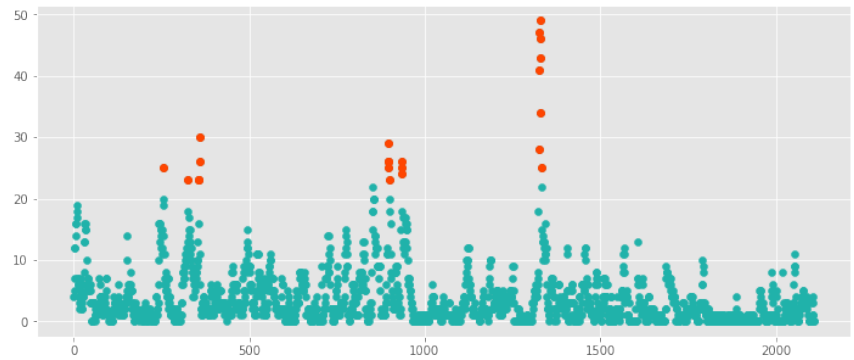


Figure 4.34: Manhattan plot of the number of recombination events overlapping discrete 1000 base-pair windows across the reference genome of Cluster 36. Windows greater than the 99th percentile of recombinations per window are plotted in orange.

been observed across the major clusters, having a recombination rate much more similar to Clusters 22, 25, and 27, slightly above average. The maximum number of recombinations per window is actually slightly lower than cluster 29, but the recombinations are much more well-distributed across the genome, 20.5% of windows, 432 out of 2110, do not have any recombination events within. The first half of the genome seems to generally have more recombinations than the second half, and selecting windows above the 99$^{\text{th}}$ percentile of recombinations per window identifies three 'hotspot' peaks in the recombination rate. The following genes were present in these regions: *piiC*, a porin, 4 copies of *pilE*, the type IV pilus, *lpxC*, involved in lipid biosynthesis, 2 copies of *lex1*, which is involved in cell membrane synthesis, *rsmE*, a methylase which acts on the the small subunit of the ribosome, *cysQ*, a nucleotidase of unclear general function, *ydfG*, a 3-hydroxy acid dehydrogenase of unclear general function, *lst*, part of the lipooligosaccharide biosynthesis pathway, *icd*, isocitrate dehydrogenase, *tbpB*, a transferrin-binding protein, *tbp1*, another transferrin-binding protein, *racE*, a protein involved in cell wall synthesis, *tsaE*, responsible for a tRNA synthesis reaction, and *amiC*, a protein which is involved peptidoglycan catabolism and therefore cell wall cleavage.

This exhaustive examination of the hotspots of recombination in the 25 major lineages of the global collection reveals two major trends. The first is the extent to which the existence of regions of high recombination and the genes present in those

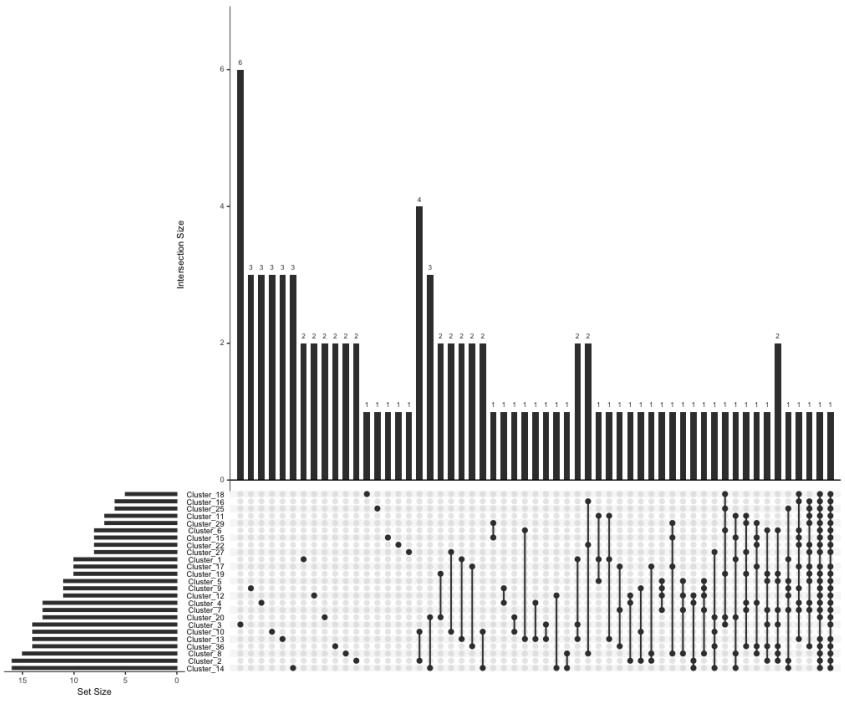


Figure 4.35: Upset plot of the overlap in genes present within 'hotspot' regions of recombination above the 99$^{\text{th}}$ percentile of recombinations per 1000 base-pair window in each cluster. The number of genes present within the hotspots of each cluster is indicated on the horizontal bar chart on the left, and the number of these genes which are either unique to a single cluster (one dot in the interaction matrix) or common between multiple clusters (2-7 dots connected by a line in the interaction matrix) is indicated in the central vertical bar chart.

regions are consistent across the diversity of *N. meningitidis*. Each cluster has clear regions where recombination far exceeds the background level (Figures 4.10-4.34), and though the intensity and distribution of these peaks is varied across the 25 major lineages, the gene content of the highest peaks of recombination remains relatively consistent across all lineages. 91 genes appeared in all lineages recombination 'hotspots' – windows which fell above the 99$^{\text{th}}$ percentile of recombinations per window in each lineage – yet only 38 of these genes (42%) were unique to any given lineage. The transferrin-binding proteins *tbp1* and *tbpB* were present in 22 and 23 of the major lineages, respectively, and *pilE*, the gene which encodes for the major type IV pilin subunit, was present in a recombination hotspot of 12 of the major lineages. 9 additional genes were present in 6 or more of the 25 lineages. The concurrence of unique genes in recombinant regions between clusters is fully shown in Figure 4.35, in an Upset plot of the overlap between the sets of genes present in the 25 major lineages. The main body of the plot shows the the number of genes which are shared between different groupings of the major lineages, ranging from unique genes on the left to those which are shared between almost all the clusters on the right. This is likely to be an underestimate of the true extent of concurrence due to the conservative method used to identify hotspots – a percentile cut-off of all windows means that tall peaks can dominate and obscure shorter peaks which are still significantly higher than the background level. Despite its probable underestimate, the extent of shared genetic content in these regions of high recombination is still surprising. What is driving these specific genes to have such high levels of diversity which is not believed to be the result of mutation? In at least one case, it may be the result of within-organism homologous recombination, or antigenic variation, particularly in the case of *pilE*, which is known to exist in several silent cassettes within the *N. meningitidis* genome which are then swapped with the actively transcribed copy [47]. This is not currently known to occur with any other genes in *Neisseria*, however, so it is unclear if this mechanism could explain the consistent appearance of other genes in recombination hotspots.

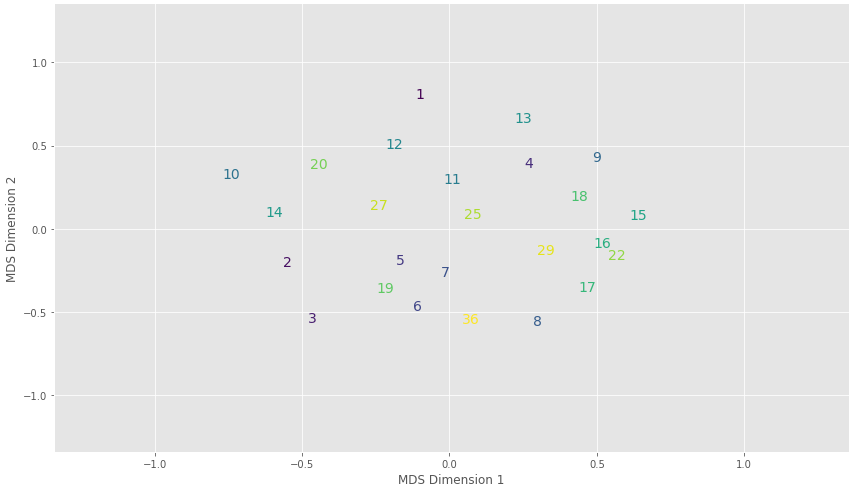


Figure 4.36: Multidimensional scaling plot of the pairwise Jacard distances $J(A, B)$ between the presence-absence matrices of genes in recombination hotspots among the 25 major lineages of *N. meningitidis* present in this collection

Phase variation with short sequence repeats (SSRs) is another mechanism of within-genome recombination in *Neisseria* [43], however, and several of the genes which are present across many of the 25 major lineages have been identified as genes which undergo phase variation, particularly the iron uptake genes, *tbp1* and *tbpB*. Phase variation with SSRs would lead to detected recombination events with a length on the 10^1 order of magnitude, however, and across all of the major clusters, the average recombination length overlapping one of the transferrin-binding genes is 4904.65 base-pairs long, with within-cluster average ranging from 1903.56 base pairs to 6975.82 base pairs. In general, these genes are known to be very diverse across species, and it is commonly believed that this is due to strong selective pressure from the immune system which causes the increased diversity [152]. It is indeed the case that most of the genes identified in recombination 'hotspot' peaks (Figures 4.10-4.34) are surface exposed and therefore likely to be subject to diversifying selection, increasing the diversity within that region and leading to many more detected recombinations, even though the true recombination rate is roughly similar across the genome. It is difficult to directly investigate whether or not this is the case, as we are obviously unable to detect recombinations which fall above some percentage of identity with the recipient genome.

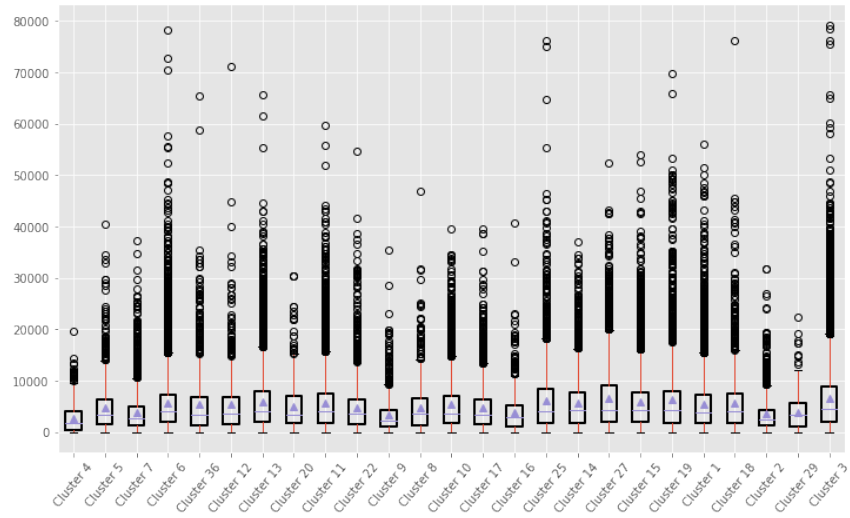


Figure 4.37: Boxplot of the distribution of inferred recombination lengths for each of the 25 major lineages in this collection.

An alternative to these patterns being driven by increased diversity in certain regions, however, would be if these regions themselves had some genetic property which caused them to be more recombinant than other regions. If this were the case, we might expect to see some structure in the differences between the genes present in the hotspots of various lineages. Figure 4.35 offers some suggestion that this may be the case as certain intersections of 2 or 3 lineages appear several times in conjunction with 3th and 4th lineages. This is not a reliable measure, however, and a more formal approach is to calculate pairwise Jacard distances (see section 2.2.2.1) between each pair of lineages based on the presence-absence matrix of hotspot genes. These distances can then be projected into two-dimensional space to create a MDS plot as per Figure 4.5. This is shown in Figure 4.36, where it is quite clear that there is no easily discernable pattern – some clusters are a little closer together than others, but in general they are well spaced and spread out. If there is a genetic factor driving increased recombinations in certain regions, it is likely not directly a results of the gene content in those regions.

The second more subtle trend evident when examining the Manhattan plots of recombination events across the genome in the 25 major lineages is the effect of recombination length on the locations of recombinations across the genome. This is

easiest to see in lineages where the maximum number of recombinations is relatively low and the average recombination length is relatively long, particularly in Cluster 20 (Figure 4.29) where shorter recombinations that overlap with longer recombinations creates a pattern of staggered "bands" of recombination across the genome. The extent to which this occurs in the lineages' Manhattan plots appears to vary – is this substantially different between clusters? Figure 4.37 shows the boxplots of the lengths of the recombinations detected in the 25 major lineages. In general, they range from a few base pairs to a maximum of around 80,000 base pairs long. The differences in recombination length are also significant between clusters (Kruskal-Wallis $H = 1963.01$, $p < 10^{-16}$), though the differences between clusters are not as widespread as those found in recombination rate (Figure 4.38 for results of Dunn's *post hoc* tests). The lengths of recombinant DNA fragments are generally believed governed entirely by stochastic processes, as a key determinant of their size is how fragmented DNA becomes in the extracellular environment before it is taken up and integrated into the chromosome of the recipient cell. That the distributions of lengths are significantly different between some clusters is surprising, and suggests that non-stochastic factors may affect the the lengths of recombinant DNA. There are a number of non-random variables which could theoretically affect the length of recombinant DNA, including the extent of synteny and homology between the DNA donor and recipient, and how likely it is that the recipient bacterium will encounter more syntentic or less syntentic DNA, all of which could play a role in causing these significant differences.

4.3 Recombination and selection

Recombination in the *N. meningitidis* genome is therefore diverse in both its frequency and location, and both of these occur in highly non-random ways – lineages are significantly different in their recombination rates, and certain regions of the genome are much more recombinant than others. The recombination rates between isolates sampled from disease and those sampled from carriage are also significantly different. Although recombi-

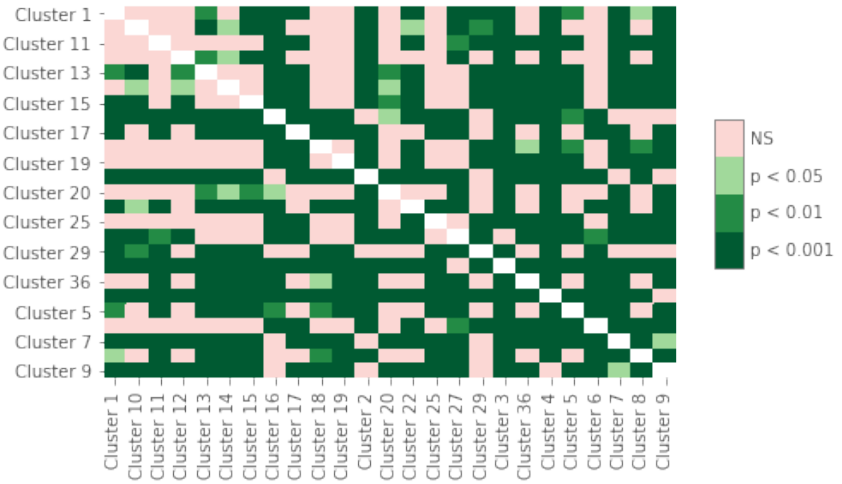


Figure 4.38: Heatmap of the significance results, with p values as indicated in the legend, of *post-hoc* Dunn's pairwise comparisons tests for between-groups between the recombination lengths of the 25 major lineages present in this collection.

nations might be expected to be randomly distributed across the genome, in all lineages there are regions which have significantly more recombinations detected than others. Selection is often invoked as the explanation for this phenomenon, as the genes which are often found present in regions of high recombination are known to be diverse and surface-exposed, and hence very likely to be subject to significant selective pressure from the human immune system [152]. This suggests that,as discussed in the previous section, selection may lead to increased diversity, and increased diversity would in turn lead to detection bias being responsible for the appearance of increased recombination in recombination 'hotspot' regions. However, studies in eukaryotic organisms have highlighted the interaction between recombination and selection, for instance by acting to increase the efficacy of directional selection, though this interaction has not been fully explored in bacteria. This is, in part, due to difficulty associated with identifying regions under selection in bacteria, when most theory and method development is focused on freely-recombining diploid organisms [9]. Could this type of interaction between recombination and selection, as opposed to simply the generation of increased diversity, be involved in creating recombination 'hotspot' regions?

As discussed in section 2.2.2.3, there are no current analysis

Cluster	SNP 1 Position	SNP 1 Gene	SNP 2 Position	SNP 2 Gene
3	1630215	*fabI*	1630217	*fabI*
	1069160	*lolB*	1069167	*lolB*
11	1675790	Hypothetical	1675792	Hypothetical
	1824421	*metK*	1824422	*metK*
15	2118239	*tufA*	2118240	*tufA*
	1032521	Hypothetical	1032522	Hypothetical
	2145198	*zraR*	2145200	*zraR*
	1037481	Hypothetical	1037487	Hypothetical
	2154536	*pyrB*	2154537	*pyrB*
	1202467	Hypothetical	1202474	Hypothetical
	587039	*btuB*	587040	*btuB*
19	201173	Hypothetical	201475	Hypothetical
	202846	Hypothetical	202852	Hypothetical
	202846	Hypothetical	203088	*kpsM*
	202852	Hypothetical	203088	*kpsM*
	202846	Hypothetical	203292	*kpsM*
	202852	Hypothetical	203292	*kpsM*
	203088	*kpsM*	203292	*kpsM*
	202992	*kpsM*	202995	*kpsM*
	202846	Hypothetical	202995	*kpsM*
	202852	Hypothetical	202995	*kpsM*
	202995	*kpsM*	203088	*kpsM*
	202995	*kpsM*	203292	*kpsM*
	202992	*kpsM*	203088	*kpsM*
	202846	Hypothetical	202992	*kpsM*
	1427796	Non-coding	1427798	Non-coding
	202852	Hypothetical	20299	*kpsM*
	203254	*kpsM*	203292	*kpsM*
	202992	*kpsM*	203292	*kpsM*
	202846	Hypothetical	203254	*kpsM*
	202852	Hypothetical	203254	*kpsM*
	203088	*kpsM*	203254	*kpsM*
	1081793	*yccS*	1081807	*yccS*
	1082011	*yccS*	1082014	*yccS*
	208518	Hypothetical	208533	Hypothetical
	203586	*kpsM*	203592	*kpsM*
	1674066	Hypothetical	1674069	Hypothetical
	1091790	*thiB*	1091793	*thiB*
	203756	*kpsT*	203765	*kpsT*
	202995	*kpsM*	203254	*kpsM*
	203010	*kpsM*	203016	*kpsM*
	203532	*kpsM*	203538	*kpsM*
	202992	*kpsM*	203254	*kpsM*
	1429382	*fhs*	1429385	*fhs*
22	1034692	Hypothetical	1034694	Hypothetical

Table 4.1: Table detailing the loci of the 45 extreme GWES outliers from a cluster-by-cluster GWES analysis for SNPs under selection

methods which are capable of detecting loci under selection across the entire collection. However, with some constraints, there are methods that will allow us to detect selection in subsets of the collection. The first of these, spydrpick [134] detects pairs of variants which are much more tightly epistatically linked than population structure or linkage disequilibrium would suggest. Table 4.1 contains all the "extreme outliers" present in the spydrpick results from each cluster – those pairs of SNPs which are much more tightly linked than background levels of linkage and population structure suggest, making it highly likely that they are under selection. Five of the 25 main lineages had pairs of variants which fell above the variable "extreme outlier" threshold, clusters 3, 11, 15, 19, and 22, across which there were 45 pairs of SNPs. These 45 pairs were found in 12 genes of known function and 9 hypothetical genes across 5 clusters. That significant signals of nucleotide variants under directional selection are found in only in 5 of the major clusters is surprising when all 25 major clusters span a substantial period of time, but this likely predominantly driven by the limited sensitivity of considering only extreme outliers when using GWES to detect selection. The GWES results do confirm that there are at least some cases of significant positive selection, though it is impossible to interpret these signals of selection when the variants concerned are found in hypothetical, predicted genes with no known annotation. Among the annotated genes, several point to interesting patterns of selection, particularly in the *kpsM* and *kpsT* genes of Cluster 19. Genes with these names have not been well-studied in any *Neisseria* species, but the have been studied extensively in *E. coli* where it is known that they encode for a transmembrane ABC transporter responsible for transporting polysialic acid across the cellular membrane. A blastx search of the uniprot protein sequence database suggests that these two genes have been mis-annotated in the automatic annotation, and instead are the genes *crtD* ('*kpsT*') and *crtC*, ('*kpsM*'). *crtC* and *crtD* are well-studied in *N. meningitidis*, and are known to be responsible for capsule transport[153]. Given the importance of capsule in infections, and particularly in causing disease, it is unsurprising that there would be such a strong signal of selection in these

genes. Evidence of evolution in the capsule genes of Cluster 19 is interesting as it is made up predominantly of Serogroup E isolates, which are generally not viewed as a disease-causing serogroup – though there have been sporadic reports of cases recently [37] – but also especially interesting due to the fact that Cluster 19 has by far the most genogroup diversity of any cluster, with 6 genogroups and capsule null isolates represented (Figure 3.25). The other genes of note which have a strong signal for being under selection in Cluster 19 are *thiB* and *fhs*. *thiB* has not been characterised in *N. meningitidis*, but research in other species has shown it and other genes in its pathway to be involved in thiamine biosynthesis[154, 155]. Interestingly, it is present in recombination hotspots in Clusters 1, 3, and 13. *fhs* is another synthase, responsible for the synthesis of 10-formyltetrahydrofolate [156]. Given that there is no evident reason to believe that these variants in particular are under selection, it is also very much a possibility that these pairs of variants could result from hitchiking on the *ctrC-ctrD* sweep, though spydrpick is deigned to avoid trying to detect such occurrences. In Cluster 3, variants in the genes *fabI* and *lolB* are tightly epistatically linked and therefore likely co-selected. *fabI* is a gene that is part of the fatty acid synthesis pathway[157, 158] and *lolB*, though not studied in *N. meningitidis* has been shown to be involved in lipoprotein localisation[159]. The related function of these genes is further evidence of genuine co-evolution between them. In Cluster 11, two single-nucleotide variants in *metK* are highly epistatically linked, suggesting that one of those two variants of the gene has, or both have recently been under strong positive selection. *metK* is the gene responsible for the synthesis of S-adenosyl-L-methionine, and has been shown to be essential for growth in some bacterial species[160]. Finally, in Cluster 15, four different non-hypothetical genes contain pairs of single nucleotide variants which seem to be under selection, *tufA*, *zraR*, *pyrB*, and *btuB*. *tufA* encondes for a translational elongation factor[161], and has been molecularly shown to exist in other *Neisseria* species[162]. *zraR* has not been studied in *N. meningitidis* but research in other species has shown it to encode for one component of the ZraPSR envelope stress

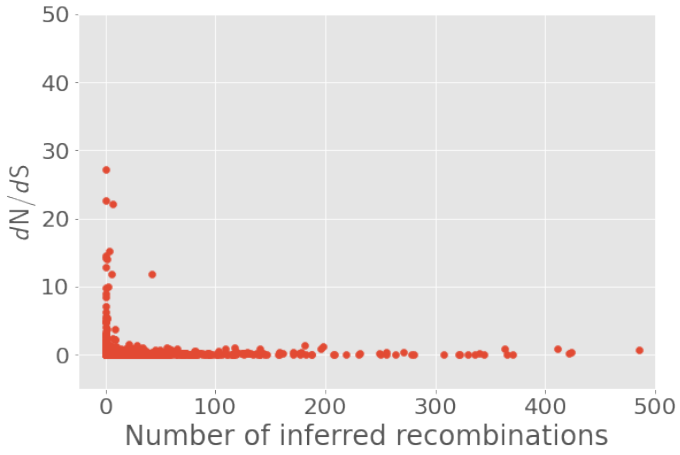


Figure 4.39: Scatter plot of number of recombination events versus $^{d_N}/_{d_S}$ for genes in the pan-genome of the entire Burkina Faso collection of carriage isolates. Genes with abnormally high estimates of $^{d_N}/_{d_S}$ were excluded from the plot as described.

response protein, which has been implicated in antimicrobial tolerance[163]. Given the importance of envelope stress response in responding to a range of external stressors, it is unsurprising that it should be found to be under selection. Given that all Cluster 15 isolates were collected from cases of invasive disease, however, it also possible that this is reflects an adaptation in response to host immune response. *pyrB* has been shown to code for aspartate transcarbamylase in *N. gonorrhoeae*[164], and described as essential in other species [165], but there is no obvious reason why it would have recently been under strong positive selection in Cluster 15. It is interesting to note, however, that another gene in the pathway, *pyrE* [166] was found in recombination hotspots in Cluster 2 and Cluster 10. Finally, *btuB* has been shown to encode for a vitamin B12 transporter in *E. coli* [167], and it is present in a recombination hotspot in Cluster 8.

