Imperial College London

PhD Thesis

# A phylogenetic method to perform genome-wide association studies in microbes

*Caitlin Collins*

PhD Candidate

Department of Infectious Disease Epidemiology
Faculty of Medicine, Imperial College London
St. Mary's Campus, Norfolk Place, London, W2 1PG
caitlin.collins12@imperial.ac.uk

supervised by
Dr. Xavier Didelot & Prof. Christophe Fraser

assessed by
Prof. Matthew Fisher & Prof. Sam Sheppard

June 14, 2019

# Declaration of Originality

I declare that the work presented in this thesis is my own. All other work is appropriately referenced in the text, and any contributions made by other parties are acknowledged in a statement at the beginning of each chapter.

# Copyright Declaration

# Acknowledgements

First and foremost, I would like to thank my supervisor, Xavier Didelot, for all of the time, effort, and care that he devoted to me throughout the course of my PhD. I owe him an great debt of gratitude for providing years of lively discussion and valuable instruction, for challenging and encouraging me, and for offering his endless patience and steady guidance at every stage of this project. I would like to extend my sincere thanks to Christophe Fraser, as well, for allowing me to pursue my doctoral research under his supervision. This project would not have been possible without his generous support. Nor could I have undertaken this work without the support of The Wellcome Trust and the Biotechnology and Biological Sciences Research Council, and I would like to thank them for their generosity in funding this work. I am very thankful for the input and insights of Daniel Wilson and David Aanensen, who were kind enough to serve as my assessors during the development of this work. I benefited greatly from our many thoughtful discussions. I also want to express my gratitude to the assessors of this thesis. Thank you, in advance, for volunteering your valuable time to review my work.

On a personal level, I am extraordinarily grateful to Nick Grassly, for the support and understanding that he has shown me. I have been overwhelmed by his generosity, and I could not have completed this work without his consideration. I feel incredibly fortunate to have been able to complete my PhD at Imperial College. It has been a pleasure to work in the Department of Infectious Disease Epidemiology, surrounded by the intellectual creativity and individual character of the students, staff, and faculty in our research community. I am immensely grateful to my colleagues, for their inspiration, collaboration, and friendship during my time in the DIDE department. In particular, I would like to thank my dearest neighbours, Daniel Laydon and Ben Lambert, for providing me with a home away from home. I also want to say thank you to Alice Ledda, my chocolate fairy, for your visits to my desk and for our chats in the kitchen. And, a special thanks to Thibaut Jombart for introducing me to population genetics and R programming in the first place, and for your moral support, alcohol consumption, and friendship ever since.

Above all else, my deepest thanks must go to my family. First, to Maher, I thank you for the boundless love and care that you have shown me, despite enduring the entirety of this process, with its ups and downs, late nights and busy weekends. Thanks are far too small for what I owe you in return, but I submit my thanks in any case, and my love. With all my heart. To Alex, my big brother, I will always be grateful for your hugs, chats, and inspiration. Thank you for coming to visit me, without hesitation, and for rescuing me with a sailing vacation. Finally, I reserve my deepest gratitude for my parents. My loving, caring, working in lab all weekend mom and dad. It is only thanks to your endless love and unwavering support that I have made it to where I am today. I dedicate this work to you.

# Abstract

Genome-Wide Association Studies (GWAS) are designed to perform an unbiased search of genetic sequence data with the intent of identifying statistically significant associations with a phenotype or trait of interest. The application of GWAS methods to microbial organisms promises to improve the way we understand, manage, and treat infectious diseases. Yet, while microbial pathogens continue to undermine human health, wealth, and longevity, microbial GWAS methods remain unable to fully capitalise on the growing wealth of bacterial and viral genetic sequence data. Clonal population structure and homologous recombination in microbial organisms make it difficult for existing GWAS methods to achieve both the precision needed to reject false positive findings and the statistical power required to detect genuine associations between microbial genotypic and phenotypic variants. In this thesis, we investigate potential solutions to the most substantial methodological challenges in microbial GWAS, and we introduce a new phylogenetic GWAS approach that has been specifically designed for use in bacterial samples. In presenting our approach, we describe the features that render it robust to the confounding effects of both population structure and recombination, while maintaining high statistical power to detect associations. Our approach is applicable to organisms ranging from purely clonal to frequently recombining, to sequence data from both the core and accessory genome, and to binary, categorical, and continuous phenotypes. We also describe the efforts taken to make our method efficient, scalable, and accessible in its implementation within the open-source R package we have created, called treeWAS. Next, we apply our GWAS method to simulated datasets. We develop multiple frameworks for simulating genotypic and phenotypic data with control over relevant parameters. We then present the results of our simulation study, and we use thorough performance testing to demonstrate the power and specificity of our approach, as compared to the performance of alternative cluster-based and dimension-reduction methods. Our approach is then applied to three empirical datasets, from *Neisseria gonorrhoeae* and *Neisseria meningitidis*, where we identify core SNPs associated with binary drug resistance and continuous antibiotic minimum inhibitory concentration phenotypes, as well as both core SNP and accessory genome associations with invasive and commensal phenotypes. These applications illustrate the versatility and potential of our method, demonstrating in each case that our approach is capable of confirming known resistance- or virulence-associated loci and discovering novel associations. Our thesis concludes with a review of the previous chapters and an evaluation of the strengths and limitations displayed by the current implementation of our phylogenetic approach to association testing. We discuss key areas for further development, and we propose potential solutions to advance the development of microbial GWAS in future work.

# Contents

# List of Figures

# List of Tables

# Acronyms

**ACCTRAN** accelerated transformation. i, 53

**BAPS** Bayesian analysis of population structure. i, 26

**BIC** Bayesian Information Criterion. i, 33, 117

**BIGSdb** Bacterial Isolate Genome Sequence Database. i, 153

**BLAST** Basic Local Alignment Search Tool. i, 153

**CMH** Cochran-Mantel-Haenszel. i, 18

**CNV** copy number variation. i, 57

**DA** discriminant analysis. i, 32

**DAPC** discriminant analysis of principal components. i, 33

**FDR** false discovery rate. i, 94

**FPR** false positive rate. i

**FWER** family-wise error rate. i

**GC** genomic control. i, 23

**GLS** generalised least squares. i

**GWAS** genome-wide association study. i, 10

**HGT** horizontal gene transfer. i, 6

**IC** independent contrast. i

**INDEL** insertion and deletion. i, 57

**LD** linkage disequilibrium. i, 11

**LMM** linear mixed model. i, 36

**MCMC** Markov chain Monte Carlo. i, 26

**MDS** metric multi-dimensional scaling. i, 32

**MIC** minimum inhibitory concentration. i, 58

**MLST** multi-locus sequence typing. i, 26

**MPR** most parsimonious reconstruction. i, 53

**NJ** Neighbour-Joining. i, 38

**OR** odds ratio. i, 28

**PBP** penicillin binding protein. i, 152

**PC** principal component. i, 18

**PCA** principal components analysis. i, 13

**PCM** phylogenetic comparative method. i

**PMM** phylogenetic mixed model. i, 56

**PPV** positive predictive value. i

**QQ** quantile-quantile. i, 23

**SEER** sequence element enrichment analysis. i, 115

**SNP** single nucleotide polymorphism. i, 11

**UPGMA** Unweighted Pair Group Method with Arithmetic Mean. i, 38

**VC** variance component. i, 55

# Chapter 1

# Introduction

**Acknowledgements** A version of some of the content in this chapter also appeared in Collins and Didelot [1]. Caitlin Collins reviewed the literature and produced the report. Xavier Didelot revised the report, and his feedback is reflected in the chapter below.

## 1.1 Microbes

No environment inhabited by man is free from cohabitation with microorganisms. Microbes from all three domains of life participate in complex mutualistic, commensal, and parasitic relationships with humans, animals, plants, and with each other. Bacteria, a diverse group of prokaryotic unicellular organisms, may aid in host metabolism [2], cause disease [3], or colonise hosts without any effect [4]. Archaea, which are similar to bacteria in morphology but genetically and metabolically distinct, have been observed in mutualistic and commensal relationships primarily with bacteria and protozoa [5]. Eukaryotic protozoa, like the malarial agent *Plasmodium falciparum*, can cause disease in both humans [6] and animals [7], but play an important role in decomposition [8]. Eukaryotic fungi are likewise known for their symbiotic role in plant growth [9], but may parasitise both plants [10] and animals [11].

The micro- and macro-organisms mentioned above can also be infected by viruses, which are made up of genetic material surrounded by a protein capsid and lipid envelope, but rely on the cellular machinery of host organisms for replication. While viruses are relatively simple, they are also common and diverse. Viruses ranging from the Flu virus *Haemophilus influenza* to the AIDS-causing HIV virus can infect human hosts. The viral bacteriophages that infect bacteria and archaea, meanwhile, may be the most abundant biological entity on earth [12]. Viral infection can result in asymptomatic carriage, disease, or the death of the host [13]. Although, mutualistic relationships do occur between viruses and host organisms or ecosystems [14, 15]. The infection of bacteria by phages, for example, regulates bacterial population growth and provides critical benefits to marine ecosystems [16] and human health [17]. Altogether, microbes form a complex network that enters, improves, and ends the lives of macro-organisms every day.

## 1.2 Bacteria

Given the extent of microbial variation, we have chosen to focus this work primarily on bacteria alone, although other microbes will receive some consideration in the pages to come. Bacteria are microscopic unicellular organisms that are able to live and reproduce independently, but form large communities that grow exponentially. The Earth is home to approximately $5x10^{30}$ bacteria [18], found living everywhere from deep-sea vents [19] to the human gut [20]. In fact, there are more bacterial cells in the human body than there are human cells [21]. Our relationship to these microorganisms is both adversarial and interdependent.

Bacteria are an essential part of our ecosystem and economy. They enrich the soil with nutrients by degrading waste and enable plant growth by mediating nitrogen fixation [22]. We rely on the metabolic and acidifying enzymes that bacteria provide to facilitate domestic animal growth and dairy production [23]. We may even engineer an end to climate change by recruiting *Clostridium thermocellum* bacteria to generate renewable energy from plant sources [24]. Meanwhile, however, long-dormant bacterial pathogens are emerging from the melting permafrost and beginning to infect humans [25]. There are 2,000 species of bacteria that commonly colonise human hosts, though 99.5% of these are either harmless commensal bacteria or beneficial mutualistic species [3]. Gut bacteria play valuable roles in metabolism. *Lactobacillus reuteri*, for example, aids in the conversion of carbohydrates into polysaccharides, while *Escherichia coli* and *Bifidobacteria* strains enable vitamin K and folate uptake [26, 27]. Among commensal bacteria, *Streptococcus pneumoniae* and *Staphylococcus aureus* are relatively frequent asymptomatic colonisers of the human nasopharynx. It is estimated that 14% of the human population carry *S. pneumoniae*, while 20% of people are persistent carriers of nasopharyngeal *S. aureus* and 60% are transient carriers [28, 29].

Some bacteria are obligate pathogens. *Bacillus anthracis*, for example, requires a human or animal host to complete its life cycle, causing typically fatal anthrax infections in host organisms [30]. But, commensal and even mutualistic bacteria, including *S. aureus*, *S. pneumoniae*, and *E. coli*, can become pathogenic and cause disease in their hosts, leading to significant morbidity and mortality. Acute respiratory infections, caused by *S. pneumoniae* as well as a number of viral pathogens, kill 4 million people annually [31]. *Mycobacterium tuberculosis* alone leads to an additional 1.5 million annual deaths by tuberculosis [32]. Diarrhoeal diseases, caused by etiological agents including *Vibrio cholerae*, *E. coli*, *Shigella*, *Campylobacter*, *Salmonella*, and *rotavirus*, take the lives of 1.3 million people per year, striking the young in developing countries most severely [33]. Worldwide, one in four deaths is attributed to infectious diseases [34], and in the developing world infectious pathogens remain a leading cause of infant mortality [35].

Existing interventions and medications stand to substantially reduce the burden of infectious disease in low- and middle-income countries [36]. However, over-reliance on antibiotics to combat poorly understood infectious disease etiologies has already sparked an arms race between pharmaceutical development and pathogenic escape mechanisms; and, unfortunately, it appears that resistance is not futile for many infectious pathogens [37,38]. Already, once readily-curable infectious diseases like gonorrhoeae must be frequently treated with “last-line” antibiotics [39]. In the United States, methicillin-resistant *S. aureus* (MRSA) now claims more lives than HIV [40]. Recent predictions suggest that by 2050, antibiotic-resistant infections could cost US$100 trillion, and take 10 million lives annually [41]. The ability to understand at a molecular level what gives rise to the phenotypic differences between bacteria is, quite literally, a matter of life and death. Certainly, the identification of genetic structures that separate commensal bacteria from their pathogenic relatives or distinguish antibiotic-susceptible bacteria from drug-resistant isolates is of substantial interest to public health.

## 1.3 Bacterial genetics

A better understanding of how genetic differences give rise to the phenotypic variation displayed by bacteria would have extensive potential applications. In biotechnology, linking genotype to phenotype may allow us to co-opt advantageous bacterial behaviours like those of *Alcanivorax borkumensis*, the alkane-degrading bacteria now used to clean up oil spills [42]. Genetic data analysis may also enhance our responsiveness to outbreaks, inform us of their origin, and allow us to determine what genetic features give rise to increased virulence [43, 44]. By capitalising on the genetic sequence data and metadata collected by disease surveillance programs, we may refine our ability to identify subtypes of disease, determine what genetic features characterise epidemic clones, and aid in the development of drugs and vaccines that will be more effective over longer periods [45–47]. Likewise, our capacity to prevent the development of drug resistance or predict the emergence of zoonoses may be determined by our ability to interrogate bacterial genomes [48–53]. As globalisation continues to increase the rate and scope of human interaction, with each other, and with animals, the evidence suggests this process will be accompanied by parallel change in the spread and evolution of infectious pathogens [43, 46, 54, 55]. Our ability to mine solutions to these problems from a growing wealth of pathogen genome sequences must continue to evolve as well.

In bacteria, genetic information is usually found in a single circular chromosome, although multiple chromosomes and linear chromosomes also occur. The genome size of bacteria range from 112 kb in *Nasuia-ALF* [56] to 14 Mb in *Sorangium cellulosum* [57], and can even vary within species like *E. coli* by over 1 Mb [58]. In contrast to eukaryotes, the protein-coding proportion of bacterial genomes is quite consistent, with around one gene per kilobase of DNA. Larger bacterial genomes reflect larger numbers of genes and more complex regulation of gene expression, and correspond to the variability of the environments to which a bacterium may be adapted [59].

Reproduction in bacteria occurs via binary fission. The single-cell organism duplicates its chromosome or chromosomes and undergoes cell division, preserving the parent cell and creating a new clone that contains one copy of each chromosome. If no errors occur during this process, the genomes of the parent cell and its clonal daughter will be identical. However, genetic replication is a relatively error-prone process, and spontaneous point mutations may occur in each generation, causing one or more bases in the daugter genome to differ from the ancestral copy. With short generation times, clonal reproduction in bacteria can lead the number of cells in a population to expand rapidly, doubling every 20 minutes in *E. coli* [60] and every 40 minutes in *Neisseria meningitidis* [61]. As a result of this imperfect replication and rapid reproduction, genetic and even phenotypic variation can easily accumulate via point mutation [62].

### 1.3.1 Recombination

In addition to the vertical inheritance of genetic variation that arises during clonal reproduction, the horizontal transfer of genetic information through a process of recombination can also generate genetic variation in bacteria. Three mechanisms enable the acquisition of exogenous genetic material by horizontal gene transfer (HGT) : transduction, conjugation, and transformation.

Transduction by bacteriophages can introduce novel genes or gene sequences. Bacteriophages are viruses that rely on bacteria for their reproduction. Infection by a bacteriophage can be fatal for the host bacterium. But, when the bacterium survives, bacteriophage infection can result in the uptake of foreign DNA, either that of the phage itself or DNA from previous bacterial hosts that the phage has been able to infect [63]. Like humans, bacteria develop mechanisms to resist viral infection. Hence, transduction only takes place between similarly phage-susceptible bacteria [64].

Conjugation between bacteria allows genetic material stored in plasmids to be transferred from a donor to a recipient cell. Plasmids are independently-replicating circularised DNA molecules that are distinct from the chromosomal DNA of a bacterium. Plasmids often contain useful accessory genes (non-essential genes found in a subset of species members [65]) that can confer selectable traits like antibiotic resistance, virulence, and secondary metabolic enzyme production [66,67]. Sharing these genes via conjugation can improve the survival of bacterial populations facing changing environmental pressures by adding functional capacities not already found in chromosomal genes [68]. Conjugation is initiated by the plasmid-containing donor cell and proceeds when the donor cell physically joins itself to the recipient by means of a pilus. The plasmid is linearised and a single strand of its DNA is transferred to the donor, whereupon both cells synthesise a complementary strand, completing the process of duplication.

Transformation allows bacteria to take up exogenous DNA from the environment. Fragments of genetic material enter the environment of bacteria upon the lysis of other bacterial cells. In certain conditions, for example during a particular growth phase [69] or under environmental stress [70, 71], many bacteria will express a set of proteins that render them "competent" for transformation. Over 80 bacterial species have been found to demonstrate such "competence" naturally, allowing them to internalise exogenous genetic material [69].

Once foreign DNA has entered the bacterial cell via HGT, gene conversion—integration of the novel genetic material into the host genome—completes the horizontal process of exchange. If gene conversion occurs through non-homologous recombination, the exogenous donor DNA fragment is simply inserted into the recipient genome. If gene conversion is mediated by the more common homologous recombination mechanism, donor DNA is integrated by replacing a segment of the recipient genome. The incorporation of recombinant DNA via homologous recombination only occurs where the exogenous DNA fragment and the existing genome sequence share a sufficient degree of homology or base-pair similarity [72]. Hence, as the genetic distance between two genomes increases and homology decreases, the probability of homologous recombination is curtailed, for example, by the inducement of mismatch repair genes or the reduced efficiency of the

RecA recombinase [73]. Recombination has been observed in most bacterial species, occuring at a wide range of rates [74] and under different constraints. In some bacteria, horizontal exchange is restricted to occur within lineages but, in others, recombination may cross species boundaries [75]. In highly recombinant organisms, like *S. pneumoniae*, recombination can drive as much as 90% of sequence diversification [76]. A ratio of $r/m = 7$ was estimated in these sampled genomes, indicating that substitutions at pneumococcal loci arose seven times more often by recombination ($r$) than by point mutation ($m$). Even in the predominantly clonal *M. tuberculosis*, a ratio of $r/m = 0.49$ implies that mutation contributes only twice as much as recombination to the generation of genetic variation [77]. At the same time, recombination events have been estimated to occur five times less frequently than point mutations in *M. tuberculosis*, though multiple loci are impacted per event. Indeed, the evolution of most bacteria is impacted to some extent by recombination as well as mutation, although the relative contributions of these horizontal and vertical forces vary between species.

Recombination plays an important role in the diversification of bacterial genomes. The variable integration of recombinant DNA fragments expands the repertoire of SNPs and generates variation within genes in the "core genome", defined as the set of genes present in all sampled genomes (99-100% of isolates) [78, 79]. Recombination also facilitates the diversification and proliferation of genes in the "accessory genome", which contains genes absent from one or more isolates [80, 81]. As a driver of bacterial sequence variation, recombination also has the capacity to endow bacteria with new phenotypic traits, enabling especially rapid adaptation in response to selection [82, 83]. Efforts to identify the genetic basis of bacterial phenotypes will, therefore, be more effective and more powerful if they can incorporate and account for the effects of recombination.

## 1.4 Linking genotype to phenotype

The reasons why bacteria differ in their traits and behaviours are often encoded in their genomes. Additional factors—host genetics, the environment, and related interactions—can also contribute to phenotypic variation, complicating the relationship between bacterial genotype and phenotype. Yet, so long as a microbial phenotype displays heritability, in that a proportion of its variation can be attributed to microbial genetic factors, genetic analyses can be used to better understand that trait. There are two major classes of approach adopted to identify the genetic basis of phenotypic variation. In molecular genetic or "reverse genetics" approaches, the genetic sequence is modified and the resulting change in phenotype is assessed. Whereas, in "forward genetics", genomes are examined or compared as a means of identifying the genetic basis of phenotypes observed in a sample.

### 1.4.1 Reverse genetics

Reverse genetics techniques have been widely used in the study of bacteria [84]. These include knockout and reversion experiments, in which the phenotypic state of a bacterium is observed when a gene is inactivated and reactivated. Gene inactivation or alteration can be accomplished by multiple means. Random mutagenesis can be induced by transposons and, if followed by selection, can enable the inactivation or incorporation of a particular gene [85, 86]. Alternatively, for bacteria like *Chlamydia trachomatis* that not amenable to this form of molecular genetic intervention in the laboratory, chemical mutagens can be used to create inactive or mutant genes across the genome [87]. However, the widespread and random generation of mutations by either approach may not be the most efficient means of inactivating a particular gene. When the gene sequence of interest is known, recombinant sequences can be specifically designed to remove or replace the target sequence upon incorporation into the recipient genome. For example, so-called suicide plasmids can be engineered to induce gene deletion via excision of the target gene or to achieve site-directed mutagenesis by replacing the wild-type gene with a mutant copy [88, 89]. Such approaches, however, are only possible in competent bacteria that are capable of undergoing recombination in the laboratory.

Of course, a prerequisite for any reverse genetics technique is the suitability of bacteria for life in the laboratory. Yet, it is estimated that only half of the major bacterial lineages have species that can be grown in the laboratory [90]. Complex polygenic traits involving epistasis between many genes will not be amenable to such approaches. In addition, many bacterial phenotypes are ill-suited to re-creation in lab settings, where environmental and host factors may be irreproducible [91]. Invasive disease can be modelled in animals, for example, but the immune response and disease susceptibility of mice and men may give rise to vastly different bacterial behaviours [92, 93]. For phenotypes like host association, laboratory-based reverse genetics approaches alone are unlikely to ever identify a genetic basis. Ultimately, when one considers the amount of time and money that would be required to thoroughly investigate via reverse genetics the phenotypic effects of all bacterial genes, let alone all polymorphic loci, one may be thankful that an alternative approach exists.

### 1.4.2 Forward genetics

Forward genetics provides another way to examine the relationship between bacterial genotype and phenotype. If naturally-occurring or randomly-generated phenotypic variation exists in a bacterial sample, forward genetics approaches can be used to identify trait-associated variation. If genetic sequences are available for a sample containing more than one phenotypic state, these sequences can be compared and genetic differences identified. Since the sequencing of the first complete bacterial genome in 1995, technological improvements have enabled rapid increases in sequencing speed and decreases in sequencing cost. As a result, the number of bacterial whole-genome sequences has exponentially increased over the past three decades [94]. Forward genetic analyses have likewise progressed over time, from structural and functional dissections of a single sequence, to comparative analyses of two or more sequences, to more systematic approaches for comparing large whole-genome samples. Today, over 100,000 bacterial genomes have been published [95]. Hence, sequence-based forward genetics approaches have gained prominence as a means of uncovering the genetic basis of phenotypic traits [91].

Comparative genomic approaches have been used to search for evidence of associations between genotype and phenotype in bacteria. The observation of co-occurring changes in genotype and phenotype is widely accepted as an indicator of allele-trait association. The strength of this evidence is increased if changes occur multiple times over the evolutionary history of the sample, which may contain closely-related isolates [96], isolates observed and sequenced over time [97], or clades of evolutionarily-related isolates [98] Large-scale statistical analyses enable even more rigorous examinations of the relationships between genotype and phenotype.

## 1.5 Genome-wide association studies

Genome-wide association studies (GWASs) have become a popular and reliable way to make inferences about the genetic basis of phenotypic variation. GWAS methods quantify the degree of statistical dependence, or "association", observed between a phenotype of interest and the genotype at every locus in a genomic dataset. Where genotypic differences at particular loci correlate systematically to differences in the phenotype, to a greater extent than might be expected by chance, these loci are deemed to be in significant association with the phenotype. Unlike hypothesis-driven laboratory-based analyses, these statistical approaches allow for an unbiased exploration of naturally-occurring genotypic and phenotypic variation. Although laboratory confirmation is required to establish causality and to ensure that a functional relationship exists between a particular allele and phenotypic state, GWAS methods provide a powerful tool for the identification of candidate loci in statistical association with a phenotypic trait. By adopting systematic genome-wide approaches and reducing the number of genetic loci that must be investigated in the laboratory by several orders of magnitude, GWAS studies can generate more comprehensive descriptions of the genetic basis of phenotypic traits while rapidly increasing the pace of discovery.

### 1.5.1 From human GWAS to microbial GWAS

Since the publication of the first GWAS studies in the early 2000s [99–102], GWAS have become a tool of choice in human genetics. Given the infeasibility of undertaking representative studies of human phenotypic variation in a laboratory context, and the health and safety restrictions preventing experimental reverse genetics in human subjects, great investment has been made in the development of bioinformatic approaches for examining existing human genotypic variation. As such, methodological approaches to GWAS have been developed primarily for the purposes of undertaking association studies in humans. To date, over 3,000 human GWAS studies have been published, leading to the discovery of over 60,000 single nucleotide polymorphisms (SNPs) associated with a wide array of phenotypes [103]. It has long been anticipated that by applying GWAS methods to microbes, similar discoveries might be made [104].

### 1.5.2 Challenges

Dramatic differences between human beings and bacteria have prevented the direct transference of GWAS methods from one organism to the other (see Table 1.1). Humans reproduce sexually, with recombination occurring at predictable intervals, while bacteria undergo asexual reproduction and exchange genetic material by recombination at different rates in a less predictable fashion. Human populations contain few genetically-identical individuals; whereas, in the absence of mutation or recombination, entire bacterial populations would consist of identical clones. Single strands of human chromosomal DNA closely resemble parental copies, but new diploid combinations of genes ensure that genotypic and phenotypic variation can increase in every generation. Copies of each chromosome are inherited during reproduction. Although linkage disequilibrium (LD) may be maintained among neighbouring genetic markers, linkage decreases predictably with physical distance along the chromosome, as chromosomal crossing over causes recombination to occur during reproduction. In bacteria, by contrast, asexual reproduction

| Trait | Humans | Bacteria |
|---|---|---|
| **Genome size** | large (3,234 Mb) | small (112 kb - 14 Mb) |
| **Ploidy** | diploid | haploid |
| **Reproduction** | sexual | clonal |
| **Recombination** | generational | variable |
| **Genetic linkage** | distance-based | not distance-based |
| **Genomic variation** | core | core and accessory |
| **Population structure** | moderate<br>GWAS confounder | variable, potentially strong<br>GWAS confounder |
| **Laboratory testing** | unethical | possible |

**Table 1.1. Comparison of human and bacterial traits.**

allows LD to persist across the entire haploid chromosome. Unless recombination or point mutation intervenes to introduce a substitution, clonal inheritance will preserve the widespread correlations between genetic loci and between genotype and phenotype. The statistical non-independence between sites that results from these bacterial characteristics presents us with a considerable challenge. As popular human GWAS methods are not equipped to deal with clonal inheritance or long-range LD, the implementation of new and creative methodological solutions will be essential for the success of bacterial GWAS.

Population structure is defined by the presence of systematic differences in allele frequencies between subpopulations in a sample. Population stratification, a related concept relevant in association studies, occurs when these ancestral allele frequencies differ systematically between the "cases" and "controls" or phenotypic groups being analysed via GWAS [105]. One may infer that the repeated observation of directional genotypic and phenotypic variance indicates a statistical or causal relationship between the two variables. If, instead, variance in both genotype and phenotype is correlated to ancestry, confounding bias can arise from the non-independence of these observations [106]. Hence, when comparing ancestral populations with different phenotypes, one may mistakenly infer that all genetic variables that distinguish these populations are associated with the difference in phenotype when, in fact, there may be no such link.

It has been well established in human GWAS that, if unaccounted for, population structure can have confounding effects on the inferences made in association studies [107]. Great care is taken within human GWAS study design to screen out close genetic relatives and to match cases and controls by ethnicity, sex, and other possible confounding factors. Human GWAS methodologists are also pursuing more effective approaches to check the spurious associations arising from "cryptic relatedness", that is, the unrecognised genetic or familial relationships between sampled individuals that give rise to subtle population structures below the level of recognised "ancestral" population clusters [108, 109]. In general, however, human ancestral relationships are well described by broad, admixed populations of variable size and genetic relatedness. Hence, human GWAS studies have been successful thus far in addressing population stratification at the level of large ethno-geographic clusters. Although this is a simplification of the true picture, the assumption that population structure is homogenous within these clusters is much more easily justified in human analyses than it would be in even semi-clonal bacteria. The most popular approach to account for the confounding effects of population structure in human GWAS uses the principal components analysis (PCA) dimension reduction method [107, 110]. PCA identifies major axes of variation that best separate these populations in multivariate space (see Section 2.2.6). Using PCA to correct for the genetic differences separating major population groups has been shown to sufficiently correct for the confounding effects of ancestry in human GWAS [107, 111].

In contrast, the clonal nature of bacteria can dramatically increase the strength of ancestral relationships. In fact, the magnitude of this potential problem was great enough to delay the advent of microbial GWAS [91]. In bacteria, the preservation and accumulation of ancestral differences increases the extent to which allele frequencies differ systematically by ancestry. Moreover, clonal inheritance encourages the formation of rigid, nested sub-population structures. The partitioning of bacterial sub-populations may also be linked to environmental differences, reflecting separation by geography, host organism, or host tissue type [79, 112, 113].

The problems posed by clonality are confounded by the observation that in association studies, phenotypic "cases" are often more closely related than phenotypic "controls". This is especially true in microbial GWAS. In human GWAS, efforts are made to distinguish "close relatives" from effectively "unrelated" individuals, and to exclude the former during sampling. Conversely, in microbial GWAS, it is not feasible to exclude genetic relatives from samples composed of clonally-related isolates. The propensity for close genetic relationships is often even higher among phenotypic “cases”, especially if these are sampled from disease outbreaks or transmission chains [114]. In addition, patterns of clonal inheritance like clonal expansion can allow a particular phenotypic state to dominate an ancestral clade or sub-population [97, 115], increasing the spurious association between the phenotype and ancestral genotype. Altogether, this combination of clonal population structure, biased sampling, and the interrelatedness of phenotypic “cases”, greatly increases the challenge of confounding population stratification in microbial association studies. Unless the bacterial clonal genealogy is obscured by widespread recombination, methods of accounting for population structure in human GWAS will not reliably control for this confounding factor in bacterial association studies [97].

Unlike the predictable recombination that accompanies human sexual reproduction, highly variable and unpredictable recombination in bacteria also poses a challenge for association studies. Specifically, recombination can distort the reconstruction of ancestral relationships and thus increase the difficulty of accounting for the confounding effects of population structure [1]. In addition, the probability of chance association with the phenotype can also be affected by recombination. In conclusion, GWAS methods that effectively prevent the identification of spurious associations at one recombination rate may fail at other levels of recombination. Recombination, therefore, poses multiple challenges in bacteria which human GWAS methods have not addressed.

### 1.5.3 Opportunities

Although, bacteria present many methodological challenges for GWAS, a number of their features encourage the undertaking of association studies. Bacterial genomes are several orders of magnitude smaller than human genomes, and a larger proportion of bacterial genomes is composed of protein-coding sites [97]. (1) Genome-wide association tests must be applied to fewer loci in bacterial analyses. This alleviates much of the burden imposed by the need to correct for multiple testing, meaning that higher discovery power can be maintained in microbial GWAS. (2) The sample sizes required to detect associations in microbes can be much smaller than the thousands of individuals required in human GWAS [116]. In humans, lack of power to detect statistically significant associations between phenotypes and genetic loci of lower effect size has caused the number of findings made by GWAS to plateau over time, even as sample sizes have increased [91]. (3) It is feasible to generate larger samples for bacterial analyses, as their smaller genomes are more easily sequenced. With larger samples, bacterial GWAS may even be capable of detecting associations with low-effect genetic loci. Indeed, for microbial phenotypes arising from combinations of low-effect alleles, it has been suggested that GWAS may be the only way of elucidating their genetic basis [117].

Homologous recombination may also offer important benefits for bacterial GWAS studies. Despite the challenges inherent in performing association studies on clonal isolates that also undergo recombination, the horizontal exchange of genetic information can be a powerful diversifying force in microbial samples. Recombination can lead to rapid evolution in bacteria and may even generate significant phenotypic differences between close relatives. Moreover, alongside point mutation, recombination provides a key mechanism for the breakdown of extensive LD in bacterial genomes. Thus, while less predictable than recombination in humans, bacterial recombination can present a critical means by which signals of association may be disentangled from noisy clonal backgrounds.

Finally, the ability to ethically manipulate bacterial genomes in the laboratory allows GWAS results to be easily confirmed or rejected. The value of GWAS studies in bacteria can thus be increased by becoming one component of a pipeline leading to the identification of statistical associations, as well as the establishment of causal links between microbial genotype and phenotype. As microbial GWAS stands to substantially increase the pace of discovery, it promises to contribute to more rapid improvements in human and veterinary medicine, and the public health management of outbreaks and antibiotic resistance. The inherent advantages presented by bacteria encourage the undertaking of association studies in these organisms. Although a number of challenges not found in human GWAS continue to complicate this endeavour, the anticipation of promising results provides significant impetus for the development of new methods that will enable the widespread application of bacterial GWAS.

## 1.6 Microbial GWAS

While the advent of microbial GWAS has been relatively recent—a decade after human GWAS—promising results can already be seen [49, 80, 106, 114, 116, 118–132] (see Table 1.2). In applying GWAS methods to bacteria and viruses, these studies have adopted a variety of methodological approaches (see Table 1.3)to address the challenges outlined above, namely population stratification, variable recombination, and the need to maintain high statistical power while rejecting false positive associations.

The microbial GWAS studies published thus far have adopted a wide range of approaches to correct for the confounding potential of population stratification. We group these approaches into three main categories: (i) cluster-based approaches, (ii) multivariate dimension reduction methods, and (iii) phylogenetic approaches.

| **Reference** | **Organism** | **Phenotype** | **Recomb Rate** | **Genetic Variant** | **Sample Size** | **Number of loci** | **Number of significant loci** |
|---|---|---|---|---|---|---|---|
| Chen and Shapiro [126] | *M. tuberculosis* | Resistance | Low | SNPs | 123 | 24,711 | 0 |
| Chewapreecha *et al.* [4 | *S. pneumoniae* | Resistance | High | SNPs | 3,701 | 392,524 | 301 |
| Laabei *et al.* [120] | *S. aureus* | Virulence | High | SNPs | 90 | 3,060 | 121 &<br>4 |
| Weinert *et al.* [124] | *S. suis* | Host association | High | SNPs, genes, k-mers | 191 | 178,979 SNPs,<br>7,675,<br>125,593 k-mers | 0 &<br>0 |
| Howell *et al.* [80] | *H. parasuis* | Clinical | High | SNPs, genes | 212 | 65,137 SNPs,<br>6,053 genes | 12 SNPs,<br>48 genes |
| Power *et al.* [116] | *HIV* | Resistance | High | SNPs | 343 | 5,100 | 8 |
| Bartha *et al.* [119] | *HIV* | Viral load | High | SNPs | 1,071 | 3,125 | 0 |
| Salipante *et al.* [125] | *E. coli* | Resistance | High | CDS (unique coding sequences) | 312 | 446,152 | 20 |
| Lees *et al.* [129] | *S. pneumoniae*<br>*& S. pyogenes* | Resistance &<br>Invasiveness | High &<br>Low | k-mers | 3,069 &<br>675 | 68,000,000 | 30,157 (9 loci) &<br>2 loci |
| Lees *et al.* [132] | *S. pneumoniae* | Carriage duration | High | SNPs, k-mers | 2,157 | 92,487 SNPs,<br>5,254,876 k-mers | 1 SNP, 320 k-mers<br>& 0 SNPs, $\geq$ 1 PCs |
| Earle *et al.* [128] | *E. coli &*<br>*K. pneumoniae &*<br>*M. tuberculosis &*<br>*S. aureus* | Resistance | High &<br>High &<br>Low &<br>Low | SNPs, k-mers<br>PCs | 241 &<br>176 &<br>1,735 &<br>992 | 263,604 &<br>417,645 &<br>654,425 &<br>107,480 | 14 &<br>20 &<br>6 &<br>28 genes |
| Maury *et al.* [133] | *L. monocytogenes* | Virulence | Low | Core genes, gene families | 104 | 1,791 | 43 |
| Alam *et al.* [122] | *S. aureus* | Resistance | Low | SNPs | 75 | 55,977 | 1 |
| Phelan *et al.* [134] | *M. tuberculosis* | Resistance | Low | SNPs | 127 | 19,248 | 7 &<br>18 |
| Coll *et al.* [135] | *M. tuberculosis* | Resistance | Low | SNPs | 6,465 | 102,160 | 43 &<br>147 |
| Farhat *et al.* [106] | *M. tuberculosis*<br>*& C. jejuni* | Resistance &<br>Host association | Low | SNPs, genes | 123 (16)<br>192 (16 | 4.4 Mb &<br>1.6 Mb | 7 &<br>107 genes |
| Brynildsrud *et al.* [127] | *S. epidermis &*<br>*S. pneumoniae* | Resistance | High &<br>High | Core genes, accessory genes | 21 &<br>3,085 | 2.2 Mb &<br>2.4 Mb | 6 &<br>1 |
| Hall *et al.* [121] | *E. coli, Shigella*<br>*& E. coli* | Virulence &<br>Host association | High &<br>High | SNPs | 68 &<br>116 | 418,500 &<br>470,806 | 97 &<br>101 |
| Desjardins *et al.* [130] | *M. tuberculosis* | Resistance | Low | SNPs | 498 | 11,704 | 12 |
| Farhat *et al.* [114] | *M. tuberculosis* | Resistance | Low | SNPs | 123 | 24,711 | 50 |
| Nebenzahl-Guimaraes *et al.* [131] | *M. tuberculosis* | Transmissibility | Low | SNPs, genes, intergenic regions | 100 &<br>143 | 4.4 Mb | 5 &<br>4 |
| Sheppard *et al.* [118] | *C. jejuni* | Host association | High | k-mers | 192 | 1.5 Mb | 7,307 k-mers<br>(7 genes) |

**Table 1.2. Microbial GWAS studies.**

| Reference | Correction for Population Structure | Association Test | Software |
|---|---|---|---|
| Chen and Shapiro [126] | Epiclusters | CMH $X^2$ test | PLINK [136] |
| Chewapreecha *et al.* [49] | BAPS clusters | CMH $X^2$ test | PLINK [136], R [137] |
| Laabei *et al.* [120] | Genomic Control & hierarchical clusters | $X^2$ test & bespoke | PLINK [136] |
| Weinert *et al.* [124] | DAPC & BAPS clusters | Logistic regression & CMH $X^2$ test | R [137], PLINK [136] |
| Howell *et al.* [80] | DAPC | Generalised linear model | R [137] |
| Power *et al.* [116] | PCA (5 PCs) | Logistic regression | PLINK [136] |
| Bartha *et al.* [119] | PCA (2 PCs) | Linear regression | PLINK [136] |
| Salipante *et al.* [125] | PCA (3 PCs) | Logistic regression | R [137] |
| Lees *et al.* [129] | MDS (1 PC) | Logistic & linear regression | SEER [129] |
| Lees *et al.* [132] | Kinship matrix & PCA (30 PCs) | LMM & linear regression | FaST-LMM [138] & SEER [129] |
| Earle *et al.* [128] | PCA (sig. PCs) | LMM | GEMMA [139], bugWAS [128] |
| Maury *et al.* [133] | Distance matrix | Generalised linear model | R [137] |
| Alam *et al.* [122] | Distance matrix & phylogenetic tree | Regression & PhyC Fisher test | (Q)ROADTRIPS [140] & – |
| Phelan *et al.* [134] | Kinship matrix & phylogenetic tree | LMM & PhyC Fisher test | EMMA [141] & – |
| Coll *et al.* [135] | Kinship matrix and PCA (5 PCs) & phylogenetic tree | LMM & PhyC Fisher test | GEMMA [139] & – |
| Farhat *et al.* [106] | Sampling strategy | Bespoke | – |
| Brynildsrud *et al.* [127] | Pairwise comparisons | Binomial test | Scoary [127] |
| Hall *et al.* [121] | Phylogenetic tree | $X^2$ test | PPFS [121] |
| Desjardins *et al.* [130] | Phylogenetic tree | Generalised least squares regression | BayesTraits [142] |
| Farhat *et al.* [114] | Phylogenetic tree (consensus) | PhyC (bespoke) | – |
| Nebenzahl-Guimaraes *et al.* [131] | Phylogenetic tree (Bayesian) | PhyC | – |
| Sheppard *et al.* [118] | Phylogenetic tree (ClonalFrame [143]) | bespoke | – |

**Table 1.3. Microbial GWAS methods.**

Cluster-based approaches like the Cochran-Mantel-Haenszel (CMH) Test [144] have become a popular means of mitigating the confounding effects of population structure [49, 124, 126]. Segregating isolates into ancestrally-related groups via clustering algorithms allows the association study to be stratified by population. Dimension reduction techniques, like PCA [110], have also been successfully applied to microbial GWAS [80, 116, 145]. These approaches represent population structure in principal components (PCs), major axes of variation that can be used to control for ancestry in regression-based association tests. These alternative GWAS approaches are described in detail in Chapter 2.

A number of phylogenetic approaches to population structure have also been explored. Some microbial GWAS studies have sought to assess the probability of chance associations due to population structure by rearranging the phenotype [49, 116, 127]. Pairwise approaches have also been implemented to examine allele-trait associations among pairs of related isolates of different phenotype [106, 127]. Regression-based approaches like ROADTRIPS [146] can take a phylogenetic tree as an input and use this to account for the effect of ancestral relationships on the association between genotype and phenotype. The most promising phylogenetic approaches have attempted to retain all information rather than extracting isolate pairs and to rearrange the genotype rather than the phenotype [114, 147]. In the following chapter, we will review how these corrective mechanisms are implemented and evaluate the effectiveness of each approach in addressing population stratification. As we will see, despite the adoption of a wide range of strategies, clonal relatedness and confounding population stratification remains a challenge for microbial GWAS.

While the confounding effects of ancestry are typically addressed by at least one of many diverse approaches in all microbial GWAS studies, methods adequately designed to account for the confounding effects of recombination are almost entirely absent from the microbial GWAS literature. GWAS analyses of microbes that undergo both low and high levels of recombination, like *M. tuberculosis* and *S. pneumoniae*, have been performed with clustering methods, dimension-reduction techniques, and pairwise or phylogenetic approaches to account for population structure (see Table 1.2 and 1.3). Sheppard *et al.* [118] use ClonalFrame [143] to account for recombination while inferring clonal relationships. However, no other bacterial GWAS method takes deliberate steps to account for the confounding effects of recombination, either during ancestry inference or when attempting to delineate between spurious and genuine associations.

A stated aim of all microbial GWAS studies has been to maintain high statistical power while rejecting the false positive associations that arise by chance and as a result of confounding factors. Each correction for the influence of ancestry in microbial GWAS is designed to reject false positive findings. However, greater stringency or less accuracy in these approaches can result in diminished power to detect genuine associations.

Several GWAS studies experiment with different methods of varying stringency, leaving the reader with mixed sets of null results and significant results of variable size [120, 122, 124]. In the absence of an overarching or unifying framework, this use of multiple methods complicates interpretation and renders the accuracy of conflicting findings unclear. The method of Earle *et al.* [128] takes deliberate steps to improve power by recapturing lineage-level associations. Most microbial GWAS methods, however, do not implement creative approaches to improve statistical power.

In summary, there are a number of challenges that have prevented the widespread, successful application of GWAS methods to microbes. We will discuss these issues in depth in the following chapter, and we will indicate how we plan to address them in developing our own microbial GWAS method.

## 1.7 Thesis structure

In **Chapter 2**, we provide a detailed review of the literature. We introduce the contending approaches available to address methodological issues in microbial GWAS. We examine solutions to the problems posed by confounding population structure, recombination, and association testing. We highlight the strengths and limitations of existing approaches and support the choices we have made in developing our method.

In **Chapter 3**, we introduce our phylogenetic method for performing microbial GWAS. We present the theoretical foundations underlying our approach, and we describe each component of our method in detail. We also present the methodology used within the simulation study that we devised as a means of testing and assessing our GWAS method. We describe the approaches used to simulate genetic and phenotypic data, and we present the range of parameters explored.

In **Chapter 4**, we present the results of our simulation study. We use simulation testing in the development and refinement of our method, for example, to identify an appropriate significance threshold and to select the best method of ancestral state reconstruction. We then assess the performance of our method under three different simulation schemes. We compare the performance of our method to that of existing GWAS approaches, and we examine how performance varies with changes in population stratification, recombination, the effect size of association, and dataset size.

In **Chapter 5**, we reveal the results of our applications to empirical data. We search for associations in *N. meningitidis* core SNPs and accessory gene presence or absence matrices. We analyse both antibiotic resistance and invasive disease phenotypes.

# Chapter 2

# Literature survey of bacterial population genetics and GWAS

**Acknowledgements** An abridged version of the review of ancestral reconstruction methods presented in the chapter below appeared in Collins and Didelot [1]. Caitlin Collins and Xavier Didelot identified relevant publications and worked together to determine the structure of this review. Collins reviewed the literature and produced the first draft of this book chapter. Didelot edited the text and provided helpful suggestions that are reflected in the published book chapter and in the chapter below.

## 2.1 Introduction

Pioneering efforts have demonstrated that GWAS studies have immense potential to inform our understanding of the molecular basis of microbial phenotypic traits [49,80,106, 114,116,118–132] (see Table 1.2). Yet, the undertaking of association studies in microbial samples has been hindered by a number of methodological challenges. If GWAS methods are to gain widespread and successful application in microbial samples, the following barriers must be overcome. First, it is necessary to counteract the confounding bias introduced by the ancestral relationships between isolates. If unaccounted for, population stratification can be a major source of spurious associations and, thus, false positive findings. Approaches to association testing in microbes must therefore take steps to reconstruct the population structure of the sample, to quantify its potential confounding effect, and to separate likely false positive findings from genuine associations with statistical support. Second, we must be able to account for the impact that recombination may have on the analysis. To remain robust in spite of the variable introduction of genetic variation via horizontal transfer between isolates, efforts must be made to control for the distortion that recombination can introduce, both when reconstructing ancestral relationships and when making inferences during association testing. Third, it is imperative that we complement efforts to eliminate false positive findings with strategies to maximise the statistical power to detect associations. Optimal power in microbial GWAS may only be achieved through the implementation of appropriate measures of association, efficient strategies for controlling confounders, and with effective use of the available data. In this chapter, we consider the potential solutions to these major challenges in microbial GWAS. We describe existing methods, including those that have been transferred from human genetics, as well as more recent approaches that have been designed for microbial samples. The strengths and limitations of each approach are evaluated, and critical gaps in the available methodology are identified. We conclude by presenting a path forward for the methodological development surrounding each key issue, indicating which strategies we will explore or which solutions we propose to implement in our own microbial GWAS method.

## 2.2 Controlling for population stratification

GWAS provides a systematic approach through which we can compare genomes and identify genetic loci that vary in association with a particular phenotype. This statistical analysis relies on the assumption that the genotype of each individual in a sample varies independently and that, therefore, the repeated observation of particular genotypic states alongside a given phenotypic state can be taken as evidence of association. This assumption of independence is often violated in GWAS, due to the presence of ancestral relationships between individuals. This is especially true with regard to microbial GWAS, where isolates may be drawn from highly clonal populations, as one can not assume that individuals represent truly independent samples. Instead, sampled isolates should be thought of as “pseudoreplicates”, with genomes possessing varying degrees of sequence similarity as a result of common ancestry [148]. When phenotypic differences coincide with the divisions between ancestral subpopulations, the confounding conditions of population stratification are established. Spurious associations between genotype and phenotype will be widespread unless the confounding impact of ancestry can be revealed and removed. GWAS approaches typically address the problem of population stratification in a two-step process: (1) Reconstruct the ancestral relationships present in a sample. (2) Compensate for the increase in spurious asociations expected, given the observed population structure.

### 2.2.1 Genomic inflation

Some techniques have, however, been devised to allow population stratification to be identified and corrected for without requiring the user to reconstruct or model the population structure of their sample. Devlin and Roeder [108] developed a metric to quantify population stratification, based on the observation that association test statistics are inflated by this form of systematic bias. The genomic inflation factor, $\lambda_{GC}$, can be computed from $X^2$ test statistics that have been obtained from a set of $i$ random genetic markers. It is calculated by dividing the observed median $X^2$ of the association of all $i$ loci by the expected median $X^2$, with $1df$ (one degree of freedom), under the null hypothesis of no association, as in equation 2.1 [108]:

$$\lambda_{\mathbf{GC}} = \frac{median(X_i^2)}{median(X_{1df}^2)} \tag{2.1}$$

This $\lambda_{GC}$ value can be used within the genomic control (GC) approach as a uniform correction for population stratification. To perform GC, one simply divides all association test statistics by the overall $\lambda_{GC}$ factor. One major limitation that arises from the uniform nature of this ancestry adjustment is that it may over- or under-correct for genomic inflation because the degree of differentiation between ancestral populations is not uniform across genetic loci [149]. The straightforward and frequently conservative approach that GC offers has nevertheless provided one solution to the problem of population stratification within the microbial GWAS literature [120, 126].

The genomic inflation factor is applied with even greater regularity as a tool to assess the effectiveness of other approaches to correcting for population stratification. Both $\lambda_{GC}$ and quantile-quantile (QQ) plots are useful for diagnosing the inflation of association test statistics and for comparing inflation when GWAS is performed with or without a particular population structure control. Where $\lambda_{GC}$ provides a quantitative measure of inflation (indicated by $\lambda_{GC} > 1.05$), a QQ plot can provide an informative qualitative alternative [150]. QQ plots allow for a visual comparison of the expected $X^2$ distributed $-log_{10}$ $P$ values with the observed $-log_{10}$ $P$ values. This can reveal systematic inflation above the $y = x$ line that may indicate population stratification [97]. On the other hand, QQ plots showing inflation above $y = x$ among only high $-log_{10}$ $P$ values may instead indicate genuine polygenic traits in small samples. Meanwhile, systematic deflation may be caused by an excess of rare variants [116]. If interpreted correctly, both the QQ plot and $\lambda_{GC}$ can improve microbial GWAS by serving as diagnostic tools. However, more complex approaches are likely to be required if we are to account sufficiently but not excessively for the confounding effects of population stratification.

### 2.2.2 Reconstructing ancestral relationships

In the microbial GWAS literature published thus far, a wide array of approaches have been adopted to tackle the problems posed by population structure. A relatively small number of microbial GWAS analyses have either implemented no correction for population stratification or have used the uniform GC correction, described above [120,126,151]. All other microbial GWAS studies have undertaken a two-step process: first, reconstructing the ancestral relationships between isolates, and then using this model of population structure to help separate genuine signals of association from signals attributable to ancestry alone.

In human genetics, a distinction is usually drawn between "ancestral" relationships that are deep-rooted in the evolutionary past and "familial" relationships that occur at present or in the recent past [150]. In bacterial populations, by contrast, clonal inheritance allows us to consider genetic relatedness at all levels and on any timescale to be a suitable target for methods attempting to reconstruct "ancestral relationships" between genomes.

Several methodological approaches have been designed to infer how genetic ancestry links bacterial isolates. Below, we introduce these contending methods, grouped into the following families: non-phylogenetic approaches, including model-based clustering methods and model-free dimension reduction techniques, and phylogenetic approaches, including standard methods and methods accounting for recombination. No single method has emerged as the universal "Gold Standard" within this domain, as the relative merits of each ultimately depend on sample characteristics and the aims of the analysis. We will highlight the strengths and limitations inherent in each approach and indicate which sample parameters might favour a particular method of reconstruction. The level of recombination observed, for example, may restrict the choice of methodological approach (see Table 2.1). Each method of ancestral reconstruction is also accompanied by a number of corrective strategies, which allow one to translate ancestry inference into a means of controlling for confounding bias. We will review these corrective mechanisms and evaluate the effectiveness of each in the pages to come.

| Typical Evolution | Optimal Method | Example of Applications |
|---|---|---|
| **Completely clonal** | Phylogenetic methods ignoring recombination | *Mycobacterium tuberculosis* [106, 114, 130, 131],<br>*Leptospira interrogans* [74] |
| **Moderate rate of recombination** | Phylogenetic methods accounting for recombination | *Escherichia coli* [152],<br>*Chlamydia trachomatis* [153],<br>*Clostridium difficile* [154],<br>*Neisseria meningitidis* [155],<br>individual lineages of *Campylobacter jejuni* [118] and *Streptococcus pneumoniae* [49, 129] |
| **High rate of recombination** | Phylogeny-independent approaches | *Helicobacter pylori* [156], HIV [116, 119],<br>species-wide datasets of *Campylobacter jejuni* [157] and *Streptococcus pneumoniae* [76] |

**Table 2.1. Optimal method of ancestral reconstruction by recombination rate.**

### 2.2.3 Non-phylogenetic methods

Non-phylogenetic methods reconstruct the population structure of a sample by separating individuals into clusters or along principal component axes (PCs) according to the variation in their genomes. In the case of model-based methods, this inference is made within a parametric framework that allows users to incorporate prior information about the population and its evolution, whereas model-free methods infer population structure from the sequence data alone. The clusters and PCs identified by these approaches often correspond to the major genealogical divisions on a phylogenetic tree [80, 128, 158, 159]. These inference methods nevertheless return a more limited representation of ancestral relationships than could be offered by a phylogeny linking all isolates back to their most recent common ancestor. Even so, in some microbes, high levels of recombination may make it impossible to reliably identify a clonal genealogy. For samples in which recombination renders phylogenetic methods inapplicable, cluster-based and dimension-reduction methods can provide valuable alternative approaches for inferring ancestral relationships.

### 2.2.4 Model-based clustering methods

Model-based non-phylogenetic methods attempt to cluster individuals into a set of genetically-similar groups that best represent the population structure of the sample. The assignment of individual genomes or genetic loci to one of these $k$ ancestral populations is accomplished within a Bayesian statistical framework. These methods allow for the exploration of a parameter space that can be shaped by the prior inputs of analysts, with the aim of identifying a population structure that gives a high likelihood of the observed genetic data. All ancestral reconstruction methods inevitably identify an approximation of the true biological reality that is influenced by the design of the approach in question and its inherent biases. Model-based clustering methods identify ancestral populations under an explicit population genetics model and a set of stated assumptions.

**STRUCTURE** is one of the older and better-known model-based clustering methods, and its approach provides a representative example of the approaches adopted within this class of methods [160, 161]. In STRUCTURE [160], a Bayesian Markov chain Monte Carlo (MCMC) algorithm is used to jointly estimate the quantities indicated in Box 2.1.

1. For each locus of each sequence, the probabilities of derivation from each of the $k$ ancestral populations.
2. For each locus and $k$ ancestral populations, the population allele frequencies.
3. Additional global parameters, such as the average length of fragments inherited from an ancestral population.

**Box 2.1. STRUCTURE parameters**

Figure 2.1) provides an illustration of typical STRUCTURE output. The linkage version [161] makes STRUCTURE more applicable to bacteria by eliminating the assumed independence of allele frequencies between loci, accounting instead for the fact that neighbouring loci are more likely to be inherited from the same ancestral population. These original STRUCTURE methods were, however, designed for multi-locus sequence typing (MLST) data, and they are not able to handle the large volumes of whole-genome sequence data typically collected for microbial GWAS analyses.

Faster alternative model-based programs have since been developed. The computationally expensive MCMC of STRUCTURE has been implemented more efficiently in Bayesian analysis of population structure (BAPS) [163], which has become one of the most popular Bayesian clustering methods. Elsewhere, the STRUCTURE MCMC mechanism is replaced with a maximum-likelihood optimisation approach in ADMIXTURE [164], a copying model in fineStructure [165], and a Bayesian change-point clustering model in BratNextGen [166]. Each of these implementations is more efficient than the original STRUCTURE program. Some computational burden remains inherent in all model-based methods, but these computational advancements have allowed model-based clustering methods to remain a feasible option for reconstructing microbial population structure.

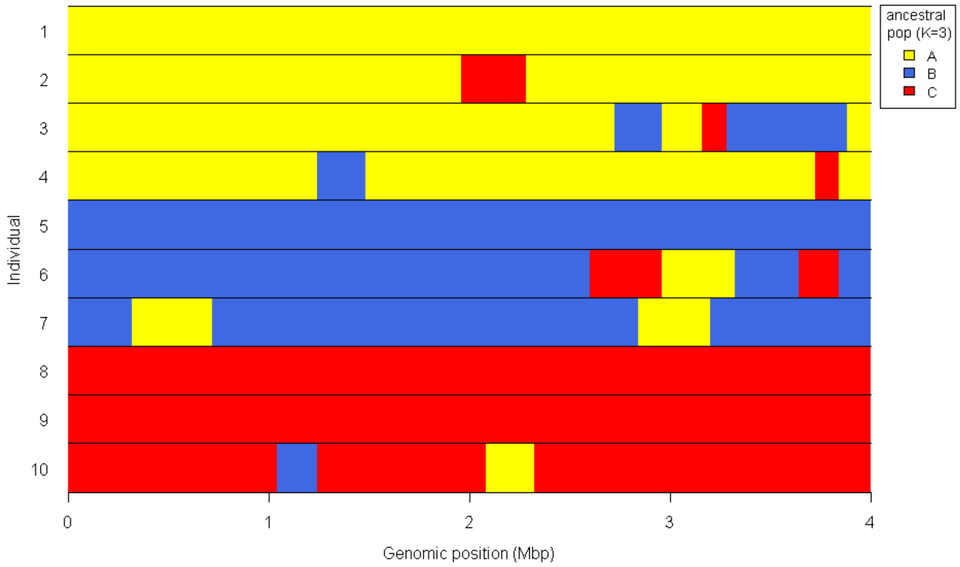


**Figure 2.1. Typical STRUCTURE output.** STRUCTURE output for an example dataset ($N = 10$) is represented graphically with STRUCTURE PLOT [162]. The STRUCTURE linkage model has assigned to each locus in each genome a set of ancestry proportions, indicating the probability of inheritance from the $k$ ancestral populations. Individual genomes are represented in rows and genomic positions are indicated along the x-axis. Linked blocks of loci are represented in columns, and their most likely ancestral population is indicated in colour, according to the legend. Sampled isolates can be clustered into populations on the basis of this ancestral inference.

These probabilistic approaches enable the objective exploration of complex multidimensional parameter spaces. The pre-specification of population genetics model parameters also allows for the incorporation of non-sequence information. Depending on the approach in question, it may be possible to permit or prohibit admixture, set upper and lower bounds on $k$, and allow or prevent hierarchical clustering. Users may also be able to include spatial information and specify prior probabilities that either favour or constrain genetic clustering to occur among physically proximate individuals.

Along with this expanded set of options, however, comes an increasing amount of subjective decision-making on the part of the user. Because the clustering procedure is guided by an explicit population genetics model, the output of these parametric methods relies on a number of assumptions. The performance of model-based methods can be highly sensitive to the assumptions made, although it is often unclear whether these assumptions can be justified or even tested. Among Bayesian clustering methods like STRUCTURE [161] and BAPS [163], for example, it is commonly assumed that Hardy-Weinberg equilibrium is upheld within populations and that representative sampling has drawn fairly unrelated individuals from the wider population [163,167]. Yet, this is rarely the case in microbial GWAS samples. Violation of these assumptions can strongly bias estimates of $k$ and make cluster memberships unreliable. In microbial GWAS, it may be difficult to support the assumptions made by model-based clustering methods [104].

### 2.2.5 Correcting for ancestry with clustering methods

The clusters identified by model-based methods like BAPS [163] can be used to control for the potential confounding effects of ancestral relatedness, either through stratification or within regression models of association. Cluster-based controls have been successfully applied within microbial GWAS [49,80,120,124,126,168]. Cluster-based corrections for population structure in microbial GWAS are usually implemented by stratifying the association test [49,120,124,126]. In stratified analyses, most commonly performed using the CMH test [144], the association test is repeated within each population cluster and the results are pooled. The CMH test [144] uses a $X^2$ test of association to examine the relationship between genotype and phenotype within a set of $k$ contingency tables, as illustrated by Table 2.2, where $i \in \{1, ..., k\}$.

| | **Genotype 1** | **Genotype 0** | Row Total |
|---|---|---|---|
| **Phenotype 1** | $A_i$ | $B_i$ | $N_{1i}$ |
| **Phenotype 0** | $C_i$ | $D_i$ | $N_{2i}$ |
| Column Total | $M_{1i}$ | $M_{2i}$ | $T_i$ |

**Table 2.2. Stratified 2x2 contingency table.**

The combined odds ratio (OR) for $k$ 2x2 contingency tables can be calculated as:

$$\mathbf{OR} = \frac{\sum_{i=1}^{K} \frac{A_i D_i}{T_i}}{\sum_{i=1}^{K} \frac{B_i C_i}{T_i}} \tag{2.2}$$

The $X^2$ distributed CMH test statistic is used to assign a p-value to the observed OR and to thus determine the statistical significance of the association being investigated. The null hypothesis of no association is rejected when $OR \neq 1$. This indicates that the genotype and phenotype are not statistically independent in each sub-population. Because the CMH test is designed to operate on contingency tables, it is applicable only to nominal variables. In practice, CMH tests have been used to control for population structure in binary analyses of antibiotic resistance in *M. tuberculosis* [126] and *S. pneumoniae* [49], and host association in *Streptococcus suis* [124].

Clusters can also be used to account for ancestry in GWAS in regression-based approaches [124, 168]. Weinert *et al.* [124], for example, use sets of hierarchical clusters identified with hierBAPS as covariates within a logistic regression model to test for association with a binary host association phenotype in *S. suis*. By including cluster membership as a fixed effect, the association between genotype and phenotype can be examined while excluding the potential confounding effects of population structure at a particular level of genetic clustering [150]. If genotype is not causally related to phenotype, but instead differences in both genotype and phenotype correspond to differences in the cluster membership of individuals, the incorrect inference of association may be avoided by including clusters as covariates in a regression model. Regression provides an alternative to the CMH test that renders cluster-based controls applicable to continuous as well as categorical variables. Regression models also make it possible to include additional covariates, to control for confounding effects that might be introduced by environmental factors or other phenotypes, where this information is known.

Although model-based clustering methods have evolved to allow for admixed or hierarchical populations [163–165], cluster-based corrections for ancestral relatedness in GWAS are not equipped to make use of the partial or nested cluster membership statuses of individual genomes. Chewapreecha *et al.* [49], for example, use a set of previously-identified [169] hierarchical BAPS clusters in a CMH test to stratify their association study of antibiotic resistance in a large sample of 3,701 *S. pneumoniae* isolates. Whether the set of 33 primary clusters or 183 secondary clusters will better control for confounding can only be determined by repeating the analysis and comparing genomic inflation. This reveals a substantial decrease with the larger $k$, from $\lambda_{GC} = 6.58$ to $\lambda_{GC} = 2.56$. However, even the reduced value remains well above the accepted $\lambda_{GC} < 1.05$ threshold, indicating that association test p-values may still be inflated considerably by population stratification [150]. In these results, confounding effects can not be sufficiently eliminated by either cluster-based control.

A major limitation of cluster-based controls is that they assume that the genetic variation remaining within each population cluster is ancestrally homogenous. In microbial samples, it can be especially challenging to identify a suitable set of $k$ distinct, internally-homogenous clusters, as clonal inheritance favours heterogeneous, hierarchical structures while recombination often blurs the boundaries between sub-populations. Furthermore, the effectiveness of these techniques depends heavily on the appropriateness of the number of clusters that are taken to represent population structure. Using too few clusters will increase the type I error rate of the GWAS by violating the assumption of genetic independence upon which association testing proceeds. Including too many clusters will, conversely, increase the type II error rate by excluding relevant genetic variation from the analysis. Methods exist to help select the $k$ that best describes the population structure of a sample [170–172]. But, without considering the phenotypic states of individuals, the $k$ clusters identified by these methods may not be those most relevant to the problem of population stratification in a GWAS study. Whether the phenotype is clustered in large clades, small groups, or not at all affects the probability of spurious association due to ancestry (see Figure 2.2). While both genotypic and phenotypic variance ought to influence the choice of $k$, there is still no widely-accepted protocol for making this decision objectively. Moreover, as cluster memberships are often unstable when $k$ changes, the effectiveness of cluster-based controls may vary unpredictably [160].

Clusters provide a clear, albeit simplified, representation of the ancestral relationships between isolates. Although more detailed reconstructions might be inferred by phylogenetic methods, reliable clonal genealogies may not be attainable for highly-recombinant microbes. Model-based clustering methods present a straightforward alternative approach to ancestry inference. Easily incorporated via the stratified CMH test or regression, cluster-based controls have become one of the most popular means of addressing population stratification in microbial GWAS [49, 80, 120, 124, 126, 168].

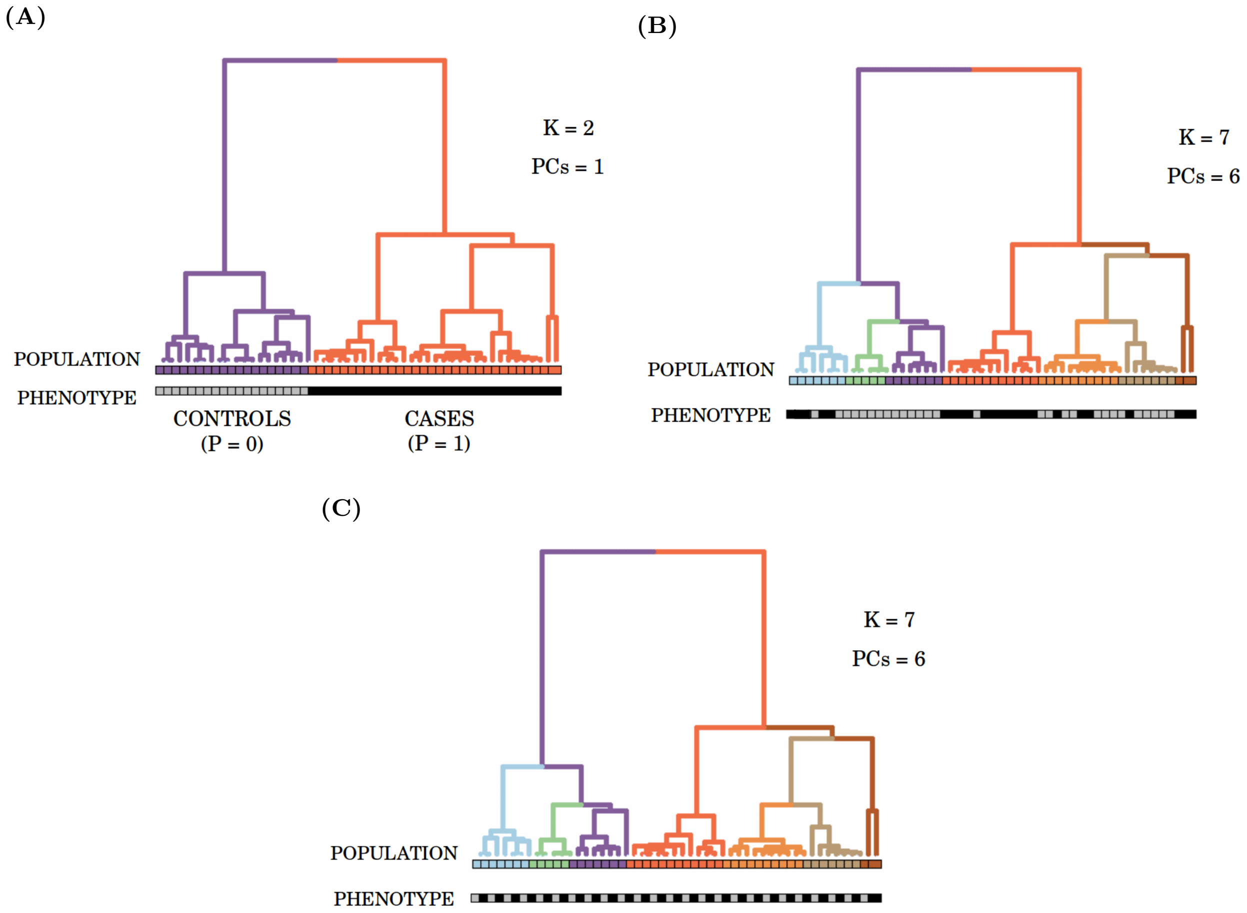


**Figure 2.2. Population stratification.** Population stratification varies with genetic differentiation and phenotypic clustering. Colours along each tree represent populations, as defined by $k$ clusters or $(k-1)$ PCs, and tip colours indicate phenotype (controls = grey, cases = black). **A:** Maximal population stratification. All of the substitutions that separate the two major clades would appear to be associated with the phenotype, even if only one of these truly caused the change. **B:** Moderate population stratification. With the tree topology of (A), the stronger population differentiation may reflect geographic or sampling differences. It may be possible to correct for this degree of population stratification. **C:** Minimal population stratification. Despite the population structure of (B), because the phenotype does not cluster within the ancestral lineages, there is no need to correct for population stratification.

### 2.2.6 Model-free dimension reduction methods

Model-free multivariate dimension reduction techniques allow for a more detailed description of population structure than the discrete clusters identified by model-based methods. Unlike Bayesian clustering algorithms, these approaches are not based on an explicit population genetics model. Instead, multivariate methods aim to summarise genetic sequence data in a set of orthogonal (linearly uncorrelated) PCs of decreasing variance, positioning all sampled individuals along these major axes of variation.

**PCA** uses the approach in Box 2.2 to summarise the variation in a genetic dataset within a reduced set of orthogonal PC axes, or weighted linear combinations of the original genetic variables whose squared coefficients sum to one [110, 173]. Figure 2.3 shows how $k-1$ PCs separate $k$ populations in PCA space [174].

1. Let $\mathbf{X}$ be a matrix containing genotypes for individuals $i$ and polymorphic loci $j$, where $j \in \{1, ..., p\}$ and $i \in \{1, ..., n\}$.
2. Compute the covariance matrix, $\mathbf{C}$, of the sample via Equation 2.3, where $\mathbf{C_{jj}}$ is the covariance of column $j$ and column $j$ of $\mathbf{X}$, and $\mathbf{C}$ has $n-1$ non-zero eigenvalues and orthogonal eigenvectors.
3. Identify the first PC axis as the eigenvector of $\mathbf{C}$ that contains the greatest variance and has the largest eigenvalue.
4. Identify the orthogonal PC with the next-largest variance as the second PC axis.
5. Repeat the previous step until all of the variance in the original dataset is summarised in the reduced set of PC axes.