None of these variants under selection, as detected by spydrpick, directly overlap at all with recombination 'hotspot' regions. Some genes with strong evidence of selection based on the results of the spydrpick output are found in recombination hotspots in other clusters, but this is insufficient evidence to suggest a direct interaction between recombination and selection. This is perhaps unsurprising as recombination would severely reduce the selective signal GWES analyses look for, but it may still be

possible to detect such an interaction by focusing our analysis on a subset of the dataset which is restricted in space and time to a local population, as previously done (Figures 4.6-4.7) to examine the extent of recombination between major lineages in a localised population. Recombinations in this case were inferred on a gene-by-gene basis on the pan-genome of the Burkina Faso dataset, which allows the use of more generalised methods of detecting genetic evidence of directional selection within a gene. Figure 4.39 is a scatterplot of the $^{d_N}/_{d_S}$ values for each gene in the pan-genome where it was possible to make such a calculation (where neither d_N nor d_S were zero) using the Nei-Gobojiri [168] algorithm for estimating $^{d_N}/_{d_S}$ as implemented in Snpgenie [169]. Figure 4.39 suggests the existence of a clear negative correlation between $^{d_N}/_{d_S}$ and recombinations, where the more recombinations a gene has, the lower its $^{d_N}/_{d_S}$. A simple non-parametric Spearman's rank correlation [150] demonstrates that this correlation is highly statistically significant returning a negative correlation coefficient r_s of -0.122, with an associated $p = 2.21 \times 10^{-6}$. This suggests that one evolutionary consequence of recombination in *N. meningitidis* is to reduce the accumulation of deleterious mutations – which would lead to a higher d_N and therefore $^{d_N}/_{d_S}$ – by increasing the efficacy of negative selection, removing deleterious alleles from the population. This has long been accepted as one of the reasons for the maintenance of recombination in eukaryotic organisms [84], and though it is perhaps unsurprising that recombination would play a similar role in prokaryotes, the significant negative correlation between the $^{d_N}/_{d_S}$ and the number of recombinations in a gene is a rare example of clear evidence of recombination playing such a role.

Figure 4.39 does, however, show the existence of a single outlying gene with a $^{d_N}/_{d_S}$ of over 10, despite having many more recombinations (42) than most genes with a similar $^{d_N}/_{d_S}$, unusual given the overall trend to which this gene, the DNA methylation modification enzyme *dpnA*, is the only exception. Analysis of the multiple sequence alignment of this gene with the BUSTED [170], FUBAR [171], and aBSREL [172] algorithms in the HYPHY [173] software package which implements molecular

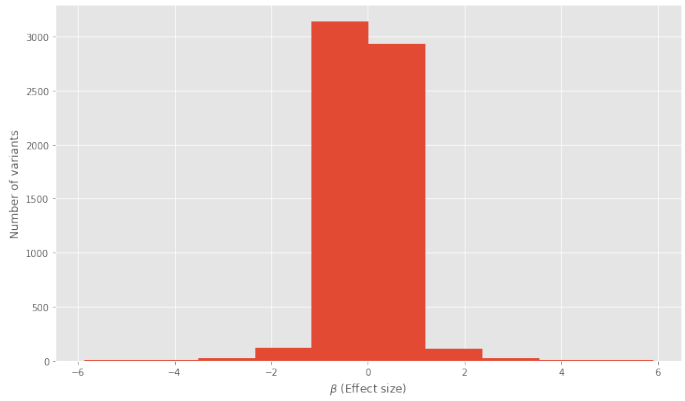


Figure 4.40: Histogram of the effect sizes (β) of the 6375 unitigs which were significantly associated with ρ/θ recombination rate, at a significance threshold of $5x10^{-8}$.

evolution methods for detecting signatures of selection accepts the alternative hypothesis that the entire gene is under positive selection (BUSTED, $p = 0.00015$), finds good evidence that there is a single amino acid site under selection (FUBAR, posterior probability=0.9039), and of the 42 branches of the gene phylogeny which contained the recombination event in this gene, at least one branch could reject the null hypothesis of not being under directional selection (aBSREL, $p = 0.00059$). Another known effect of recombination in eukaryotes is speeding the effect of positive selection [85], and the combined evidence of *dpnA* being under positive selection with its relatively high number of recombinations suggests that recombination may be playing a similar role in this case. DNA methylation is known to have significant consequences for a number of processes in *Neisseria*, including recombination itself [146], and it is possible that *dpnA* was therefore under selection in the Burkina Faso population after the turbulence introduced into the population through the mass vaccination campaign started shortly after the fist data of collection of the Burkina Faso isolates.

4.4 Genetic factors underlying recombination rates

As discussed in the methods section of Chapter 2, specifically Section 2.2.2.5, advances in bacterial GWAS methodology have

Unitig	Unitig Frequency	LRT-adjusted p-value	β
1: GGCAATCAATCCTGCCGCTTCGCGCCGCATCACCTCTTG	0.0755	4.840000×10^{-88}	5.87
2: ATGCGGCGCGAAGCGGCAGGATTGATTGCCGCCG	0.0755	4.840000×10^{-88}	5.87
3: CTTCGGCGGCAATCAATCCTGCCGCTTCGCGCC	0.0755	4.840000×10^{-88}	5.87
4: TTTACACCTACGCGCAAGAGGTGATGCGGCGCGAAGCGGCAG	0.0755	4.840000×10^{-88}	5.87
5: TCCTGCCGCTTCGCGCCGCATCACCTCTTGCG	0.0755	4.840000×10^{-88}	5.87
6: GGAATAGGCATATCCGACAACAATGCCGTCCGAAGATTCAGACGGCAT	0.1340	1.340000×10^{-85}	5.87
7: AAAAAACAAAAAGCCGAACCCAAAGCCACCGTCGG	0.0755	9.570000×10^{-89}	5.90
8: GGCACCAAAAAACAAAAAGCCGAACCCAAAGCCACC	0.0755	9.570000×10^{-89}	5.90

Table 4.2: Table of the unitigs which are significantly associated with ρ/θ recombination rate in a GWAS, and have high effect size $\beta \geq 4$. LRT-adjusted p-values are p values adjusted with a likelihood ratio test to account for the nested effect of population structure. β reflects the effect size of the variant on the phenotype.

allowed for increasingly sensitive association studies between phenotypes of interest and different types of genomic variation. The ρ/θ recombination rate is calculated for each isolate in the 25 major lineages, and is therefore also isolate phenotype for all isolates in the 25 major lineages, and we can therefore directly asses the extent to which genetic variation affects the recombination rate. I do so here using unitigs [174] as the genetic variants, as they are an extension of k-mer based associations using the nodes compacted De Bruijn graphs to reduce redundancy and therefore make the computation associated with a dataset of this size much more feasible. In total, 4693555 unique unitgs were discovered in the population, and I therefore filter the results with a significance threshold of 5×10^{-8}, the threshold typically used for associations studies where millions of variants are tested[175].

6375 unitigs were significantly associated with recombination rate, but as Figure 4.40 shows, the effect sizes of most of these variants are extremely small. The variants with the largest positive effect sizes ($\beta \geq 4$) are reported in Table 4.2. Unitigs 1-5, 7, and 8 are found in all Cluster 4 isolates, and also at low frequency in Cluster 2 (Unitigs 1-5) and Cluster 22 (Unitigs 7 and 8). Unitg 6 is found at high frequency in Cluster 2, and at low frequency in Cluster 6. Remarkably, unitigs 1,2, and 5 are not found through a blastn search of the NCBI non-redundant nucleotide database, unitigs 3 and 4 primarily are found in *N. gonorrhoeae* isolates, and unitigs 7 and 8 in other *Neisseria*

species, primarily *N. mucosa*, *N. sicca*, and *N. bacilliformis*. Unitig 6 is found primarily in *N. meningitidis*, where it matches the 5' end of a prophage-associated protein in the FAM18[43] reference genome.

The fact that most unitigs, apart from 7, seem to be primarily found in other *Neisseria* species or not found at all in the nr database would typically suggest that these sequences likely represent some form of contamination, but the fact that these unitigs are present in all isolates in at least one of the major clusters (Clusters 2 and 4), indicates that some other phenomenon is at work. The clusters these unitigs are present in at high frequency is additionally confusing, as they both of have below-average recombination rates (Figure 4.4). The only explanation which is consistent with these associations, for unitigs which appear to have been an import from other species, is that unitigs 1-5, 7, and 8 were gained ancestrally in Cluster 4, but their recent gain in Clusters 2 and 22 has lead to an enormous increase in recombination rate, which has lead to their strong association with recombination rate. A similar event has likely occurred involving unitig 6, which maps to a prophage-associated gene which is present at high frequency in Cluster 2 and at low frequency in Cluster 6. Looking at the isolates with these unitigs at low frequency in Clusters 2, 22, and 6 confirms that this is definitely the case – ρ/θ in these isolates is the maximum value found across all the isolates for which these values were calculated, with $\rho/\theta = 6$ in all cases.

Although these variants with high effect sizes ultimately appear to be driven by a small number of very recombinant isolates, the general fact that 6375 unitigs are significantly associated with recombination rate confirms that there is a substantial genetic component affecting the observed rates of recombination in *N. meningitidis*. These unitigs range from 31 base pairs long to 182 base pairs long, and surprisingly, the mean effect size is actually slightly negative (-0.01006), but very close to 0. In general, that the unitigs are so balanced in their effect sizes raises an important issue: although these uitigs are strongly associated with recombination rate, that does not necessarily mean that the association is causative, as illustrated

by the unitigs found in Table 4.2. Can we confirm that at least some of these unitigs are associated with recombination rate and show some evidence of causality? A simple way of doing this is to see if any of the unitigs contain sequences which are known to be associated with recombination rate. The 10 base-pair DNA uptake sequence has been shown to be directly related to DNA uptake in *Neisseria*[47, 176] and specifically *N. meningitidis*[177]. 200 unitigs significantly associated with recombination contained the 10 base-pair *Neisesria* DNA uptake sequence, and the mean effect size of these unitigs is low, but positive at 0.0354145. A Mann-Whitney U test[149] confirms that this is significantly different from the mean among unitigs without the DNA uptake sequence.

4.5 Concluding remarks

Since long before the advent of whole-genome sequencing, it has been well-known that recombination is a common occurance in *N. meningitidis* [46]. Here, we demonstrate that its frequency is not consistent between lineages or across the genome of any given isolate – significant differences exist between lineages, and the rates of recombination across the genome are highly unequal. Even the positions of 'hotspot' regions, which have elevated recombination rates, differ between different lineages, although certain genes which have long been shown to be highly recombinant are commonly found in hotspots across multiple lineages. Even when focusing our analysis on a highly localised subset of the global dataset, these significant differences persist between lineages, confirming that recombination is a frequent, ongoing process. These significant differences also suggest that the rates of recombination in *N. meningitidis* are not simply governed by stochastic factors, but must have a genetic component. This is confirmed by the GWAS of unitigs versus recombination rates, where 6375 unitigs were very highly significantly associated with recombination rate. Although not all these significant associations are directly causal, as shown by the unitigs with the highest effect size, but at least some, which contain a clear mechanism for affecting recombination rate in the form of a DNA uptake

sequence, appear to be. Interestingly, these DUS-containing unitigs have very small effect sizes, suggesting that the genetic component controlling recombination is perhaps not large, but definitely affected by a significant number of loci.

Recombination within the Burkina Faso collection also demonstrates the extent to which recombination occurs primarily within the species. Although there is much evidence of sporadic recombination between *N. meningitidis* and other *Neisseria* species[70, 178, 179], and indeed further evidence found here, in Table 4.2, Figure 4.7 demonstrates that, at least in one regional population, this represents a very small fraction of recombination events, most of which occur between distantly related lineages of *N. meningitidis*. The study of recombination in the Burkina Faso population also provides some evidence for the evolutionary consequences of recombination in that population, which seems to be be primarily preventing the accumulation of deleterious mutations[105], a well-known evolutionary function of recombination[84]. Evidence of selection within major clusters of the global population also exists, with evidence of a major selective sweep in the capsule transport genes of Cluster 19, the most diverse cluster in terms of genogroups. Though predominantly composed of genogroup E isolates, 11.67% of its isolates are typed as genogroup B, and smaller numbers are identified as genogroups W, X, Y, and Z, and also some non-groupable, likely capsule null. Although the *ctrD* and *ctrC* genes are not directly responsible for the shifts in genogroup, their loci are proximate to proximate to the biosynthesis genes[153], and a large capsule-switching recombination event could be responsible for the change in genogroup and the introduction of these variants into the cluster and spydrpick has detected the subsequent strong selection on the entire capsule locus which has led to its expansion, potentially of the genogroup B isolates which make up just under 12% of sampled isolates from Cluster 19. Although this single event is not enough to make any general conclusions about the interaction between recombination and selection within the global population, the existence of identical or similar in genes detected by spydrpick and in gene recombination hotspots – though in different clusters – combined with

the strong example of selection in the capsule transport genes suggests that there may be some general entanglement of the two evolutionary forces during the course of adaptation in *N. meningitidis*. The existence of repeated recombination hotspots across different lineages has previously been used as indirect evidence of this phenomenon [152], but in the Burkina Faso collection I find direct evidence of this phenomenon in the form of a gene with a high number of recombination events and a high $^{d_N}/_{d_S}$, and exception to the overall trend.

Although the combination of recombination and selection inference performed in this chapter has suggested that there are occasional episodes of selection which interact with recombination within lineages, no current methods or data are able to determine how selection acts on lineages or between lineages as a whole. Chapter 3 suggests that the expansion and bottlenecking of different lineages is a frequent occurrence, and these demographic changes may well be linked to selection at certain loci, and recombination may disrupt the demographic shift in such cases by allow loci under selection to jump from their initial lineage into other lineages present in the background diversity in the region. No methods can test that hypothesis, because they are unable to account for the extent of genome flexibility between lineages, and no data provides a representative enough sample from the population to allow for the disentangling of demographic shifts caused by selection and drift. In this chapter, I used an analysis of the pan-genome of a regional population to partially account for the former challenge by reducing the whole genome sequences of the Burkina Faso dataset to their coding regions and a network describing their syntenic relationships in each isolate combined across the population. This approach allowed me to assess how recombination and selection were interacting in the entire population, but much is still unknown about how pan-genomes themselves evolve [79], with investigation on the limited number of datasets on the scale of 10^4 isolates providing surprising and novel insight into the distribution of genes in their species' pan-genomes[81]. What can this global collection of *N. meningitidis* reveal about the whole species' pan-genome?

CHAPTER 5

The *Neisseria meningitidis* Pan-genome

The Pan-Genome – the union of the sets of all genes present within the individuals of a species – is a concept which serves two main purposes in the context of this thesis. The first is to enable the comparative study and categorisation of species with extremely flexible genomes that do not have anywhere close to 100% synteny between individuals. The second is to identify homologous regions across individual members of diverse species to enable downstream analyses without the use of a reference genome when the diversity in a species exceeded what could be captured in a linear reference. When the concept was first introduced, most research involving pan-genomics focused on trying to address the first issue – trying to understand and classify the genomic flexibility of strains and species using pan-genomic methods. The staggering extent of diversity, however, quickly made researchers rethink the utility of such study – could any extent of sampling accurately capture a diversity which was so vast and, in the case of some species, changing so quickly? Attention then turned to primarily using the pan-genome and pan-genomic methods primarily as a tool to study other topics of interest.

Recent methodological advancements in the speed, and particularly the accuracy, of pan-genome inference have led to a resurgence of interest in understanding pan-genomes themselves. In this chapter, we will primarily focus on developing this un-

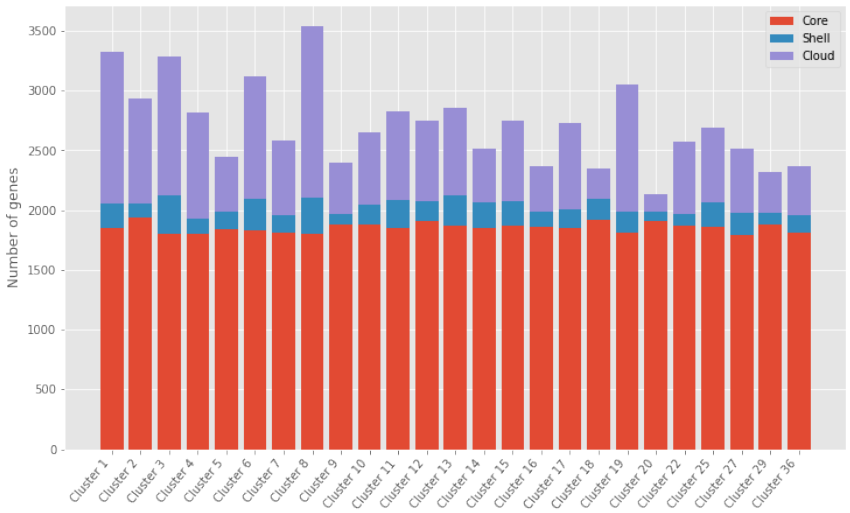


Figure 5.1: Stacked bar graph indicating the distribution of genes within each clusters' individual pan-genome, for the 25 major clusters. Core genes are those which are present in 95% of the isolates in a cluster, shell genes between 95% and 15% of isolates, and cloud genes are those genes which are present in less than 15% of the cluster's isolates.

derstanding in *Neisseria meningitidis*. *N. meningitidis* is one of a number of species which are particularly interesting from the perspective of pan-genome analysis, for a number of reasons. As we have seen from the analysis of the population structure in Chapter 3, *N. meningitidis* is a species with a deeply divergent and complex population structure, with substantial diversity present at low frequency globally, and also in local populations. Yet, as we have seen in Chapter 4, it is also very recombinogenic, though to varying degrees between different lineages in the population. These characteristics of its evolutionary genetics, combined with its host-obligate niche, should combine to create a pan-genome structure quite different from species where some isolates are capable of independent survival, or are similarly host-restricted, but less recombinogenic.

5.1 Structure of the pan-genome

The pan-genome in this collection of *N. meningitidis* consists of 10626 genes in total, unsurprisingly significantly greater than previous estimates using two orders of magnitude fewer isolates [180, 181]. Of these genes, a relatively small proportion are 'core'

genes, with only 1533 of 10626, or 14.43% of these genes being present in more than 95% of isolates in the collection. This is smaller than early estimates of the size of the *N. meningitidis* core genome by around 200 core genes [180] but consistent with more recent estimates of the pangenome size [181]. Most genes, 8396 of 10626 – 79.01% – are cloud genes present in less than 15% of the isolates in collection, with the remaining 697 accessory genes present in between 15 and 95% of the isolates in the collection. Figure 5.1 shows the size and structure of the pan-genome of each lineage, where the above criteria are applied within each lineage. Though some variation in the total size of the pan-genome exists in each cluster, it is mostly driven by variation in the number of cloud genes, with the core genomes of all lineages in particular maintaining a relatively consistent size. Differences in the number of shell and cloud genes are also likely to be subject to sampling bias, making it difficult to draw any firm conclusions about the sizes of the different clusters' pan-genomes. As for the core genes, apart from their relatively consistent size across clusters, it is notable that their absolute sizes are surprisingly high. These range from 1791 genes to 1936 genes, compared to an overall species core of 1533, making the core of each cluster 250-400 genes larger, suggesting the existence of cluster-specific genes which are core genes in that cluster, either unique to each cluster or shared between subsets of the 25 major clusters. The total sizes of each cluster's pan-genome, range from 2129 genes to 3537 – much smaller than the species-level pan-genome of the entire collection. This reflects a relatively high number of unique genes within the cloud portion of each cluster's pan-genome. Figure 5.2 shows the number of unique genes per cluster, and indeed most have > 100 unique genes, with the average being 270.92 unique genes per cluster, with the number of unique genes ranging from 60 to 880 among the clusters.