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}\mathbf{X}' \tag{2.3}$$

**Box 2.2. PCA protocol**

A large number of alternatives to PCA can be used to perform similar ordinations in reduced space, but under different optimisation criterion [175]. PCA aims to separate individual datapoints in multidimensional space by maximising the overall variance, or squared Euclidean distance. The related metric multi-dimensional scaling (MDS) [176] approach operates on any Euclidean distance, for example, enabling a PCA-like solution for a distance matrix constructed from k-mers [129]. By contrast, discriminant analysis (DA) [177] adopts a supervised approach that aims to maximise the distances between groups only. Yet, DA can rarely be performed on genetic sequence data, because it

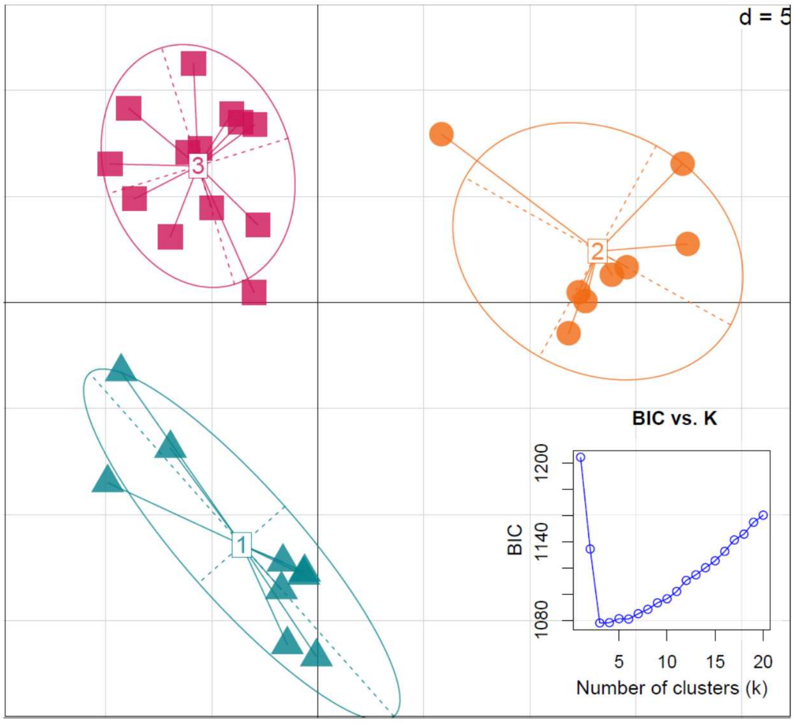


**Figure 2.3. Typical PCA output.** This figure contains a two-dimensional representation of the output of a two-step procedure: (i) applying k-means clustering, and (ii) PCA to an example dataset ($N = 30$, $k = 3$). The inset plot at the bottom right displays the Bayesian Information Criterion (BIC) curve used to select the $k$ that best fits the data. A distinct optimal (minimum) BIC value is achieved at $k = 3$, after which point BIC values begin to climb again. The individual composition of the three population clusters is defined by k-means clustering with $k = 3$. In the main plotting area, all 30 individuals are projected onto the first two PCs (the most significant axes of variation). Each datapoint represents a bacterial isolate. The shape and colour of these datapoints distinguish isolates by population (k-means cluster). Each population is accompanied by an ellipse whose area corresponds to its variance. Because PCA maximises overall variation, PCA plots capture the variance that exists both between clusters and within them. If DAPC were used instead of PCA, the within-group component of overall variance would be minimised and the resulting plot would show tighter clusters.

requires that there be fewer columns of genetic variables in a dataset than there are rows of individuals. To overcome this obstacle, the discriminant analysis of principal components (DAPC) [145] approach can be used to render DA applicable to highly multivariate genome-wide data, by first transforming genetic data into a reduced set of PCs with PCA. The reconstructions of population structure generated by each of these multivariate methods can offer valuable insights into the population structure of microbial samples.

Multivariate methods return output that reflects genetic sequence variation alone, untainted by the potentially poorly-selected priors of model-based clustering methods. Whereas model validation entails the computationally-intensive generation and likelihood-based comparison of multiple models, model-free inference can be made from the single set of eigenvectors and eigenvalues that reliably result from the application of a given dimension reduction method to a particular dataset. Multivariate methods are consequently computationally efficient and scale well to large genetic datasets.

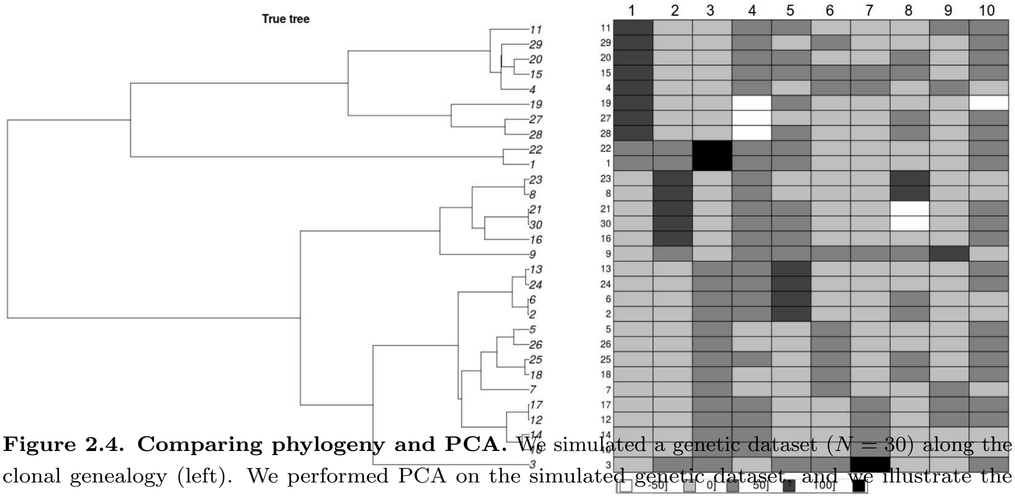


**Figure 2.4. Comparing phylogeny and PCA.** We simulated a genetic dataset ($N = 30$) along the clonal genealogy (left). We performed PCA on the simulated genetic dataset, and we illustrate the coordinates of each individual (terminal node) along the ten PCs with the largest eigenvalues (right) according to the shades of grey indicated in the legend (below). The genealogical interpretation of many PCs can be inferred from their coordinates, particularly in the most significant PCs. However, it becomes increasingly difficult to predict, if not to interpret, the relationship between PC and tree structure as we move right-ward into higher PC dimensions.

Dimension reduction can reconstruct ancestral relationships with greater resolution than model-based clustering methods. By positioning all individuals at coordinates along each synthetic PC axis, multivariate methods can not only indicate clusters, but describe the relationships between clusters and among individuals within clusters, potentially revealing clades, admixture, and clines [145]. The PC axes identified often have genealogical interpretations, as Figure 2.4 demonstrates [158]. Yet, relevant population structuring variation is not always neatly captured within a set of significant PCs [178]. Aside from the differences between ancestral populations, PCs may be shaped by:

- How many and which genetic markers are included [174].
- Local LD, rather than genome-wide population structure [165, 179].
- Assay artefacts, variable data quality [180].

- Artefacts of sampling, sub-population sample size [158].
- Within-population variance [181].
- Variation introduced by recombination (see 2.3.2).
- Variation associated with a phenotype [182].

Nevertheless, as it is applicable to recombining organisms, and it can offer a more informative description of population structure than clusters alone, PCA has become the dominant means of controlling for population structure in human GWAS [150] and an increasingly prevalent approach in microbial GWAS [116, 119, 125, 128, 183]. Related dimension reduction methods have also been adopted in microbial GWAS [80, 124, 129], but how they are used and with what effectiveness remains far from settled.

### 2.2.7 Correcting for ancestry with dimension reduction methods

In microbial GWAS, multivariate methods can be used to control for the confounding effects of population structure by regressing along significant axes of variation. Unlike corrections made with GC, the corrections applied to genetic markers via PCA are not uniform and, instead, vary site by site, such that greater control is exerted over loci that exhibit large differences in allele frequencies across ancestral populations [150].

The approach adopted in the original EIGENSTRAT software [107] and used in human GWAS is to directly adjust the original genotypes of individuals by the ancestry proportions reflected in significant PCs. First, PCA is performed, and the set of PCs that are identified as significant are taken to represent the population structure of the sample. The human GWAS literature stresses the importance of performing PCA on unlinked, phenotype-independent genetic markers [107, 150, 174]. Yet, this issue has been overlooked repeatedly in PCA-based microbial GWAS studies [116, 119, 122, 125]. A linear regression is performed, modelling the genotype at each locus $g_{ij}$ as a function of the population-structuring axes of variation. This allows for the identification of regression coefficients specifying the degree to which genotypes $g_{ij}$ are predicted by ancestry alone. The original genotypes $g_{ij}$ are "corrected" by subtracting this ancestry proportion. The same transformation process may also be used to adjust the phenotypic values by ancestry [107]. Association testing is then carried out on the corrected genotypes and phenotypes. Alternatively, a linear or logistic regression can be used to model the phenotype as a function of both the genotype and the set of significant PCs, which are included as fixed effects [129]. Regression with PCA [116, 119, 122, 125], MDS [129, 132] and DAPC [80] have all been used to correct for ancestry in microbial GWAS.

Multivariate approaches can reduce the false positive rate attributed to clonal population structure, although they are often less effective than phylogenetic approaches [97, 132]. Unfortunately, these approaches also exclude potentially-relevant variation contained in the set of significant PCs, which can substantially reduce the power to detect genuine associations [128]. This problem is aggravated in microbial GWAS when selective pressures acting on the phenotype also impact the population structure, entangling phenotypically-associated polymorphisms and population-stratified variation. Additionally, because the population structure inferred by PCA reflects variation due to both mutation and recombination, correcting genetic data with PCA may eliminate trait-associated variation instead of ancestral differences.

Like $k$ selection in clustering methods, selecting the optimal number of PCs to control for population structure in GWAS can be difficult, although a multitude of PC selection methods exist [172, 184]. Approaches from human GWAS recommend that the set of PCs account for a significant proportion of the total genomic variation [107, 174]. In practice, most PCA-based microbial GWAS studies have selected the number of PCs subjectively, by visually inspecting eigenvalue distributions or PCA plots. Moreover, with any number of PCs, multivariate approaches often struggle to both maintain power and eliminate false positive findings in microbial GWAS [97]. Using PCs as fixed effects has been found to be effective in analyses of phenotypes under strong selection in single HIV lineages with high levels of recombination and weak population structure [116, 119]. But, where population differentiation is strong or the phenotype clusters within ancestral lineages, PCA-based regression may only detect associations occurring at the tips of a phylogenetic tree while overlooking associations that arise over the evolutionary history of a sample [97, 132].

Recent proposals have attempted to recover the power lost by dimension reduction methods. Instead of using a set of significant PCs as fixed effects within a regression model, Earle *et al.* [128] use the entire correlation matrix between strains as a set of random effects within a linear mixed model (LMM). This dramatically reduces the type I error caused by population structure by capturing the extent to which genetically similar strains are phenotypically similar, but also increases the type II error rate [150]. Associations that may be genuine but also correlated with the population structure of the sample are captured within the set of background random effects. These associations are thus excluded from the set of “locus effects” identified by mixed models. Earle *et al.* [128] attempt to compensate for this behaviour by recovering population-stratified signals of association in a secondary set of “lineage effects”, in the form of trait-associated PCs. It does not necessarily seem desirable, however, to segregate associations into these two sets of effects, and we believe this can be avoided.

Both clustering and dimension reduction methods provide useful alternatives to phylogenetic methods for counteracting biases resulting from the genetic relationships between isolates. For organisms that undergo very high levels of recombination, especially in diverse species-wide samples, even phylogenetic methods that account for recombination are unlikely to reconstruct accurate genealogical trees [185]. Where minimal clonality remains, non-phylogenetic approaches may be essential. In such cases, PCA- or cluster-based solutions are likely to be sufficient to control the confounding bias attributable to ancestry.

On the other hand, for organisms where the clonal genealogy is not obscured by recombination, the spurious associations generated by strong ancestral relationships are unlikely to be kept in check by the simplified representations of population structure provided by a set of clusters or significant PCs [97]. In a large proportion of microbial GWAS studies, therefore, a phylogenetic approach may be more effective.

### 2.2.8 Phylogenetic methods ignoring recombination

When attempting to reconstruct the ancestral relationships between bacterial genomes, phylogenetic methods are the most obvious choice of approach. Phylogenetic methods generate detailed reconstructions of population structure at all levels. Whether describing genetic relationships on an evolutionary timescale or revealing genealogical links between close relatives, phylogenetic trees can provide an intuitive representation of the ancestral relationships between microbes. Among the older and more commonly-encountered phylogenetic methods are those that ignore recombination. These methods assume that the evolutionary history of all loci in the genomes of sampled isolates can be adequately described by a single clonal genealogy. Hence, in modelling bacterial population structure, standard phylogenetic methods are most useful for predominantly clonal organisms that undergo minimal recombination (see Table 2.1).

**Distance-based methods** aim to identify the tree that results from progressive agglomerative clustering of similar individuals, taking the steps outlined in Box 2.3.

1. Define the distance $d_{i,j}$ between sampled individuals $i$ and $j$ as the proportion of genetic loci at which they differ.
2. Compute the distance matrix, $D_{i,j}$, containing the distances $d_{i,j}$ between all pairs of sampled individuals.
    (a) For NJ method, let $D_{i,j} = d_{i,j} - \sum_{k \neq i} d_{i,k}/(n-2) - \sum_{k \neq j} d_{j,k}/(n-2)$.
3. Cluster together the two individuals or clusters separated by the smallest $d_{i,j}$.
4. Update $D_{i,j}$ to reflect the grouping of individuals $i$ and $j$ into the cluster $n$. Define the distance $d_{k,n}$ between an individual or cluster $k$ and cluster $n$ according to the method you are using by selecting the appropriate formula from Table 2.3.
5. Repeat Steps 3 – 4 until all individuals have been merged into a single cluster.

**Box 2.3. Distance-based phylogenetic method**

An array of distance-based methods exist, each relying on a unique criterion to identify the clustering order, including Complete-Linkage [188], Single-Linkage Agglomerative Clustering [187], Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [187], and Neighbour-Joining (NJ) [186]. Distance-based methods run in polynomial time, enabling rapid analyses even for large datasets (e.g., N=1,000) [189].

The most popular distance-based method is the NJ approach [186] and its extensions (e.g., BIONJ [190], FastME [191]). Many studies use NJ as a first approach to show relationships between bacterial pathogen genomes, before more complex methods are applied, because it is easy to apply and has well known properties [118, 147]. By contrast to the linkage clustering procedures, which assume a molecular clock [192], the NJ algorithm accounts for heterogeneous evolutionary rates. Consequently, linkage procedures such as UPGMA output rooted, ultrametric trees where the distance from the root to any leaf is identical, whereas, NJ outputs unrooted, non-ultrametric trees (see Figure 2.5).

| |
|---|
| **NJ** |
| Neighbour-Joining [186] |
| $d_{k,n} = (d_{k,i} + d_{k,j} - d_{i,j})/2$ (2.4) |
| **UPGMA** |
| Unweighted Pair Group Method<br>with Arithmetic Mean [187] |
| $d_{k,n} = (N_i d_{k,i} + N_j d_{k,j})/N_n$ (2.5) |
| **Complete** |
| Complete Linkage<br>Agglomerative Clustering [188] |
| $d_{k,n} = max(d_{k,i}, d_{k,j})$ (2.6) |
| **Single** |
| Single Linkage<br>Agglomerative Clustering [187] |
| $d_{k,n} = min(d_{k,i}, d_{k,j})$ (2.7) |

**Table 2.3. Combinatorial formulas for distance-based trees.** $d_{k,n}$ is the distance from an observation/cluster $k$ to a cluster $n$ containing observations/clusters $i$ and $j$.

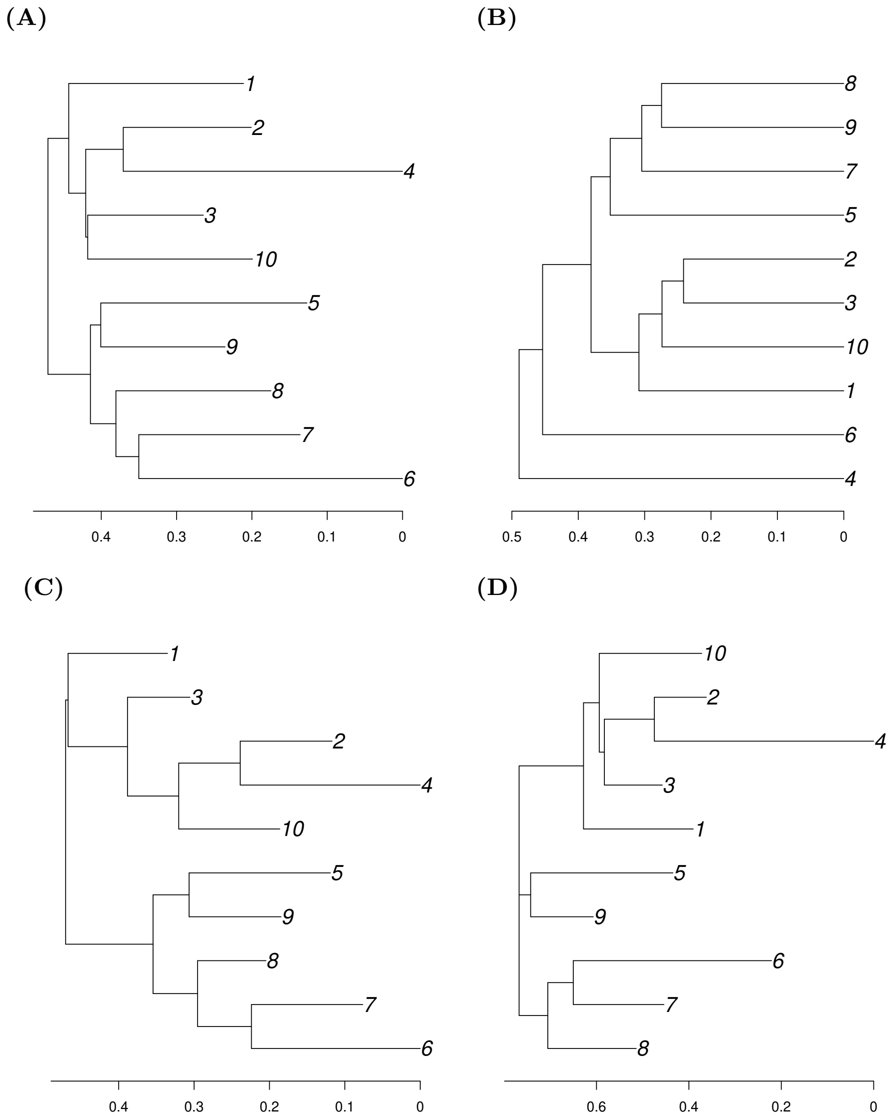


**Figure 2.5. Typical phylogenetic output.** Four standard phylogenetic reconstruction methods have been applied to the same genetic dataset ($N = 10$). Their output is shown, with an axis indicating branch length in units of substitutions per site. **A:** NJ outputs an unrooted, non-ultrametric tree. A rooted representation is shown for comparative purposes. The NJ tree suggests that the molecular clock hypothesis does not hold true for this dataset: compare the branch lengths connecting nodes 2 and 4 to their common ancestor. **B:** UPGMA outputs a rooted, ultrametric tree. UPGMA assumes that evolution occurs at a constant rate, and its estimation is not robust to violation of the molecular clock. **C:** Parsimony, with edge lengths estimated by ACCTRAN [193], outputs the shortest possible tree which, here, is similar to the NJ reconstruction. **D:** ML estimates a phylogeny that is topologically similar, in most respects, to the NJ and parsimony trees. The ML tree has comparatively long branch lengths, which are based on evolutionary rates rather than genetic distance or homoplasy.

**Maximum-parsimony methods** aim to identify the tree that requires the smallest number of substitutions to explain the data, using the procedure in Box 2.4.

1. Select an initial tree topology.
2. Compute parsimony cost.
3. Apply a random modification to the tree.
4. Compute new parsimony cost.
5. Accept new tree if it reduces the parsimony cost; else, keep previous tree.
6. Repeat Steps 3 – 5 until no further improvements can be found.

**Box 2.4. Maximum-parsimony phylogenetic method**

The dnapars algorithm in PHYLIP [194] is among the most popular implementations of parsimony. In practice, a reasonable starting tree is often estimated with a fast approach like NJ [186]. We describe the methods used to assign branch lengths to parsimony trees later in this chapter (see Box 2.9). Although the parsimonious phylogenetic reconstruction method is not the most frequently cited in the bacterial genomics literature, many examples of its successful application can be found, especially for the study of closely related genomes within genetically monomorphic pathogens, such as *M. tuberculosis* lineage Beijing [195], *Y. pestis* [196, 197], and *S. enterica* serovar Agona [198].

Parsimony methods operate according to a single criterion: minimisation of the parsimony cost of the tree. This approach requires few assumptions and does not attempt to overcomplicate the model of evolution. Parsimony trees represent a “minimum evolution” scenario and provide the simplest phylogenetic explanation of the data. Although this may not perfectly reconstruct the genuine evolutionary history of the sample, it can nevertheless offer a useful representation of the relationships between sampled genomes.

**Maximum-likelihood methods** aim to simultaneously estimate a phylogenetic tree and set of evolutionary model parameters. They use the approach described in Box 2.5 to select those which achieve the highest probability of observing the genomic data [199]. Among the most popular maximum-likelihood methods for bacterial pathogen genome analysis are PhyML [200], RAxML [201], GARLI [202] and FastTree [203, 204].

1. Select an initial tree and parameters of the model of sequence evolution.
2. Compute likelihood.
3. Propose changes to the tree and parameters.
4. Compute new likelihood.
5. Accept proposed changes if likelihood increased; else, reject changes.
6. Repeat Steps 3 – 5 until no further improvements can be found.

**Box 2.5. Maximum-likelihood phylogenetic method**

ML methods take a more complex approach than parsimony methods and require a model of sequence evolution and evaluate the probability associated with different substitution rates, which are allowed to vary between sites and across the tree. This rate-based approach may estimate longer branch lengths than standard parsimony or distance-based methods, as repeated substitutions are permitted (see Figure 2.5. ML reconstructions may offer a more realistic representation of the true evolutionary history of a sample, which is often less straightforward than a parsimonious reconstruction might suggest. ML methods are more computationally intensive than the methods described above. ML methods are often used to reconstruct bacterial phylogenies, with many appearing in high profile studies, for example, in *V. cholerae* [205], *S. aureus* [206], or *C. trachomatis* [153].

**Bayesian methods** aim to simultaneously estimate a phylogenetic tree and evolutionary model parameters, selecting a sample of trees from the posterior probability distribution. Bayesian phylogenetic methods use the approach outlined in Box 2.6.

1. Select an initial location in the parameter space (defined by tree topology, branch lengths, and parameters of the model of sequence evolution).
2. Propose changes to the parameters according to a proposal distribution.
3. Compute the Metropolis-Hastings ratio [207, 208], $R$, between the previous and proposed parameter values.
4. If $R > 1$, adopt the proposed changes. If $R \leq 1$, accept the changes only if $u < R$, where the value of $u$ is randomly sampled from $U(0, 1)$.
5. Repeat Steps 2 – 4 until convergence is achieved.

**Box 2.6. Bayesian phylogenetic method**

Standard Bayesian methods include those offered by MrBayes [209], BEAST [210], and BEAST2 [211]. Like ML methods, Bayesian approaches require a model of sequence evolution to reconstruct a phylogenetic tree, although Bayesian evolutionary models can be more complex. In addition, Bayesian methods allow for the specification of prior probability distributions, which permit the user to either inform or bias the outome [212]. In bacterial population genomics, BEAST is a popular choice for reconstructing a timed phylogeny, where leaves are aligned with their known sampling dates and the age of ancestors is estimated. In bacterial GWAS, Bayesian methods may be useful for studies carried out longitudinally [132]. The MCMC approach allows Bayesian methods to identify a point estimate of the phylogeny, or to sample a set of possible trees from the posterior probability distribution. Where phylogenetic uncertainty is a concern, this feature may be useful in either quanitifying the uncertainty present or as a means of comparing alternative representations of the phylogenetic relationships between isolates.

### 2.2.9 Phylogenetic methods accounting for recombination

Recombination disrupts the pattern of clonal inheritance on which traditional phylogenetic methods rely. Failing to adequately account for recombination when constructing a phylogeny can obscure the true clonal relationships between isolates [213–215]. Even very low levels of recombination can cause standard phylogenetic methods to produce trees that are topologically inaccurate [215] and have distorted branch lengths [213, 216]. The evidence suggests that, if recombination is present, the conclusions drawn from standard phylogenetic inference should be questioned [217].

Some authors have suggested that by removing recombinant regions, the clonal frame [218] can be revealed using standard phylogenetic methods [196, 219, 220]. But, the evidence suggests that this removal can, in fact, intensify the distortive effect that recombination has on the tree [216]. Phylogenetic methods that explicitly account for recombination offer a more appropriate solution in the case of clonally-related recombinant organisms. As clonality persists across the genome wherever recombination is not taking place, it remains possible to infer the clonal genealogy even in the presence of recombination.

**ClonalFrameML** aims to reconstruct the clonal genealogy, while accounting for recombination and identifying the location of recombinant regions, as in Box 2.7

1. Construct an initial maximum-likelihood tree.
2. Reconstruct ancestral sequences using maximum-likelihood.
3. Estimate branch lengths, recombination parameters (rate, length of events, average donor/recipient distance), and locations of recombination events for each branch via Baum-Welch Expectation-Maximisation algorithm.
4. Estimate uncertainty using a bootstrapping procedure.

**Box 2.7. ClonalFrameML recombination-aware phylogenetic method**

ClonalFrame [143, 221] is one of the most frequently-used phylogenetic approaches that explicitly models and accounts for recombination. The original ClonalFrame [143] worked well for MLST data or very few genomes, identifying clonal genealogies in samples of *C. trachomatis* [222, 223] and *E. coli* [152], which were found to recombine at low to moderate rates, respectively, across or within lineages. ClonalFrameML [221] offers a more efficient alternative that scales well to hundreds of genomes. This implementation has been applied to organisms ranging from the relatively clonal *M. tuberculosis* [224] to the moderately- and highly-recombinogenic *C. jejuni* [225] and *S. pneumoniae* [226]. ClonalFrame [143] is the only method we are aware of that has been used to reconstruct a recombination-aware phylogenetic tree in microbial GWAS. It is used to identify associations with host source in one *C. jejuni* dataset, in the simulation-based approach of Sheppard *et al.* [118] and in the phylogenetic samping strategy of Farhat *et al.* [106].

**Gubbins** aims to reconstruct the clonal genealogy, while finding and excluding regions of likely recombination, using the approach in Box 2.8.

1. Construct initial maximum-likelihood tree.
2. Reconstruct ancestral sequences using maximum-likelihood (FastML [227]).
3. Identify putative recombinant regions (i.e. clusters of substitutions unlikely to have arisen through point mutation) using a sliding window scan.
4. Remove putative recombinant sites.
5. Iterate through Steps 1-4 until convergence occurs.

**Box 2.8. Gubbins recombination-aware phylogenetic method**

Gubbins [185] has been used many times to examine the population structure and dynamics of *S. pneumoniae* [169, 228–231], the pathogen for which it was initially designed, as well as in analyses of *C. trachomatis* [153] and *L. monocytogenes* [232]. However, in their *H. parasuis* GWAS, Howell *et al.* [80] find that the need to remove recombinant sites prevents Gubbins from identifying a core-genome phylogeny.

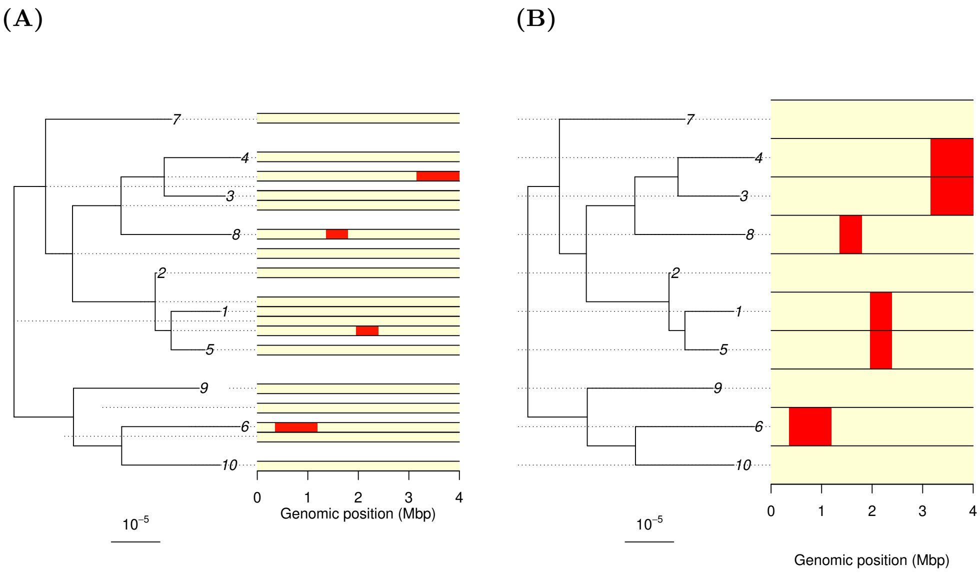


**Figure 2.6. Typical recombination-aware phylogenetic output.** These figures show the output obtained from two phylogenetic reconstruction methods that account for recombination, **A:** ClonalFrameML and **B:** Gubbins, when applied to the same example dataset ($N = 10$). Each figure contains the inferred clonal genealogy (left) and a representation of the inferred genomic locations of recombination (right). Coloured regions represent recombinant loci occurring at positions in the genome indicated by the scale that runs along the x-axis. A key difference between the two methods can be seen in the right-hand panels of each figure. In (A), ClonalFrameML indicates recombination among both terminal and ancestral nodes; whereas, in (B), Gubbins indicates recombination in terminal nodes only, whether recombination events have occured on terminal or internal branches.

Both Gubbins [185] and ClonalFrameML [221] return rooted, non-ultrametric trees and indicate any recombination events inferred in the sampled genomes (see Figure 2.6). Instead of accounting for recombination like ClonalFrameML [221], Gubbins [185] eliminates recombinant regions when reconstructing the clonal genealogy. In addition, ClonalFrameML [221] identifies recombinant regions in both the sampled genomes of terminal nodes and the un-sampled genomes of ancestral nodes, focusing on processes between internal and terminal nodes. Gubbins [185], on the other hand, delimits recombinant regions only in the set of sampled genomes, focusing on outcomes at the terminal nodes of the genealogy.

The use of recombination-aware phylogenetic methods in GWAS has been recommended since the undertaking of association studies in bacteria was first considered [91]. These approaches provide a high-resolution reconstruction of the ancestral relationships between isolates. Critically, they make it possible to identify clonal relationships while accounting for the fact that recombination can introduce multiple polymorphisms in single events and can transfer DNA between close relatives, distant lineages, and even separate species. By using recombination-aware phylogenetic approaches in GWAS, we can ensure tight control over the confounding effects of ancestral relationships while revealing instances of phenotypic differentiation between close relatives, which are of great relevance to association studies.

### 2.2.10 Correcting for ancestry with phylogenetic methods

Although clonality poses a major challenge to microbial GWAS, it also enables the adoption of phylogenetic solutions [106,114,118,121,127,130,131]. Phylogenetic trees allow for the detailed identification of genetic relationships, not only at the level of population clusters, but also at the resolution of subpopulations and individual relationships. Where clonal relationships among isolates are apparent, phylogenetic methods are likely to offer better control over type I errors arising from ancestry in microbial GWAS than non-phylogenetic approaches [97,183]. And, thanks to recombination-aware phylogenetic methods [143,185,233], the adoption of a tree-based approach in GWAS does not require evolution to be treated as purely clonal, nor that recombination be ignored, since the effects of recombination events can be considered within a phylogenetic framework. Phylogenetic approaches are by far the most popular method to describe microbial population structure, and therefore they are a natural option to control for population structure when performing GWAS in microbes. The literature displays a variety of ways in which phylogenetic information can be incorporated in microbial GWAS.

A number of microbial GWAS studies permute the phenotype along the tips of the tree to identify a significance threshold [49,116]. However, these methods rely on an assumption of exchangeability that is violated by the varying degrees of genetic relatedness between individuals. Hence, phenotypic permutation approaches provide a poor solution to the problem of population stratification in microbial GWAS, because they do not account for genetic covariance. Within-cluster label switching has been proposed as a means of rendering permutation techniques more robust to genetic non-independence; but this approach may, conversely, be overly conservative [183].

Farhat *et al.* [106] and Brynildsrud *et al.* [127] have, respectively, proposed a phylogeny-based sampling strategy and a pairwise comparisons approach [234]. Both select a reduced set of phenotypically discordant but genetically proximate pairs of isolates from within the wider phylogeny. This removes ancestral correlations between genotype and phenotype, and makes it possible to search for genotypic and phenotypic homoplasies that may indicate convergent evolution. By design, however, only $\leq N/2$ comparisons can be made, so a substantial proportion of the dataset is exclued from the association test. This information loss reduces the statistical power to detect associations. The association score adopted by Brynildsrud *et al.* [127] further restricts their approach to binary phenotypes only. As phylogenetic corrections for population stratification can be achieved without eliminating valuable data [114, 118, 123, 130, 235], alternative approaches are preferable to sampling strategies and pairwise approaches.

More sophisticated approaches further capitalise on the phylogenetic framework, by using it to infer ancestral states and thus to expand the association test into ancestral lineages. The approach of Hall *et al.* [121] is relatively straightforward. The authors apply a $X^2$ test of association, first, to the genotypic and phenotypic ancestral states inferred at internal nodes and, second, to the substitutions inferred along the branches of the tree. Because these tests are performed sequentially, they successively narrow down the set of significant findings. To be deemed significant by this approach, an association must be broadly upheld throughout the evolutionary history of the sample, and it must also display correlated evolution. Desjardins *et al.* [130] use BayesTraits [142], an approach developed for use in phylogenetic comparative methods, to test for association with antibiotic resistance in *M. tuberculosis*. BayesTraits [142] uses MCMC to estimate evolutionary rates for the genotype and the phenotype. It then performs a tree-based test for correlated evolution by calculating the likelihood of models of the dependent and independent evolution of both variables and comparing the two hypotheses with a likelihood ratio test [236]. The approaches of Hall *et al.* [121] and Desjardins *et al.* [130] are able to extend the association test across the evolutionary history of the sample. Both approaches rely heavily on the accuracy of their reconstructions of the phylogeny and set of ancestral states or evolutionary rates, which may render them highly sensitive to uncertainties in these estimates [237].

Two other bacterial GWAS methods use simulation-based approaches, inspired by phylogenetic comparative methods [238, 239], to obtain phylogenetically-correct null distributions of association test statistics [114, 118]. Sheppard *et al.* [118] use simulations to generate null genetic data by randomly assigning substitutions to branches of the phylogenetic tree. Association testing is performed along the tips of the tree in both the simulated and empirical datasets to identify alleles that are over-represented alongside either binary phenotypic state. Empirical association tests statistics that fall within the simulated null distribution are inferred to be made possible by the tree structure alone and are, therefore, rejected. The "PhyC" method of Farhat *et al.* [114] uses simulations to test for convergent evolution, examining substitutions along the phylogenetic tree. Association is defined by the number of genotypic substitutions that occur alongside or after a change in phenotype. PhyC does not simulate a null genetic dataset, nor does it identify a genome-wide null distribution. Instead, it creates one null distribution for each number of genotypic substitutions inferred. Empirical associations that have undergone $N$ mutations are evaluated against hypothetical associations between the ancestral phenotype and random samples of $N$ substitutions. This establishes a conditional significance for each locus, assessing the probability that one would observe such an extreme association by chance, given the number of substitutions observed. Sheppard *et al.* [118], by contrast, work with a single null distribution to establish genome-wide significance, by measuring the probability that one would encounter such an exteme deviation *somewhere* along the genome. As this criterion aligns more closely with the goals of a GWAS study, it is probably the better approach. The simulation-based framework allows both approaches to capitalise on the high resolution reconstruction of ancestral relationships provided by the phylogenetic tree. While these simulation-based phylogenetic approaches have generated promising results thus far, this foundational work leaves open many opportunities for improvement.

### 2.2.11 Rationale for our approach

In developing our own approach to microbial GWAS, we want to build on the foundations identified by our search of the literature. This ensures that methodological development in microbial GWAS moves thoughtfully forward, rather than occurring in parallel. We determined that a simulation-based phylogenetic approach would allow our microbial GWAS method to most effectively address the biases introduced by ancestral relatedness as well as other confounding factors. A number of additional procedures were designed to overcome critical limitations and gaps in the methods proposed thus far. We aimed to distinguish our approach from existing phylogenetic methods by adding the flexibility needed to appropriately analyse a wider range of organisms, including recombinant as well as clonal bacteria. Towards this end, we planned to use a recombination-aware tree-building method and to implement new procedures for addressing recombination and mutation rates, as discussed below. A more informed data simulation procedure was developed, to allow for the estimation and incorporation of additional empirical parameters, drawn from each dataset. We wished to account for features like the empirical distribution, and mutation and recombination rates. These impact the probability of spurious association and the discovery power of association tests, yet they are overlooked by existing approaches [91, 240–242]. Since available simulation-based methods have been criticised for being too stringent, we pursued greater discovery power by implementing a new approach to association testing, as described below [183]. As phylogenetic GWAS approaches have thus far been restricted to binary phenotypes, we also sought to extend our association tests to categorical and continuous phenotypic data types. Finally, we implemented our approach in effective and accessible software, taking additional steps to improve the efficiency of our software package. This set our method apart from other simulation-based pipelines, which have been restricted in practice by burdensome computational demands on time and memory. Ultimately, we were able to release the first user-friendly, simulation-based phylogenetic GWAS tool that was tailor made for use in microbial data.

## 2.3 Accounting for recombination

While the clonal structures produced by vertical inheritance are a central feature of bacterial population genetics, we must not overlook the impact of horizontal mechanisms, which can shape evolution via recombination. Once underestimated in frequency, extent, and significance, genetic recombination has been revealed by research in recent decades to be a widespread and powerful force [243, 244]. Recombination has now been observed to varying degrees in most bacterial species [74, 245]. It can be a crucial driver of genetic sequence differentiation, accelerating evolution and accentuating phenotypic variation, while serving to break up genome-wide LD in the process. Recombination can, therefore, make it easier for GWAS methods to separate genuine signals of association from a confounding clonal background that otherwise generates widespread dependence between genetic loci [91]. Of course, it remains critical that microbial GWAS methods account for the clonal relationships between isolates. It is, however, evident that being additionally suitable for use in recombinant organisms will enable the accurate identification of associations in a much larger range of datasets.

In organisms with a natural competence for the uptake of exogenous DNA by HGT, recombination rates are known to vary across lineages or genomic regions in response to selective pressures acting on bacterial phenotypes [96, 246]. Under selection, the genetic polymorphisms introduced by recombination may face rapid expulsion or spread quickly in a population [72]. This can conflate population-stratified variation with genuine signals of association in competent organisms. In the two most frequently-studied traits in bacterial GWAS, drug resistance and virulence, phenotypic differences have been found to drive lineage-level variation in recombination rates [247, 248]. When recombination is present but not accounted for, GWAS methods may make faulty inferences about population structure or the evidence for association. Association studies will clearly benefit if they can accurately account for both clonal inheritance and recombination. The GWAS methods developed thus far, however, have not been equipped with the tools necessary to reliably address both of these fundamental features of bacteria.

### 2.3.1 Measuring recombination

Before performing an association study in bacteria, it is recommended that the recombination rate of the organism under analysis be assessed. The recombination rates of many bacterial species have been previously characterised [74]. However, recombination rates are known to vary within species, and even within genomes, particularly as a result of selection [72]. Isolates sampled from an outbreak, for example, might reflect the results of a clonal expansion; hence, a species-level estimate of recombination may provide an over-estimate of the sample recombination rate [249]. As the rate of homologous recombination in a particular sample may not align with estimates taken from the wider species or other samples, recombination is best estimated in the sample being submitted to GWAS analysis.

Recombination is often indicated by a high prevalence of genetic homoplasy. Homoplasy is inferred to occur when the distribution of states at a particular genetic locus can only be explained by a mutation or recombination event, given a particular phylogenetic tree [206, 249]. Many methods of estimating recombination rates rely on this principle, including measures of the phylogenetic congruence between loci [250], and measures of the extent of genome-wide LD, via the four-gamete test [251], D' measure [252], or r-squared measure [253]. Alternatively, more sophisticated model-based approaches can be used to infer recombination rates [166, 233, 254]. We recommend using ClonalFrameML [221] to estimate recombination rates before performing GWAS. This will ensure an accurate assessment of recombination and, unless a phylogenetic approach is contraindicated by excessive recombination, the tree reconstructed by ClonalFrameML can then be fed directly into our phylogenetic GWAS approach.

### 2.3.2 Confounding by recombination

Recombination can act to confound association studies in two major ways. First, recombination can interfere with the inferences made by most methods of ancestral reconstruction, preventing accurate adjustments for the confounding effects of population stratification. Second, recombination can alter the probability of spurious association, directly impacting the inferences made during association testing.

### Confounding of ancestry inference

Recombination can act to confound reconstructions of the ancestral relationships between isolates. As the accurate inference of clonal ancestry is necessary to reject spurious population-stratified associations, it is critical in bacterial GWAS that recombination is accounted for during this preliminary stage of any method. Yet, aside from the use of ClonalFrame [143] in the analysis of *C. jejuni* by Sheppard *et al.* [118] and re-analysis by Farhat *et al.* [106], we are unaware of any bacterial GWAS studies that have taken steps to account explicitly for the effects of both clonal inheritance and recombination during ancestry inference.

Although non-phylogenetic clustering and dimension reduction methods are applicable in the presence of recombination, the population structure inferences made by these methods will nevertheless reflect both the vertical and horizontal transfer of genetic information. Hence, whether standard phylogenetic methods or lower-resolution clustering and dimension reduction methods are used, recombination can render genuine clonal relationships unclear. For instance, a recombination event might encourage these methods to separate a clade of true clonal relatives into two populations. As a result, GWAS methods might be presented with spurious evidence for the presence of two distinct hypervirulent lineages when in fact virulence has arisen only once among these genetic relatives. On the other hand, the interference of recombination might cause these methods to group together clonally-distinct lineages [91]. As a result, GWAS methods might mistakenly identify population-stratified variants as significant associations or reject genuine associations.

Moreover, because phenotypically adaptive variants often spread rapidly via recombination [72], standard phylogenetic methods, clustering techniques, and PCA-based approaches may mistake trait-associated variants for population-structuring alleles. As a result, they may encourage both the removal of phenotypically-relevant genetic variation and the preservation of confounding ancestral variation. Hence, the use of recombination-naive ancestral reconstruction methods is likely to be particularly problematic in bacterial GWAS. Both the power and false positive rate of microbial GWAS methods are likely to be improved by explicitly accounting for recombination when reconstructing ancestral relationships.

### Confounding of association inference

The second major challenge, which has been largely overlooked by bacterial GWAS methods, is that recombination can more directly confound the inferences made in association testing. The number of substitutions, $N_{sub}$, that occur in the evolutionary history of a sample affects the probability of spurious association. We can illustrate this point with the aid of Figure 2.7. The terminal distribution of the phenotype in Figure 2.7A is the result of seven substitutions across the tree. Suppose that in a genetic dataset, substitutions randomly occur along the tree according to the homoplasy distribution Figure 2.7C. It is easy to see how the distribution of binary genotypic states at many of these loci could, by chance, take on a similar pattern to the phenotype. Likewise, as 17 substitutions give rise to the phenotype in Figure 2.7B, chance associations would be expected with larger numbers of substitutions at genetic loci, as in the homoplasy distribution in Figure 2.7D. Conversely, we would expect to see considerably lower probabilities of spurious association between (A) and (D), or (B) and (C). Thus, by modulating the substitution rate, recombination changes the probability of chance association with the phenotype.

Even if the clonal genealogies of two samples are topologically identical, as in Figure 2.7, a difference in recombination rate can alter the probability of spurious association. Variation in mutation rates can have a similar, if usually smaller, effect. Therefore, unless the impact of recombination and mutation on chance association is accounted for, GWAS methods will be unable to reliably reject false positive findings.

## 2.3.3 Controlling for recombination

With respect to ancestry inference, we have already introduced the problem posed by recombination, and we have presented potential solutions above. Unless clonal relationships are entirely obscured by recombination, we should be able to inhibit the inflation of association statistics due to ancestry by using a recombination-aware phylogenetic method in our simulation-based approach [91]. We choose to use ClonalFrameML [221] within our bacterial GWAS method. This allows us to identify high-resolution phylogenies while accounting for the impact of recombination, which, in turn, enables accurate corrections for the clonal relationships linking isolates.

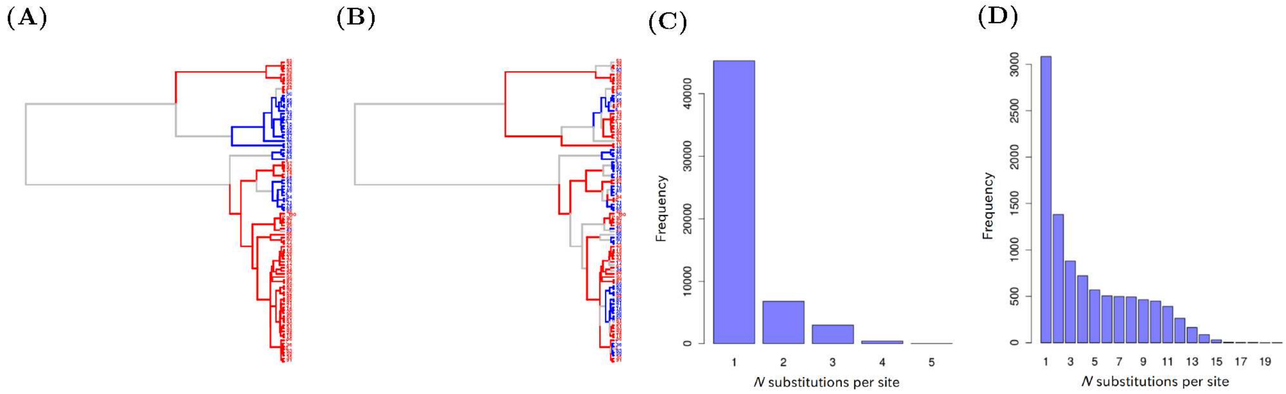


**Figure 2.7. Recombination affects the probability of spurious association.** Two phylogenetic trees with the same topology are shown, each with a distinct binary phenotype along its branches (blue = 0, red = 1, grey = substitution). **A:** A phenotype that clusters within lineages, where $N_{sub} = 7$. **B:** A phenotype showing a more fragmented terminal distribution, where $N_{sub} = 17$. Two homoplasy distributions are presented as histograms, showing the number of genetic loci (y-axis) undergoing a given number of substitutions (x-axis) per site along the tree. **C:** A homoplasy distribution with low $N_{sub}$ values, due to mutation in a clonal organism. **D:** A homoplasy distribution with high $N_{sub}$ values at many sites, typical of a recombinant organism with $R = 0.05$, as well as mutation. Consider the probability of chance association between phenotype (A) and genetic loci with $N_{sub}$ as in (C), or phenotype (B) and genotypic $N_{sub}$ (D). Compare this to (A) and (D), or (B) and (C). From this, we can see why the probability of spurious association increases with greater similarity between the numbers of phenotypic and genotypic substitutions.

How we should account for the more direct impact of recombination on association inference was less clear. This problem has received very limited attention within the bacterial GWAS literature. In the phylogenetic approach of Sheppard *et al.* [118], the authors simulate loci along the tree with numbers of substitutions drawn from a Poisson distribution with $\lambda = 1$. The simulated dataset is able to approximately maintain the empirical tree structure, but it does not reflect the impact of recombination and mutation on the probability of chance association in the empirical dataset. In developing our own method, we overcome this limitation by estimating the real number of substitutions per site from the empirical dataset. This estimation is achieved via ancestral state reconstruction, which can be performed by maximum parsimony or ML methods, as described below.

**Maximum-parsimony methods** aim to identify the ancestral states that require the smallest number of substitutions to explain the data, minimising the "parsimony cost". Parsimonious ancestral state reconstruction proceeds as described in Box 2.9.

1. Postorder traversal (from tips to root) to assign parsimonious character states to ancestors.
2. If the root's direct descendants differ in state, assign one of these states to the root at random.
3. Preorder traversal (from root to tips) to modify descendants' states where they do not match direct ancestor's state.
4. If the root's direct descendants had differed in state in Step (2), repeat Step (3) with the other state and compare the parsimony costs.
5. Return the set or sets of ancestral states associated with the lowest parsimony cost.

**Box 2.9. Maximum-parsimony ancestral state reconstruction method**

Popular implementations include Fitch's parsimony [255], whose steps are outlined in Box 2.9, and Wagner parsimony [193]. Parsimony methods can take either a one-pass or two-pass approach. One-pass approaches, like the most parsimonious reconstruction (MPR) method, reconstruct ancestral states sequentially, from the tips to the root, selecting the most parsimonious states at each ancestral node based on its direct descendants only. Two-pass approaches, including Fitch [255] and the accelerated transformation (ACCTRAN) approach in Wagner parsimony [193], instead infer ancestral states from tips to root and back again. Two-pass procedures remain fast, and they are more likely to resolve ties and less likely to identify sub-optimal solutions.

Parsimony methods are computationally efficient, straightforward in application and intuitive in interpretation. By design, parsimony methods assume that changes between all states are equally probable [256]. Although, if this assumption is known to be incorrect, weighted parsimony algorithms [257] can improve accuracy by altering the relative cost of state changes. Parsimonious reconstructions are informed by terminal states and tree topology, but branch lengths are not taken into account when making inferences about ancestral states. Ultimately, as their name indicates, parsimonious approaches aim to identify the sparsest "minimum evolution" scenario. This can provide a useful reconstruction of the states and state changes among ancestral isolates, although it may or may not reflect the genuine evolutionary history of the sample.

**Maximum-likelihood methods** treat the set of ancestral states as parameters and aim to select those that give the highest probability of observing the data, by taking the steps in Box 2.10.

1. Select an initial set of ancestral states and parameters of the model of evolution.
2. Compute likelihood.
3. Propose changes to the ancestral states and parameters.
4. Compute new likelihood.
5. Accept the proposed changes if the new likelihood is higher than the previous one; else, reject the changes.
6. Repeat Steps 3 – 5 until no further improvements can be found.

**Box 2.10. Maximum-likelihood ancestral state reconstruction method**

The initial ancestral states in Step 1 can be selected at random; although, in practice, these are often selected with more efficient distance-based or parsimony methods. ML methods can perform either marginal or joint reconstructions. Marginal reconstructions proceed upward, from tips to root, identifying the most likely state at each ancestral node with reference only to its direct descendants. Joint reconstructions take a more holistic approach and attempt to identify the set of ancestral states at all internal nodes that, collectively, maximise the likelihood of the data. Joint ML reconstructions are more computationally intensive, but they are less likely to get trapped at local optima [256].

In general, ML methods propose ancestral states and model the probabilities of transition across each branch of the phylogeny. Unlike parsimony, ML methods take branch length into account and are not penalised for proposing larger numbers of evolutionary transitions. Evaluating the amount of evolutionary time available may lead to more realistic inferences of ancestral states and substitution rates. Yet, ML estimates of evolutionary change can be easily skewed when substitutions are very frequent or exceedingly rare, or if they occur on very short or even zero-length branches [258]. Like parsimony, ML methods identify a single "best" set of ancestral states. However, ML methods explicitly acknowledge and quantify uncertainty in the reconstruction. Parsimonious reconstructions may reveal uncertainty in some equally-parsimonious ties. ML reconstructions, by contrast, can quanitfy their overall likelihood and, for discrete variables, the relative likelihood of every possible state at each ancestral node.

With the aid of ML or parsimonious reconstruction methods, we will be able to infer ancestral states and to estimate the corresponding genome-wide homoplasy distribution of $N_{sub}$ values. Our approach can then use these inferences to estimate how the unobserved processes of recombination and mutation may have shaped the distribution of association statistics observed. While other GWAS approaches might struggle to incorporate this information, our simulation-based approach provides a natural opportunity to do so. By using the homoplasy distribution to inform our simulation of the null genetic dataset along the tree, we can maintain both the empirical population structure and the $N_{sub}$ variation observed. In this way, we can account for the confounding effects of mutation and recombination in association testing, as well as during ancestry inference.

Furthermore, in accounting for mutation, recombination, and tree structure within our simulation procedure, we expect to find that our simulated allele frequencies will approximate their empirical counterparts [259]. Previous phylogenetic simulation procedures [118], despite faithfully recreating tree topology, may make inappropriate inferences about branch lengths and population allele frequencies and their impact on the probability of chance association [260]. We hope that by extending our simulation procedure to account for these inter-related empirical characteristics and evolutionary processes, the incorporation of additional parameters will enhance the power and precision of our GWAS method.

## 2.4 Detecting associations

### 2.4.1 Estimating heritability

GWAS methods will only be able to identify associations with a phenotypic trait that is heritable. Broad-sense heritability ($H^2$) measures the proportion of phenotypic variation that is attributable to genetic factors alone, relative to the collective contributions of both genetic and environmental factors [261].

$$\mathbf{H^2} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} \tag{2.8}$$

$H^2$ includes both additive and epistatic effects. By contrast, narrow-sense heritability ($h^2$) measures the proportion of $H^2$ that can be ascribed to additive genetic variation, relative to dominant or epistatic effects.

Often, the literature contains sufficient evidence to support a claim of heritability. If not, several methods can be used to estimate this parameter. First, the variance component (VC) approach allows $H^2$ to be estimated by comparing phenotypic covariance to genetic relatedness [262]. VC can be transferred from humans to microbes by replacing family pedigrees with matrices of pairwise genetic similarity [261]. Second, ANOVA can be used to estimate $H^2$ by comparing phenotypic variance within and among clonal lineages [263]. Third, if enough donor-recipient transmission pairs are available, $H^2$ can be estimated via linear regression [264]. Fourth, using SNP-based genomic relatedness as random effects, LMM can perform $H^2$ estimation, enabling the inclusion of additional potential confounders as covariates [132]. Finally, phylogenetic approaches like Pagel's $\lambda$ [265] and phylogenetic mixed models (PMMs) [266] can also be used to assess whether phenotypic variation is correlated to evolutionary history. While GWAS methods are hypothetically applicable to any phenotype with $H^2 > 0$, the success of GWAS will require large sample sizes if heritability is low.

### 2.4.2 The power to detect associations

In GWAS, statistical power depends on both the dataset and the method of analysis. The power of an association study increases with the effect size and the MAF of associated loci. As natural characteristics of the data, however, these two parameters are beyond the control of the analyst. Increasing sample size, on the other hand, provides one avenue by which we can increase statistical power. Indeed, even with a sensitive GWAS method, larger samples may be needed to detect small-effect loci that contribute to weakly-heritable, probabilistically-determined phenotypes. It is therefore important that we develop a GWAS method that scales efficiently to larger datasets. Even when sample size is within our control, we hope that our GWAS method can extract the highest statistical power possible, without compromising the specificity of the association test.

#### Genotypic data

Discovery power in bacterial GWAS can be enhanced by capitalising on all available information. As such, we aim to develop a bacterial GWAS method that does not squander valuable genetic data by ignoring isolates, like existing pairwise comparisons methods and phylogenetic sampling strategies, or exclude potentially-relevant genetic variation, like PCA-based regression methods. To capitalise on the available sequence data, our GWAS method should test for associations across the pan-genome. How best to explore this variation remains an open question in bacterial GWAS [183].

In microbial GWAS, the pan-genome is typically either explored by performing GWAS on two smaller datasets of core SNPs and accessory genes, or by performing GWAS on one much larger dataset of pan-genome k-mers. In the first approach, data pre-processing steps must be completed prior to association testing. First, sequence assembly must be undertaken to compile and align genomic data for all sampled isolates, with the aid of dedicated software [267–269]. This is usually accomplished by mapping short contiguous sequence reads to a reference genome; or, a reference-free alignment can be achieved through the more intensive de novo assembly process of mapping sequence reads to one another. Second, sites must be assigned to either the core or accessory genome, according to their inclusion in all ($> 99\%$) or a subset ($\leq 99\%$) of sampled genomes. Finally, SNPs must be identified within the core genome and gene presence or absence must be indicated in the accessory genome.

The second approach provides an alignment-free alternative, using k-mers, sub-sequence strings of length $k$. K-mer datasets span the pan-genome and contain the set of all possible, overlapping $k$-length sub-strings. K-mers are designed to facilitate association testing on non-SNP variation, like insertions and deletions (INDELs) or copy number variations (CNVs), although these variants can also be incorporated into more traditional SNP-based approaches with relative ease [49, 120, 155, 270, 271]. A distinct benefit of k-mers, however, is their ability to detect associations with non-gene variation in the accessory genome, enabling the extension of GWAS to SNPs in accessory genes, and to variation in promoters and inter-genic regions within the accessory genome.

Thus far in the microbial GWAS literature, however, few associations have been identified with k-mers that could not have otherwise been identified with more efficient parallel analyses of core SNPs and accessory genes [129, 132]. Furthermore, the marginal benefits that may be offered by k-mers are also accompanied by 50-fold increases in the number of loci that must be tested for association [132]. Given the need to correct for multiple testing in genome-wide analyses, this difference can have a substantial impact on the statistical power of association studies. The the scale of k-mer-based approaches also drastically reduces the efficiency and scalability of GWAS studies. Yet, because the optimal value of $k$ is difficult to estimate in advance, this computational burden is often compounded by repeating the analysis with multiple values of $k$. The added computational expense of k-mer-based approaches may be difficult to justify in many cases, especially for more intensive methods, like simulation-based approaches. The interpretation of GWAS results is also more challenging in k-mer-based analyses. Although k-mers can be identified without an alignment, additional efforts must be spent on pinpointing the source of association in any k-mers identified as significant by GWAS studies. Overall, the costs imposed by k-mer-based approaches—on statistical power, computational efficiency, and ease of interpretation—appear to outweigh the potential benefits. In light of the available evidence, we prefer not to focus our efforts on a k-mer-based approach. Although we will aim to develop a GWAS method that is also applicable to k-mer data, we intend to work with core SNP and accessory gene presence-or-absence data when performing association studies.

### Phenotypic data

Phenotypic data also comes in many forms, and preserving the richness of carefully collected phenotypic data can improve both the statistical power and accuracy of association testing methods. Generally speaking, phenotypes can be binary, categorical, or continuous. Binary phenotypes take on one of two states, for example, “drug resistant” or “susceptible”. Categorical phenotypes can take on any number of discrete states or factor levels. These may represent numbers whose relative values are meaningful, for example, a number of flagella. Alternatively, levels may correspond to unordered non-numeric categories, as in the analysis of multiple hosts, such as “cow”, “chicken”, and “pig”. Continuous phenotypes, by contrast, are numeric variables that can take on any value in a given range. The minimum inhibitory concentration (MIC) values used to quantify drug resistance, virulence metrics like set point viral load in HIV, and measures of toxicity are all examples of continuous variables. These phenotypic data types differ in information content, increasing from binary (least informative), to categorical, to continuous (most informative).

Many microbial association tests have only been designed to be applicable to binary variables. Indeed, the majority of microbial GWAS studies have sought to identify associations with binary drug resistance phenotypes that are highly determined by a relatively small number of high-effect genetic loci (see Table 1.2). It is possible for both continuous and categorical phenotypes to be reduced to binary variables, if required, although information loss results. Alam *et al.* [122], for example, reclassify continuous MIC values into a “high” and “low” resistance phenotype. This artificial binary distinction prevents any perception of the difference between the median and maximum MIC values. If such differences represent a considerable component of the relationship between genotype and phenotype, this information loss might undermine the explanatory power of a GWAS analysis. On the other hand, recoding continuous phenotypic values into binary or categorical phenotypic states can generate artefactual class differences between truly similar individuals. Hence, reclassifying data not only reduces the information available for association testing, but it also makes this information less reliable. We therefore aim to develop an approach that is applicable to binary, categorical, and continuous phenotypes. Users may choose to transform or re-categorise their phenotypic data, but this choice should not be dictated by methodology too narrow to properly explore the variation that microbial life displays. Instead, by developing a strategy that preserves the integrity and complexity of both phenotypic and genotypic data, and fully explores the associations between both variables, we aim to increase both the power and precision of our GWAS method.

### 2.4.3 Measures of association

The strength of an association between genotype and phenotype can be measured in a number of ways. The association tests used in microbial GWAS to date can generally be broken down into "allele-counting" and "homoplasy-counting" approaches [126].

#### Allele-counting measures

Allele-counting approaches define association as the over-representation of a particular allele alongside a given phenotypic state. The majority of association scores used thus far in microbial GWAS have been allele-based measures. Popular allele-based tests of association include the Fisher's exact test and the $X^2$ test, as well as the CMH test, multivariate regression approaches, and the association test performed at terminal nodes in the approach of Sheppard *et al.* [118].

Allele-based measures of association work only with the data observed. A broad sample-wide relationship between genotype and phenotype is required to achieve a high allele-based association score. In the proper evolutionary context this may indicate widespread support for association. But, because it can be difficult to separate genuine allele-based association scores from those due to population stratification alone, moderate allele-based scores must often be rejected to avoid false positive findings. Suppose a phenotype is entirely determined by either SNP 1 or SNP 2, but neither locus accounts for the majority of variation in the terminal phenotype. An allele-based measure, like that used by Sheppard *et al.* [118], may overlook both truly-associated loci. Because they do not rely on inferred states or substitutitons among unseen ancestors, allele-counting approaches incorporate less uncertainty and are less error prone. On the other hand, without considering ancestral states, allele-based measures can overlook signals of association that can be revealed when the evolutionary history of isolates is examined.

#### Homoplasy-counting measures

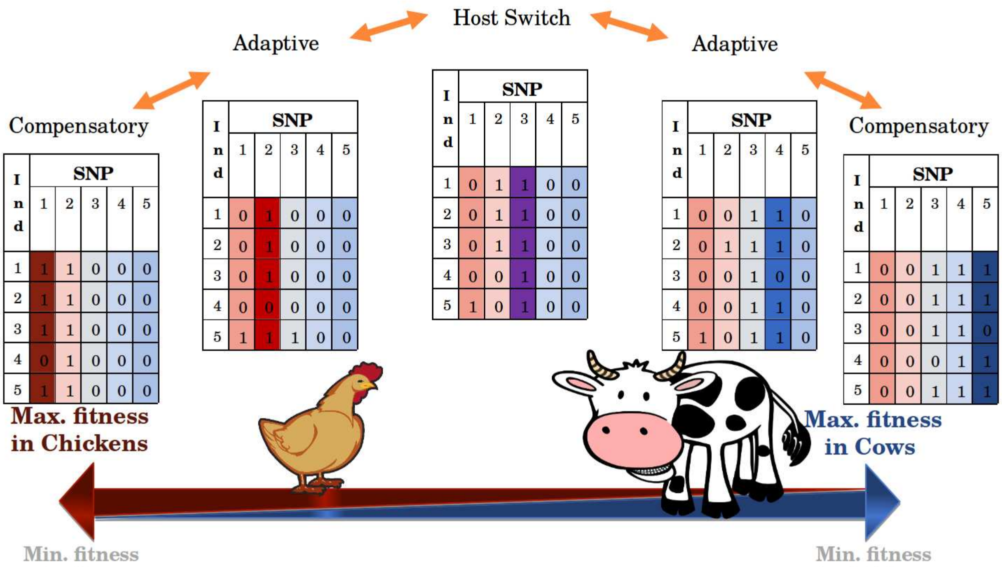
The homoplasy-counting framework appears only within phylogenetic GWAS approaches. Homoplasy-based GWAS methods seek to identify signals of convergent evolution and positive selection. Traditional measures of selection, like those used in the dN/dS method [272] or haplotype-based tests [77], perform poorly in clonally-related samples. In GWAS, however, homoplasy-based measures have been able to take as evidence for association the repeated and independent emergence of substitutions in both genotype and phenotype along the branches of a phylogenetic tree. Phylogenetic GWAS methods that adopt a homoplasy-counting approach include the PhyC method [114], BayesTraits [142], the tree-based $X^2$ test of Hall *et al.* [121], and pairwise comparative methods [106, 127].

Homoplasy-based methods make it possible to expand the association study into the inferred evolutionary history of the sample. Working within a phylogenetic framework allows them to evaluate the evolutionary support for association. If phylogenetic or ancestral reconstructions are unreliable, homoplasy-counting methods may exaggerate estimation errors and draw incorrect conclusions about associations. But, if homoplasy-based metrics are applied to reliably estimated ancestral states, they may be able to draw greater insight from a dataset. Homoplasy-counting methods can pick up on repeated patterns of association occuring in a subset of the data. If association is not upheld across the phylogenetic tree, but it repeats across several substitutions in both genotype and phenotype, homoplasy-counting measures may nevertheless detect a strong association. Then again, their requirement for simultaneous substitutions can allow homoplasy scores to overlook associations upheld across all observed isolates.

Overall, it is clear that both the allele-based and homoplasy-counting approaches make distinct, potentially-valuable contributions to association testing, and that their strengths and limitations counterbalance one another. Instead of arguing the merits of either the allele-based or homoplasy-counting approaches to association testing, we believe that the most sensible solution may be to draw on both frameworks. In our efforts to improve the power of our stringent phylogenetic GWAS approach, we will therefore consider both indicators of association.

### Alternative measures

There is not always a straightforward, deterministic relationship between genotype and phenotype. Less strictly heritable and more probabilistically determined phenotypes may be associated with less frequent or less penetrant alleles. Phenotypes like host association may be partially influenced by multiple alleles, as well as non-genetic factors like host immunity. Figure 2.8 shows how complex phenotypes may be favoured through the accumulating impact of genotypic substitutions that can occur before, during, and after substitution in the phenotypic state, posing a challenge for both allele-based and homoplasy-based methods. Allele-based measures may fail to detect such weak associations, if they are not widely upheld across the sample. Homoplasy-based measures may also struggle, if genotypic and phenotypic substitutions do not occur repeatedly on the same branches. The PhyC [114] homoplasy score modifies this requirement, detecting loci that undergo substitutions either alongside or after a substitution in the phenotype.



| Ind | SNP | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 | 0 |

| Ind | SNP | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 1 | 0 | 0 |

| Ind | SNP | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 1 | 0 | 0 |
| 3 | 0 | 1 | 1 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 |
| 5 | 1 | 0 | 1 | 0 | 0 |

| Ind | SNP | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 0 | 1 | 1 | 0 |
| 2 | 0 | 1 | 1 | 1 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 1 | 1 | 0 |
| 5 | 1 | 0 | 1 | 1 | 0 |

| Ind | SNP | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 0 | 0 | 1 | 1 | 1 |
| 2 | 0 | 0 | 1 | 1 | 1 |
| 3 | 0 | 0 | 1 | 1 | 0 |
| 4 | 0 | 0 | 0 | 1 | 1 |
| 5 | 0 | 0 | 1 | 1 | 1 |

**Figure 2.8. Probabilistic association with host.** Probabilistic association is illustrated by a multi-stage process of host switching between chickens (red) and cows (blue). The spectrum (bottom) indicates relative fitness in chicken and cow hosts. In the centre of the image, we see five examples of genotypes for five SNPs and five individuals. In each frame, the highlighted polymorphism either occurs or reverses, increasing or decreasing fitness according to the mechanism stated above the frame. **Host switching** occurs in the centre of the figure, where the preference for either cow or chicken hosts is exerted by the substitution, shifting the probabilistic association towards one end of the spectrum. **Adaptive substitutions** make an isolate more fit in a particular environment. For example, just right of centre, an adaptive substitution enabling vitamin B5 biosynthesis in *C. jejuni* isolates increases fitness in cattle by facilitating adaptation to host diet (as in Sheppard *et al.* [118]). **Compensatory substitutions** make up for the fitness costs of previous adaptive substitutions. For example, at right, if B5 biosynthesis is metabolically expensive for the isolates to maintain, an alternate metabolic pathway or another fitness-enhancing trait could be activated by a compensatory substitution. This new trait-associated locus would likely then be taken up throughout the population as selective pressures are exerted by the present cattle host environment.

Unfortunately, this criteria allows both false positives and false negatives to slip through. Suppose, for example, that a phenotypic change occurs on one of two branches descending from the root node. Repeated genotypic substitutions occurring at any locus in the descendant clade may be inferred as significant evidence of association, even though these might be associated with any other single-origin phenotype or genotype that distinguishes the ancestor of this clade from the ancestor of the other clade. This theoretical limitation has already been observed in practice, requiring users of PhyC [114] to remove deep phylogenetic mutations by hand and to manually comb through significant findings for sites truly associated with lineage rather than phenotype [134]. On the other hand, suppose a genotypic substitution from SNP 0 to 1 falls on the branch preceding a phenotypic change from chicken to cow host. Using the PhyC homoplasy score, even if SNP 1 and the cow phenotype were maintained in all subsequent descendants, the only association measured in this case would be between SNP 1 and the chicken phenotype [114]. Hence, when applied to probabilistic, complex phenotypes, like host association and virulence, both allele-based and homoplasy-based measures may fail to accurately identify associations. We will, therefore, consider whether any additional approaches to association testing may improve our ability to detect the weaker signals of association that accompany some of the most interesting relationships between genotype and phenotype.

### Multiple measures

Although the majority of microbial GWAS methods have relied on a single measure of association, methods using multiple measures have also been proposed. Chen and Shapiro [126] suggest a sequential two-step procedure, in which an allele-based approach selects genomic regions broadly correlated with the phenotype, and then a targeted homoplasy-based method extracts only those loci that also demonstrate convergent evolution with the phenotype along the tree. Hall *et al.* [121] implement a version of this approach, by performing a phylogeny-wide correlation test on ancestral states, and then subsequently reducing these initial findings by requiring significant correlation among substitutions as well. Brynildsrud *et al.* [127], likewise, adopt a sequential implementation to narrow down results through increasingly stringent association tests. We recognise the potential value of using multiple measures of association, yet we question whether they are best applied in succession or if we might, instead, explore a new approach by adopting a parallel implementation. Instead of narrowing down the pool of associated loci identified, this could serve to expand the set of findings detected by GWAS. Whereas running multiple association tests in succession might increase specificity, a parallel implementation might allow us to achieve greater statistical power and, if we can ensure high specificity through other components of our method, better performance overall.

## 2.5 Objectives

In our review of the literature, we examined a number of major challenges that have prevented the widespread, successful application of association testing to microbial samples. In developing our own microbial GWAS method, we aim to address the following methodological issues. Each of these objectives represents a significant limitation of existing approaches and presents a substantial opportunity for improvement.

### 2.5.1 Account for clonal ancestry

First, we aim to address the problems of confounding population structure and homologous recombination. We set out to develop a method that can distinguish true signals of association from the spurious associations that arise as a result of ancestral relationships in clonal organisms. We intend to use a recombination-aware phylogenetic reconstruction method, like ClonalFrameML [221], to correct for the confounding influence of ancestry. If integrated within a simulation-based parametric bootstrap approach, a phylogenetic tree should allow us to achieve the greatest precision and the most robust check on bias due to population stratification.

### 2.5.2 Account for variable recombination

We also wish to address the less predictable, more variable confounding effects of homologous recombination. We aim for the microbial GWAS method that we develop to be applicable to organisms ranging from purely clonal to moderately recombinant. The method that we develop should not only adjust for recombination during ancestry inference, but it should actively account for the impact of recombination on the inferences made during association testing. Once we have developed a GWAS method that is able to address both clonality and recombination, we will compare the performance of our method to that of competing cluster-based and dimension reduction methods by testing each on simulated datasets.