The general trend in the numbers of unique genes is for larger clusters (ie Clusters 1-4) to have more unique genes, while smaller clusters, such as 20-36, tend to have fewer unique genes. This loose trend is consistent with the number of unique genes being governed primarily by non-adaptive processes, here

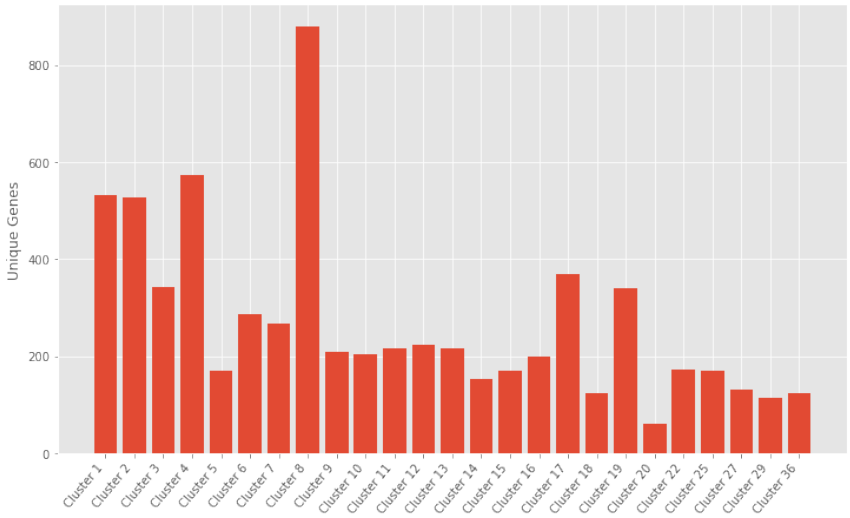


Figure 5.2: Bar graph showing the number of unique genes per cluster. The average number of unique genes per cluster is 270.92.

determined either by the number of isolates sampled or possibly the relative population sizes of each cluster. The significant exceptions to that trend, however, in for example Cluster 3, Cluster 8, Cluster 17, and Cluster 19 which have fewer, in the case of Cluster 3, or greater numbers of genes than this simple trend would suggest. It may be the case that the numbers of unique genes present within a cluster is governed instead by recombination rate, and a simple way to test this is to assess the correlation between the average ρ/θ recombination rate of each cluster, calculated as described in Chapter 4, and the number of unique genes per cluster. This correlation (Scatter plot in Figure 5.3) is surprisingly slightly negative $r_s = -0.03616$ but is highly non-significant, with $p = 0.8637$. Based on these 25 clusters, it seems that recombination rate has little to no effect on the number of unique genes per cluster. This is somewhat surprising given that the acquisition of unique genes requires the uptake and integration of external DNA, both of which should be tightly correlated with the overall recombination rate, but this lack of correlation primarily the fact that gene gain is likely to be considerably rarer than recombination. Within the time frame of the phylogenies of the 25 major clusters, we have observed many thousands of recombination events, but the number of gene gain events, judging by the number of unique genes, is at least an order of magnitude lower, and unique gene acquisition

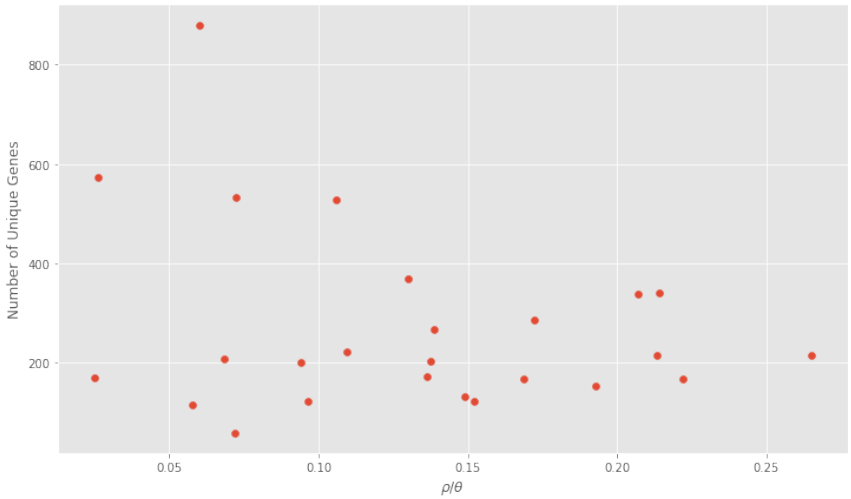


Figure 5.3: Scatter plot of the recombination rates of the 25 major clusters vs the number of unique genes in that cluster. No significant correlation: Spearman correlation coefficient: $r_s = -0.03616$, $p = 0.8637$

is not restricted to the time span of the phylogeny in the same was as recombination inference, therefore making factors other than the genetic (and therefore physiological) properties of the bacteria in each lineage much more important in determining whether or not a unique gene acquisition takes place. The lack of correlation may also reflect the existence of other stochastic and non-stochastic factors governing not only the acquisition of novel genes, but also their persistence after the initial acquisition. Both selection and genetic drift are known to play important roles in determining whether or not a novel mutation goes extinct, and particularly given the likely pronounced effect of drift in the evolution of *N. meningitidis* noted in Chapter 3, it seems likely that these effects on the persistence of a novel gene acquisition will be far more influential in determining the number of unique genes in a cluster. The spurious minor negative correlation observed between recombination rate and the number of unique genes seems likely to be driven by Cluster 8 in particular, which has the fourth lowest average recombination rate, but by far the greatest number of unique genes.

Although the relative differences in the distribution of unique genes between clusters is interesting, it offers no insight into what function these genes might be serving, and whether or not that differs significantly between clusters. However, even

in clusters with relatively few unique genes, it is impossible to meaningfully interpret a list of gene names without any additional context. Fortunately, Gene Ontology (GO) [182, 183] enrichment analyses provide a basic but scaleable method of comparing large sets of genes with some background population. Here, we use the hmmer2go software package to annotate the entire pan-genome with GO terms by searching the pan-genome reference sequences against Pfam with Hmmer, followed by a Parent-Child-Union enrichment analysis [184] of each cluster's unique genes against the background of the entire pan-genome, using the Ontologizer GO enrichment software package [185].

All clusters had GO terms which were significantly enriched in their set of unique genes after initial tests, but after multiple testing correction only four clusters contained unique genes with a statistically higher representation of certain GO terms at the $\alpha = 0.05$ significance threshold, with an additional cluster, Cluster 8, having a significantly enriched GO term at the $\alpha = 0.1$ significance threshold. These results are summarised in Table 5.1, but the most notable feature of the results is that every single GO term that was enriched among the unique genes of the various clusters is related to binding either DNA or RNA, with three clusters having the same enriched GO term, GO:0032196, 'Transposition', any process involved in mediating the movement of discrete segments of DNA between non-homologous sites. Although all these terms are still very general and do not point to a direct similarity in the various clusters' unique genes, the slightly unexpected nature of the terms, having to do with DNA binding as opposed to directly affecting some cellular process, combined with the fact that there is such a consistent signal across five different clusters does suggest that there may be some common overall function in the unique genes present within each cluster, as it is not necessarily expected for there to be any overlap in GO terms between any of the clusters. In particular, DNA and RNA binding GO term enrichment is consistent with previous research on *N. meningitidis* population, though at a much smaller scale, which identified unique restriction-modification systems within major lineages [103], possibly leading to reproductive isolation.

Cluster	GO Term	Holm-Bonferroni p	GO Category	Name
4	GO:0034061	0.000657	Molecular Function	DNA polymerase activity
8	*GO:0003723*	*0.074373*	*Molecular Function*	*RNA polymerase*
16	GO:0032196	0.000053	Biological Process	Transposition
16	GO:0140097	0.007361	Molecular Function	Catalytic activity, acting on DNA
29	GO:0032196	0.016678	Biological Process	Transposition
36	GO:0032196	0.016678	Biological Process	Transposition

Table 5.1: Table of significantly enriched Gene Ontology terms in clusters with significantly enriched gene ontology terms. Cluster 8 is presented in italics as it does not meet the $\alpha = 0.05$

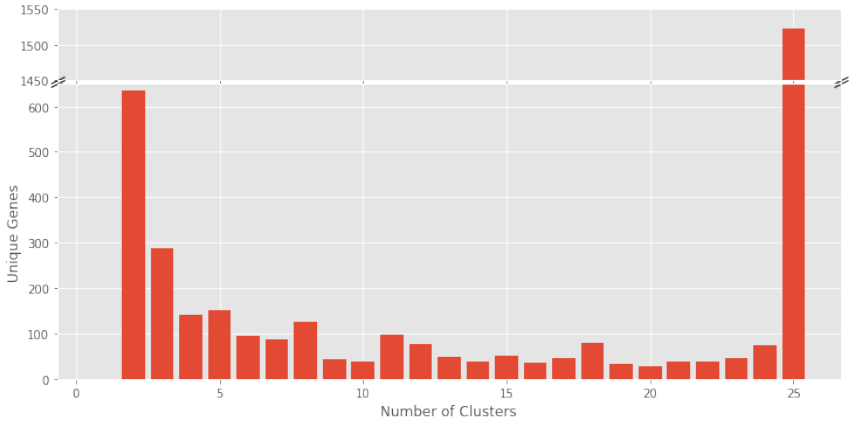


Figure 5.4: Histogram depicting the numbers of non-unique genes which span 2-25 clusters.

Cluster-unique genes, the most substantial part of the lineage-specific pan-genome in *N. meningitidis*, represent around 63.74% of the pan-genome. In total there are only 6773 genes which are unique to single clusters, and even when combined with the core genes, this only represents 8306 of 10626 genes. The remaining 2320 genes in the species-wide pan-genome are those genes distributed at low frequency across multiple clusters. What form does the distribution of these genes take? Figure 5.4 shows the distribution of non-core, non-unique genes across all clusters, and from it we can see that the distribution seems to be primarily as we would expect if the association were mostly driven by chance – genes which are present in multiple clusters are either present in all clusters, most of which are core genes, or they are present in relatively few clusters, with fewer and fewer genes being present the greater the number of clusters they are expected to span. This is largely as we would expect, but these genes are also crucial in providing additional information regarding how isolates are related in terms of their pan-genome. While PopPUNK determines relatedness between two isolates by calculating the Jacard distance (section 2.2.2.1) between the sets of their k-mers, it is also possible to perform a similar pairwise distance calculation between isolates based instead on the sets of present and absent genes within each isolate. Then, by using the t-distributed stochastic neighbour embedding (t-SNE) dimension reduction method to project these series of high-dimension distances into two dimensions, we can visually assess how isolates are cluster based on the patterns of gene presence and absence in their genomes.

Figure 5.5 shows the resulting plot of isolates in two dimensional space from applying a t-SNE to the pairwise distance matrix of Jacard distances based on gene sets, coloured by PopPUNK cluster. Some of the 25 major clusters, it can be seen, do form relatively tight, discrete groupings of isolates in this projected pan-genome space, in particular, Clusters 9, 12, 13, 18, 20, and 25 show evidence of relatively straightforward structure in their pan-genomes. Most clusters, however, show at least some evidence of either major structural splits, or smaller clusters of isolates which are very distant in the pan-genome space

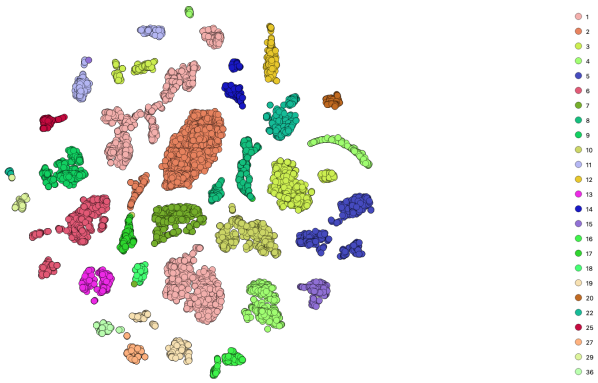


Figure 5.5: Plot of the t-SNE-reduced pan-genome pairwise Jaccard distance matrix between all isolates in the global *N. meningitidis* collection, coloured by PopPUNK cluster as indicated in the legend. An interactive version of this figure with additional metadata is available at the following uniform resource locator: `https://microreact.org/project/oqkHTv7N2CgJbNkUW3NoqP`

from the rest of the isolates. The lineage splits of small numbers of isolates are relatively straightforward to interpret; they most likely represent a recent gene gain from outside the cluster of origin. In many cases, these single or handful of outlier isolates are overlapping isolates from other clusters, suggesting that the exchange of genes within the pan-genome is largely between different lineages of *N. meningitidis*, consistent with finding from the network of recombination events between clusters between clusters in the Burkina Faso population 2009-2012 (Figure 4.7). The splits between large groupings within clusters are more difficult to interpret, as like with any dimensional reduction, we do not know precisely which factors lead to the grouping together, or not, of isolates. Nonetheless, at a basic level it is simply evidence for the existence of structure within the pan-genome of each cluster. In some clusters, such as Cluster 4, this seems to broadly follow the phylogenetic structure within the lineage – the two major groupings of Cluster 4 isolates roughly correspond to the two branches of the phylogeny, but further investigation reveals that this correspondence is not absolute, and isolates from the opposite lineage occasionally are grouped with very distant isolates within the cluster based on patterns in their pan-genomes. The same is true for many other clusters, indicating that Cluster 4 is not an exception, and the evolution of pan-genomes is in many cases not wholly determined by the

whole-genome phylogeny, even if there is some correlation.

5.2 Pan-genome dynamics in different lineages

Taking a step back from examining the structure of the pan-genome itself, we might seek to understand how the pan-genome changes over time, and how that relates to our understanding, developed over the last two and half chapters, of the evolution of *N. meningitidis*. Historically, studies of pan-genome dynamics relied primarily on inferring trends based on statistics regarding the pan-genome simply based on its inference. This approach, similar to the analysis performed in the first section of this chapter, suffered due to its reliance on over-extrapolation from statistics such as the size of an inferred pan-genome, which were prone to error and could be strongly biased by sampling. Panaroo implements a number of more modern approaches to understanding pan-genome dynamics, which rely on inferring parameters based on an explicit model of pan-genome change over time. We will use one of these inference implementations [79], based on the 'Finitely many genes' (FMG) model of pan-genome evolution, to infer gene gain and gene loss rates from the dated phylogenies produced for each cluster in Chapter 3 and each cluster's gene presence-absence matrix created in the process of pan-genome inference [186].

The finitely many genes model of pan-genome change over time models the gain and loss of individual genes over time with four key parameters: the gene gain rate a, the rate of change from absent to present on a branch of the phylogeny; the gene loss rate v, the rate of change of the reverse; M the size of the pool of genes from which the phylogeny can gain or lose genes; and finally G, the number of genes present in the average genome in the phylogeny. The finitely many genes model offers a key advantage compared to the infinitely many genes model of pan-genome evolution[187, 188], as the FMG model allows for the same gene to be gained multiple times, which is not accounted for in other models. With the exception of collections of closely-related outbreak isolates, repeated gain and loss is

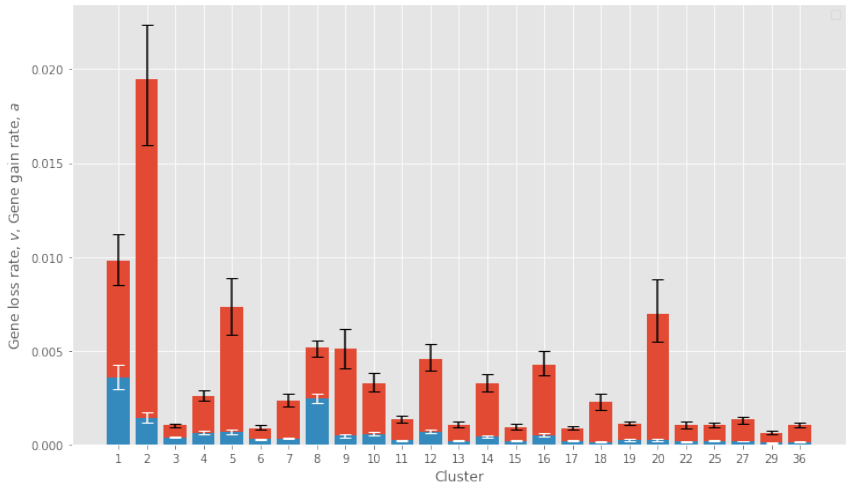


Figure 5.6: Bar chart of the inferred gene gain rate (in red) and gene loss rate(in blue) for the 25 major clusters. Error bars indicate the 95% confidence interval of the estimates, obtained through bootstrapping the estimate.

essential for the accurate modelling of pan-genome change in diverse phylogenies spanning several decades of sampling.

Figure 5.6 shows the initial results from the gain and loss rate parameter inference. Reassuringly, the values are extremely small, with gene gain rates ranging from being on the order of 10^{-3} to the order of 10^{-2}, with gene loss rates being roughly an order of magnitude less. The differences between some clusters are extremely pronounced, though in general there seems to be around three different levels of clusters with overlapping confidence intervals. Finally, although the largest two clusters have the greatest rates of a, gene gain, and some of the greatest rates of v, gene loss, in general cluster size does not seem to correlate well with either rate. As it is difficult to interpret the net magnitude of pan-genome change over time based on these two rates, Figure 5.7 displays the relative gene gain rate versus loss rate. Here, the trend becomes very clear. Though gene gain rate ranges from 2-fold to 30-fold more frequent than gene loss, most clusters have a relative rate of gene gain around the average – 7.85, gene gain generally being between 5 and 15 fold more often than gene loss. A handful of clusters have substantially lower gene gain rates (Clusters 1, 3, 4, 5, 8), and a couple of clusters much higher gene gain rates, particularly Clusters 18 and 20. Similar to the pattern of unique genes across clusters,

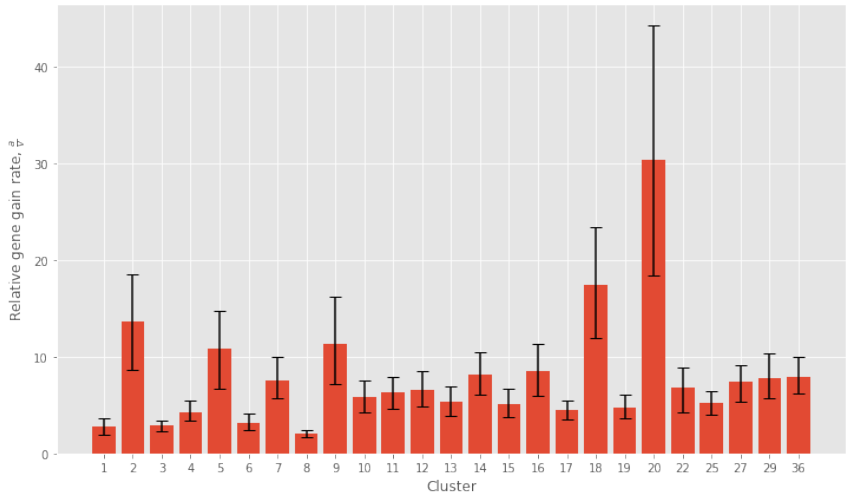


Figure 5.7: Bar chart of the relative gene gain versus loss rate, a/v for the 25 major clusters. Error bars indicate the combined 95% confidence interval of the estimates, originally obtained through independent bootstrapping of both estimate.

the general invariance of the relative gene gain rate between clusters suggests that gene gain and loss in the pan-genome is primarily shaped by factors other than recombination rate, such as selection or frequency of mixed-lineage infection.

If this were to be the case, we would expect to see no correlation between the clusters' a/v relative gene gain rates and the average ρ/θ recombination rates, as the two variables are governed by different phenomena and should therefore be independent. Figure 5.8 is the scatter plot of recombination rate versus relative gene gain rate, and we indeed do see that there is no obvious correlation between the two, as we see in Figure 5.8. A Spearman's [150] rank correlation confirms this, with $r_s = -0.2892$ and $p = 0.1608$. The absence of significant correlation is not unexpected given the observations regarding the general structure of pan-genome in the first section of this chapter. It confirms that the rate of gene gain the pan-genome is primarily determined by factors other than recombination, likely directional selection and genetic drift, though stochastic factors such as the frequency of lineage-lineage interaction doubtlessly also play a role.

Surprisingly, counter to the intuitive notion that a higher relative gene rain rate should lead to more unique genes, there is a significant negative correlation between the a/v relative gene

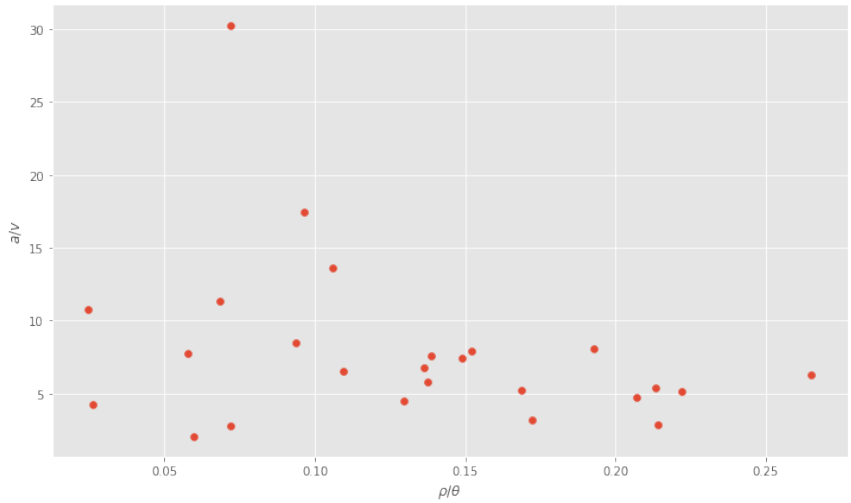


Figure 5.8: Scatter plot of each cluster's average ρ/θ recombination rate versus its relative gene gain versus loss rate, a/v for the 25 major clusters. No significant correlation, $r_s = -0.2892$ **and** $p = 0.1608$.

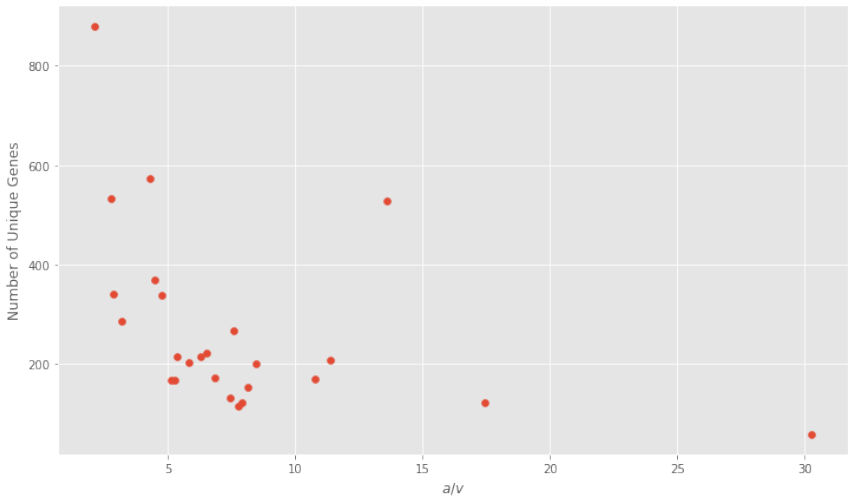


Figure 5.9: Scatter plot of the relative gene gain versus loss rate, a/v of, versus the number of unique genes present in the 25 major clusters. Spearman's rank correlation indicates that they are negatively correlated $r_s = -0.6544$ **and** $p = 0.0003874$.

gain rate and the number of unique genes (Figure 5.9). This is a surprising and somewhat counter-intuitive result given that our expectation is that lineages with higher rates of gene gain should more quickly accumulate unique genes. While that supposition would make sense if the pattern in gene gain rates behaved like recombination rate, governed by a combination of genetic factors and stochasticity, we have seen evidence which suggests that is not the case. If both the relative gene gain loss rate and the number of unique genes are governed by selection, but also crucially, the opportunity to acquire genes through stochastic interaction with other clusters, it is possible that clusters which are, for whatever reason, interacting more with other clusters (in our sampling) are much more likely to gain genes, but as that interaction is necessarily both ways, much less likely to maintain their genes uniquely. Cluster 20 therefore, which has the highest relative gene gain rate, seems to be interacting much more with other clusters – at least in our sampling – and therefore has the fewest unique genes. Cluster 8, on the other hand, which has by far the most unique genes, is interacting least with other clusters and therefore has the lowest gene gain rate. Unexpectedly, this interpretation is also consistent with the minor (and non-significant) negative correlation observed between the number of unique genes and average ρ/θ recombination rate, as the lack of opportunity to gain genes would imply also a lack of opportunity to recombine. This trend has two clear implications – first, that most gene gain does actually occur from other lineages of *N. meningitidis*. This is unsurprising, though not guaranteed by any means, and difficult to examine directly with this dataset, making this indirect signal in the data very valuable. Secondly, this suggests that some of the stochastic factors governing interaction between lineages may not, in fact, be truly random. Cluster 8 and Cluster 20 are the exact opposite lineages one would naively expect to have the gene gain rates and the number of unique genes that they do, with Cluster 8 containing isolates with more geographical diversity, a wider sampling window, and evidence of some degree of complex population structure. Cluster 20 contains only isolates from Europe and North America, and has one of the most narrow

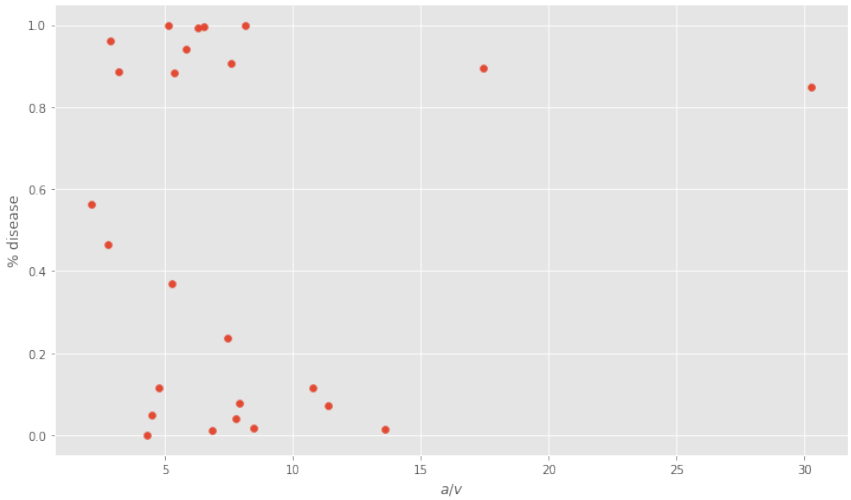


Figure 5.10: Scatter plot of the relative gene gain versus loss rate, a/v of, versus the disease isolates in the 25 major clusters. **No correlation,** $r_s = -0.1473$ **and** $p = 0.4822$**.**

sampling windows and appears to have a simple clonal outbreak population structure, yet more evidence of recent interaction with other lineages of *N. meningitidis*. How different lineages disperse, migrate, and subsequently then interact in a structured bacterial population remains very much unknown, but this is one indication that different lineages do seem to differ in at least part of that process. This is further consistent with the evidence [103], discussed earlier in this chapter, that there are mechanisms for reproductive isolation between different lineages of *N. meningitidis*.

Finally, though we are not necessarily expecting there to be a significant correlation, it is worth assessing whether or not the relative gene gain rate is associated with disease. Figure 5.10 shows a scatter plot of each cluster's gene gain rate plotted against the proportion of disease isolates in that cluster, and it is very clear from both the scatter plot and a Spearman's rank correlation ($r_s = -0.1473$, $p = 0.4822$) that no correlation exists.

Inference using the finitely many genes model estimates one other parameter of interest, M, the size of the pool of genes from which the isolates in the phylogeny gain and lose genes. In a sense, this is an inference of the true or effective size of the pan-genome, accounting for the imbalances and biases in sampling which restricted our ability to draw conclusions from the number

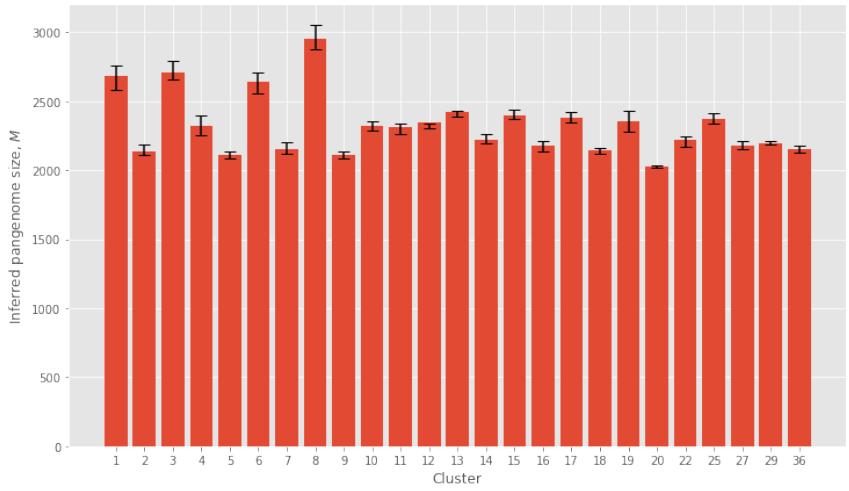


Figure 5.11: Bar chart of the inferred pangenome size, M, under the FMG model of pangenome change, for the 25 major clusters. Error bars indicate the 95% confidence interval of the estimates, originally obtained through bootstrapping.

of genes present in each cluster after in our inferred pan-genome (Figure 5.1). M, as inferred, (Figure 5.11) generally does reflect the trends in the raw counts of the number of genes (Figure 5.1), but the extent of the variance between clusters is somewhat decreased from the raw counts, likely as a result of accounting for the differences in the extent of diversity sampled between different clusters. Combined with the overall consistency of pan-genome size across clusters (Figure 5.1), the decrease in variance suggests that the size of a cluster's pan-genome is likely under considerable evolutionary constraint, and therefore does not diverge, even between deeply divergent lineages of the global *N. meningitids* population.

A Spearman's rank correlation [150] between each cluster's average ρ/θ recombination rate and M, the inferred pan-genome size (Figure 5.12), surprisingly found a significant, though not particularly strong correlation ($r_s = 0.4077$, $p = 0.04308$). Recombination rate has previously been found to be not significantly correlated with neither the number of unique genes nor a/v , the net gene gain rate, so its significant correlation with the estimated size of the pan-genome is rather unexpected. Our understanding of how recombination rate is independent from gene gain and the number of unique genes, which are negatively correlated, can also explain this, however. Relative gene gain

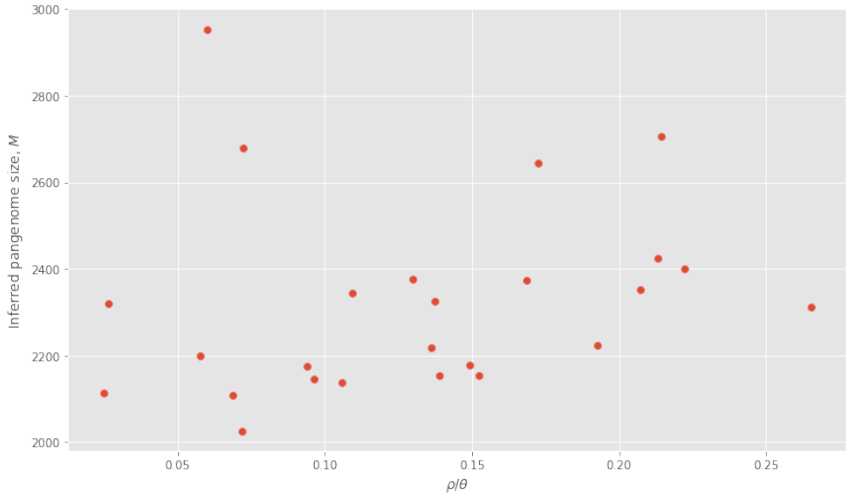


Figure 5.12: Scatter plot of each cluster's average ρ/θ recombination rate versus their inferred pan-genome size, M. A Spearman's rank correlation finds a moderate, but significant correlation between the two. $r_s = 0.4077$, $p = 0.04308$

rates seem to be primarily determined by selection or genetic drift and the frequency of interaction between lineages, and negatively correlated with the number of unique genes because of the inverse relationship between the number of unique genes in a cluster and the frequency of its interaction with other lineages. Overall pan-genome size, however, is not driven entirely by unique genes, but also non-unique accessory genes shared between clusters. Given the aforementioned evidence in this chapter that a substantial portion, if not most of the genes gained in *N. meningitidis* are – as we might expect *a priori* – from other lineages of *N. meningitidis*, the total number of genes must therefore depend upon the frequency of interaction between lineages to maintain or increase these numbers of genes, though it is unlikely to be the predominant component in determining the pan-genome size. Similarly, recombination rate is somewhat, though not primarily, determined by the frequency of interaction between bacteria of different lineages, and the co-variation between recombination rate and pan-genome size is likely what is driving their weak yet significant correlation here. The observation of this correlation was also found in a similar study in *Streptococcus pneumoniae* [79], and though the two bacteria are very distantly related evolutionarily, their similar life histories may explain this observation in both species.

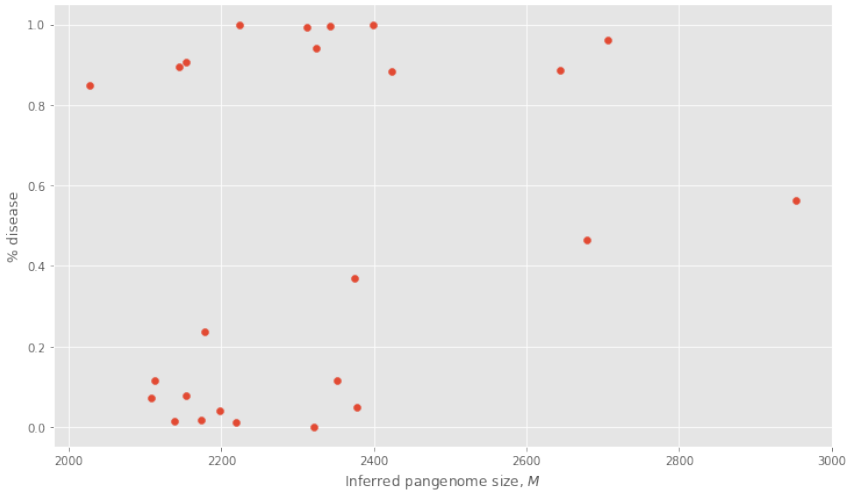


Figure 5.13: Scatter plot of the inferred size of each cluster's pan-genome, M, versus the proportion of disease isolates in each Cluster. **No significant correlation exists.**

Previous results in *S. pneumoniae* also found a significant correlation between M and the proportion of isolates which were isolated from cases of invasive disease [79]. We find no such correlation in *N. meningitidis* however, with a Spearman's rank correlation: $r_s = 0.3131$, $p = 0.1275$. Figure 5.13 suggests a likely cause for this, however, in the form of eleven clusters with a very high proportion of disease isolates, greater than 80%, and in some cases consisting entirely of disease isolates. It is, to our knowledge, impossible for any cluster to be composed entirely of disease isolates, and the high proportions of disease isolates in these clusters likely reflects a strong sampling bias for cases of invasive disease, a known issue in the sampling of much of this collection (Chapter 2). Unfortunately, this means that we are unable to say with certainty if there exists or does not exist a correlation between pan-genome size and disease.

Though recombination rate proved to not be correlated with the number of unique genes yet significantly correlated to the pan-genome size M, it would seem very unusual for pan-genome size not to be correlated with the number of unique genes. Figure 5.14 shows a scatter plot of these two variables, and consistent with the appearance of some degree of correlation, the results of a Spearman's rank correlation indicate that the two are significantly correlated, with $r_s = 0.5347$ and $p = 0.0058890$.

Finally, we might wonder what the relationship is between

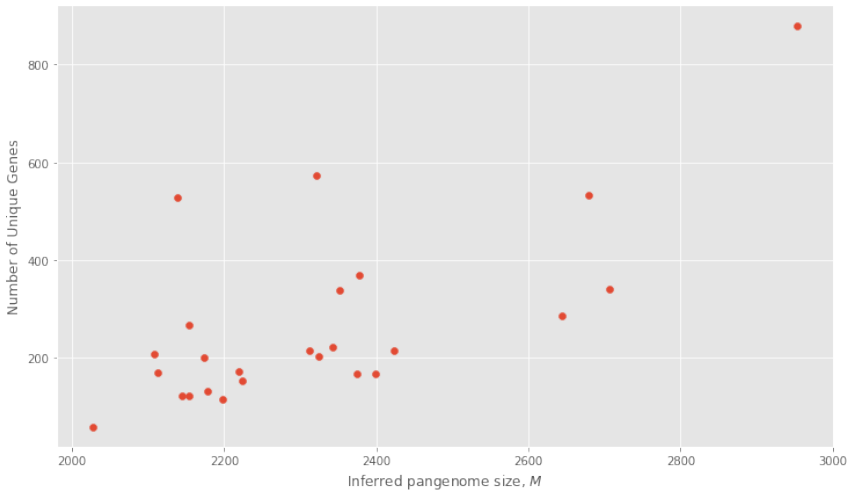


Figure 5.14: Scatter plot of the inferred size of each cluster's pan-genome, M, versus the number of unique genes present within each cluster. Spearman's rank correlation finds a **significant correlation,** $r_s = 0.5347$**,** $p = 0.0058890$**.**

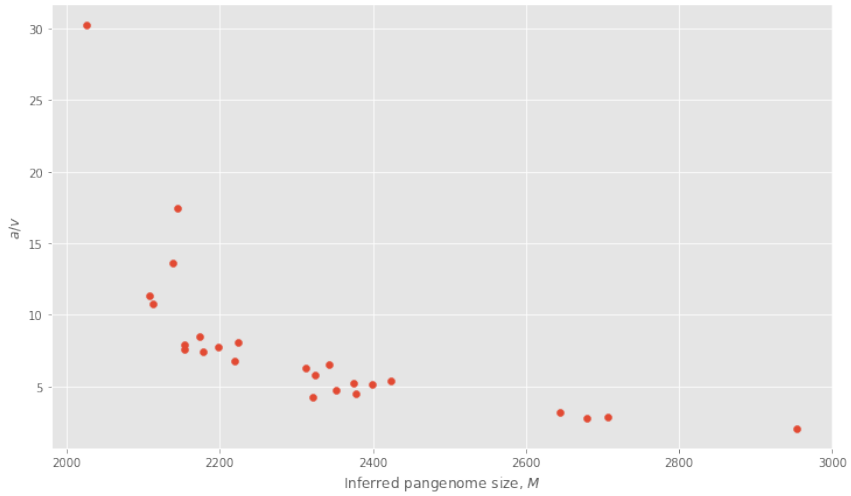


Figure 5.15: Scatter plot of the inferred size of each cluster's pan-genome, M, versus the relative gene gain rate, a/v. Spearman's rank correlation finds a strong, significant negative correlation, $r_s = -0.9362$, $p = 6.331 \times 10^{-12}$.

the relative rate of gene gain and the total size of the pan-genome. Though the intuitive answer would be to suggest that the two ought to be directly proportional, Figure 5.15 shows that the exact opposite is the case. The two are very clearly negatively correlated, and a Spearman's rank correlation confirms that the two are highly correlated and significantly so, with $r_s = -0.9362$ and $p = 6.331 \times 10^{-12}$. This counter-intuitive result again demonstrates the complex and highly variable nature of pan-genome evolution, and indeed how little is truly understood about the long-term evolutionary dynamics at the pan-genome level. This significant negative correlation can only be explained if the overwhelming majority of gene gain in each of the 25 major lineages' pan-genomes are from within *N. meningitidis*, and not elsewhere, as this would lead to those major lineages of *N. meningitidis* which maintained a larger pan-genome not detectably gaining genes due to fact that the genes which could be gained – from other lineages of *N meningitidis* – would already be present. If gene gain in *N. meningitidis* were not primarily from distant lineages within the species, but instead equally or primarily from other species, major lineages of *N. meningitidis* with large pan-genomes should just as readily be able to gain genes from other species as major lineages of *N. meningitidis* with smaller pan-genomes. This would suggest that there should be no correlation between the two variables, unlike what is shown in Figure 5.15, which therefore counter-factually suggests, consistent with other patterns of pan-genome evolution observed in this chapter, that the majority of gene gains within lineages of *N. meningitidis* originate from other *N. meningitidis* lineages.

5.3 Pan-genome association studies

Pan-genome inference reduces a set of diverse bacterial genomes to a set of genes and their presence and absence across the various genomes input into the inference. In this chapter thus far, I have used those data to explore the structure and evolution of the *N. meningitidis* pan-genome, but these data can also be used for another purpose. Like associating the presence or ab-

Gene	Frequency	LRT-adjusted p-value	β
mafB17	0.133	1.810000×10^{-13}	-0.981
pilX/fimA	0.0579	9.910000×10^{-9}	-0.463
group_2782	0.0154	1.710000×10^{-07}	0.418

Table 5.2: Table of the genes in the pan-genome which are significantly associated with ρ/θ recombination rate in a pan-GWAS. LRT-adjusted p-values are p values adjusted with a likelihood ratio test to account for the nested effect of population structure. β reflects the effect size of the variant on the phenotype. Isolates with multiple annotations have both displayed, separated by a forward slash.

sence of nucleotide variants, k-mers, or unitigs with phenotypes present in each isolate (as in section 4.4), it is also possible to perform an association study between the gene presence and absence patterns and isolate phenotypes. This approach, coined pan-GWAS[189], has also been extended to associate structural changes in the pan-genome graph (Section 2.2.2.4) where genes are inserted, deleted, or swapped, with phenotypic outcomes[79]. As such, it is possible to associate the ρ/θ recombination rate phenotypes for each isolate determined in Chapter 4 with both changes in the accessory genomes of isolates as well as with changes in the structure of the pan-genome in isolates. As these results are testing substantially fewer variants than nucleotide polymorphism, k-mer, or unitig-based approaches – only 10626 genes and 48878 structural variants – I use a higher, less stringent, α threshold of 5×10^{-5} to filter out non-significant results.

This significance threshold resulted in 52 genes whose presence or absence was significantly associated with recombination rate, and 81 significant pan-genome structural variants. The effect size of most of these variants is very small, however, with almost all most absolute values of β being below 0.3, and the majority being below 0.1. Filtering significant associations for only those with the absolute value of their effect size coefficient, β greater than 0.3 returns three genes and nine structural variants which are significantly associated with ρ/θ, reported in Table 5.2 and Table 5.3, respectively.

As with any result of an association study, our ability to draw conclusions heavily depends on how well-annotated the variants which were input into the association are. In the case

Structural Variant	Variant Frequency	LRT-adjusted p-value	β
fabF2/dxs/rpmA–fabF2–fabF2	0.1330	$2.060000 \times 10^{-}19$	-1.020
group_1007–*mafB17*–group_1836	0.1330	1.640000×10^{-7}	-0.452
group_2386–group_2385–*ubiH/ubiF*	0.0778	8.050000×10^{-9}	-1.720
pilJ/pilW–pilK–pilX/fimA	0.0575	6.600000^{08}	-0.437
group_1837–group_1879–group_2405	0.1310	6.520000×10^{-8}	-0.333
cycA–group_1880–group_1904	0.0192	1.080000×10^{-8}	-0.868
group_267–*fabF2/dxs/rpmA–fabF2*	0.1330	1.190000×10^{-25}	-1.050
groES–*frpD5*–group_2330	0.0101	1.010000×10^{-05}	-0.339
group_1533–*frpD5*–group_2330	0.0100	2.840000×10^{-05}	-0.336

Table 5.3: Table of the structural variants in the pan-genome graph which are significantly associated with ρ/θ recombination rate in a sv-pan-GWAS. LRT-adjusted p-values are p values adjusted with a likelihood ratio test to account for the nested effect of population structure. β reflects the effect size of the variant on the phenotype. Isolates with multiple annotations have both displayed, separated by a forward slash.

of gene presence-absence, these results are relatively straightforward to interpret as genes are typically well-annotated and their presence/absence is a straightforward binary phenotype. The gene annotated as *mafB17* is the most significant association, with a p-value of 1.810000×10^{-13}, and also the highest absolute value of β among the three significantly associated genes with substantial β, with $\beta = -0.981$, suggesting that the presence of *mafB17* is strongly associated with decreased or lower recombination rates. This is unsurprising as *mafB17* is a gene from the Maf family of genes which are often found in multiple copies in the pathogenic *Neisseria* and have been characterised as a toxin-antitoxin system for inhibiting the growth of proximal bacteria which lack the antitoxin[190, 191]. This neatly explains why the presence of the gene is so strongly associated with a lower recombination rate, the growth-inhibitory effect on bacteria without *mafB17* must prevent bacteria which carry the gene from picking up donor DNA from bacteria which do not have a copy of *mafB17*. Interestingly, *mafB17* is unique to Cluster 2, a cluster with a recombination rate slightly below the average, where it is present in over 99% of isolates, and only absent in five isolates randomly distributed across the phylogeny.