### 2.5.3 Balance sensitivity and specificity

Second, it follows that we face a challenge of effective inference-making. We must not only eliminate false positives that arise from confounding population structure and recombination, but we must do so while maintaining high power to detect associations. We need to determine which measures of association will give our approach the greatest statistical power. We aim to make evidence-based decisions by comparing the performance of approaches through applications to simulated data. We intend to explore diverse approaches to association testing, including measures of allele-trait correlation and indicators of convergent evolution. In addition, we will consider whether some combination of these approaches may improve sensitivity further still. Finally, in our efforts to develop a robust yet high-powered approach, we will address open questions in microbial GWAS about how best to assess significance and how to control for multiple testing [183].

### 2.5.4 Capitalise on data diversity

Third, we would like to develop a microbial GWAS method that will be applicable to as many forms of genotypic and phenotypic data as possible. This should include core SNPs, accessory gene presence-or-absence data, and k-mers. Our method should be computationally efficient and scalable to large genetic datasets, spanning the entire pan-genome. To make the most effective use of available phenotypic data, it will also be beneficial if our method is able to test for associations with binary, categorical, and continuous phenotypes, and those drawn from longitudinal samples. In order to preserve phenotypic information, we will need to develop measures of association that are versatile and generalisable. Most of the bacterial GWAS methods to date have investigated binary phenotypes under strong selection, like drug resistance, that are highly-determined by a small number of high effect loci. We are motivated to develop methods of association testing that have the flexibility and power to detect lower-effect associations that give rise to more probabilistic phenotypes, like host association.

### 2.5.5 Build an effective, accessible tool

Finally, we want to consolidate our ideas within a coherent, overarching approach to microbial GWAS, and we want to package our approach in effective, user-friendly software that will render our approach most useful to others. We hope that our contribution in this area will, more generally, expand the reach and popularity of the phylogenetic approach in microbial GWAS, which has thus far been hindered by a dearth of usable software (see Table 1.3). Our method should be able to provide objective solutions that do not require the user to make subjective assessments about population structure, as in cluster-based and multivariate approaches. Our approach should generate reproducible results that have clear, meaningful interpretations. The tool that we develop should be user-friendly, well-documented, and accessible to users from across a wide range of biological and computational backgrounds. Our approach should also be implemented efficiently, and it should offer the flexibility needed to address a variety of GWAS problems in a range of datasets.

# Chapter 3

# A new phylogenetic method to perform microbial GWAS

**Acknowledgements** A version of some of the written content in this chapter and Figure 3.3 also appeared in the original publication of our method [155]. Caitlin Collins was responsible for the design of our GWAS method and its implementation in the treeWAS R package. Xavier Didelot helped to guide the development of our method, providing insight through regular discussions with Collins. Didelot also helped to create Figure 3.5 and provided valuable feedback throughout the writing process We would like to thank Daniel Wilson and David Aanensen for taking the time to review this work during its development, for our many thoughtful discussions, that prompted more thorough explorations of key issues and led to improvements in our method. We are grateful to the two anonymous reviewers who provided helpful suggestions that we incorporated in our paper. My thanks to Richard Fitzjohn for offering his advice on how to assess and improve the efficiency of the treeWAS R package.

## 3.1 Introduction

We have developed a novel method to perform GWAS in microbial samples that overcomes many of the limitations in available methodology, discussed above. In this chapter, we introduce our phylogenetic GWAS approach and explain its components in detail. We demonstrate how our approach meets the aims outlined in Chapter 2, simultaneously tackling the confounding influences of both clonal relatedness and variable recombination, while achieving high power yet maintaining a low false positive rate. Our approach makes effective use of the available diversity of both genotypic and phenotypic data, and operates within a user-friendly platform. We also describe the implementation of our method in the R programming language, in our R package, treeWAS. Finally, we highlight features of treeWAS that improve the computational efficiency of our implementation and that render our approach flexible and accessible to users.

## 3.2 Overview of the method

The central aim of our GWAS approach is to delineate true signals of association from a noisy background of spurious associations. As a GWAS method, our approach adopts a systematic statistical approach that requires no prior hypotheses about potential associations at candidate loci. We perform an unbiased search, assessing the degree of association between each locus in a genetic dataset and a phenotype of interest.

As clonal relatedness and homologous recombination may act to increase or decrease the probability of spurious association in microbial GWAS, it is essential that we evaluate both the statistical and evolutionary support for association. To address these confounding factors, we adopt a simulation-based approach. First, we characterise the features of the empirical dataset that may bias the inferences made during association testing. Then, we use these parameters to guide the simulation of a genetic dataset that recreates these confounding factors but does not have any true association with the phenotype. The purpose of generating this simulated genetic dataset is to estimate the null distribution of association score statistics expected under the null hypothesis of “no association”. We use the null distribution to determine which association score values in the empirical dataset are likely to be truly significant and which may, in fact, arise by chance as a result of confounding factors. Our approach maintains strict control over the number of false positive findings by rejecting nearly all empirical associations that fall within the null distribution.

To ensure that our approach is powerful, as well as robust, we have devised a strategy to increase sensitivity through the parallel application of multiple measures of association. To this end, we have developed three association scores that draw on distinct, complementary signals of association. First, an allele-based measure is applied to the genotypes and phenotypes observed along the tips of the phylogenetic tree, returning high scores in the presence of broad, sample-wide association. Second, a homoplasy-based approach is used to expand the association test into the reconstructed evolutionary past, where it confers high scores if inferred substitutions in genotype and phenotype occur on the same branch of the phylogenetic tree. Third, an integral-based measure draws on both the maintenance of allelic states and the presence of homoplasies to determine whether association is widely indicated across the branches of the tree. Each of these scores is capable of identifying associations with either binary, categorical, or continuous phenotypes. These three scores are applied in parallel, so that they can collectively ensure a thorough exploration of the potentially diverse landscape of association signals. By identifying only associations with strong evolutionary and statistical support, but from multiple complementary measures, our approach is able to pair conservative control over confounding factors with a high-powered test of association.

| Symbol | Description |
|---|---|
| $p_i^{anc}$ | Phenotypic state at the ancestral node of branch $i$ |
| $p_i^{des}$ | Phenotypic state at the descendant node of branch $i$ |
| $g_i^{anc}$ | Genotypic state at the ancestral node of branch $i$ |
| $g_i^{des}$ | Genotypic state at the descendant node of branch $i$ |
| $l_i$ | The length of branch $i$ of the phylogenetic tree |
| $N_{branch}$ | Number of branches in the phylogenetic tree |
| $N_{ind}$ | Number of individuals in the empirical genetic dataset (or terminal nodes in the tree) |
| $N_{loci}$ | Number of loci in the empirical genetic dataset |
| $N_{assoc}$ | Number of phenotypically-associated sites in the genetic dataset |
| $N_{sim}$ | Number of simulated loci in the null genetic dataset |
| $N_{sub}$ | Number of substitutions |
| $N_{score}$ | Number of association scores measured |
| $S_i$ | The association score value at empirical locus $i$ |
| $s_j$ | The association score value at simulated locus $j$ |
| $P_i$ | The empirical p-value for an association at empirical locus $i$ |
| $\alpha_{base}$ | The base p-value specifying the overall significance level |
| $\alpha$ | The p-value specifying the per-test significance level |

**Table 3.1. Mathematical notation.**

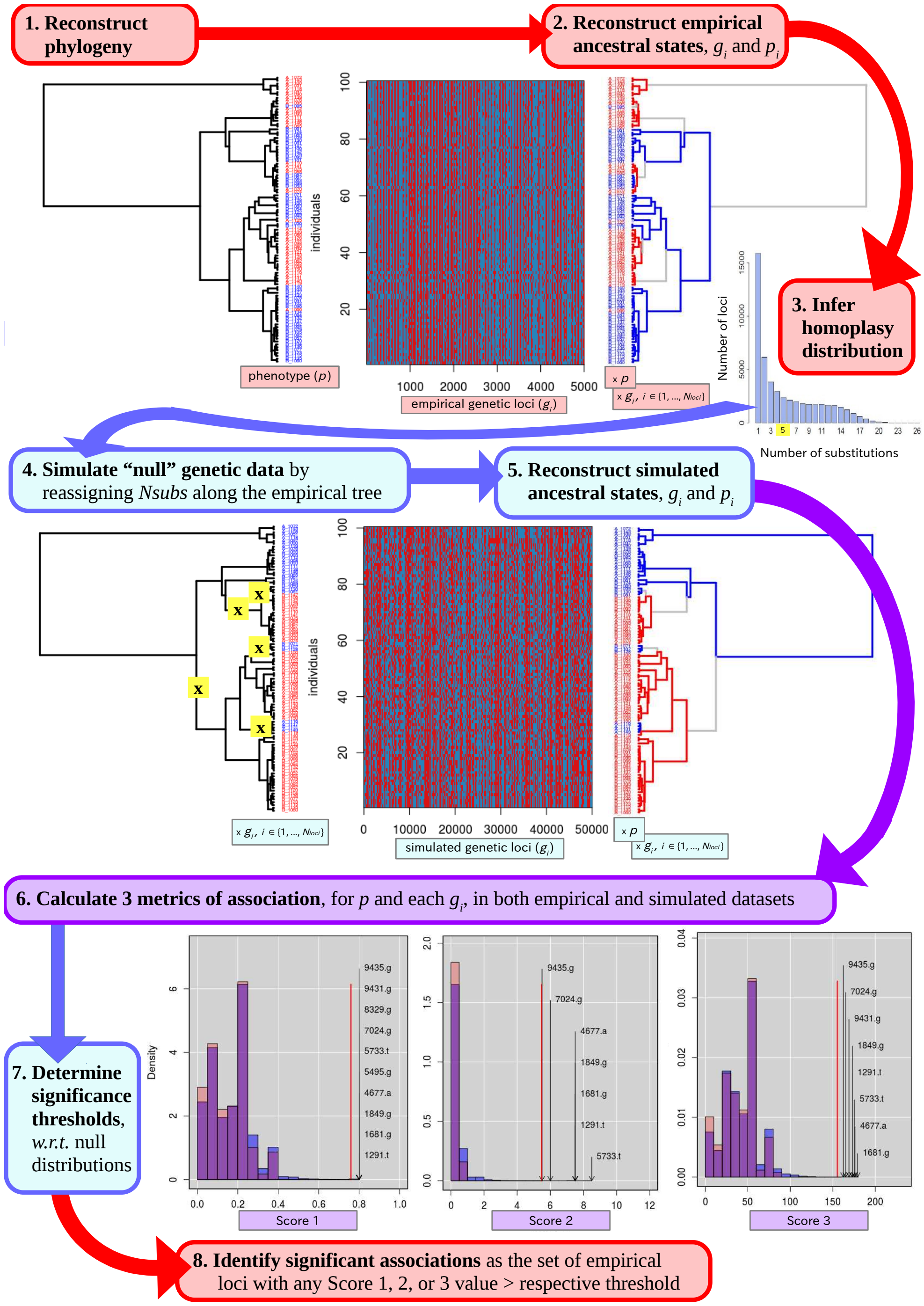

1. Reconstruct phylogeny
2. Reconstruct empirical ancestral states, $g_i$ and $p_i$
3. Infer homoplasy distribution
phenotype ($p$)
individuals
100
80
60
40
20
1000
2000
3000
4000
5000
empirical genetic loci ($g_i$)
x $p$
x $g_i$, $i \in \{1, ..., N_{loci}\}$
Number of loci
15000
10000
5000
0
1
3
5
7
9
11
14
17
20
23
26
Number of substitutions
4. Simulate "null" genetic data by reassigning *Nsubs* along the empirical tree
5. Reconstruct simulated ancestral states, $g_i$ and $p_i$
x
x
x
x
x
x $g_i$, $i \in \{1, ..., N_{loci}\}$
individuals
0
10000
20000
30000
40000
50000
simulated genetic loci ($g_i$)
x $p$
x $g_i$, $i \in \{1, ..., N_{loci}\}$
6. Calculate 3 metrics of association, for $p$ and each $g_i$, in both empirical and simulated datasets
7. Determine significance thresholds, *w.r.t.* null distributions
Density
9435.g
9431.g
8329.g
7024.g
5733.t
5495.g
4677.a
1849.g
1681.g
1291.t
0.0
0.2
0.4
0.6
0.8
1.0
Score 1
9435.g
7024.g
4677.a
1849.g
1681.g
1291.t
5733.t
0
2
4
6
8
10
12
Score 2
9435.g
7024.g
9431.g
1849.g
1291.t
5733.t
4677.a
1681.g
0
50
100
150
200
Score 3
8. Identify significant associations as the set of empirical loci with any Score 1, 2, or 3 value > respective threshold

**Figure 3.1. treeWAS method pipeline.** This figure summarises our bacterial GWAS method in a series of illustrated steps, showing the key procedures implemented in treeWAS, as described below, in the Method Protocol (Section 3.3) and in the remainder of Chapter 3. The analysis begins, at the top left of the figure, with one phenotype ($p$) and $N_{loci}$ genotypes ($g_i$) for the $N_{ind}$ individuals that make up the sample. In **Step 1**, we reconstruct the clonal genealogy via phylogenetic estimation. Ancestral character estimation of $p$ and $g_i$ in **Step 2** enables the identification of the homoplasy ($N_{sub}$) distribution in **Step 3**. The inferences made in Steps 1-3 inform **Step 4** in the simulation of a "null" genetic dataset, whose only associations to the phenotype will arise from the confounding effects of ancestry, mutation and recombination. The reassignment of substitutions (e.g., $N_{sub} = 5$) is represented by the presence of yellow "x"s along the tree. Once all $N_{sim}$ loci have been simulated, we proceed with **Step 5** by reconstructing the ancestral states of all simulated $g_i$. In **Step 6**, we use Score 1, 2, and 3 to quantify all $\{p, g_i\}$ associations in both the real and the simulated datasets, generating three empirical and three null distributions of association statistics. The significance threshold for each metric is drawn in **Step 7** at the upper $(1 - \alpha)$ tail of the respective null distribution. Finally, in **Step 8**, any empirical locus that exceeds this significance threshold, in Score 1, 2, or 3, is identified as a significant association.

## 3.3 Method protocol

Our approach is implemented in the following steps, which are described in greater detail in the sections below. The mathematical notation used is defined in Table 3.1.

1. **Phylogenetic reconstruction** is performed to identify the ancestral relationships between sampled isolates. As the clonal genealogy must be reconstructed without interference from homologous recombination, we prefer to reconstruct the phylogeny with a recombination-aware approach like ClonalFrameML [221]. If recombination can be ruled out, standard phylogenetic methods can also be used.

2. **Estimation of empirical ancestral states** is performed to identify a marginal reconstruction of the ancestral states of both genotype and phenotype at all internal nodes of the phylogenetic tree. Parsimony methods [255] are used to infer the ancestral states of the genotype and the phenotype, if it is a binary variable. ML reconstruction methods [236, 273] are used if the phenotype is a continuous or categorical variable.

3. **Inference of the homoplasy distribution** is performed with the Fitch parsimony algorithm [255]. The homoplasy distribution describes the minimum number of substitutions, $N_{sub}$, that must occur at each genetic locus in the empirical dataset. This includes substitutions arising by both mutation and recombination.

4. **Simulation of null genetic data** facilitates the delineation of true associations from spurious associations. Simulated under $H_0$, the null hypothesis of “no association”, the “null” genetic dataset embodies potentially confounding features of the empirical dataset, but does not contain any genuine associations with the phenotype. The simulation procedure is guided by three parameters: (i) the phylogenetic tree, (ii) the homoplasy distribution, and (iii) the number of loci to be simulated, $N_{sim}$. The empirical phenotype is also maintained across the tips of the tree. Each of the $N_{sim}$ loci is simulated along the phylogenetic tree, from root to tips, undergoing a number of substitutions drawn from the homoplasy distribution on branches selected randomly with probabilities proportional to branch length. The resulting simulated dataset resembles the empirical dataset in its sample size, ancestral relationships, and terminal phenotypic distribution, thereby recreating the level of confounding population stratification observed. The simulated dataset also approximates the observed $N_{sub}$ and MAF distributions, allowing us to account for the effects of both mutation and recombination, as well as the strength of LD in the empirical dataset. By comparing the simulated and empirical datasets, we can separate genuine signals of association from associations caused by chance and confounding factors.

5. **Estimation of simulated ancestral states** is required prior to association testing, in each analysis performed by treeWAS, so that associations in the simulated dataset can be calculated across branches of the phylogenetic tree. To ensure a fair comparison can be made between the empirical and simulated datasets, we infer ancestral states in the same way in both datasets. The phenotype has already been reconstructed in Step (2), so we need only to use parsimony [255] to reconstruct the binary genotype.

6. **Association testing** is carried out in both the empirical and simulated datasets. Associations between simulated loci and the empirical phenotype are measured to allow for the identification of a null distribution of association score statistics under $H_0$. Associations between empirical loci and the phenotype are then measured and evaluated with reference to this null distribution. We use three independent tests to measure association with the phenotype at each locus in both datasets The use of these multiple measures improves statistical power by expanding the test to capture different signals of association.

7. **Identification of the significance threshold and associations** is achieved by drawing a threshold in the upper tail of the null distribution, at the value corresponding to a base p-value (e.g., $\alpha_{base} = 0.01$) that has been corrected for multiple testing via Bonferroni correction to account for both $N_{loci}$ and $N_{score}$. Among the set of empirical association scores, all values that exceed this threshold are deemed to be statistically significant associations and, thus, candidates for true biological association, pending subsequent confirmatory analyses.

## 3.4 Data processing

Two sources of data are required by our GWAS approach: genetic sequence data and a phenotype. The evolutionary history of the sample, its ancestral states, and relevant association metrics are all inferred from these data. Before performing the analytical steps outlined in our method protocol, some measures are taken to ensure that the data is appropriately organised.

### 3.4.1 Genetic data

Our GWAS method is applicable to any form of binary genetic data, including SNPs, genes, INDELs, and k-mers We prefer to analyse genome-wide data by applying our approach to core SNPs and accessory gene presence-or-absence data (see Chapter 2). We recommend this approach to ensure good coverage of pan-genome variation. Genetic datasets must be aligned, with individuals in their rows and loci in their columns, and with unique identifiers labelling each. As no associative relationship can be inferred at a non-polymorphic locus, we exclude fixed loci (MAF = 0) from the analysis, once phylogenetic estimation is completed.

We let all genetic variants be represented by binary states. In the treeWAS R package, all loci are encoded in matrices of logical values, which require less memory than numeric or character objects. Any tri- or tetrallelic sites are redefined in terms of biallelic loci. Suppose the bases A, C, and G appear at a given position (i.e., column) in the sequence alignment. To express this in terms of binary variables, we replace the original column with three new dummy variables, such that the original triallelic column is represented by three biallelic columns. At the first of these new new biallelic sites, individuals with allele A at the original locus will be assigned state 1, and all other individuals will be assigned state 0. This procedure will be repeated for allele C, at the second dummy locus, and for allele G at the third. Submitting these three biallelic sites to analysis by our GWAS approach allows us to identify associations between the phenotype and any of the three original alleles.

Our approach permits missing values in genetic data, which arise due to sequencing errors or low-quality samples. Where necessary, the inferences required for data simulation and association testing can be made on the basis of partial information. Any genome or genetic locus entirely composed of missing values is, however, removed prior to analysis. We also recommend the exclusion of any locus whose states are known among $\leq 25\%$ of individuals. As these loci are unlikely to generate sufficient support for association, their removal is likely to improve computational efficiency without sacrificing power.

Less efficient or scalable GWAS methods often attempt to further alleviate their computational burden by restricting GWAS to non-synonymous SNPs in coding regions, perceiving the additional erosion of statistical power as a marginal and acceptable cost. Although protein-coding regions account for the majority of microbial GWAS discoveries made to date, we are conscious of the mounting evidence linking non-coding and synonymous SNPs to phenotypes of interest, often exposing important regulatory relationships [274, 275]. We therefore prefer to retain synonymous SNPs and variants in non-coding regions within the genetic datasets submitted to our GWAS approach. To facilitate this, we have pursued other means of minimising the computational expense of our approach (see Section 3.11.2).

### 3.4.2 Phenotypic data

Our approach is applicable to most phenotypic data types. The phenotype may be binary, for example, indicating whether an isolate is "pathogenic" or "commensal". Categorical phenotypes are also permitted, although these must be ordered, interval variables It would be inappropriate to use our method, for instance, to identify associations with a tertiary host association phenotype containing "cows", "chickens", and "pigs", as these levels do not differ in magnitude. Our approach could, however, be applied to the toxicity phenotype examined by Laabei *et al.* [120], as the authors have classified initially-continuous data into categories of "low", "medium", and "high" toxicity. The phenotype may also be a continuous variable, like MIC values measuring drug resistance on a continuous scale. The phenotype submitted to analysis by the treeWAS R package must be a vector or factor whose names correspond to the row names of the genetic dataset. Any individual whose phenotype is unknown is excluded from the analysis.

#### Imbalanced phenotypes

The accuracy of association tests can be reduced by the imbalanced sampling of phenotypic states (e.g., an extreme ratio of "cases" to "controls") [276]. If the phenotypic distribution is clearly skewed or significant phenotypic outliers are present in a sample, relevant variation between the phenotypes of most individuals can be overshadowed by the large differences between the bulk of the values and the rarer extreme phenotypes.

We recommend that users consider the distribution of phenotypic states prior to analysis. Efforts should be made during sampling to ensure adequate representation of the phenotype at each of its levels or across its range. For skewed continuous variables, it may also be possible to improve the analysis by transforming the phenotype. Transforming the phenotype by rank, for example, will lead to a more uniform distribution of phenotypic values, spreading the initially-skewed values more evenly within their range. Association inference may be improved by analysing these relative phenotypic values, as greater weight will be given to a larger proportion of the variation contained in a dataset.

## 3.5 Phylogenetic reconstruction

Our analytical approach begins by identifying the phylogenetic tree that describes the ancestral relationships linking the $N_{ind}$ isolates under analysis. To obtain a reliable estimate of ancestry, phylogenetic reconstruction is performed on aligned whole-genome sequence data, regardless of the type of genetic data under analysis. If, for example, GWAS is being performed on accessory gene presence-or-absence data, we still wish to analyse patterns of gene gain and loss along the clonal evolutionary history of the sample, as inferred from genome-wide data. Because the tree topology and the relative length of branches shape the data simulated within our method and inform the inferences made by our approach, a reliable reconstruction of the phylogenetic tree is important.

To uphold the evolutionary model embodied by our phylogenetic tree, we must reconstruct only the vertical process of genetic inheritance. If the absence of recombination has been empirically established for a given dataset, standard phylogenetic methods should be able to reliably reconstruct ancestral relationships. For purely clonal organisms, tree-building can be performed within treeWAS via maximum parsimony [277] or distance-based methods [186, 190]. Under these circumstances, our R package allows users to simply set the "tree" argument of the treeWAS function to one of "parsimony", "NJ", or "BIONJ".

In any sample where recombination has not been ruled out, we encourage users to account for recombination during phylogenetic reconstruction. To this end, the treeWAS R package provides tools for integration with ClonalFrameML, facilitating conversion between the output of ClonalFrameML and the format required by our method. We use a standard parsimony method, which does not enforce ultrametricity, to identify an initial estimate of the clonal genealogy. We submit the initial phylogenetic estimate to a recombination-aware phylogenetic method, using the ClonalFrameML [221] software. This resolves inaccuracies in the initial tree caused by homologous recombination, which can distort branch lengths and tree topology [213, 215, 216].

We measure branch lengths in units of the expected number of substitutions per site across the tree. By excluding the effects of recombination from the clonal genealogy, we let branch lengths reflect the number of substitutions introduced by mutation. These branch lengths do not include substitutions due to recombination; but, instead, they represent the amount of evolutionary opportunity available for the introduction of substitutions by recombination events. For parsimony trees, we use the ACCTRAN [193] procedure to infer the minimum $N_{sub}$ per site on each branch $b$, where $b \in \{1, ..., N_{branch}\}$. We define the length of a branch $l_b$ by the $N_{sub}$ on that branch at all $i$ loci, where $i \in \{1, ..., N_{loci}\}$, such that:

$$l_b = \frac{1}{N_{loci}} \sum_{i=1}^{N_{loci}} N_{sub(i,b)} \tag{3.1}$$

When calculating the matrix of pairwise distances for distance-based approaches, we use the Jukes and Cantor [278] substitution model to specify equal substitution rates across sites. If NJ [186] or BIONJ [190] give any $l_b < 0$, we set $l_b = 0$, as negative branch lengths have no clear biological interpretation and may have an undesirable impact on data simulation along the tree. Our simulation procedure is time-reversible. Our GWAS method is applicable to both rooted and unrooted phylogenies, as well as ultrametric and non-ultrametric trees.

## 3.6 Reconstructing ancestral states

To better inform our data simulation and association testing procedures, we estimate the ancestral states of the genotype and phenotype at all internal nodes. For the binary genotypic data, we use parsimony [255] to infer the most probable state at ancestral nodes, for each locus (see Chapter 2, Box 2.9). We assume a minimum evolution model and use the two-pass ACCTRAN procedure [193] to infer substitutions between genotypic states along the branches of the tree only where these are required to explain the data observed. A binary state is assigned to each internal node or, in a small number of ambiguous cases, a value of 0.5 may indicate equally probable states.

ML methods can also be used to reconstruct the ancestral genotype (see Chapter 2, Box 2.10). The ML approach is more computationally intensive, but it allows uncertainty in the estimates to be quantified. To incorporate this source of uncertainty into the inferences made in association testing, we work with the marginal likelihoods of the binary genotype, instead of working with the point estimates, as in parsimony. In Chapter 4, through applications to simulated data, we compare the performance of our approach with parsimonious and ML reconstructions (results in Tables 4.3, 4.4, and 4.5). No significant difference in the overall performance of our approach is observed with either reconstruction method. Users of the treeWAS R package may use ML methods to reconstruct ancestral genotypes, if they prefer. However, we recommend parsimony, as it can achieve similar performance in minimal computational time.

For the phenotype, the appropriate ancestral reconstruction method depends on the data type. Binary phenotypes can be reconstructed via parsimony, using the same approach as for the genotypic data [193, 255]. Continuous phenotypes should be reconstructed with continuous ML methods [236, 273]. Categorical phenotypes can be reconstructed by either parsimony or ML methods and treated as either discrete or continuous variables. A parsimonious or a discrete ML reconstruction should be selected if it is illogical for ancestral nodes to have taken on intermediate phenotypic states. These methods will treat the transitions between all states as equally likely, unless relative weights are specified by the user [257] If the levels of a categorical phenotype represent classifications of a naturally continuous trait, like host age in years, a continuous ML reconstruction is likely to be more appropriate. Point estimates of the ancestral phenotypic states are analysed by our association tests.

## 3.7 Accounting for recombination and mutation

### 3.7.1 Impact on spurious association

In Chapter 2, we discussed how variable rates of substitution due to recombination and mutation can alter the probability of spurious association and thus confound the inferences made during association testing (see Figure 2.7). Both evolutionary processes can produce genuine associations with a phenotype. Recombination, in particular, can drive rapid phenotypic change by repeatedly introducing substitutions in adaptive alleles. If non-associated sites undergo far fewer substitutions, associated sites may stand out more clearly against this background. But, if substitutions occur frequently across the genome due to widespread mutation and recombination, some proportion of phenotypically-neutral alleles may correlate with the phenotype by chance alone. To distinguish between spurious and genuine associations, we therefore need a way to account for the impact of mutation and recombination in microbial GWAS. Without accounting for this source of bias, existing GWAS methods are likely to over- or underestimate the support for association. Somewhat surprisingly, microbial GWAS methodology has largely overlooked this issue. In contrast, we anticipate that by accounting for the number of substitutions observed, our GWAS approach will achieve greater accuracy and reliability when making inferences about association in a range of datasets.

### 3.7.2 The homoplasy distribution

After reconstructing ancestral states, we identify the homoplasy distribution, comprising the number of substitutions per site at all empirical loci. We use Fitch's parsimony procedure [255] to calculate $N_{sub}$ at each site, defining $N_{sub}$ as the parsimony cost, or the minimum number of substitutions that must have occured across the clonal genealogy. The parsimony procedure can be applied to any type of binary genetic data, whether these are core SNPs or accessory gene presence-or-absence data; only the unit of substitution changes. Ancestral states and substitutions are inferred at recombinant as well as non-recombining sites. Though recombination events are excluded from the clonal genealogy, we do not remove recombinant regions from the sequences under analysis. We can thus identify substitutions along the tree due to both vertical and horizontal evolutionary mechanisms.

### 3.7.3 Implementation and rationale

Within our simulation procedure, we account for the effects of both mutation and recombination at the level of individual substitutions. The simulation procedure is described in detail below. Simply stated, however, at each simulated locus, $N_{sub}$ is drawn from the homoplasy distribution and $N_{sub}$ substitutions are redistributed with uniform probability along the phylogenetic tree. As we reassign individual substitutions to new branches on the phylogenetic tree, the link between our simulation procedure and the process of mutation is straightforward. The ability of our approach to simulate the effects of recombination is perhaps less intuitive.

An alternative approach might attempt to more closely recreate the genuine process of recombination, in which substitutions tend to be generated at multiple adjacent sites through the introduction of an exogenous sequence fragment. Farhat *et al.* [106], for example, suggest that each recombinant fragment should be treated as a "single site". In such an approach, instead of reassigning individual substitutions to new branches of the phylogeny, one might try to reassign contiguous sequence fragments. To do this, one would need to estimate the distribution of recombinant sequence fragment lengths in the evolutionary history of the sample. Attempting to infer these unobserved events from the data would be computationally expensive and it would introduce an additional source of error into the simulation procedure. Critically, we do not believe this would improve the simulation procedure or our estimation of the null distribution.

To explain why, we can compare two hypothetical scenarios, occurring in the evolutionary history of a sample: (A) A 5 bp sequence fragment has recombined into the ancestral genome at three locations along the phylogenetic tree, or (B) 5 adjacent loci have undergone three mutations on the same three branches of the tree. Both (A) and (B) would produce the same pattern in the genome sequences of sampled individuals. In the homoplasy distribution, both would be represented by five values of $N_{sub} = 3$. We have two options for how to simulate these events in the null genetic dataset. For the mutation events in (B), we could simulate five separate loci by randomly reassigning three substitutions to new branches of the phylogeny in each case. For the recombination event in (A), we could either do the same and treat the five loci separately, or we could reassign the three substitutions at these five loci together to the same randomly-selected branches of the phylogeny.

In the latter case, the five simulated loci would achieve identical association scores, resulting in five identical contributions to the null distribution. In the former case, by contrast, the association scores at these five loci would probably differ, and five separate values would be added to the null distribution. Recall, however, that we simulate $\geq 10N_{loci}$ sites. Moreover, most values of $N_{sub}$ repeat at hundreds or thousands of empirical loci. Hence, whether we make five identical contributions to the null distribution or five separate contributions in this instance, over many such randomisations we will still converge on the same null distribution.

Our simulation procedure is therefore able to recreate the impact of recombination without having to precisely recreate the process of recombination. We can use the homoplasy distribution to summarise both mutation and recombination. And, by simulating the randomised outcome of both processes, our approach can generate the appropriate distribution of substitutions-per-site and accurately estimate the resulting probability of spurious association.

## 3.8 Data simulation

The purpose of the data simulation procedure implemented within our association testing approach is to generate a “null” genetic dataset that represents the null hypothesis of “no association”. Our simulation procedure recreates the following features of the empirical dataset, each of which influence the underlying probability of chance association:

1. The sample size.
2. The population structure.
3. The phenotypic distribution.
4. The mutation rate.
5. The recombination rate.
6. The genetic composition (MAF).
7. The strength of linkage disequilibrium.

Each of these potentially confounding factors affects the probability of spurious association in the dataset under analysis. Yet, these spurious associations cannot be immediately distinguished from genuine associations. Our motivation for simulating data, therefore, is to reproduce the degree of association observed due to confounding factors alone. Then, by comparing the empirical dataset to its simulated counterpart, we can determine which associations are likely to be caused by chance or confounding bias and which garner sufficient support to suggest a true biological link.

We reconstruct the phylogeny, ancestral states, and homoplasy distribution as described above. Our estimates of these empirical parameters are then used to inform the simulation process. We simulate $N_{sim}$ loci, such that $N_{sim} \geq 10N_{loci}$, where the optimal value of $N_{sim}$ was determined through applications to simulated data (see Chapter 4). This ensures that the null distribution is estimated with sufficient accuracy and that it accounts for the variability inherent in our simulation-based approach. Finally, we maintain the empirical distribution of phenotypic states along the terminal nodes of the tree. This allows us to account for population stratification, considering both the strength of population structure and the extent to which phenotypic states cluster in ancestral lineages. We assume that, under the null model, mutations and recombination events happen at a constant rate along branches of the clonal genealogy. Substitution events occur independently of one another, constrained only by the tree structure and homoplasy distribution. We generate each simulated genetic locus $g_i$, where $i \in \{1, ..., N_{sim}\}$, by following the simulation procedure in Box 3.1.

1. The number of substitutions, $N_{sub}$, to occur along the tree at genetic locus $i$ is randomly drawn from the homoplasy distribution.
2. $N_{sub}$ branches of the phylogenetic tree are randomly sampled without replacement, and the probability of selecting branch $j$ depends on its relative length:

   $$Pr(branch_j) = \frac{l_j}{\sum_{j=1}^{N_{branch}} l_j} \tag{3.2}$$

   Sampling without replacement ensures that the per-branch $N_{sub} \leq 1$ at each locus, to uphold the minimum evolution model assumed by the parsimonious homoplasy distribution.
3. The state to be assigned to the root of the phylogeny is randomly sampled from a discrete uniform distribution of the possible $g_i$ values, where $g_i \in [0, 1]$.
4. The states of $g_i$ at descendant nodes are set to be the same as their direct ancestor, unless the branch $j$ connecting $g_i^{anc}$ to $g_i^{des}$ was assigned a substitution in Step 2. If this is the case, then we let $g_i^{des} = |g_i^{anc} - 1|$.
5. Step 4 is repeated until the simulation of $g_i$ completes its journey from root to tips, and the states of $g_i$ have been defined at all internal and terminal nodes.
6. Calculate the MAF of $g_i$ at terminal nodes. If $g_i$ is not polymorphic, repeat Steps 2 to 5 until MAF $\geq 1/N_{term}$, to ensure that some genetic variation is present at each simulated locus, as in the empirical dataset.

,

**Box 3.1. Simulation of genetic loci.**

We ground all stochastic processes in pseudorandom number generation, which was, in fact, developed for use in similar Monte Carlo simulation procedures [279]. So, in Box 3.1, Steps 1-3 truly contain pseudorandom sampling processes. This ensures that effectively random results can be achieved for each simulated locus and with each run of the algorithm. At the same time, it gives users the opportunity to specify a seed, so that identical results can be reproduced in subsequent analyses, if desired.

The simulation process allows the observed relationship between genotype and phenotype to vary, but only within constraints established by the empirical dataset. Our simulations recreate the empirical phylogenetic structure and approximately maintain the observed LD strength and MAF distribution. This allows us to estimate the statistical non-independence between isolates due to ancestry and between loci due to linkage. Hence, whereas clonal relationships undermine the precision of most existing GWAS methods, our approach is better prepared to delineate genuine associations from correlated spurious findings. By incorporating the homoplasy distribution, our simulation process recreates the variation introduced by both mutation and recombination. With a recombination-aware phylogeny, this renders our GWAS method applicable to recombinant as well as clonal organisms. As we demonstrate in Figure 3.2, preserving the $N_{sub}$ distribution and tree structure improves our ability to recreate the empirical population structure and distribution of allele frequencies [259]. The strength of association and the probability of chance correlation with a phenotype are influenced by the MAF of empirical loci and the underlying processes of mutation and recombination [241, 280]. The ability to simulate these parameters and account for their effects is a valuable addition to our GWAS method.

Altogether, our simulation-based approach allows us to acount for the effects of seven properties of the empirical dataset, each of which influences the potential for confounding bias in association inference. Only one or two of these are addressed by most existing microbial GWAS approaches. In addition, our approach inherently allows us to account for interactions between these factors. Alternative approaches may struggle to accurately quantify the individual and collective impact of this complex web of confounding variables. Our Monte Carlo method provides a natural solution to this problem, replacing direct calculation with estimation through simulation. As the simulated dataset captures known confounders found in the empirical dataset, but does not recreate any genuine associations with the phenotype, we proceed under the assumption that this dataset represents the null hypothesis. Therefore, by comparing the distributions of association score statistics in the empirical and simulated datasets, we should be able to separate true signals of association from the noisy background of associations due to chance and confounding factors.

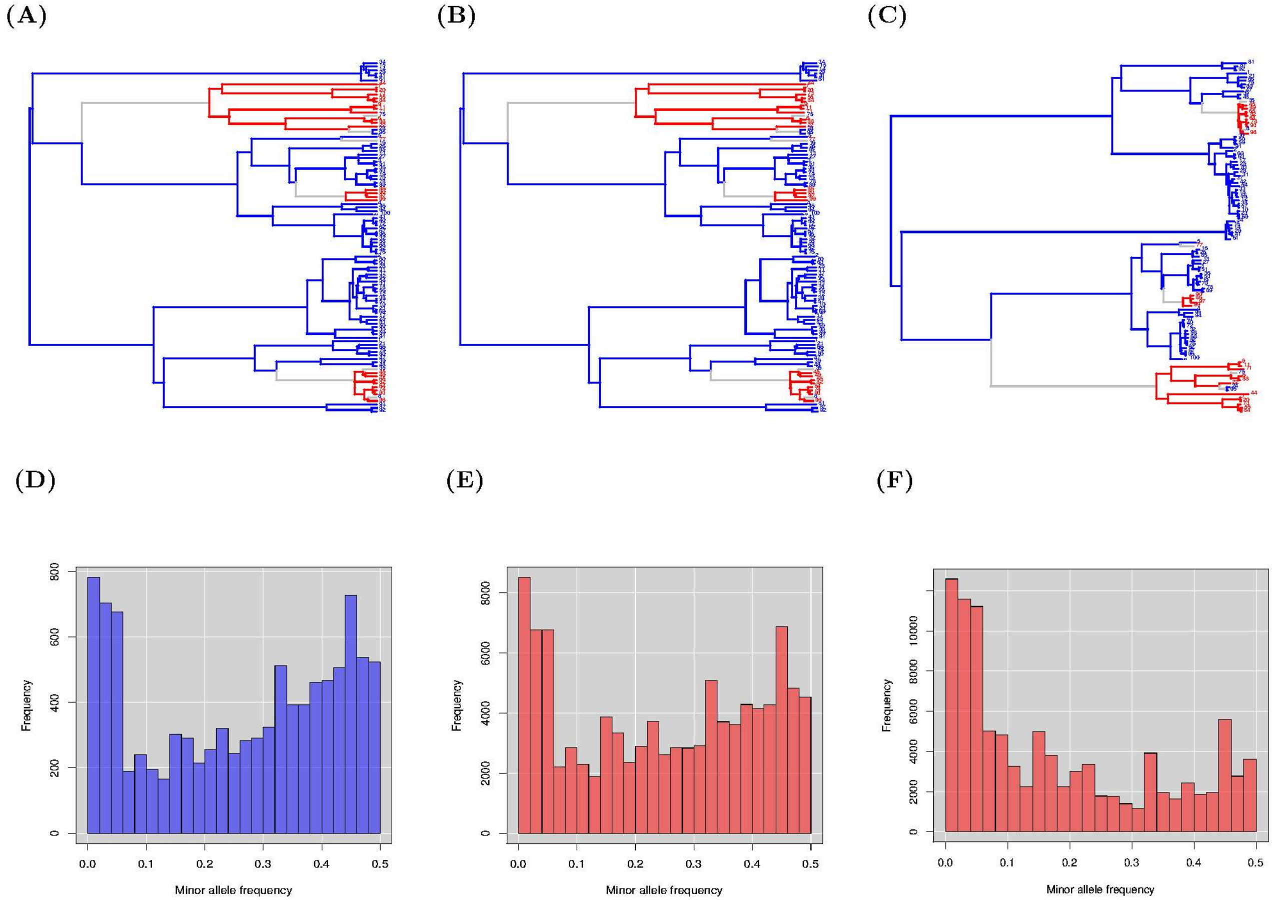


**Figure 3.2. Accurate simulation of population structure and MAF.** We compare an empirical dataset ($N_{ind} = 100, R = 0$) to two simulated datasets, and we evaluate the similarity between each simulation and the original dataset. Both synthetic datasets were simulated along the empirical phylogenetic tree. However, the simulation of $N_{sub}$ was based on two different estimation procedures, with (i) the empirical $N_{sub}$ inferred as in Box 3.1, and (ii) $N_{sub}$ drawn from a Poisson distribution ($\lambda = 1$), as previously proposed [118]. **A:** The empirical tree is clearly more faithfully re-created by **B:** the tree with empirical $N_{sub}$ than by **C:** the tree with $N_{sub} \sim Pois(1)$. Likewise, a much better approximation to **D:** the empirical allele frequency (MAF) distribution is provided by **E:** MAF simulated with empirical $N_{sub}$ than by **F:** MAF simulated with $N_{sub} \sim Pois(1)$. These observations show that our new simulation procedure is able to reproduce key confounding features of empirical datasets, like the population structure and genetic composition, with far greater accuracy than existing approaches.

## 3.9 Tests of association

An effective GWAS method must (i) eliminate false positive findings, and (ii) retain the statistical power to identify genuine associations. Our data simulation procedure, above, facilitates a robust estimation of the empirically-shaped null distribution. This addresses our first marker of effectiveness by enabling strict control over FPR in a wide range of datasets, with any measure of association. In GWAS, control over type I error often comes at the expense of control over type II error. Yet, the low FPR of our simulation-based approach, in fact, presents an opportunity to augment the discovery power of our GWAS method. We can attempt to increase the sensitivity of our approach by implementing multiple independent tests of association. While GWAS methods with weaker control over FPR would be unable to benefit from this strategy, we can draw on the findings of multiple association tests to increase power without excessively compounding the FPR of our approach as a whole.

We have designed three separate association scores to capture distinct, if overlapping, signals of association. Each score is described in detail below (for notation, see Table 3.1). Figure 3.3 summarises the purpose of each score, alongside an illustration of the type of association it aims to detect. Our three measures are designed to complement one another. By expanding the search for associations, our three measures should increase the probability of detecting genuine associations. Collectively, these measures should equip our GWAS method with greater power and flexibility, which will be a valuable asset when facing unknown association landscapes in each new dataset analysed.

As opposed to existing phylogenetic GWAS methods, all three of our association tests can be applied to binary, categorical, and continuous phenotypic data, and any type of genetic data. This prevents the information loss and artificial categorisation required when phenotypes must be reclassified as binary or categorical variables. In addition, unlike the *ad hoc* or comparative applications of multiple association tests seen in some microbial GWAS studies [120, 122, 132, 134], our three association tests have been specifically designed to work together within a single, cohesive approach.

For reasons disussed below, and tested in Chapter 4, we apply our three association scores in parallel. We quantify the relationship between each genetic locus and the phenotype, in both the simulated and empirical datasets, with each measure of association. We estimate a null distribution for each score from the values calculated for the simulated dataset. This enables the identification of significant associations in the empirical dataset, with each measure. Finally, each test contributes a set of well-supported associations to the collective set of significant findings for our approach overall.

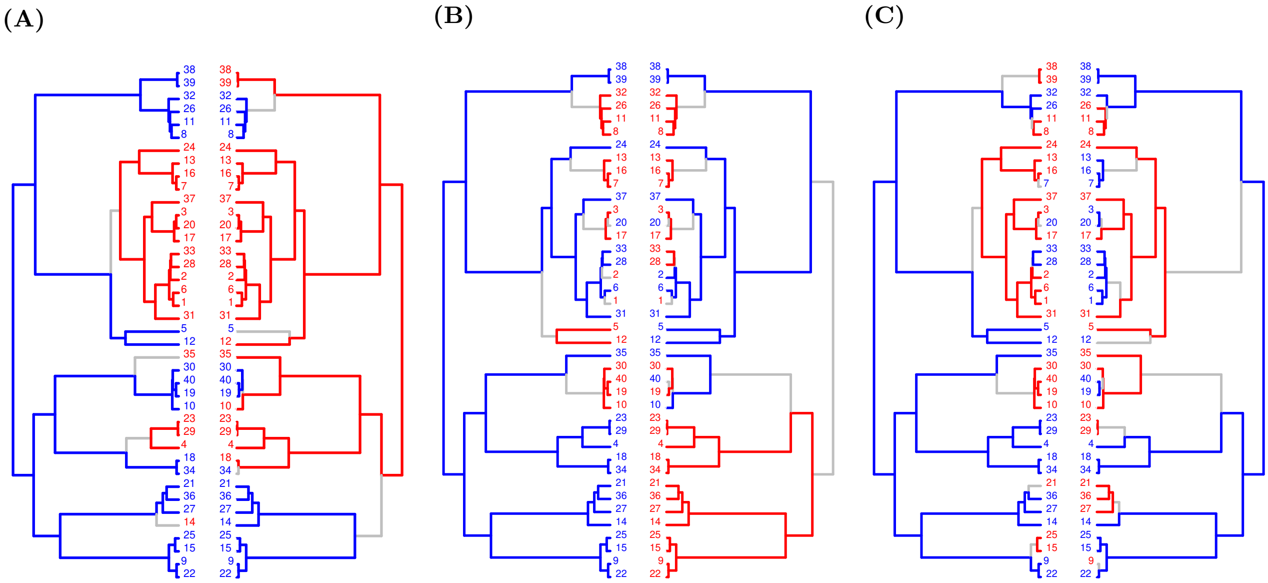


**Figure 3.3. Evolutionary scenarios detected by our association scores.** The three complementary tests of association in treeWAS assign high scores to different patterns of association, examples of which are illustrated above. Each panel displays the phenotype (left) and the genotype of one associated locus (right), with binary states plotted along the tips of the phylogenetic tree ($N_{term} = 40$) and reconstructed ancestral states indicated along the branches of the tree (blue = 0, red = 1, grey = substitution). **A:** Score 1 aims to detect association among terminal nodes and assigns a relatively high value of 0.7 to this terminal configuration of phenotypic and genotypic states. **B:** Score 2 measures association by counting how many branches contain a substitution in both genotype and phenotype, assigning this pattern a score of 5. **C:** Score 3 is designed to find associations maintained loosely across the phylogenetic tree, resulting in a Score 3 value of 10 in this scenario.

### 3.9.1 Score 1

Score 1, the "Terminal Score", measures sample-wide association across the leaves of the phylogenetic tree. With Score 1, we perform a straightforward allele-based test of association on the basis of observed data only. This measure operates on the same theoretical principles as the Quadrant Score described by Kruskal [281], rendering it applicable to association testing, as previously proposed by Sheppard *et al.* [147]. For a binary phenotype, Score 1 is calculated on the basis of a 2x2 contingency table containing the possible combinations of terminal states, with and without the phenotype, and with and without the genotype. Generalizing to continuous phenotypes gives Equation 3.4.

$$\textbf{Score 1} = \left| \sum_{i=1}^{N_{term}} \frac{1}{N_{term}} (p_i^{des} g_i^{des} + (1 - p_i^{des})(1 - g_i^{des}) - (1 - p_i^{des}) g_i^{des} - p_i^{des}(1 - g_i^{des})) \right| \tag{3.4}$$

The $N_{term}$ denominator in Equation 3.4 ensures that Score 1 values always fall in [-1,1]. The null and empirical distributions for Score 1 (as for Scores 2 and 3) are expected to be approximately symmetric around zero. This allows us to take the absolute value of each association statistic when estimating the phylogenetically-correct significance level of empirical associations, enabling the identification of a single significance threshold for Score 1 (as described below). The treeWAS R package reports the initial sign of each association, to inform users of the directionality of all quantified relationships between genotype and phenotype. We use the same convention to express directionality for Scores 1, 2, and 3, where $S_i$ is the value of one of our three measures of association at locus $i$:

$$S_i \begin{cases} < 0 & \text{net negative association } (g_i^{des} p_i^{des} \in \{01, 10\}, \text{ predominantly}) \\ > 0 & \text{net positive association } (g_i^{des} p_i^{des} \in \{00, 11\}, \text{ predominantly}) \end{cases} \tag{3.5}$$

With respect to Score 1, association at a given locus is defined as the over-representation of an allele among individuals with a particular phenotypic state, or according to the value of the phenotype. To achieve a high Score 1 value, the relationship between genotype and phenotype must be upheld across a relatively large proportion of terminal nodes. In Figure 3.3A, for example, 85% of terminal nodes (34/40) are positively associated, while only 15% (6/40) display negative association. As a result of this relatively consistent pattern of association, Equation 3.4 produces a high Score 1 value of 0.7 ($= 0.85 - 0.15$), suggesting that the genotype and phenotype do not vary independently.

Score 1 considers observed data only, ignoring reconstructed ancestral information. Measurements of association strength made by Score 1 will, therefore, be robust to errors in phylogenetic or ancestral state reconstruction. Score 1 will also retain the capacity to identify terminal associations even in the absence of homoplastic substitutions. For example, in Figure 3.3A, the inferred ancestral states show that no branch of the phylogenetic tree contains a coincident change in genotype and phenotype. According to the homoplasy-counting framework for association testing, this scenario would provide no evidence for association. Yet, by focusing only on the observed genotypic and phenotypic states of sampled individuals, we can clearly see widespread allele-based association along the tips of the tree. Score 1 aims to ensure that such relationships are not overlooked.

The null distribution allows us to evaluate the evolutionary case for association. For example, the perfect co-distribution of terminal states in Figures 3.4A and B produces the maximum possible Score 1 value of 1.0 in Figure 3.4C. Yet, with a single coincident origin of allele 1 and phenotype 1 on the tree, there is no reason to believe that these variables have a functional or adaptive relationship. This is an extreme case of “pseudoreplication”, a problem articulated in phylogenetic comparative methods [148] and inherited in microbial GWAS. For example, because the phylogenetic PhyC approach [114] does not estimate a genome-wide null distribution, it has been found to attribute significance to pseudoreplicates [134]. Thankfully, for sound evolutionary reasons, the null distribution simulated by treeWAS places this association below the significance threshold. Hence, by assessing the statistical strength of association and evaluating the empirical context through data simulation, our approach can identify allele-based associations while accounting for pseudoreplication and population stratification.

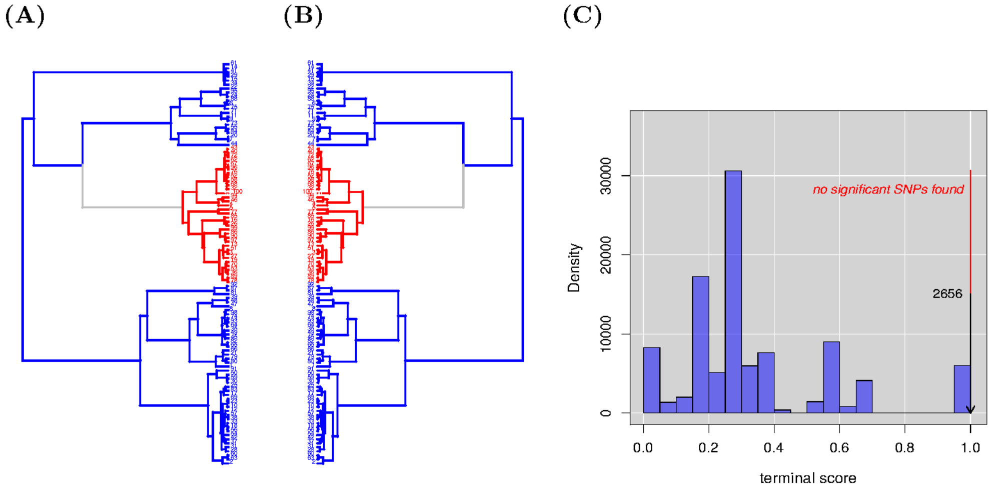


**Figure 3.4. Pseudoreplication.** We have simulated a dataset in which only genetic locus 2656 is in association with the phenotype ($N_{term}$ = 100, $N_{loci}$ = 10,000). The effect size of this association is 100%. The binary states of **A:** the phenotype, and **B:** the genotype at associated locus 2656 are indicated along the phylogenetic tree (blue = 0, red = 1, grey = substitution). **C:** The null distribution for Score 1 is plotted and we indicate the Score 1 value for the “associated” locus, although it falls below the significance threshold. Locus 2656 attains a Score 1 value of 1.0, indicating perfect terminal association. In isolation, the observed degree of association seems statistically improbable by chance alone. Yet, the null distribution encourages us to reject this finding by revealing that, in fact, all loci that undergo a substitution on the same single branch would achieve an equally extreme association score value. Even when focused on terminal association, as in Score 1, both statistical and evolutionary significance are required for association.

### 3.9.2 Score 2

Score 2, the "Simultaneous Score", takes an orthogonal approach to our first association score, adopting a homoplasy-based approach that defines association as the over-representation of repeated, independently-emerging, coincident state changes in both genotype and phenotype. With Score 2, we extend the association test back into the evolutionary history of the sample, where it can capitalise on the additional information inferred during phylogenetic and ancestral state reconstruction. Equation 3.6 quantifies the amount of association between the phenotype, $p_i$, and the genotype at a given locus, $g_i$, across each branch $i$ of the tree.

$$\textbf{Score 2} = \left| \sum_{i=1}^{N_{branch}} (p_i^{anc} - p_i^{des})(g_i^{anc} - g_i^{des}) \right| \tag{3.6}$$

For a binary phenotype with a parsimonious ancestral state reconstruction, Equation 3.6 is equivalent to counting the number of branches containing a simultaneous substitution in genotype and phenotype. Score 2 was initially designed to count substitutions, though Equation 3.6 now achieves the same goal for all phenotypes. To preserve this information, we do not include a denominator as we had done in Equation 3.4. We also deliberately exclude branch length from Equation 3.6, as we assume that all simultaneous substitutions provide equally valid evidence of association.

Unlike Score 1, Score 2 does not require a sample-wide relationship to infer association. In Figure 3.3B, for example, there is no net association across the tips of the tree (Score 1 = 0). But, as five simultaneous substitutions do provide evidence of association, Score 2 achieves a relatively high value of 5. By measuring distinct signals of association, therefore, our second and first association tests act to complement one another. In clonal populations, strong population stratification may prevent Score 1 from detecting associations. Yet, if some loci undergo homoplastic mutations that deviate from the pattern of ancestral correlation, Score 2 may be able to detect associations by separating these loci from the genome-wide LD. In recombinogenic organisms, the horizontal introduction of many trait-associated variants may modulate the phenotype to varying degrees across the sample. If association at any one variant is insufficient for detection by Score 1, it may nevertheless be caught by Score 2.

#### Complementary pathways

Our second association test is also designed to detect loci that give rise to the phenotype through complementary pathways. Equation 3.6 imparts a cumulative character to Score 2. Simultaneous substitutions increase the score, but branches with substitutions in one or neither variable do not decrease its value. Equally significant values of Score 2 might be achieved by a strong association upheld in half of the phylogeny, or a weaker association upheld across the tree. Returning to Figure 3.3B, we can see that Score 2 finds evidence for association only in the upper-most of the two major clades, at this locus. Other loci may be responsible for the phenotype in the lower clade. Score 2 may detect elements of both complementary pathways, whereas Score 1 would reject each on a sample-wide basis. Hence, by pairing our first allele-based measure with this homoplasy-counting approach, we expand the scope of our search for associations.

### 3.9.3 Score 3

Score 3, the "Subsequent Score", aims to quantify the association between genotype and phenotype across the entire phylogenetic tree. It combines elements of both the allele-based and homoplasy-counting frameworks. Score 3 attempts to identify any associations that may have been overlooked, if any gaps have been left by Scores 1 and 2.

Score 3 is the mathematical solution to the integral of an association score, $C_x$, that is measured at all points along the tree. Let $P_x$ and $G_x$ represent, respectively, the probable value of the phenotype and genotype at a point $x$ on a branch $i$ of length $l_i$. $P_x$ and $G_x$ are identified using a linear interpolation between the known or reconstructed states at the ancestral and descendant nodes of branch $i$.

$$P_x \;= \frac{1}{l_i}(p_i^{anc}(l_i - x) \;+\; p_i^{des}x) \tag{3.7}$$

$$G_x = \frac{1}{l_i}(g_i^{anc}(l_i - x) \;+\; g_i^{des}x) \tag{3.8}$$

Let the initial "correlation score", $C_x$, represent the degree of association between the phenotype and genotype at point $x$. We can express $C_x$ in terms of $P_x$ and $G_x$:

$$C_x \;= P_xG_x + (1 - P_x)(1 - G_x) - P_x(1 - G_x) - (1 - P_x)G_x \tag{3.9}$$

Fully describing $C_x$ in terms of its components, $P_x$ and $G_x$, gives us the following expanded form of $C_x$, which can be measured at any point $x$ along branch $i$.

$$\begin{aligned} C_x \ = & \frac{p_i^{anc}(l_i - x) \ + \ p_i^{des}x}{l_i}\frac{g_i^{anc}(l_i - x) \ + \ g_i^{des}x}{l_i} + \\ & \left(1 - \frac{p_i^{anc}(l_i - x) \ + \ p_i^{des}x}{l_i}\right)\frac{g_i^{anc}(l_i - x) \ + \ g_i^{des}x}{l_i} - \\ & \frac{p_i^{anc}(l_i - x) \ + \ p_i^{des}x}{l_i}\left(1 - \frac{g_i^{anc}(l_i - x) \ + \ g_i^{des}x}{l_i}\right) - \\ & \left(1 - \frac{p_i^{anc}(l_i - x) \ + \ p_i^{des}x}{l_i}\right)\frac{g_i^{anc}(l_i - x) \ + \ g_i^{des}x}{l_i} \end{aligned} \tag{3.10}$$

Let Score 3 be defined as the absolute sum, for all branches $i$, of the integral of $C_x$, where the point $x$ takes all positions along branch $i$, from 0 to $l_i$.

$$\mathit{Score}\ 3 = | \sum_{i=1}^{N_{branch}} \int_0^{l_i} C_x dx \ | \tag{3.11}$$

Instead of solving the integral of $C_x$ numerically upon each calculation of Score 3, we have been able to achieve greater computational efficiency by solving the integral mathematically:

$$\begin{aligned} \mathit{Score}\ 3 = | \sum_{i=1}^{N_{branch}} \ & (-1 + 2p_i^{anc})(-1 + 2g_i^{anc})l_i - \\ & (-p_i^{anc} + p_i^{des} - g_i^{anc} + 4p_i^{anc}g_i^{anc} - \\ & 2p_i^{des}g_i^{anc} + g_i^{des} - 2p_i^{anc}g_i^{des})l_i \ + \\ & \frac{4}{3}(p_i^{anc} - p_i^{des})(g_i^{anc} - g_i^{des})l_i \ | \end{aligned} \tag{3.12}$$

Simplifying this gives Equation 3.13

$$\begin{aligned} \mathit{Score}\ 3 = | \sum_{i=1}^{N_{branch}} \ & l_i \frac{4}{3}p_i^{anc}g_i^{anc} \ + \ \frac{2}{3}p_i^{anc}g_i^{des} \ + \ \frac{2}{3}p_i^{des}g_i^{anc} \ + \ \frac{4}{3}p_i^{des}g_i^{des} \\ & - \ p_i^{anc} \ - \ p_i^{des} \ - \ g_i^{anc} \ - \ g_i^{des} \ + \ 1 \ | \end{aligned} \tag{3.13}$$

Our evaluation of Score 3 through applications to simulated data (Section 4.3.1) indicates that the performance of Score 3 is improved by removing the branch length term, $l_i$. This produces the final form of Equation 3.14, which gives all edges equal weight. We calculate Score 3 as follows, for each genetic locus and across all branches of the tree:

$$\begin{aligned}\textbf{Score 3} = |\sum_{i=1}^{N_{branch}} & \frac{4}{3}p_i^{anc}g_i^{anc} + \frac{2}{3}p_i^{anc}g_i^{des} + \frac{2}{3}p_i^{des}g_i^{anc} + \frac{4}{3}p_i^{des}g_i^{des} \\ & - p_i^{anc} - p_i^{des} - g_i^{anc} - g_i^{des} + 1 \,|\end{aligned} \tag{3.14}$$

The contour plot in Figure 3.5 illustrates how Equation 3.14 responds to change in the genotype and phenotype. As the figure legend explains, Score 3 achieves its highest values whenever a particular genotypic allele and phenotypic state are maintained across a branch of the phylogeny.

If a substitution in one variable is followed, on a subsequent branch, by a substitution in the other variable, Score 3 will incur no penalty for the lack of simultaneous change and will capture the downstream association in so far as it is maintained. In Figure 3.3C, for example, a substitution changes the genotype from 0 to 1 on the branch leading to the largest of four clades in the right-hand phylogeny. A phenotypic substitution then follows on a descendant branch in the left-hand phylogeny. Association is subsequently maintained across many descendant branches within this subtree, though substitutions near its tips disrupt the pattern of association among terminal nodes. This clade and, indeed, the entire Figure 3.3C phylogeny illustrate how Score 3 can detect relationships between genotype and phenotype that emerge along the tree, without requiring substitutions in both variables to occur on the exact same branch. Even in the absence of terminal association (Score 1 = 0) and simultaneous substitution (Score 2 = 0), Score 3 retains the ability to infer association through its more flexible phylogeny-wide measure.

Score 3 may be able to identify probabilistic patterns of association, which fluctuate across the evolutionary history of the sample. In host association, for example, genetic adaptation may contribute to host switching by increasing affinity for a different host or by offering compensatory fitness advantages once in a new environment [147]. Figure 2.8 in Chapter 2 illustrates how these more complex associations may emerge through staggered substitutions and persist following the transition between phenotypic states. Score 3 aims to detect these less deterministic relationships. While Score 1 may capture broad, sample-wide associations and Score 2 may detect convergent evolution or complementary associations, weaker phylogeny-wide associations overlooked by our first two measures may yet be captured by Score 3.

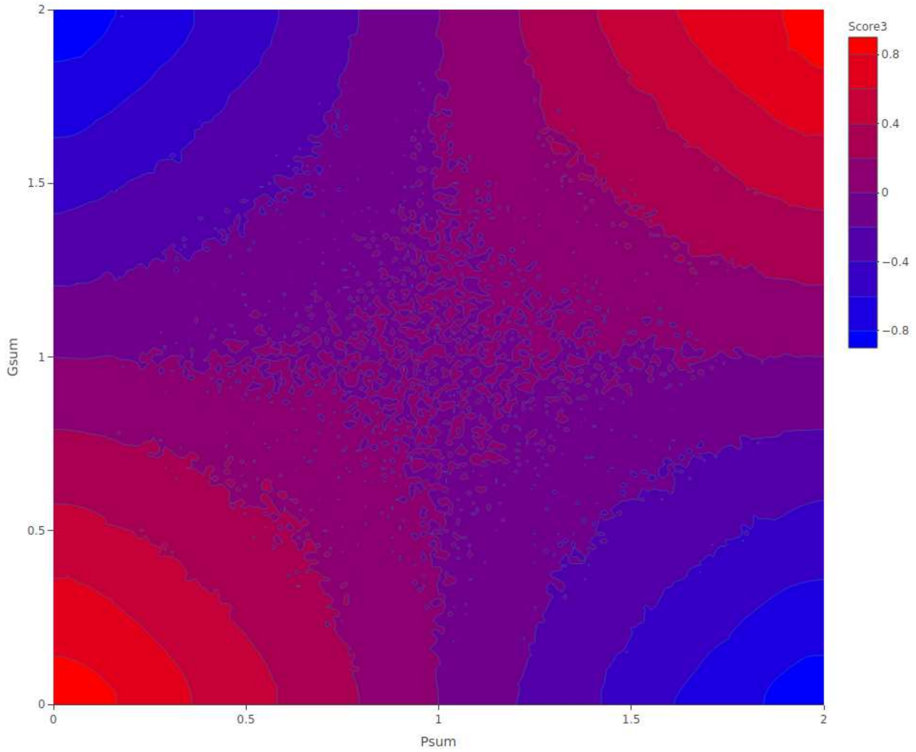


**Figure 3.5. Score 3 contour plot.** This contour plot shows how Score 3 varies as a function of the change in phenotypic and genotypic states across a branch of the phylogenetic tree. The figure is plotted along two summary variables, $Psum$ and $Ssum$, where $Psum = p_i^{anc} + p_i^{des}$ and $Gsum = g_i^{anc} + g_i^{des}$. This allows us to illustrate in two dimensions how the four parameters, $p_i^{anc}$, $p_i^{des}$, $g_i^{anc}$, $g_i^{des}$, affect the value of Score 3. The contour plot uses color to represent the Score 3 values that result from each set of original variables (red = +, purple $\simeq$ 0, blue = −). Towards the corners of the figure, where both phenotype and genotype maintain states near 0 or 1 across a branch ($Psum \in \{0, 2\}$, $Gsum \in \{0, 2\}$), we see regions of increasingly large positive and negative Score 3 values, approaching ±1. Smaller values (Score 3 $\simeq \pm 1/3$) result from simultaneous substitutions, owing to uncertainty in the proximity of the two substitutions along the branch in question. Towards the central region of the figure, values begin to approach zero, reflecting either a state change in only one variable or the maintenance of intermediate states ($\simeq$ 0.5) in both genotype and phenotype.

### 3.9.4 Pooling results

The three association scores described above have been designed to work together. As Figure 3.3 illustrates, each measure may pick up on signals of association that are overlooked by one or both of the other scores. To capture the benefits of each score, we therefore adopt a parallel implementation of our three tests of association. To the best of our knowledge, this has not been done elsewhere in microbial GWAS. Yet, the simulation study in Section 4.5.4 provides ample support for our parallel implementation.

With each score, we quantify the association between the phenotype and the genotype at each locus, in both the simulated and empirical datasets. We select a significance threshold for each score, with reference to the relevant null distribution (as detailed below). One set of significant associations is identified by each test, with a high degree of statistical and evolutionary support. Each of these findings, therefore, constitutes a suitable candidate for further investigation. Although it may provide additional support for a finding, identification by a second or third association test is not required for overall significance.

The findings from our three tests are pooled together to give one larger, collective set of findings for our approach as a whole. Instead of merging our three well-defined measures into one less informative aggregate score, we report three separate scores and p-values for each locus. This improves the interpretability of our findings, by pairing measures of significance with meaningful insight into the nature of each association.

## 3.10 The threshold of significance

We use the null distribution to assess the significance of associations in the empirical dataset. We distinguish significant findings from probable chance associations by selecting a threshold of significance in the upper tail of the null distribution. Identification of the significance threshold involves the following elements:

1. A base p-value, $\alpha_{base}$.
2. A correction for multiple testing.
3. A means of estimating empirical p-values from the null distribution.
4. A number of simulated loci, $N_{sim}$, to make up the null distribution.

Each of these four elements can be achieved in several ways. In Chapter 4 (Section 4.4), we use applications to simulated data to guide our selection of the most appropriate approach in each case. We introduce each approach below, as we have implemented these in the treeWAS R package to offer users greater flexibility. We indicate, at the end of this section, which approach is recommended by our analysis.

### 3.10.1 The base p-value

The base p-value, $\alpha_{base}$, allows us to specify the acceptable type I error rate for a given association study. It represents the probability of incorrectly rejecting the null hypothesis when, in fact, no association is present at the locus in question. With a base p-value of $\alpha_{base}$, we permit a $\alpha_{base} * 100\%$ chance of identifying a false positive association. It is typical to see $\alpha_{base} = 0.05$ in significance testing, allowing a 5% FPR, although this is an arbitrary threshold. We view this as the minimally stringent threshold for significance. To compare the standard $\alpha_{base}$ to a range of more conservative base p-values, we examine results with: $\boldsymbol{\alpha_{\textbf{base}} = 0.05, 0.01, 0.001, 0.0001}$.

### 3.10.2 Multiple testing correction

The multiple comparisons problem states that, as the number of statistical inferences increases, the number of incorrect inferences will also increase. This problem is especially pronounced in GWAS, where the number of statistical tests can number in the millions. Multiple testing correction mechanisms compensate for the increased probability of incorrectly rejecting $H_0$ due to multiple comparisons.

**The Bonferroni correction** provides a simple, conservative solution to the multiple comparisons problem [282]. The Bonferroni correction controls the total number of type I errors, to ensure that the significance of findings from each test remains valid when all tests are considered. With no correction, if the overall significance level $\alpha_{base} = 0.01$, then $\leq 1\%$ of all tests should return false positives. Instead, we work with a per-test significance level, $\alpha$, letting $Pr(T_i = +|H_0) \leq \alpha$, such that $Pr(\text{some } T_i = +|H_0) \leq \alpha_{base}$. We identify $\alpha$ by adjusting $\alpha_{base}$ to account for the total number of independent statistical tests performed:

$$\alpha = \frac{\alpha_{base}}{N_{score}N_{loci}} \tag{3.15}$$

We define the number of tests by multiplying the $N_{loci}$ tested by $N_{score}$, the number of association scores measured at each locus, where $N_{score} = 3$ by default.

| | **Condition Positive**<br>($H_A$) | **Condition Negative**<br>($H_0$) |
|---|---|---|
| **Test Positive**<br>($T_i = +$) | **T**rue **P**ositive<br>(+) | **F**alse **P**ositive<br>(Type I Error) |
| **Test Negative**<br>($T_i = -$) | **F**alse **N**egative<br>(Type II Error) | **T**rue **N**egative<br>($-$) |

**Table 3.2. Significance testing outcomes.** $T_i$ is a test of association performed at genetic locus $i$. Under $H_0$, the null hypothesis, no association exists between locus $i$ and the phenotype. Under $H_A$, the alternative hypothesis, locus $i$ and the phenotype are in association.

**The False Discovery Rate** provides a less stringent correction for multiple testing. It may be appropriate, as many correlated tests are expected in microbial GWAS. The Benjamini-Hochberg procedure [283] controls the false discovery rate (FDR) at the significance level $\alpha_{base}$, such that if $\alpha_{base} = 0.01$, then $\leq 1\%$ of test positives should be false positives. We can define FDR, using the notation in Table 3.2, as:

$$FDR = E\left[\frac{FP}{(TP + FP)}\right] = Pr(H_0 = \text{ true } \mid \text{reject } H_0) \tag{3.16}$$

We take the following steps to control FDR:

1. Identify the per-test significance level, $\alpha$, by adjusting $\alpha_{base}$ to account for the total number of independent statistical tests performed:

$$\alpha = \frac{\alpha_{base}}{N_{score}N_{loci}} \tag{3.17}$$

2. Calculate a p-value, $P_j$, from the null distribution, for each simulated association score, $s_j$, where $j \in \{1, ..., N_{sim}\}$.
3. List $P_j$ in ascending order, using an index $k$, such that $P_{k=1}$ is the smallest p-value.
4. Find the largest $k$ such that:

$$P_k \leq k\alpha \tag{3.18}$$

5. Let the corresponding value of $s_{j=k}$ be the location of the significance threshold.

### 3.10.3 Empirical p-value estimation

The null distribution estimates the distribution of association score values expected under $H_0$, “no association”, given confounding factors in the empirical dataset. For each empirical association, we can derive a phylogenetically-correct empirical p-value from the null distribution. For score $S$ at locus $i$, we define the empirical p-value, $P_i$, as the probability of observing an association as extreme as $S_i$ by chance under $H_0$:

$$P_i = Pr(S \geq S_i|H_0) \tag{3.19}$$

**The count-based approach** is the most straightforward means of estimating empirical p-values from the null distribution. Let $S_i$ represent the association score at empirical locus $i$ and $s_j$ represent the association score at simulated locus $j$, where $i \in \{1, ..., N_{loci}\}$ and $j \in \{1, ..., N_{sim}\}$. For each association score, we define the null distribution as the histogram of all $s_j$ values. We estimate the upper tail of the null distribution to extend only up to $max(s_j)$.

Let $C_{ij}$ be a binary count indicating the relative positions of a pair of scores $s_j$ and $S_i$:

$$C_{ij} = \begin{cases} 0 & \text{if } s_j < S_i \\ 1 & \text{if } s_j \geq S_i \end{cases} \tag{3.20}$$

We define the empirical p-value $P_i$ as the proportion of $s_j$ falling at or above $S_i$.

$$P_i = \frac{1}{N_{sim}} \sum_{j=1}^{Nsim} C_{ij}, \tag{3.21}$$

For any $P_i = 0$, we state the empirical p-value as $P_i < 1/N_{sim}$, as this more accurately reflects the extent to which we can estimate the probability of spurious association.

**The kernel density approach** attempts to improve our estimation by smoothing out the shape of the null distribution that was inferred directly from the $s_j$ values. We define the null distribution as the kernel density estimate constructed from $s_j$. Kernel density estimation replaces the discrete count-based histogram with a continuous function, whose density is expressed by:

$$f(s) = \frac{1}{N_{sim}} \sum_{j=1}^{Nsim} K_h(s - s_j), \tag{3.22}$$

where $K_h$ is a Gaussian kernel function with smoothing bandwidth $h$ [284]. For empirical association score $S_i$, we define the empirical p-value $P_i$ as the area under the curve sketched by $f(s)$ that lies above $S_i$:

$$P_i = 1 - \int_0^{S_i} f(s)ds \tag{3.23}$$

Again, we state any $P_i = 0$ as $P_i < 1/N_{sim}$.

While the count-based procedure is intuitive and easy to apply, the kernel density approach may provide a more refined estimate of the null distribution. The smoothing procedure is expected to extend the estimated upper tail of the null distribution into higher association score values. If the counts of $s_j$ under-estimate the variance of the null distribution, the kernel density approach may reduce type I error. Alternatively, the density function may over-extend the upper tail of the null distribution and increase type II error, rejecting genuine associations when there is, in fact, little evidence to suggest that similar values are likely to occur by chance. Because the count-based approach does not involve further estimation procedures, it is not susceptible to this type of error. Hence, if the simulated data accurately estimates the shape and variance of the null distribution, the count-based procedure may be more reliable.

### 3.10.4 The number of simulated loci

Our choice of $N_{sim}$ will affect the null distribution and our assessment of significance. If we let $\mathbf{N_{sim} = N_{loci}}$, then for each empirical locus, we generate only one estimate of the potential association score value at such a site under $H_0$. Yet, by simulating across many sites with the same $N_{sub}$ value, we may still be able to estimate the shape and variance of the null distribution with adequate resolution with $N_{sim} = N_{loci}$.

If we let $\mathbf{N_{sim} = 10N_{loci}}$, we should achieve a more refined estimate of the null distribution. The less prevalent, more extreme chance association score values may be better represented, and the upper tail of the null distribution may be better defined. If so, this will improve our estimation of the significance threshold. We anticipate that as $N_{sim}$ increases, the accuracy of inferences made with respect to the null distribution will also increase. However, increasing $N_{sim}$ will also increase the computational cost of the simulation procedure. Past a point, diminishing returns in improved accuracy will be outweighed by computational efficiency considerations. This trade-off will also depend on the other components of the threshold-selection mechanism.

### 3.10.5 Selecting the significance threshold

In Section 4.4, we compare the threshold-selection mechanisms above on simulated data. Based on this evidence, we use the following procedures in our GWAS approach:

1. Set $\alpha_{base} = 0.01$.
2. Account for multiple testing with the Bonferroni correction.
3. Estimate $P_i$ with the count-based procedure.
4. Simulate $N_{sim} = 10N_{loci}$.

Altogether, we take the following steps to identify the significance treshold and empirical p-values, for Scores 1, 2, and 3. First, specify the overall significance level as $\alpha_{base} = 0.01$.

Then, use the Bonferroni correction to get the per-test significance level, $\alpha$:

$$\alpha = \frac{\alpha_{base}}{N_{score}N_{loci}} = \frac{0.01}{3N_{loci}}, \tag{3.24}$$

Let $C_{xj}$ record whether $s_j$ in the simulated “null” dataset falls at or below $s_x$:

$$C_{xj} = \begin{cases} 0 & \text{if } s_j > s_x \\ 1 & \text{if } s_j \leq s_x \end{cases} \tag{3.25}$$

Let $F(s)$ be the empirical cumulative distribution function describing the simulated null distribution. Then, $F(s_x)$ is the probability that the null distribution will take a value less than or equal to $s_x$:

$$F(s_x) = \frac{1}{N_{sim}} \sum_{j=1}^{Nsim} C_{xj} \tag{3.26}$$

We draw the significance threshold at the $s_x$ below $\alpha * N_{sim}$ null distribution values:

The significance threshold is drawn in the upper tail of the null distribution, at the $s_x$ below At the $s_x$ value below $\alpha * 100\%$ of the null distribution above $s_x$ $\alpha$ Solve for $s_x$ to get the location of the significance threshold:

$$F(s_x) = 1 - \alpha \tag{3.27}$$

Then, calculate the empirical p-value $P_i$ for association $S_i$, for all $i \in \{1, ..., N_{loci}\}$. Let $C_{ij}$ record whether $s_j$ in the simulated “null” dataset exceeds $S_i$, as in Equation 3.20.

$$C_{ij} = \begin{cases} 0 & \text{if } s_j < S_i \\ 1 & \text{if } s_j \geq S_i \end{cases} \tag{3.28}$$

Finally, define $P_i$ as the proportion of null distribution values that fall at or above $S_i$:

$$P_i = \frac{1}{N_{sim}} \sum_{j=1}^{Nsim} C_{ij}, \tag{3.29}$$

where $P_i < 1/N_{sim}$ for any $P_i = 0$.

## 3.11 Implementation in the treeWAS R package

The key aims of this project encompassed not only the development of a new approach to microbial GWAS, but also the implementation, improvement, and application of this method. Our next objective was thus to translate our proposed approach from theory into practice, through the implementation of an effective, efficient, and user-friendly software package. On a pragmatic note, embedding our GWAS method within a stable computational tool would be necessary for us to test, assess, and refine our approach. More importantly, however, the development of a dedicated software package could facilitate the uptake and wider application of our method. The treeWAS R package implements our tree-based microbial GWAS method in the R programming language [137]. It is freely available at `https://github.com/caitiecollins/treeWAS`. The treeWAS package aims to offer both flexibility and accessibility to users from a variety of scientific backgrounds and skill levels. In its simplest form, treeWAS requires only two arguments: a genetic dataset and a phenotype to be analysed. Supported by the broader architecture of the package, our GWAS approach can be run with one core function, *treeWAS*. However, sixteen optional arguments give users control over elements ranging from phylogenetic and ancestral reconstruction, to memory usage and output visualisation. Forty-one additional functions are implemented to execute the procedures described above, either within *treeWAS* or independently.

### 3.11.1 Presentation of results

#### Output returned

Upon completion of the association analysis, we report the following:

**Significant findings** are presented, giving the column names and sequence positions of all genetic loci identified as significantly associated with the phenotype. We first present the entire pooled set of findings identified by treeWAS, as a whole. Then, we list the three sets of significant findings, identified by Scores 1, 2, and 3 individually.

**Association statistics** are included for each association score, specifying the $S_i$ and $s_j$ score values that were calculated at each empirical and simulated genetic locus. Each empirical locus is also accompanied by an empirical p-value. The lowest estimable value of $P_i$ is noted, to represent any $P_i = 0$. For each association score, the threshold of significance is indicated. For each association score, a summary table gives the names, positions, $P_i$, and $S_i$ values of all significant loci identified, alongside the cell counts of a $g_i$x$p_i$ contingency table, if the genotype and phenotype are binary.

**The data** that was used in the analysis is provided to the user. This includes data generated by treeWAS, like the simulated dataset, the homoplasy distribution, and phylogenetic and ancestral state reconstructions. The cleaned genetic dataset and phenotype are returned in the exact form analysed. This improves transparency, and it enables reproducibility and further analysis, if desired.

### Visualisation

The high dimensionality of genome-wide data can make it difficult to interpret GWAS results. With appropriate visual summaries, however, we can dramatically improve the interpretability of our findings. The treeWAS R package implements several customisable plotting functions for this reason. In each analysis, three separate graphical representations are produced by treeWAS, unless otherwise directed.

First, we generate a plot of the estimated or input phylogenetic tree, as in Figure 3.6A. We represent the phenotype along the tree with a blue-to-red colour scale, which is binary, discrete, or continuous to match the phenotype. We indicate the phenotypic states observed at terminal nodes, as well as those inferred by treeWAS at ancestral nodes and across branches (grey = substitution). Our rapid data cleaning and reconstruction procedures allow this figure to be generated within seconds, providing users with a simple initial visual check of the data. Furthermore, it enables an examination of the phenotypic and ancestral genotypic variation in the dataset. In addition, when the analysis reports its findings, this figure will serve as a valuable reference, allowing users to trace the evolutionary origins of the significant associations identified by each measure.

Two other types of plot are produced upon completion of the analysis. One set contains Manhattan plots, showing the empirical association score values, $S_i$, for all $i \in \{1, ..., N_{loci}\}$, as measured by Scores 1, 2, and 3 (as in Figures 3.6B, D, and F). Each point represents one empirical association, $S_i$, at the sequence position indicated on the x-axis and with a score value shown along the y-axis. The colour of points serve only to visually break up the x-axis. The significance threshold is drawn in red, horizontally, at $y = s_x$. Points above this line represent significant associations. Alternatively, the *manhattan.plot* function can be used to plot the negative log-transformed p-values along the y-axis, as is commonly done in human GWAS. We find our original association score values to be more informative, especially as they can be compared to values in the null distribution. Figure 3.6 contains an illustration of the plots produced by treeWAS during a typical analysis. In real empirical analyses, these plots can show much greater complexity and may indicate linkage or reveal relationships among associated sites.

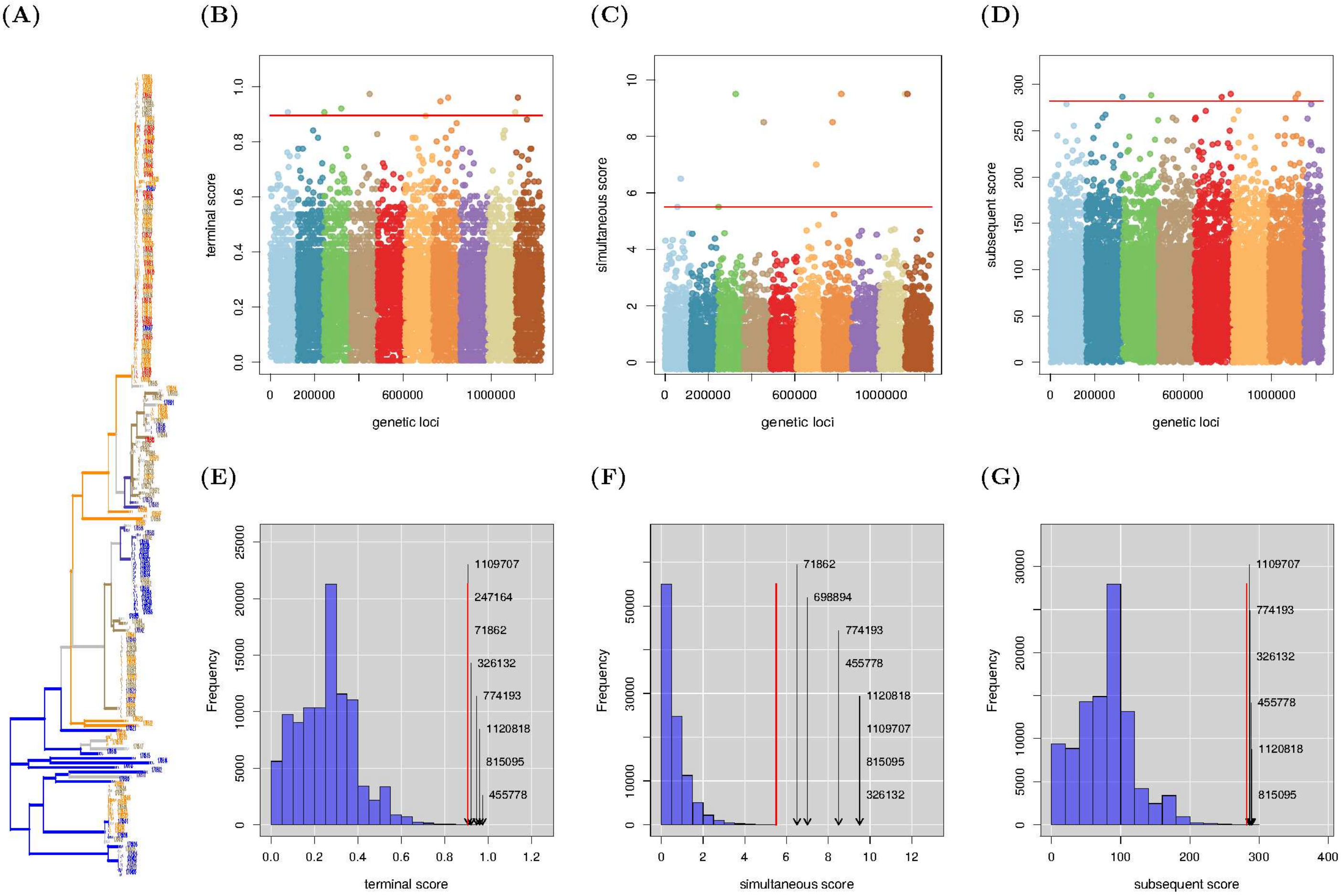


**Figure 3.6. Visualising treeWAS results.** These plots are generated by treeWAS, under default settings, to illustrate the dataset under analysis and to enable a visual interpretation of the results of the three association tests performed. These particular plots were produced by applying treeWAS to a dataset simulated with $N_{ind} = 200$, $N_{loci} = 10,000$ polymorphic sites, a non-coalescent tree, and a continuous phenotype. At left, **A:** a plot of the phylogenetic tree showing observed and inferred phenotypic states in colour (continuous spectrum: blue = minimum, red = maximum). The upper row contains Manhattan plots for **B:** Score 1, **C:** Score 2, **D:** Score 3. Each point shows the association score (y-axis) achieved at one locus (x-xis). Note that the colours of the points are meaningless and serve only to visually break up the x-axis. A horizontal red line indicates the significance threshold, and points above this line indicate significant associations. The lower row contains plots of the null distribution of association score values for **E:** Score 1, **F:** Score 2, **G:** Score 3. A vertical red line indicates the significance threshold, and significant associations are indicated by labelled arrows pointing to the significant score values attained along the x-axis.

Third, the null distribution is plotted as a histogram of the simulated association scores, $s_j$, for $j \in \{1, ..., N_{sim}\}$, for Scores 1, 2, and 3, as in Figures 3.6C, E, and G. In its upper tail, a significance threshold is drawn, at $s_x$. Above this threshold, significant empirical associations are represented with labelled arrows pointing to the x-axis at $S_i$, for each $S_i > s_x$. This plot function can also be used to visualise the empirical distribution, or to overlay the null and empirical distributions (as in Figure 4.12).

### 3.11.2 Optimisation and efficiency

The treeWAS R package provides an efficient implementation of our phylogenetic GWAS method. treeWAS can perform GWAS on a typical dataset ($N_{ind} = 100$, $N_{loci} = 100,000$) on a standard laptop computer in under two minutes (Figure 4.18). Over time, treeWAS has undergone considerable improvements in computational efficiency. Compared to previous versions, the current implementation of treeWAS achieves a four-fold reduction in the run time required for analysis and a three-fold decrease in memory demands.

These efficiency gains were attained through an iterative process of development, testing, and refinement. We reduced the memory burden of our approach by encoding all genetic variation in binary sites and storing it in logical matrices rather than numeric or character-based sequences. By excluding fixed loci, following phylogenetic reconstruction, we were able to further decrease the time and memory required to run treeWAS, without impacting the power or performance of our GWAS approach (as $Pr(H_0 = true|MAF = 0) = 0$). At the same time, we were able to achieve efficiency without having to eliminate synonymous SNPs or SNPs in non-coding regions. A growing body of evidence indicates that, by removing these sites, GWAS methods may be overlooking genuine functional relationships, for example, in the regulation of gene expression [285–288].

Additional streamlining was achieved at polymorphic sites. Instead of repeating all procedures at every empirical locus $i$ and simulated locus $j$, for *all* $i \in \{1, ..., N_{loci}\}$ and $j \in \{1, ..., N_{sim}\}$, we found we could restrict most procedures to the subset of unique sites only. To ensure accurate estimation of the null distribution, simulation is independently performed at all $N_{sim}$ loci. But, we reconstruct ancestral states and compute association scores in the empirical and simulated datasets at unique sites only, as these calculations will be identical for duplicate patterns. We retain one representative of each unique column pattern $g_i \in \{0, 1\}^{N_{ind}}$, but any identical site $g_j$ ($D_{ij} = 0$) is indexed and set aside. Before using the simulated null distribution to make any inferences about the significance of empirical associations, we expand all calculations according to the index, to preserve the original distribution of column patterns. This procedure reduces redundancy by eliminating potentially large proportions of the $N_{loci}$ sites to be analysed, which consistently and often dramatically improves the efficiency of treeWAS.

Line-by-line profiling allowed us to identify and address lingering redundancies and bottlenecks in the code. Any loops implemented over the $N_{loci}$ columns, for example, were reoriented along a smaller dimension (e.g., $N_{ind}$, $N_{branch}$) or, wherever possible, they were reconfigured to perform all procedures simultaneously. For example, our approach to data simulation is described in Box 3.1 as a step-wise procedure, repeated $N_{sim}$ times. In practice, these steps are performed only once, and all $N_{sim}$ loci are independently but simultaneously simulated. Likewise, the $N_{sub}$ substitutions to occur at each simulated site are not drawn $N_{sim}$ times; instead, a more efficient implementation randomly draws $N_{sim}$ samples of $N_{sub}$ together, in a single step. Efforts to optimise treeWAS methodology, in terms of its sensitivity and specificity, were also pursued through comparative applications to simulated datasets (see Chapter 4).

With large microbial genome-wide datasets, some computers may still find it challenging to carry out our simulation-based approach to GWAS, despite the improvements in efficiency offered by the current version of treeWAS. Insufficient available memory is most often the limiting factor in these conditions. Memory constraints can, however, be overcome if the initial large volumes of sequence data can be broken down and analysed in more manageable fragments. In a genetic dataset with $N_{ind}$ rows and $N_{loci}$ columns, we consider each sequence to be a concatenation of smaller chunks of sequence, $C_i$, where $i \in \{1, ..., N_{chunk}\}$ and $\sum_{i=1}^{N_{chunk}} C_i = N_{loci}$. Chunk size, $C_i$, is the same for all $i$, with the possible exception of $C_{i=N_{chunk}}$. The optimal value of $C_i$ depends on $N_{loci}$, $N_{sim}$, the number of unique column patterns, and the memory available on the computer at the time of analysis. Our phylogenetic GWAS procedure can be carried out across any number of chunks, as we propose in Box 3.2.

1. Select $C_i$, the size of each chunk, in one of two ways:
   (a) Specify $C_i$ to be an integer between 1 and $N_{loci}$ (e.g., $C_i = N_{loci}/2$).
   (b) Set treeWAS argument *mem.lim* = TRUE to determine the maximum $C_i$ possible without breaching memory limits.
2. Perform phylogenetic and reconstruction and data cleaning on all input sites.
3. Reconstruct ancestral states at all $N_{loci}$ polymorphic sites and identify the homoplasy distribution.
4. Get $N_{chunk}$ contiguous subsets of the $N_{loci}$ sites, s.t. chunk 1 spans loci $g_{i \in \{1,...,C_i\}}$.
5. For chunk $i$, where $i \in \{1, ..., N_{chunk}\}$:
   (a) Define $N_{sim}$ for chunk $i$ as $N_{sim(i)} = C_i * N_{sim}/N_{loci}$.
   (b) Simulate $N_{sim(i)}$ sites along the tree, drawing $N_{sub}$ from the genome-wide homoplasy distribution.
   (c) Reconstruct ancestral states at each simulated locus.
   (d) Measure associations at each epirical and simulated locus, storing $S_i$ and $s_j$ for all $i \in \{1, ..., C_i\}$ and $j \in \{1, ..., N_{sim(i)}\}$, for Scores 1,2, and 3.
   (e) Remove any data that was used in chunk $i$ but is no longer needed.
6. Repeat Step 5 for all chunks, until association scores $S_i$ and $s_j$ have been calculated for all $i \in \{1, ..., N_{loci}\}$ and $j \in \{1, ..., N_{sim}\}$.
7. Using all $s_j$ to estimate the null distribution, follow the procedure in Section 3.10.5 to identify the significance threshold and significant values of $S_i$ for each measure.

**Box 3.2. Chunk-by-chunk procedure.**

The chunk-by-chunk procedure provides a valuable alternative implementation of our GWAS method. Naturally, computational time increases as a function of $N_{chunk}$. Hence, where memory limits are not restrictive, the default behaviour of treeWAS prioritises efficiency in computational time by setting $C_i = N_{loci}$ and following the standard procedure. However, on machines with insufficient memory, our ability to control the trade-off between time and memory requirements allows treeWAS to escape the constraints imposed by prohibitive memory limits. Time permitting, therefore, treeWAS should be able to analyse datasets of any size on almost any computer.

### 3.11.3 Accessibility and user resources

To ensure that users get the most out of the treeWAS R package, we provide detailed examples, tutorials, and documentation. Each function in treeWAS is accompanied by a thorough description of its purpose, arguments, and output. Useful information is also printed out during the execution of key functions, to inform the user of changes made during data cleaning, to provide updates on processes underway, and to offer suggestions in case any argument is contraindicated by the data under analysis. Worked examples and sample data are included to illustrate useful applications of treeWAS functions, and to provide a practical demonstration for users unfamiliar with expected inputs and outputs, argument usage and data formatting. A more extensive tutorial is presented in vignettes, available within treeWAS and in our online Wiki. Users can follow this documentation through each stage of the analytical process, from the installation of treeWAS, to data cleaning, conversion, and integration with ClonalFrameML, through the treeWAS association testing pipeline, to the visualisation of output and the interpretation of results. We encourage users to interact with us on our online forum, where we provide detailed explanations and implement new features in response to user questions and requests. The treeWAS R package has been released under version $\geq 3$ of the GNU General Public License. All code and documentation in the treeWAS R package can be viewed online at `https://github.com/caitiecollins/treeWAS/issues`. We are pleased that, in addition to meeting our fundamental aims of efficacy and efficiency, we have been able to implement our GWAS method in software that is open-source, user-friendly, and freely available to the public.

# Chapter 4

# Application to Simulated Data

**Acknowledgements** This chapter includes some written and illustrated content that appeared, in a related form, in the publication of our method [155]. Caitlin Collins and Xavier Didelot designed the structure of the simulation study. Collins developed and implemented the simulation processes, performed the analyses, generated the figures, and produced the report. Didelot contributed to the development of the simulation processes for generating associations. Didelot also provided useful feedback throughout the writing and editing process, and his input is reflected in our publication and in the chapter below. We are grateful to Daniel Wilson and David Aanensen for providing useful feedback during the development of this work. We also thank the two anonymous reviewers who provided helpful suggestions that we incorporated in our paper. Finally, we would like to extend our gratitude to the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project, for generously providing us with access to their computing cluster.

## 4.1 Simulation study

In this section, we apply our GWAS method to a large number of simulated datasets and we examine how its performance varies. Our motivation for performing this simulation study is two-fold. First, we use analyses of simulated data to guide the development of our GWAS method. We make isolated changes to our approach and, by comparing the resulting performance in a controlled, simulated setting, we are able to make evidence-based decisions about which methodological choices are likely to produce more accurate and reliable results in real analyses. Second, once we have optimised our approach and settled upon a stable version of our GWAS method, we use analyses of simulated data to evaluate the performance of our method and to compare it to alternative approaches.

We apply treeWAS and six comparator methods to over 600 unique synthetic datasets. Our protocols for simulating genotypic, phenotypic, and associative data are described below and implemeted in the treeWAS R package. We evaluate the performance of our method, we describe how it changes as we vary parameters of the simulated datasets, and we compare its performance to that of existing GWAS methods. Altogether, through applications to simulated data, we aim to refine our GWAS method and to provide an assessment of its capacity to identify associations across a diverse array of datasets.

### 4.1.1 Honing methodology

We first apply our GWAS approach to simulated datasets as a means of improving our developing method. We observe how performance varies when different methodological choices are made, presenting a detailed examination of these components:

- **Ancestral state reconstruction method:** parsimony or maximum-likelihood.
- **Score 3 calculation:** including or excluding branch length (Eqn 3.13 or 3.14)
- **Significance threshold selection:** combinations of (i) base p-value, (ii) multiple testing correction, (iii) p-value estimation, (iv) number of simulated sites.
- **Association testing:** single score or multiple scores, in parallel or sequentially.

In each case, we use the evidence generated by the simulation study to identify the optimal approach, which we then either permanently encode in the treeWAS R package or establish as the recommended option for users of our GWAS method.

### 4.1.2 Parameters explored

We then use this simulation study to evaluate the performance our method. We assess the performance of our method, as a whole and in terms of each association score, and we compare this to the performance of alternative GWAS methods. Simulated data allows us, for example, to see how the sensitivity and specificity of GWAS methods vary as a function of the relative strength of simulated associations and confounding factors. We assess how performance varies as a function of the following parameters:

- **Simulation framework:** Sets A, B, and C.
- **Recombination rate:** 0, 0.01, 0.05, 0.1.
- **Number of individuals:** 50 to 200.
- **Number of genetic loci:** 10,000 to 100,000.
- **Accessory genome sim.:** $N_{ind} = 100$, $N_{loci} = 5,000$, $R = 0.2$.

Unless otherwise indicated, the following parameters remain fixed throughout the study. We set $N_{ind} = 100$, $N_{loci} = 10,000$, $N_{assoc} = 10$, phenotypic $N_{sub} \sim Pois(15)$, we require the frequency of the minor phenotype to be $\geq 25\%$, and we simulate indiviuals along coalescent phylogenetic trees.

### 4.1.3 Simulating non-associated loci

The vast majority of sites in our simulated genetic sequences are not associated with the phenotype. Genetic variation at these loci is characterised by the ancestral relationships between individuals and shaped by mutation and recombination. Before simulating the 10 trait-associated sites, we simulate these 9,990 background loci as follows:

1. Define the clonal genealogy linking isolates by simulating a binary, ultrametric, coalescent tree with $N_{ind} = 100$ terminal nodes.
2. Define the homoplasy distribution by using SimBac to estimate $N_{sub}$ for particular rates of mutation and recombination (see below, Figure 4.1).
3. Simulate neutral evolution along the tree according to the procedure described in Box 3.1, such that mutation and recombination events at each site occur at a constant rate across the tree.

### 4.1.4 Simulating recombination

To assess performance as a function of the recombination rate among non-associated loci, we simulate datasets at four values of $R$, varying by an order of magnitude. We used SimBac [289], software specifically designed to simulate the effect of homologous recombination on bacterial evolution, to estimate the effect of recombination on the homoplasy distribution. We simulated four genetic datasets with SimBac, setting the recombination rate parameter, $-R$, to 0, 0.01, 0.05, and 0.1, where these site-specific rates specify $E[N_{sub}]$ across the evolutionary history of simulated genomes due to within-species homologous recombination. We selected these $-R$ values to simulate the range of recombination rates that we expect to encounter in bacterial association studies performed in organisms for which a clonal genealogy can be inferred [221].

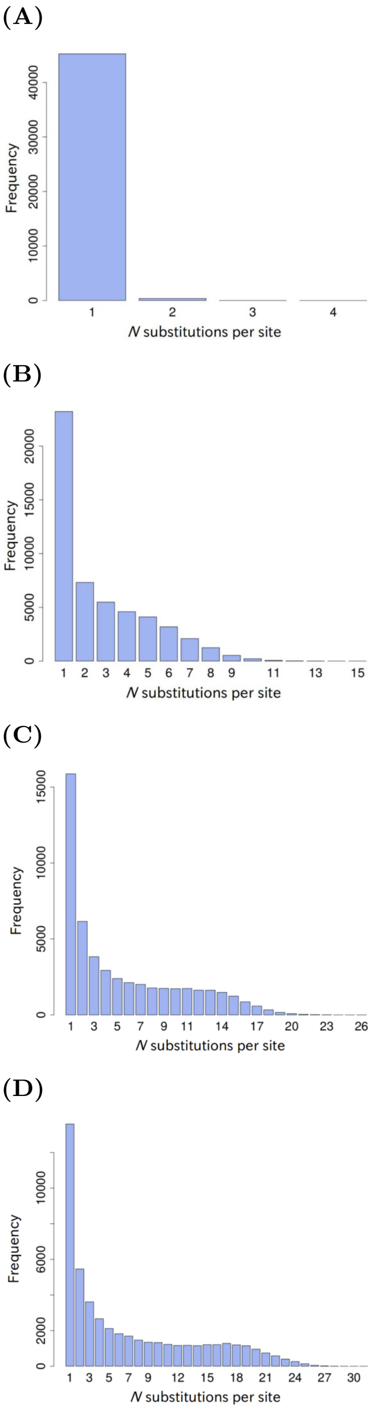


**Figure 4.1. SimBac homoplasy distributions by recombination rate.** $x = N_{sub}$, $y =$ Frequency. **A:** $R = 0$ **B:** $R = 0.01$ **C:** $R = 0.05$ **D:** $R = 0.1$.

We fixed all other SimBac parameter values. In each case, we simulated $-N = 100$ individuals, and a large number of loci in each dataset, specifying $-B = 1,000,000$ sites with no gaps between them ($-G = 0$). We set the average length of a within-species recombinant interval to $-D = 500$. We specified no between-species recombination ($-r = 0$), with no variation ($-m = 0, -M = 0$), and we set the site-specific mutation rate to $-T = 0.01$. At the four within-species recombination rates examined, $r/m$ $= 0$, 1, 5, and 10, ranging from ratios observed in clonal *L. interrogans* [74] and *M. tuberculosis* [290] to those of recombinant *C. jejuni* [157] and *N. meningitidis* [74].

We estimated the homoplasy distribution of each SimBac simulated dataset. First, we inferred the clonal genealogy, using ClonalFrameML [221] to account for recombination, with an initial tree reconstructed with the dnapars algorithm in PHYLIP [194] We used the Fitch parsimony algorithm [255] to calculate the minimum $N_{sub}$ per site. This allowed us to identify homoplasy distributions characteristic of the four recombination rates we wished to investigate (see Figure 4.1). We use these homoplasy distributions to simulate recombination among the 9,990 non-associated loci in each simulated dataset. The three simulation sets described below are each used to simulate 80 datasets for performance testing, with 20/80 datasets simulated at each of the four recombination rates above.

### 4.1.5 Simulating trait-associated loci

To test the performance of our GWAS method, we devised three different protocols for simulating phenotypically-associated genetic variables. The synthetic datasets analysed in this study are primarily grouped into three main panels, termed Set A, B, and C, which are defined by the simulation framework used to generate the trait-associated columns in each genetic dataset. Each set operates on a different definition of “association” and takes a unique approach to simulate the relationship between the $N_{assoc} = 10$ associated loci and the phenotype. We analyse 240 simulated datasets, comprising 80 datasets from each set, in the comparative performance evaluation at the core of this simulation study. An additional 394 unique datasets are simulated under the sophisticated Set C framework to facilitate a sensitivity analysis of performance and run time variation.

### 4.1.6 Set A

Set A was designed with the same conception of association as Score 1, in that associations are defined and quantified primarily at the terminal nodes of the tree. In Set A, the ten genotype-phenotype associations are generated in a three-step procedure:

1. Simulate the phenotype.
   (a) Let phenotypic $N_{sub}$ $Pois(\lambda = 15)$, such that $E[N_{sub}] = 15$.

   (b) Sample one value from $Pois(15)$ to determine how many phenotypic $N_{sub}$ will be simulated along the tree.

   (c) Distribute the $N_{sub}$ phenotypic substitutions along the tree by sampling $N_{sub}$ branches such that $Pr(branch_j) \propto l_j$.

   (d) Select $p^{root}$ by sampling one value from $U(0, 1)$.

   (e) Determine the state of all nodes. From root to tip, set $p_j^{des} = p_j^{anc}$, unless $branch_j$ contains a substitution, in which case, $p_j^{des} = (1 - p_j^{anc})$.

2. Generate perfect association.
   (a) Select the sequence positions of the 10 associated loci by sampling 10 values from $U(1, 10000)$.

   (b) Define the genotype at each associated locus $i$ on branch $i$ as $g_{i,j} = p_j$, producing perfect correlation with the phenotype at all nodes.

3. "Dilute" association at terminal nodes.
   (a) Specify dilution factor $\delta = 0.1$ and randomly sample $\delta * N_{ind}$ terminal nodes, for each associated locus.

   (b) At each associated locus, dilute the association by redefining sampled nodes $g_d^{term}$ as $(1 - g_d^{term})$.

This procedure weakens the initially-perfect association simulated between the ten genetic variables and the phenotype. Note that setting $\delta = 0.1$ will give Score 1 = 0.8 (= (90 - 10)/100) at each associated locus in Set A. Whereas, because nodes $d$ will differ among the ten associated loci, Scores 2 and 3 will vary.

### 4.1.7 Set B

Set B was designed to test, in particular, the potential of Score 2 to identify associated genetic loci that give rise to the phenotype through two non-overlapping complementary pathways. In Set B, complementary associations between each of the ten genetic loci and the phenotype are created as follows:

1. Simulate phenotypic states $p_j$ as in Set A, Step 1.
2. Generate perfect association at sites $g^{assoc}_{i\in[1,10]}$, as in Set A, Step 2, s.t. $g_{i,j} = p_j$.
3. Create two complementary pathways and assign five associations to each.
   (a) Bisect the phylogeny into $K = 2$ major clades by identifying two subtrees.
   (b) If $N_k < \frac{1}{3}N_{ind}$ in either clade $k = 1$ or $k = 2$, transfer sub-clades from the larger to the smaller major clade until $\frac{1}{3}N_{ind} < N_k < \frac{2}{3}N_{ind}$.
   (c) To generate complementarity, maintain perfect association at $g^{assoc}_{i\in\{1,\ldots,5\}}$ in one subtree but set $g^{assoc}_{i\in\{1,\ldots,5\}} = 0$ in all genomes in the other subtree. Repeat, with the opposite subtrees, for $g^{assoc}_{i\in\{6,\ldots,10\}}$.

Our purpose in generating these strong associations in subtrees alone is to test the ability of Score 2 to detect associations that give rise to the phenotype through complementary pathways. We do not expect that any of the other tests of association, in treeWAS or competing approaches, will perform particularly well in this simulation set.

### 4.1.8 Set C

Set C was designed to generate the most complex and subtle associations of our three simulation sets, through a simulation process that more closely recapitulates genuine evolutionary processes. Similar to the conceptual framework adopted in Score 3, Set C conceives of associations as probabilistic relationships in a constant state of flux across the phylogenetic tree. In Sets A and B, a pre-determined number of phenotypic and genotypic substitutions are assigned to the branches of the tree, with perfect associations generated and then subsequently modified. By contrast, in Set C, the processes of substitution and association at the ten associated loci are stochastically generated, according to an instantaneous transition rate matrix, $Q$, in a time reversible Markov chain.

In Set C, association is simulated as follows:

1. Let $Q$ control the rate of transition between all four possible combinations of a binary genotype, $g_i$, and phenotype, $p_i$ across branch $i$ of the tree.

2. Specify $Q$ with two parameters: $s$, the baseline substitution rate, which applies to all columns of $Q$; and $a$, the association factor, which encodes the preference for positive association $\{(0,0),(1,1)\}$ over negative $\{(0,1),(1,0)\}$.

3. Let $Q$ be a matrix whose cells $Q_{ij}$ specify the instantaneous rate of transition between the ancestral genotypic and phenotypic states $(g^{anc}, p^{anc})$ in row $i$ and descedant states $(g^{des}, p^{des})$ in column $j$. Because we assume that transitions do not occur instantaneously in both variables, let $Q_{ij} = 0$ along the antidiagonal. Along the main diagonal, set $Q_{ij}$ such that $\sum_{j=1}^{4} Q_{ij} = 0$ in each row.

$$Q = (g^{anc}, p^{anc}) \text{ x } (g^{des}, p^{des}) = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 1,0 \\ 1,1 \end{array} \begin{array}{c} \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \begin{pmatrix} -2s & s & s & 0 \\ sa & -2sa & 0 & sa \\ sa & 0 & -2sa & sa \\ 0 & s & s & -2s \end{pmatrix} \end{array} \tag{4.1}$$

4. Parameterise $Q$ to create a dependent relationship between genotype and phenotype at associated sites.

   (a) Set $s = 20$ and $a = 10$ to get $E[N_{sub}] \simeq 15$ for the phenotype. This will generate moderate population stratification in a sample of $N_{ind} = 100$ clonally-related individuals, because phenotypic states will tend to cluster along ancestral lines.

   (b) To account for the total branch length of the tree, divide $s$ by $\sum_{i=1}^{N_{branch}} l_i$.

   With a tree whose branch lengths sum to 8.48, setting $s = 2.35$ $(= 20/8.48)$ and $a = 10$ gives:

$$Q = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 1,0 \\ 1,1 \end{array} \begin{array}{c} \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \begin{pmatrix} -4.717 & 2.358 & 2.358 & 0.000 \\ 23.585 & -47.169 & 0.000 & 23.585 \\ 23.585 & 0.000 & -47.169 & 23.585 \\ 0.000 & 2.358 & 2.358 & -4.717 \end{pmatrix} \end{array} \tag{4.2}$$

5. Convert the stationary rate matrix, $Q$, into a matrix of probabilities, $P$. Define the $P$ matrix as $(g^{anc}, p^{anc})$ x $(g^{des}, p^{des})$. Let the cells of $P_i$ specify $Pr(g_i^{des}, p_i^{des} | g_i^{anc}, p_i^{anc})$ for a branch of length $l_i$.

6. Create the $P$ matrix for branch $i$ by using matrix exponentiation to account for the length, $l_i$, of the branch in question.

$$P_i = exp(Ql_i) \tag{4.3}$$

Shorter branches reduce the probability of transition, favouring the initial states of both genotype and phenotype. For example, if $l_i = 0.0001$, $P_i$ is calculated as:

$$P_i = \begin{array}{c|cccc} & 0,0 & 0,1 & 1,0 & 1,1 \\ \hline 0,0 & 0.999 & 0.000 & 0.000 & 0.000 \\ 0,1 & 0.004 & 0.993 & 0.000 & 0.004 \\ 1,0 & 0.004 & 0.000 & 0.993 & 0.004 \\ 1,1 & 0.000 & 0.000 & 0.000 & 0.999 \end{array} \tag{4.4}$$

Longer branches favour transition, increasing the probability of substitution in one or both of the genotype and phenotype, depending on their initial states and the relative preference for association established in $Q$. For example, if $l_i = 1$, $P_i$ is calculated as:

$$P_i = \begin{array}{c|cccc} & 0,0 & 0,1 & 1,0 & 1,1 \\ \hline 0,0 & 0.459 & 0.045 & 0.045 & 0.450 \\ 0,1 & 0.455 & 0.045 & 0.045 & 0.455 \\ 1,0 & 0.455 & 0.045 & 0.045 & 0.455 \\ 1,1 & 0.450 & 0.045 & 0.045 & 0.459 \end{array} \tag{4.5}$$

7. Progressing from root to tips, define $P_i$ for each branch $i \in \{1, ..., N_{branch}\}$, and draw the genotypic and phenotypic states at the descendant node, letting $\Pr(g_i^{des}$, $p_i^{des})$ be a function of $(g_i^{anc}$, $p_i^{anc})$ and $l_i$.

   (a) Because we simulate one phenotype but ten associated loci, we simulate all genotypes simultaneously. Using the $P$ matrix to guide the selection of the state of $p_i^{des}$, we draw the descendant states of all ten associated loci simultaneously from the probability distribution of its possible states, conditional on the state of $p_i^{des}$.

8. Repeat this process of transition across all branches $i \in \{1, ..., N_{branch}\}$, until genotypic and phenotypic states have been selected at all nodes, from root to tips, for all trait-associated sites.

### 4.1.9 Comparator GWAS methods

To benchmark the performance of our approach, we carried out a comparative analysis of GWAS approaches, by applying multiple methods to the same simulated datasets. In addition to our own phylogenetic GWAS method, we also applied six alternative approaches to each of the 240 datasets simulated. We aim to compare the performance of different GWAS methods under the three simulation set frameworks for generating associated loci and while varying the parameters controlling non-associated loci. Table 4.1 presents, for our approach and the six comparator GWAS methods, which tests of association are used and which corrective measures are taken to counteract confounding population structure. A detailed description of these alternative GWAS approaches can be found in Chapter 2.

We use the Fisher's exact test, and the $X^2$ test available in PLINK as benchmarks, to demonstrate what results would be found by two of the most popular tests of association when no correction for population stratification was applied [136, 291]. The PLINK $X^2$ test with Genomic Control (GC) provides the simplest solution to population stratification. As high $\lambda_{GC}$ values are likely to be calculated for our simulated clonal populations, we expect that the uniform correction by $\lambda_{GC}$ may reduce the power of GC [136].

We include two multivariate approaches, Principal Components Analysis (PCA) and the Discriminant Analysis of Principal Components (DAPC). These approaches summarise the variation contained in a genetic dataset by identifying synthetic PC dimensions that represent major axes of overall (within- and between-group) variation, in PCA, and that maximise between-group variation in DAPC [110, 145]. PCA is the method most commonly used to correct for population stratification in human GWAS [107, 174] and a prevalent approach in microbial GWAS [116,119,122,125]. DAPC has more recently been proposed as a potential improvement on PCA, and it has also been applied in bacterial GWAS [80, 145]. We followed the protocol used in human genetics and corrected for ancestry by regressing along the significant PCs of PCA or DAPC, continuously adjusting the genetic data by the amounts attributable to ancestry, according to the method in question [107]. We then identified significant associations via $X^2$ test. Both PCA and DAPC were implemented in the R programming language [137], using functions encoded in the *adegenet* R package [292]. Our implementation of PCA is similar to the sequence element enrichment analysis (SEER) method of Lees *et al.* [129]. As in SEER, we use regression to model the phenotype as a function of both the genotype and the set of significant PCs, incorporating these descriptor variables as fixed effects. However, as we decribe below, we use additional optimisation procedures to select the number of PC axes retained in each application of both PCA and DAPC in our comparative analysis of

GWAS methods.

The sixth and final comparator GWAS method included in our simulation study is the Cochran-Mantel-Haenszel (CMH) test. The CMH test implements a cluster-based control for population structure by stratifying the association test along $k$ population clusters within a 2x2x$k$ design [144]. The CMH test is among the most popular microbial GWAS methods [49, 124, 126]. We implemented the CMH test with functions from the *stats* R package in the R software [137]. The number of clusters, $k$, was objectively selected through the k-means clustering procedure described below [292, 293].

| Method | Association Test | Population Stratification Correction | Reference |
|---|---|---|---|
| **Fisher** | Fisher's exact test | None. | [291] |
| **PLINK** | $X^2$ test | None. | [136] |
| **GC** | $X^2$ test | Adjusts all association statistics by a factor, $\lambda_{GC}$, that quantifies overall inflation due to population stratification. | [136] |
| **PCA** | Analysis of Variance (ANOVA) | Corrects the genetic data matrix by regressing along the significant PCs of PCA. | [110] [294] |
| **DAPC** | Analysis of Variance (ANOVA) | Corrects the genetic data matrix by regressing along the significant PCs of DAPC. | [145] [294] |
| **CMH** | Stratified $X^2$ test | Stratifies the association test by population. | [144] |
| **treeWAS** | Scores 1, 2, 3 | Compares associations to a null distribution generated by simulating genetic data along the tree. | [295] |

**Table 4.1. GWAS methods compared.**

### Choosing the number of clusters and PC axes:

To use the CMH test, PCA, or DAPC to correct for confounding population structure in microbial GWAS, one must first determine how many clusters or PC dimensions will be used to represent ancestral populations or significant axes of ancestral variation, before these can be used to stratify the analysis or included as fixed effects in the regression model. How best to make this decision remains a topic of debate, with many proposed solutions [170, 172, 184]. To eliminate this potential source of subjective bias and variation from our comparative analysis of GWAS methods, we chose to use one objective approach, k-means clustering, as the basis for selecting the number of significant clusters or dimensions for all three approaches.

We use the *find.clusters* algorithm implemented in the *adegenet* R package to apply k-means clustering to the PCA-transformed genetic data [292, 293]. The optimal $k$ to describe the population structure of a dataset is inferred by identifying the value of $k$ associated with the lowest BIC, as shown in Figure 4.2. When performing GWAS with the CMH test, we stratify the association analysis according to this configuration of $k$ population clusters. In the PCA analysis, we correct the genetic data by regressing along $(k-1)$ significant PCs, the number of dimensions required to separate the $k$ populations. Likewise, we set $(k-1)$ PCs as the number of discriminant functions to be retained in the DA component of the DAPC analysis.

We carry out a separate procedure to objectively estimate the optimal number of PCs to be retained in the initial PCA step of the DAPC analysis. We perform this optimisation procedure via stratified cross-validation, using our own implementation of the procedure in the *xvalDapc* algorithm, which is available in the *adegenet* R package [292]. At each level of PC retention, from 5 to 45 PCs, DAPC is performed on 30 different "training sets", comprising stratified random samples of 90% of the data from each cluster. The results of each training set analysis are used to predict the cluster memberships of individuals in the remaining 10% test set, and predictive success is plotted as a function of the number of PCs, as in Figure 4.3. The number of PCs associated with the minimum mean squared error in prediction, and usually also the maximum mean predictive success, is inferred to be the optimal number of dimensions to retain in the PCA step of DAPC. By using this estimate to set the level of PC retention in the final DAPC analysis, we ensure that the best fitting model of population structure is selected.

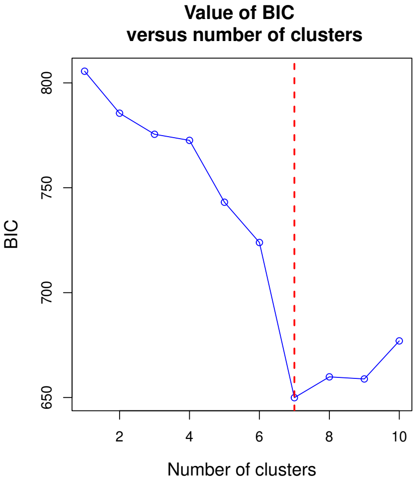


**Figure 4.2. K-means clustering.** Example output, plotting BIC values by number of clusters, $k$. The red line at $min(BIC)$ indicates the optimal value of $k = 7$.

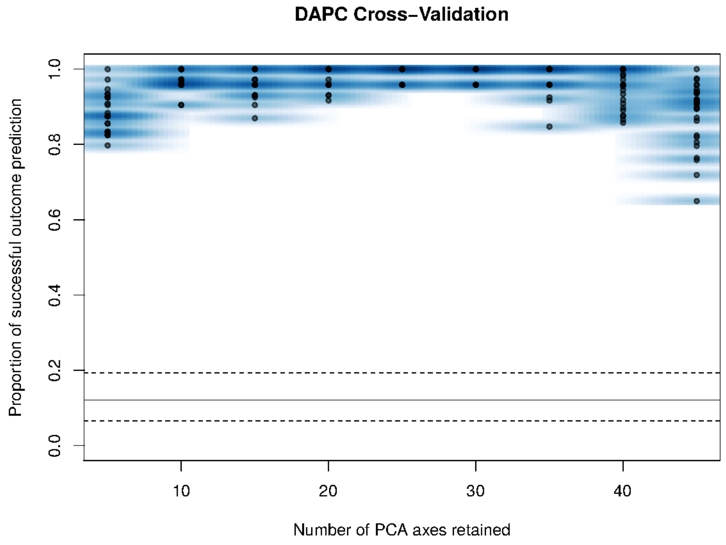


**Figure 4.3. DAPC cross-validation.** The output of cross-validation is plotted, showing proportion of individuals whose population membership was correctly identified on the y-axis, as the number of PCs increases along the x-axis. The x-axis value associated with the minimum mean squared error in cluster membership prediction, here at 25 PCs, indicates the optimal number of PCs to be retained in DAPC.

## 4.1.10 Assessing performance

To achieve a balanced appraisal of the performance of our approach and each association test, we evaluate performance along four metrics, which are defined in Table 4.2. Owing to the large class imbalance between the number of associated and non-associated loci (10 vs. 9,990) in our simulated datasets, we include PPV as a practical complement to FPR. A numerically "low" FPR value can obscure an unacceptably high number of false positives, whereas PPV values provide a more intuitive assessment of performance that is most relevant to the user, quantifying the proportion of "significant" findings are truly associated. In evaluating performance, we therefore focus primarily on sensitivity and PPV. The composite F1 score [296] is the harmonic mean of sensitivity and PPV. It summarises both metrics, while tending towards the lower of the two. The F1 score therefore provides a useful, conservative estimate of overall performance within a single measure that we can use to evaluate and compare the overall performance of GWAS methods.

| |
|---|
| **FPR** |
| (False Positive Rate) |
| $\frac{(FP)}{(FP+TN)} = (1 - specificity)$ (4.6) |
| $Pr(T_i = + \mid H_0)$ |
| **Sensitivity** |
| $\frac{(TP)}{(TP+FN)}$ (4.7) |
| $Pr(T_i = + \mid H_A)$ |
| **PPV** |
| (Positive Predictive Value) |
| $\frac{(TP)}{(TP+FP)}$ (4.8) |
| $Pr(H_A \mid T_i = +)$ |
| **F1.score** |
| (F1 Score) |
| $2 * \frac{(PPV * sensitivity)}{(PPV + sensitivity)}$ (4.9) |
| *Composite measure of* ***overall*** *performance* |

**Table 4.2. Metrics of evaluation.**

## 4.2 Optimising our approach

Our first purpose for undertaking this simulation study was to assess and improve upon our analytical approach. We simulated data as described above and analysed these datasets with two or more variants of our microbial GWAS method. Below, we present the results of these applications to simulated data, which allowed us to make evidence-based decisions about how we could most effectively reconstruct ancestral states, calculate Score 3, select the threshold of significance, and peform association testing.

## 4.3 Ancestral state reconstruction

Within our GWAS approach, we reconstruct the ancestral states of the phenotype and the genotype, at each locus in both the empirical and simulated datasets, and we infer the locations of substitutions across the tree. Ancestral states can be inferred by both parsimony-based and ML reconstruction methods. As explained in Chapter 2 (see Box 2.9 and 2.10), parsimony and ML methods operate on distinct principles and make different assumptions. The estimates made by both methods may agree when the genuine evolutionary history of a trait is defined by a small number of unambiguous transitions betwen states [258]. Yet, when faced with more complex patterns of ancestral substitution, parsimonious and ML reconstructions often deviate from one another, as uncertainty and inaccuracies increase in one or both approaches. Our implementation of these two methods also differs, as described in Chapter 3 With the parsimonious reconstructions, we perform association testing on the point estimates of the ancestral states. By contrast, with the ML reconstructions, we work directly with the marginal likelihoods of binary states, so that the uncertainty quantified by the ML method can be incorporated into our association test.

To determine objectively which method of ancestral state reconstruction would improve the performance of our approach, we applied our GWAS method twice, to each simulated dataset ($N = 240$). In the first instance, we performed a parsimonious reconstruction of the ancestral states of the genotype and phenotype. In the second, we performed a ML reconstruction and let the ancestral states of both variables be defined by the marginal likelihood of binary state zero.

For both analyses of each simulated dataset, we calculated the four performance metrics in Table 4.2, generating matched pairs of each statistic. We then performed a two-sample Wilcoxon rank sum test to assess the change across these matched pairs. The results of this non-parametric test are presented below, stratified by simulation set, in Tables 4.3, 4.4, and 4.5. Note that results are not included for Score 1 as the reconstructed ancestral states do not inform its calculation or performance. The sign of the median difference indicates which reconstruction method improves performance along metric $x$, such that:

$$Median(x) \begin{cases} < 0 & \text{ML supersedes parsimony along } x \\ > 0 & \text{parsimony supersedes ML along } x \end{cases} \tag{4.10}$$

Rows containing a statistically significant difference ($p < 0.05$) are highlighted in yellow.

| | | | | **Δ (Parsimony – ML)** | | |
|---|---|---|---|---|---|---|
| | **Association Score** | **Statistic** | **P-value** | **$\text{C.I.}_{\text{Lower}}$** | **Median** | **$\text{C.I.}_{\text{Upper}}$** |
| 1 | Score 2 | F1.score | 0.0102 | -0.0819 | -0.0527 | -0.0125 |
| 2 | Score 2 | PPV | 0.0454 | -0.1458 | -0.0909 | -0.0001 |
| 3 | Score 2 | sensitivity | 0.0237 | -0.1500 | -0.1000 | 0.0000 |
| 4 | Score 2 | FPR | 0.3045 | -0.0001 | 0.0001 | 0.0001 |
| 5 | Score 3 | F1.score | 0.0000 | 0.3766 | 0.4616 | 0.5641 |
| 6 | Score 3 | PPV | 0.0340 | 0.0101 | 0.3920 | 0.7046 |
| 7 | Score 3 | sensitivity | 0.0000 | 0.3000 | 0.4000 | 0.5001 |
| 8 | Score 3 | FPR | 0.6078 | -0.0001 | 0.0000 | 0.0001 |
| 9 | treeWAS (all) | F1.score | 0.0533 | -0.0793 | -0.0455 | 0.0000 |
| 10 | treeWAS (all) | PPV | 0.0519 | -0.0974 | -0.0520 | -0.0001 |
| 11 | treeWAS (all) | sensitivity | 0.1486 | -0.1500 | -0.1000 | 0.0499 |
| 12 | treeWAS (all) | FPR | 0.0242 | 0.0000 | 0.0000 | 0.0001 |

**Table 4.3. Wilcoxon Test: Reconstruction Method (Set A).**

| | | | | **Δ (Parsimony – ML)** | | |
|---|---|---|---|---|---|---|
| | **Association Score** | **Statistic** | **P-value** | **$\text{C.I.}_{\text{Lower}}$** | **Median** | **$\text{C.I.}_{\text{Upper}}$** |
| 1 | Score 2 | F1.score | 0.1510 | -0.3150 | -0.1428 | 0.0030 |
| 2 | Score 2 | PPV | 0.2363 | -0.2262 | -0.0378 | 0.1050 |
| 3 | Score 2 | sensitivity | 0.1200 | -0.5000 | -0.5000 | 0.0000 |
| 4 | Score 2 | FPR | 0.5184 | -0.0001 | 0.0000 | 0.0001 |
| 5 | Score 3 | F1.score | 0.2342 | -0.0001 | 0.6458 | 0.6667 |
| 6 | Score 3 | PPV | 1.0000 | NA | NA | NA |
| 7 | Score 3 | sensitivity | 0.1294 | NA | NA | NA |
| 8 | Score 3 | FPR | 1.0000 | NA | NA | NA |
| 9 | treeWAS (all) | F1.score | 0.1565 | -0.2569 | -0.0576 | 0.0029 |
| 10 | treeWAS (all) | PPV | 0.2158 | -0.1666 | -0.0596 | 0.0238 |
| 11 | treeWAS (all) | sensitivity | 0.2986 | -0.5000 | -0.4999 | 0.0000 |
| 12 | treeWAS (all) | FPR | 0.5811 | 0.0000 | 0.0001 | 0.0001 |

**Table 4.4. Wilcoxon Test: Reconstruction Method (Set B).**

|  |  |  |  | **Δ (Parsimony – ML)** |  |  |
|---|---|---|---|---|---|---|
|  | **Association Score** | **Statistic** | **P-value** | **C.I.$_{\textbf{Lower}}$** | **Median** | **C.I.$_{\textbf{Upper}}$** |
| 1 | Score 2 | F1.score | 0.0675 | -0.1541 | -0.0783 | 0.0064 |
| 2 | Score 2 | PPV | 0.1151 | -0.2374 | -0.0950 | 0.0354 |
| 3 | Score 2 | sensitivity | 0.0634 | -0.1999 | -0.1000 | 0.0000 |
| 4 | Score 2 | FPR | 0.3045 | 0.0000 | 0.0001 | 0.0001 |
| 5 | Score 3 | F1.score | 0.0516 | 0.0000 | 0.1056 | 0.2137 |
| 6 | Score 3 | PPV | 0.6089 | -0.2251 | 0.0417 | 0.3590 |
| 7 | Score 3 | sensitivity | 0.0656 | -0.0001 | 0.1000 | 0.2000 |
| 8 | Score 3 | FPR | 0.6598 | -0.0001 | 0.0000 | 0.0001 |
| 9 | treeWAS (all) | F1.score | 0.4652 | -0.0952 | -0.0292 | 0.0555 |
| 10 | treeWAS (all) | PPV | 0.2416 | -0.1454 | -0.0514 | 0.0416 |
| 11 | treeWAS (all) | sensitivity | 0.2403 | -0.1500 | -0.0500 | 0.0499 |
| 12 | treeWAS (all) | FPR | 0.2522 | 0.0000 | 0.0000 | 0.0001 |

**Table 4.5. Wilcoxon Test: Reconstruction Method (Set C).**

The only significant differences observed between the two reconstruction methods are found in Table 4.3, for the simple Set A simulations. All metrics except FPR show significant variation in both Scores 2 and 3. Yet, the signs of the median differences indicate that the reconstruction methods have opposite effects of the performance of the two individual association scores. ML improves the performance of Score 2, while parsimony improves the performance of Score 3. It is notable, however, that with parsimony the composite F1 score measure experiences a nearly 50% increase in Score 3, which is almost ten times the magnitude of the increase conferred by ML to the overall performance of Score 2. In Sets B and C, the signs of the median differences follow a similar opposing trend in the performance of Scores 2 and 3, although none of these differences is significant.

Overall, it appears that the relative benefits and disadvantages of each ancestral state reconstruction method for Scores 2 and 3 cancel each other out in the performance of our approach as a whole. Even in Set A, where performance differences for individual association scores are significant, Table 4.3 reveals that neither reconstruction method improves the performance of our approach overall. A significant difference is observed for FPR in Set A, but no median difference within four significant digits is observed in either direction. Even the upper 95% confidence interval indicates that ML led to, at most, one fewer false positive finding with $N_{loci} = 10,000$. We conclude, therefore, that there is no clear advantage conferred by adopting either reconstruction method over the other.

One other variable that must be considered is computational time. We found that parsimonious reconstructions were consistently completed in a matter of seconds, whether we were working with $N_{loci} = 100$ to $N_{loci} = 1,000,000$. ML reconstructions, on the other hand, demanded more computational time and took several minutes to produce reconstructions for genetic datasets with $N_{loci} \geq 100,000$. Hence, computational efficiency considerations favour parsimony over ML.

We make both methods available to users of the treeWAS R package, for the reconstruction of both genotypic and phenotypic ancestral states. In light of the evidence provided by this analysis and the practical value of computational efficiency and scalability, however, we choose parsimony to be the default method for reconstructing the ancestral states of the genotype within our GWAS approach. We work with parsimonious reconstructions in all analyses below.

### 4.3.1 Excluding branch length improves Score 3 performance

In designing our third measure of association, we wished to explore whether Score 3 would give better results with the branch length term $l_i$ included, as in Equation 3.13, or excluded, as in Equation 3.14. To determine which formulation of Score 3 would provide the more effective measure of association, we compared the performance of the two versions in applications to simulated data. When analysing each of these simulated datasets ($N = 240$), we repeated the calculation of Score 3, substituting Equation 3.14 without branch length (Score3$_{NoBL}$) with Equation 3.13 with branch length (Score3$_{BL}$). Following each analysis, we recorded the performance of each version of Score 3 along the four evaluation metrics defined in Table 4.2.

We ran a two-sample Wilcoxon rank sum test on matched pairs of our four performance statistics under the two conditions. We present the results below, for simulation sets A, B, and C, in Tables 4.6, 4.7, and 4.8. The sign of the median difference indicates which formulation of Score 3 improves performance along metric $x$, such that:

$$Median(x) \begin{cases} < 0 & \textit{including} \text{ branch length improves performance along } x \\ > 0 & \textit{excluding} \text{ branch length improves performance along } x \end{cases} \tag{4.11}$$

Rows containing a statistically significant difference ($p < 0.05$) are highlighted in yellow.

| | | | $\Delta$ ($\mathbf{Score3_{NoBL}} - \mathbf{Score3_{BL}}$) | | |
|---|---|---|---|---|---|
| | **Statistic** | **P-value** | **C.I.$_{\mathbf{Lower}}$** | **Median** | **C.I.$_{\mathbf{Upper}}$** |
| 1 | F1.score | 0.0002 | 0.0978 | 0.1685 | 0.2321 |
| 2 | PPV | 0.5698 | -0.1548 | 0.0379 | 0.2046 |
| 3 | sensitivity | 0.0189 | 0.0000 | 0.1499 | 0.2500 |
| 4 | FPR | 0.7447 | -0.0001 | 0.0001 | 0.0001 |

**Table 4.6. Wilcoxon test: Score 3 (Set A).**

| | | | $\Delta$ ($\mathbf{Score3_{NoBL}} - \mathbf{Score3_{BL}}$) | | |
|---|---|---|---|---|---|
| | **Statistic** | **P-value** | **C.I.$_{\mathbf{Lower}}$** | **Median** | **C.I.$_{\mathbf{Upper}}$** |
| 1 | F1.score | 1.0000 | NA | NA | NA |
| 2 | PPV | 1.0000 | -0.8333 | -0.0595 | 0.8333 |
| 3 | sensitivity | 0.5653 | 0.0000 | 0.0000 | 0.5000 |
| 4 | FPR | 0.6600 | 0.0000 | 0.0000 | 0.0001 |

**Table 4.7. Wilcoxon test: Score 3 (Set B).**

| | | | $\Delta$ ($\mathbf{Score3_{NoBL}} - \mathbf{Score3_{BL}}$) | | |
|---|---|---|---|---|---|
| | **Statistic** | **P-value** | **C.I.$_{\mathbf{Lower}}$** | **Median** | **C.I.$_{\mathbf{Upper}}$** |
| 1 | F1.score | 0.0147 | 0.0123 | 0.1041 | 0.1628 |
| 2 | PPV | 0.5261 | -0.3214 | -0.0461 | 0.1917 |
| 3 | sensitivity | 0.0137 | 0.0001 | 0.1000 | 0.2000 |
| 4 | FPR | 0.2986 | 0.0000 | 0.0000 | 0.0001 |

**Table 4.8. Wilcoxon test: Score 3 (Set C).**

Significant results were found for Set A and Set C. In both Table 4.6 and Table 4.8, we find that treating all branch lengths as equal improves the sensitivity of Score 3, leading to a median increase of 10% - 15%. No significant impact on FPR or PPV is observed, so there is no concomitant cost of excluding the branch length term. As a result, we see an increase in the F1 score that corresponds to the increase in sensitivity. No change in any variable is observed in Table 4.7, but this is in line with expectation for Score 3 in Set B, where associations arise through complementary pathways, as only Score 2 is explicitly designed to detect such associations. Overall, based on this evidence, we choose to adopt this version of Score 3 within our approach. We implement Equation 3.14 in treeWAS and we calculate Score 3 without the branch length term in all analyses below.

## 4.4 Selecting the threshold of significance

The central aim of GWAS is to accurately delineate between significant and insignificant association score values. As discussed in Chapter 3, our approach relies on the simulation of an empirically-shaped null distribution to estimate the appropriate location of the significance threshold. To identify associations for each of our association scores, our method must draw a significance threshold within the upper tail of the relevant null distribution.

How best to select this significance threshold remains unresolved. In Chapter 3, we outlined four components of the threshold-selection mechanism used within our approach, and we introduced various implementation strategies for each component. Here, we used simulated data to make an evidence-based decision about the optimal approach. In applying our approach to each of the simulated datasets ($N = 240$) generated in our three simulation sets, we selected the threshold of significance in 32 different ways, accounting for all unique combinations of the four parameters that control the mechanism of threshold-selection.

| | **Base p-value** | **Multiple testing** | **P-value estimation** | **N sim (x $N_{loci}$)** |
|---|---|---|---|---|
| 1 | 0.0001 | Bonferroni | Count | 1 |
| 2 | 0.0001 | Bonferroni | Count | 10 |
| 3 | 0.0001 | Bonferroni | Density | 1 |
| 4 | 0.0001 | Bonferroni | Density | 10 |
| 5 | 0.0001 | FDR | Count | 1 |
| 6 | 0.0001 | FDR | Count | 10 |
| 7 | 0.0001 | FDR | Density | 1 |
| 8 | 0.0001 | FDR | Density | 10 |
| 9 | 0.001 | Bonferroni | Count | 1 |
| 10 | 0.001 | Bonferroni | Count | 10 |
| 11 | 0.001 | Bonferroni | Density | 1 |
| 12 | 0.001 | Bonferroni | Density | 10 |
| 13 | 0.001 | FDR | Count | 1 |
| 14 | 0.001 | FDR | Count | 10 |
| 15 | 0.001 | FDR | Density | 1 |
| 16 | 0.001 | FDR | Density | 10 |
| 17 | 0.01 | Bonferroni | Count | 1 |
| 18 | 0.01 | Bonferroni | Count | 10 |
| 19 | 0.01 | Bonferroni | Density | 1 |
| 20 | 0.01 | Bonferroni | Density | 10 |
| 21 | 0.01 | FDR | Count | 1 |
| 22 | 0.01 | FDR | Count | 10 |
| 23 | 0.01 | FDR | Density | 1 |
| 24 | 0.01 | FDR | Density | 10 |
| 25 | 0.05 | Bonferroni | Count | 1 |
| 26 | 0.05 | Bonferroni | Count | 10 |
| 27 | 0.05 | Bonferroni | Density | 1 |
| 28 | 0.05 | Bonferroni | Density | 10 |
| 29 | 0.05 | FDR | Count | 1 |
| 30 | 0.05 | FDR | Count | 10 |
| 31 | 0.05 | FDR | Density | 1 |
| 32 | 0.05 | FDR | Density | 10 |

**Table 4.9. Threshold-selection mechanisms.**

These four parameters took on the following values:

- **Base p-value ($\alpha_{base}$):** 0.0001, 0.001, 0.01, 0.05.
- **Multiple testing correction:** Bonferroni correction [282], FDR correction [283].
- **P-value estimation:** count-based approach, kernel density estimation.
- **Number of simulated loci ($N_{sim}$):** $N_{loci}$, $10N_{loci}$.

The resulting set of 32 unique threshold-selection mechanisms are listed in Table 4.9. The box plots below compare the performance of these threshold-selection mechanisms in each of our three main simulation sets. The mean of each metric is indicated in red. Each mechanism in the figures below is labelled with a number that corresponds to the row numbers in Table 4.9.

We aim to select the optimal threshold-selection mechanism, balancing these criteria:

1. PPV should be as close to the maximum as possible (i.e., PPV $\simeq 1$, FPR $\simeq 0$).
2. Sensitivity should not be unjustifiably reduced for minor improvements in PPV.