The second gene significantly negatively associated with recombination rate, a variant copy of *pilX* with a p-value of

9.910000×10^{-9} and an effect size of -0.463, is less straightforward to interpret. *pilX* is a type IV pilus gene in *N. meningitidis*, which was previously implicated in cell adhesion and aggregation[192], though it has recently been shown to be periplasmic and hence affect the type IV pilus in a regulatory manner, possibly by affecting its biogenesis and therefore the number of pili expressed on the surface of the cell [193]. The variant of *pilX* strongly associated with a decrease in recombination rate here is found as a core gene in Cluster 6, and at low frequency in Cluster 9. In the remaining clusters, a different copy of *pilX* is present as a core gene. The copy which is predominant in the collection is 204 nucleotides longer than the version negatively associated with recombination rate, and a pairwise alignment confirms that although they share substantial homology in the central region of both proteins, the version negatively associated with recombination rate is truncated and has some nucleotide substitutions at the 5' end of the gene. Given the importance of components of the type IV pilus in the recognition of exogenous *Neisseria* DNA[49] and its uptake[50], the negative association between the truncation of an important type IV pilus regulatory gene and recombination rate is unsurprising.

The final gene significantly associated with recombination rate, and the only one with a positive effect size is also unfortunately the most difficult to interpret. Group_2782, as identified by panaroo, is computationally annotated as a phage-associated protein, and manual annotation does not reveal anything further, other than that it is found in multiple *Neisseria* species. In this collection, it is a core gene in Cluster 17, and present at low frequency in Clusters 1, 3, 6, 11, and 27. The fact that it is present in multiple clusters does suggest that this is not a spurious association caused by the insertion of a phage protein, but without knowing the function of group_2782, it is impossible to comment further on why it is associated with recombination rate.

Unlike genes, and particularly their presence or absence, structural variants are not well-annotated or easy to interpret. They can often represent one of several paths in the pan-genome

graph, which in particular makes their interpretation much less straightforward. In general, it is interesting to note that not a single structural variant, represented by triplet gene patters at forks in the pan-genome graph, is positively associated with recombination rate. Combined with the fact that two of the genes whose presence was significantly associated with recombination also show in structural variants (*mafB17* and *pilX*), it is not clear that in any case it is specifically the variation in the genome structure which is leading to these associations. Manual inspection of the pan-genome graph at the positions of these various structural variants brings no additional clarity, and the graph in all these locations is convoluted and does not present a binary option between two structural variants. As such, the extent of what can be concluded regarding these variants is quite limited. Given the *pilX* variant previously discussed, and the fact that these structural variants are all negatively associated with recombination rate, it seems very likely that these variants reflect a disruption in the pathways responsible for DNA sensing, uptake, or integration.

5.4 Concluding remarks

The pan-genome in *N. meningitidis*, as represented in this collection, consists of 10626 genes. In one sense, this is a surprisingly small number given the extent of the temporal and geographical diversity present in this collection. Collections on the same order of magnitude in other species have found pan-genomes around five times the size of the one here[81], but the number of studies of this scale remain few in number, and of species with very different life-histories, so it is impossible to say how the pan-genome of *N. meningitidis* relates to a mean or median across species. Also notable is the limited number of core genes present in this collection. A decade of research on *N. meningitidis* genomics has found that, in general, isolates possess around two thousand genes[6, 7, 43, 194]. That only 1533 are present in over 95% of isolates in this collection suggests that the true 'core' genome of *N. meningitidis*, is lower than analysis of a single lineage would suggest [195] though the proportion of core

genes in a typical *N. meningitidis* genome remains high compared to other species[81]. This is particularly interesting due to the fact that the accessory genome shared between lineages is relatively small, at 2320 genes, and most of the accessory genome is primarily composed of genes which are unique to clusters – the remaining 6773 genes. Cluster-unique genes make up 74.49% of the accessory genome, and appear to have some overlap in their functionality based on a Gene Ontology enrichment analysis. Given that all these major clusters must have a common ancestor, of which some claim must be relatively recent due to the disease having been first describe medically within recent centuries [196], it is surprising that they have diverged so substantially in terms of their accessory genome content, and this may reflect a more ancient origin for the species, which some evidence has pointed toward [197].

The various relationships between the inferred gene gain rate, loss rate, pangenome size, the number of unique genes, and recombination rate suggests that this has been primarily driven by selective forces, rather than the accumulation of differences over time. This is demonstrated in the significant correlation between recombination rate and pan-genome size, the significant negative correlation between gene gain rate and the number of unique genes, and the lack of any significant correlation between gene gain rate and recombination rate. This pattern suggests that while recombination, pan-genome size, and the number of unique genes are related, as we would expect, what actually governs the rate of change over time cannot covary with those traits. This is consistent with, and explains, the high degree of cluster-specific uniqueness within the accessory genome.

What remains unexplained is how despite this divergence, clusters remain very much distinct within a dimensionally-reduced space of pan-genome content, as shown in Figure 5.5. It is possible that the extent of evolutionary constraint after a major shift in the pan-genome of a lineage is such that there is a relative stability, and major shifts in the pan-genome are rare. Elements of Figure 5.5, are consistent with this view, as many clusters are split into a few tight groupings within the pan-genome space. Another notable feature in Figure 5.5 is

the existence of small numbers of isolates which cluster not with their own cluster (based on PopPUNK core and accessory distances) but rather within the pan-genome space of another cluster. It has recently been suggested that it is possible to use pan-genome patterns to infer evolutionary history in species, such as bacteria, where the genome is so small that multiple-hits and other biases render direct molecular methods unusable[198]. Lineages which are split in their pan-genome space in Figure 5.5 may reflect an early stage in a speciation process into multiple lineages, where a rapidly or abruptly evolving pan-genome diverges into distinct groups, whereas the more gradually changing molecular signal in the core genome shows no sign of such divergence. The lack of congruence in many clusters between their grouping in pan-genome space and isolates which form monophyletic clades on their whole-genome phylogenies in Chapter 3 gives reason to doubt this hypothesis. Alternately, continued gene flow between phylogenetically distinct groups, such as the two main clades which make up Cluster 4, may be what is preventing the two clades from diverging beyond the PopPUNK threshold for delineating lineages. Regardless, as the isolates sampled in each lineage increase, however, and methods which leverage pan-genome data to study relatedness advance, this line of questions could form the next stage of research into bacterial evolution – understanding not only the structure of bacterial populations, but also how that structure has come into being.

CHAPTER 6

Conclusion

This thesis aimed to use a large collection of over fifteen thousand whole-genome sequenced *N. meningitidis* isolates to explore some fundamental questions regarding the evolution of the species. How is its population structured, globally? How is the evolution of the species shaped by recombination and natural selection? How is the genic content of the species structured across the population? These questions are central to a number of important problems regarding the management of the disease, including the design and administration of vaccinations and mass vaccination campaigns, but also the role of disease surveillance in monitoring and responding to potential outbreaks of disease (and how outbreak response should inform vaccine design), strategies for vaccination, and optimal methods of sampling for surveillance. To investigate these questions, I have sequentially undertaken three strands of analysis on these data: First, using PopPUNK to determine a threshold at which to delineate lineages of *N. meningitidis* and build recombination-free whole-genome phylogenies of those lineages to analyse the finer details of the population structure within each lineage. This was done with particular regard for how that structure compares with other data about the lineage, especially each isolates' nation and date of isolation. I then used whole-genome analysis methods to detect recombinations and evidence of selective sweeps within each lineage, comparing lineages both in terms of the rates of recombination on branches of their phylogenies and how that recombination is distributed across the

genome in each lineage, and whether or not that overlaps with evidence of recent directional selection. I separately repeated and extended this analysis in a very densely sampled region to demonstrate that trends in recombination rate, though observed globally, persist even at the local level. Finally, I inferred the pan-genome of the entire global collection, and compared the distribution and patterns of evolution in *N. meningitidis*'s genic content to the population structure determined from their whole-genome sequences.

6.1 Insights into the global population structure

The existence of a determinable population structure in *N. meningitidis*, given its extensive recombination rate, is not guaranteed. Indeed, some have suggested that any signal of structure may not reflect identity by descent but instead reflects differences in recombination rate[199]. Figure 3.2 indicates that in a *N. meningitidis* dataset of this scale and diversity, there is no clear threshold of distance where the distribution of distances is disjoint and therefore can be used to delineate major lineages. This is in contrast to previous work performed on smaller datasets of *N. meningitidis* [80, 105], where using PopPUNK to determine the population structure has always found a clear threshold at which to group isolates into lineages. The problems encountered in such a determination in this dataset are a result of the increased sampling, filling in the gaps which are left when sampling isolates only from a particular region or at a particular time. Despite the non-disjoint distances in Figure 3.2, PopPUNK was still able to determine a threshold which recapitulates many of the main lineages determined using other typing methods [10], and is predominantly congruent with the core genome distance phylogeny, in Figure 3.3. This is due to the network-property based refinement step of PopPUNK (Section 2.2.2.1), which, despite the completely continuous distances present, finds a threshold where across the entire population, the network graph is most well-formed. The existence of such a threshold is consistent with the thresholds identified in previ-

ous work where there was an exclusive disjunction between the distributions of within-lineage distances and between-lineage distances [80, 105]. Though seemingly innocuous, this result demonstrates that despite the extensive recombination in a relatively small genome, there is a detectable and biologically meaningful population structure within the species at the global level, and population structures determined on a local level also reflect the true, global population structure.

The structure of the global collection, as identified by PopPUNK, also reveals something which can only be seen by taking a global perspective on the *N. meningitidis* population structure, namely that it is dominated by a small number of lineages. 25 major clusters, out of a total of 1262 clusters, or strains, make up 78.54% of the population, a considerably lower number of clusters making up a higher proportion of the global population than currently discovered in any of the other species for which a global dataset on the scale of 10^4 isolates has been analysed [81, 82]. Although this may be partially driven by relatively shallow sampling in Asia and South America – excluding the three major US datasets, for example, causes 20 lineages to make up a similar proportion of the population – the difference between *N. meningitidis* and *S. pneumoniae*, where 35 major lineages make up only 62% of the population suggests that even though the two organisms have very similar life-histories, their population structures are not necessarily analogous. The dominance of a small number of lineages and the low frequency of a large number of lineages echoes the results of studies on regional populations of *N. meningitidis* [34, 98, 105], which have repeatedly found that a handful of lineages make up the majority of the population, and a large number of lineages exist at much lower frequency. A global perspective further demonstrates that many of the lineages which exist at low frequency at one time in one region are actually part of a lineage which in a different region or time is dominant. This is not an observation that could be made without the large-scale aggregation of multiple datasets, as presented in this thesis. It also highlights the extent to which migration is consistent and frequent. Only two of the 25 major clusters were restricted to a single continent,

and in many cases, clusters' phylogenies contained many clades that had a mosiac of geographical origins, even among closely related isolates sampled within a few years of one another. This suggests not only that presently, migration over great distances is commonplace in *N. meningitids*, but also that it happens over short timescales, as we might expect given the speed and convenience of modern travel.

The low-frequency isolates found in various parts of the world that are part of dominant lineages elsewhere also indicates the nature of *N. meningitidis* population dynamics, which appears to consist of many lineages evolving at low frequency, some of which expand in a specific region, before population collapse though a bottleneck, whereupon the lineage becomes undetectable or once again exists at a low frequency. This view of *N. meningitidis* population dynamics, with recurrent clonal expansions and population bottlenecks, has long been the accepted view of population dynamics in *N. meningitidis* [200, 201], however even in more rigorous formulations of this model [55] the primary subject of study remains the clonal expansion followed by a population bottleneck[10], rather than how the lineage evolves before the expansion and after the bottleneck. The number of minor lineages, along with the existence of many distant outgroups and minor clades within the phylogenies of most of the 25 major clusters suggests that the persistence of lineages between large-scale clonal expansion and subsequent bottleneck plays an equal role in generating the population structure of *N. meningitidis* as the cycles of expansion and bottlenecks. Estimates of the carriage rate of *N. meningitidis* in the general population range from around 2% to 20% in certain age groups [88–90], which, if consistently true around the globe, suggests that the number of active infections at any any given time is on the order of 10^9. Even if 80% of the population is dominated by a small number of lineages – around 2% of all lineages in this dataset, though it is possible that the true percentage may be slightly higher due to the lack of sampling in some regions (Figure 2.3) – the remaining lineages still constitute an enormous population in absolute terms. Minor lineages may expand and become major lineages, and major lineages may be displaced, as

we often see regionally [201]. The persistence of minor lineages in the population is a necessary precursor for their expansion, and which evolutionary forces govern the maintenance of such diversity at low frequency – if it is indeed maintained over long time periods – is an important open question. Addressing this question of what causes lineages of *N. meningitids* to persist at low frequency without expanding will be essential to establishing a complete understanding of the population dynamics and structure in not only *N. meningitidis*, but all bacteria with similar life-histories and population structure, and may give insight into optimal strategies for managing the diseases they cause.

6.2 Recombination, Selection, and Evolution in *N. meningitidis*

Neisseria, including *N. meningitidis* have long been known to be extremely recombinant relative to many other bacteria[47], so much so that it has been suggested in the past that they are "freely recombining" [55, 56] or "panmictic" [54]. The existence of a detectable clonal population structure has generally put such views to rest in the case of *N. meningitidis*[10], and I find good evidence to suggest that is accurate, as discussed in the previous section. Despite that evidence of a clonal population structure, assessing the level of recombination in the major lineages of this collection also confirms the fact that recombination is in general very frequent, occasionally occurring more than six times as frequently as mutation in the phylogenetic branch leading to a single isolate (Figure 4.2). Most notable in this global population, however, is the extent to which recombination – as a phenotype in its own right – is diverse. This diversity takes two forms – the rate of recombination is significantly different between lineages, (Figure 4.3, Figure 4.5), and also the distribution of recombination hotspots across the genome is different in each major lineage (Section 4.2, Figure 4.35, Figure 4.36). These differences are caused, in part, by many different genetic factors, including genes and repeat sequences, which affect the recombination rate of *N. meningitidis* (This is described fully in

Section 4.4 and Section 5.3). Most of these loci have very small effect sizes, however, and, taken together, these results suggest that recombination rate, as a phenotype changes in response to evolutionary forces, but not in a simple stepwise fashion like other traits of interest such as antimicrobial resistance. The fact that most genetic variants associated with recombination rate are negatively associated with it (Section 4.4, Table 5.2, Table 5.3) and in one well-annotated case represent the truncation and potential loss of function in a gene (*pilX*, section 5.3) there may be a strong selective pressure for *N. meningitidis* to maintain a high level of recombination, and most genetic variation which affects recombination rate reflects deleterious mutation. Recently, considerable interest has developed in studying the evolutionary effects of variation in recombination rate [202, 203], although in many contexts this is complicated by the difficulty associated with estimating differences in recombination rate [202]. *N. meningitidis* provides a clear example of a species in which such genetically-caused differences exist, including toxin antitoxin systems and pilin genes (Section 5.3), and lead to significant differences between different lineages in the species (Figure 4.2. As the amount of *N. meningitidis* genomic data continues to increase, and the major lineages are studied in more detail, it will offer an opportunity to further explore the consequences of variation in recombination rate. Finally, the existence of such significant differences between different lineages of *N. meningitidis* in terms of their recombination rate provides another data point for assessing the relative risk of a lineage to cause potential outbreak – with more recombinant lineages seemingly more likely to escape intervention though a vaccination campaign, as has happened in other species [41], and possibly go on to cause an outbreak of disease.

The fact that recombination in *Neisseria* is variable, and yet generally persists at such a high level that it occurs approximately $1/10^{\text{th}}$ as frequently as mutation, (Figure 4.4) and has been shown to contribute more diversity in terms of nucleotide variants than mutation [75] is unexpected given that the maintenance of the various adaptations to maintain it is costly [47]. This problem, the evolution and persistence of recombination

was a significant historical conundrum in evolutionary theory. It has been largely been resolved via a number of different theoretical mechanisms, including: avoiding the accumulation of deleterious mutations [84], facilitating the efficacy of directional selection [85], and the generation of additional diversity upon which selection can act [86], among others. Studies in bacteria have traditionally favoured the generation of diversity at key loci as the primary reason for the evolutionary maintenance of recombination [87], though evidence also exists for the prevention of the accumulation deleterious mutations [204, 205]. The in-depth analysis of recombination between lineages in the Burkina Faso isolates finds good evidence to suggest that, in general, recombination in *N. meningitidis* acts to prevent the accumulation of deleterious mutations (Figure 4.39), though in the same analysis I also find an example of recombination facilitating the action of positive selection. Despite not being conclusive evidence, the overlap in genes between those found in recombination hotspots and those containing nucleotide variants which have recently undergone a selective sweep (Table 4.1) suggests that this phenomenon may be more widespread. Finally, the analysis of recombination within the Burkina Faso population also confirms that substantial recombination occurs between the distantly-related lineages which make up the local population (Figure 4.7) and that the primary determinant of the likelihood of being involved in a recombination event, particularly as the DNA donor, is the relative population size of the lineage within the local population at that time. Lineages which are undergoing a population expansion, as per the model of population dynamics suggested by the structures observed in Chapter 3, are therefore likely to accumulate diversity from distant lineages at minor frequency in the population while their relative population size remains large. Carriage surveillance at a high enough depth of sampling to be able to identify the makeup of a local *N. meningitidis* population could prove instrumental in predicting what lineages are more likely to receive diverse recombinations from distantly related lineages existing in the population at minor frequency, and therefore be more likely to go on to cause disease (Figure 4.9).

6.3 The pan-genome perspective on *N. meningitidis* population structure and evolution

Inferring the pan-genome of the entire global *N. meningitidis* collection allows for a different perspective on the interaction between species, by directly comparing patterns of gene presence/absence in lineages across the collection, instead of analysing each lineage independently against different reference genomes. This approach highlights the importance of stochastic factors, along with selection and drift in determining the realised extent of gene transfer between lineages. The fact that the inferred relative gene gain rates for each cluster are not at all correlated with the relative recombination rates determined for each cluster in Chapter 4 (Figure 5.8) in particular demonstrates how the likelihood of taking up and integrating a gene – the process of which we assume must be governed primarily by a bacterium's effective 'recombination rate' – does not actually influence whether or not the lineage as a whole will gain a gene in its pan-genome. This suggests that observed large-scale changes in the gene content of a lineage, such as the acquisition of alleles from *N. gonorrhoeae* observed in Cluster 10 [70], the acquisition of capsule and virulent genes in non-groupable/capsule null isolates [206], or the substantial capsule switching observed in Cluster 19 of this collection (Figure 3.25) are likely to be the result of directional selection. In the latter case, this is further evidenced by the presence of several nucleotide variants within the capsular transport genes in the highly significant results of the spydrpick method for discovering co-evolving genes (Table 4.1). Some previous research has given selection a privileged role in the evolution of *N. meningitidis*, with selective pressures from the immune system and inter-lineage competition, implicated in the maintenance of population structure in *N. meningitidis* [73]. Although more recent work emphasises the contribution of other factors [55], this lack of correlation between gene gain rate and recombination rate suggests that in general, the existence of recombination events large enough to affect the pan-genome that

are detectable in our data are driven primarily by selection and potentially drift, and are not routine, stochastic recombination events.

The pan-genome of the entire global *N. meningitidis* collection also offers a different perspective on the relationships between isolates within lineages – many of which form clusterings that do not match the relationships set out in their whole-genome phylogenies (Figure 5.5). In some lineages which contain deep phylogenetic divergences, such as Cluster 4 (Figure 3.10), the phylogenetic tree and the clustering in pan-genome space are correlated, but contain a signififcant number of exceptions. A reduction in gene flow and horizontal gene transfer in bacteria has been repeatedly identified as a prerequisite to speciation [83, 207, 208], and the disparate pan-genome clustering of isolates, which fall under the PopPUNK lineage definition but into separate monophyletic clades in their cluster phylogenies, may reflect the early stages of the of a single cluster of *N. meningitidis* speciating into two. Models also predict that in addition to reduced or restricted gene flow, the process of speciation must also be driven by directional selection – positive or negative – acting on the diverged isolates [83, 208]. Of the five clusters we observe with recent evidence of a major selective sweep (Table 4.1) four have most recent common ancestors dated unreliably in the distant past (Figure 3.32), and contain deep divergences in the date of the tree. How new lineages arise from old lineages in *N. meningitidis* remains an open problem, and the suggestion has been made that it is largely driven by genetic drift [209]. The results of this work do not rule out that possibility as there exist clusters which have deep divergences yet no evidence of a recent selective sweep (Cluster 4, Figure 3.10). Cluster 4's phylogeny is one of the most deeply split between two sister clades and its estimated most recent common ancestor is beyond the threshold of reliable estimation. However, the overlap between clusters with strong evidence of recent selection (Table 4.1, Clusters 3, 11, 15, 19, 22) and those clusters with deep divergences between isolates rendering the estimated dates of their most recent common ancestor in the distant past (Figure 3.32, Clusters 3, 4, 8, 15, 19, 22), together

with the fact that these clusters all show some divergence in the projection of their pan-genome distances into two-dimensional space (Figure 5.5) suggests that a reduction in gene flow and episodes of directional selection may be sufficient, though not necessary, for the creation of new lineages in *N. meningitidis*, and possibly in the genus *Neissiera* as a whole. The number of unique genes in the pan-genome of each cluster is consistent with this view, with each of the 25 main clusters' pan-genome containing an average of 270.92 unique genes, suggesting that barriers to gene flow are present. The further fact that these unique genes are enriched for DNA-binding genes in four major clusters (Table 5.1), and previous research has shown that lineages often contain unique restriction-modification systems [103], reinforces the view that major lineages of *N. meningitidis* can diverge through reductions in gene flow – possibly due to active mechanisms to reduce it [103] – and selection.