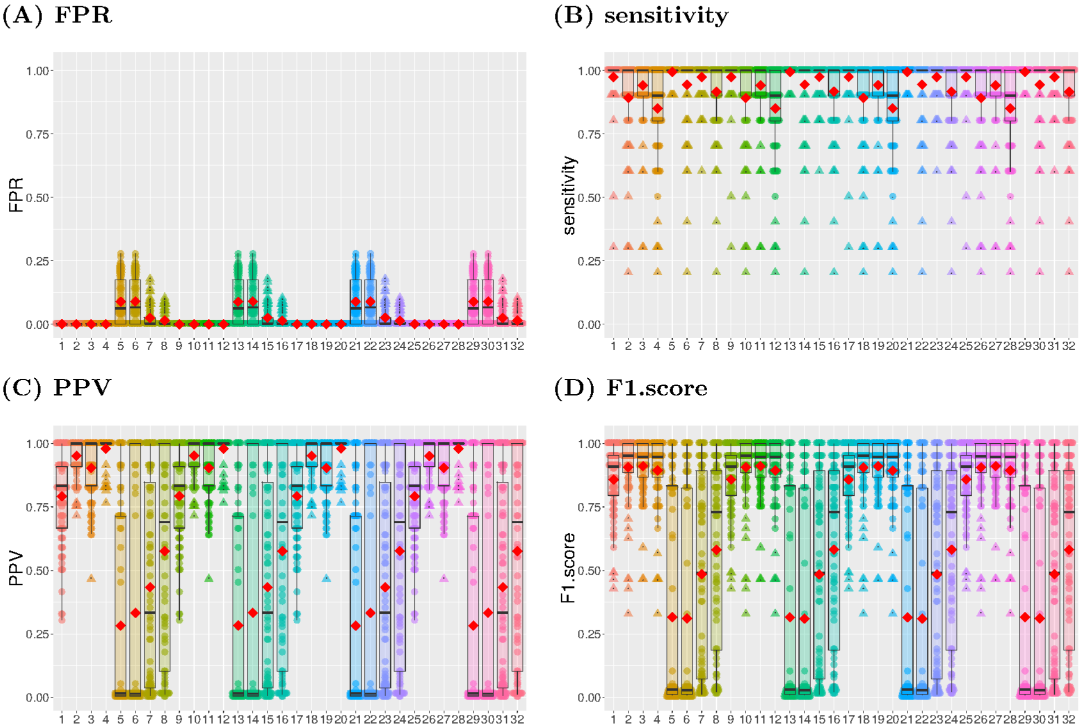


**Figure 4.4. Performance by threshold-selection mechanism (Set A).** treeWAS performance is shown with the significance threshold defined via 32 different mechanisms. X-axis labels correspond to the rows of Table 4.9 that describe each mechanism. Box plots show the median and interquartile range, red diamonds indicate the mean, and individual dots represent results for one analysis ($N = 80$). **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

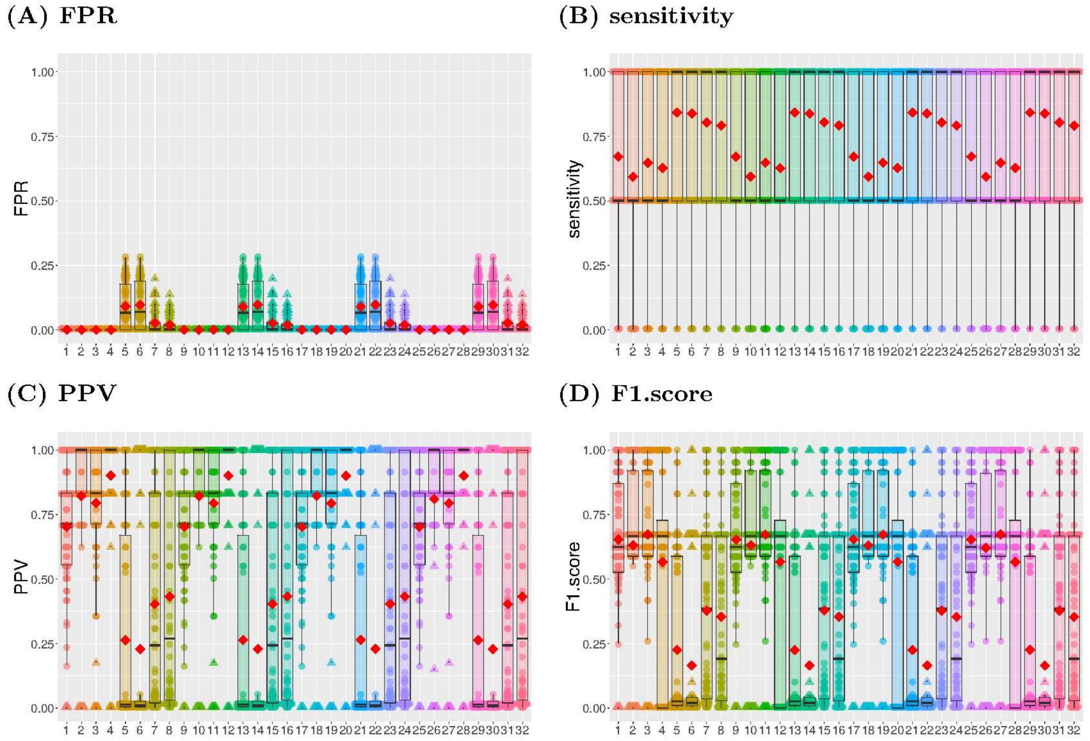


**Figure 4.5. Performance by threshold-selection mechanism (Set B).** treeWAS performance is shown with the significance threshold defined via 32 different mechanisms. X-axis labels correspond to the rows of Table 4.9 that describe each mechanism. Box plots show the median and interquartile range, red diamonds indicate the mean, and individual dots represent results for one analysis ($N = 80$). **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

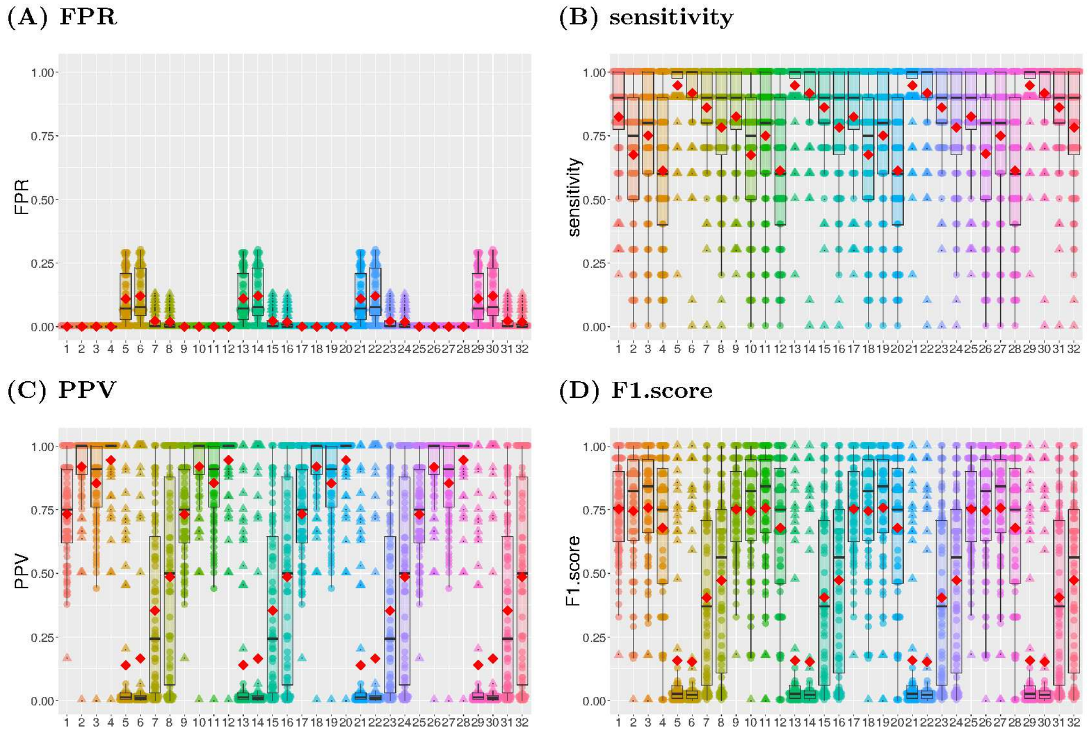


**Figure 4.6. Performance by threshold-selection mechanism (Set C).** treeWAS performance is shown with the significance threshold defined via 32 different mechanisms. X-axis labels correspond to the rows of Table 4.9 that describe each mechanism. Box plots show the median and interquartile range, red diamonds indicate the mean, and individual dots represent results for one analysis ($N = 80$). **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

The results above reveal the relative dominance of threshold-selection mechanisms that:

1. use a Bonferroni correction
2. use the count-based approach to estimate p-values
3. estimate the null distribution with $10N_{loci}$ simulated sites.

Among the threshold mechanisms examined in Figures 4.4, 4.5, and 4.6, we find that threshold 18 ensures the best performance. It combines each of the aforementioned approaches and, in doing so, achieves among the highest F1 scores in each simulation set and consistently strikes the most favourable balance between high PPV and high sensitivity. In light of these findings, we make these three parameter values the default settings within treeWAS to ensure optimal selection of the significance threshold. In all analyses carried out below, these three parameters remain fixed at these optimal values.

Variation in the fourth parameter, the base p-value, had no additional effect on the performance of the threshold-selection mechanism. Figures 4.4, 4.5, and 4.6 show repeating patterns of four-way ties, reflecting the fact that all four base p-values examined ($\alpha_{base}$ =0.0001, 0.001, 0.01, 0.05) resulted in the same performance. This remains true when the performance statistics are separated by association score. Owing to the correction for multiple testing, the effective difference between the four base p-values is minimal. With the Bonferroni correction, they become 3.3x10$^{-}$9, 3.3x10$^{-}$8, 3.3x10$^{-}$7, and 1.7x10$^{-}$6, in this simulation study. Without definitive evidence in support of any particular base p-value, we have opted for a moderate value of 0.01 in all analyses. To ensure that our comparative performance assessment is as fair as possible, we also set the significance level to $\alpha_{base} = 0.01$ when performing GWAS via each of the six alternative methods examined in the pages below.

## 4.5 Evaluating performance

### 4.5.1 Set A

In Figure 4.7, we examine the performance of our approach and its individual association scores, and compare it to the performance of the six alternative GWAS methods that were applied to Set A simulations. Along all four of our evaluation metrics, our approach performs very well. treeWAS demonstrates greater precision and stronger overall performance than any other GWAS method in Set A.

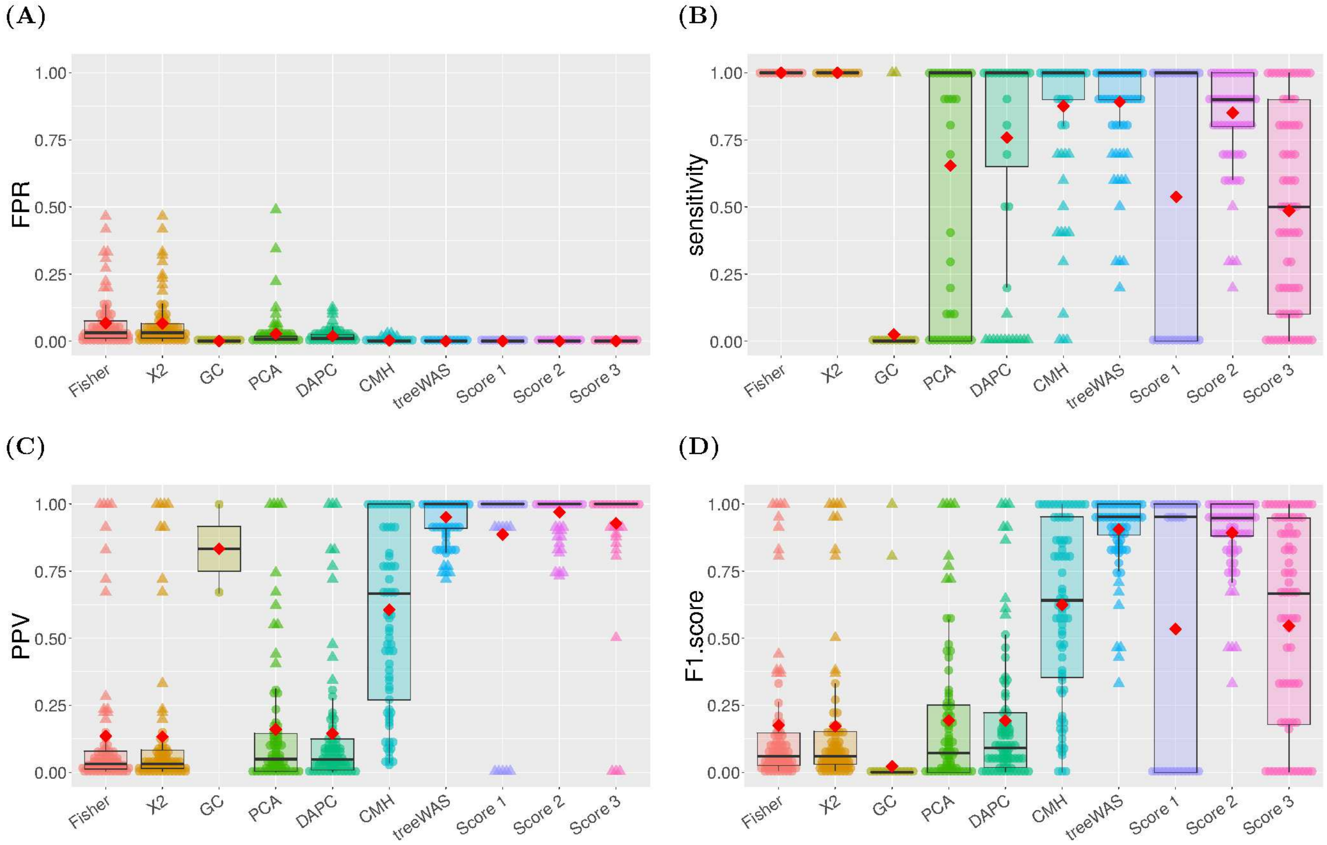


**Figure 4.7. Performance by GWAS method (Set A).** The performance on simulated datasets for the six comparator GWAS methods and our approach, alongside its three association tests individually, is summarised along the four metrics of evaluation. Box plots display the median and interquartile range, red diamonds indicate the mean, and individual dots represent results for one of the 80 simulated datasets. **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

Figure 4.7A shows that Scores 1, 2, and 3 each achieve FPR values of approximately zero, with each score identifying a median of zero false positives ($mean = 0.13, 0.31, 0.18$). The high degree of precision obtained by all three measures across Set A provides valuable support for our pooled, multi-measure approach, as too many false positive findings from any score could have had a compounding effect on the collective FPR of treeWAS. Our results reveal that the cumulative FPR of treeWAS is still near zero (5x10$^{-5}$), which translates to an average of 0.5 false positive findings per analysis (5x10$^{-5}$x9990).

In Figure 4.7B, we find moderate sensitivities for Scores 1 and 3, but a high sensitivity for Score 2. As the strength of associations is defined at terminal nodes in both Set A simulations and in Score 1, we had anticipated stronger Score 1 sensitivity in Set A. But, because it assigns the same value to all Set A associations (Score $1 = 0.9 - 0.1 = 0.8$), Score 1 sensitivity is binary in Set A, dropping to zero when the null distribution shifts its significance threshold above 0.8. Score 3 obtains a similar sensitivity to Score 1 in Set

A, albeit on a discrete scale, by measuring associations on a phylogeny-wide basis. Both our sample- and tree-wide measures are reduced by the "dilution" of signal at terminal nodes in Set A. Score 2 achieves the strongest sensitivity by uncovering many of the simultaneous substitutions that were used to establish association in Set A. Associations missed by Score 2 are recovered by Score 1 in 31 cases and Score 3 in 8. By combining their findings, treeWAS achieves greater sensitivity than Score 1, 2, or 3 alone. Even when the nature of associations favours one score over others, pooling multiple scores improves the discovery power of our approach.

Scores 1, 2, and 3 achieve high PPVs in Figure 4.7C, as each rejects all or nearly all false positive findings. The collective precision of treeWAS is, thus, very high. Critically, while each score enhances the sensitivity of treeWAS, the accumulation of false positives does not undermine its PPV in Set A. The F1 scores in Figure 4.7D confirm that the incorporation of multiple metrics improves the overall performance of our approach.

Our approach stands out against the six other GWAS methods examined in Figure 4.7. The only approaches to exceed treeWAS in sensitivity are the population-naive Fisher's exact and $X^2$ tests, both of which identify all ten truly-associated loci in in Figure 4.7B. Yet, Figure 4.7C reveals that these uncorrected tests found 30-65 false positives for every true positive identified. Our approach clearly represents a substantial improvement over this baseline error rate. At the other extreme, the uniform correction applied by GC consistently reduces FPR to zero in Figure 4.7A. But, in the clonal and semi-clonal populations simulated, this highly specific approach appears to sharply over-correct for population stratification, repeatedly giving GC zero sensitivity in Figure 4.7B.

Among the ancestry-aware GWAS methods, the CMH test displayed the strongest sensitivity in Set A, aside from treeWAS. Though with a slightly lower mean and more variation, the cluster-based test achieves similar power to our tree-based approach in Figure 4.7B, as both regularly find nine or ten of the ten associated sites. DAPC and PCA have sensitivities well above GC and just above our sample-wide Score 1 and phylogeny-wide Score 3, yet clearly below the CMH test and treeWAS as a whole. Greater performance gaps are exposed in Figure 4.7C, where PCA and DAPC show surprisingly poor precision, with PPV values closer to the uncorrected Fisher and $X^2$ tests than to our approach. In PPV, the CMH test is again the closest competitor to treeWAS. But, where CMH and treeWAS sensitivities differed slightly in Figure 4.7B, their PPVs diverge significantly in Figure 4.7C. In Set A, treeWAS consistently found one or zero false positives; whereas, the CMH test regularly found as many false positives as true positives. Ultimately, Figure 4.7D sets treeWAS starkly apart from all other GWAS methods, with a mean F1 score $>$ 50% above the next-best CMH test. We conclude that, with high sensitivity and unmatched precision, our phylogenetic approach was able to achieve the strongest overall performance in Set A.

### 4.5.2 Set B

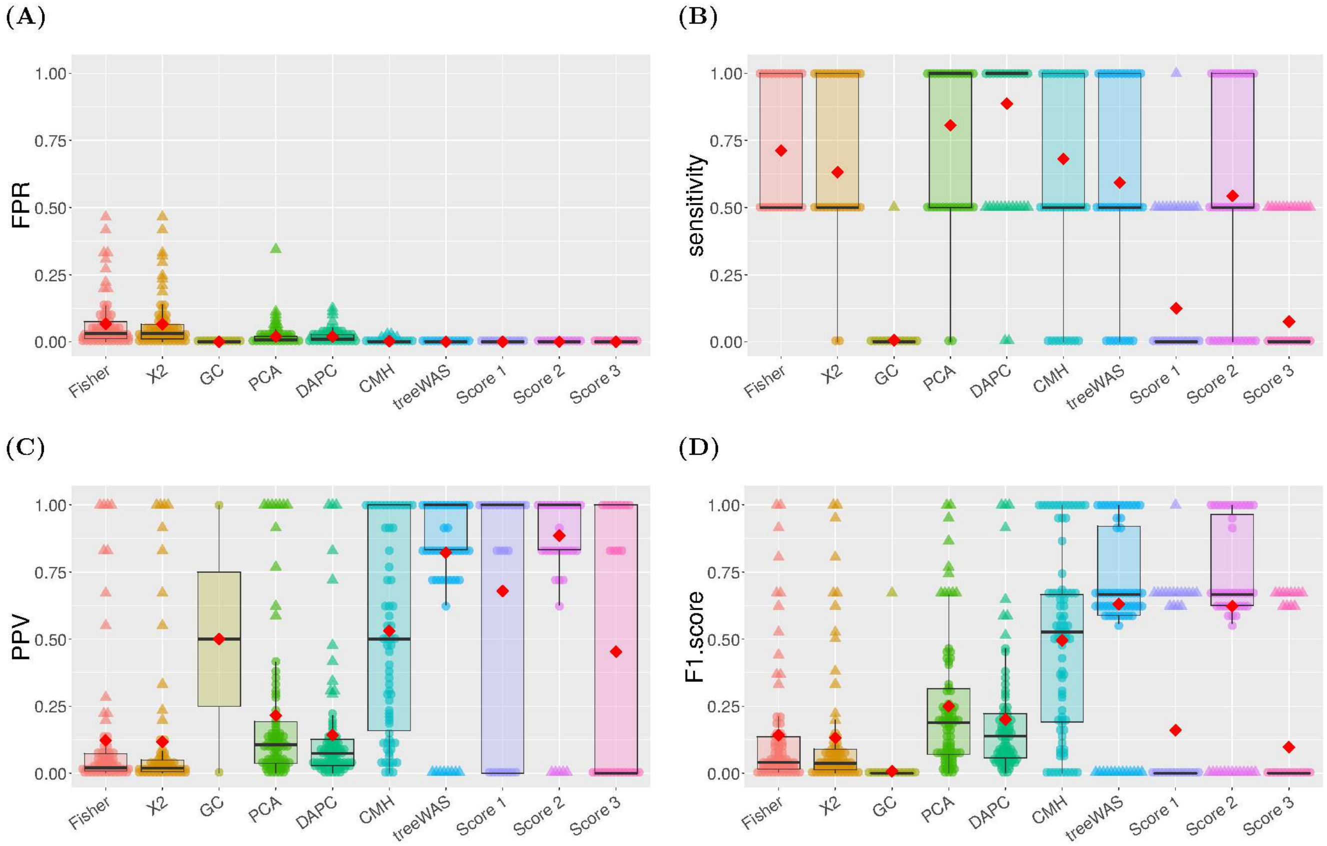


**Figure 4.8. Performance by GWAS method (Set B).** The performance on simulated datasets for the six comparator GWAS methods and our approach, alongside its three association tests individually, is summarised along the four metrics of evaluation. Box plots display the median and interquartile range, red diamonds indicate the mean, and individual dots represent results for one of the 80 simulated datasets. **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

We developed simulation Set B to determine whether Score 2 is able to identify genetic associations with phenotypes that arise through complementary pathways. Figure 4.8A shows that Scores 1-3 and treeWAS overall achieve FPR values of approximately zero, with treeWAS finding on average 0.6 false positives per analysis. In fact, we find the same relative FPR values for all GWAS methods in Set B as we had in Set A.

Figure 4.8B confirms our expectations for Set B. As Score 2 is the only one of our three tests that is designed to detect associations indicated by homoplasy alone, it is also the only score capable of repeatedly detecting complementary associations that are not upheld sample-wide or phylogeny-wide. The sensitivity of treeWAS closely matches that of Score 2 alone, typically identifying six of ten associations simulated. Nevertheless, even in Set B, all three scores added to the power of our approach.

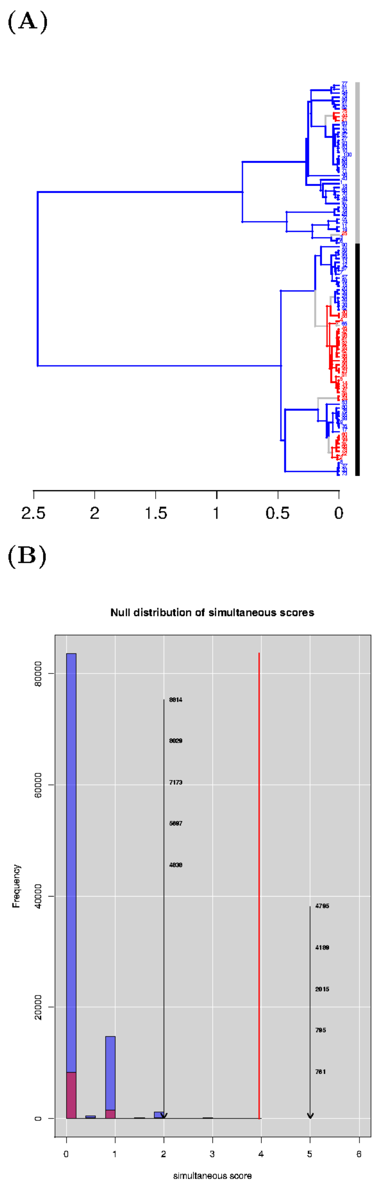


**Figure 4.9. Complementary association (Set B). A:** Phylogenetic tree, coloured by phenotype (blue = 0, red = 1). Terminal marker indicates individuals with association at loci 4838, 5697, 7173, 8029, 8814 (grey), or at loci 761, 795, 2015, 4189, 4795 (black). **B:** Null distribution for Score 2 showing the locations of the two sets of simulated associated loci, both above and below the significance threshold.

In many analyses, PCA, DAPC, and treeWAS are actually able to improve upon the power of the uncorrected tests while also improving upon their PPV, demonstrating that correcting for population stratification can reduce both type I and type II error. Interestingly, PCA and DAPC rose from among the least powerful methods in Set A to the most powerful in Set B, while treeWAS fell from most powerful to least powerful. In context, however, it is clear that the small gain in sensitivity by PCA and DAPC is overshadowed by the PPV benefits offered by treeWAS. Note that the low PPVs of Scores 1 and 3 are due to their sensitivities (see Equation 4.8); the mean number of false positives found in Set B (0.15, 0.34, 0.23) is in line with Set A (0.13, 0.31, 0.18). Figure 4.8C shows treeWAS to be unique in its ability to eliminate false positives from Set B analyses.

Although Score 2 does not achieve maximum sensitivity in Set B, the pursuit of further increases in sensitivity is not, in fact, desireable in this context. To explain why, we must examine the data underlying Figure 4.8C. As an example, compare the two clades in Figure 4.9A. Two sets of complementary associations were simulated in perfect association across half of the tree. But, clearly, the evidence for association in the smaller, upper clade is weak. Our homoplasy score recognises that this association arises through only two substitutions, giving Score 2 = 2 at this locus. Accounting for the inferred tree and $N_{sub}$ distribution, our simulation-based approach places this Score 2 value well within the null distribution in Figure 4.9B, as many loci with $N_{sub} \geq 2$ could achieve a similar degree of association by chance. None of the other GWAS methods accounts for $N_{sub}$, or its impact on the genome-wide probability of false positive findings, in their assessment of significance. As such, they may increase sensitivity by detecting loci like this one; but, in doing so, they are prone to accepting volumes of spurious associations that obtained similar scores by chance. The low and variable PPVs in Figure 4.8C and poor overall performance in Figure 4.8D confirm that unfavourable sensitivity-specificity trade-offs were made by all six other GWAS methods in Set B. By drawing on additional empirical information and evolutionary inference, our approach was able to more carefully evaluate the evidence for association and to strike a more appropriate balance between power and precision.

### 4.5.3 Set C

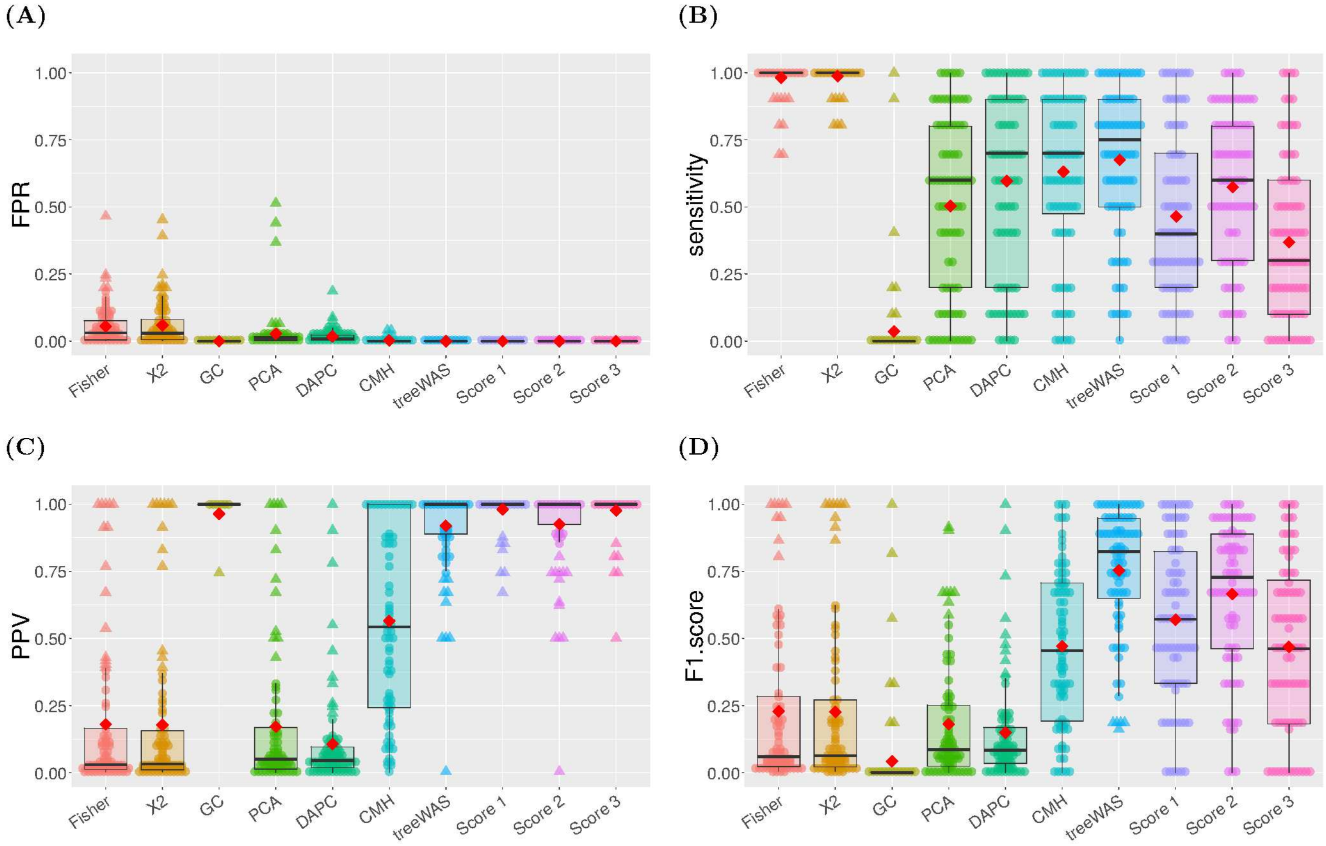


**Figure 4.10. Performance by GWAS method (Set C).** The performance on simulated datasets for the six comparator GWAS methods and our approach, alongside its three association tests individually, is summarised along the four metrics of evaluation. Box plots display the median and interquartile range, red diamonds indicate the mean, and individual dots represent results for one of the 80 simulated datasets. **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

The trends in performance of the seven GWAS approaches compared in Set C largely conform to the patterns observed above, in Set A and, to a lesser extent, in Set B. In Set C, we observe the greatest consistency in power and precision across all three of our association scores. As opposed to Sets A and B, where the power of treeWAS was due in large part to Score 2, in Set C, no single score stands out as definitively responsible for the sensitivity of treeWAS overall. In Figure 4.10B, Scores 1-3 each display a moderate sensitivity, discovering on average 4.7, 5.7, and 3.7 true positive findings. These findings show that our three scores were reasonably successful in detecting the more complex patterns of associations simulated in Set C. And, as further explored below (see Figure 4.11), we find that separate scores repeatedly picked up on different signals and identified distinct sets of truly-associated loci. Individually, the sensitivities of our three scores fell below all methods except GC. But, together, they achieved a 20% increase on our most powerful Score 2, discovering a mean 6.8 true positive findings. Their collective power pushed treeWAS above all population-aware alternatives in Figure 4.10B. We are especially pleased to find that the high power of our multi-measure approach is upheld in our most sophisticated and biologically-realistic simulation set. These results suggest that the flexibility and sensitivity gained by pooling our three scores will be an equal or greater asset to treeWAS in real GWAS studies than in the three simulation sets above.

The PPVs in Figure 4.10C take on a familiar pattern. Scores 1-3 achieve high PPV values, contributing 0.09, 0.46, and 0.09 false positives, on average. The collective precision of treeWAS, thus, remains very high ($median = 0$, $mean = 0.6$ false positives). Among comparator methods, the precision of GC in Figure 4.10C is robbed of its utility without any power in Figure 4.10B. PCA and DAPC, meanwhile, sacrifice considerable sensitivity in Figure 4.10B for only modest FPR reductions in Figure 4.10A. This produces ratios of true and false positive findings similar to the uncorrected Fisher and $X^2$ tests in Figure 4.10C. In fact, the F1 scores in Figure 4.10D suggest that, in Set C, correcting for ancestry via GC, PCA, or DAPC may be more detrimental to overall performance than making no correction at all. The CMH test is clearly more precise than any of these methods. Still, in Set C, as in Sets A and B, one could not be any more certain that a CMH test finding was a true positive than a false positive. This major weakness in precision, suffered by even our strongest competitor, is overcome by treeWAS once again in Set C. At the same time, treeWAS displays greater power in Figure 4.10B than any ancestry-aware alternative. The result, in Figure 4.10D, is that the overall performance of treeWAS exceeds all other methods by the largest margin of any simulation set.

### 4.5.4 Association scores are most informative when pooled

One of our aims in this study was to determine how our three association scores could be used most effectively within our GWAS method. We considered three possible approaches:

1. Single score: use only the best-performing score.
2. Multiple scores implemented in parallel: treeWAS = (Score 1 $\cup$ Score 2 $\cup$ Score 3).
3. Multiple scores performed sequentially: treeWAS = (Score 1 $\cap$ Score 2 $\cap$ Score 3).

The strongest single measure in Sets A, B, and C was Score 2. However, we repeatedly found that all three scores improved the collective power and overall performance of our approach. Score 3 made the smallest marginal contribution to treeWAS sensitivity, though this was not surprising. Its purpose, where possible, was to compensate for any gaps left between our primary allele-based Score 1 and homoplasy-based Score 2. Given the comparatively modest power of Scores 1 and 3, additional analyses, including those in Chapter 5, may help us determine whether allele-based metrics are less sensitive by nature or if bias in our simulations simply favoured the homoplasy-counting scheme. Notably, the relative contributions of our three measures varied both between and within simulation sets. The dominance of Score 2 was not universal, as Scores 1 or 3 achieved the greatest power and/or precision in many analyses. We cannot expect to predict which score will be most useful in each new empirical analysis. Thankfully, the performance of our collective approach was rarely exceeded by any measure on its own. Therefore, rather than relying on a single score, we have chosen to incorporate multiple measures of association within our GWAS method.

The results of this study indicate that, in parallel, our three measures produce good statistical power and very low type I error, despite the accumulation of false positives with each score. A sequential implementation would progressively narrow down the set of significant findings. Our simulation study suggests that this could reduce type I error, but only very minimally. However, it would likely increase type II error to a considerable degree. Indeed, we suspect that both Approach 1 and 3 will be under-powered in most analyses. We expect that our GWAS method will achieve greater power, similar precision, and better overall performance by pursuing Approach 2 and pooling the findings of our three tests of association.

The Venn Diagram in Figure 4.11 lets us compare the proposed parallel implementation to the single-score and sequential implementations used elsewhere in microbial GWAS [114, 121, 126, 127, 147]. Given the true positives in Figure 4.11A, the power we can derive via Approach 1 with Score 2 is 57% (459/800). If we pool findings via Approach 2, power reaches 68% (540/800). Conversely, if we take their intersection via Approach 3, power drops to 29% (235/800). Of course, as all three scores never misidentify a spurious association in Figure 4.11B, Approach 3 eliminates all false positive findings. Still, Approach 1 finds just 0.46 false positives per analysis, and Approach 2 only increases this to 0.6. The parallel implementation gives a mean F1 score of 0.78; whereas, this decreases to 0.67 with Score 2 alone, and it drops to 0.45 with the sequential implementation. Therefore, rather than relying on a single score or a sequential implementation, we prefer to adopt a parallel implementation of our three scores in treeWAS.

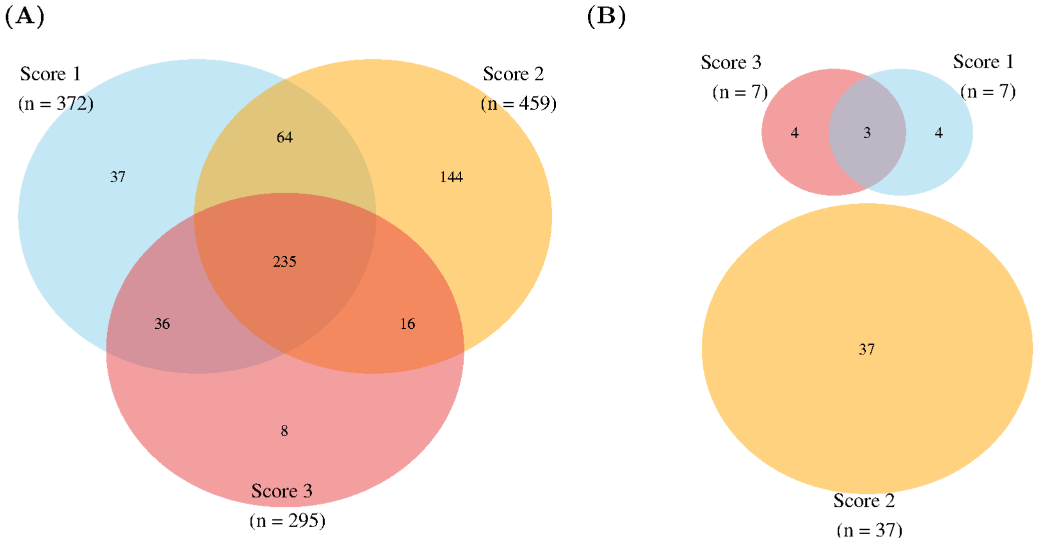


**Figure 4.11. Venn Diagram: association scores (Set C).** Two Venn Diagrams show the findings of Scores 1, 2, and 3 across the 80 Set C datasets, each of which contained ten associated loci. **A:** True positive findings for Scores 1, 2, and 3, which collectively identified 540 associations, out of a possible 800. Of these, 35% (189/540) are uniquely contributed by one score, 21% (116/540) by two scores, and 44% (235/540) by all three scores. **B:** False positive findings for Scores 1, 2, and 3, which collectively misidentified 48 loci as significant, out of a total of 799,200 truly non-associated sites. Of these, 94% (45/48) are uniquely contributed by one score, 77% (37/48) by Score 2 alone, and 6% (3/48) by two scores, Scores 1 and 3. No spurious findings are collectively misidentified by all three scores.

### 4.5.5 Summary

The results of this simulation study provide strong support for the strategy adopted by our phylogenetic approach to microbial GWAS. In Sets A, B, and C, our approach consistently achieved a low FPR, identifying a median of zero false positive findings in each set ($mean = 0.51, 0.64, 0.60$). The simulated genetic loci contained a diverse distribution of association statistics, reflecting multiple independent and interacting factors, including variaton in population structure, the genotypic homoplasy distribution due to variable recombination and mutation, the number of phenotypic substitutions, the stratification of phenotypic states along ancestral lines, and the variable nature and effect size of associations. Yet, across this complex and shifting association landscape, treeWAS remained able to distinguish between spurious and sufficient association signals. These results attest to the robust and responsive nature of our simulation-based approach. They show that treeWAS was able to maintain effective control over the confounding factors that we simulated in these datasets and which we expect to encounter in real microbial GWAS analyses. Our findings highlight the precision of our three measures of association, each of which upheld $H_0$ correctly at all 9,990 non-associated sites in $\geq 75\%$ of the above analyses. And, they provide support for our multi-measure association testing strategy, demonstrating that treeWAS can maintain high specificity even when the results of Scores 1-3 are pooled. In fact, ours was the only GWAS method that demonstrated a reliable ability to separate true from false positive findings in this study.

While the precision of treeWAS relied on the unwavering stringency of each association test, its high collective power was supported by more variable sensitivities in our three scores. Our conservative approach to significance testing produced moderate sensitivities in each score, often below most competing approaches. But, with robust findings from each score, our pooled approach attained high levels of statistical power. The collective power of treeWAS regularly exceeded any of its individual components, obtaining the strongest cumulative effect in our most realistic, final simulation set. Remarkably, while maintaining far greater precision than any other method, our approach also achieved greater power than any population-aware alternative in Sets A and C, with power approaching most competitors in Set B. Altogether, treeWAS was able to achieve the highest F1 scores and the best overall performance of any method in each simulation set.

### 4.5.6 Comparison with other GWAS methods

In this simulation study, five of the six comparator approaches repeatedly failed to reject large numbers of false positive findings. Only treeWAS and the strict GC approach regularly avoided type I errors. PCA, DAPC, and the CMH test reduced FPR below the level incurred with no correction for population structure, but, for every genuine association they detected, respectively, 25, 19, and 3 spurious associations were deemed significant. Our results suggest that these popular dimension-reduction and cluster-based methods may not correct sufficiently for confounding population structure in microbial GWAS. In practice, the type I error rates displayed by PCA, DAPC, and CMH would cause a lot of time and money to be wasted following up on truly insignificant leads.

Each of the non-phylogenetic controls for population structure simplify the extensive genetic relationships between isolates. Population stratification is summarised by GC in the single $\lambda_{GC}$ statistic, by the CMH test in $k$ clusters, and by PCA and DAPC in $(k-1)$ PCs. These approaches work well in a human genetics context, where an explicit distinction can be made between ancestral "population structure" and recent "cryptic relatedness" or unknown family relationships [140, 150]. However, this paradigm does not naturally apply to the ancestral relationships in many microbial samples. The k-means clusters used in CMH, and the PCs of PCA and DAPC often correspond to the lineages of a phylogenetic tree [80, 128, 159]; although, as we showed in Figure 2.4, this relationship is not always intuitive. But, using $(k-1)$ PCs or $k$ clusters as fixed effects in microbial GWAS requires the analyst to make a conceptual delineation at a given height on the genealogical tree, between what will and will not be considered parts of the population structure. Our approach, by contrast, retains relevant information at all levels of the clonal population structure by working directly with the full phylogeny. And, in maintaining the phenotype along the tips of the tree, our approach makes sure to account for stratification as a function of both phenotypic and genotypic covariance. Most of the non-phylogenetic methods examined in this simulation study are plagued by type I errors because they rely on an assumption that is often unjustifiable in microbial GWAS, that, within $k$ clusters or beyond $(k-1)$ PCs, genetic variation is ancestrally homogenous and confounding bias is not responsible for statistically significant associations [297]. Our approach avoids this pitfall by using inference of the full evolutionary history to determine which associations are unlikely to arise by chance.

Whereas the power of treeWAS was enhanced by combining multiple precise association tests, this study revealed the non-phylogenetic approaches to be reliant on a zero-sum trade-off between sensitivity and specificity. No comparator method offered both high PPV and high sensitivity. GC paired high PPV with low-to-zero power, rarely producing any findings, correct or incorrect. PCA and DAPC achieved only moderate power, typically below treeWAS. Both methods sacrifice power as they exclude major axes of variation that correlate with the phenotype [128, 297]. DAPC restricts this cost to higher-order lineage effects by maximising between-group variation, while PCA sensitivity is further weakened by the within-group variation in its $(k-1)$ PCs [145,181]. Nevertheless, confounding variation persisted, particularly among lower levels of the population structure, leaving both methods with high FPRs and low PPVs. Increasing the number of PCs in an effort to reduce FPR in this study regularly resulted in a complete loss of sensitivity. Our results suggest that when PCA and DAPC are used in microbial GWAS, depending on the population structure and the effect size of associations, a satisfactory trade-off between sensitivity and specificity may be unattainable. Compared to PCA and DAPC, the CMH test more effectively managed the sensitivity-specificity trade-off by performing a more stringent stratified test, without regressing out relevant information. In fact, CMH sensitivity fell just below treeWAS in Sets A and C. But, while it was more precise than PCA and DAPC, CMH precision fell well below treeWAS. Whereas $\geq 90\%$ of treeWAS test positives were true positives, almost half of CMH test results were false positives in Sets A, B, and C. In terms of overall performance, CMH F1 scores were similar to the weakest individual association test in treeWAS. Furthermore, the CMH test attained these F1 scores by adopting the less conservative approach of favouring high sensitivity at the expense of precision. In practice, it may be preferable to pursue a precision-led strategy, even at some cost to sensitivity, to ensure that results remain reliable and follow-up studies worthwhile.

Though we implemented optimisation procedures for PCA, DAPC, and CMH, it is possible that additional optimisation efforts could produce further improvements in performance. For example, we could attempt to improve the model of population structure or alter $k$ or the number of PCs, through visual inspection or by comparing the results of repeated analyses, guided by metrics like $\lambda_{GC}$. In this way, we might strike a better balance between sensitivity and specificity, although it would have been impractical to perform this process for each of the 240 analyses in this study. This is a clear limitation of these alternative approaches, as it demands more time and human effort, requires considerable user experience, and increases the subjectivity and variability of GWAS results. Our approach removes these barriers by carrying out any necessary optimisation procedures automatically within treeWAS. This ensures that treeWAS can objectively identify the significance threshold with precision, without squandering sensitivity.

In comparing methods across Figures 4.7,4.8, and 4.10, we found that the performance of the non-phylogenetic methods was limited by multiple factors: the focus on higher-order population structure, insufficient control for the confounding effects of ancestry, lack of control for the impact of substitution on spurious association, and the need to make inefficient trade-offs between sensitivity and PPV. By avoiding these pitfalls, the design of treeWAS is able to produce stronger performance on these simulated datasets.

## 4.6 Evaluating performance by recombination rate

One potentially significant limitation of existing microbial GWAS methods is that they do not account for the impact of varying substitution rates on the probability of spurious association. We designed our simulation study to explore how the performance of treeWAS and existing GWAS methods varies as a function of recombination rate. In each of our three simulation sets ($N = 80$), we generated 20 datasets at each of four recombination rates ($R = 0.01, 0.05, 0.1$) by defining $N_{sub}$ according to the homoplasy distributions in Figure 4.1.

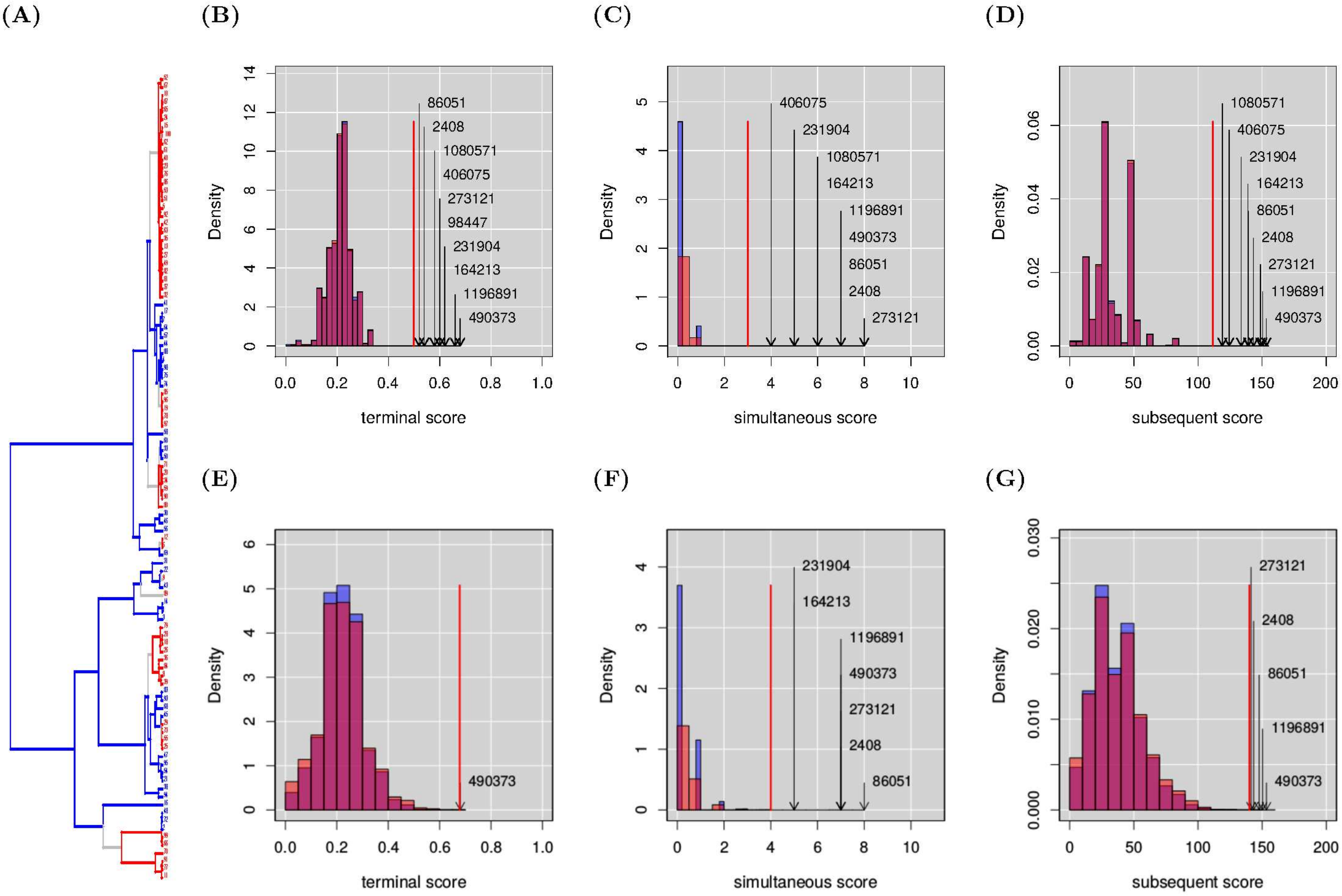


**Figure 4.12. Recombination impacts the probability of spurious association.** Two genetic datasets have been simulated with the same **A:** phylogenetic tree and phenotype (blue = 0, red = 1, grey = $p_i^{anc} \neq p_i^{des}$), and with the same 10 associated sites ($N_{loci}$ = 10,000). The only difference is that Dataset 1 was simulated with mutation alone ($R = 0$, Figure 4.1A), while Dataset 2 simulates mutation and recombination ($R = 0.1$, Figure 4.1D). We applied treeWAS to both datasets. The empirical (red) and null (blue) distributions are shown, with significant findings for Dataset 1 ($R = 0$) in **B:** Score 1, **C:** Score 2, and **D:** Score 3, and for Dataset 2 ($R = 0.1$) in **E:** Score 1, **F:** Score 2, and **G:** Score 3. In Dataset 2, the empirical and null distributions for each score extend further right and the significance threshold ($\alpha_{base} = 0.01$) falls at a higher value. Recombination alone has increased the probability of spurious association, in this case. Of the 10 associations in each dataset, far fewer can be identified in Dataset 2, without accepting many more false positives. By accounting for $N_{sub}$, our approach adjusts for the increased confounding effect of recombination on association inference.

Figure 4.12 shows how an increase in recombination alone can impact the probability of spurious association by inflating the association score values obtained by chance. Comparing Figures 4.12B-D and E-G, we see that the increase from $R = 0$ to $R = 0.1$ has produced a right-ward shift in the distributions of Scores 1, 2, and 3. Figure 4.12 also shows that treeWAS remains robust despite this change in $R$, responding by increasing the significance threshold for each Score. As our simulations uphold $H_0$ at 9,990/10,000 empirical sites, we are pleased to see considerable overlap between our simulated null distribution (blue) and the empirical distribution (red) in each panel of Figure 4.12B-G.

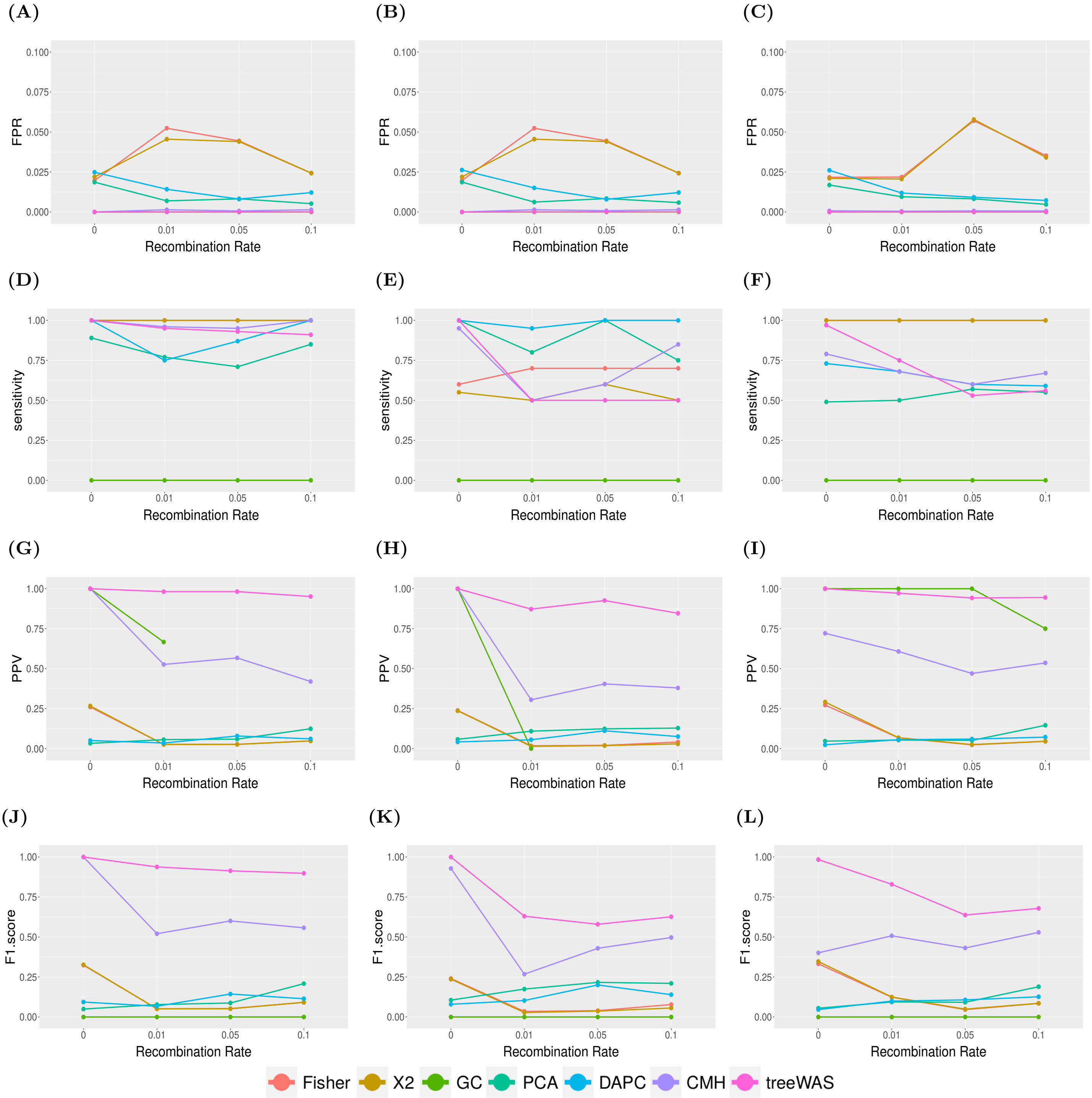


**Figure 4.13. Performance by recombination rate.** Interquartile mean performance by GWAS method and recombination rate is plotted along four statistics (by row), presenting **A-C:** False Positive Rate, **D-F:** Sensitivity, **G-I:** Positive Predictive Value, and **J-L:** F1 Score values across three simulation sets (by column), with Set A first (A, D, G, J), Set B second (B, E, H, K), and Set C third (C, F, I, L). Each plot contains average values of the relevant statistic (y-axis) at four recombination rates (x-axis), showing performance trends for treeWAS and six comparator GWAS methods (legend at bottom).

Figure 4.13 shows how the performance of our approach and the six comparator GWAS methods responds as the background recombination rate varies from clonal to increasingly recombinant. The FPR values of the Fisher's exact test and $X^2$ test in Figures 4.13A-C vary noticeably as the recombination rate changes along the x-axis. As neither test corrects for population stratification or recombination, these FPR values represent the baseline number of false positive findings expected at particular recombination rates, given the relative numbers of phenotypic and genotypic substitutions in each dataset.

Looking back at Figure 4.1, we can see that as $R$ increases from 0 to 0.01, 0.05, and 0.1, the upper tail of the homoplasy distribution among non-associated sites approaches, reaches, and then exceeds the mean of 15 phenotypic substitutions. Figures 4.13A-C show a corresponding trend in the Fisher and $X^2$ FPR values, as the number of false positive findings increases, plateaus, and finally declines as the recombination rate increases above. Striking a clear contrast, treeWAS maintains a stable FPR at zero and consistently eliminates false positives as the recombination rate increases, demonstrating a distinct ability to control for the confounding effects of mutation and recombination.

In Figures 4.13D-F, while the sensitivity of competing approaches vary noisily, the sensitivity of treeWAS appears to decrease with increasing recombination. As Figure 4.12 shows, our sensitivity declines because treeWAS can no longer attribute significance to some more weakly associated loci when similar patterns of association are likely to occur by chance. This data-dependent behaviour varies by context, reducing sensitivity in Sets B and C more than in Set A, acting only where necessary to suppress FPR.

We do see a slight decline in the PPV of treeWAS as recombination increases in Figures 4.13G-I. But, in practice, this implies only a small shift from an average of zero to one false positive finding between $R = 0$ and $R = 0.1$. On the contrary, the PPV trends of PCA and DAPC reveal a major weakness in the use of PCs as fixed effects to control for ancestry. In Figures 4.13A-C, the FPRs of both multivariate approaches clearly improve with increasing recombination, decreasing by half between $R = 0$ and $R = 0.1$. This is reflected in a slight PPV increase in Figures 4.13G-I. While PCA and DAPC may eliminate more false positives in more recombinant organisms, however, it is clear from their PPV trends that if signals of clonality remain (whether at $R = 0$ or $R = 0.1$), the vast majority of test positives identified by either multivariate approach may be false positives nonetheless. The behaviour of the CMH test also differs from that of treeWAS. The CMH test maintains relatively stable sensitivity as $R$ increases in Figures 4.13D-F, whereas Figures 4.13G-I show a sharp decline in CMH PPV in response to increasing recombination. This considerable increase in the number of false positives identified by CMH indicates a lack of control for recombination.

Although the F1 scores of treeWAS and the CMH test in Figures 4.13K and L appear to narrow with increasing recombination, it is important to note the practical implications of the trade-offs being made by both methods. For example, while the sensitivity of the CMH test exceeds that of treeWAS at the highest recombination rate ($R = 0.1$) in Figure 4.13F, this sensitivity benefit corresponds only to the identification of one additional true positive finding on average. A comparison of the PPV values in Figure 4.13I, however, reveals that the CMH test has paid a disproportionate penalty in precision. Conversely, treeWAS uses its estimation-by-simulation procedures to identify a more useful trade-off, instead exchanging a marginal sensitivity cost for a substantial specificity benefit. As a result, even where the F1 score of CMH comes closest to that of treeWAS, at $R = 0.1$ in Set C, treeWAS finds less than one false positive on average, while the CMH test results contain as many false positives as true positives.

In the presence of recombination, the microbial GWAS literature recommends the use of dimension reduction or cluster-based controls for population structure. Yet, at all recombination rates in this study, our results indicate that equal or greater precision and overall performance can be achieved by our recombination-aware phylogenetic approach. The cluster-based CMH test, which is evidently the strongest non-phylogenetic method examined, repeatedly achieves its greatest precision and overall performance in purely clonal conditions (Figures 4.13G-K; $R = 0$). Moreover, we find that the CMH test is not robust to the introduction of even minimal recombination. Ultimately, none of the non-phylogenetic alternatives to treeWAS is able to recognise the variable influence of recombination or to respond appropriately to the changing probability of spurious association. Across four recombination rates in three simulation sets, our approach strikes a more sensible balance between power and precision than any other method.

As the F1 scores in Figures 4.13J-L demonstrate, our approach not only produces the strongest overall performance, but by accounting for recombination, it is able to maintain this advantage across a range of backgrounds, from purely clonal to recombinant. In addition to providing a more thorough control over population structure, treeWAS was the only method capable of accounting explicitly and appropriately for the variable confounding effects of recombination in this study.

## 4.7 Evaluating performance by dataset size

The dimensions of the datasets analysed in the simulation study above were selected to be within the typical range for microbial GWAS and to give representative results without being so large as to reduce the number of repetitions performed. Naturally, however, we expect that in empirical analyses our GWAS method will encounter both larger and smaller microbial samples, with larger or smaller genome sizes. To determine whether the performance of treeWAS is robust to such variations in dataset size, we simulated three additional panels of datasets. In each case, $N = 100$ datasets were simulated along randomly-generated coalescent trees, and the phenotype and associations at ten genetic loci were simulated according to the Set C framework.

We first aimed to simulate accessory gene presence-or-absence data which, compared to core SNP data, usually contains fewer genetic variables that often undergo more frequent substitutions, especially when facing the selective pressures common in GWAS. In the first panel of datasets, therefore, we set $N_{ind} = 100$ and we reduced $N_{loci}$ to 5,000 sites. We used SimBac [289] to simulate a genetic dataset with a recombination rate of $R = 0.2$, twice as high as any rate examined above, and we estimated the homoplasy distribution in Figure 4.14 with Fitch parsimony [255]. Accesory genomes were simulated along randomly-generated coalescent trees, using the $R = 0.2$ homoplasy distribution to simulate the evolution of accessory genome variation with frequent gain and loss of genetic elements.

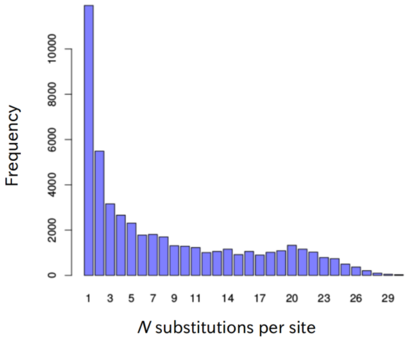


**Figure 4.14. SimBac homoplasy distribution (R = 0.2).** A histogram of the number of substitutions per site when $R = 0.2$.

Our aim in simulating the second and third set of additional datasets was to examine how the change in dataset size alone might impact the performance of treeWAS and competing GWAS methods. In the second panel of datasets, we set $N_{ind} = 100$, and we let $N_{loci}$ increase from 10,000 to 100,000, by selecting 100 uniformly-distributed values, such that $N_{loci} \in \{10000, 10909, ..., 99091, 100000\}$. In the third panel, we set $N_{loci} = 10,000$, and we let $N_{ind}$ vary uniformly in $[50, 200]$, such that $N_{ind} \in \{50, 52, ..., 198, 200\}$. We applied treeWAS and the six comparator GWAS methods to each of the 100 datasets in all three panels. Figure 4.15 compares the performance of each method on accessory genome datasets. Figure 4.16 plots performance as a function of the number of genetic loci in a dataset, and Figure 4.17 shows how performance varies with the number of individuals.

### 4.7.1 Performance on accessory genome data

To determine whether the typically small genomes and more frequent recombination of accessory genome datasets would impact performance, we applied treeWAS and the six comparator methods to 100 accessory gene presence-or-absence datasets, simulated with $N_{ind} = 100$, $N_{loci} = 5,000$, and $R = 0.2$. Figure 4.15 presents the results, revealing trends broadly similar to those observed above.

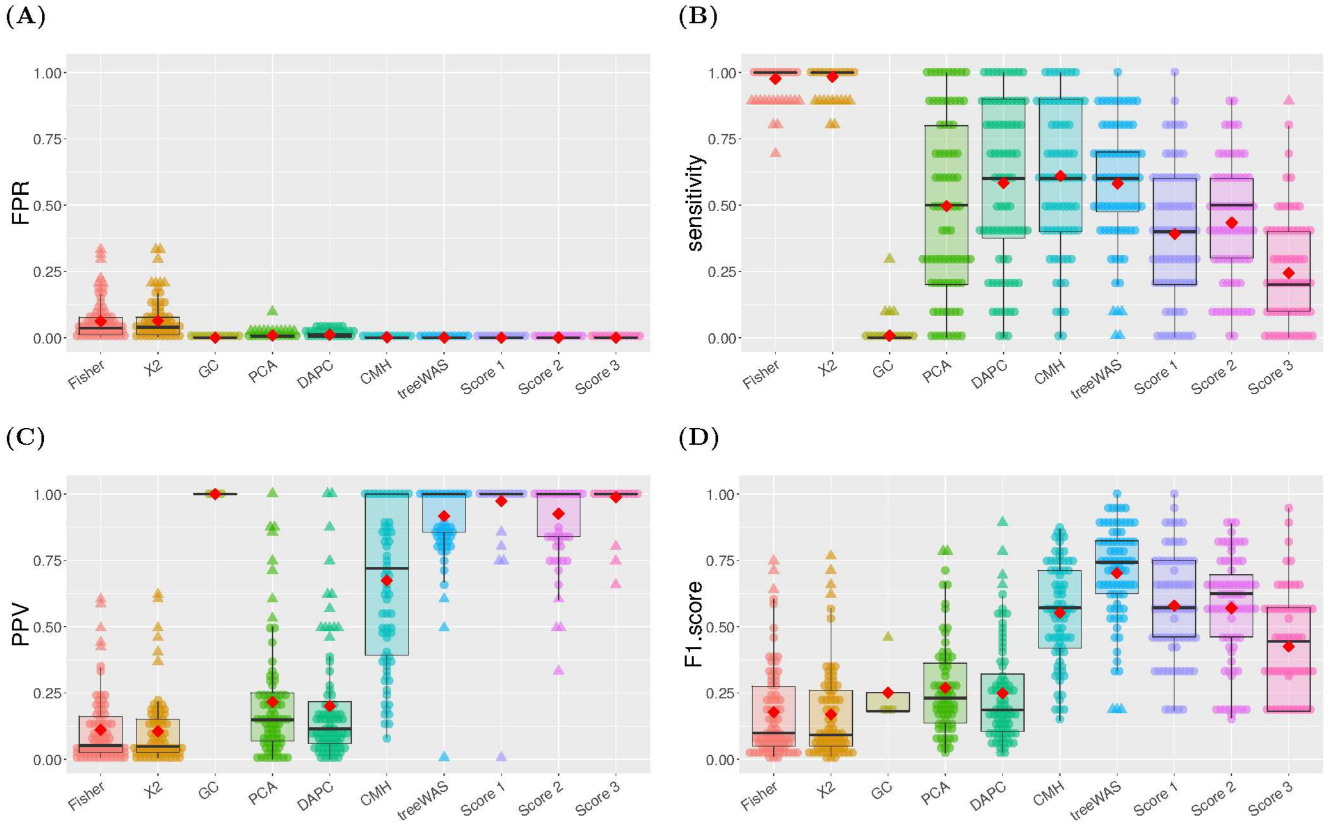


**Figure 4.15. Performance on accessory genome data.** Performance on simulated accessory gene presence-or-absence datasets ($N = 80$) is summarised along four metrics of evaluation. Box plots display the median and interquartile range, red diamonds indicate the mean, and each dot shows the result for one dataset. **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

In Figure 4.15A, our approach consistently achieves near-zero FPR, finding a median of zero and a mean of 0.6 false positives in the pooled findings of its three association scores. Sensitivity in Scores 1-3 is relatively low in Figure 4.15B, but not unexpectedly so, given the high rate of confounding recombination simulated among non-associated loci. Moreover, with each score contributing to our collective set of findings, the sensitivity of treeWAS is brought into line with that of our ancestry-aware competitors. The relative PPV values of all seven GWAS approaches resemble the distribution observed in Figure 4.10 for Set C. Altogether, we can see from the F1 scores in Figure 4.15D that the overall performance of treeWAS continues to exceed the performance of alternative approaches. These results encourage us to expect that treeWAS will achieve similar power and specificity, whether it is used to perform GWAS on core SNP datasets or accessory gene presence-or-absence data.

### 4.7.2 Performance by genome size

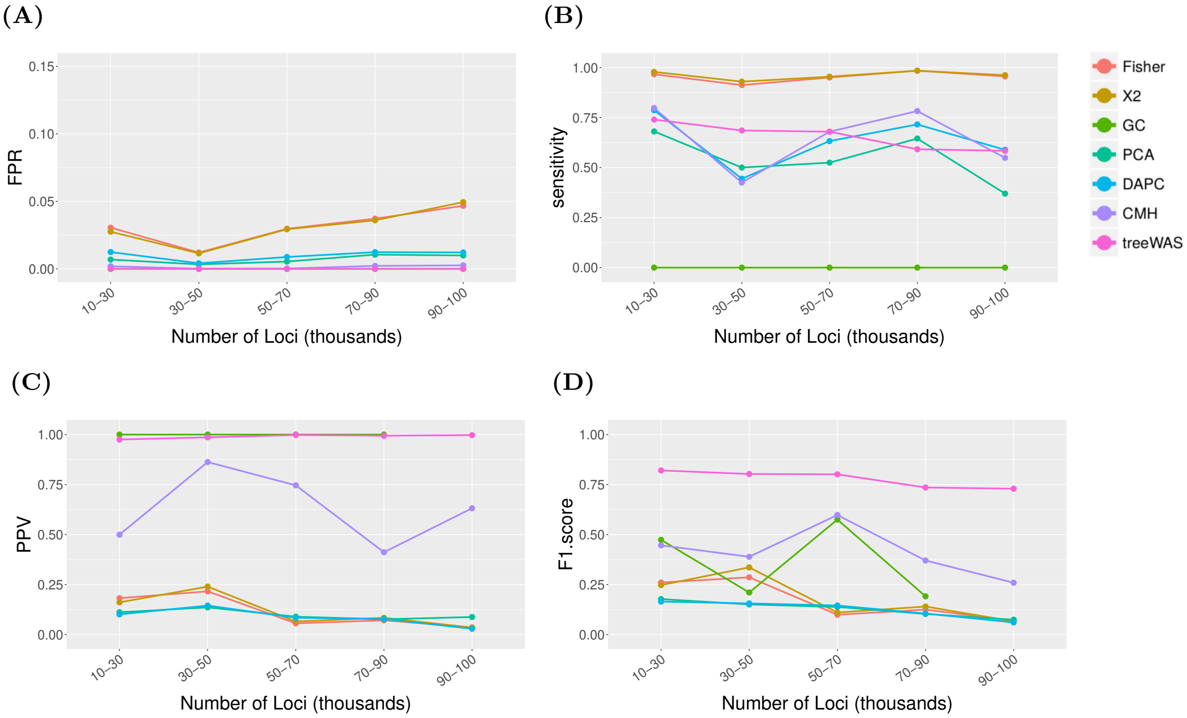


**Figure 4.16. Performance by number of genetic loci.** The interquartile mean performance of treeWAS and six comparator methods is presented as a function of the number of genetic loci simulated per dataset ($N = 100$), where $N_{loci} \in \{10000, 10909, ..., 99091, 100000\}$. Performance statistics are aggregated over 20,000-loci intervals and presented along four metrics: **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

The results presented in Figure 4.16 suggest that the number of genetic variables in a dataset may have a modest effect on the performance of GWAS methods. As $N_{loci}$ increases along the x-axes in Figure 4.16, while the number of associated sites and all other parameters remain unchanged, the ratio of associated to non-associated sites declines, and the genome-wide signal to noise ratio decreases with it. In Figure 4.16A, the FPR of the uncorrected Fisher and $X^2$ tests appears to increase with $N_{loci}$, with a linear increase between $N_{loci} = 30,000$ and $N_{loci} = 100,000$. PCA, DAPC, and, to a greater extent, the CMH test, dampen this trend by moving FPR closer to zero at each point. But, only treeWAS and the GC approach keep FPR at zero at all $N_{loci}$.

Figure 4.16B reveals that, in achieving zero FPR, GC also sacrifices all discovery power. The sensitivities of most other methods vary noisily, although all population-aware approaches appear to lose some power as $N_{loci}$ increases. Unlike the variable PCA, DAPC, and CMH approaches, treeWAS uses $N_{loci}$ to inform its estimation of the null distribution. So, although it experiences a similar decrease with increasing $N_{loci}$, the sensitivity of treeWAS displays greater stability and undergoes more incremental, linear change than competing methods, indicating a more controlled underlying process.

Our approach stands out most distinctly in Figure 4.16C. The uncorrected tests and both dimension reduction methods, which achieve little precision to begin with, lose precision as $N_{loci}$ increases, with the uncorrected tests experiencing the sharpest decline in PPV. The more moderate PPV of the CMH test varies noisily and shows no clear trend corresponding to the change in $N_{loci}$. We do, however, observe that CMH precision maintains an approximately inverse proportional relationship to its sensitivity. Meanwhile, at all $N_{loci}$, treeWAS keeps its precision at a maximum. In fact, our PPV even rises slightly with genome size. We attribute this to the fact that, as $N_{sim}$ increases proportionally with $N_{loci}$, treeWAS can make ever more refined estimates of the location of the significance threshold in larger datasets.

The F1 scores in Figure 4.16D show that no alternative approach is able to offer a combination of sensitivity and precision that is more effective than treeWAS. Competing approaches obtain lower overall performance than treeWAS when genomes are small. As $N_{loci}$ increases, the F1 scores of comparator methods only decline further, displaying unpredictable variation along the way. The F1 scores of treeWAS vary only marginally, between 0.82 and 0.74. Therefore, across all values of $N_{loci}$ explored, treeWAS maintains strong overall performance with limited variation.

### 4.7.3 Performance by sample size

Figure 4.17 reveals that the number of individuals in a dataset can have a substantial impact on the performance of all GWAS methods examined. The steady rise in FPR experienced by the uncorrected Fisher and $X^2$ tests in Figure 4.17A indicates that, all else being equal, increasing $N_{ind}$ in our simulated datasets increases the probability of spurious association. Indeed, while the expected phenotypic $N_{sub}$ and genotypic $N_{sub}$ distribution are controlled by fixed parameters, population stratification should increase with $N_{ind}$, as each substitution is inherited by an ever-larger clade of descendants. Because ours is the only approach that accounts for each of these factors, it is the only method equipped to make an informed assessment about the changing probability of spurious association in this context. Figure 4.17A confirms that only treeWAS and GC prevent FPR from increasing with $N_{ind}$; although, once again, GC accomplishes this by eliminating all true and false positive findings. All other methods reduce FPR below the

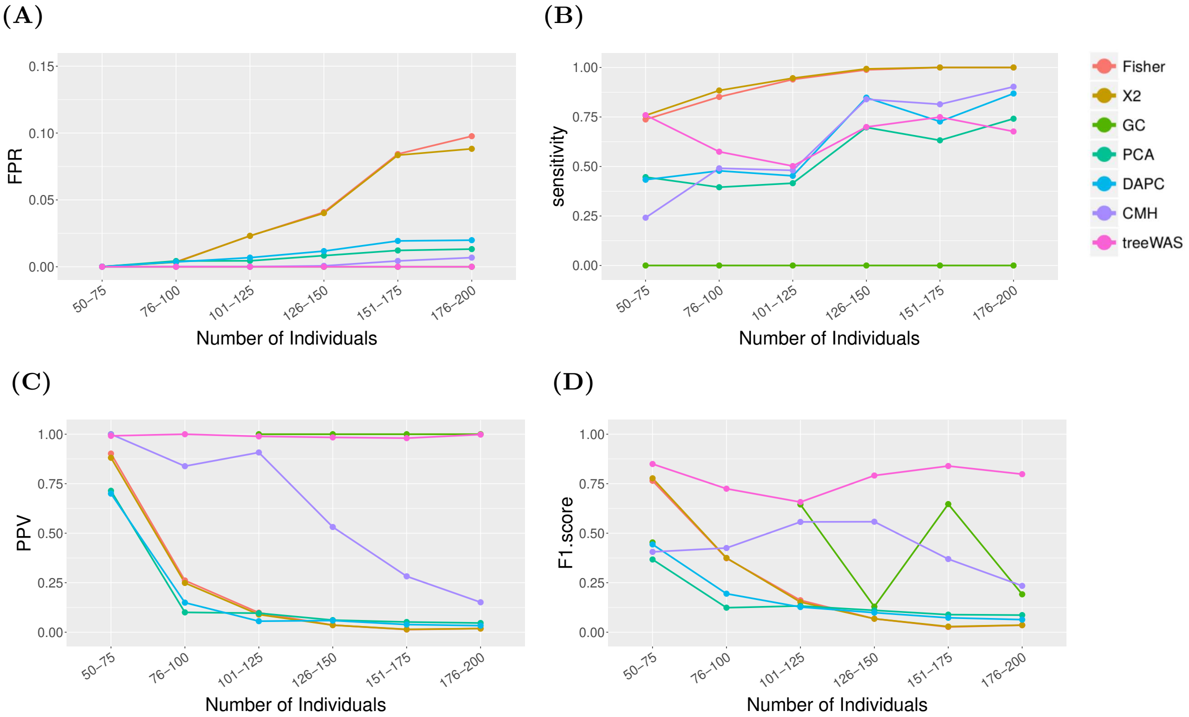


**Figure 4.17. Performance by number of individuals.** The interquartile mean performance of treeWAS and six comparator methods is presented as a function of the number of individuals simulated per dataset ($N = 100$), where $N_{ind} \in \{50, 52, ..., 198, 200\}$. Performance statistics are aggregated over 20,000-loci intervals and presented along four metrics: **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

uncorrected baseline level. But, DAPC, PCA, and CMH each fail to prevent FPR from increasing with $N_{ind}$.

Figure 4.17B indicates that, alongside FPR, sensitivity increases with $N_{ind}$. The smooth, increasing trend displayed by the uncorrected tests is replicated by all population-aware approaches, except GC, though they introduce more noise and begin with only half the power of population-naive approaches where $N_{ind} \in [50, 75]$. It is not surprising that larger sample sizes increase the power of GWAS approaches. It is, however, notable that treeWAS power deviates from all ancestry-aware alternatives in smaller samples, where $N_{ind} \leq 100$. At the smallest sample size in Figure 4.17B, for example, our approach is able to match the sensitivity of both uncorrected tests, tripling the sensitivity of the CMH test. Where $N_{ind} > 100$, treeWAS joins comparator methods in experiencing a similar positive sensitivity trend as $N_{ind}$ increases.

Figure 4.17C shows a stark difference between the PPV behaviour of treeWAS and competing approaches. Ours is the only method to maintain high precision (PPV $\simeq 1$) as $N_{ind}$ increases. In the five other sensitive GWAS methods, by contrast, the proportion of true positive findings approaches zero as $N_{ind}$ increases. CMH PPV remains high where $N_{ind} \leq 125$ but declines precipitously thereafter, falling to near zero with other methods as $N_{ind}$ approaches 200. This has serious implications for our assessment of the CMH test, especially because our evaluation thus far has been based on simulations with $N_{ind} = 100$. These findings suggest that the performance of our closest competitor may not remain near that of treeWAS when the CMH test is applied to samples with $N_{ind} > 125$, which is already smaller than many samples examined in microbial GWAS.

Ultimately, the F1 scores in Figure 4.17D demonstrate that treeWAS is able to achieve the strongest overall performance at all $N_{ind}$. Moreover, driven by the divergence in PPV, this performance advantage increases as sample size grows. Our empirically-parameterised GWAS method offers a clear benefit in this regard, one which we anticipate will only increase in relevance as microbial GWAS progresses and the average sample size increases.

### 4.7.4 Computational time

Our final aim in performing this simulation study was to provide an estimate of the amount of time required to run our approach on a standard computer and to assess the scalability of treeWAS with increasing dataset size. We applied our approach to two panels of simulated datasets: first, varying the number of individuals, as above, such that $N_{ind} \in \{50, 52, ..., 198, 200\}$ $(N = 100)$; and, second, varying the number of genetic loci, such that $N_{loci} \in \{10000, 11919, ..., 198081, 200000\}$ $(N = 100)$. Each dataset was analysed on a standard computer with default arguments, and the amount of computational time required to complete the analysis was recorded. In Figure 4.18, we plot treeWAS run time as a function of $N_{ind}$ and $N_{loci}$.

Figure 4.18A shows that, for genetic datasets with 10,000 sites and up to 200 individuals, the treeWAS R package can execute our GWAS approach in under thirty seconds. Meanwhile, Figure 4.18B shows that our approach can be performed in under four minutes for any dataset with 100 individuals and up to 200,000 sites. These results also show that treeWAS run time scales approximately linearly with both $N_{ind}$ and $N_{loci}$. Hence, we can extrapolate from the line of best fit to estimate treeWAS run times for a wider range of hypothetical datasets. Even in large datasets, for example, with $N_{ind} = 1,000$ and $N_{loci} = 100,000$, we would still expect treeWAS to complete the analysis in under 20 minutes on a standard personal computer.

The line of best fit in Figure 4.18A suggests that treeWAS will take approximately 12.2

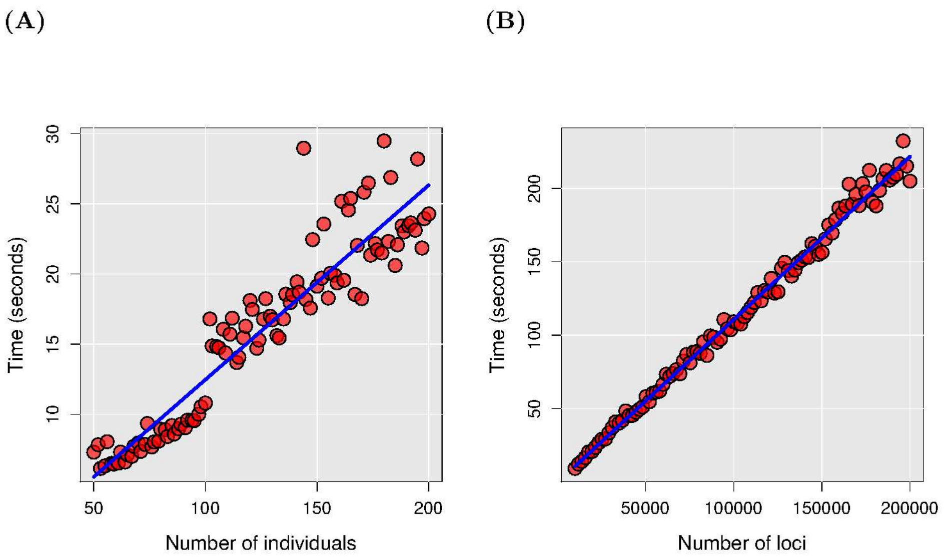


**Figure 4.18. Computational time to run treeWAS.** The amount of time, in seconds, required to run treeWAS is plotted for simulated datasets of increasing size. One dimension is varied in each plot, while all other parameters are fixed at default values. **A:** treeWAS run time is plotted as a function of sample size, for $N_{ind} \in [50, 200]$. **B:** treeWAS run time is plotted as a function of genome size, for $N_{loci} \in [10000, 200000]$.

seconds to analyse a dataset of the size simulated in Sets A, B, and C, where $N_{ind} = 100$ and $N_{loci} = 10,000$. Run time analysis performed throughout the above simulation study confirms this finding with a similar result, showing that treeWAS required, on average, 13.1 seconds per analysis. GWAS was only performed in less time by the uncorrected Fisher's exact test (mean = 9.4 seconds), though a speed similar to treeWAS was achieved by the CMH test (mean = 13.8 seconds). PLINK [136] provided comparably inefficient implementations of relatively straightforward GWAS methods, taking 25% longer than treeWAS to run the uncorrected $X^2$ test (mean = 16.2 seconds) and taking 2.6 times as long to perform GWAS with Genomic Control (mean = 34.2 seconds). Finally, although regression-based GWAS methods have been praised for their computational speed [183], we found that treeWAS was, on average, 3.3 times faster than the PCA-based approach (mean = 42.7 seconds) and 7.7 times faster than the DAPC-based approach (mean = 1 minute, 41.3 seconds), including model validation and clustering steps. We can, therefore, conclude that treeWAS required the least computational time to perform GWAS of any ancestry-aware approach examined in this simulation study. These results provide a valuable demonstration of the computational efficiency and scalability of our phylogenetic GWAS method, as implemented in the treeWAS R package.

### Concluding remarks

In setting out to perform this simulation study, we developed multiple simulation frameworks to provide variable parameter and association landscapes for these analyses. These methods of data generation allowed us to determine, among other things, that our approach could achieve optimal performance with a parsimony based method of ancestral state reconstruction, and that Score 3 would offer a better combination of performance and efficiency if we excluded the branch length term. We also made an evidence-based choice of threshold selection mechanism to be implemented in the treeWAS R package and in subsequent empirical analyses. Having implemented these simulation tools in the treeWAS R package, we hope they will be useful to others, for example, in facilitating additional sensitivity studies, performance assessments, or comparative analyses of our own or other GWAS methods in the future.

We had hoped that our method would perform well in these applications to simulated data, but we had nevertheless anticipated that other GWAS approaches would have the upper hand in some circumstances or parameter ranges. We were therefore very pleased with our results, as they demonstrated that our phylogenetic GWAS method was able to achieve the strongest performance of any GWAS method examined in nearly all of the above analyses. In terms of overall performance, our approach was able to dominate all competitors in all three simulation sets and across all parameter ranges explored. In exploring multiple parameter ranges, we found that our approach was both robust and flexible, maintaining stronger performance than other GWAS methods as the size, scope, and complexity of datasets varied.

Our results revealed that treeWAS was able to achieve unmatched precision, as well as comparable sensitivity to alternative approaches by pooling our three association scores. We were especially impressed with the power of our homoplasy-counting Score 2. However, we were somewhat surprised by the comparatively modest performance of Score 1 and the relatively poor performance of Score 3. We acknowledge that some of this performance differential may be attributable to the nature of the simulations examined in this chapter. We will therefore re-examine the relative contributions of our three scores when we complete the analyses of empirical data examined in the next chapter.

# Chapter 5

# Applications to Empirical Data

**Acknowledgements** This chapter includes some written and illustrated content that appeared, in a related form, in the publication of our method [155]. Xavier Didelot guided us in our search for interesting phenotypes and useful datasets to analyse and provided feedback during the writing of our publication and the chapter below. We would like to extend our gratitude to the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project, for generously providing us with access to their computing cluster. We are grateful to Daniel Wilson and David Aanensen for providing useful feedback during the development of this work. We also thank the two anonymous reviewers who provided helpful suggestions that we incorporated in our paper.

## 5.1 Introduction

In this chapter, we present the results of multiple applications of our GWAS approach to empirical data. We aim to demonstrate that treeWAS can accurately identify trait-associated loci in biological sequence data and to confirm that the power and precision observed in our simulation study translates to empirical analyses. We analyse three datasets from *N. gonorrhoeae* and *N. meningitidis*, to examine the performance of our GWAS method in organisms known to display both clonal population structure and homologous recombination [72, 298]. Our association testing method is applied to both accessory and core genome variation. Phenotypes are examined as both binary and continuous variables, where appropriate, to further explore the versatility of our approach.

First, we analyse a previously-characterised phenotype of resistance to the antibiotic drug cefixime in a *N. gonorrhoeae* dataset ($N = 200$) that was published by Grad *et al.* [96]. This allows us to appraise our results by making direct comparisons to the findings of the authors. Second, we use treeWAS to test for associations with a related pencillin resistance phenotype, in serogroup B *N. meningitidis* isolates ($N = 171$). To evaluate our findings in this original analysis, we refer to the available literature, in which the genetic basis of penicillin resistance is well described. Third, we apply treeWAS to a more complex invasive disease phenotype in serogroup C *N. meningitidis* isolates ($N = 129$). Despite the public health relevance of invasive meningococcal disease, the genetic basis of this pathogenic phenotype remains incompletely-characterised [47]. We hope that treeWAS can confirm associations previously-identified in the literature, and that it may also identify novel candidate loci whose link to meningococcal virulence has yet to be established. All analyses were run on a standard laptop computer, with a 4-core Intel processor, a CPU clock speed of 2.60GHz, and 16 GB of RAM available.

In applying our approach to these three datasets, we aim to demonstrate that the novel design features implemented in treeWAS allow our method to identify associations effectively in empirical analyses. Given the results of our simulation study, we hope to find few spurious findings and to confirm that our approach accounts appropriately for the confounding factors present in real datasets. As we encounter unfamiliar association landscapes, shaped by mutation, recombination, and selection, we hope to find that our multiple measures of association equip treeWAS with the power to confirm known associations and detect novel relationships between genotype and phenotype.

## 5.2 Cefixime resistance in Neisseria gonorrhoeae

We begin by applying our GWAS method to a previously-characterised phenotype in a published empirical dataset, so that we can validate our results and make intial estimates of the performance of our method. We examined cefixime resistance in 200 *N. gonorrhoeae* genomes drawn from the genomic epidemiology study of Grad *et al.* [96] (see Appendix, Tables A.1, A.2, A.3). *N. gonorrhoeae* are obligate human pathogens that typically colonise the genito-urinary tract and spread easily through sexual contact between hosts [299]. Recent increase in both gonorrhoea incidence and antibiotic resistance brings renewed public health relevance to the study of *N. gonorrhoeae* [44].

Cefixime is a member of the cephalosporin class of $\beta$-lactam antibiotics. It is one of only two remaining effective first-line gonorrhoea treatments [300]. Like penicillin, cefixime uses $\beta$-lactams to bind gonococcal penicillin binding proteins (PBPs). This interferes with the essential contribution that PBPs make to peptidoglycan synthesis and cell wall biogenesis, which prevents the survival of *N. gonorrhoeae* bacteria. Over time, however, cefixime susceptbility has declined in *N. gonorrhoeae* [301]. The previous analysis performed by Grad *et al.* [96] attributes cefixime resistance primarily to *penA* alleles that encode PBP variants capable of escaping $\beta$-lactam binding. We set out to determine whether our GWAS approach could confirm the association status of these previously-identified cefixime-associated loci. We also wish to assess whether the precision, power, and efficiency observed in our simulation study remains consistent when treeWAS is applied to empirical data.

We downloaded aligned whole-genome sequence data for the 200 *N. gonorrhoeae* isolates from the *Neisseria* Bacterial Isolate Genome Sequence Database (BIGSdb) [302]. We assembled two genetic datasets to be analysed by our GWAS method:

1. The **core SNP** dataset contains all polymorphic loci found in the core genome, amounting to 23,932 binary SNPs.
2. The **accessory gene** dataset notes the presence-or-absence status of 3,036 genes.

The core and accessory genome datasets were produced with the Basic Local Alignment Search Tool (BLAST) [303]. BLAST uses a sliding window approach was to identify the locations of probable genes, evaluating the protein-coding potential of the transcribed genome, in 20-nucleotide sequence fragments ("words"). The repeated observation of particular genomic features across multiple sequences may, further, indicate genetic homology [304]. We set a 95% threshold to distinguish core from accessory genome variation, such that any gene or intergenic region that is present in $\geq$ 95% of the isolates in a given sample is included in the core genome dataset [78, 79]. Conversely, any gene absent from 5% of sampled genomes is added as a column of the accessory genome dataset. When testing for association with the phenotype, core genome variation is examined at the level of individual nucleotides. Accessory genome variation is evaluated with respect to the presence or absence of the gene as a unit.

Phenotypic information was extracted from the published meta-data [96]. We analysed the cefixime resistance phenotype in two ways:

1. The **binary phenotype** categorised isolates as "sensitive", if MIC $\leq 0.25\mu$g/mL, or "resistant", if cefixime MIC $> 0.25\mu$g/mL.
2. The **continuous phenotype** was defined by ranking cefixime MIC values.

The rank-transformed phenotype was more uniformly distributed and contain more useful information than the original MIC values (see Figure A.1). This allowed our association tests to give greater weight to the differences observed between a larger number of individuals.

Neisseria provide a representative example of semi-clonal organisms [305]. In this *N. gonorrhoeae* dataset, the recombination to mutation ratio was $r/m = 1.9$. This amount of recombination will warp the inferences of traditional phyogenetic methods [1]. Moreover, the presence of large, resistance-associated genomic islands will have negative implications for both ancestry inference and association inference, if inappropriately addressed [306]. Yet, the underlying clonal relationships likewise remain strong enough to introduce bias in association inference and clear enough to permit the use of recombination-aware tree-building methods. We reconstructed the phylogenetic tree from whole-genome sequence data. An initial parsimony tree was estimated with the dnapars algorithm in PHYLIP [194], and we provided this as an input to ClonalFrameML [221], which we used to infer the clonal genealogy while accounting for recombination (Figure 5.1A).

We applied treeWAS to the accessory and core genome datasets, and to both the binary resistance and continuous MIC phenotypes. No accessory genes were found to be significantly associated with either the binary or continuous cefixime phenotypes. treeWAS did, however, identify many core SNPs as significantly associated with both forms of phenotypic data. The core SNP GWAS was performed by treeWAS in just over 2 minutes. The accessory genome analysis was completed in 15 seconds.

### 5.2.1 Identifying associations with cefixime resistance

The application of treeWAS to the set of core SNPs resulted in the identification of 132 SNPs significantly associated with the binary cefixime resistance phenotype (Table 5.1). Of these, 129 SNPs were located in the NEIS1753 (*penA*) gene, indicating that changes in PBP2 affect resistance. This confirms the primary finding of Grad *et al.* [96]. We also found three significant SNPs in the neighbouring NEIS1751 (*murE*) gene. The previous hypothesis-driven analysis did not investigate this locus. Yet, additional evidence implicates *murE* in the same cell wall biosynthesis pathways as *penA* [307]. Experimental analyses in *N. gonorrhoeae* have shown that cefixime resistance correlates with variation in *murE* and that *murE* regularly accompanies *penA* in resistance-associated recombinant fragments [308, 309]. In this analysis, all three of the association scores in treeWAS found associated loci within the same two genes. The findings of Scores 1-3 and the overlap between these findings, however, varies across subsequent empirical analyses, as discussed in Section 5.5.

| Gene | N.SNPs | SNPs | Gene product |
|---|---|---|---|
| NEIS1751 (*murE*) | 3 | 1125522.a 1126060.c 1126111.c | UDP-N-acetylmuramoyl-alanyl-D-glutamate–2, 6-diaminopimelate ligase |
| NEIS1753 (*penA*) | 129 | 1126892.g 1126940.g 1126976.g 1127054.a 1127077.t 1127116.a 1127120.g 1127177.a 1127225.t 1127234.c 1127238.c 1127240.g 1127258.c 1127264.c 1127267.c 1127303.c 1127306.c 1127312.c 1127315.c 1127333.c 1127339.c 1127354.c 1127399.t 1127411.c 1127414.g 1127429.c 1127434.c 1127444.c 1127450.c 1127453.t 1127461.g 1127468.c 1127470.g 1127487.c 1127591.g 1127619.c 1127630.g 1127792.g 1127795.t 1127818.a 1127819.t 1127828.g 1127833.g 1127834.g 1127835.c 1127836.c 1127837.a 1127849.g 1127852.g 1127867.c 1127879.g 1127886.c 1127900.c 1127909.c 1127911.c 1127912.a 1127913.c 1127918.g 1127921.c 1127929.t 1127930.t 1127934.c 1127940.c 1127951.g 1127954.g 1127957.g 1127969.g 1127970.c 1127978.a 1127982.a 1127984.a 1127985.t 1127992.a 1127993.a 1127996.g 1127999.c 1128002.g 1128003.c 1128004.g 1128005.c 1128006.g 1128008.g 1128011.a 1128015.a 1128017.t 1128018.c 1128032.a 1128035.c 1128038.g 1128039.c 1128041.c 1128041.g 1128047.c 1128056.g 1128059.g 1128068.g 1128071.c 1128080.t 1128089.a 1128089.t 1128107.c 1128110.a 1128111.c 1128116.c 1128119.t 1128122.c 1128122.t 1128125.c 1128128.t 1128131.g 1128134.c 1128135.a 1128146.c 1128148.g 1128149.t 1128152.g 1128152.t 1128155.t 1128170.c 1128173.c 1128182.c 1128182.g 1128206.c 1128212.a 1128221.c 1128222.c 1128230.t 1128234.g 1128239.c | penicillin-binding protein 2 |

**Table 5.1. SNPs associated with cefixime resistance in *N. gonnorrhoeae*.** These 132 SNPs were identified as significantly associated with the binary cefixime resistance phenotype when treeWAS was applied to core SNPs from 200 *N. gonorrhoeae* isolates.

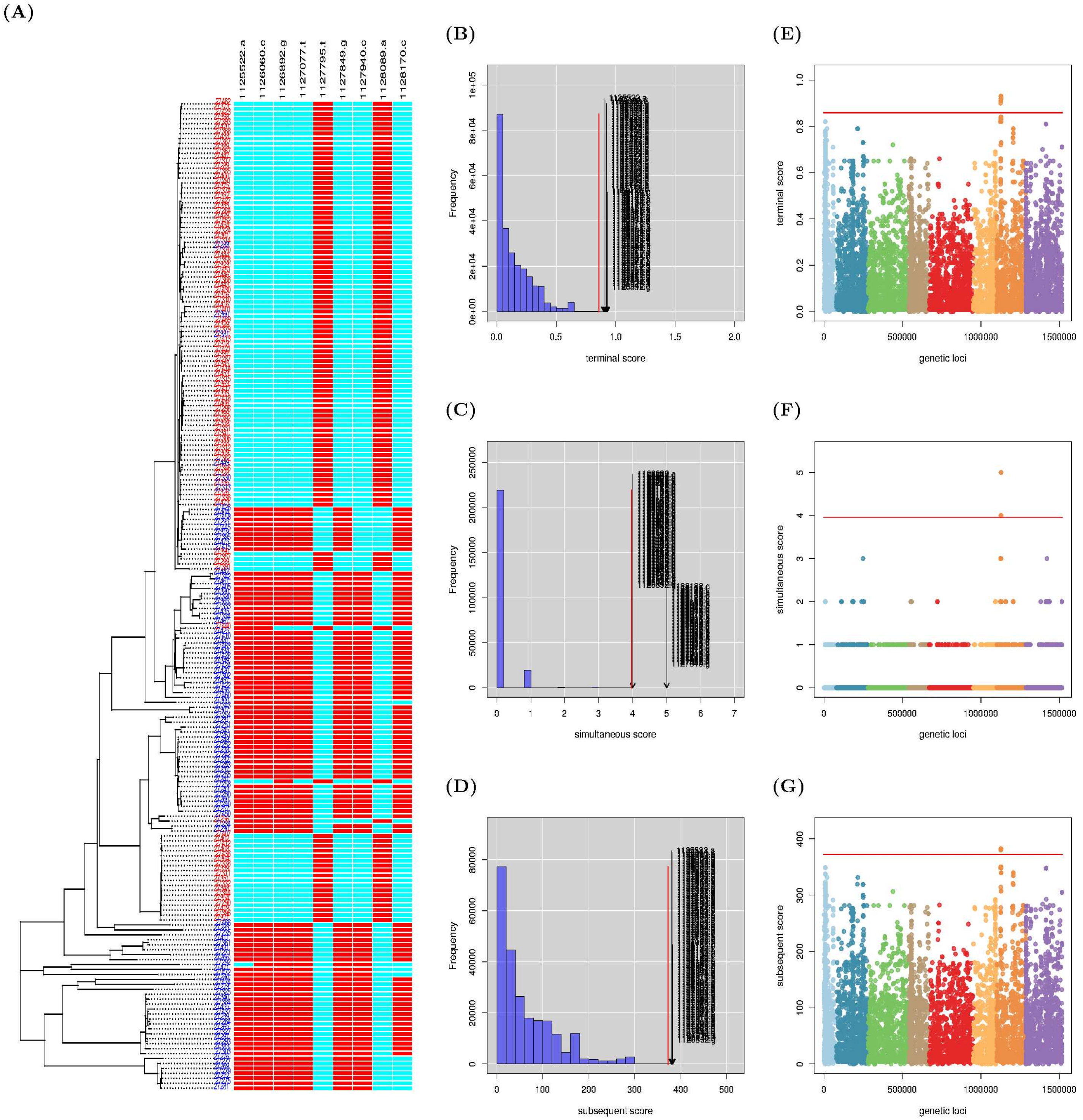


**Figure 5.1. Application to cefixime resistance in *N. gonorrhoeae* core SNPs.** treeWAS identified 132 SNPs associated with the binary cefixime resistance phenotype, when applied to core SNPs from a sample of 200 *N. gonorrhoeae* isolates. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = susceptible; red = resistant). At right, an alignment of the nine unique SNPs column patterns (blue = allele 0; red = allele 1) that were observed among the 132 significant SNPs. **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red line), above which associated SNPs are indicated. **E-G:** Manhattan plots for (E) Score 1, (F) Score 2, (G) Score 3 showing association score values for all SNPs, a significance threshold (red line), above which points indicate significant associations.

Visualising our findings along the phylogeny in Figure 5.1A, we can see that ancestral reconstruction has enabled our approach to identify multiple origins of the cefixime resistance phenotype. It also reveals the concomitant introduction of polymorphisms in *penA* and *murE*. The physical proximity and strong correlation among significant SNPs (see Figure A.3), and the near-uniformity (98% identity) among significant loci in resistant isolates indicate strong LD among identified sites. Considering the repetition of genotypic patterns across different clades, we can infer that recombination has introduced existing mosaic *penA* variants into the genomes of ancestral isolates. This agrees with the mechanism described in the literature and identified in the previous study [96, 310]. Notably, although our findings clearly indicate that resistance-associated polymorphisms do not emerge independently through point mutations, our approach has nevertheless been able to trace cefixime resistance to variation within known associated genes by modeling all mutation and recombination events at the level of individual substitutions. Indeed, treeWAS has been able to distinguish significant SNPs, despite strong LD across the recombinant region. These SNPs represent promising candidates for experimental analyses, which may reveal a functional link to resistance. They may also be useful biomarkers, which, in rapid PCR- or sequence-based diagnostic tests can dramatically reduce the time required to assess resistance, compared to existing laboratory assays relying on the growth of bacterial cultures.

Considerable population stratification was generated by recombination. The introduction of linked blocks of resistance-associated polymorphisms produced correlation between ancestral clades and phenotypic clusters, as seen in Figure 5.1A. In fact, these population clusters may themselves be shaped by antimicrobial selective pressures [306]. We inferred 11 phenotypic substitutions along the tree and estimated the genotypic $N_{sub}$ distribution as in Figure 5.2. Accounting for the strong population stratification, and the proximity of genotypic and phenotypic $N_{sub}$ values, our approach inferred that high levels of association could occur by chance alone in a large proportion of the loci in this dataset. Especially because these conditions pose such a challenge to GWAS anlayses, we are pleased to report that our results (Figures 5.1E-G) nonetheless show that treeWAS is able to distinguish 132 well-supported associations within the mosaic *penA* region, without permitting the identification of spurious findings at other, population-stratified loci, in any association score. Our approach also thus rejected any significant association between cefixime resistance and variation in the *mtrR*, *porB*, or *pilQ* genes, corroborates additional findings from the Grad *et al.* study [96].

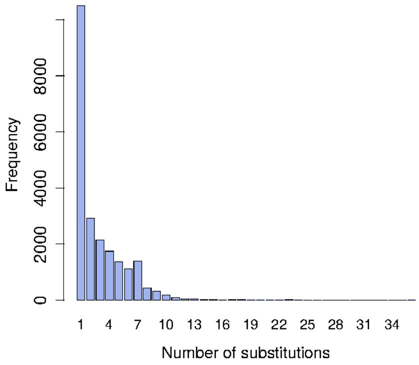


**Figure 5.2. Homoplasy distribution (*N. gonorrhoeae*, SNPs).**

Having identified the candidate loci in Table 5.1, additional laboratory work would be needed to confirm the structural and functional role of all significant *murE* and *penA* SNPs on cefixime resistance. Experimental studies have investigated some of the polymorphisms in *penA*, and our results contain many of these loci [299, 300, 306]. For example, site-directed mutagenesis has been used to introduce SNPs 1128089.a/t, 1128107.c, 1128125.c, and 1128230.t into *N. gonorrhoeae* genomes. This results in amino acid substitutions F504L, A510V, A516G, and P551S, and leads to a five-fold reduction in $\beta$-lactam binding [311]. A role in cefixime resistance has also been experimentally confirmed for 1128071.c (T498) [312], 1128080.t (A501V) [300], 1128122.c/t (I515V) [311], and 1128212.a (G545S) [313]. Interestingly, some of these sites have been found to impact cefixime MIC only in the presence of other residues [312]. Yet, thanks to selective pressures acting on epistatically-interacting sites, our site-by-site approach has been able to identify these loci. Although further laboratory investigation is needed to evaluate the remaining SNPs in Table 5.1, our ability to confirm the previously-identified *penA* and *murE* genes, and known functional SNPs within these regions, provides strong support for the performance of our approach.

Overall, through its use of the clonal framework to simulate a neutral distribution of $N_{sub}$ across the tree, our approach has been able to demonstrate robust control over the confounding effects of population stratification, genetic linkage, mutation and recombination. One limitation of these stringent efforts to reject spurious findings is that our approach may have rejected additional SNPs, likely in *penA* and *murE*, whose weaker relationship with the phenotype may still have been genuine. On the other hand, with only nine unique SNP column patterns among the 132 SNPs identified, we do not have enough information to separate epistatic effects from LD within the mosaic *penA* gene [314]. Finding small sets of perfectly-linked SNPs is not surprising, nor particularly problematic. In comparison, human GWAS requires fine mapping to hone in on causal SNPs that may be linked to a marker SNP detected kilobases away [315]. To confirm and refine our findings, we could repeat the analysis with a larger sample size, perform a meta-analysis, or interrogate SNPs experimentally.

### 5.2.2 Results from other GWAS methods

| | | **Found by treeWAS** | | **In *murE-penA*** | |
|---|---|---|---|---|---|
| **GWAS method** | **Total** | **Yes** | **No** | **Yes** | **No** |
| treeWAS | 132 | 132 | 0 | 132 | 0 |
| Fisher | 4,001 | 132 | 3,869 | 304 | 3,697 |
| $X^2$ | 3,688 | 132 | 3,556 | 302 | 3,386 |
| GC | 0 | 0 | 0 | 0 | 0 |
| PCA | 384 | 109 | 275 | 276 | 108 |
| DAPC | 382 | 109 | 273 | 276 | 106 |
| CMH | 403 | 119 | 284 | 314 | 89 |

**Table 5.2. Comparing associations with cefixime resistance by GWAS method.**

For comparative purposes, we repeated the analysis of cefixime resistance with the GWAS methods examined in our simulation study. Table 5.2 suggests that similar relative levels of performance are obtained in empirical analyses, even with this straightforward, highly heritable phenotype. While all methods except GC found most or all of the SNPs identified by treeWAS, they also found large numbers of additional loci. Aside from the null results of GC, alternative approaches deemed many other sites in *murE-penA* "significant", though many of these are likely atttributable to LD alone. Furthermore, each method found many loci with no known connection to the phenotype, adding many probable false positives.

The evidence suggests that our method was able to achieve better precision by using simulations to evaluate the impact of ancestry at all levels of the population structure. Meanwhile, the performance of competing approaches was undercut by false positive findings Owing to the phenotypic uniformity of some sub-populations, the CMH test was only able to account for $k = 4$ clusters in this dataset, leaving correlations between phenotype and clade were left unchecked in other sub-populations. PCA and DAPC more thoroughly accounted for the population structure, with $k = 10$. Yet, neither clusters nor PCs adjusted for the true clonal ancestry of the sample, as widespread recombination was not accounted for. Altogether, by accounting for the clonal genealogy, recombination, mutation, and the phenotypic distribution, our approach more effectively counteracted confounding factors in this empirical dataset. We also balanced stringency with power, by drawing on our three association scores. Our GWAS method thus achieved similar sensitivity to competing approaches, while excluding volumes of their likely false positive findings.

### 5.2.3 Identifying associations with cefixime MIC

Unlike many existing GWAS methods, our appraoch makes it possible to identify associations with continuous phenotypes. This allowed us to analyse the underlying cefixime MIC phenotype from which the binary categories above were derived. In our analysis of the continuous rank-transformed MIC phenotype, treeWAS identified 222 significant SNPs. All significant SNPs fell in *penA* and *murE*, as in the binary analysis, with strong LD between significant sites (see Figure A.4). Although additional laboratory analyses would be needed to support a causal association at individual SNPs in Table 5.3, we have already presented evidence confirming that both genes are functionally related to cefixime resistance [299, 306–308]. Indeed, variants of *penA* have been found to induce 8- to 500-fold changes in cefixime MIC in *N. gonorrhoee* [300].

Our analyses of the binary and continuous phenotypes were both performed on the same genetic dataset, using the same phylogenetic model of population structure. And, while the binary analysis was successful, the analysis of the continuous phenotype provided a distinct perspective and enabled a more detailed examination of the data. Whereas

| Gene | N.SNPs | SNPs | Gene product |
|---|---|---|---|
| NEIS1751 (*murE*) | 6 | 1125247.a 1125258.a 1125266.c 1125522.a 1126060.c 1126111.c | UDP-N-acetylmuramoyl-alanyl-D-glutamate–2, 6-diaminopimelate ligase |
| NEIS1753 (*penA*) | 216 | 1126892.g 1126901.t 1126940.g 1126964.c 1126976.g 1127009.t 1127015.c 1127054.a 1127077.t 1127116.a 1127120.g 1127177.a 1127195.c 1127196.c 1127199.t 1127203.g 1127204.c 1127206.a 1127210.c 1127213.c 1127222.t 1127225.t 1127234.c 1127238.c 1127240.g 1127258.c 1127264.c 1127267.c 1127282.a 1127303.c 1127306.c 1127312.c 1127315.c 1127333.c 1127339.c 1127354.c 1127399.t 1127411.c 1127414.g 1127429.c 1127434.c 1127444.c 1127450.c 1127453.t 1127461.g 1127468.c 1127470.g 1127487.c 1127492.c 1127498.t 1127510.t 1127522.t 1127528.g 1127531.a 1127534.c 1127537.a 1127543.c 1127544.g 1127545.t 1127546.g 1127552.g 1127555.g 1127565.g 1127567.g 1127574.a 1127575.c 1127576.c 1127576.g 1127580.t 1127581.t 1127582.g 1127584.a 1127588.a 1127589.c 1127590.g 1127591.g 1127592.c 1127594.g 1127600.g 1127602.a 1127609.t 1127618.a 1127619.c 1127622.t 1127625.c 1127630.g 1127636.c 1127645.t 1127651.c 1127654.c 1127655.t 1127681.g 1127687.g 1127699.c 1127702.a 1127705.c 1127711.g 1127717.g 1127718.c 1127719.g 1127720.t 1127723.c 1127724.g 1127725.g 1127726.c 1127727.g 1127729.c 1127730.g 1127741.t 1127744.c 1127753.t 1127756.a 1127759.g 1127762.c 1127763.a 1127765.c 1127768.t 1127774.t 1127780.c 1127783.g 1127792.g 1127795.g 1127795.t 1127818.a 1127819.t 1127828.g 1127833.g 1127834.g 1127835.c 1127836.c 1127837.a 1127837.c 1127849.g 1127852.g 1127867.c 1127879.g 1127886.c 1127900.c 1127909.c 1127911.c 1127912.a 1127913.c 1127918.g 1127921.c 1127929.t 1127930.t 1127934.c 1127940.c 1127951.g 1127954.g 1127957.g 1127969.g 1127970.c 1127978.a 1127982.a 1127984.a 1127985.t 1127992.a 1127993.a 1127996.g 1127999.c 1128002.g 1128003.c 1128004.g 1128005.c 1128006.g 1128008.g 1128011.a 1128015.a 1128017.t 1128018.c 1128032.a 1128035.c 1128038.g 1128039.c 1128041.c 1128041.g 1128047.c 1128056.g 1128059.g 1128068.g 1128071.c 1128080.t 1128089.a 1128089.t 1128107.c 1128110.a 1128111.c 1128116.c 1128119.t 1128122.c 1128122.t 1128125.c 1128128.t 1128131.g 1128134.c 1128135.a 1128146.c 1128148.g 1128149.t 1128152.g 1128152.t 1128155.t 1128170.c 1128173.c 1128179.c 1128182.c 1128182.g 1128206.c 1128212.a 1128221.c 1128222.c 1128227.c 1128230.t 1128234.g 1128239.c | penicillin-binding protein 2 |

**Table 5.3. SNPs associated with cefixime MIC in *N. gonnorrhoeae*.** These 222 SNPs were identified as significantly associated with the continuous ranked cefixime MIC phenotype when treeWAS was applied to core SNPs from 200 *N. gonorrhoeae* isolates.

the binary phenotype in Figure 5.1A falls mainly into four large clades, two susceptible and two resistant, the continuous MIC phenotype in Figure 5.3A produces a less rigid correlation between ancestral population and phenotype, by introducing moderate values into the major phenotypic clades. The MIC phenotype also resolves artefacts of the binary phenotypic categorisation. In Figure 5.1A, significant SNP genotypes offered no explanation for the six "sensitive" isolates in the upper-most resistant clade, or "sensitive" isolate 27241 in the central clade. As each of these isolates conversely receives a high MIC rank in Figure 5.3A, the continuous phenotype allows our approach to make more appropriate inferences about association in these genomes. Furthermore, especially as resistance phenotypes are known to evolve through the accumulation of substitutions in associated loci [307, 316], the analysis of continuous MIC data provided a valuable opportunity to examine phenotypic changes across a more refined gradient than the binary analysis could allow.

With the more informative continuous phenotype, treeWAS appears able to make a stronger distinction between population-associated loci and probable phenotypically-associated sites. As the estimated probabilities of spurious association are reduced across the genome, the null distributions shift left-ward from Figures 5.1B-D to 5.3 B-D, and the significance thresholds for Scores 1, 2, and 3 are reduced by 50%, from 0.9, 4, and 372 with the binary phenotype to 0.5, 1.7, and 176 with the continuous phenotype. Meanwhile, the initial set of resistance-associated SNPs remain significantly associated with the MIC phenotype, and an additional 90 core SNPs in *penA* and *murE* are found to be associated with cefixime resistance. Some of these have also been functionally validated [317], but further experimental analyses will be needed to investigate the impact of each candidate SNP on cefixime MIC. If their roles can be confirmed, we will be able to infer that, in this *N. gonorrhoeae* dataset, our approach achieved greater power with the continuous phenotype while maintaining the precision of the binary analysis.

We recommend that GWAS be applied to both binary and continuous phenotypes, where possible. Subsequent analyses will allow us to more thoroughly discuss how treeWAS performance varies as a function of phenotypic data type (see Section ). In this case, the continuous analysis appeared to reduce population stratification and add power to association inference. Both analyses can be informative, however, and additional insight may be gleaned from a comparison of their findings. Overall, in this application, our approach demonstrated the ability to identify phenotypically-associated recombinant regions and to select relevant loci within them, when applied to both binary and continuous phenotypic data. Similar applications of our GWAS method will be useful in areas ranging from drug development to surveillance.

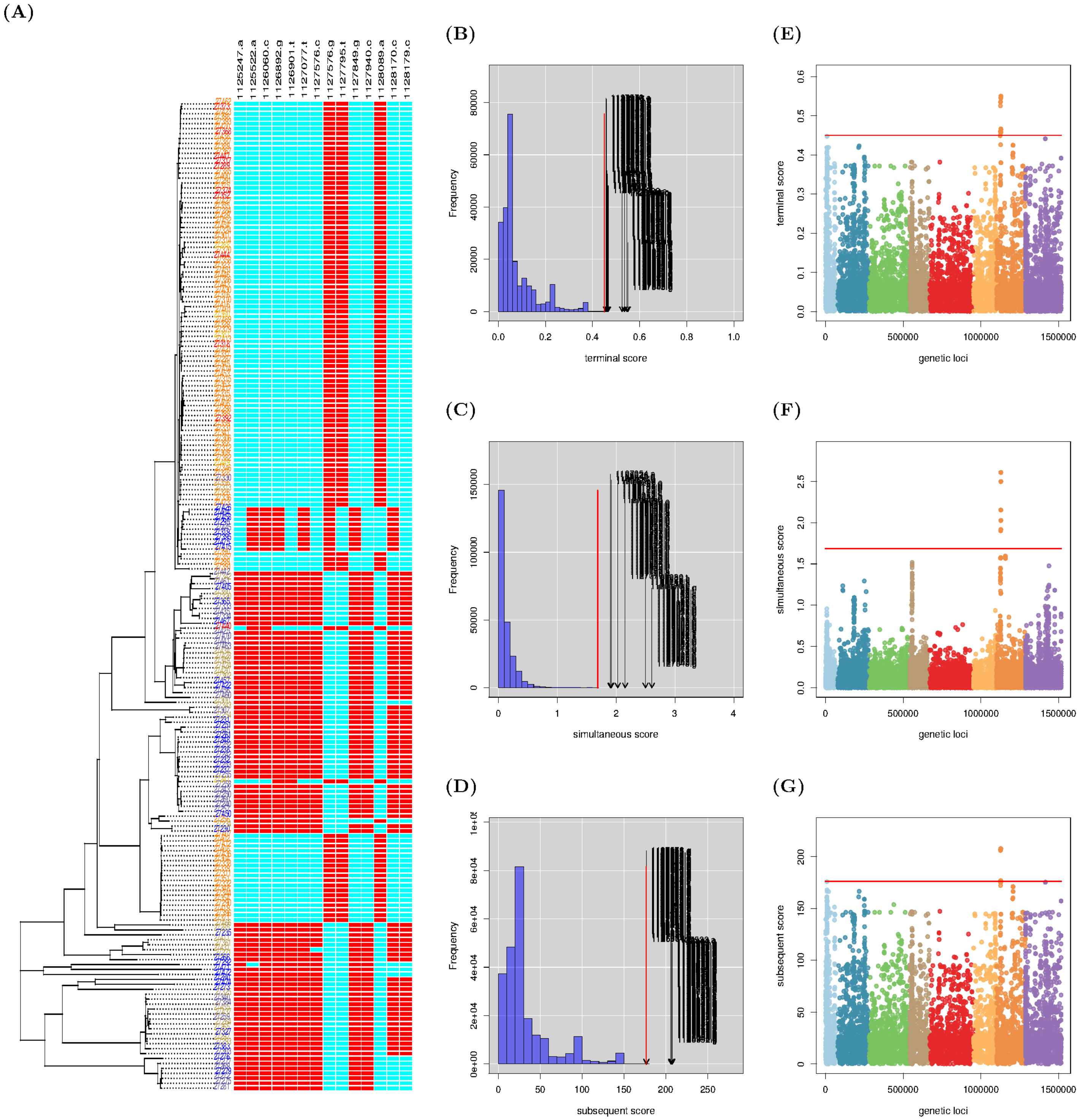


**Figure 5.3. Application to cefixime MIC in *N. gonorrhoeae* core SNPs.** treeWAS identified 222 SNPs associated with the continuous rank-transformed cefixime MIC phenotype, when applied to core SNPs from a sample of 200 *N. gonorrhoeae* isolates. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = lowest MIC; red = highest MIC). At right, an alignment of the 14 unique SNPs column patterns (blue = allele 0; red = allele 1) that were observed among the 222 significant SNPs. **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red line), above which associated SNPs are indicated. **E-G:** Manhattan plots for (E) Score 1, (F) Score 2, (G) Score 3 showing association score values for all SNPs, a significance threshold (red line), above which points indicate significant associations.

## 5.3 Penicillin resistance in Neisseria meningitidis

In our second analysis, we applied our approach to a dataset of *N. meningitidis* isolates with a penicillin resistance phenotype. *N. meningitidis* isolates typically inhabit the human nasopharynx, where they can be carried asymptomatically. Less often, meningococci invade host tissues, causing septicemia by infecting the bloodstream or giving rise to meningococcal meningitis by invading the central nervous system. Young children and immunocompromised individuals are especially at risk. Thirty million cases of these two diseases occur annually, alongside other manifestations of meningococcal infection [6]. Penicillin has mitigated the risks from meningococcal infection since it first became publicly available in 1942 [318]. Like cefixime, penicillin is a member of the $\beta$-lactam class of antibiotics. It interrupts the life cycle of *N. meningitidis* by preventing peptidoglycan synthesis and inhibiting meningococcal cell division and survival. In the past thirty years, however, *N. meningitidis* has become increasingly resistant to this life-saving drug. Although no previous studies have examined penicillin resistance in the dataset below, this phenotype is thought to be well characterised in the literature. Empirical evidence indicates that recombination plays a larger role in the evolution of the $\beta$-lactam resistance in *N. meningitidis* than in *N. gonorrhoeae* [310]. Nevertheless, many of the same genes are known to be involved in both gonococcal cefixime resistance and meningococcal penicillin resistance [319]. We test that, as in the above analysis, our GWAS method can confirm the identity of known resistance loci. We will, again, compare the results of our binary and continuous analyses of the resistance phenotype, to investigate how performance varies between empirical datasets.

We downloaded 171 serogroup B *N. meningitidis* whole-genome sequences from the *Neisseria* BIGSdb database [302] (see Tables A.4, A.5, A.6). We assembled two genetic datasets to be analysed by our GWAS method:

1. The **core SNP** dataset contained 166,848 binary SNPs.
2. The **accessory gene** dataset notes the presence-or-absence status of 2,808 genes.

*N. meningitidis* has a fairly high recombination rate, with previous estimates indicating $r/m \simeq 2$ (95% C.I. 0 - 5) [320] and suggesting that 40% of core genes are impacted by recombination [47]. Our approach remains applicable in this context, provided that recombination is accounted for during phylogenetic inference [1]. As above, we reconstructed a recombination-aware phylogenetic tree from whole-genome sequence data by building an intial tree with PHYLIP [194] and using ClonalFrameML [221] to identify the clonal genealogy (Figure 5.5A).

The penicillin resistance phenotype was analysed in two ways:

1. The **binary phenotype** categorised isolates as either penicillin "sensitive", if MIC $\leq 0.06\mu$g/mL, or "resistant" if MIC $> 0.06\mu$g/mL.
2. The **continuous phenotype** was defined as the ranks of the MIC values (see Figure A.2).

We applied treeWAS to both the accessory genome and core SNP datasets, with both the binary and continuous phenotypes. No accessory genes were found to be significantly associated with binary resistance or continuous MIC. Many core SNP associations were, however, identified with both phenotypic variables. treeWAS completed the accessory and core genome analyses in 23 seconds and 24 minutes, respectively.

### 5.3.1 Identifying associations with penicillin resistance

The application of treeWAS to the set of *N. meningitidis* core SNPs resulted in the identification of 162 SNPs significantly associated with the binary penicillin resistance phenotype. Table 5.4 shows that the majority, 126 SNPs, were in the well-known NEIS1753 (*penA*) gene encoding PBP-2, while 36 SNPs were located in the NEIS1751 (*murE*) gene. The connection between variation in *penA* and penicillin resistance in *N. meningitidis* is supported by ample evidence [299,321,322]. Even at the level of individual polymorphisms, a causal relationship has already been established experimentally for many of the *penA* SNPs we identified, in Table 5.4 [299,323]. The role of *murE* in penicillin resistance has also been confirmed in *N. meningitidis* [324]. Additional experimental analyses are still needed, however, to ascertain the impact of particular *murE* SNPS. In fact, while we had expected *penA* to feature in both our cefixime and penicillin analyses in these separate Neisseria species, we had not predicted the recurrence of *murE*. Given the under-representation of muropeptides like *murE* in the $\beta$-lactam resistance literature, we hope that our identification of *murE* candidate loci may provide motivation and direction for future work of this kind [155].

More *murE* and *penA* diversity underlies penicillin resistance in *N. meningitidis* than cefixime resistance in *N. gonorrhoeae*, above. In Figure 5.5A, we see 67 unique column patterns among the 162 significantly-associated polymorphic sites. The associated loci display both recurring linkage blocks and site-specific variation, with the bulk of this variation occuring in the genomes of resistant isolates. We infer that resistance arises primarily via recombination. The variation observed among associated sites may be due to mutation or the integration of exogenous DNA at variable sites. Compared to cefixime-associated loci, significant sites display more moderate and less uniform correlation (see Figure A.5). The literature attributes the extensive mosaicism in meningococcal PBPs, as compared to gonococcal variants, to more frequent intra- and inter-specific recombination events [310]. Indeed, hundreds of mosaic *penA* variants have been identified in *N. meningitidis* [299]. In this dataset, it would have been impractical and inappropriate to treat recombinant regions as single units, as suggested by Farhat *et al.* [106], as resistance has not evolved in durable, well-defined recombinant blocks. By operating at the level of individual sites, our approach was better suited to the identification of associated variants in this analysis.

| Gene | N.SNPs | SNPs | Gene product |
|---|---|---|---|
| NEIS1751 (*murE*) | 36 | 1241046.c 1241048.c 1241050.a 1241050.c 1241056.t 1241135.a 1241135.c 1241146.g 1241154.a 1241154.c 1241172.c 1241259.c 1241289.a 1241321.a 1241349.a 1241349.t 1241370.c 1241388.a 1241391.t 1241422.g 1241451.t 1241454.a 1241472.t 1241478.a 1241520.t 1241522.a 1241523.a 1241529.c 1241589.c 1241595.a 1241598.c 1241599.c 1241625.c 1241637.c 1241670.c 1241958.c | UDP-N-acetylmuramoyl-alanyl-D-glutamate–2, 6-diaminopimelate ligase |
| NEIS1753 (*penA*) | 126 | 1243506.a 1243510.t 1243515.c 1243542.c 1243575.g 1243581.c 1243587.c 1243605.g 1243608.t 1243612.g 1243615.g 1243617.c 1243635.c 1243644.g 1243668.c 1243671.g 1243683.c 1243688.c 1243692.a 1243695.t 1243696.t 1243702.a 1243704.a 1243704.c 1243713.c 1243715.g 1243716.g 1243774.c 1243788.c 1243788.t 1243797.c 1243799.c 1243800.a 1243800.c 1243801.c 1243806.c 1243806.g 1243809.c 1243817.t 1243818.t 1243822.c 1243845.g 1243848.g 1243854.c 1243854.g 1243857.g 1243858.c 1243866.a 1243869.c 1243870.a 1243872.a 1243872.c 1243873.t 1243880.a 1243881.a 1243884.g 1243887.c 1243890.g 1243891.c 1243892.g 1243893.a 1243893.c 1243894.a 1243894.g 1243896.g 1243899.a 1243902.c 1243903.a 1243905.t 1243906.c 1243920.t 1243923.c 1243926.g 1243927.c 1243929.c 1243929.g 1243935.c 1243944.g 1243947.c 1243947.g 1243953.g 1243956.g 1243959.c 1243968.t 1243977.a 1243977.g 1243986.c 1243995.c 1244001.c 1244004.c 1244010.a 1244010.g 1244010.t 1244013.a 1244013.t 1244016.t 1244018.c 1244022.c 1244023.a 1244032.a 1244034.c 1244037.t 1244040.c 1244040.t 1244043.t 1244058.c 1244058.g 1244061.c 1244067.c 1244070.c 1244070.g 1244094.c 1244094.t 1244109.c 1244110.c 1244118.t 1244134.g 1244139.t 1244143.c 1244144.c 1244152.a 1244155.a 1244163.c 1244166.c 1244185.a 1244190.c | penicillin-binding protein 2 |

**Table 5.4. SNPs associated with penicillin resistance in *N. meningitidis*.** These 162 SNPs were identified as associated with the binary penicillin resistance phenotype when treeWAS was applied to core SNPs from 171 *N. meningitidis* serogroup B isolates.

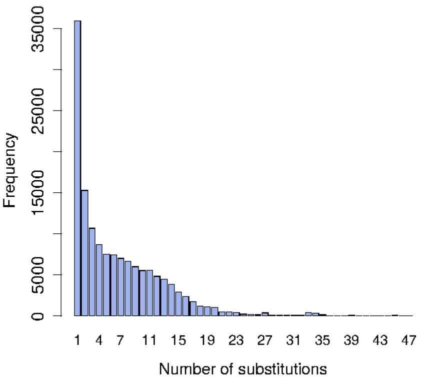


**Figure 5.4. Homoplasy distribution (*N. meningitidis* B, SNPs).**

We estimated the $N_{sub}$ distribution in Figure 5.4 among core SNPs, and 41 phenotypic substitutions in the evolutionary history of this sample. The high rate of recombination among associated sites, inferred above, aligns with the rapid pace of phenotypic change here observed in meningococcal penicillin resistance. Yet, the background rate of recombination, relative to mutation, is no greater in *N. meningitidis* ($r/m = 1.9$ [305]) than in *N. gonorrhoeae* ($r/m \simeq 2$) [320] . Evaluating mutation, recombination, and trait variation in a phylogenetic context, as in Figure 5.5A, our approach was able to estimate a relatively low probability of confounding bias, due to weak population stratification and LD. Through the simulation of null genetic data, treeWAS predicted limited inflation of association score statistics under the null hypothesis of "no association", identifying relatively low significance thresholds for all three scores, as compared to the cefixime resistance analysis above. As all significant SNPs are in confirmed resistance-associated genes, we can conclude that the FPR of treeWAS is low. And, with a larger proportion of unique column patterns among significant sites, the potential for false positives due to perfect linkage is reduced. At the same time, the sensitivity and discriminatory power of treeWAS also appears greater than it was in the previous analysis. In Figure 5.5E-G, treeWAS has been able to separate even moderate signals of association from the non-associated background. There nevertheless remain some individuals whose binary phenotypic state can not be explained by the significant SNPs identified. These may be better explained by their continuous MIC values or by non-genetic factors.

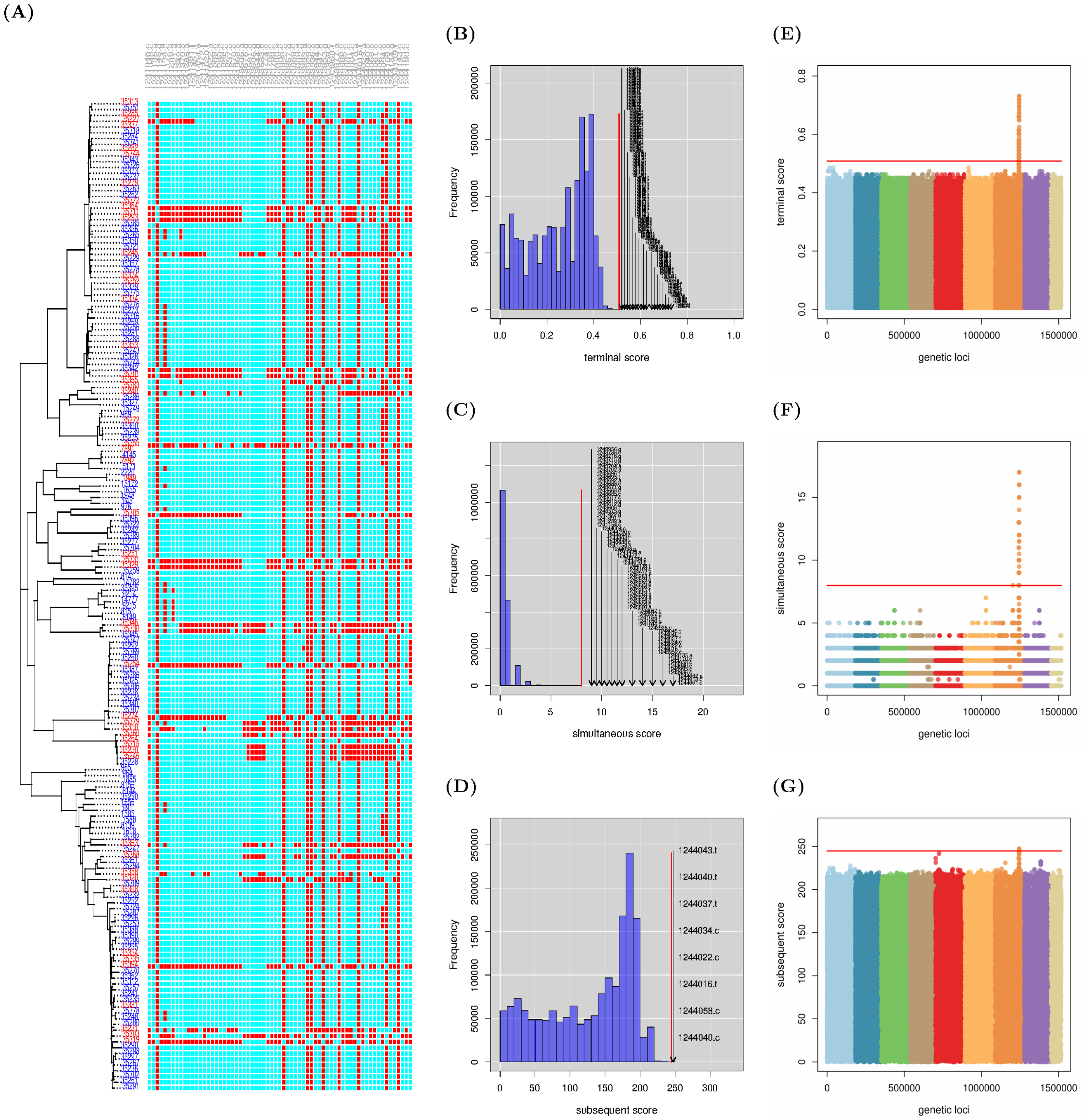


**Figure 5.5. Application to penicillin resistance in *N. meningitidis* core SNPs.** treeWAS identified 162 SNPs associated with the binary penicillin resistance phenotype. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = susceptible; red = resistant). At right, an alignment of the 67 unique SNPs column patterns (blue = allele 0; red = allele 1) that were observed among the 162 significant SNPs . **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red line), above which real associated SNPs are indicated. **E-G:** Manhattan plots for (E) Score 1, (F) Score 2, (G) Score 3 showing association score values for all SNPs, a significance threshold (red line), above which points indicate significant associations.

### 5.3.2 Identifying associations with penicillin MIC

Analysis of the continuous penicillin MIC phenotype identified 30 significant SNPs. Table 5.5 indicates that the majority were located in the *penA* gene, as expected. Of the 23 *penA* SNPs identified, at least six have already been confirmed experimentally [299, 323]. The identification of additional SNPs in *murE* further suggests that the role played by this locus may merit greater attention in research and drug development. SNPs in two additional genes were also found to be associated with penicillin MIC. The transcription factor *nusA* facilitates gene expression and is known to play a role in the resistance of *N. meningitidis* to other antibiotics [325, 326]. The NEIS0367 nucleotidyltransferase may participate in DNA damage repair and stress response, as similar gene products do in other Gram-negative bacteria [327]. If the gene does not impact penicillin resistance directly, it may increase MIC values by conferring a marginal fitness advantage in the presence of antibiotics [117]. All candidate SNPs, especially the two in novel genes, should be subjected to experimental validation to determine whether a causal link to penicillin MIC can be established, beyond the broader gene-level association.

| Gene | N.SNPs | SNPs | Gene product |
|---|---|---|---|
| NEIS0376 | 1 | 270097.g | putative sugar-phosphate nucleotidyl transferase |
| NEIS1556 (*NusA*) | 1 | 1093364.c | transcription elongation factor |
| NEIS1751 (*murE*) | 5 | 1241046.c 1241050.a 1241056.t 1241721.c 1241724.a | UDP-N-acetylmuramoyl-alanyl-D-glutamate–2, 6-diaminopimelate ligase |
| NEIS1753 (*penA*) | 23 | 1243959.c 1243977.g 1243986.c 1243995.c 1244001.c 1244010.g 1244016.t 1244018.c 1244022.c 1244032.a 1244034.c 1244037.t 1244040.c 1244040.t 1244043.t 1244058.c 1244067.c 1244109.c 1244110.c 1244118.t 1244134.g 1244185.a 1244190.c | penicillin-binding protein 2 |

**Table 5.5. SNPs associated with penicillin MIC in *N. meningitidis*.** These 30 SNPs were identified as significantly associated with penicillin MIC when treeWAS was applied to core SNPs from 171 serogroup B *N. meningitidis* isolates.

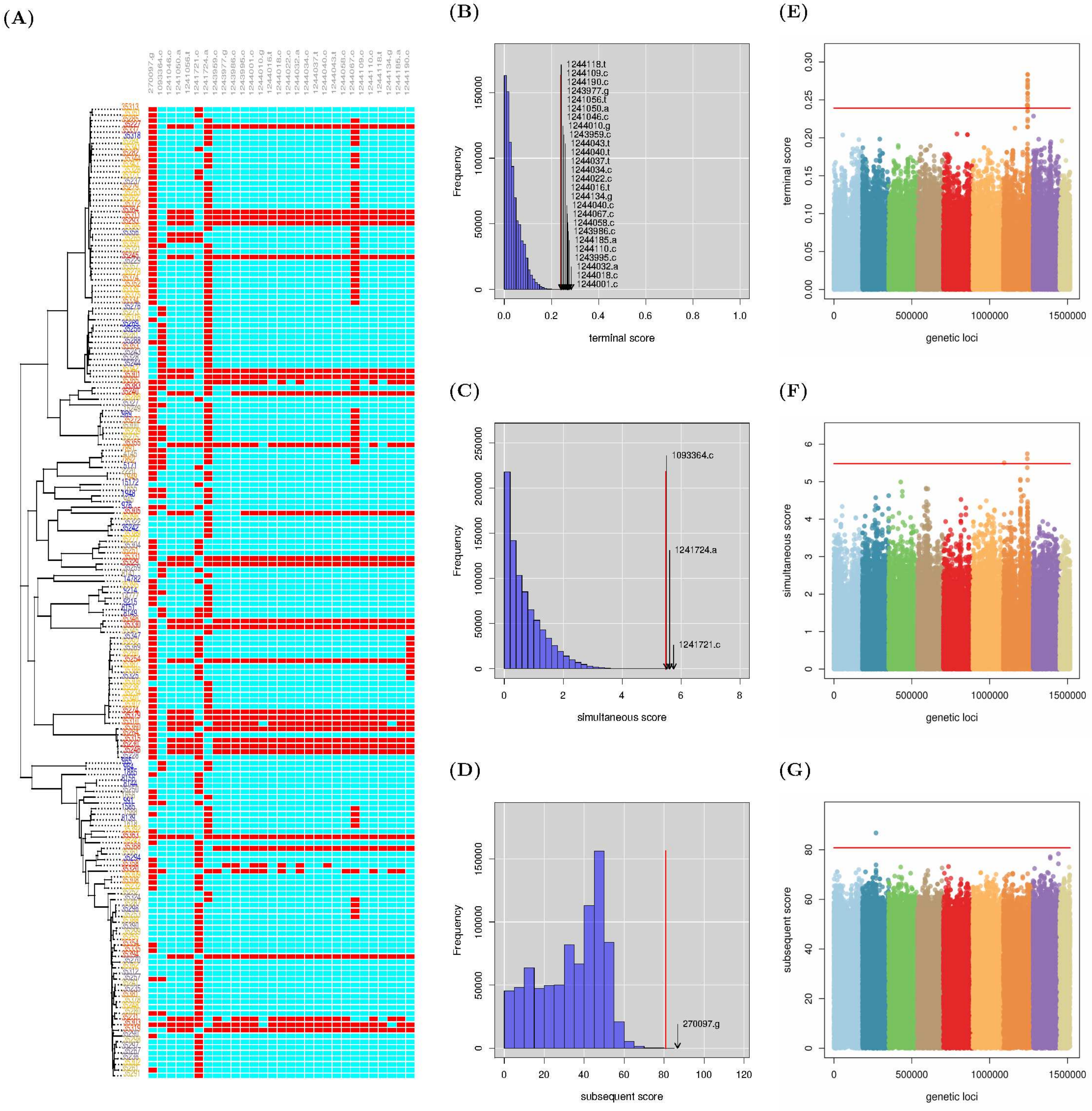


**Figure 5.6. Application to penicillin MIC in *N. meningitidis* core SNPs.** treeWAS identified 30 SNPs associated with the rank-transformed penicillin MIC phenotype. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (continuous: blue = lowest, yellow = moderate, red = highest MIC ranks). At right, an alignment of the 30 significant SNPs (blue = allele 0; red = allele 1). **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red line), above which real associated SNPs are indicated. **E-G:** Manhattan plots for (E) Score 1, (F) Score 2, (G) Score 3 showing association score values for all SNPs, a significance threshold (red line), above which points indicate significant associations.

Inverting the trend observed in our analyses of gonococcal cefixime resistance, our analyses of meningococcal penicillin resistance identified fewer associations with the continuous phenotype than with the binary phenotype. The binary phenotype was, again, more clustered and population-stratified than the continuous variable. But, with far more variation in the penicillin resistance phenotype, there was limited opportunity for population stratification to arise, even with the binary variable. As in the cefixime analysis, the significance thresholds for Scores 1-3 were reduced with the continuous phenotype. But, in the penicillin analysis, this did not lead to improvements in power, as power was not limited by population structure to begin with. Instead, penicillin MIC produced weaker association scores at significant loci. Yet, most associations with penicillin MIC were found with the same strongly-resistant isolates as in the binary analysis. To achieve the power needed to detect associations with intermediate penicillin MIC values, we would likely require a larger sample size. In this MIC GWAS, however, we even lost the variation at significant SNPs in resistant genomes that was permitted in the binary analysis. With few exceptions, MIC-associated SNPs were thus strongly correlated with one another (see Figure A.6). Several SNPs identified in the binary analysis, including two functionally-validated SNPs (1243956.g (T483S) and 1244139.t (P551S) [299]), were subsequently overlooked in the continuous analysis. Overall, whereas MIC values had improved sensitivity in the cefixime analysis by "diluting" a rigidly population-stratified phenotype, in the penicillin analysis, MIC values added noise to an already noisy relationship and drowned out many signals of association.

Across our analyses of binary and continuous $\beta$-lactam resistance phenotypes, the phenotypic data type alone was not predictive of the strength of confounding bias, nor the power or precision of our approach. But, despite considerable variation in the confounding effects encountered across the four analyses above, our approach was consistenty able to eliminate all or nearly all false positive findings. Although the number of significant SNPs varied, our approach was able to identify significant variants in known resistance-associated genes in each analysis.

Our results in these analyses of Neisseria suggest that association inference in treeWAS is accurate, efficient, and robust to the confounding influences of both vertical and horizontal evolutionary forces. By detecting new resistance-associated genetic variants, similar applications of treeWAS may be used to update and improve surveillance schemes, or to discover valuable molecular targets for drug development. As drug resistance continues to rise in many bacteria, evolving most rapidly in recombinant species, the suitability of treeWAS to both clonal and recombinant organisms, and to both binary and continuous phenotypes, will substantially increase the scope of our GWAS method and improve its capacity to uncover solutions to this and other public health challenges.

## 5.4 Invasive disease in Neisseria meningitidis

Having demonstrated the ability of our approach to identify associations with strongly genetically-determined resistance phenotypes, we wanted to attempt a more ambitious application of our method. We applied treeWAS to a sample of serogroup C *N. meningitidis* isolates, to search for associations with an invasive disease *versus* carriage phenotype, whose etiology is more complex than antibiotic resistance and less well understood. The invasive potential of meningococcal isolates is determined more probabilistically than resistance, arising from the contributions of multiple pathogen genetic factors, as well as external factors, like host immunity [328]. Whereas resistance-associated SNPs were repeatedly introduced in a recombinant genomic island, virulence factors are known to evolve through point mutation, intragenic recombination, and the gain and loss of whole genes and genomic islands [47]. Pathogenicity may emerge via recombination in otherwise commensal populations [329], or it may be restricted to hyper-virulent lineages [330]. Recombination should help to disrupt the extensive linkage of loci within the clonal frame, which may improve our ability to distinguish virulence-associated loci. The conflicting influences of clonal inheritance and recombination are nevertheless expected to confound the analysis, unless appropriately addressed. We aim to demonstrate the capacity of our GWAS method to disentangle and account for both factors, and to show that the design of treeWAS equips it with the precision, power, and flexibility needed to successfully identify both genes and SNPs associated with such subtle and complex phenotypes. We hope to identify both previously-characterised and novel invasiveness factors.

From the *Neisseria* BIGSdb database [302], we downloaded 129 European *N. meningitidis* sequences from serogroup C (see Tables A.7, A.8). We assembled two genetic datasets:

1. The **core SNP** dataset contains 115,386 binary SNPs.

2. The **accessory gene** dataset indicates the presence or absence of 2,809 genes.

As above, our approach will have to contend with the competing influences of clonal inheritance, mutation, and recombination in this application. We reconstructed the phylogenetic tree from whole-genome sequences, estimating an initial tree with PHYLIP [194] and using ClonalFrameML [221] to account for recombination while inferring the clonal genealogy (Figure 5.9A).

The phenotype was analysed as a binary variable. We extracted metadata from the BIGSdb database and used this to assign either an “invasive” or “carriage” phenotype to each isolate, according to the clinical status of the human host from which each isolate was sampled. Carriage isolates were sampled by taking throat swabs from healthy individuals, and invasive isolates were sampled from the blood or  of unhealthy patients. Virulence was not quantified, so the phenotype was not analysed as a continuous variable. We applied treeWAS to the accessory and core genome datasets, completing analyses in 17 seconds and 7 minutes, respectively. Our approach identified significant associations in both cases.

### 5.4.1 Identifying core SNPs associated with invasiveness

The tree and phenotype in Figure 5.8A do not differ from the accessory genome analysis. The homoplasy distribution in Figure 5.7 indicates that core SNPs have undergone frequent substitutions in the evolutionary history of these 129 sampled isolates. Among this variation, our approach was able to pick out seven SNPs significantly associated with invasive disease (Table 5.6).

| Locus | Gene | Gene product |
|---|---|---|
| 247279.t | NEIS0343 (*NAGS*) | N-acetylglutamate synthase |
| 248345.g | NEIS0344 | putative cell-surface protein |
| 445706.g | NEIS0614 (*ligA*) | DNA ligase |
| 653954.g | NEIS0361 (*mdaB*) | modulator of drug activity B |
| 944923.c | NEIS1348 | hypothetical protein |
| 945508.a | NEIS1364 (*porA*) | porin A, class 1 outer membrane protein |
| 1507935.c | NEIS2137 (*gapA2*) | glyceraldehyde 3-phosphate dehydrogenase C |

**Table 5.6. SNPs associated with invasive disease in *N. meningitidis*.** These 7 SNPs were identified as significantly associated with invasive disease when treeWAS was applied to core SNPs from 129 serogroup C *N. meningitidis* isolates.