6.4 Consequences for the management of meningicoccal disease

Although *N. meningitidis* is an interesting organism for study in its own right, a significant driver of the research interest in the species, particularly the widespread use of whole-genome sequencing, is its ability to inform important decisions regarding the public health management and medical treatment of invasive meningococcal disease [10]. Understanding the evolution of a pathogen is of fundamental importance to designing an effective vaccination strategy in the long-term, particularly in species where mass vaccination is not likely to lead to rapid extinction [92]. *N. meningitidis*, which is primarily controlled worldwide through vaccination [210], neatly fits these criteria. Currently available vaccinations cover five of the capsular groups defined in *N. meningitidis* [17, 153] and there is growing concern that mass vaccination campaigns which target a subset of all capsules may result in an increase in disease caused by serogroups which are not targeted by the vaccine [211], a phenomenon which has been clearly observed in another meningitis-causing bacteria, *S. pnuemoniae* [41, 92]. The results of this thesis – particularly

the insights into the population dynamics of *N. meningitidis* and the findings regarding recombination rate and gene flow between lineages – further our understanding of how the global *N, meningitidis* population is likely to respond to continued mass vaccination campaigns using vaccines which target a subset of the capsule diversity. The existence of an enormous well of diversity which is not routinely sampled, combined with frequent migration and the rapid regional expansion and bottlenecking of lineages highlights the substantial risk that mass vaccination campaigns will cause a shift in the lineages which cause disease if a strong negative selection pressure – such as a vaccine – is applied against a subset of lineages. This has already been observed in parts of the meningitis belt, where vaccination with a vaccine targeting serogroup A *N. meningitidis* was co-incident with an expansion in the disease caused by *N. meningitidis* isolates from other serogroups [212–215], and also in the recent detection of cases of invasive disease caused by non-groupable [38, 39] and serogroup E isolates [37] elsewhere in the world. In contrast to lineage replacement, the results of this thesis also suggest that the switching of a capsule within a lineage is rare. Across the 25 major clusters, all of which span at least a decade of sampling, no more than a half dozen genogroup-switching events were observed in any given lineage (Figures 3.7-3.31). However, an important caveat is that, as discussed above, capsule switching events appear to be strongly driven by selection. Mass vaccination campaigns which cover only a subset of capsule locus types will produce a strong selection pressure against specific variants. This pressure may lead to a higher rate of capsule switching in the future compared with what has been observed in this collection, in a similar vein to what has been observed in *S. pneumoniae* [42]. Though we cannot be certain of the cause of the capsule variation in Cluster 19, it demonstrates that more frequent capsule switching is possible. Finally, an important theme across all three results chapters in this thesis has been how the existence of minor, low-frequency lineages and their interaction with major lineages is an important determinant in *N. meningitidis* evolution. The study of the Burkina Faso population confirms that as lineages increase in size, their

interactions with minor lineages increases (Figure 4.7), lending further support to a vaccine strategy which rapidly responds to the growth of the population size of a lineage, or an outbreak, with a vaccination campaign in the affected area [216–218]. This will not only directly reduce the spread of the lineage and cases of invasive disease, but also reduce the chances of encountering *N. meningitidis* from distantly related lineages. This will also reduce the ability of any potential outbreak lineage to recombine and gain genes, ultimately reducing the speed at which a lineage can adapt to different selection pressures. Indeed, if routine carriage surveillance were to become the norm, assuming there were no significant advances in vaccine technology, a sensible strategy would be to design vaccination campaigns to target lineages undergoing rapid increases in population size, regardless of whether or not that lineage had been previously implicated in causing disease. All of these considerations are particularly true of the protein-based serogroup B vaccines, as the available vaccines are protein-based, and do not target the capsule directly [219], but instead variants of surface-exposed proteins common in serogroup B lineages. Adaptation in proteins targeted by the serogroup B vaccines could potentially lead to vaccine evasion even without a large-scale gene gain event changing the capsule type. Being able to reformulate and vaccinate against any potential escape lineages with a serogroup B capsule may be necessary for long-term control of serogroup B invasive disease.

Although *N. meningitidis* is an interesting organism for study in its own right, a significant driver of the research interest in the species, particularly the widespread use of whole-genome sequencing, is its ability to inform important decisions regarding the public health management and medical treatment of invasive meningococcal disease [10]. Understanding the evolution of a pathogen is of fundamental importance to designing an effective vaccination strategy in the long-term, particularly in species where mass vaccination is not likely to lead to rapid extinction [92]. *N. meningitidis*, which is primarily controlled worldwide through vaccination [210], neatly fits these criteria. Currently available vaccinations cover five of the capsular groups defined in *N. meningitidis* [17, 153] and there is growing concern that

mass vaccination campaigns which target a subset of all capsules may result in an increase in disease caused by serogroups which are not targeted by the vaccine [211], a phenomenon which has been clearly observed in another meningitis-causing bacteria, *S. pnuemoniae* [41, 92]. The results of this thesis – particularly the insights into the population dynamics of *N. meningitidis* and the findings regarding recombination rate and gene flow between lineages – further our understanding of how the global *N, meningitidis* population is likely to respond to continued mass vaccination campaigns using vaccines which target a subset of the capsule diversity. The existence of an enormous well of diversity which is not routinely sampled, combined with frequent migration and the rapid regional expansion and bottlenecking of lineages highlights the substantial risk that mass vaccination campaigns will cause a shift in the lineages which cause disease if a strong negative selection pressure – such as a vaccine – is applied against a subset of lineages. This has already been observed in parts of the meningitis belt, where vaccination with a vaccine targeting serogroup A *N. meningitidis* was co-incident with an expansion in the disease caused by *N. meningitidis* isolates from other serogroups [212–215], and also in the recent detection of cases of invasive disease caused by non-groupable [38, 39] and serogroup E isolates [37] elsewhere in the world. In contrast to lineage replacement, the results of this thesis also suggest that the switching of a capsule within a lineage is rare. Across the 25 major clusters, all of which span at least a decade of sampling, no more than a half dozen genogroup-switching events were observed in any given lineage (Figures 3.7-3.31). However, an important caveat is that, as discussed above, capsule switching events appear to be strongly driven by selection. Mass vaccination campaigns which cover only a subset of capsule types will produce a strong selection pressure against specific variants. This pressure may lead to a higher rate of capsule switching in the future compared with what has been observed in this collection, in a similar vein to what has been observed in *S. pneumoniae* [42]. Though we cannot be certain of the cause of the capsule variation in Cluster 19, it demonstrates that more frequent capsule switching is possible. Finally, an

important theme across all three results chapters in this thesis has been how the existence of minor, low-frequency lineages and their interaction with major lineages is an important determinant in *N. meningitidis* evolution. The study of the Burkina Faso population confirms that as lineages increase in size, their interactions with minor lineages increases (Figure 4.7), lending further support to a vaccine strategy which rapidly responds to the growth of the population size of a lineage, or an outbreak, with a vaccination campaign in the affected area [216–218]. This will not only directly reduce the spread of the lineage and cases of invasive disease, but also reduce the chances of encountering *N. meningitidis* from distantly related lineages. This will also reduce the ability of any potential outbreak lineage to recombine and gain genes, ultimately reducing the speed at which a lineage can adapt to different selection pressures. Indeed, if routine carriage surveillance were to become the norm, assuming there were no significant advances in vaccine technology, a sensible strategy would be design vaccination campaigns to target lineages undergoing rapid increases in population size, regardless of whether or not that lineage had been previously implicated in causing disease. All of these considerations are particularly true of the protein-based serogroup B vaccines, which do not target the capsule directly, but instead variants of surface-exposed proteins common in serogroup B lineages. Adaptation in proteins targeted by the serogroup B vaccines could potentially lead to vaccine evasion even without a large-scale gene gain event.

The field of biology has been enormously enriched by the availability of whole genome sequencing and genomic methods to researchers, leading some pundits to make claims that the 21st century would be the "century of biology" [220]. Just over 20 years after the first draft of the human genome was announced, there is good reason to believe that, eighty years from now, this optimistic projection will indeed be realised. In terms of research into *N. meningitidis*, particularly its genomics, the past 20 years have seen the number of whole-genome sequenced bacteria rise by a factor of roughly 10,000. Studying the genomics of bacteria at this scale – on the order of magnitude of 10^4 – is increasingly common [81, 82], but method development has not yet taken

maximum advantage of this new scale of data. Most methods, including those used in this thesis, rely upon partitioning large-scale datasets into smaller, monophyletic chunks which are then analysed using methods capable of handling that smaller scale of data. Analytical methods which are capable of utilising datasets of this scale to their full potential could lead to the direct study of problems which currently remain just out of reach, such as the relationships between the major lineages of a global population of bacteria. The continued and increasing availability of affordable whole-genome sequencing should allow these large-scale datasets to continue to grow, and hopefully, become more representative of the global population. Over time, the accumulation of sequencing projects will create a significantly longer time period of relatively deep sampling, which has only become possible in the past ten years (Figure 2.4). This will allow for the direct study of the global population dynamics of bacterial species like *N. meningitidis*, which is currently generally infeasible yet has extremely important consequences for the control of the diseases they cause. Datasets of this scale will also allow for additional testing of theories regarding important characteristics of their evolution, particularly concerning the mechanisms of speciation [198, 208], and the nature of selection within and between lineages. In *N. meningitidis* specifically, future large-scale genomic research should aim to take into account the entire genus. It is well-known that bacteria of the genus *Neisseria* frequently exchange DNA [70, 178, 179, 221], yet they are studied largely from the perspective of individual species. Methods which are able to scale to 10^4 isolates [79, 80] should be applied at a genus-level perspective to study the relationships between species, as well as to investigate their evolutionary history. In a similar vein, research which takes into account the entire microbiome of the nasopharynx has the potential to significantly increase our understanding of *N. meningitidis* transmission and disease, although novel methods would need to be developed for these analyses. The importance of continued study into the evolution of *N. meningitidis* and global surveillance of cases cannot be understated. Surveillance on a regional level has already been crucial in guiding the response to epidemic outbreaks of dis-

ease [32], and its utility in this regard will continue to grow. Although it is possible that vaccine technology may one day prevent any and all infection of humans with *N. meningitidis*, the massive diversity present within *N. meningitidis* renders this a remote possibility. Developing continuously effective vaccination strategies relies upon understanding what is out there and how it is changing. Finally, all of this research will not be possible, and this thesis would not have been possible, without the aggressive public data-sharing of those involved in whole-genome sequencing *N. meningitidis* worldwide. Efforts are underway to co-ordinate such efforts in the future [222], and the extent to which the systems which emerge are able to efficiently handle such data at scale will enormously affect what is possible in the future.

Bibliography

[1] E Marchiafava and A Celli. Spra i micrococchi della meningite cerebrospinale epidemica. *Gazz degli Ospedali*, 5:59, 1884.

[2] Anton Weichselbaum. *Ueber die Aetiologie der akuten Meningitis cerebro-spinalis*. na, 1887.

[3] Charles Darwin. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859. or the Preservation of Favored Races in the Struggle for Life.

[4] Ernst Mayr. *The growth of biological thought: Diversity, evolution, and inheritance*. Harvard University Press, 1982.

[5] Ronald A Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[6] H. Tettelin, N. J. Saunders, J. Heidelberg, A. C. Jeffries, K. E. Nelson, J. A. Eisen, K. A. Ketchum, D. W. Hood, J. F. Peden, R. J. Dodson, W. C. Nelson, M. L. Gwinn, R. DeBoy, J. D. Peterson, E. K. Hickey, D. H. Haft, S. L. Salzberg, O. White, R. D. Fleischmann, B. A. Dougherty, T. Mason, A. Ciecko, D. S. Parksey, E. Blair, H. Cittone, E. B. Clark, M. D. Cotton, T. R. Utterback, H. Khouri, H. Qin, J. Vamathevan, J. Gill, V. Scarlato, V. Masignani, M. Pizza, G. Grandi, L. Sun, H. O. Smith, C. M. Fraser, E. R. Moxon, R. Rappuoli, and J. C. Venter. Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. *Science*, 287(5459):1809–1815, Mar 2000.

[7] J. Parkhill, M. Achtman, K. D. James, S. D. Bentley, C. Churcher, S. R. Klee, G. Morelli, D. Basham, D. Brown,

T. Chillingworth, R. M. Davies, P. Davis, K. Devlin, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Leather, S. Moule, K. Mungall, M. A. Quail, M. A. Rajandream, K. M. Rutherford, M. Simmonds, J. Skelton, S. Whitehead, B. G. Spratt, and B. G. Barrell. Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491. *Nature*, 404(6777):502–506, Mar 2000.

[8] W. P. Hanage. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harb Perspect Biol*, 8(7), 07 2016.

[9] B. J. Shapiro, L. A. David, J. Friedman, and E. J. Alm. Looking for Darwin's footprints in the microbial world. *Trends Microbiol*, 17(5):196–204, May 2009.

[10] D. A. Caugant and O. B. Brynildsrud. Neisseria meningitidis: using genomics to understand diversity, evolution and pathogenesis. *Nat Rev Microbiol*, 18(2):84–96, 02 2020.

[11] S. N. Ladhani, J. Lucidarme, S. R. Parikh, H. Campbell, R. Borrow, and M. E. Ramsay. Meningococcal disease and sexual transmission: urogenital and anorectal infections and invasive disease due to Neisseria meningitidis. *Lancet*, 395(10240):1865–1877, 06 2020.

[12] N. E. Rosenstein, B. A. Perkins, D. S. Stephens, T. Popovic, and J. M. Hughes. Meningococcal disease. *N Engl J Med*, 344(18):1378–1388, May 2001.

[13] M. S. Edwards and C. J. Baker. Complications and sequelae of meningococcal infections in children. *J Pediatr*, 99(4):540–545, Oct 1981.

[14] K. J. Olbrich, D. MÃijller, S. Schumacher, E. Beck, K. Meszaros, and F. Koerber. Systematic Review of Invasive Meningococcal Disease: Sequelae and Quality of Life Impact on Patients and Their Caregivers. *Infect Dis Ther*, 7(4):421–438, Dec 2018.

[15] A. Vyse, A. Anonychuk, A. JÃďkel, H. Wieffer, and S. Nadel. The burden and impact of severe and long-term sequelae of meningococcal disease. *Expert Rev Anti Infect Ther*, 11(6):597–604, Jun 2013.

[16] S. R. Parikh, H. Campbell, J. A. Bettinger, L. H. Harrison, H. S. Marshall, F. Martinon-Torres, M. A. Safadi, Z. Shao, B. Zhu, A. von Gottberg, R. Borrow, M. E. Ramsay, and S. N. Ladhani. The everchanging epidemiology of meningococcal disease worldwide and the potential for prevention through vaccination. *J Infect*, 81(4):483–498, 10 2020.

[17] Odile B Harrison, Heike Claus, Ying Jiang, Julia S Bennett, Holly B Bratcher, Keith A Jolley, Craig Corton, Rory Care, Jan T Poolman, Wendell D Zollinger, et al. Description and nomenclature of neisseria meningitidis capsule locus. *Emerging infectious diseases*, 19(4):566, 2013.

[18] E. C. Gotschlich, T. Y. Liu, and M. S. Artenstein. Human immunity to the meningococcus. 3. Preparation and immunochemical properties of the group A, group B, and group C meningococcal polysaccharides. *J Exp Med*, 129(6):1349–1365, Jun 1969.

[19] J. M. Griffiss, B. L. Brandt, P. L. Altieri, G. B. Pier, and S. L. Berman. Safety and immunogenicity of group Y and group W135 meningococcal capsular polysaccharide vaccines in adults. *Infect Immun*, 34(3):725–732, Dec 1981.

[20] D. Toneatto, S. Ismaili, E. Ypma, K. Vienken, P. Oster, and P. Dull. The first use of an investigational multicomponent meningococcal serogroup B vaccine (4CMenB) in humans. *Hum Vaccin*, 7(6):646–653, Jun 2011.

[21] P. H. Mäkelä, H. Käyhty, P. Weckström, A. Sivonen, and O. V. Renkonen. Effect of group-A meningococcal vaccine in army recruits in Finland. *Lancet*, 2(7941):883–886, Nov 1975.

[22] R. Borrow, M. K. Taha, M. M. Giuliani, M. Pizza, A. Banzhoff, and R. Bekkat-Berkani. Methods to evaluate serogroup B meningococcal vaccines: From predictions to real-world evidence. *J Infect*, 81(6):862–872, 12 2020.

[23] Caroline L Trotter, Clément Lingani, Katya Fernandez, Laura V Cooper, André Bita, Carol Tevi-Benissan, Olivier Ronveaux, Marie-Pierre Préziosi, and James M Stuart. Impact of menafrivac in nine countries of the african meningitis belt, 2010–15: an analysis of surveillance data. *The Lancet infectious diseases*, 17(8):867–872, 2017.

[24] David S Stephens, Brian Greenwood, and Petter Brandtzaeg. Epidemic meningitis, meningococcaemia, and neisseria meningitidis. *The Lancet*, 369(9580):2196–2210, 2007.

[25] World Health Organization. Defeating meningitis by 2030: a global road map, 2020.

[26] Y. L. Tzeng and D. S. Stephens. Epidemiology and pathogenesis of Neisseria meningitidis. *Microbes Infect*, 2(6):687–700, May 2000.

[27] Anna M Molesworth, Madeleine C Thomson, Stephen J Connor, Mark P Cresswell, Andrew P Morse, Paul Shears, C Anthony Hart, and Luis E Cuevas. Where is the meningitis belt? defining an area at risk of epidemic meningitis in africa. *Transactions of the royal society of tropical medicine and hygiene*, 96(3):242–249, 2002.

[28] Caroline L Trotter and Brian M Greenwood. Meningococcal carriage in the african meningitis belt. *The Lancet infectious diseases*, 7(12):797–803, 2007.

[29] J. M. Okwo-Bele, F. M. LaForce, R. Borrow, and M. P. Preziosi. Documenting the Results of a Successful Partnership: A New Meningococcal Vaccine for Africa. *Clin Infect Dis*, 61 Suppl 5:S389–390, Nov 2015.

[30] D. M. Daugla, J. P. Gami, K. Gamougam, N. Naibei, L. Mbainadji, M. NarbÃĭ, J. Toralta, B. Kodbesse,

C. Ngadoua, M. E. Coldiron, F. Fermon, A. L. Page, M. H. Djingarey, S. Hugonnet, O. B. Harrison, L. S. Rebbetts, Y. Tekletsion, E. R. Watkins, D. Hill, D. A. Caugant, D. Chandramohan, M. Hassan-King, O. Manigart, M. Nascimento, A. Woukeu, C. Trotter, J. M. Stuart, M. Maiden, and B. M. Greenwood. Effect of a serogroup A meningococcal conjugate vaccine (PsA-TT) on serogroup A meningococcal meningitis and carriage in Chad: a community study [corrected]. *Lancet*, 383(9911):40–47, Jan 2014.

[31] Fabien VK Diomandé, Mamoudou H Djingarey, Doumagoum M Daugla, Ryan T Novak, Paul A Kristiansen, Jean-Marc Collard, Kadidja Gamougam, Denis Kandolo, Nehemie Mbakuliyemo, Leonard Mayer, et al. Public health impact after the introduction of PsA-TT: the first 4 years. *Clinical Infectious Diseases*, 61(suppl_5):S467–S472, 2015.

[32] N. Topaz, D. A. Caugant, M. K. Taha, O. B. Brynildsrud, N. Debech, E. Hong, A. E. Deghmane, R. OuÃľdraogo, S. Ousmane, K. Gamougame, B. M. Njanpop-Lafourcade, S. Diarra, L. M. Fox, and X. Wang. Phylogenetic relationships and regional spread of meningococcal strains in the meningitis belt, 2011-2016. *EBioMedicine*, 41:488–496, Mar 2019.

[33] R. Borrow, D. A. Caugant, M. Ceyhan, H. Christensen, E. C. Dinleyici, J. Findlow, L. Glennie, A. Von Gottberg, A. Kechrid, J. VÃązquez Moreno, A. Razki, V. Smith, M. K. Taha, H. Tali-Maamar, and K. Zerouali. Meningococcal disease in the Middle East and Africa: Findings and updates from the Global Meningococcal Initiative. *J Infect*, 75(1):1–11, 07 2017.

[34] Ola Brønstad Brynildsrud, Vegard Eldholm, Adelina Rakhimova, Paul A Kristiansen, and Dominique A Caugant. Gauging the epidemic potential of a widely circulating non-invasive meningococcal strain in africa. *Microbial genomics*, 5(8), 2019.

[35] B. Wang, R. Santoreneos, L. Giles, H. Haji Ali Afzali, and H. Marshall. Case fatality rates of invasive meningococcal disease by serogroup and age: A systematic review and meta-analysis. *Vaccine*, 37(21):2768–2782, 05 2019.

[36] O. Xie, A. J. Pollard, J. E. Mueller, and G. Norheim. Emergence of serogroup X meningococcal disease in Africa: need for a vaccine. *Vaccine*, 31(27):2852–2861, Jun 2013.

[37] D. Thangarajah, C. J. D. Guglielmino, S. B. Lambert, G. Khandaker, B. R. Vasant, J. A. Malo, H. V. Smith, and A. V. Jennison. Genomic Characterization of Recent and Historic Meningococcal Serogroup E Invasive Disease in Australia: A Case Series. *Clin Infect Dis*, 70(8):1761–1763, 04 2020.

[38] L. Willerton, J. Lucidarme, H. Campbell, D. A. Caugant, H. Claus, S. Jacobsson, S. N. Ladhani, P. MÃűlling, A. Neri, P. Stefanelli, M. K. Taha, U. Vogel, and R. Borrow. Geographically widespread invasive meningococcal disease caused by a ciprofloxacin resistant non-groupable strain of the ST-175 clonal complex. *J Infect*, 81(4):575–584, 10 2020.

[39] L. A. McNamara, C. C. Potts, A. Blain, N. Topaz, M. Apostol, N. B. Alden, S. Petit, M. M. Farley, L. H. Harrison, L. Triden, A. Muse, T. Poissant, X. Wang, and J. R. MacNeil. Invasive Meningococcal Disease due to Nongroupable Neisseria meningitidis-Active Bacterial Core Surveillance Sites, 2011-2016. *Open Forum Infect Dis*, 6(5):ofz190, May 2019.

[40] W. H. Chen, K. M. Neuzil, C. R. Boyce, M. F. Pasetti, M. K. Reymann, L. Martellet, N. Hosken, F. M. LaForce, R. M. Dhere, S. S. Pisal, A. Chaudhari, P. S. Kulkarni, R. Borrow, H. Findlow, V. Brown, M. L. McDonough, L. Dally, and M. R. Alderson. Safety and immunogenicity of a pentavalent meningococcal conjugate vaccine containing serogroups A, C, Y, W, and X in healthy adults: a phase 1, single-centre, double-blind, randomised, con-

trolled study. *Lancet Infect Dis*, 18(10):1088–1096, 10 2018.

[41] A. B. Brueggemann, R. Pai, D. W. Crook, and B. Beall. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathog*, 3(11):e168, Nov 2007.