Many of the core SNPs identified by treeWAS occur within well-known virulence genes. The Neisserial porin gene, *porA*, encodes an outer membrane protein that has been shown to translocate into eukaryotic cell membranes, to selectively interfere with polymorphonuclear neutrophils, to inhibit phagocytosis, and to facilitate epithelial cell invasion [331, 332]. It is already a meningococcal vaccine target [333]. The *mdaB* gene is associated with pathogenicity in many bacteria [334], and it has been shown to impact virulence in experimental studies [335]. Even components of core metabolic pathways are now known to play important roles in colonisation and the development of invasive disease in *N. meningitidis* [336]. Hence, although *NAGS* and *gapA-2* encode enzymes essential for arginine biosynthesis [337] and glycolysis [338], the evidence suggests that variation at these loci can impact invasivity, either by acting as a virulence factor or by providing a compensatory fitness advantage [336]. Comparative and experimental analyses of gene expression data have shown that *mdaB* [339], *NAGS* [340], *gapA-2* [336, 341], and NEIS0344 [342, 343] are up-regulated upon contact with eukaryotic epithelial or blood cells, providing additional evidence that these genes play an active role in invasive disease. We recommend that experimental follow-up studies be undertaken to elucidate the mechanisms by which polymorphisms in these genes may impact virulence, and to ascertain whether a biological link to virulence can be established at the novel candidates, *ligA* and NEIS1348. Because the genetic basis of meningococcal virulence is complex and incompletely understood, we anticipate that future work will clarify the roles of these loci. Indeed, existing evidence already suggests or supports a functional role in invasive disease for a majority of our findings.

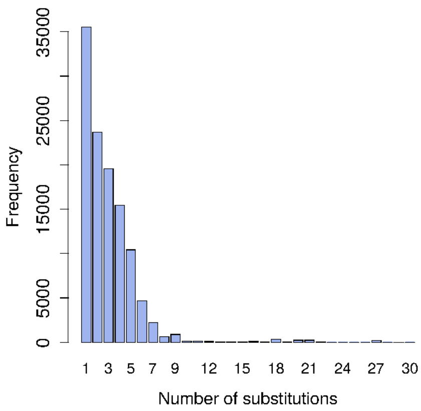


**Figure 5.7. Homoplasy distribution (*N. meningitidis* C, SNPs).**

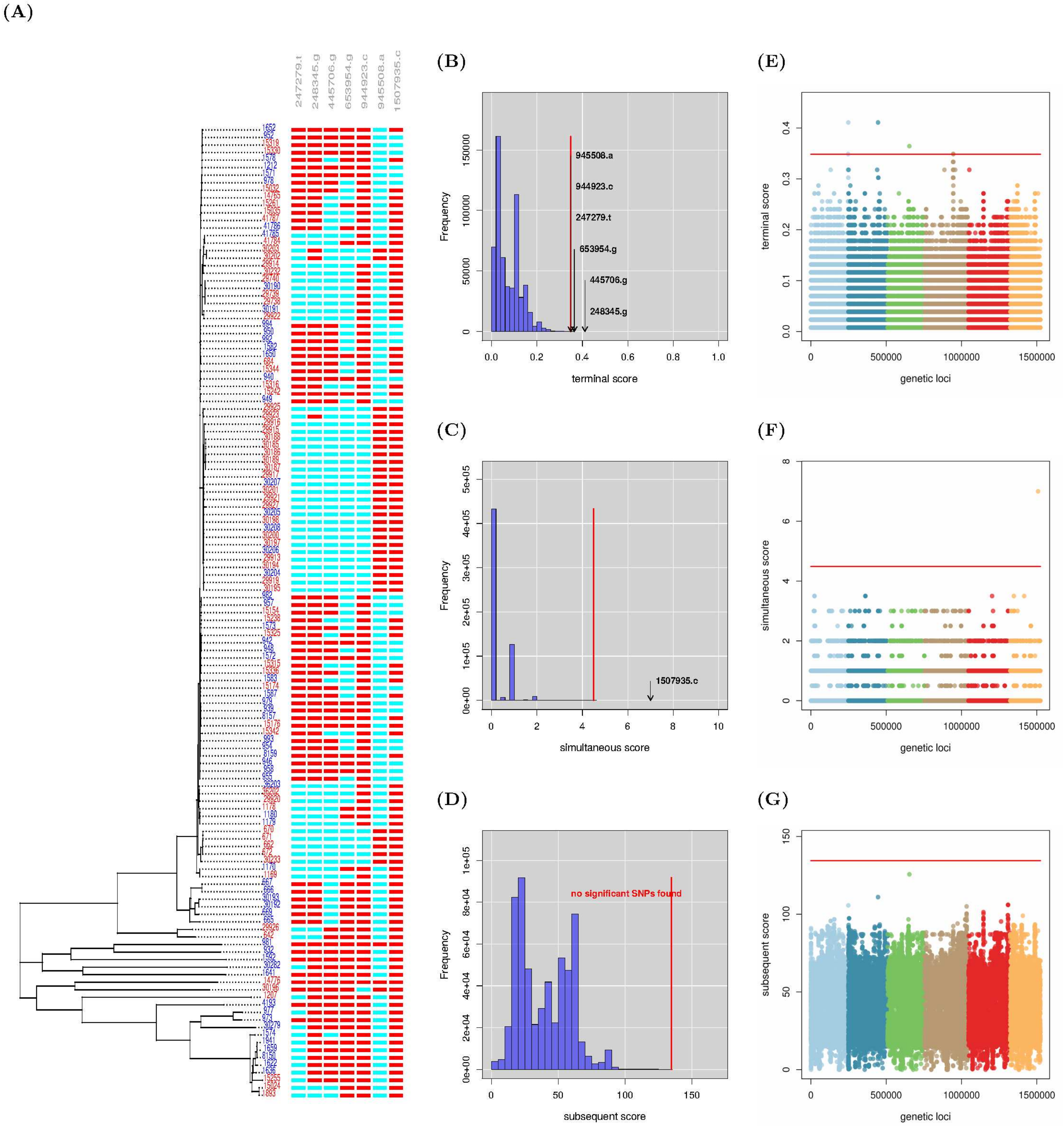


**Figure 5.8. Application to invasive disease in *N. meningitidis* core SNPs.** treeWAS identified 7 SNPs associated with invasive disease. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = carrier, red = invasive). At right, an alignment of the 7 significant SNPs (blue = allele 0; red = allele 1). **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red line), above which real associated SNPs are indicated. **E-G:** Manhattan plots for (E) Score 1, (F) Score 2, (G) Score 3 showing association score values for all SNPs, a significance threshold (red line), above which points indicate significant associations.

### 5.4.2 Identifying accessory genes associated with invasiveness

| Gene | Gene product |
|---|---|
| NEIS1969 (*NadA*) | Neisseria adhesin A |
| *hmbR* | Haemoglobin receptor protein |
| igr_NEIS0405_0406 | intergenic region between NEIS0405 and NEIS0406 |
| NEIS0596 (*MafA-2*) | Multiple adhesin family A 2 |
| NEIS0832 | hypothetical protein |
| NEIS0956 | cell-surface protein |
| NEIS0975 | putative outer membrane protein |
| NEIS1124 | hypothetical protein |
| NEIS1574 (*comEA-2*) | DNA transport competence protein |
| NEIS1880 | DNA transport competence protein |
| NEIS1996 (*comE*) | DNA transport competence protein |
| NEIS2072 | putative periplasmic protein |

**Table 5.7. Genes associated with invasive disease in *N. meningitidis*.** These 12 genes were identified as significantly associated with invasive disease when treeWAS was applied to 129 accessory genome gene presence-or-absence sequences from *N. meningitidis* serogroup C.

Unlike the resistance GWASs above, this analysis uncovered significant associations in the accessory genome, linking 12 genes to carriage or invasive disease (Table 5.7, Figure 5.9). It is encouraging to see that our method is able to identify both known and novel associations in the accessory genome, when they are predicted to exist [47]. Notably, none of the six comparator GWAS methods examined in Chapter 4 were able to find any significant associations to virulence among either core SNPs or accessory genes.

Three of the 12 genes were found to be positively associated with invasive disease. We were able to confirm the functional role of each by consulting the literature. *NadA* is an adhesin with well-characterised roles in virulence, enabling adhesion, colonisation, and invasion of host mucosal cells [344, 345]. It is one of the antigen targets of the Bexsero meningococcal vaccine [333]. *MafA2*, another adhesin, facilitates adhesion to human cells via glycolipid binding [346] and plays a similar role to *NadA* in pathogenic *Neisseria* [347]. *hmbR* is a haemoglobin receptor protein, which facilitates iron acquisition and haeoglobin uptake [348]. These processes have been shown to enhance the growth of invasive meningococci within the bloodstream [349]. As the *hmbR* gene is highly conserved [350], it may represent a good target for vaccine development.

We also identified nine accessory genes whose presence was associated with Neisserial carriage. Virulence factors that improve commensal fitness rather than pathogenic invasiveness are, in fact, widely known to be enriched in the accessory genome, where they mediate interaction and communication with host cells and facilitate adaptation to selective pressures induced by host immune responses, competition from other nasopharyngeal microbes, and environmental conditions [47,351]. Three *comEA*-like competence proteins were identified, encoded by the NEIS1574, NEIS1880, and NEIS1996 genes. Analyses in *S. pneumoniae* have recently established A similar link between competence genes and carriage duration has recently been established in *S. pneumoniae* [132,352]. These genes may have enduring, if non-specific, effects on the phenotype, as an enhanced capacity for recombination can improve the flexibility and responsiveness of Neisseria [337,353].

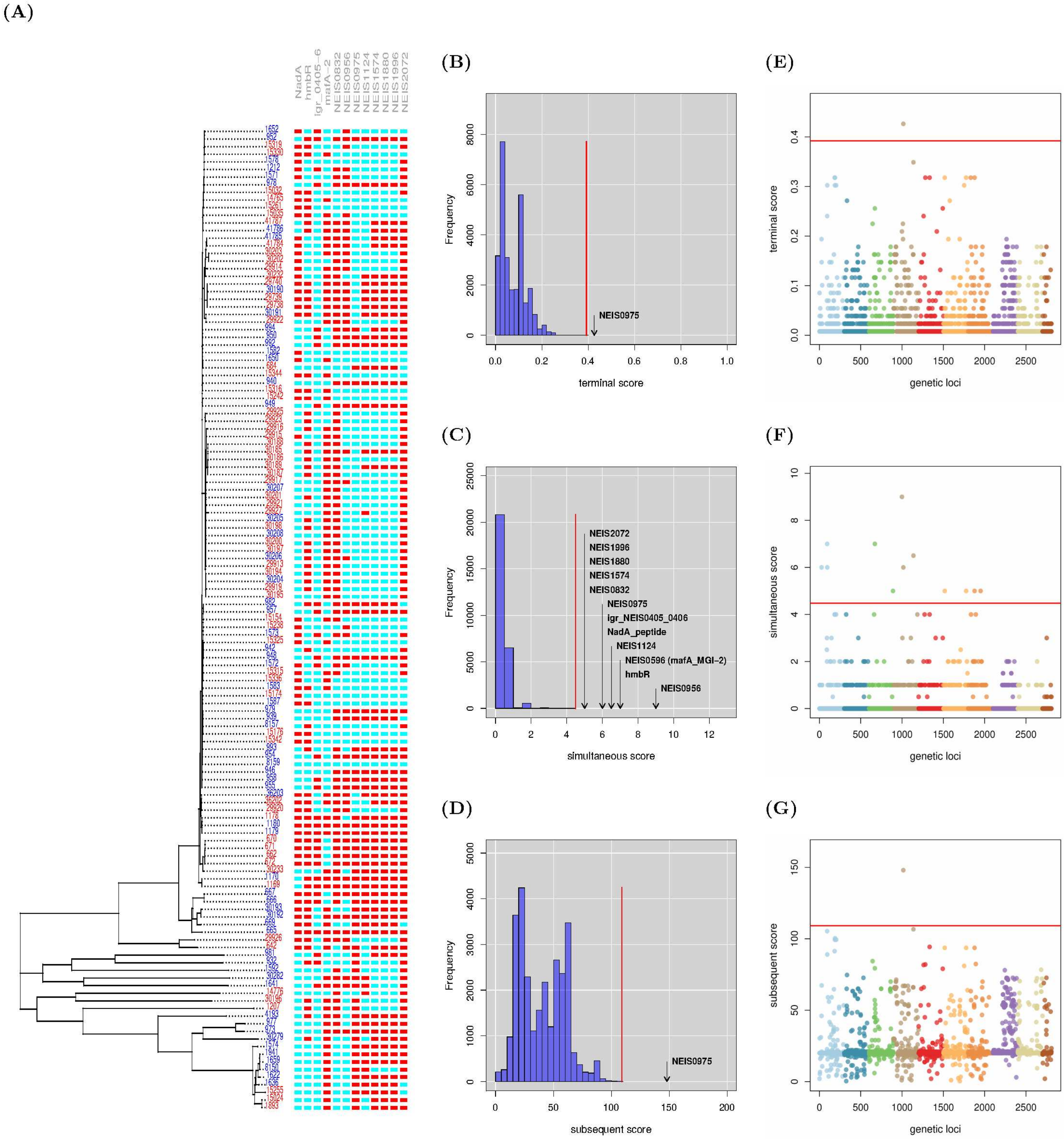


**Figure 5.9. Application to invasive disease in the *N. meningitidis* accessory genome.** treeWAS identified 12 genes associated with invasive disease. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = carrier, red = invasive). At right, an alignment of the 12 significant genes (blue = gene absence, red = gene presence). **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red line), above which real associated genes are indicated. **E-G:** Manhattan plots for (E) Score 1, (F) Score 2, (G) Score 3 showing association score values for all genes, a significance threshold (red line), above which points indicate significant associations.

An adaptive advantage to *N. meningitidis* may also be provided by the membrane-associated proteins encoded by NEIS0956, NEIS0975 , and NEIS2072, which may enable immune evasion via surface modulation and, thus, favour colonisation and survival in the nasopharyngeal niche [354, 355]. Table 5.7 contains three additional loci, but as their gene products remain unknown, we are not yet able to characterise or confirm these findings. Among them, our GWAS method has detected one virulence-associated intergenic region, between NEIS0405 and NEIS0406. If we had not chosen to retain synonymous and non-coding genetic variation in this GWAS study, this association would have been overlooked. We note that hundreds of similar intergenic and non-coding regions have been linked to pathogenicity through potential regulatory roles [274]. Although additional experimental analyses would be needed to establish causality at these three novel genes, the specificity of our approach thus far gives us confidence in our results. Indeed, in each case where the protein product is known, we have been able to find direct evidence linking the significant genes identified by treeWAS to the virulence phenotype.

Significant virulence genes display more variation and have a higher $N_{sub}$ than most accessory genes in Figure 5.10. The repeated gain and loss of multiple genes along the tree in Figure 5.9A, especially on very short branches, suggests that virulence factors are being introduced through the horizontal exchange of pathogenicity associated islands, as well as individual accessory genes [353, 356]. But, because the phenotype undergoes no less than 35 substitutions in the evolutionary history of these 129 isolates, the confounding effect of population stratification is estimated to be minimal. As the "invasive" or "commensal" phenotype represents a snapshot at the time of sampling, however, some of this variation may not reflect the true invasive potential of the genome in another host or sample. We expect that some association strength and discovery power will be lost to this phenotypic noise. On the other hand, though some accessory genes undergo enough substitutions (Figure 5.10) to achieve high association scores under $H_0$ in theory, the simulations within treeWAS show that only limited association scores are likely to arise by chance. Ultimately, by carefully ruling out spurious associations due to reproducible confounding factors, and by enhancing power with our homoplasy score, our method was able to attribute significance to a dozen accessory genes with both known and novel associations to virulence.

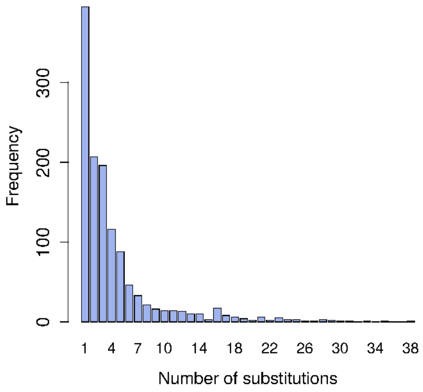


**Figure 5.10. Homoplasy distribution (*N. meningitidis* C accessory genes).**

### Considering epistatic interactions associated with invasive disease

Evidence from the literature indicates that individual meningococcal virulence genes are rarely necessary or sufficient to independently convert commensal isolates into pathogenic ones, or vice versa [357]. Comparing the patterns of accessory gene presence and absence in Figure 5.9A to the terminal distribution of “invasive” and “commensal” phenotypic states, it appears that none of the 12 significant genes has a deterministic effect on virulence in isolation. Accordingly, in Figures 5.9B and D, all but one of the 12 virulence-associated genes displays weak sample-wide association.

On the other hand, virulence is known to be probabilistically determined through the contributions and interactions of multiple loci [358]. Strong pairwise correlations between many of the virulence-associated accessory genes may, in fact, represent significant interactions (see Appendix, Figure A.8). Notably, although resistance-associated loci displayed even stronger correlations, the physical proximity of these significant sites along the chromosome meant that we had insufficient evidence to distinguish any potential epistasis from physical linkage. By contrast, as the putative virulence genes identified above are separated by considerable distance along the chromosome, the correlations among these sites are less likely to reflect LD and more likely to indicate epistatic interactions. For example, the perfect correlations and significant interactions ($p < 0.001$) observed among the three competence genes in Table 5.7 (NEIS1574 *comEA-2*, NEIS1880, NEIS1996 *comE*) fit with our understanding of the epistatic interaction between genes and the integration of gene products that enable bacterial transformation [359,360]. This network of significant interactions also appears to extend to NEIS0956, NEIS0975, and NEIS1124, prompting an expanded view of the protein-protein interactions that may be involved in recombination. More broadly, this preliminary analysis enhances our understanding of the individual actions and epistatic interactions displayed by accessory genes associated with meningococcal colonisation and invasive disease.

## 5.5 Concluding remarks

In each of the empirical analyses above, we used treeWAS to search for potential trait-associated variation throughout the pan-genome. Investigating resistance to $\beta$-lactam antibiotics in our first two analyses, our GWAS method uncovered significant associations only within the core genomes of *N. gonorrheae* and *N. meningitidis*. Yet, owing to the efficiency of the treeWAS R package, we were additionally able to test all accessory genes for association to the phenotype in a matter of seconds. Our two-pronged association testing pipeline ensured that a wider array of potentially relevant variation was considered. This analytical strategy was ultimately rewarded in our third and final application to empirical data, where treeWAS identified novel and known associations with virulence among both core SNPs and accessory genes.

We report that the performance of our association testing method exceeded our expectations when treeWAS was applied to biological sequence data. The empirical analyses above showcase the reliability and adaptability of our GWAS method. Our results display the reliable capacity of treeWAS to reject false positive findings, and they demonstrate that treeWAS is able to achieve the power needed to discover many true positive findings as well as novel putative associations. By contrast, when we applied alternative GWAS methods to the same empirical datasets, we found that their results were most often characterised by either volumes of probable false positives or the absence of test positives.

The Venn Diagrams in Figures 5.11 and 5.12 summarise the contributions of Scores 1, 2, and 3 to our findings from the six empirical analyses above. Confirming our observations from the simulation study, we find that each measure adds to the power of treeWAS. But, whereas Score 2 had been the major driver of sensitivity in each simulation set, the results of the empirical analyses reveal greater diversity. Here, we find that the relative contributions of Scores 1, 2, and 3 vary substantially between analyses. Together, Scores 1-3 improve our ability to uncover genuine associations across diverse association landscapes, enhancing the scope and versatility of or GWAS method.

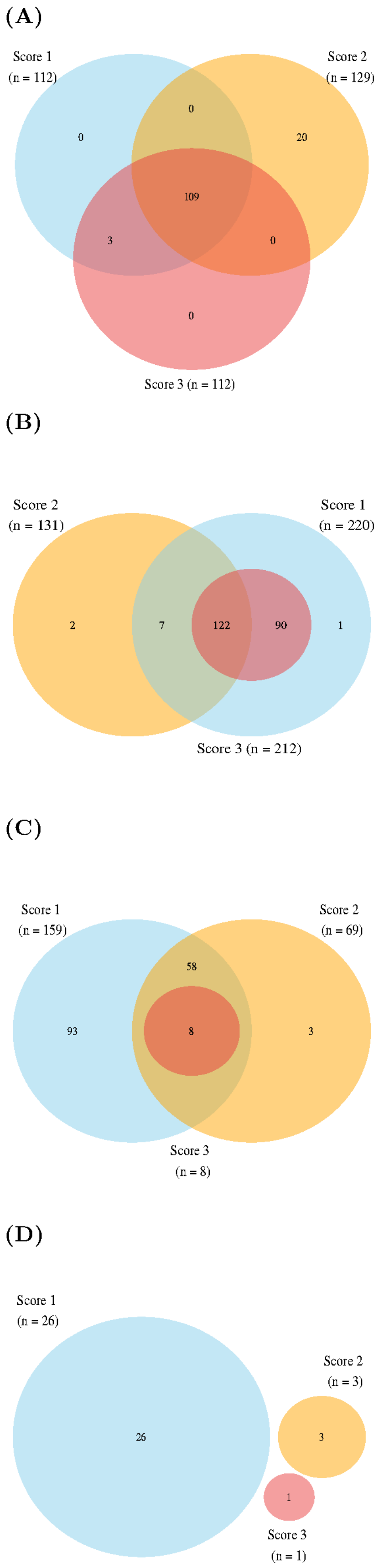


**Figure 5.11. Venn Diagram I. A:** cefixime resistance, **B:** cefixime MIC, **C:** penicillin resistance, **D:** penicillin MIC.

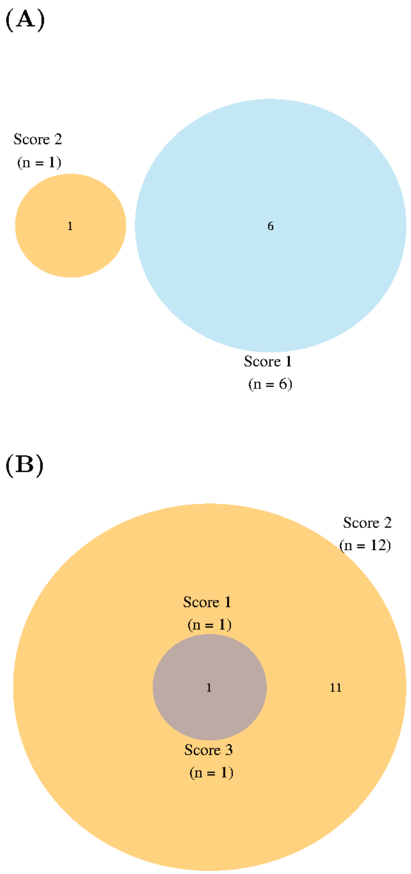


**Figure 5.12. Venn Diagram III.** Invasive disease **A:** core SNPs, **B:** accessory genes.

Overall, our results from these empirical analyses illustrate the power and flexibility that can be achieved by the multiple measures implemented in treeWAS. They suggest that the value added by pooling Scores 1-3 is likely to be even greater in empirical analyses than we could have predicted from our simulation study alone. Equally important, the above applications to empirical data show that, while maintaining high power, our approach was consistenty able to adapt to new confounding biases and to eliminate all or nearly all false positive findings. Altogether, these results provide strong suppport for the performance and potential of our approach to microbial GWAS.

# Chapter 6

# Discussion

## 6.1 Summary of previous chapters

We opened this thesis with an introduction to microbial genetics and microbial GWAS, more specifically. In Chapter 1, we described the comparative analytical approach adopted in association studies. We argued that the efficiency, broad applicability, unbiased nature, and genome-wide design of GWAS methods offered clear advantages over traditional laboratory techniques, particularly in the "omics" era [48]. This was followed by a review of the GWAS studies that have been carried out thus far in microbial samples. We underscored the vast potential that association studies have to improve our understanding of infectious pathogens. However, we noted that the undertaking of microbial GWAS studies has been hampered by a lack of purpose-built and thoroughly-tested methodology.

Chapter 2 was devoted to a review of the literature and a thorough examination of the major methodological challenges in microbial GWAS. We focussed on three main areas, namely population stratification, recombination, and association testing. Potential solutions to each issue were considered, and their strengths and limitations compared. This allowed us to lay out our motivations for the choices we made when developing our own microbial GWAS method in the next chapter.

In Chapter 3, we introduced our new phylogenetic, multi-measure, Monte Carlo approach to microbial GWAS. We proposed the following procedure. First, a recombination-adjusted phylogenetic tree is reconstructed from whole-genome sequence data, and ancestral states and substitutions are reconstructed via maximum parsimony. A Monte Carlo simulation process is then carried out along the inferred phylogeny, while preserving additional empirical parameters, including the homoplasy distribution. This allows us to model the neutral evolutionary process expected under the null hypothesis of "no association", while accounting for confounding bias due to clonal ancestry, variable mutation and recombination rates, and variation in other empirical parameters. To ensure that our robust approach could also achieve good statistical power, we developed a strategy to enhance sensitivity through the parallel application of three measures of association, including both allele-based and homoplasy-counting scores. A comparison of the null and empirical distributions of each association statistic then allows us to distinguish statistically and evolutionarily significant relationships between genotype and phenotype from a noisy background of spurious associations. We concluded this chapter by describing the implementation of our GWAS method in the treeWAS R package, highlighting features that enhance its accessibility, efficiency, and effectiveness.

In Chapter 4, we presented the results of over 600 applications of our GWAS method to simulated datasets. We began by introducing multiple simulation frameworks, which were developed for these analyses and implemented in the treeWAS R package. We then described how genotypic and phenotypic datasets were simulated for this simulation study, varying parameters governing the dimensions of simulated datasets, the phylogenetic relationships among individuals, the recombination rate among genomic loci, and the frequency, nature, and strength of associations between genotype and phenotype. In the first set of analyses presented, we applied treeWAS to these synthetic datasets to evaluate elements of our GWAS method, prompting evidence-based decisions in favour of parsimony for ancestral state reconstruction and against scaling Score 3 by branch length. We also determined that optimal performance could be achieved with a count-based, Bonferroni-corrected significance threshold, with $\alpha_{base} = 0.01$ and $N_{sim} = 10N_{loci}$. Next, applications to simulated data were used to evaluate the performance of treeWAS and compare it to six popular alternative approaches, including uncorrected association tests and approaches with uniform, dimension reduction, and cluster-based adjustments for ancestry. Exceeding expectations, treeWAS achieved by far the greatest precision of any method while reaching comparable sensitivity. As a result, our approach displayed the strongest overall performance of any approach in Sets A, B, and C, respectively identifying simple, complementary, and more complex associations. Finally, we applied treeWAS and alternative approaches to datasets simulated while sample size, genome size, and background noise were varied across wide parameter ranges. Here, again, our approach consistently out-performed the other GWAS methods examined. Our simulation study thus provided strong support for our GWAS method. Indeed, these results suggested that treeWAS would be more effective than alternative approaches at identifying genuine associations in analyses of real biological sequence data.

In Chapter 5, we applied our GWAS method to empirical datasets from *N. gonorrhoeae* and *N. meningitidis*. In examining both antibiotic resistance and virulence phenotypes, treeWAS was able to both confirm well-known loci and to uncover novel associations. We were pleased to find that a putative functional link, or a homologous relationship in another species, could be gleaned from the literature for a large majority of the previously-unreported associations identified by our approach. Furthermore, these analyses confirmed that the performance of treeWAS observed in our simulation study was reasonably representative of its true power and precision in empirical analyses. If anything, the variable contributions of Scores 1-3 to our empirical findings suggest that our analyses of simulated data may have underestimated the flexibility, utility, and perhaps even the power of our GWAS method. These applications to empirical data also provided a practical demonstration of the versatility of treeWAS, as we were able to identify significant findings in both the core and accessory genome, with both deterministic and complex associations, examined as both binary and continuous variables. Altogether, these applications to empirical data provided a powerful demonstration of the potential of our GWAS approach in real genetic data analyses.

## 6.2 Strengths

Through a process of development, testing, and refinement, we created a new approach to GWAS that was tailor-made for use in microbial samples. We implemented new solutions to each of the three major methodological challenges examined in our review of the literature. Moreover, we were able to achieve each of the primary objectives identified in Chapter 2, by developing a method that was able to:

1. Address the confounding effects of clonal ancestry.
2. Account for the variable confounding influence of homologous recombination.
3. Augment discovery power without sacrificing precision.
4. Improve accuracy and expand reach by capitalising on all available data.
5. Ensure efficiency and accessibility by implementing the treeWAS R package.

The results of our analyses of simulated data attest to the merits of our approach. Furthermore, our applications to empirical datasets provide a powerful case for the potential of treeWAS to lead to new discoveries that will advance our understanding of microbial genetics and infectious disease. Ultimately, through a process of development, testing, and refinement, we were able to overcome a number of critical limitations in existing GWAS methods, by producing an approach that was:

**Designed for clonal and recombinant microbes** Our approach derives its greatest benefits from the fact that it was specifically designed for use in microbes, unlike most of the GWAS methods that have been applied to microbes thus far. For example, we know from empirical evidence that most organisms display both clonality and recombination to some degree, and that strong clonal relationships and variable homologous recombination are two of the most significant sources of confounding bias in microbial GWAS studies [74, 97, 183, 245]. We were able to ensure that the confounding effects of ancestral relationships could be sufficiently addressed, even in strictly clonal organisms, by adopting a phylogenetic approach. Our simulation study showed that uniform, PC-based, and cluster-based corrections for population structure were, conversely, unable to offer comparable precision and power when applied to the same datasets [97, 183]. Our simulation study also raised new questions about the advertised benefits of dimension reduction techniques and cluster-based methods in the presence of recombination. We were pleasantly surprised to find that the precision and performance of our simulation-based approach remained well above all competing approaches at all recombination rates examined. Despite these clear strengths, treeWAS is still the only phylogenetic GWAS method to explicitly account for recombination as well as clonality during ancestry and association inference. We hope that these positive results may encourage further research into recombination-aware, tree-based approaches.

**Powerful and versatile** Another advantage of our approach is that it incorporates multiple, complementary association testing frameworks within a single overarching method. In addition to the allele-based measure that we first developed, in Score 1, our tree-based approach enabled the implementation of our homoplasy-counting Score 2 and allowed us to extend our allele-based Score 3 into the evolutionary past. By drawing on three distinct measures, treeWAS is likely to identify a broader range of associations with more diverse evolutionary backgrounds, as one might expect to encounter in analyses of diverse microbial genotypic and phenotypic datasets. The results of multiple applications to simulated and empirical data allowed us to confirm this in practice and to state confidently that our multi-measure approach improves power without undermining performance. The parallel application of our three measures proved additionally useful in providing more detailed insights into the nature of each association identified. Whereas most alternative approaches produce a binary significant/insignificant result at each locus, our method offers three clear perspectives on the association present at every significant locus. This improves the interpretability of our findings, for example, as compared to approaches that merge measures from multiple tests [130] or that separate findings into lineage and locus effects [128]. As demonstrated in our empirical analyses, treeWAS can identify associations between any form of binary genetic data and any binary, continuous or ordered categorical phenotype, unlike many other methods. The modular structure of our R package will also ensure that treeWAS can accommodate additional or alternative association scores, in future. This will allow treeWAS to be updated and expanded as association testing measures are improved and extended to meet new analytical aims.

**Efficient and accessible** The treeWAS R package itself offers a number of advantages over alternative software tools. The computational efficiency and scalability of the treeWAS R package will be an asset in any analysis, whether our GWAS method is being applied to accessory genome data, core SNPs, or evem k-mer datasets. Indeed, our efforts to optimise the efficiency of treeWAS managed to alleviate the computational burden often associated with simulation-based phylogenetic methods [183]. The treeWAS R package also eliminates much of the subjectivity inherent in most other GWAS methods. PCA, DAPC, and the CMH test, for example, require users to perform separate analyses of the population structure and to select $k$, through visual inspection, with reference to one of many goodness-of-fit measures, or by choosing a particular $k$-selection procedure. We avoid these non-standardised decision-making procedures, by implementing objective, automated optimisation procedures within the treeWAS R package. By eliminating this source of variation, our approach reduces subjectivity and error.

Perhaps the greatest testament to the benefits of this project has been provided by the positive reception of our publication [155] and the treeWAS R package. We are proud to report that the treeWAS R package has developed an active community of users, who have applied our method to a wide variety of organisms. Melnyk *et al.* [361], for example, have used the treeWAS R package to successfully identify genes associated with pathogenicity in the plant pathogen *Pseudomonas fluorescens.* We have certainly been encouraged to see that, since the publication of our paper [155] earlier this year, it has been downloaded over 3,800 times and viewed by over 11,000 people. Our work has generated much interest among members of the scientific community from over 30 countries, sparking more discussion on social media than 98% of papers published during the same period. In this thesis, we demonstrated the power and potential of treeWAS through applications to simulated and empirical data. But, the widespread uptake and popularity of treeWAS may provide an even stronger endorsement of the utility of our new GWAS method and the value of this work.

## 6.3 Limitations

Although our results showed that the design of our microbial GWAS method gives treeWAS a considerable performance advantage over alternative GWAS approaches, a number of areas remain open to improvement.

**Phylogenetic uncertainty** One potential limitation of our approach is that it assumes that there is no uncertainty about the phylogenetic tree. An unreliable tree topology or incorrect branch lengths will impact the association scores calculated and the data simulated within our method, increasing our chances of making incorrect inferences about the nature and significance of associations. Thankfully, we do not expect phylogenetic reconstruction to be a major source of uncertainty in our GWAS analyses. Because our samples must contain multiple phenotypic substitutions, they tend to be fairly diverse, often spanning an entire lineage or species. The availability of genome-wide data also improves phylogenetic accuracy. Still, some uncertainty will inevitably occur, especially in samples that undergo recombination.

**Limits on recombination** We designed our approach to account for both clonal inheritance and recombination during phylogenetic reconstruction and association inference. While our approach should be able to handle clonal population structure of any strength, we acknowledge that there is an upper limit on the acceptable level of recombination. When recombination exceeds mutation to the extent that it obscures the clonal genealogy, the essential ancestry and association inference components of our method will become unreliable and alternative approaches may be more appropriate. In the publication and documentation of our method, we encourage users to estimate $r/m$ and to question the applicability of treeWAS in each analysis [155]. However, at present, we cannot state precisely at what relative levels of recombination and mutation treeWAS will be rendered inappropriate. In lieu of our current call for caution, it would be better if we could provide users with an evidence-based $r/m$ cutoff, to provide a clear upper bound on the level of recombination that is acceptable in treeWAS analyses.

**Skewed continuous phenotypes** We indicate in Chapter 3, and demonstrate in Chapter 5, that our method can be applied to skewed continuous phenotypes when suitable transformations are applied to the data. We have, thus far, been able to provide advice on this topic and to aid individual users through the treeWAS forum, available online at `https://github.com/caitiecollins/treeWAS/issues`. A treeWAS tutorial now describes the visual assessment of continuous phenotypic distributions, as well as the transformation by rank of continuous MIC values that we used in our MIC analyses [155], which has since been adopted elsewhere [362]. Nevertheless, treeWAS would benefit from stronger guidelines about what constitutes a problematic level of skew and what steps might be taken to address this. We may want to recommend particular measures of skewness and indicate acceptable upper bounds [363]. We could also make clearer recommendations about what data transformations would be appropriate in different conditions.

**Score 3** Our simulation study revealed that Scores 1, 2, and 3 were each a net benefit to treeWAS, as each contributed more true positives than false positives to our pooled set of results. But, in empirical analyses, whereas Scores 1 and 2 made many unique contributions to the findings of treeWAS, Score 3 clearly remained the weakest contributor in applications to both simulated and empirical datasets. We were not too surprised by this, as Score 3 was designed to fill in any remaining gaps between Scores 1 and 2. But, I still wonder if we can improve upon the current implementation of Score 3. It may be worth pursuing either improvements to Score 3 or the development of one or more additional association scores, to see if we can identify any measures that out-perform Score 3 or otherwise improve the performance of treeWAS. The modular design of our R package makes it straightforward to incorporate new measures of association, which can be used either instead of or in addition to any of the three association scores currently implemented in treeWAS.

**Additional confounders** We know that variation in the host genome or the environment can also affect bacterial phenotypes. However, at this point, our GWAS approach does not account for variables outside of the bacterial genome. At present, if unknown or unmeasured external confounders were a concern, we would recommend either increasing the sample size or repeating the analysis in a separate sample to validate results [49]. If we were concerned about a particular external factor, we could stratify the analysis by the potentially-confounding variable, applying treeWAS to separate samples, each containing little to no variation in the confounding factor. Alternatively, we could repeat the treeWAS analysis to test for association with the confounding variable as well as the phenotype. We remain somewhat cautious about modifying treeWAS to account for additional measured or unmeasured confounders, as analyses on simulated data have shown that these approaches can increase error without improving performance [364]. However, we note that regression models have been successfully used in microbial GWAS to account for covariates, like environment, serotype, and host source. Controlling for correlations with these variables has been shown to reduce confounding bias and improve precision in some cases [134,135]. Additional variables have also been incorporated in joint analyses, by testing for associations either with multiple correlated phenotypes [365, 366] or with interacting host and pathogen genotypes [119, 367, 368]. As these approaches have been shown to improve precision and power in certain circumstances, they merit further consideration as well.

Analyses of simulated data could help us to estimate the impact of external confounders and to guide any future efforts to account for additional variables in treeWAS.

## 6.4 Future directions

### 6.4.1 Application to viruses

Although our focus in this thesis has been on bacterial GWAS, our approach should also be able to test for phenotypic associations with viral genetic variation. Compared to bacterial samples, viral genomic data is often characterised by shorter sequences, higher rates of mutation and recombination, and more variation. Therefore, we expect that our GWAS method will encounter a lower multiple testing burden, less population stratification and lower rates of confounding bias in analyses of viral sequence data. On the other hand, the higher rates of recombination exhibited by many viruses may be problematic for tree building and thus for treeWAS as well. Both Bartha *et al.* [119] and Power *et al.* [116], used PCA to correct for population structure in GWAS analyses of HIV. The success of these applications has been mixed, however, with no significant findings being identified in the former study. The applications of treeWAS to synthetic accessory genome data in Chapter 4 provide our closest approximation to viral genomic analyses. Given the positive results of these analyses, especially as compared to cluster-based and PCA-based methods, it seems reasonable to expect that treeWAS should be applicable to some viral datasets. It would be interesting to confirm this hypothesis in future, by applying treeWAS to empirical sequence data from real viral samples, preferably beginning with analyses of well-characterised phenotypes. Overall, provided recombination does not vastly exceed mutation, we may find that treeWAS can be a powerful method of analysis in the smaller and more diverse genomes of many viral samples.

### 6.4.2 Detecting rare variants and impenetrant associations

GWAS has been a valuable addition to the field, but it nevertheless remains a fairly blunt instrument. Approaches to association testing, including our own, have become increasingly proficient at identifying strong associations with Mendelian traits, like antibiotic resistance. It is, of course, useful to identify new genes that are simply correlated with or co-evolve with a phenotype. These may account for much of the variation in some phenotypes. But, if the relationships among genes and phenotypes can be as complex as we suspect, then GWAS methods have a lot of room for improvement. A logical next step is to expand the scope of treeWAS by developing association tests that will be attuned to the larger number of rare variants and impenetrant associations that underlie complex traits, like virulence and host association.

Our current measures of association, especially Scores 1 and 3, assume that causal sites will have allele frequencies similar to the phenotype. We assume, therefore, that relevant causal sites have been exposed to selective pressures and that their relative frequency of their alleles (i.e., MAFs) have been altered as a result. This makes treeWAS less likely to identify rare variants (which have arisen recently in few individuals) as associated with a phenotype that is not similarly rare in the sample. It also makes treeWAS less likely identify incompletely penetrant associations (alleles that are prevalent in the sample, but which are only responsible for the phenotype in a subset of individuals). These associated sites may have arisen through random neutral mutations that have only become adaptive due to a change in environment or a new epistatic mutation [369, 370].

Our current Score 2 is somewhat successful in detecting these types of associations, as demonstrated by the identification of associations arising through complementary pathways in Set B. But, to detect marginal associations with allele-based scores, we will need to consider alternative measures of association. A suitable allele-based score would presumably have to be scaled by the relative frequencies of the phenotype and genotype in question [371]. Available allele-based association tests that are scaled by frequency include the burden test, $X^2$ test and Fisher's exact test. Although we found that our current Score 1 out-performed the Fisher test when both were implemented in treeWAS and applied to the simulated datasets analysed in Chapter 3 (data not shown), the reverse may be true in different circumstances. We may want to consider implementing some proportional measures of association and testing their efficacy by simulating data with weaker or less frequent associations to the phenotype. Especially given the widespread linkage in bacterial genomes, it will not be trivial to tease apart all of the rare and impenetrant associations underlying bacterial phenotypes [314]. But, if we can make both absolute and relative measures of association strength available in treeWAS, we may be able to enhance future efforts to fully characterise the genetic basis of microbial traits.

### 6.4.3 Accounting for phylogenetic uncertainty

The phylogenetic tree forms the backbone of our GWAS method. Yet, we can never be certain that we have inferred the true tree. This concern will be heightened in samples where phylogenetic signal must contend with the conflicting influence of recombination. So, instead of assuming that we can rely on the inferred tree, it may be preferable to ensure that phylogenetic uncertainty is acknowledged and addressed. It would be useful to perform a sensitivity analysis in future, to get a better idea of how uncertainty or bias in the estimated phylogeny impacts the performance of treeWAS.

Future work will be needed to fully develop and implement a method of incorporating phylogenetic uncertainty into the treeWAS approach. However, we can sketch out a reasonable proposal here. It would be natural to begin by estimating the amount of phylogenetic uncertainty associated with a given sample or tree. If the current methods of phylogenetic reconstruction implemented in treeWAS were to be used, we would need to add an additional estimation step. Resampling procedures like the bootstrap or jackknife would allow us to quantify the uncertainty associated with an inferred tree [372–374]. Better yet, we could use a Bayesian phylogenetic approach, like BEAST or MrBayes [210, 375], to simultaneously estimate the tree and associated uncertainty.

Bayesian reconstruction methods would also allow us to replace the point estimate of a tree produced by other phylogenetic methods with a representative distribution of possible trees. To address uncertainty in treeWAS, we could then select a set of likely trees by sampling from the posterior distribution. We could then repeat the treeWAS analysis over this set of trees, avoiding the bias inherent in assuming that any single "best" tree was reliable. This would, of course, produce multiple sets of results. Whereas related approaches often take the mean of multiple sets of findings, in our case, a more sensible approach might be to accept only significant sites present in the findings of a large proportion (e.g., $\geq 80\%$) of the analyses performed on different trees. By pooling these significant findings together, we should be able to identify associations without falling prey to the biases or errors inherent in any particular reconstruction of the sample phylogeny. Instead, this approach may allow us to identify a set of consensus findings that is representative of the space of probable trees.

### 6.4.4 Testing for epistatic interactions

GWAS brings new insights with the discovery of each independent, trait-associated locus. Nevertheless, our understanding of most microbial phenotypes will remain incomplete until we consider the interactions between sites. Indeed, epistatic relationships, protein-protein interactions, and compensatory mechanisms are widely known to shape bacterial evolution and phenotypic variation, especially for complex traits [376–378].

We suspect that we may also be able to use treeWAS to test for epistatic relationships between two sites, much in the way that we have used it to test for associations between genotype and phenotype. In fact, Melnyk *et al.* [361] have already used treeWAS to identify associations between genes, by using the presence or absence of a pathogenicity-associated island as the "phenotype" in their GWAS study. This analysis was more narrowly focussed than the epistasis analysis we envision, but their work nonetheless confirms that treeWAS can be used to identify interactions between genes.

Preceding the epistasis test with a filtering step may be necessary, as the number of potential interactions between pairs of loci can be overwhelming, even in the relatively small genomes of bacteria. A genome of just 10,000 sites will amount to nearly 50 million pairwise tests. Cordoner *et al.* [379] have shown that reducing the number of sites submitted to pairwise testing can dramatically improve both the power and precision of epistasis analyses. They recommended filtering sites by MAF, or by a biological parameter like hydrophobicity [379]. Even an uncorrected $X^2$ test can be useful to eliminate sites with a low probability of significance in the epistasis analysis. In our case, sites could be filtered by their individual significance in treeWAS analysis, performed prior to epistasis testing.

As a proof of concept, we used a modified version of treeWAS to expand our exploratory analysis from Section 5.4.2, where we had examined interactions among invasive disease loci in *N. meningitidis*. This time, we tested for epistasis between every accessory gene and each of the 12 previously-identified significant virulence-associated genes (see Table 5.7). We measured the strength of interactions by calculating Score 2, quantifying the association between pairs of genes. To estimate the null distribution of gene-gene associations, while accounting for confounding factors, we also computed Score 2 for pairs of sites drawn from the simulated dataset. To reduce the number of comparisons required in this exploratory analysis, we restricted $N_{sim}$ to 1,000 (amounting to nearly 500,000 pairwise tests in the simulated dataset). The empirical distribution of interactions between genes and virulence-associated genes was then compared the null distribution of interactions between gene-gene pairs from the simulated dataset. We found significant Score 2 values for six additional loci (see Table A.9) with significant positive interactions to our initial set of virulence-associated genes. With the exception of one hypothetical protein (NEIS1357), each of these genes has a known or previously-proposed connection to pathogenicity. The competence protein encoded by NEIS0041 is known to interact with the three previously-identified *comEA-like* loci, to facilitate transformation, and to promote virulence [359, 380]. The three O-antigen *rfb* genes promote invasiveness through their roles in capsular synthesis, lipooligosaccharide biosynthesis, and adhesion to host cells [381–384]. Likewise, the outer membrane pilus protein encoded by NEIS0213 (*pglA*) participates in invasive disease by facilitating interaction with and adhesion to host cells [384,385]. Overall, we were delighted with these results. We had hoped to show that, in theory, treeWAS could be adapted to test for epistasis as well as association. In fact, we found that treeWAS was not only capable of testing for interactions, but that it may have already proved useful in uncovering epistatic relationships between significant genes as well as genes not identified in the initial association analysis.

This preliminary examination proved to be a powerful demonstration of the potential benefits that might be gained by extending and applying treeWAS to the analysis of epistasis. Of course, much additional work will be needed to tailor treeWAS to this purpose and to test any proposed modifications. Among other things, we will need to consider which measures of association will be best suited to this new purpose. For example, in the above analysis, Scores 1 and 3 were excluded because the widespread correlations among genetic loci produced atypical behaviour and left-tailed null distributions. As in our discussion of rare and impenetrant variants, above, we may find that frequency-adjusted measures of association are better suited to this task than our current measures of allele-based association. Another consideration that will need to be addressed is the relationship between correlation and LD. Our phylogenetic, simulation-based approach is ideally tailored to control for chance associations due to physical linkage, in purely clonal oganisms. As we have noted, however, the short-range LD within recombination fragments may generate small numbers of false positive findings, especially in smaller or less diverse samples. Relevant in association studies, this concern will only be heightened in epistasis analyses. Finally, if we wish for treeWAS to be broadly applicable to epistasis testing, we will need to undertake additional efforts to tackle the considerable computational burden that comes with even filtered-down pairwise testing. In time, we expect that these concerns and many others can be addressed by appropriate modifications and additions to the treeWAS method. Our first attempt at epistasis testing certainly encourages us to make this a focus of future work. We have little doubt that expanding our GWAS framework to consider the interactions between sites will make it possible to produce deeper insights into the true complex molecular nature of microbial phenotypes.

## 6.5 Concluding remarks

Genetic data analysis has revolutionised the way we understand and respond to the microbes that surround us. Over the past century, the knowledge base established with experimental methods and observational studies has significantly improved our ability to mitigate the health risks posed by infectious diseases, by informing public health campaigns and guiding drug development. Today, we face new pressures, as globalisation drives up the rate and extent of pathogen transmission, and rapid evolution continues to equip microbial pathogens with novel mechanisms of immune escape and antibiotic resistance. But, we face new opportunities, too. Improvements in sequencing technologies have produced an increasing abundance of genetic data for an ever-widening range of microbial organisms. The analysis of sequence data is already leading to discoveries that will help us to improve treatment efficacy, identify new drug targets, and design surveillance schemes to detect emerging threats. Furthermore, genetic data analysis can substantially increase the pace of discovery over conventional laboratory-based techniques alone. To capitalise on the opportunities presented by the sequencing era, however, it is essential that the accumulation of whole-genome sequences does not outpace the development of statistical and computational tools for their analysis.

GWAS studies represent a valuable, versatile addition to the genetic data analysis toolbox. Association studies will not supplant experimental analyses as a means of discovery. Instead, by focusing experimental efforts towards appropriate candidate genes, GWAS methods will enhance the benefits that can be derived from laboratory work. The power of GWAS methods, established in human genetics, is now becoming clear in the microbial domain. Although the development of appropriate microbial GWAS methodology is still in its early days, microbial GWAS studies have already uncovered novel associations with critical phenotypes, like antibiotic resistance [49, 114, 116, 121–123, 125–130], transmissibility [131], host- and tissue-specificity [106, 118, 124, 131], toxicity [120], and virulence [80, 119, 155]. As association studies are applied to additional phenotypes and larger samples, we expect to see the number of discoveries expand. Even greater progress will be achieved if microbial GWAS methods can better account for ancestry and recombination, incorporate new measures of association, apply to a wider variety of organisms, adjust for uncertainty, and consider epistatic interactions. As the first purpose-built microbial GWAS methods emerge, the evidence suggests that GWAS methods will have great power to provide insight into the genetic mechanisms that give rise to the phenotypic diversity observed across the microbial world. We hope that our contributions to microbial GWAS methodology will ultimately help to advance our understanding of health-relevant microbial traits and that the research presented in this thesis will be useful to others who share in this pursuit.

*My sincerest thanks to each of you for reading all the way through this thesis!*

# Appendix A

# Appendix

| | ID | ENA Accession | Cefixime Resistance | Cefixime MIC | Cefixime MIC Rank | Country | Clinic | Year | Sexual Orientation | MLST |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 27429 | ERR191730 | R | 0.25 | 145 | USA | CHI | 2009 | MSM | 1901 |
| 2 | 27457 | ERR191731 | S | 0.015 | 17.5 | USA | CHI | 2009 | MSM | 1901 |
| 3 | 27315 | ERR191732 | R | 0.25 | 145 | USA | CHI | 2009 | MSM | 1901 |
| 4 | 27422 | ERR191733 | S | 0.015 | 17.5 | USA | CHI | 2009 | MSM | 1901 |
| 5 | 27318 | ERR191734 | R | 0.5 | 195 | USA | CHI | 2009 | MSM | 1901 |
| 6 | 27467 | ERR191735 | S | 0.015 | 17.5 | USA | CHI | 2009 | MSM | 8110 |
| 7 | 27367 | ERR191736 | R | 0.25 | 145 | USA | CHI | 2009 | MSM | 1901 |
| 8 | 27276 | ERR191737 | S | 0.015 | 17.5 | USA | CHI | 2009 | MSM | 1588 |
| 9 | 27256 | ERR191738 | R | 0.25 | 145 | USA | DEN | 2009 | MSM | 8126 |
| 10 | 27230 | ERR191739 | S | 0.015 | 17.5 | USA | LVG | 2009 | MSM | 1580 |
| 11 | 27431 | ERR191740 | S | 0.06 | 82 | USA | DTR | 2009 | MSW | 1893 |
| 12 | 27411 | ERR191741 | S | 0.06 | 82 | USA | DTR | 2009 | MSW | 1893 |
| 13 | 27375 | ERR191746 | R | 0.25 | 145 | USA | HON | 2009 | MSM | 1901 |
| 14 | 27238 | ERR191747 | S | 0.015 | 17.5 | USA | HON | 2009 | MSM | 9363 |
| 15 | 27379 | ERR191748 | R | 0.25 | 145 | USA | HON | 2009 | MSM | 1901 |
| 16 | 27405 | ERR191749 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1579 |
| 17 | 27414 | ERR191750 | R | 0.25 | 145 | USA | HON | 2009 | MSM | 1901 |
| 18 | 27325 | ERR191751 | S | 0.06 | 82 | USA | SDG | 2009 | MSM | 1901 |
| 19 | 27391 | ERR191752 | R | 0.25 | 145 | USA | HON | 2009 | MSW | 1901 |
| 20 | 27343 | ERR191753 | S | 0.06 | 82 | USA | HON | 2009 | MSW | 7823 |
| 21 | 27337 | ERR191754 | R | 0.25 | 145 | USA | HON | 2009 | MSM | 1901 |
| 22 | 27455 | ERR191755 | S | 0.06 | 82 | USA | SDG | 2009 | MSM | 1901 |
| 23 | 27301 | ERR191756 | R | 0.25 | 145 | USA | HON | 2009 | MSM | 1901 |
| 24 | 27404 | ERR191757 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 25 | 27291 | ERR191758 | R | 0.25 | 145 | USA | HON | 2009 | MSM | 1901 |
| 26 | 27319 | ERR191759 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 27 | 27374 | ERR191760 | R | 0.5 | 195 | USA | HON | 2009 | MSM | 1901 |
| 28 | 27259 | ERR191761 | S | 0.03 | 51.5 | USA | SDG | 2009 | MSM | 1901 |
| 29 | 27473 | ERR191762 | R | 0.25 | 145 | USA | LAX | 2009 | MSW | 1901 |
| 30 | 27466 | ERR191763 | S | 0.06 | 82 | USA | ORA | 2009 | MSW | 7371 |
| 31 | 27357 | ERR191764 | R | 0.25 | 145 | USA | LAX | 2009 | MSW | 1901 |
| 32 | 27456 | ERR191765 | S | 0.03 | 51.5 | USA | ORA | 2009 | MSW | 1901 |
| 33 | 27444 | ERR191766 | R | 0.5 | 195 | USA | LAX | 2009 | MSM | 1901 |
| 34 | 27237 | ERR191767 | S | 0.015 | 17.5 | USA | LAX | 2009 | MSM | 9363 |
| 35 | 27341 | ERR191768 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 36 | 27307 | ERR191769 | S | 0.03 | 51.5 | USA | LVG | 2009 | MSW | 7823 |
| 37 | 27469 | ERR191770 | R | 0.25 | 145 | USA | LVG | 2009 | MSM | 1901 |
| 38 | 27380 | ERR191771 | S | 0.06 | 82 | USA | LVG | 2009 | MSM | 1901 |
| 39 | 27306 | ERR191772 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 40 | 27265 | ERR191773 | S | 0.06 | 82 | USA | LVG | 2009 | MSW | 1901 |
| 41 | 27336 | ERR191774 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 42 | 27313 | ERR191775 | S | 0.06 | 82 | USA | LVG | 2009 | MSW | 1901 |
| 43 | 27312 | ERR191776 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 44 | 27330 | ERR191777 | S | 0.03 | 51.5 | USA | LVG | 2009 | MSW | 1901 |
| 45 | 27356 | ERR191778 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 46 | 27292 | ERR191779 | S | 0.015 | 17.5 | USA | LVG | 2009 | MSW | 8154 |
| 47 | 27326 | ERR191780 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 48 | 27442 | ERR191781 | S | 0.03 | 51.5 | USA | LVG | 2009 | MSW | 1901 |
| 49 | 27390 | ERR191782 | R | 0.25 | 145 | USA | LVG | 2009 | MSW | 1901 |
| 50 | 27320 | ERR191783 | S | 0.06 | 82 | USA | LVG | 2009 | MSW | 1901 |
| 51 | 27358 | ERR191784 | R | 0.25 | 145 | USA | PHI | 2009 | MSM | 1901 |
| 52 | 27437 | ERR191785 | S | 0.06 | 82 | USA | PHI | 2009 | MSM | 1901 |
| 53 | 27418 | ERR191786 | R | 0.25 | 145 | USA | PHX | 2009 | MSM | 1901 |
| 54 | 27410 | ERR191787 | S | 0.03 | 51.5 | USA | PHX | 2009 | MSM | 1901 |
| 55 | 27370 | ERR191788 | R | 0.25 | 145 | USA | PHX | 2009 | MSM | 1901 |
| 56 | 27264 | ERR191789 | S | 0.06 | 82 | USA | PHX | 2009 | MSM | 1901 |
| 57 | 27447 | ERR191790 | R | 0.5 | 195 | USA | POR | 2009 | MSM | 1901 |
| 58 | 27231 | ERR191791 | S | 0.015 | 17.5 | USA | POR | 2009 | MSM | 1580 |
| 59 | 27298 | ERR191792 | R | 0.25 | 145 | USA | POR | 2009 | MSM | 1901 |
| 60 | 27271 | ERR191793 | S | 0.06 | 82 | USA | POR | 2009 | MSM | 1893 |
| 61 | 27359 | ERR191794 | R | 0.25 | 145 | USA | POR | 2009 | MSM | 1901 |
| 62 | 27415 | ERR191795 | S | 0.015 | 17.5 | USA | LVG | 2009 | MSM | 1901 |
| 63 | 27406 | ERR191796 | R | 0.25 | 145 | USA | SDG | 2009 | MSM | 1901 |
| 64 | 27475 | ERR191797 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 65 | 27328 | ERR191798 | R | 0.25 | 145 | USA | SDG | 2009 | MSW | 1901 |
| 66 | 27460 | ERR191799 | S | 0.03 | 51.5 | USA | SDG | 2009 | MSW | 1901 |
| 67 | 27371 | ERR191800 | R | 0.25 | 145 | USA | SDG | 2009 | MSM | 1580 |

**Table A.1. Isolates in the *N. gonorrhoeae* dataset (I/III).**

| | ID | ENA Accession | Cefixime Resistance | Cefixime MIC | Cefixime MIC Rank | Country | Clinic | Year | Sexual Orientation | MLST |
|---|---|---|---|---|---|---|---|---|---|---|
| 68 | 27471 | ERR191801 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 6712 |
| 69 | 27419 | ERR191802 | R | 0.25 | 145 | USA | SDG | 2009 | MSM | 1580 |
| 70 | 27235 | ERR191803 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1583 |
| 71 | 27389 | ERR191804 | R | 0.25 | 145 | USA | SDG | 2009 | MSM | 1901 |
| 72 | 27395 | ERR191805 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 73 | 27401 | ERR191806 | R | 0.25 | 145 | USA | SDG | 2009 | MSW | 1901 |
| 74 | 27383 | ERR191807 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSW | 1901 |
| 75 | 27385 | ERR191808 | R | 0.25 | 145 | USA | SEA | 2009 | MSM | 1901 |
| 76 | 27365 | ERR191809 | S | 0.015 | 17.5 | USA | SEA | 2009 | MSM | 1901 |
| 77 | 27400 | ERR191810 | R | 0.25 | 145 | USA | SEA | 2009 | MSM | 1901 |
| 78 | 27285 | ERR191811 | S | 0.06 | 82 | USA | SEA | 2009 | MSM | 1902 |
| 79 | 27245 | ERR191812 | S | 0.03 | 51.5 | USA | SFO | 2009 | MSM | 9363 |
| 80 | 27251 | ERR191813 | S | 0.015 | 17.5 | USA | SFO | 2009 | MSM | 9363 |
| 81 | 27452 | ERR191814 | R | 0.25 | 145 | USA | SFO | 2009 | MSM | 1901 |
| 82 | 27435 | ERR191815 | S | 0.03 | 51.5 | USA | SFO | 2009 | MSM | 1901 |
| 83 | 27346 | ERR191816 | R | 0.25 | 145 | USA | SFO | 2009 | MSW | 1901 |
| 84 | 27449 | ERR191817 | S | 0.125 | 98 | USA | SFO | 2009 | MSW | 1901 |
| 85 | 27443 | ERR191818 | R | 0.25 | 145 | USA | SFO | 2009 | MSM | 1901 |
| 86 | 27279 | ERR191819 | S | 0.015 | 17.5 | USA | SFO | 2009 | MSM | 1588 |
| 87 | 27316 | ERR191820 | R | 0.25 | 145 | USA | CHI | 2010 | MSM | 1901 |
| 88 | 27254 | ERR191821 | S | 0.06 | 82 | USA | CHI | 2010 | MSM | 1580 |
| 89 | 27310 | ERR191822 | R | 0.25 | 145 | USA | CHI | 2010 | MSM | 1901 |
| 90 | 27461 | ERR191823 | S | 0.03 | 51.5 | USA | CHI | 2009 | MSM | 1901 |
| 91 | 27453 | ERR191824 | R | 0.25 | 145 | USA | CHI | 2010 | MSM | 1901 |
| 92 | 27260 | ERR191825 | S | 0.03 | 51.5 | USA | CHI | 2010 | MSM | 9363 |
| 93 | 27373 | ERR223603 | R | 0.5 | 195 | USA | CHI | 2010 | MSM | 1901 |
| 94 | 27240 | ERR223604 | S | 0.03 | 51.5 | USA | CHI | 2010 | MSM | 9363 |
| 95 | 27440 | ERR223605 | R | 0.5 | 195 | USA | CLE | 2010 | MSW | 1901 |
| 96 | 27361 | ERR223606 | S | 0.06 | 82 | USA | DTR | 2010 | MSW | 1893 |
| 97 | 27470 | ERR223607 | R | 0.25 | 145 | USA | DEN | 2010 | MSM | 1901 |
| 98 | 27288 | ERR223608 | S | 0.125 | 98 | USA | ALB | 2010 | MSM | 1901 |
| 99 | 27462 | ERR223609 | R | 0.25 | 145 | USA | DEN | 2010 | MSM | 1901 |
| 100 | 27381 | ERR223610 | S | 0.125 | 98 | USA | ALB | 2010 | MSM | 1901 |
| 101 | 27299 | ERR223611 | R | 0.25 | 145 | USA | DEN | 2010 | MSW | 1901 |
| 102 | 27450 | ERR223612 | S | 0.015 | 17.5 | USA | PHX | 2010 | MSW | 8126 |
| 103 | 27345 | ERR223613 | R | 0.25 | 145 | USA | GRB | 2010 | MSM | 1901 |
| 104 | 27463 | ERR223614 | S | 0.03 | 51.5 | USA | BAL | 2009 | MSM | 1901 |
| 105 | 27421 | ERR223615 | R | 0.25 | 145 | USA | HON | 2010 | MSM | 1901 |
| 106 | 27258 | ERR223616 | S | 0.03 | 51.5 | USA | SFO | 2010 | MSM | 9363 |
| 107 | 27287 | ERR223619 | R | 0.25 | 145 | USA | HON | 2010 | MSM | 1901 |
| 108 | 27295 | ERR223620 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 109 | 27352 | ERR223621 | R | 0.5 | 195 | USA | HON | 2010 | MSM | 1901 |
| 110 | 27327 | ERR223622 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 111 | 27284 | ERR223623 | R | 0.25 | 145 | USA | HON | 2010 | MSW | 8129 |
| 112 | 27255 | ERR223624 | S | 0.03 | 51.5 | USA | HON | 2010 | MSW | 9363 |
| 113 | 27355 | ERR223625 | R | 0.25 | 145 | USA | HON | 2010 | MSM | 8129 |
| 114 | 27261 | ERR223626 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 9363 |
| 115 | 27368 | ERR223627 | R | 0.5 | 195 | USA | LAX | 2010 | MSM | 1901 |
| 116 | 27428 | ERR223628 | S | 0.06 | 82 | USA | LAX | 2010 | MSM | 1901 |
| 117 | 27438 | ERR223629 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 118 | 27360 | ERR223630 | S | 0.125 | 98 | USA | LAX | 2010 | MSM | 1901 |
| 119 | 27387 | ERR223631 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 120 | 27241 | ERR223632 | S | 0.125 | 98 | USA | LAX | 2010 | MSM | 9363 |
| 121 | 27331 | ERR223633 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 122 | 27242 | ERR223634 | S | 0.03 | 51.5 | USA | LAX | 2010 | MSM | 9363 |
| 123 | 27324 | ERR223635 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 124 | 27441 | ERR223636 | S | 0.03 | 51.5 | USA | SDG | 2009 | MSM | 1901 |
| 125 | 27413 | ERR223637 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 126 | 27342 | ERR223638 | S | 0.06 | 82 | USA | SDG | 2009 | MSM | 1901 |
| 127 | 27363 | ERR223639 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 128 | 27340 | ERR223640 | S | 0.03 | 51.5 | USA | SDG | 2009 | MSM | 1901 |
| 129 | 27317 | ERR223641 | R | 0.5 | 195 | USA | LAX | 2010 | MSM | 1901 |
| 130 | 27253 | ERR223642 | S | 0.03 | 51.5 | USA | LAX | 2010 | MSM | 9363 |
| 131 | 27314 | ERR223643 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 132 | 27232 | ERR223644 | S | 0.015 | 17.5 | USA | LAX | 2010 | MSM | 9363 |
| 133 | 27398 | ERR223645 | R | 0.25 | 145 | USA | LAX | 2010 | MSM | 1901 |
| 134 | 27424 | ERR223646 | S | 0.06 | 82 | USA | LAX | 2010 | MSM | 1901 |

**Table A.2. Isolates in the *N. gonorrhoeae* dataset (II/III).**

| | ID | ENA Accession | Cefixime Resistance | Cefixime MIC | Cefixime MIC Rank | Country | Clinic | Year | Sexual Orientation | MLST |
|---|---|---|---|---|---|---|---|---|---|---|
| 135 | 27388 | ERR223647 | R | 0.25 | 145 | USA | LVG | 2010 | MSW | 1901 |
| 136 | 27273 | ERR223648 | S | 0.015 | 17.5 | USA | LVG | 2009 | MSW | 1579 |
| 137 | 27309 | ERR223649 | R | 0.25 | 145 | USA | LVG | 2010 | MSW | 1901 |
| 138 | 27234 | ERR223650 | S | 0.06 | 82 | USA | LVG | 2009 | MSW | 1580 |
| 139 | 27386 | ERR223651 | R | 0.25 | 145 | USA | LVG | 2010 | MSM | 1901 |
| 140 | 27281 | ERR223652 | S | 0.03 | 51.5 | USA | LVG | 2009 | MSM | 1588 |
| 141 | 27382 | ERR223653 | R | 0.25 | 145 | USA | LVG | 2010 | MSW | 1901 |
| 142 | 27334 | ERR223654 | S | 0.03 | 51.5 | USA | SFO | 2009 | MSW | 1901 |
| 143 | 27396 | ERR223655 | R | 0.25 | 145 | USA | LVG | 2010 | MSW | 1580 |
| 144 | 27474 | ERR223656 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSW | 1903 |
| 145 | 27427 | ERR223657 | R | 0.25 | 145 | USA | LVG | 2010 | MSW | 8129 |
| 146 | 27300 | ERR223658 | S | 0.06 | 82 | USA | LVG | 2010 | MSW | 7822 |
| 147 | 27412 | ERR223659 | R | 0.25 | 145 | USA | LVG | 2010 | MSW | 1580 |
| 148 | 27280 | ERR223660 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSW | 8130 |
| 149 | 27430 | ERR223661 | R | 0.25 | 145 | USA | MIN | 2010 | MSM | 1901 |
| 150 | 27250 | ERR223662 | S | 0.03 | 51.5 | USA | MIN | 2010 | MSM | 9363 |
| 151 | 27329 | ERR223663 | R | 0.25 | 145 | USA | NYC | 2010 | MSM | 1901 |
| 152 | 27425 | ERR223664 | S | 0.03 | 51.5 | USA | PHI | 2009 | MSM | 1901 |
| 153 | 27393 | ERR223665 | R | 0.25 | 145 | USA | ORA | 2010 | MSM | 1901 |
| 154 | 27426 | ERR223666 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 1901 |
| 155 | 27408 | ERR223667 | R | 0.25 | 145 | USA | ORA | 2010 | MSM | 1580 |
| 156 | 27472 | ERR223668 | S | 0.015 | 17.5 | USA | SDG | 2009 | MSM | 8152 |
| 157 | 27323 | ERR223669 | R | 0.25 | 145 | USA | PHI | 2010 | MSM | 1901 |
| 158 | 27303 | ERR223670 | S | 0.03 | 51.5 | USA | PHI | 2009 | MSM | 1901 |
| 159 | 27451 | ERR223671 | R | 0.25 | 145 | USA | PHX | 2010 | MSM | 1901 |
| 160 | 27458 | ERR223672 | S | 0.015 | 17.5 | USA | PHX | 2009 | MSM | 1901 |
| 161 | 27339 | ERR223673 | R | 0.25 | 145 | USA | PHX | 2010 | MSM | 1580 |
| 162 | 27297 | ERR223674 | S | 0.06 | 82 | USA | PHX | 2010 | MSM | 1901 |
| 163 | 27311 | ERR223675 | R | 0.25 | 145 | USA | PHX | 2010 | MSM | 1580 |
| 164 | 27293 | ERR223676 | S | 0.06 | 82 | USA | PHX | 2010 | MSM | 1901 |
| 165 | 27351 | ERR223677 | R | 0.25 | 145 | USA | PHX | 2010 | MSM | 1580 |
| 166 | 27266 | ERR223678 | S | 0.015 | 17.5 | USA | PHX | 2010 | MSM | 1893 |
| 167 | 27338 | ERR223679 | R | 0.25 | 145 | USA | PHX | 2010 | MSM | 1580 |
| 168 | 27243 | ERR223680 | S | 0.03 | 51.5 | USA | PHX | 2010 | MSM | 9363 |
| 169 | 27304 | ERR223681 | R | 0.25 | 145 | USA | PHX | 2010 | MSM | 1580 |
| 170 | 27248 | ERR223682 | S | 0.015 | 17.5 | USA | PHX | 2010 | MSM | 9363 |
| 171 | 27454 | ERR223683 | R | 0.25 | 145 | USA | PHX | 2010 | MSW | 1901 |
| 172 | 27445 | ERR223684 | S | 0.03 | 51.5 | USA | PHX | 2010 | MSW | 1901 |
| 173 | 27289 | ERR223685 | R | 0.5 | 195 | USA | POR | 2010 | MSM | 1901 |
| 174 | 27392 | ERR223687 | R | 0.5 | 195 | USA | POR | 2010 | MSM | 1901 |
| 175 | 27417 | ERR223689 | R | 0.25 | 145 | USA | POR | 2010 | MSM | 1901 |
| 176 | 27399 | ERR223691 | R | 0.25 | 145 | USA | POR | 2010 | MSM | 1901 |
| 177 | 27277 | ERR223692 | S | 0.03 | 51.5 | USA | LVG | 2009 | MSM | 1588 |
| 178 | 27335 | ERR223693 | R | 0.25 | 145 | USA | POR | 2010 | MSM | 1901 |
| 179 | 27269 | ERR223694 | S | 0.03 | 51.5 | USA | LVG | 2009 | MSM | 1588 |
| 180 | 27416 | ERR223695 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1901 |
| 181 | 27384 | ERR223696 | S | 0.03 | 51.5 | USA | SDG | 2010 | MSM | 1901 |
| 182 | 27348 | ERR223697 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1580 |
| 183 | 27257 | ERR223698 | S | 0.03 | 51.5 | USA | SDG | 2010 | MSM | 9363 |
| 184 | 27436 | ERR222892 | R | 0.25 | 145 | USA | SDG | 2010 | MSW | 1580 |
| 185 | 27423 | ERR222894 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1901 |
| 186 | 27333 | ERR222895 | S | 0.03 | 51.5 | USA | SDG | 2010 | MSM | 1901 |
| 187 | 27344 | ERR222896 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1580 |
| 188 | 27290 | ERR222898 | R | 0.25 | 145 | USA | SDG | 2010 | MSW | 1580 |
| 189 | 27402 | ERR222900 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1580 |
| 190 | 27252 | ERR222901 | S | 0.03 | 51.5 | USA | SDG | 2009 | MSM | 9363 |
| 191 | 27407 | ERR222902 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1580 |
| 192 | 27446 | ERR222904 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1901 |
| 193 | 27354 | ERR222906 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1580 |
| 194 | 27332 | ERR222908 | R | 0.25 | 145 | USA | SDG | 2010 | MSM | 1580 |
| 195 | 27275 | ERR222909 | S | 0.03 | 51.5 | USA | PHX | 2009 | MSM | 1588 |
| 196 | 27272 | ERR222911 | S | 0.03 | 51.5 | USA | SDG | 2009 | MSM | 1588 |
| 197 | 27434 | ERR222931 | S | 0.06 | 82 | USA | SEA | 2010 | MSM | 1901 |
| 198 | 27349 | ERR222933 | S | 0.06 | 82 | USA | SEA | 2010 | MSM | 1901 |
| 199 | 27420 | ERR222935 | S | 0.06 | 82 | USA | SEA | 2010 | MSM | 1901 |
| 200 | 27302 | ERR222937 | S | 0.06 | 82 | USA | SEA | 2009 | MSM | 8129 |