[42] N. J. Croucher, J. A. Finkelstein, S. I. Pelton, P. K. Mitchell, G. M. Lee, J. Parkhill, S. D. Bentley, W. P. Hanage, and M. Lipsitch. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet*, 45(6):656–663, Jun 2013.

[43] S. D. Bentley, G. S. Vernikos, L. A. Snyder, C. Churcher, C. Arrowsmith, T. Chillingworth, A. Cronin, P. H. Davis, N. E. Holroyd, K. Jagels, M. Maddison, S. Moule, E. Rabbinowitsch, S. Sharp, L. Unwin, S. Whitehead, M. A. Quail, M. Achtman, B. Barrell, N. J. Saunders, and J. Parkhill. Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet*, 3(2):e23, Feb 2007.

[44] K. Jyssum and S. Jyssum. Variation in density and transformation potential in deoxyribonucleic acid from Neisseria meningitidis. *J Bacteriol*, 90(6):1513–1519, Dec 1965.

[45] C. Johnston, B. Martin, G. Fichant, P. Polard, and J. P. Claverys. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat Rev Microbiol*, 12(3):181–196, Mar 2014.

[46] S. Jyssum and K. Jyssum. Specific uptake of homologous DNA accompanying transformation in Neisseria meningitidis. *Acta Pathol Microbiol Scand B Microbiol Immunol*, 78(2):140–148, 1970.

[47] Kyle P Obergfell and H Steven Seifert. Mobile DNA in the pathogenic neisseria. *Microbiology spectrum*, 3(3), 2015.

[48] H. O. Smith, M. L. Gwinn, and S. L. Salzberg. DNA uptake signal sequences in naturally transformable bacteria. *Res Microbiol*, 150(9-10):603–616, 1999.

[49] A. Cehovin, P. J. Simpson, M. A. McDowell, D. R. Brown, R. Noschese, M. Pallett, J. Brady, G. S. Baldwin, S. M. Lea, S. J. Matthews, and V. Pelicic. Specific DNA recognition mediated by a type IV pilin. *Proc Natl Acad Sci U S A*, 110(8):3065–3070, Feb 2013.

[50] C. Hepp and B. Maier. Kinetics of DNA uptake during transformation provide evidence for a translocation ratchet mechanism. *Proc Natl Acad Sci U S A*, 113(44):12467–12472, 11 2016.

[51] Y. Zhang, N. Heidrich, B. J. Ampattu, C. W. Gunderson, H. S. Seifert, C. Schoen, J. Vogel, and E. J. Sontheimer. Mol CellProcessing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. *Mol Cell*, 50(4):488–503, May 2013.

[52] E. A. Stohl and H. S. Seifert. The recX gene potentiates homologous recombination in Neisseria gonorrhoeae. *Mol Microbiol*, 40(6):1301–1310, Jun 2001.

[53] E. J. Feil, M. C. Maiden, M. Achtman, and B. G. Spratt. The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Mol Biol Evol*, 16(11):1496–1502, Nov 1999.

[54] R. C. Read. Neisseria meningitidis and meningococcal disease: recent discoveries and innovations. *Curr Opin Infect Dis*, 32(6):601–608, 12 2019.

[55] M. Tibayrenc and F. J. Ayala. How clonal are Neisseria species? The epidemic clonality model revisited. *Proc Natl Acad Sci U S A*, 112(29):8909–8913, Jul 2015.

[56] J. M. Smith, N. H. Smith, M. O'Rourke, and B. G. Spratt. How clonal are bacteria? *Proc Natl Acad Sci U S A*, 90(10):4384–4388, May 1993.

[57] M. C. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. Multilocus sequence typing: a portable approach to the

identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*, 95(6):3140–3145, Mar 1998.

[58] Martin CJ Maiden, Jane A Bygraves, Edward Feil, Giovanna Morelli, Joanne E Russell, Rachel Urwin, Qing Zhang, Jiaji Zhou, Kerstin Zurth, Dominique A Caugant, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, 1998.

[59] K. A. Jolley, J. E. Bray, and M. C. J. Maiden. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*, 3:124, 2018.

[60] D. A. Caugant, K. BÃÿvre, P. Gaustad, K. Bryn, E. Holten, E. A. HÃÿiby, and L. O. FrÃÿholm. Multilocus genotypes determined by enzyme electrophoresis of Neisseria meningitidis isolated from patients with systemic disease and from healthy carriers. *J Gen Microbiol*, 132(3):641–652, Mar 1986.

[61] M. C. Maiden. Multilocus sequence typing of bacteria. *Annu Rev Microbiol*, 60:561–588, 2006.

[62] M. C. Maiden. High-throughput sequencing in the population analysis of bacterial pathogens of humans. *Int J Med Microbiol*, 290(2):183–190, May 2000.

[63] J. Jelfs, R. Munro, F. E. Ashto, and D. A. Caugant. Genetic characterization of a new variant within the ET-37 complex of Neisseria meningitidis associated with outbreaks in various parts of the world. *Epidemiol Infect*, 125(2):285–298, Oct 2000.

[64] L. W. Mayer, M. W. Reeves, N. Al-Hamdan, C. T. Sacchi, M. K. Taha, G. W. Ajello, S. E. Schmink, C. A. Noble, M. L. Tondella, A. M. Whitney, Y. Al-Mazrou, M. Al-Jefri, A. Mishkhis, S. Sabban, D. A. Caugant, J. Lingappa, N. E. Rosenstein, and T. Popovic. Outbreak of W135

meningococcal disease in 2000: not emergence of a new W135 strain but clonal expansion within the electophoretic type-37 complex. *J Infect Dis*, 185(11):1596–1605, Jun 2002.

[65] A. C. Retchless, F. Hu, A. S. OuÃľdraogo, S. Diarra, K. Knipe, M. Sheth, L. A. Rowe, L. SangarÃľ, A. Ky Ba, S. Ouangraoua, D. Batra, R. T. Novak, R. OuÃľdraogo TraorÃľ, and X. Wang. The Establishment and Diversification of Epidemic-Associated Serogroup W Meningococcus in the African Meningitis Belt, 1994 to 2012. *mSphere*, 1(6), 2016.

[66] R. Abad, E. L. López, R. Debbag, and J. A. Vázquez. Serogroup W meningococcal disease: global spread and current affect on the Southern Cone in Latin America. *Epidemiol Infect*, 142(12):2461–2470, Dec 2014.

[67] S. Mowlaboccus, K. A. Jolley, J. E. Bray, S. Pang, Y. T. Lee, J. D. Bew, D. J. Speers, A. D. Keil, G. W. Coombs, and C. M. Kahler. Clonal Expansion of New Penicillin-Resistant Clade of Neisseria meningitidis Serogroup W Clonal Complex 11, Australia. *Emerg Infect Dis*, 23(8):1364–1367, 08 2017.

[68] M. Krone, S. Gray, R. Abad, A. SkoczyÅĎska, P. Stefanelli, A. van der Ende, G. Tzanakaki, P. MÃűlling, M. JoÃčo SimÃţes, P. KÅŹÃŋÅ"ovÃą, S. Emonet, D. A. Caugant, M. Toropainen, J. Vazquez, I. WaÅŻko, M. J. Knol, S. Jacobsson, C. Rodrigues Bettencourt, M. Musilek, R. Born, U. Vogel, and R. Borrow. Increase of invasive meningococcal serogroup W disease in Europe, 2013 to 2017. *Euro Surveill*, 24(14), Apr 2019.

[69] M. K. Taha, A. E. Deghmane, M. Knol, and A. van der Ende. Whole genome sequencing reveals Trans-European spread of an epidemic Neisseria meningitidis serogroup W clone. *Clin Microbiol Infect*, 25(6):765–767, Jun 2019.

[70] Adam C Retchless, Cécilia B Kretz, How-Yi Chang, Jose A Bazan, A Jeanine Abrams, Abigail Norris Turner, Laurel T Jenkins, David L Trees, Yih-Ling Tzeng, David S

Stephens, et al. Expansion of a urethritis-associated neisseria meningitidis clade in the united states with concurrent acquisition of n. gonorrhoeae alleles. *BMC genomics*, 19(1):176, 2018.

[71] A. Lamelas, S. R. Harris, K. RÃűltgen, J. P. Dangy, J. Hauser, R. A. Kingsley, T. R. Connor, A. Sie, A. Hodgson, G. Dougan, J. Parkhill, S. D. Bentley, and G. Pluschke. Emergence of a new epidemic Neisseria meningitidis serogroup A Clone in the African meningitis belt: high-resolution picture of genomic changes that mediate immune evasion. *mBio*, 5(5):e01974–01914, Oct 2014.

[72] H. Claus, A. Friedrich, M. Frosch, and U. Vogel. Differential distribution of novel restriction-modification systems in clonal lineages of Neisseria meningitidis. *J Bacteriol*, 182(5):1296–1303, Mar 2000.

[73] C. O. Buckee, K. A. Jolley, M. Recker, B. Penman, P. Kriz, S. Gupta, and M. C. Maiden. Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. *Proc Natl Acad Sci U S A*, 105(39):15082–15087, Sep 2008.

[74] K. A. Jolley, D. J. Wilson, P. Kriz, G. McVean, and M. C. Maiden. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Mol Biol Evol*, 22(3):562–569, Mar 2005.

[75] E. J. Feil, M. C. Enright, and B. G. Spratt. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidis and Streptococcus pneumoniae. *Res Microbiol*, 151(6):465–469, 2000.

[76] E. J. Feil, J. E. Cooper, H. Grundmann, D. A. Robinson, M. C. Enright, T. Berendt, S. J. Peacock, J. M. Smith, M. Murphy, B. G. Spratt, C. E. Moore, and N. P. Day. How clonal is Staphylococcus aureus? *J Bacteriol*, 185(11):3307–3316, Jun 2003.

[77] D. A. Caugant and M. C. Maiden. Meningococcal carriage and disease–population biology and evolution. *Vaccine*, 27 Suppl 2:64–70, Jun 2009.

[78] D. Golparian and M. Unemo. Will Genome Analysis Elucidate Evolution, Global Transmission and Virulence of Neisseria Meningitidis Lineages? *EBioMedicine*, 2(3):186–187, Mar 2015.

[79] Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A Lees, Rebecca A Gladstone, Stephanie W Lo, Christopher Beaudoin, R Andrés Floto, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*, 21(1):180, 07 2020.

[80] John A Lees, Simon R Harris, Gerry Tonkin-Hill, Rebecca A Gladstone, Stephanie W Lo, Jeffrey N Weiser, Jukka Corander, Stephen D Bentley, and Nicholas J Croucher. Fast and flexible bacterial genomic epidemiology with poppunk. *Genome research*, 29(2):304–316, 2019.

[81] G. Horesh, G. A. Blackwell, G. Tonkin-Hill, J. Corander, E. Heinz, and N. R. Thomson. A comprehensive and high-quality collection of Escherichia coli genomes and their genes. *Microb Genom*, 7(2), 02 2021.

[82] R. A. Gladstone, S. W. Lo, J. A. Lees, N. J. Croucher, A. J. van Tonder, J. Corander, A. J. Page, P. Marttinen, L. J. Bentley, T. J. Ochoa, P. L. Ho, M. du Plessis, J. E. Cornick, B. Kwambana-Adams, R. Benisty, S. A. Nzenze, S. A. Madhi, P. A. Hawkins, D. B. Everett, M. Antonio, R. Dagan, K. P. Klugman, A. von Gottberg, L. McGee, R. F. Breiman, and S. D. Bentley. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, 43:338–346, May 2019.

[83] B. J. Shapiro and M. F. Polz. Microbial Speciation. *Cold Spring Harb Perspect Biol*, 7(10):a018143, Sep 2015.

[84] Joseph Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2):737–756, 1974.

[85] N. H. Barton. A general model for the evolution of recombination. *Genet Res*, 65(2):123–145, Apr 1995.

[86] S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nat Rev Genet*, 3(4):252–261, Apr 2002.

[87] M. Vos. Why do bacteria engage in homologous recombination? *Trends Microbiol*, 17(6):226–232, Jun 2009.

[88] L. V. Cooper, P. A. Kristiansen, H. Christensen, A. Karachaliou, and C. L. Trotter. Meningococcal carriage by age in the African meningitis belt: a systematic review and meta-analysis. *Epidemiol Infect*, 147:e228, 01 2019.

[89] Hannah Christensen, Margaret May, Leah Bowen, Matthew Hickman, and Caroline L Trotter. Meningococcal carriage by age: a systematic review and meta-analysis. *The Lancet infectious diseases*, 10(12):853–861, 2010.

[90] S. P. Yazdankhah and D. A. Caugant. Neisseria meningitidis: an overview of the carriage state. *J Med Microbiol*, 53(Pt 9):821–832, Sep 2004.

[91] H. Broutin, S. Philippon, G. Constantin de Magny, M. F. Courel, B. Sultan, and J. F. GuÃľgan. Comparative study of meningitis dynamics across nine African countries: a global perspective. *Int J Health Geogr*, 6:29, Jul 2007.

[92] S. D. Bentley and S. W. Lo. Global genomic pathogen surveillance to inform vaccine strategies: a decade-long expedition in pneumococcal genomics. *Genome Med*, 13(1):84, May 2021.

[93] G. D. Biswas, T. Sox, E. Blackman, and P. F. Sparling. Factors affecting genetic transformation of Neisseria gonorrhoeae. *J Bacteriol*, 129(2):983–992, Feb 1977.

[94] O. Ali, A. Aseffa, A. Bedru, T. Lema, T. Moti, Y. Tekletsion, A. Worku, H. G. Xabher, L. Yamuah, R. M. Boukary,

J. M. Collard, I. D. Dano, I. Habiboulaye, B. Issaka, J. F. Jusot, S. Ousmane, I. Rabe, D. M. Daugla, J. P. Gami, K. Gamougam, L. Mbainadji, N. Naibei, M. NarbÃĭ, J. Toralta, A. Berthe, K. Diallo, M. Keita, U. Onwuchekwa, S. O. Sow, B. Tamboura, A. Traore, A. Toure, T. Clark, L. Mayer, M. Amodu, O. Beida, G. Gadzama, B. Omotara, Z. Sambo, S. Yahya, D. Chandramohan, B. M. Greenwood, M. Hassan-King, O. Manigart, M. Nascimento, J. M. Stuart, A. Woukeu, N. E. Basta, X. Bai, R. Borrow, H. Findlow, S. Alavo, H. Bassene, A. Diallo, M. Dieng, S. DoucourÃĭ, J. F. Gomis, A. Ndiaye, C. Sokhna, J. F. Trape, B. Akalifa, A. Forgor, A. Hodgson, I. Osei, S. L. Quaye, J. Williams, P. Wontuo, T. Irving, C. L. Trotter, J. Bennett, D. Hill, O. Harrison, M. C. Maiden, L. Rebbetts, and E. Watkins. The Diversity of Meningococcal Carriage Across the African Meningitis Belt and the Impact of Vaccination With a Group A Meningococcal Conjugate Vaccine. *J Infect Dis*, 212(8):1298–1307, Oct 2015.

[95] O. Ali, A. Aseffa, A. Bedru Omer, T. s. Lema, T. Moti Demissie, Y. Tekletsion, A. Worku, H. Guebre Xabher, L. Yamuah, R. M. Boukary, J. M. Collard, I. D. Dano, I. Habiboulaye, B. Issaka, J. F. Jusot, S. Ousmane, I. Rabe, D. M. Daugla, J. P. Gami, K. Gamougam, L. Mbainadji, N. Naibei, M. NarbÃĭ, J. Toralta, A. Berthe, K. Diallo, M. Keita, A. Coulibaly, U. Onwuchekwa, S. O. Sow, B. Tamboura, A. Traore, A. Toure, T. Clark, L. Mayer, M. Amodu, O. Beida, G. Gadzama, B. Omotara, S. Zailani, S. h. Yahya, D. Chandramohan, B. M. Greenwood, M. Hassan-King, O. Manigart, M. Nascimento, J. M. Stuart, A. Woukeu, N. E. Basta, X. Bai, R. Borrow, H. Findlow, S. Alavo, H. Bassene, A. Diallo, M. Dieng, S. DoucourÃĭ, J. F. Gomis, A. Ndiaye, C. h. Sokhna, J. F. Trape, A. Bugri, A. Forgor, A. Hodgson, I. Osei, S. L. Quaye, J. Williams, P. Wontuo, T. Irving, C. L. Trotter, A. Karachaliou, J. Bennett, D. Hill, O. Harrison, M. C. Maiden, L. Rebbetts, and E. Watkins. Household transmission of Neisseria meningitidis in the African meningitis

belt: a longitudinal cohort study. *Lancet Glob Health*, 4(12):e989–e995, 12 2016.

[96] Guro K Bårnes, Ola Brønstad Brynildsrud, Bente Børud, Bereket Workalemahu, Paul A Kristiansen, Demissew Beyene, Abraham Aseffa, and Dominique A Caugant. Whole genome sequencing reveals within-host genetic changes in paired meningococcal carriage isolates from ethiopia. *BMC genomics*, 18(1):407, 2017.

[97] M. J. Whaley, S. J. Joseph, A. C. Retchless, C. B. Kretz, A. Blain, F. Hu, H. Y. Chang, S. A. Mbaeyi, J. R. MacNeil, T. D. Read, and X. Wang. Whole genome sequencing for investigations of meningococcal outbreaks in the United States: a retrospective analysis. *Sci Rep*, 8(1):15803, 10 2018.

[98] S. J. Joseph, N. Topaz, H. Y. Chang, M. J. Whaley, J. T. Vuong, A. Chen, F. Hu, S. E. Schmink, L. T. Jenkins, L. D. Rodriguez-Rivera, J. D. Thomas, A. M. Acosta, L. McNamara, H. M. Soeters, S. Mbaeyi, and X. Wang. Insights on Population Structure and Within-Host Genetic Changes among Meningococcal Carriage Isolates from U.S. Universities. *mSphere*, 5(2), 04 2020.

[99] J. A. Lees, P. H. C. Kremer, A. S. Manso, N. J. Croucher, B. Ferwerda, M. V. SerÃşn, M. R. Oggioni, J. Parkhill, M. C. Brouwer, A. van der Ende, D. van de Beek, and S. D. Bentley. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom*, 3(1):e000103, 01 2017.

[100] A. Lamelas, J. Hauser, J. P. Dangy, A. M. Hamid, K. RÃűltgen, M. R. Abdul Sater, A. Hodgson, A. Sie, T. Junghanss, S. R. Harris, J. Parkhill, S. D. Bentley, and G. Pluschke. Emergence and genomic diversification of a virulent serogroup W:ST-2881(CC175) Neisseria meningitidis clone in the African meningitis belt. *Microb Genom*, 3(8):e000120, 08 2017.

[101] C. Schoen, J. Blom, H. Claus, A. Schramm-GlÃijck, P. Brandt, T. MÃijller, A. Goesmann, B. Joseph, S. Konietzny, O. Kurzai, C. Schmitt, T. Friedrich, B. Linke, U. Vogel, and M. Frosch. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. *Proc Natl Acad Sci U S A*, 105(9):3473–3478, Mar 2008.

[102] J. R. Piet, R. A. Huis in 't Veld, B. D. van Schaik, A. H. van Kampen, F. Baas, D. van de Beek, Y. Pannekoek, and A. van der Ende. Genome sequence of Neisseria meningitidis serogroup B strain H44/76. *J Bacteriol*, 193(9):2371–2372, May 2011.

[103] S. Budroni, E. Siena, J. C. Dunning Hotopp, K. L. Seib, D. Serruto, C. Nofroni, M. Comanducci, D. R. Riley, S. C. Daugherty, S. V. Angiuoli, A. Covacci, M. Pizza, R. Rappuoli, E. R. Moxon, H. Tettelin, and D. Medini. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci U S A*, 108(11):4494–4499, Mar 2011.

[104] J. Klughammer, M. Dittrich, J. Blom, V. Mitesser, U. Vogel, M. Frosch, A. Goesmann, T. MÃijller, and C. Schoen. Comparative Genome Sequencing Reveals Within-Host Genetic Changes in Neisseria meningitidis during Invasive Disease. *PLoS One*, 12(1):e0169892, 2017.

[105] N. MacAlasdair, M. Pesonen, O. Brynildsrud, V. Eldholm, P. A. Kristiansen, J. Corander, D. A. Caugant, and S. D. Bentley. The effect of recombination on the evolution of a population of *Neisseria meningitidis*. *Genome Res*, Jun 2021.

[106] H. Sichtig, T. Minogue, Y. Yan, C. Stefan, A. Hall, L. Tallon, L. Sadzewicz, S. Nadendla, W. Klimke, E. Hatcher, M. Shumway, D. L. Aldea, J. Allen, J. Koehler, T. Slezak, S. Lovell, R. Schoepp, and U. Scherf. FDA-ARGOS is a database with public quality-controlled reference genomes

for diagnostic use and regulatory science. *Nat Commun*, 10(1):3313, 07 2019.

[107] Emanuele Bosi, Beatrice Donati, Marco Galardini, Sara Brunetti, Marie-France Sagot, Pietro Lió, Pierluigi Crescenzi, Renato Fani, and Marco Fondi. Medusa: a multi-draft based scaffolder. *Bioinformatics*, 31(15):2443–2451, 2015.

[108] Andrew J Page, Nishadi De Silva, Martin Hunt, Michael A Quail, Julian Parkhill, Simon R Harris, Thomas D Otto, and Jacqueline A Keane. Robust high-throughput prokaryote de novo assembly and improvement pipeline for illumina data. *Microbial genomics*, 2(8), 2016.

[109] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.

[110] Michael Inouye, Harriet Dashnow, Lesley-Ann Raven, Mark B Schultz, Bernard J Pope, Takehiro Tomita, Justin Zobel, and Kathryn E Holt. Srst2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome medicine*, 6(11):90, 2014.

[111] Lennard Epping, Andries J Van Tonder, Rebecca A Gladstone, Stephen D Bentley, Andrew J Page, Jacqueline A Keane, Global Pneumococcal Sequencing Consortium, et al. Seroba: rapid high-throughput serotyping of streptococcus pneumoniae from whole genome sequence data. *Microbial genomics*, 4(7), 2018.

[112] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*, 17(1):132, 06 2016.

[113] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[114] Guido van Rossum and Jelke de Boer. Interactively testing remote servers using the python programming language. *CWi Quarterly*, 4(4):283–303, 1991.

[115] Fernando Pérez and Brian E Granger. Ipython: a system for interactive scientific computing. *Computing in science & engineering*, 9(3):21–29, 2007.

[116] Charles R. Harris, K. Jarrod Millman, StÃľfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime FernÃąndez del RÃŋo, Mark Wiebe, Pearu Peterson, Pierre GÃŀrard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357âĂŞ362, 2020.

[117] Wes McKinney et al. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[118] John D Hunter. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3):90, 2007.

[119] Guangchuang Yu, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 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, 2017.

[120] Hadley Wickham. *Ggplot2: Elegant graphics for data analysis*. Use R! Springer International Publishing, Cham, Switzerland, 2 edition, June 2016.

[121] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[122] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272, 2020.

[123] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[124] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[125] S. ArgimÃşn, K. Abudahab, R. J. E. Goater, A. Fedosejev, J. Bhai, C. Glasner, E. J. Feil, M. T. G. Holden, C. A. Yeats, H. Grundmann, B. G. Spratt, and D. M. Aanensen. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom*, 2(11):e000093, 11 2016.

[126] C. M. Stott and L. M. Bobay. Impact of homologous recombination on core genome phylogenies. *BMC Genomics*, 21(1):829, Nov 2020.

[127] J. Corander, P. Waldmann, and M. J. SillanpÃdÃd. Bayesian analysis of genetic differentiation between populations. *Genetics*, 163(1):367–374, Jan 2003.

[128] G. Tonkin-Hill, J. A. Lees, S. D. Bentley, S. D. W. Frost, and J. Corander. RhierBAPS: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res*, 3:93, 2018.

[129] Nicholas J Croucher, Andrew J Page, Thomas R Connor, Aidan J Delaney, Jacqueline A Keane, Stephen D Bentley,

Julian Parkhill, and Simon R Harris. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins. *Nucleic acids research*, 43(3):e15–e15, 2014.

[130] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[131] Rafal Mostowy, Nicholas J Croucher, Cheryl P Andam, Jukka Corander, William P Hanage, and Pekka Marttinen. Efficient inference of recent and ancestral recombination within bacterial populations. *Molecular biology and evolution*, 34(5):1167–1182, 2017.

[132] Y. Cui, X. Yang, X. Didelot, C. Guo, D. Li, Y. Yan, Y. Zhang, Y. Yuan, H. Yang, J. Wang, J. Wang, Y. Song, D. Zhou, D. Falush, and R. Yang. Epidemic Clones, Oceanic Gene Pools, and Eco-LD in the Free Living Marine Pathogen Vibrio parahaemolyticus. *Mol Biol Evol*, 32(6):1396–1410, Jun 2015.

[133] S. Puranen, M. Pesonen, J. Pensar, Y. Y. Xu, J. A. Lees, S. D. Bentley, N. J. Croucher, and J. Corander. SuperDCA for genome-wide epistasis analysis. *Microb Genom*, 4(6), 06 2018.

[134] J. Pensar, S. Puranen, B. Arnold, N. MacAlasdair, J. Kuronen, G. Tonkin-Hill, M. Pesonen, Y. Xu, A. Sipola, L. SÃąnchez-BusÃş, J. A. Lees, C. Chewapreecha, S. D. Bentley, S. R. Harris, J. Parkhill, N. J. Croucher, and J. Corander. Genome-wide epistasis and co-selection study using mutual information. *Nucleic Acids Res*, 47(18):e112, 10 2019.

[135] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35, Feb 1974.

[136] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano.

ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, Mar 2006.

[137] John W Tukey et al. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

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

[139] H. A. Thorpe, S. C. Bayliss, S. K. Sheppard, and E. J. Feil. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience*, 7(4):1–11, 04 2018.

[140] W. Ding, F. Baumdicker, and R. A. Neher. panX: pan-genome analysis and exploration. *Nucleic Acids Res*, 46(1):e5, 01 2018.

[141] S. L. Salzberg. Next-generation genome annotation: we still struggle to get it right. *Genome Biol*, 20(1):92, 05 2019.

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

[143] D. Hyatt, G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene

recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, Mar 2010.

[144] J. A. Lees, M. Galardini, S. D. Bentley, J. N. Weiser, and J. Corander. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, 34(24):4310–4312, 12 2018.

[145] Dong Yu, Yuan Jin, Zhiqiu Yin, Hongguang Ren, Wei Zhou, Long Liang, and Junjie Yue. A genome-wide identification of genes undergoing recombination and positive selection in neisseria. *BioMed research international*, 2014, 2014.

[146] Won Jong Kim, Dustin Higashi, Maira Goytia, Maria A Rendón, Michelle Pilligua-Lucas, Matthew Bronnimann, Jeanine A McLean, Joseph Duncan, David Trees, Ann E Jerse, et al. Commensal neisseria kill neisseria gonorrhoeae through a DNA-dependent mechanism. *Cell host & microbe*, 26(2):228–239, 2019.

[147] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[148] Olive Jean Dunn. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.

[149] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[150] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[151] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[152] C. Chewapreecha, S. R. Harris, N. J. Croucher, C. Turner, P. Marttinen, L. Cheng, A. Pessia, D. M. Aanensen, A. E.

Mather, A. J. Page, S. J. Salter, D. Harris, F. Nosten, D. Goldblatt, J. Corander, J. Parkhill, P. Turner, and S. D. Bentley. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet*, 46(3):305–309, Mar 2014.

[153] Y. L. Tzeng, J. Thomas, and D. S. Stephens. Regulation of capsule in Neisseria meningitidis. *Crit Rev Microbiol*, 42(5):759–772, Sep 2016.

[154] Y. Zhang and T. P. Begley. Cloning, sequencing and regulation of thiA, a thiamin biosynthesis gene from Bacillus subtilis. *Gene*, 198(1-2):73–82, Oct 1997.

[155] A. M. Dietl, Z. Meir, Y. Shadkchan, N. Osherov, and H. Haas. Riboflavin and pantothenic acid biosynthesis are crucial for iron homeostasis and virulence in the pathogenic mold Aspergillus fumigatus. *Virulence*, 9(1):1036–1049, 2018.

[156] S. Sah, S. Aluri, K. Rex, and U. Varshney. One-carbon metabolic pathway rewiring in Escherichia coli reveals an evolutionary advantage of 10-formyltetrahydrofolate synthetase (Fhs) in survival under hypoxia. *J Bacteriol*, 197(4):717–726, Feb 2015.

[157] P. Rana, S. M. Ghouse, R. Akunuri, Y. V. Madhavi, S. Chopra, and S. Nanduri. FabI (enoyl acyl carrier protein reductase) - A potential broad spectrum therapeutic target and its inhibitors. *Eur J Med Chem*, 208:112757, Dec 2020.

[158] J. Yao, D. F. Bruhn, M. W. Frank, R. E. Lee, and C. O. Rock. Activation of Exogenous Fatty Acids to Acyl-Acyl Carrier Protein Cannot Bypass FabI Inhibition in Neisseria. *J Biol Chem*, 291(1):171–181, Jan 2016.

[159] N. Yokota, T. Kuroda, S. Matsuyama, and H. Tokuda. Characterization of the LolA-LolB system as the general lipoprotein localization mechanism of Escherichia coli. *J Biol Chem*, 274(43):30995–30999, Oct 1999.

[160] Y. Wei and E. B. Newman. Studies on the role of the metK gene product of Escherichia coli K-12. *Mol Microbiol*, 43(6):1651–1656, Mar 2002.

[161] P. H. van der Meide, E. Vijgenboom, A. Talens, and L. Bosch. The role of EF-Tu in the expression of tufA and tufB genes. *Eur J Biochem*, 130(2):397–407, Feb 1983.

[162] B. P. Goldstein, G. Zaffaroni, O. Tiboni, B. Amiri, and M. Denaro. Determination of the number of tuf genes in Chlamydia trachomatis and Neisseria gonorrhoeae. *FEMS Microbiol Lett*, 51(3):305–309, Aug 1989.

[163] K. Rome, C. Borde, R. Taher, J. Cayron, C. Lesterlin, E. Gueguen, E. De Rosny, and A. Rodrigue. The Two-Component System ZraPSR Is a Novel ESR that Contributes to Intrinsic Antibiotic Tolerance in Escherichia coli. *J Mol Biol*, 430(24):4971–4985, 12 2018.

[164] E. N. Shinners and B. W. Catlin. Arginine and pyrimidine biosynthetic defects in Neisseria gonorrhoeae strains isolated from patients. *J Bacteriol*, 151(1):295–302, Jul 1982.

[165] B. P. Burns, S. L. Hazell, G. L. Mendz, T. Kolesnikow, D. Tillet, and B. A. Neilan. The Helicobacter pylori pyrB gene encoding aspartate carbamoyltransferase is essential for bacterial survival. *Arch Biochem Biophys*, 380(1):78–84, Aug 2000.

[166] V. Časaitė, R. Stanislauskienė, J. Vaitekūnas, D. Taurait e, R. Rutkienė, R. Gasparavičiūtė, and R. Meškys. Microbial Degradation of Pyridine: a Complete Pathway in Arthrobacter sp. Strain 68b Deciphered. *Appl Environ Microbiol*, 86(15), 07 2020.

[167] T. Pieńko and J. Trylska. Extracellular loops of BtuB facilitate transport of vitamin B12 through the outer membrane of E. coli. *PLoS Comput Biol*, 16(7):e1008024, 07 2020.

[168] Masatoshi Nei and Takashi Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5):418–426, 1986.

[169] Chase W Nelson, Louise H Moncla, and Austin L Hughes. Snpgenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics*, 31(22):3709–3711, 2015.

[170] Ben Murrell, Steven Weaver, Martin D Smith, Joel O Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P Martin, Davey M Smith, et al. Gene-wide identification of episodic selection. *Molecular biology and evolution*, 32(5):1365–1371, 2015.

[171] Ben Murrell, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L Kosakovsky Pond, and Konrad Scheffler. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution*, 30(5):1196–1205, 2013.

[172] Martin D Smith, Joel O Wertheim, Steven Weaver, Ben Murrell, Konrad Scheffler, and Sergei L Kosakovsky Pond. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular biology and evolution*, 32(5):1342–1353, 2015.

[173] Sergei L Kosakovsky Pond and Spencer V Muse. HyPhy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution*, pages 125–181. Springer, 2005.

[174] M. Jaillard, L. Lima, M. Tournoud, P. MahÃĭ, A. van Belkum, V. Lacroix, and L. Jacob. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet*, 14(11):e1007758, 11 2018.

[175] J. A. Lees, M. Vehkala, N. VÃďlimÃďki, S. R. Harris, C. Chewapreecha, N. J. Croucher, P. Marttinen, M. R. Davies, A. C. Steer, S. Y. Tong, A. Honkela, J. Parkhill,

S. D. Bentley, and J. Corander. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*, 7:12797, 09 2016.

[176] P. M. Duffin and H. S. Seifert. DNA uptake sequence-mediated enhancement of transformation in Neisseria gonorrhoeae is strain dependent. *J Bacteriol*, 192(17):4436–4444, Sep 2010.

[177] O. H. Ambur, S. A. Frye, M. Nilsen, E. Hovland, and T. TÃÿnjum. Restriction and sequence alterations affect DNA uptake sequence-dependent transformation in Neisseria meningitidis. *PLoS One*, 7(7):e39742, 2012.

[178] B. A. Al Suwayyid, L. Rankine-Wilson, D. J. Speers, M. J. Wise, G. W. Coombs, and C. M. Kahler. Meningococcal Disease-Associated Prophage-Like Elements Are Present in Neisseria gonorrhoeae and Some Commensal Neisseria Species. *Genome Biol Evol*, 12(2):3938–3950, 02 2020.

[179] M. Chen, C. Zhang, X. Zhang, and M. Chen. Meningococcal Quinolone Resistance Originated from Several Commensal Neisseria Species. *Antimicrob Agents Chemother*, 64(2), 01 2020.

[180] J. C. D. Hotopp, R. Grifantini, N. Kumar, Y. L. Tzeng, D. Fouts, E. Frigimelica, M. Draghi, M. M. Giuliani, R. Rappuoli, D. S. Stephens, G. Grandi, and H. Tettelin. Comparative genomics of Neisseria meningitidis: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology (Reading)*, 152(Pt 12):3733–3749, Dec 2006.

[181] Q. F. Lu, D. M. Cao, L. L. Su, S. B. Li, G. B. Ye, X. Y. Zhu, and J. P. Wang. Genus-Wide Comparative Genomics Analysis of Neisseria to Identify New Genes Associated with Pathogenicity and Niche Adaptation of Neisseria Pathogens. *Int J Genomics*, 2019:6015730, 2019.

[182] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver,

A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[183] The Gene Ontology Consortium.

[184] S. Grossmann, S. Bauer, P. N. Robinson, and M. Vingron. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, 23(22):3024–3031, Nov 2007.

[185] S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson. BioinformaticsOntologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, Jul 2008.

[186] S. A. Zamani-Dahaj, M. Okasha, J. Kosakowski, and P. G. Higgs. Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Mol Biol Evol*, 33(7):1843–1857, 07 2016.

[187] F. Baumdicker, W. R. Hess, and P. Pfaffelhuber. The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol*, 4(4):443–456, 2012.

[188] R. E. Collins and P. G. Higgs. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol Biol Evol*, 29(11):3413–3425, Nov 2012.

[189] O. Brynildsrud, J. Bohlin, L. Scheffer, and V. Eldholm. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol*, 17(1):238, 11 2016.

[190] A. Jamet, A. B. Jousset, D. Euphrasie, P. Mukorako, A. Boucharlat, A. Ducousso, A. Charbit, and X. Nassif. A new family of secreted toxins in pathogenic Neisseria species. *PLoS Pathog*, 11(1):e1004592, Jan 2015.

[191] A. Jamet and X. Nassif. Characterization of the Maf family of polymorphic toxins in pathogenic Neisseria species. *Microb Cell*, 2(3):88–90, Mar 2015.

[192] S. HÃľlaine, E. Carbonnelle, L. Prouvensier, J. L. Beretti, X. Nassif, and V. Pelicic. PilX, a pilus-associated protein essential for bacterial aggregation, is a key to pilus-facilitated attachment of Neisseria meningitidis to human cells. *Mol Microbiol*, 55(1):65–77, Jan 2005.

[193] A. F. Imhaus and G. DumÃľnil. The number of Neisseria meningitidis type IV pili determines host cell interaction. *EMBO J*, 33(16):1767–1783, Aug 2014.

[194] Christoph Schoen, Hervé Tettelin, Julian Parkhill, and Matthias Frosch. Genome flexibility in neisseria meningitidis. *Vaccine*, 27:B103–B111, 2009.

[195] O. B. Harrison, J. E. Bray, M. C. Maiden, and D. A. Caugant. Genomic Analysis of the Evolution and Global Spread of Hyper-invasive Meningococcal Lineage 5. *EBioMedicine*, 2(3):234–243, Mar 2015.

[196] P. Domingo, V. Pomar, A. Mauri, and N. Barquet. Standing on the shoulders of giants: two centuries of struggle against meningococcal disease. *Lancet Infect Dis*, 19(8):e284–e294, 08 2019.

[197] J. W. Eerkens, R. V. Nichols, G. G. R. Murray, K. Perez, E. Murga, P. Kaijankoski, J. S. Rosenthal, L. Engbring, and B. Shapiro. A probable prehistoric case of meningococcal disease from San Francisco Bay: Next generation sequencing of Neisseria meningitidis from dental calculus and osteological evidence. *Int J Paleopathol*, 22:173–180, 09 2018.

[198] A. A. DavÃŋn, E. Tannier, T. A. Williams, B. Boussau, V. Daubin, and G. J. SzÃűllÅŚsi. Gene transfers can date the tree of life. *Nat Ecol Evol*, 2(5):904–909, 05 2018.

[199] T. Sakoparnig, C. Field, and E. van Nimwegen. Whole genome phylogenies reflect the distributions of recombi-

nation rates for many bacterial species. *Elife*, 10, Jan 2021.

[200] A. Lamelas, A. M. Hamid, J. P. Dangy, J. Hauser, M. Jud, K. RÃűltgen, A. Hodgson, T. Junghanss, S. R. Harris, J. Parkhill, S. D. Bentley, and G. Pluschke. Loss of Genomic Diversity in a Neisseria meningitidis Clone Through a Colonization Bottleneck. *Genome Biol Evol*, 10(8):2102–2109, 08 2018.

[201] S. J. Gray, C. L. Trotter, M. E. Ramsay, M. Guiver, A. J. Fox, R. Borrow, R. H. Mallard, and E. B. Kaczmarski. Epidemiology of meningococcal disease in England and Wales 1993/94 to 2003/04: contribution and experiences of the Meningococcal Reference Unit. *J Med Microbiol*, 55(Pt 7):887–896, Jul 2006.

[202] J. V. Peñalba and J. B. W. Wolf. From molecules to populations: appreciating and estimating recombination rate variation. *Nat Rev Genet*, 21(8):476–492, 08 2020.

[203] J. Stapley, P. G. D. Feulner, S. E. Johnston, A. W. Santure, and C. M. Smadja. Recombination: the good, the bad and the variable. *Philos Trans R Soc Lond B Biol Sci*, 372(1736), Dec 2017.

[204] R. E. Michod, H. Bernstein, and A. M. Nedelcu. Adaptive value of sex in microbial pathogens. *Infect Genet Evol*, 8(3):267–285, May 2008.

[205] O. H. Ambur, J. EngelstÃďdter, P. J. Johnsen, E. L. Miller, and D. E. Rozen. Steady at the wheel: conservative sex and the benefits of bacterial transformation. *Philos Trans R Soc Lond B Biol Sci*, 371(1706), 10 2016.

[206] O. B. Brynildsrud, V. Eldholm, J. Bohlin, K. Uadiale, S. Obaro, and D. A. Caugant. Acquisition of virulence genes by a carrier strain gave rise to the ongoing epidemics of meningococcal disease in West Africa. *Proc Natl Acad Sci U S A*, 115(21):5510–5515, 05 2018.

[207] J. G. Lawrence and A. C. Retchless. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. *Methods Mol Biol*, 532:29–53, 2009.

[208] B. J. Shapiro, J. B. Leducq, and J. Mallet. What Is Speciation? *PLoS Genet*, 12(3):e1005860, Mar 2016.

[209] T. J. Straub and O. Zhaxybayeva. A null model for microbial diversification. *Proc Natl Acad Sci U S A*, 114(27):E5414–E5423, 07 2017.

[210] A. W. Dretler, N. G. Rouphael, and D. S. Stephens. Progress toward the global control of Neisseria meningitidis: 21st century vaccines, current guidelines, and challenges for future vaccine development. *Hum Vaccin Immunother*, 14(5):1146–1160, 05 2018.

[211] Y. L. Tzeng and D. S. Stephens. A Narrative Review of the W, X, Y, E, and NG of Meningococcal Disease: Emerging Capsular Groups, Pathotypes, and Global Control. *Microorganisms*, 9(3), Mar 2021.

[212] Paul A Kristiansen, Fabien Diomandé, Stanley C Wei, Rasmata Ouédraogo, Lassana Sangaré, Idrissa Sanou, Denis Kandolo, Pascal Kaboré, Thomas A Clark, Abdoul-Salam Ouédraogo, et al. Baseline meningococcal carriage in burkina faso before the introduction of a meningococcal serogroup a conjugate vaccine. *Clin. Vaccine Immunol.*, 18(3):435–443, 2011.

[213] Paul A Kristiansen, Absatou Ky Ba, Idrissa Sanou, Abdoul-Salam Ouédraogo, Rasmata Ouédraogo, Lassana Sangaré, Fabien Diomandé, Denis Kandolo, Jennifer Dolan Thomas, Thomas A Clark, et al. Phenotypic and genotypic characterization of meningococcal carriage and disease isolates in burkina faso after mass vaccination with a serogroup a conjugate vaccine. *BMC infectious diseases*, 13(1):363, 2013.

[214] J. M. Collard, B. Issaka, M. Zaneidou, S. Hugonnet, P. Nicolas, M. K. Taha, B. Greenwood, and J. F. Jusot. Epidemiological changes in meningococcal meningitis

in Niger from 2008 to 2011 and the impact of vaccination. *BMC Infect Dis*, 13:576, Dec 2013.

[215] World Health Organization et al. Meningococcal disease control in countries of the african meningitis belt, 2014. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 90(13):123–131, 2015.

[216] C. L. Trotter, L. Cibrelus, K. Fernandez, C. Lingani, O. Ronveaux, and J. M. Stuart. Response thresholds for epidemic meningitis in sub-Saharan Africa following the introduction of MenAfriVac ®. *Vaccine*, 33(46):6212–6217, Nov 2015.

[217] S. Vuocolo, P. Balmer, W. C. Gruber, K. U. Jansen, A. S. Anderson, J. L. Perez, and L. J. York. Vaccination strategies for the prevention of meningococcal disease. *Hum Vaccin Immunother*, 14(5):1203–1215, 05 2018.

[218] J. Alderfer, R. E. Isturiz, and A. Srivastava. Lessons from mass vaccination response to meningococcal B outbreaks at US universities. *Postgrad Med*, 132(7):614–623, Sep 2020.

[219] R. Rappuoli, M. Pizza, V. Masignani, and K. Vadivelu. Meningococcal B vaccine (4CMenB): the journey from research to real world experience. *Expert Rev Vaccines*, 17(12):1111–1121, 12 2018.

[220] Clive Cookson. 'Century of biology' takes time to bear fruit. *Financial Times*, 2009.

[221] J. Zhou, L. D. Bowler, and B. G. Spratt. Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of Neisseria meningitidis and commensal Neisseria species. *Mol Microbiol*, 23(4):799–812, Feb 1997.

[222] E. Rodgers, S. D. Bentley, R. Borrow, H. B. Bratcher, S. Brisse, A. B. Brueggemann, D. A. Caugant, J. Findlow, L. Fox, L. Glennie, L. H. Harrison, O. B. Harrison, R. S. Heyderman, M. J. van Rensburg, K. A. Jolley,

B. Kwambana-Adams, S. Ladhani, M. LaForce, M. Levin, J. Lucidarme, N. MacAlasdair, J. Maclennan, M. C. J. Maiden, L. Maynard-Smith, A. Muzzi, P. Oster, C. M. C. Rodrigues, O. Ronveaux, L. Serino, V. Smith, A. van der Ende, J. VÃązquez, X. Wang, S. Yezli, and J. M. Stuart. The global meningitis genome partnership. *J Infect*, 81(4):510–520, 10 2020.