**Table A.3. Isolates in the *N. gonorrhoeae* dataset (III/III).**

| | ID | ENA Accession | Sero-group | Penicillin Resistance | Penicillin MIC | Penicillin MIC Rank | Disease | Source | Country | Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 945 | ERR133700 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 2 | 964 | ERR133711 | B | S | 0.008 | 5 | carrier | throat | Czech | 1993 |
| 3 | 965 | ERR133712 | B | S | 0.008 | 5 | carrier | throat | Czech | 1993 |
| 4 | 969 | ERR133716 | B | S | 0.008 | 5 | carrier | throat | Czech | 1993 |
| 5 | 976 | ERR133723 | B | S | 0.008 | 5 | carrier | throat | Czech | 1993 |
| 6 | 991 | ERR133734 | B | S | 0.008 | 5 | carrier | throat | Czech | 1993 |
| 7 | 1585 | ERR133743 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 8 | 1588 | ERR133745 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 9 | 1618 | ERR137097 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 10 | 1655 | ERR137132 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 11 | 1656 | ERR137133 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 12 | 1885 | ERR137138 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 13 | 1948 | ERR137144 | B | S | 0.008 | 5 | carrier | throat | Czech | 1993 |
| 14 | 1949 | ERR137145 | B | R | 0.063 | 121 | carrier | throat | Czech | 1993 |
| 15 | 2220 | ERR137149 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 16 | 4145 | ERR310532 | B | S | 0.032 | 56.5 | invasive | CSF | Czech | 2000 |
| 17 | 5171 | ERR310530 | B | S | 0.016 | 17.5 | invasive | CSF | Czech | 2000 |
| 18 | 7891 | ERR310539 | B | R | 0.063 | 121 | invasive | blood | Czech | 2004 |
| 19 | 7892 | ERR310540 | B | R | 0.063 | 121 | invasive | blood | Czech | 2004 |
| 20 | 8139 | ERR137154 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 21 | 8141 | ERR137156 | B | S | 0.032 | 56.5 | carrier | throat | Czech | 1993 |
| 22 | 8144 | ERR137159 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 23 | 8149 | ERR137164 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 24 | 8151 | ERR137166 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 25 | 8156 | ERR137170 | B | S | 0.016 | 17.5 | carrier | throat | Czech | 1993 |
| 26 | 9214 | ERR133685 | B | S | 0.016 | 17.5 | invasive | CSF | Czech | 1993 |
| 27 | 9215 | ERR133746 | B | S | 0.016 | 17.5 | invasive | CSF | Czech | 1993 |
| 28 | 14777 | ERR133688 | B | S | 0.032 | 56.5 | invasive | – | Czech | 1993 |
| 29 | 14782 | ERR133690 | B | S | 0.016 | 17.5 | invasive | – | Czech | 1993 |
| 30 | 15172 | ERR137134 | B | S | 0.016 | 17.5 | invasive | – | Czech | 1993 |
| 31 | 15249 | ERR137173 | B | S | 0.032 | 56.5 | invasive | – | Czech | 1993 |
| 32 | 35227 | ERR847079 | B | R | 0.25 | 164 | invasive | blood | UK | 2009 |
| 33 | 35228 | ERR847080 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 34 | 35229 | ERR847081 | B | S | 0.03 | 41 | invasive | CSF | UK | 2009 |
| 35 | 35230 | ERR847082 | B | R | 0.25 | 164 | invasive | blood | UK | 2009 |
| 36 | 35231 | ERR847083 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2009 |
| 37 | 35232 | ERR847084 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |
| 38 | 35234 | ERR847086 | B | S | 0.06 | 100 | invasive | CSF | UK | 2009 |
| 39 | 35235 | ERR847087 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 40 | 35236 | ERR847088 | B | S | 0.03 | 41 | invasive | CSF | UK | 2009 |
| 41 | 35237 | ERR847089 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 42 | 35238 | ERR847090 | B | S | 0.06 | 100 | invasive | CSF | UK | 2009 |
| 43 | 35239 | ERR847091 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |
| 44 | 35240 | ERR847092 | B | R | 0.25 | 164 | invasive | blood | UK | 2009 |
| 45 | 35241 | ERR847093 | B | S | 0.045 | 71 | invasive | blood | UK | 2009 |
| 46 | 35242 | ERR847094 | B | S | 0.003 | 1 | invasive | blood | UK | 2009 |
| 47 | 35243 | ERR847095 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 48 | 35244 | ERR847096 | B | S | 0.023 | 27 | invasive | blood | UK | 2009 |
| 49 | 35245 | ERR847097 | B | R | 0.25 | 164 | invasive | blood | UK | 2009 |
| 50 | 35246 | ERR847098 | B | R | 0.375 | 169.5 | invasive | blood | UK | 2009 |
| 51 | 35247 | ERR847099 | B | S | 0.06 | 100 | invasive | CSF | UK | 2009 |
| 52 | 35248 | ERR847100 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |
| 53 | 35250 | ERR847102 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 54 | 35251 | ERR847103 | B | R | 0.08 | 123 | invasive | blood | UK | 2009 |
| 55 | 35252 | ERR847104 | B | S | 0.045 | 71 | invasive | blood | UK | 2009 |
| 56 | 35253 | ERR847105 | B | S | 0.045 | 71 | invasive | blood | UK | 2009 |
| 57 | 35254 | ERR847106 | B | R | 0.18 | 154 | invasive | blood | UK | 2009 |
| 58 | 35255 | ERR847107 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |
| 59 | 35256 | ERR847108 | B | S | 0.012 | 9.5 | invasive | blood | UK | 2009 |
| 60 | 35257 | ERR847109 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 61 | 35259 | ERR847111 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 62 | 35260 | ERR847112 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |
| 63 | 35261 | ERR847113 | B | S | 0.045 | 71 | invasive | blood | UK | 2009 |
| 64 | 35262 | ERR847114 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |
| 65 | 35263 | ERR847115 | B | S | 0.06 | 100 | invasive | other | UK | 2009 |
| 66 | 35264 | ERR847116 | B | R | 0.125 | 149 | invasive | blood | UK | 2009 |
| 67 | 35265 | ERR847117 | B | S | 0.06 | 100 | invasive | blood | UK | 2009 |

**Table A.4. Isolates in the serogroup B *N. meningitidis* dataset (I/III).**

|  | ID | ENA Accession | Sero-group | Penicillin Resistance | Penicillin MIC | Penicillin MIC Rank | Disease | Source | Country | Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 68 | 35267 | ERR847119 | B | S | 0.03 | 41 | invasive | blood | Iceland | 2009 |
| 69 | 35269 | ERR847121 | B | S | 0.008 | 5 | invasive | blood | UK | 2009 |
| 70 | 35270 | ERR847122 | B | S | 0.03 | 41 | invasive | blood | UK | 2009 |
| 71 | 35272 | ERR847252 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2010 |
| 72 | 35273 | ERR847125 | B | S | 0.045 | 71 | invasive | blood | UK | 2010 |
| 73 | 35274 | ERR847126 | B | R | 0.18 | 154 | invasive | blood | UK | 2010 |
| 74 | 35275 | ERR847127 | B | S | 0.045 | 71 | invasive | joint fluid | UK | 2010 |
| 75 | 35276 | ERR847128 | B | S | 0.023 | 27 | invasive | blood | UK | 2010 |
| 76 | 35277 | ERR847129 | B | S | 0.06 | 100 | invasive | blood | UK | 2010 |
| 77 | 35278 | ERR847130 | B | S | 0.06 | 100 | invasive | blood | UK | 2010 |
| 78 | 35279 | ERR847132 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2010 |
| 79 | 35280 | ERR847133 | B | S | 0.045 | 71 | invasive | blood | UK | 2010 |
| 80 | 35281 | ERR847134 | B | S | 0.045 | 71 | invasive | blood | UK | 2010 |
| 81 | 35282 | ERR847135 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2010 |
| 82 | 35284 | ERR847137 | B | S | 0.047 | 80 | invasive | blood | UK | 2010 |
| 83 | 35285 | ERR847138 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2010 |
| 84 | 35286 | ERR847139 | B | S | 0.06 | 100 | invasive | blood | UK | 2010 |
| 85 | 35287 | ERR847140 | B | S | 0.045 | 71 | invasive | CSF | UK | 2010 |
| 86 | 35288 | ERR847141 | B | S | 0.023 | 27 | invasive | blood | UK | 2010 |
| 87 | 35290 | ERR847143 | B | S | 0.03 | 41 | invasive | blood | UK | 2010 |
| 88 | 35291 | ERR847144 | B | S | 0.045 | 71 | invasive | blood | UK | 2010 |
| 89 | 35292 | ERR847145 | B | S | 0.06 | 100 | invasive | blood | UK | 2010 |
| 90 | 35293 | ERR847146 | B | R | 0.18 | 154 | invasive | blood | UK | 2010 |
| 91 | 35294 | ERR847147 | B | S | 0.012 | 9.5 | invasive | blood | UK | 2010 |
| 92 | 35296 | ERR847149 | B | S | 0.023 | 27 | invasive | blood | UK | 2011 |
| 93 | 35297 | ERR847150 | B | S | 0.03 | 41 | invasive | blood | UK | 2010 |
| 94 | 35298 | ERR847151 | B | S | 0.045 | 71 | invasive | blood | UK | 2011 |
| 95 | 35299 | ERR847152 | B | S | 0.045 | 71 | invasive | blood | UK | 2011 |
| 96 | 35300 | ERR847153 | B | S | 0.04 | 62 | invasive | CSF | UK | 2011 |
| 97 | 35301 | ERR847154 | B | R | 0.18 | 154 | invasive | blood | UK | 2011 |
| 98 | 35302 | ERR847155 | B | S | 0.06 | 100 | invasive | blood | UK | 2011 |
| 99 | 35303 | ERR847156 | B | R | 0.18 | 154 | invasive | blood | UK | 2011 |
| 100 | 35304 | ERR847157 | B | S | 0.03 | 41 | invasive | blood | UK | 2011 |
| 101 | 35305 | ERR847158 | B | R | 0.375 | 169.5 | invasive | CSF | UK | 2011 |
| 102 | 35306 | ERR847159 | B | S | 0.06 | 100 | invasive | blood | UK | 2011 |
| 103 | 35307 | ERR847160 | B | S | 0.06 | 100 | invasive | blood | UK | 2011 |
| 104 | 35308 | ERR847161 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2011 |
| 105 | 35309 | ERR847162 | B | S | 0.06 | 100 | invasive | CSF | UK | 2011 |
| 106 | 35310 | ERR847163 | B | R | 0.12 | 141.5 | invasive | CSF | UK | 2011 |
| 107 | 35311 | ERR847164 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2011 |
| 108 | 35312 | ERR847165 | B | S | 0.03 | 41 | invasive | blood | UK | 2011 |
| 109 | 35313 | ERR847166 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2011 |
| 110 | 35315 | ERR847168 | B | R | 0.18 | 154 | invasive | blood | UK | 2011 |
| 111 | 35316 | ERR847169 | B | S | 0.06 | 100 | invasive | blood | UK | 2011 |
| 112 | 35318 | ERR847171 | B | S | 0.015 | 11 | invasive | CSF | UK | 2011 |
| 113 | 35319 | ERR847172 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2011 |
| 114 | 35320 | ERR847173 | B | R | 0.25 | 164 | invasive | blood | UK | 2011 |
| 115 | 35321 | ERR847174 | B | S | 0.045 | 71 | invasive | blood | UK | 2011 |
| 116 | 35322 | ERR847175 | B | S | 0.03 | 41 | invasive | blood | UK | 2011 |
| 117 | 35324 | ERR847178 | B | S | 0.03 | 41 | invasive | blood | UK | 2011 |
| 118 | 35325 | ERR847179 | B | S | 0.023 | 27 | invasive | blood | UK | 2011 |
| 119 | 35326 | ERR847180 | B | S | 0.06 | 100 | invasive | CSF | UK | 2011 |
| 120 | 35327 | ERR847181 | B | S | 0.03 | 41 | invasive | blood | UK | 2011 |
| 121 | 35328 | ERR847182 | B | S | 0.03 | 41 | invasive | blood | UK | 2011 |
| 122 | 35329 | ERR847183 | B | R | 0.25 | 164 | invasive | blood | UK | 2012 |
| 123 | 35330 | ERR847184 | B | R | 0.25 | 164 | invasive | blood | UK | 2012 |
| 124 | 35331 | ERR847185 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2012 |
| 125 | 35334 | ERR847188 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2012 |
| 126 | 35335 | ERR847189 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2012 |
| 127 | 35337 | ERR847191 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2012 |
| 128 | 35339 | ERR847193 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 129 | 35340 | ERR847194 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 130 | 35341 | ERR847195 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 131 | 35342 | ERR847196 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 132 | 35343 | ERR847197 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 133 | 35344 | ERR847198 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2013 |
| 134 | 35345 | ERR847199 | B | S | 0.045 | 71 | invasive | CSF | UK | 2013 |

**Table A.5. Isolates in the serogroup B *N. meningitidis* dataset (II/III).**

| | ID | ENA Accession | Sero-group | Penicillin Resistance | Penicillin MIC | Penicillin MIC Rank | Disease | Source | Country | Year |
|---|---|---|---|---|---|---|---|---|---|---|
| 135 | 35346 | ERR847200 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2013 |
| 136 | 35347 | ERR847201 | B | S | 0.023 | 27 | invasive | blood | UK | 2013 |
| 137 | 35350 | ERR847204 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 138 | 35351 | ERR847205 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 139 | 35352 | ERR847206 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2013 |
| 140 | 35353 | ERR847207 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2013 |
| 141 | 35354 | ERR847208 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2013 |
| 142 | 35355 | ERR847209 | B | R | 0.18 | 154 | invasive | CSF | UK | 2013 |
| 143 | 35356 | ERR847210 | B | S | 0.023 | 27 | invasive | CSF | UK | 2012 |
| 144 | 35357 | ERR847211 | B | S | 0.045 | 71 | invasive | blood | UK | 2012 |
| 145 | 35358 | ERR847212 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2012 |
| 146 | 35360 | ERR847214 | B | R | 0.12 | 141.5 | invasive | blood | UK | 2012 |
| 147 | 35361 | ERR847215 | B | S | 0.045 | 71 | invasive | blood | UK | 2012 |
| 148 | 35362 | ERR847216 | B | S | 0.06 | 100 | invasive | blood | UK | 2012 |
| 149 | 35363 | ERR847217 | B | R | 0.25 | 164 | invasive | blood | UK | 2012 |
| 150 | 35364 | ERR847218 | B | R | 0.19 | 159 | invasive | blood | UK | 2012 |
| 151 | 35365 | ERR847219 | B | R | 0.094 | 134 | invasive | blood | UK | 2012 |
| 152 | 35368 | ERR847222 | B | R | 0.18 | 154 | invasive | CSF | UK | 2012 |
| 153 | 35369 | ERR847223 | B | S | 0.03 | 41 | invasive | blood | UK | 2012 |
| 154 | 35372 | ERR847226 | B | R | 0.09 | 128.5 | invasive | CSF | UK | 2012 |
| 155 | 35373 | ERR847227 | B | S | 0.06 | 100 | invasive | CSF | UK | 2012 |
| 156 | 35374 | ERR847228 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2012 |
| 157 | 35375 | ERR847229 | B | S | 0.06 | 100 | invasive | blood | UK | 2012 |
| 158 | 35378 | ERR847232 | B | S | 0.045 | 71 | invasive | CSF | UK | 2009 |
| 159 | 35379 | ERR847233 | B | R | 0.25 | 164 | invasive | blood | UK | 2009 |
| 160 | 35381 | ERR847235 | B | R | 0.09 | 128.5 | invasive | blood | UK | 2013 |
| 161 | 35383 | ERR847237 | B | R | 0.45 | 171 | invasive | blood | UK | 2013 |
| 162 | 35385 | ERR847239 | B | S | 0.06 | 100 | invasive | CSF | UK | 2013 |
| 163 | 35386 | ERR847240 | B | S | 0.06 | 100 | invasive | CSF | UK | 2013 |
| 164 | 35387 | ERR847241 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 165 | 35388 | ERR847242 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 166 | 35389 | ERR847243 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 167 | 35390 | ERR847244 | B | S | 0.03 | 41 | invasive | blood | UK | 2013 |
| 168 | 35392 | ERR847246 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 169 | 35394 | ERR847248 | B | R | 0.18 | 154 | invasive | blood | UK | 2013 |
| 170 | 35395 | ERR847249 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |
| 171 | 35396 | ERR847250 | B | S | 0.06 | 100 | invasive | blood | UK | 2013 |

**Table A.6. Isolates in the serogroup B *N. meningitidis* dataset (III/III).**

| | ID | ENA Accession | Sero-group | Disease | Source | Country | Year |
|---|---|---|---|---|---|---|---|
| 1 | 642 | ERS006926 | C | invasive | – | UK | 1996 |
| 2 | 662 | ERR063503 | C | invasive | – | UK | 1997 |
| 3 | 665 | ERR063495 | C | carrier | – | UK | 1997 |
| 4 | 666 | ERR063497 | C | carrier | – | UK | 1997 |
| 5 | 667 | ERR063498 | C | carrier | – | UK | 1997 |
| 6 | 669 | ERR063494 | C | carrier | – | UK | 1997 |
| 7 | 670 | ERR063500 | C | invasive | – | UK | 1997 |
| 8 | 671 | ERR063496 | C | invasive | – | UK | 1997 |
| 9 | 672 | ERR063493 | C | invasive | – | UK | 1997 |
| 10 | 684 | ERR036073 | C | invasive | – | Czech | 1993 |
| 11 | 932 | ERR133693 | C | carrier | throat | Czech | 1993 |
| 12 | 939 | ERR036099 | C | carrier | throat | Czech | 1993 |
| 13 | 940 | ERR036103 | C | carrier | throat | Czech | 1993 |
| 14 | 942 | ERR036104 | C | carrier | throat | Czech | 1993 |
| 15 | 946 | ERR036105 | C | carrier | throat | Czech | 1993 |
| 16 | 948 | ERR036106 | C | carrier | throat | Czech | 1993 |
| 17 | 949 | ERR036107 | C | carrier | throat | Czech | 1993 |
| 18 | 950 | ERR036108 | C | carrier | throat | Czech | 1993 |
| 19 | 952 | ERR036109 | C | carrier | throat | Czech | 1993 |
| 20 | 954 | ERR036110 | C | carrier | throat | Czech | 1993 |
| 21 | 955 | ERR036100 | C | carrier | throat | Czech | 1993 |
| 22 | 957 | ERR036101 | C | carrier | throat | Czech | 1993 |
| 23 | 958 | ERR036102 | C | carrier | throat | Czech | 1993 |
| 24 | 973 | ERR133720 | C | carrier | throat | Czech | 1993 |
| 25 | 977 | ERR133724 | C | carrier | throat | Czech | 1993 |
| 26 | 978 | ERR036112 | C | carrier | throat | Czech | 1993 |
| 27 | 979 | ERR036116 | C | carrier | throat | Czech | 1993 |
| 28 | 981 | ERR133726 | C | carrier | throat | Czech | 1993 |
| 29 | 982 | ERR036118 | C | carrier | throat | Czech | 1993 |
| 30 | 992 | ERR036119 | C | carrier | throat | Czech | 1993 |
| 31 | 993 | ERR036120 | C | carrier | throat | Czech | 1993 |
| 32 | 994 | ERR036121 | C | carrier | throat | Czech | 1993 |
| 33 | 1169 | ERR522738 | C | invasive | CSF | Greece | 1996 |
| 34 | 1170 | ERR522746 | C | carrier | throat | Greece | 1996 |
| 35 | 1178 | ERR522777 | C | invasive | CSF | Greece | 1997 |
| 36 | 1179 | ERR522785 | C | carrier | throat | Greece | 1997 |
| 37 | 1180 | ERR522793 | C | carrier | throat | Greece | 1997 |
| 38 | 1207 | ERR557644 | C | invasive | – | UK | 1999 |
| 39 | 1212 | ERR036122 | C | carrier | throat | Czech | 1993 |
| 40 | 1571 | ERR036113 | C | carrier | throat | Czech | 1993 |
| 41 | 1572 | ERR036114 | C | carrier | throat | Czech | 1993 |
| 42 | 1573 | ERR036115 | C | carrier | throat | Czech | 1993 |
| 43 | 1574 | ERR133736 | C | carrier | throat | Czech | 1993 |
| 44 | 1578 | ERR036060 | C | carrier | throat | Czech | 1993 |
| 45 | 1582 | ERR036064 | C | carrier | throat | Czech | 1993 |
| 46 | 1583 | ERR036065 | C | carrier | throat | Czech | 1993 |
| 47 | 1587 | ERR036066 | C | carrier | throat | Czech | 1993 |
| 48 | 1592 | ERR133754 | C | carrier | throat | Czech | 1993 |
| 49 | 1622 | ERR137101 | C | carrier | throat | Czech | 1993 |
| 50 | 1636 | ERR137115 | C | carrier | throat | Czech | 1993 |
| 51 | 1641 | ERR137120 | C | carrier | throat | Czech | 1993 |
| 52 | 1650 | ERR036067 | C | carrier | throat | Czech | 1993 |
| 53 | 1652 | ERR036068 | C | carrier | throat | Czech | 1993 |
| 54 | 1659 | ERR137137 | C | carrier | throat | Czech | 1993 |
| 55 | 1893 | ERR133683 | C | invasive | CSF | Czech | 1993 |
| 56 | 1941 | ERR137143 | C | carrier | throat | Czech | 1993 |
| 57 | 4193 | ERR522813 | C | carrier | throat | UK | 1999 |
| 58 | 8150 | ERR137165 | C | carrier | throat | Czech | 1993 |
| 59 | 8157 | ERR036069 | C | carrier | throat | Czech | 1993 |
| 60 | 8159 | ERR036070 | C | carrier | throat | Czech | 1993 |
| 61 | 14765 | ERR036061 | C | invasive | – | Czech | 1993 |
| 62 | 14776 | ERR133687 | C | invasive | – | Czech | 1993 |
| 63 | 15024 | ERR133747 | C | invasive | blood | Czech | 1993 |
| 64 | 15032 | ERR036062 | C | invasive | – | Czech | 1993 |
| 65 | 15035 | ERR036063 | C | invasive | – | Czech | 1993 |
| 66 | 15154 | ERR036078 | C | invasive | – | Czech | 1993 |
| 67 | 15174 | ERR036079 | C | invasive | – | Czech | 1993 |

**Table A.7. Isolates in the serogroup C *N. meningitidis* dataset (I/II).**

| | ID | ENA Accession | Sero-group | Disease | Source | Country | Year |
|---|---|---|---|---|---|---|---|
| 68 | 15176 | ERR036080 | C | invasive | – | Czech | 1993 |
| 69 | 15238 | ERR036081 | C | invasive | – | Czech | 1993 |
| 70 | 15242 | ERR036082 | C | invasive | – | Czech | 1993 |
| 71 | 15255 | ERR137175 | C | invasive | CSF | Czech | 1993 |
| 72 | 15261 | ERR036084 | C | invasive | – | Czech | 1993 |
| 73 | 15315 | ERR036074 | C | invasive | – | Czech | 1993 |
| 74 | 15316 | ERR036075 | C | invasive | – | Czech | 1993 |
| 75 | 15319 | ERR036076 | C | invasive | – | Czech | 1993 |
| 76 | 15325 | ERR036086 | C | invasive | – | Czech | 1993 |
| 77 | 15330 | ERR036090 | C | invasive | – | Czech | 1993 |
| 78 | 15336 | ERR036091 | C | invasive | – | Czech | 1993 |
| 79 | 15342 | ERR036093 | C | invasive | – | Czech | 1993 |
| 80 | 15344 | ERR036094 | C | invasive | – | Czech | 1993 |
| 81 | 29738 | ERR558124 | C | invasive | – | UK | 1997 |
| 82 | 29739 | ERR558125 | C | invasive | – | UK | 1997 |
| 83 | 29740 | ERR558126 | C | invasive | – | UK | 1997 |
| 84 | 29913 | ERR1134940 | C | invasive | – | UK | 1999 |
| 85 | 29914 | ERR557947 | C | invasive | – | UK | 1999 |
| 86 | 29915 | ERR1134942 | C | invasive | – | UK | 1999 |
| 87 | 29916 | ERR1134944 | C | invasive | – | UK | 1999 |
| 88 | 29917 | ERR1134946 | C | invasive | – | UK | 1999 |
| 89 | 29919 | ERR1134948 | C | invasive | – | UK | 1999 |
| 90 | 29920 | ERR557949 | C | invasive | – | UK | 1999 |
| 91 | 29921 | ERR1134950 | C | invasive | – | UK | 1999 |
| 92 | 29922 | ERR557950 | C | invasive | – | UK | 1999 |
| 93 | 29923 | ERR1134952 | C | invasive | – | UK | 1999 |
| 94 | 29925 | ERR1134954 | C | invasive | – | UK | 1999 |
| 95 | 29926 | ERR557952 | C | invasive | – | UK | 1999 |
| 96 | 29927 | ERR1134956 | C | invasive | – | UK | 1999 |
| 97 | 30185 | ERR557632 | C | invasive | – | UK | 1996 |
| 98 | 30186 | ERR1134901 | C | invasive | – | UK | 1996 |
| 99 | 30187 | ERR1134903 | C | invasive | – | UK | 1996 |
| 100 | 30188 | ERR557634 | C | invasive | – | UK | 1996 |
| 101 | 30189 | ERR1134905 | C | invasive | – | UK | 1996 |
| 102 | 30190 | ERR557637 | C | carrier | – | UK | 1996 |
| 103 | 30191 | ERR557638 | C | carrier | – | UK | 1996 |
| 104 | 30192 | ERR557639 | C | carrier | – | UK | 1996 |
| 105 | 30193 | ERR557641 | C | carrier | – | UK | 1996 |
| 106 | 30194 | ERR1134907 | C | invasive | – | UK | 1999 |
| 107 | 30195 | ERR1134909 | C | invasive | – | UK | 1999 |
| 108 | 30196 | ERR557645 | C | invasive | – | UK | 1999 |
| 109 | 30197 | ERR1134911 | C | invasive | – | UK | 1999 |
| 110 | 30198 | ERR1134913 | C | invasive | – | UK | 1999 |
| 111 | 30200 | ERR1134917 | C | invasive | – | UK | 1999 |
| 112 | 30201 | ERR1134919 | C | invasive | – | UK | 1999 |
| 113 | 30202 | ERR557648 | C | invasive | – | UK | 1999 |
| 114 | 30203 | ERR1134921 | C | invasive | – | UK | 1999 |
| 115 | 30204 | ERR1134923 | C | carrier | – | UK | 1999 |
| 116 | 30205 | ERR1134925 | C | carrier | – | UK | 1999 |
| 117 | 30206 | ERR557653 | C | carrier | – | UK | 1999 |
| 118 | 30207 | ERR1134927 | C | carrier | – | UK | 1999 |
| 119 | 30208 | ERR1134929 | C | carrier | – | UK | 1999 |
| 120 | 30232 | ERR557668 | C | invasive | – | UK | 1996 |
| 121 | 30233 | ERR1134949 | C | invasive | – | UK | 1996 |
| 122 | 30279 | ERR557631 | C | carrier | – | Portugal | 2012 |
| 123 | 30282 | ERR557633 | C | carrier | – | Portugal | 2012 |
| 124 | 36202 | ERR976804 | C | invasive | CSF | Greece | 1996 |
| 125 | 36203 | ERR976805 | C | carrier | throat | Greece | 1996 |
| 126 | 41784 | ERR063501 | C | invasive | – | UK | 1997 |
| 127 | 41785 | ERR063502 | C | carrier | – | UK | 1997 |
| 128 | 41786 | ERR036071 | C | carrier | throat | Czech | 1993 |
| 129 | 41787 | ERR036083 | C | invasive | – | Czech | 1993 |

**Table A.8. Isolates in the serogroup C *N. meningitidis* dataset (II/II).**

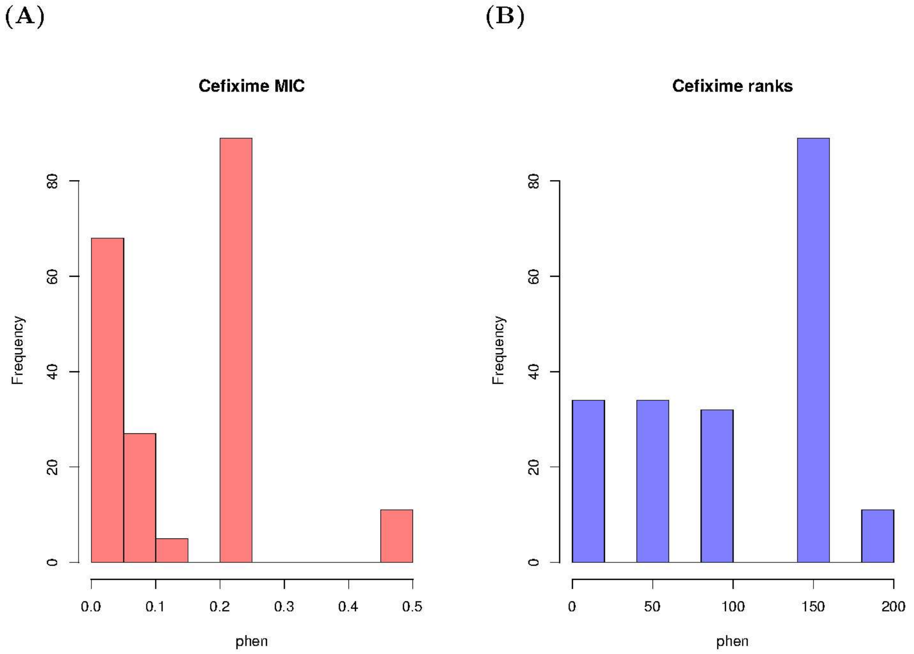


**Figure A.1. Cefixime MIC and rank distributions.** Histograms show the distribution of continuous values for the *N. gonorrhoeae* cefixime resistance phenotype, plotting phenotypic values along the x-axis and counts along the y-axis. **A:** The original MIC values have a highly skewed distribution. **B:** The rank-transformed MIC values are more evenly distributed across the x-axis range. We applied treeWAS to the ranks to give greater weight to the relative differences between individuals.

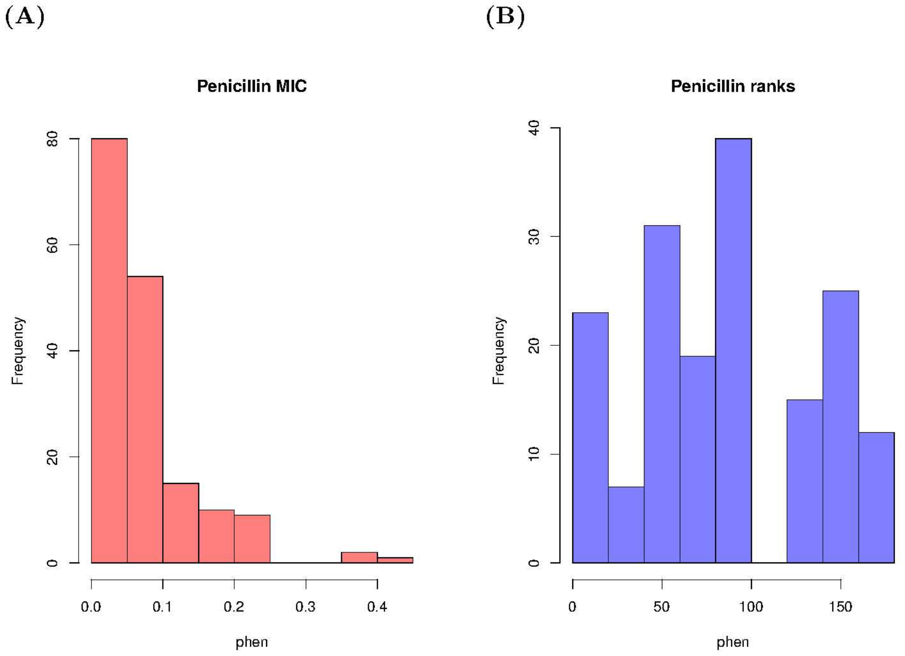
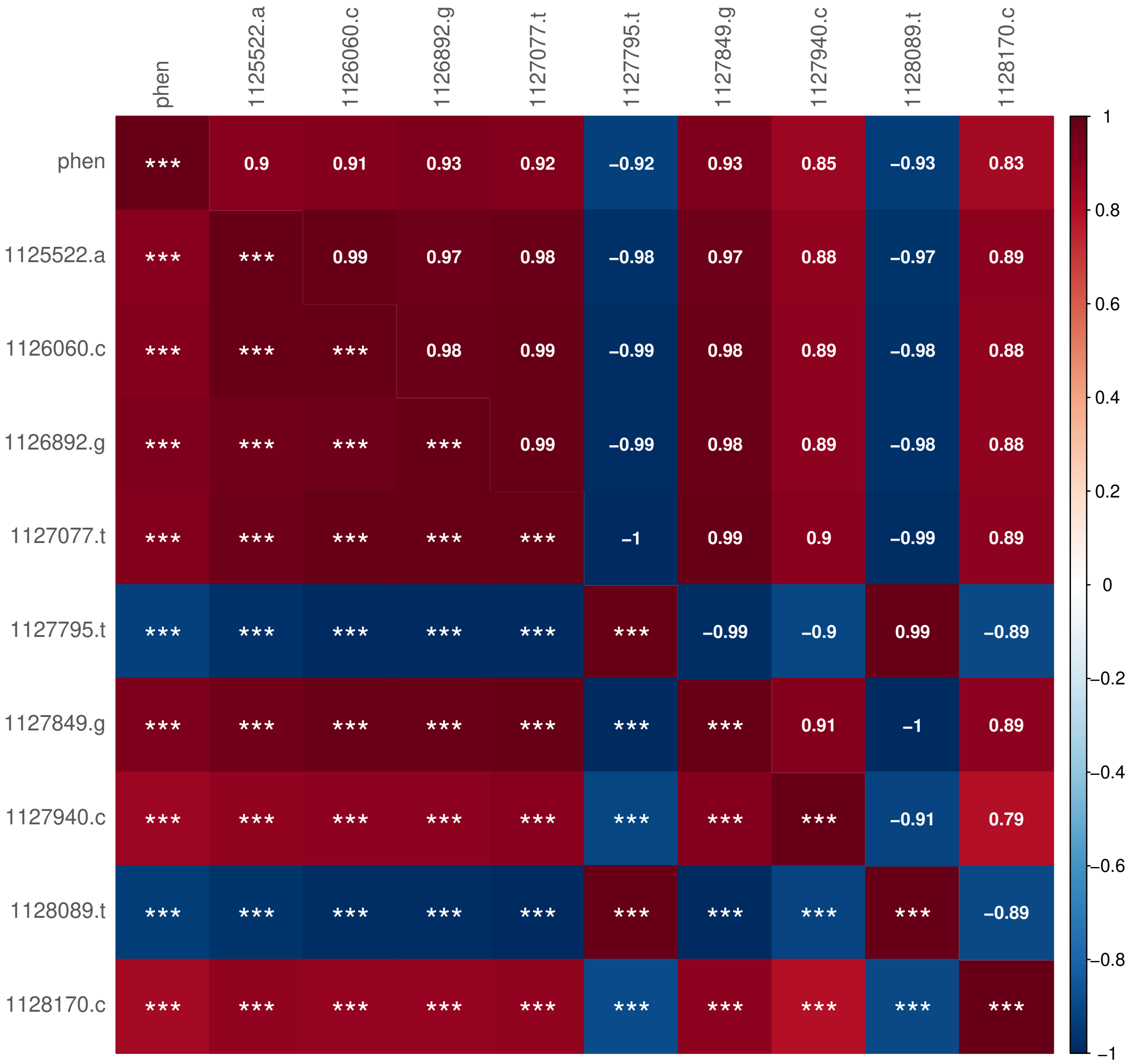


**Figure A.2. Penicillin MIC and rank distributions.** Histograms show the distribution of continuous values for the *N. meningitidis* penicillin resistance phenotype, plotting phenotypic values along the x-axis and counts along the y-axis. **A:** The original MIC values have a highly skewed distribution. **B:** The rank-transformed MIC values are more evenly distributed across the x-axis range. We applied treeWAS to the ranks to give greater weight to the relative differences between individuals.



| | phen | 1125522.a | 1126060.c | 1126892.g | 1127077.t | 1127795.t | 1127849.g | 1127940.c | 1128089.t | 1128170.c |
|---|---|---|---|---|---|---|---|---|---|---|
| phen | *** | 0.9 | 0.91 | 0.93 | 0.92 | −0.92 | 0.93 | 0.85 | −0.93 | 0.83 |
| 1125522.a | *** | *** | 0.99 | 0.97 | 0.98 | −0.98 | 0.97 | 0.88 | −0.97 | 0.89 |
| 1126060.c | *** | *** | *** | 0.98 | 0.99 | −0.99 | 0.98 | 0.89 | −0.98 | 0.88 |
| 1126892.g | *** | *** | *** | *** | 0.99 | −0.99 | 0.98 | 0.89 | −0.98 | 0.88 |
| 1127077.t | *** | *** | *** | *** | *** | −1 | 0.99 | 0.9 | −0.99 | 0.89 |
| 1127795.t | *** | *** | *** | *** | *** | *** | −0.99 | −0.9 | 0.99 | −0.89 |
| 1127849.g | *** | *** | *** | *** | *** | *** | *** | 0.91 | −1 | 0.89 |
| 1127940.c | *** | *** | *** | *** | *** | *** | *** | *** | −0.91 | 0.79 |
| 1128089.t | *** | *** | *** | *** | *** | *** | *** | *** | *** | −0.89 |
| 1128170.c | *** | *** | *** | *** | *** | *** | *** | *** | *** | *** |

**Figure A.3. Correlation between SNPs associated with cefixime resistance.** A visualisation of the correlation matrix describing the similarity between the unique column patterns observed among the 132 core SNPs identified as significantly associated with the cefixime resistance phenotype in *N. gonorrhoeae.* The Pearson's correlation is given above the diagonal and the significance level of each correlation is indicated by the number of asterisks below the diagonal (* = 0.05, ** = 0.01, *** = 0.001).

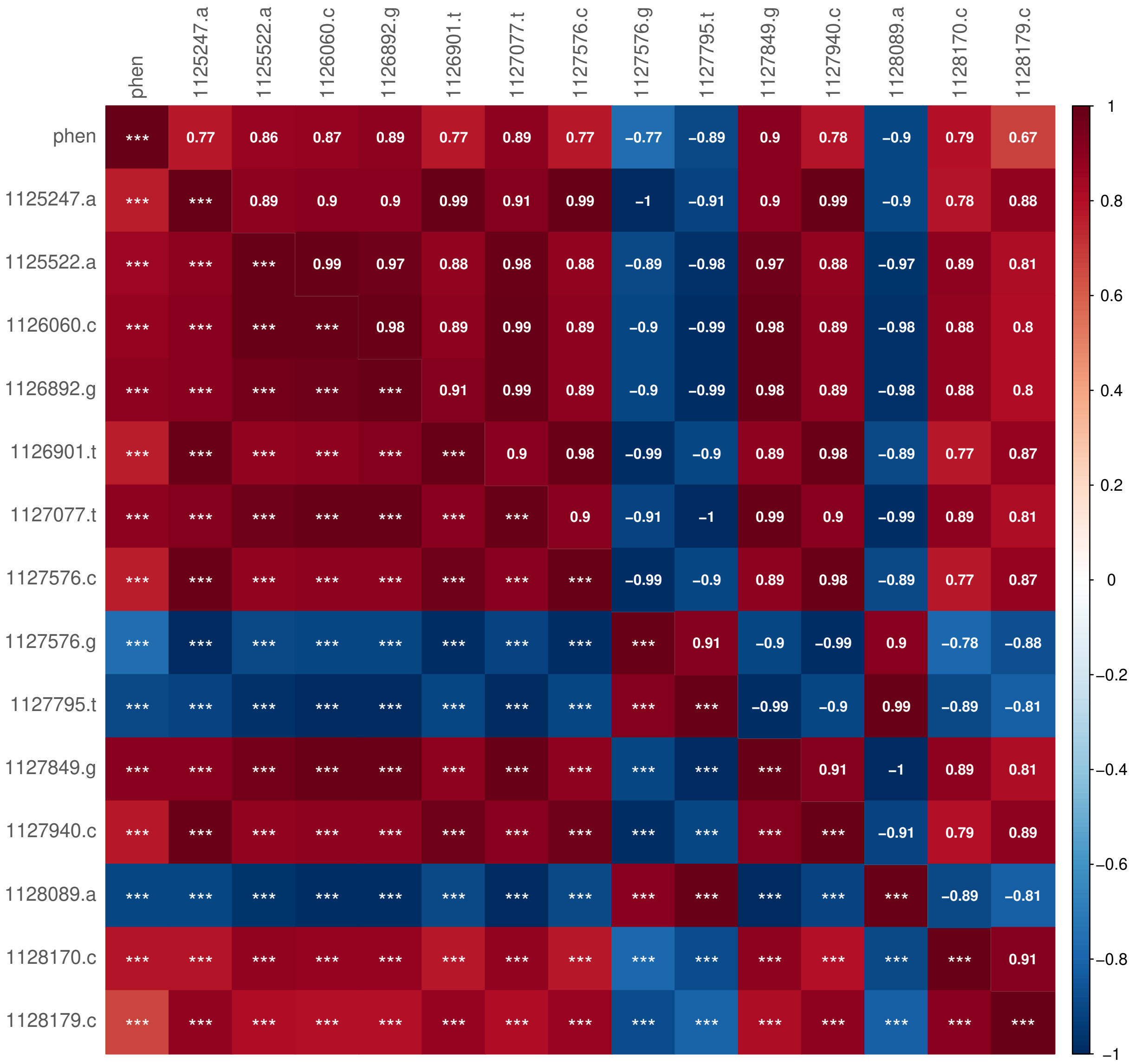


**Figure A.4. Correlation between SNPs associated with cefixime MIC.** A visualisation of the correlation matrix describing the similarity between the 14 unique column patterns observed among the 222 core SNPs identified as significantly associated with the cefixime MIC phenotype in *N. gonorrhoeae*. The Pearson's correlation is given above the diagonal and the significance level of each correlation is indicated by the number of asterisks below the diagonal (* = 0.05, ** = 0.01, *** = 0.001).

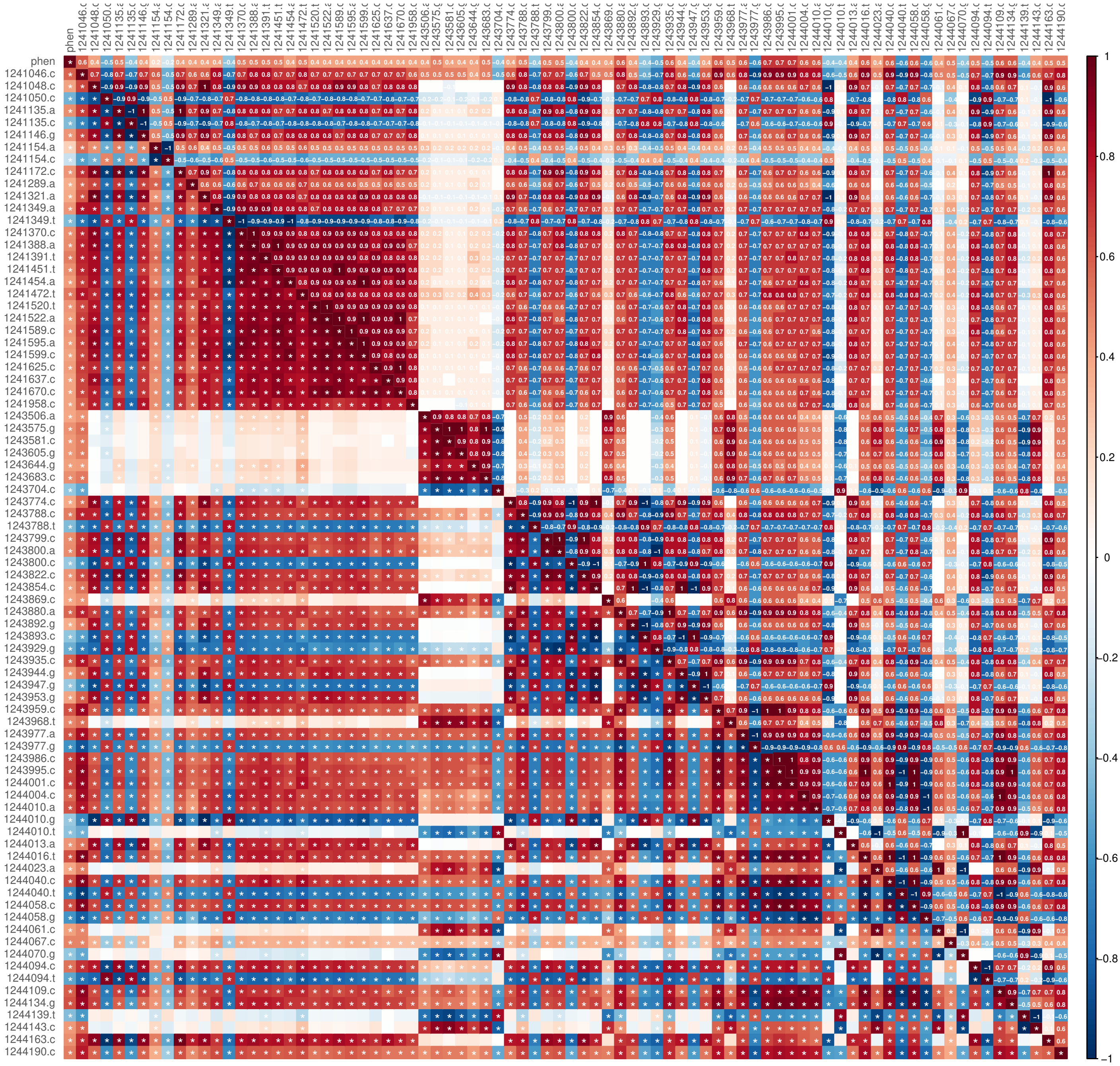


**Figure A.5. Correlation between SNPs associated with penicillin resistance.** A visualisation of the correlation matrix describing the similarity between the unique column patterns observed among the 162 core SNPs identified as significantly associated with the penicillin resistance phenotype in *N. meningitidis*. The Pearson's correlation is given above the diagonal and the significance of each correlation is indicated by the presence of an asterisk below the diagonal ($* \leq 0.05$).

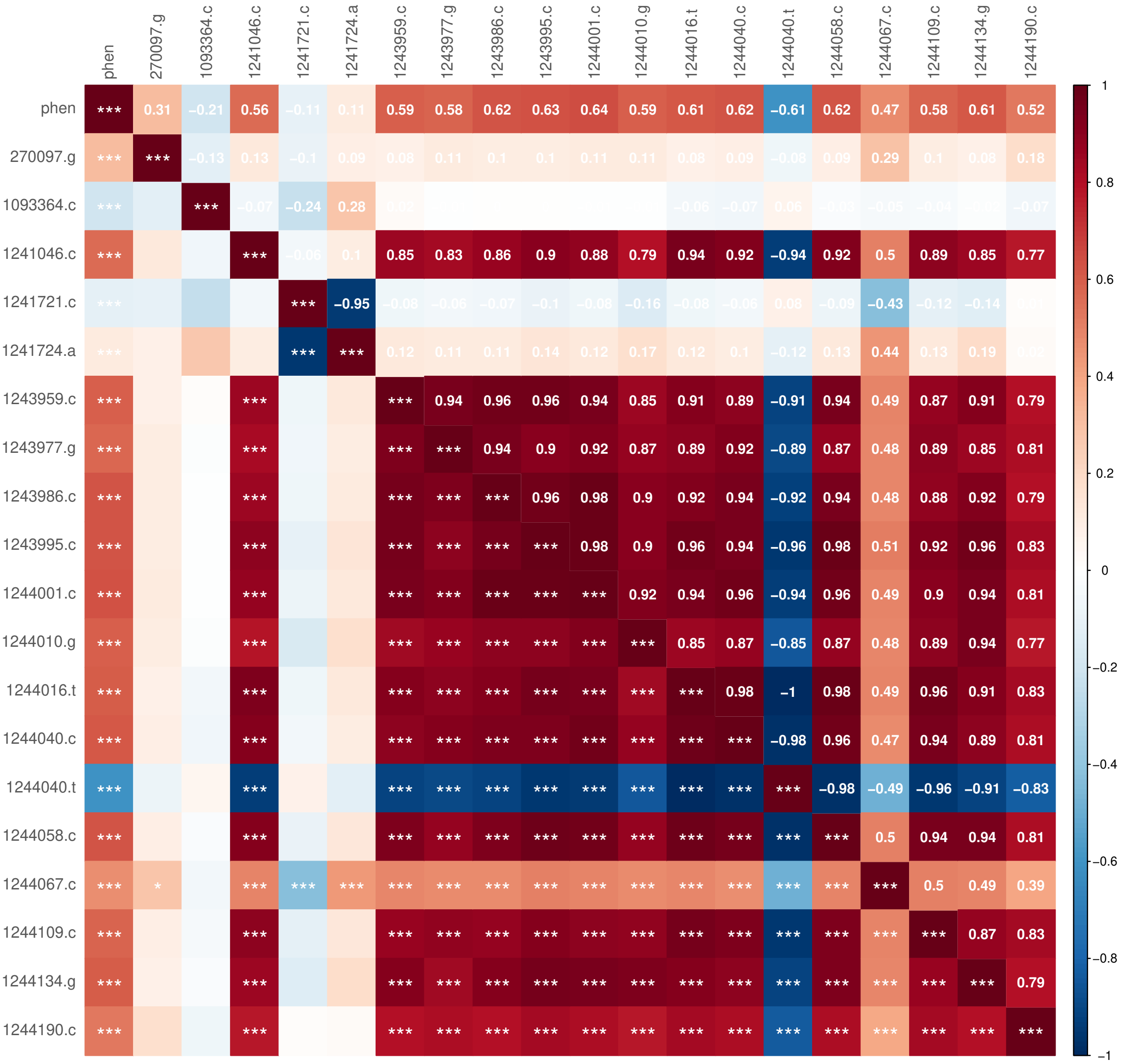


**Figure A.6. Correlation between SNPs associated with penicillin MIC.** A visualisation of the correlation matrix describing the similarity between the 19 unique column patterns observed among the 30 core SNPs identified as significantly associated with the penicillin MIC phenotype in *N. meningitidis*. The Pearson's correlation is given above the diagonal and the significance level of each correlation is indicated by the number of asterisks below the diagonal (* = 0.05, ** = 0.01, *** = 0.001).

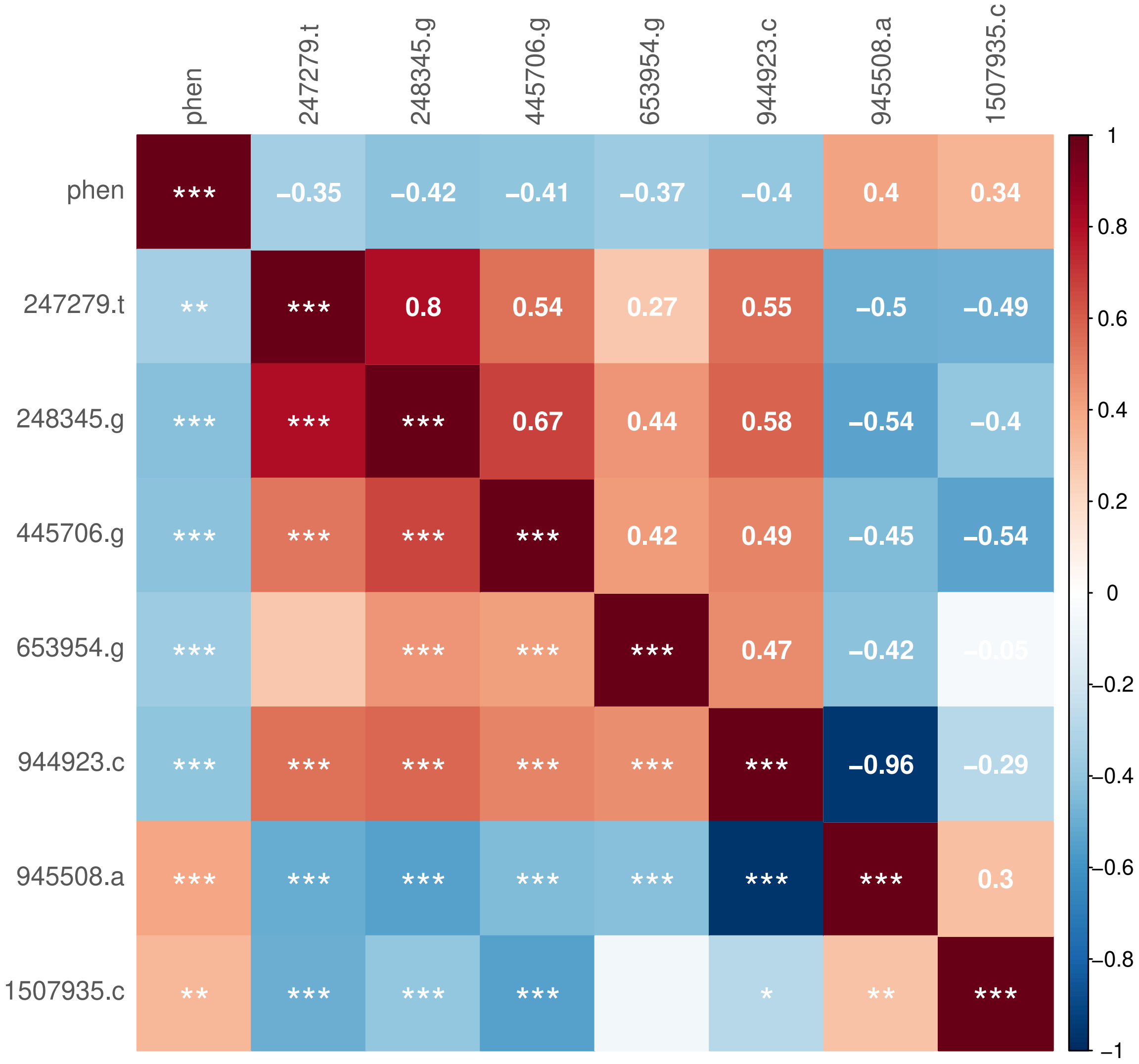
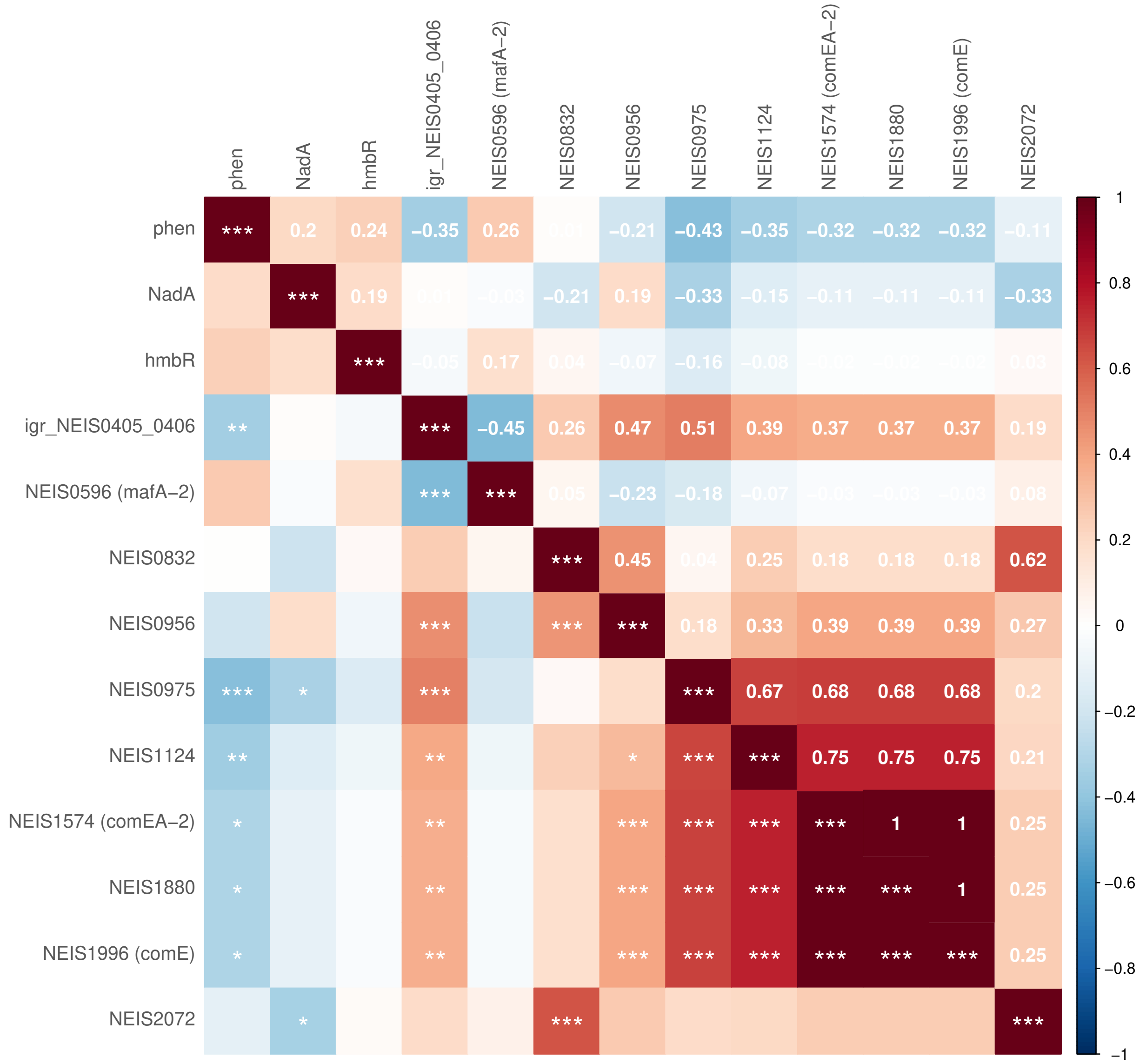


**Figure A.7. Correlation between SNPs associated with invasive disease.** A visualisation of the correlation matrix describing the similarity between the 7 core SNPs identified as significantly associated with the invasive disease phenotype in *N. meningitidis*. The Pearson's correlation is given above the diagonal and the significance level of each correlation is indicated by the number of asterisks below the diagonal (* = 0.05, ** = 0.01, *** = 0.001).



| | phen | NadA | hmbR | igr_NEIS0405_0406 | NEIS0596 (mafA–2) | NEIS0832 | NEIS0956 | NEIS0975 | NEIS1124 | NEIS1574 (comEA–2) | NEIS1880 | NEIS1996 (comE) | NEIS2072 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| phen | *** | 0.2 | 0.24 | –0.35 | 0.26 | 0.01 | –0.21 | –0.43 | –0.35 | –0.32 | –0.32 | –0.32 | –0.11 |
| NadA | | *** | 0.19 | 0.01 | –0.03 | –0.21 | 0.19 | –0.33 | –0.15 | –0.11 | –0.11 | –0.11 | –0.33 |
| hmbR | | | *** | –0.05 | 0.17 | 0.04 | –0.07 | –0.16 | –0.08 | –0.02 | –0.02 | –0.02 | 0.03 |
| igr_NEIS0405_0406 | ** | | | *** | –0.45 | 0.26 | 0.47 | 0.51 | 0.39 | 0.37 | 0.37 | 0.37 | 0.19 |
| NEIS0596 (mafA–2) | | | | *** | *** | 0.05 | –0.23 | –0.18 | –0.07 | –0.03 | –0.03 | –0.03 | 0.08 |
| NEIS0832 | | | | | | *** | 0.45 | 0.04 | 0.25 | 0.18 | 0.18 | 0.18 | 0.62 |
| NEIS0956 | | | | *** | | *** | *** | 0.18 | 0.33 | 0.39 | 0.39 | 0.39 | 0.27 |
| NEIS0975 | *** | * | | *** | | | | *** | 0.67 | 0.68 | 0.68 | 0.68 | 0.2 |
| NEIS1124 | ** | | | ** | | | * | *** | *** | 0.75 | 0.75 | 0.75 | 0.21 |
| NEIS1574 (comEA–2) | * | | | ** | | | *** | *** | *** | *** | 1 | 1 | 0.25 |
| NEIS1880 | * | | | ** | | | *** | *** | *** | *** | *** | 1 | 0.25 |
| NEIS1996 (comE) | * | | | ** | | | *** | *** | *** | *** | *** | *** | 0.25 |
| NEIS2072 | | * | | | | *** | | | | | | | *** |

**Figure A.8. Correlation between accessory genes associated with invasive disease.** A visualisation of the correlation matrix describing the similarity between the 12 accessory genes identified as significantly associated with the invasive disease phenotype in *N. meningitidis*. The Pearson's correlation is given above the diagonal and the significance level of each correlation is indicated by the number of asterisks below the diagonal (* = 0.05, ** = 0.01, *** = 0.001).

| **Gene** | **Gene product** |
|---|---|
| NEIS0041 | DNA transport competence protein |
| NEIS0045 (*rfbC*) | O-antigen capsular region D |
| NEIS0047 (*rfbB*) | O-antigen capsular region D |
| NEIS0065 (*rfbC2*) | O-antigen capsular region D' |
| NEIS1357 | hypothetical protein |
| NEIS0213 (*pglA*) | Pilin glycosyltransferase |

**Table A.9. Genes interacting with invasive disease genes in *N. meningitidis*.** These 6 accessory genes were identified as having significant interactions with one or more of the 12 significant virulence-associated genes identified in Table 5.7. Using a modified version of treeWAS, Score 2 was measured between each of the 12 putative virulence genes and all other accessory genes ($N_{loci} = 2,809$) in a dataset containing gene presence-or-absence sequences ($N_{ind} = 129$) from *N. meningitidis* serogroup C.

# Bibliography

1. Collins C, Didelot X. Reconstructing the Ancestral Relationships Between Bacterial Pathogen Genomes. Methods Mol Biol. 2017;1535:109–137.

2. Zoetendal EG, Vaughan EE, de Vos WM. A microbial world within us. Mol Microbiol. 2006 Mar;59(6):1639–1650.

3. McFall-Ngai M. Adaptive immunity: care for the community. Nature. 2007 Jan;445(7124):153.

4. Young BC, Golubchik T, Batty EM, et al. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc Natl Acad Sci U S A. 2012 Mar;109(12):4550–4555.

5. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. Science. 2005 Jun;308(5728):1635–1638.

6. Vos T, Allen C, Arora M, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016 Oct;388(10053):1545–1602.

7. Bartel RA, Oberhauser KS, De Roode JC, Altizer SM. Monarch butterfly migration and parasite transmission in eastern North America. Ecology. 2011 Feb;92(2):342–351.

8. Sherr BF, Sherr EB, Berman T. Decomposition of organic detritus: A selective role for microflagellate Protozoa1. Limnol Oceanogr. 1982 Jul;27(4):765–769.

9. Bonfante P. Plants, Mycorrhizal Fungi and Endobacteria: a Dialog Among Cells and Genomes. Biol Bull. 2016 Sep;.

10. Paoletti M, Buck KW, Brasier CM. Selective acquisition of novel mating type and vegetative incompatibility genes via interspecies gene transfer in the globally invading eukaryote Ophiostoma novo-ulmi. Mol Ecol. 2006 Jan;15(1):249–262.

11. Kauffman CA. Histoplasmosis: a clinical and laboratory update. Clin Microbiol Rev. 2007 Jan;20(1):115–132.

12. Batinovic S, Wassef F, Knowler SA, et al. Bacteriophages in Natural and Artificial Environments. Pathogens. 2019 Jul;8(3).

13. De Chiara M, Hood D, Muzzi A, et al. Genome sequencing of disease and carriage isolates of nontypeable Haemophilus influenzae identifies discrete population structure. Proc Natl Acad Sci U S A. 2014 Apr;111(14):5439–5444.

14. Xu P, Chen F, Mannas JP, et al. Virus infection improves drought tolerance. New Phytol. 2008 Sep;180(4):911–921.

15. Roossinck MJ. The good viruses: viral mutualistic symbioses. Nat Rev Microbiol. 2011 Jan;9(2):99–108.

16. Suttle CA. Viruses in the sea. Nature. 2005 Sep;437(7057):356–361.

17. Górski A, Miedzybrodzki R, Borysowski J, et al. Bacteriophage therapy for the treatment of infections. Curr Opin Investig Drugs. 2009 Aug;10(8):766–774.

18. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998 Jun;95(12):6578–6583.

19. Marteinsson VT, Birrien JL, Reysenbach AL, et al. Thermococcus barophilus sp. nov., a new barophilic and hyperthermophilic archaeon isolated under high hydrostatic pressure from a deep-sea hydrothermal vent. Int J Syst Bacteriol. 1999 Apr;49 Pt 2:351–359.

20. Saxena R, Sharma VK. Chapter 9 - A Metagenomic Insight Into the Human Microbiome: Its Implications in Health and Disease. In: Kumar D, Antonarakis S, editors. Medical and Health Genomics. Oxford: Academic Press; 2016. p. 107–119.

21. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016 Aug;14(8):e1002533.

22. Gilbert JA, Neufeld JD. Life in a World without Microbes. PLoS Biol. 2014 Dec;12(12):e1002020.

23. Motta V, Trevisi P, Bertolini F, et al. Exploring gastric bacterial community in young pigs. PLoS One. 2017 Mar;12(3):e0173029.

24. Akinosho H, Yee K, Close D, Ragauskas A. The emergence of Clostridium thermocellum as a high utility candidate for consolidated bioprocessing applications. Front Chem. 2014 Aug;2:66.

25. Fox-Skelly J. There are diseases hidden in ice, and they are waking up; 2017. Accessed: 2017-8-13. `http://www.bbc.com/earth/story/20170504-there-are-diseases-hidden-in-ice-and-they-are-waking-up`.

26. Bentley R, Meganathan R. Biosynthesis of vitamin K (menaquinone) in bacteria. Microbiol Rev. 1982 Sep;46(3):241–280.

27. Pompei A, Cordisco L, Amaretti A, et al. Folate production by bifidobacteria as a potential probiotic property. Appl Environ Microbiol. 2007 Jan;73(1):179–185.

28. Bayer M, Aslan G, Emekdaş G, Kuyucu N, Kanik A. [Nasopharyngeal carriage of Streptococcus pneumoniae in healthy children and multidrug resistance]. Mikrobiyol Bul. 2008 Apr;42(2):223–230.

29. Peacock SJ, de Silva I, Lowy FD. What determines nasal carriage of Staphylococcus aureus? Trends Microbiol. 2001 Dec;9(12):605–610.

30. Turnbull P. Anthrax in humans and animals. World Health Organization. 2008;4:43.

31. Ferkol T, Schraufnagel D. The global burden of respiratory disease. Ann Am Thorac Soc. 2014 Mar;11(3):404–406.

32. Raviglione M, Sulis G. Tuberculosis 2015: Burden, Challenges and Strategy for Control and Elimination. Infect Dis Rep. 2016 Jun;8(2):6570.

33. GBD Diarrhoeal Diseases Collaborators. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. Lancet Infect Dis. 2017 Jun;.

34. WHO. World Health Statistics. Global Health Indicators: Cause-specific mortality and morbidity. World Health Organisation. 2015;p. 72.

35. Liu L, Johnson HL, Cousens S, et al. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. Lancet. 2012 Jun;379(9832):2151–2161.

36. Bhutta ZA, Sommerfeld J, Lassi ZS, Salam RA, Das JK. Global burden, distribution, and interventions for infectious diseases of poverty. Infect Dis Poverty. 2014 Jul;3:21.

37. Friedman ND, Temkin E, Carmeli Y. The negative impact of antibiotic resistance. Clin Microbiol Infect. 2016 May;22(5):416–422.

38. Holden MTG, Hsu LY, Kurt K, et al. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. 2013 Apr;23(4):653–664.

39. O'Neill J. Antimicrobial Resistance: Tackling a crisis for the health and wealth of nations. Review on AMR. 2014;.

40. Monina Klevens R, Morrison MA, Nadle J, et al. Invasive Methicillin-Resistant Staphylococcus aureus Infections in the United States. JAMA. 2007 Oct;298(15):1763–1771.

41. O'Neill J. Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. The Review on Antimicrobial Resistance. 2016;.

42. Biello D. Man-Made Genetic Instructions Yield Living Cells for the First Time. Scientific American. 2010 May;.

43. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med. 2011 Jan;364(1):33–42.

44. Didelot X, Dordel J, Whittles LK, et al. Genomic Analysis and Comparison of Two Gonorrhea Outbreaks. MBio. 2016 Jun;7(3).

45. Antimicrob Chemother J, 143. Dominance of EMRSA-15 and -16 among MRSA causing nosocomial bacteraemia in the UK: analysis of isolates from the European Antimicrobial Resistance Surveillance System (EARSS). J Antimicrob Chemother. 2001;48:141–156.

46. Lowder BV, Guinane CM, Ben Zakour NL, et al. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. Proc Natl Acad Sci U S A. 2009 Nov;106(46):19545–19550.

47. Joseph B, Schwarz RF, Linke B, et al. Virulence evolution of the human pathogen Neisseria meningitidis by recombination in the core and accessory genome. PLoS One. 2011 Apr;6(4):e18441.

48. Kiechle FL, Zhang X, Holland-Staley CA. The -omics era and its impact. Arch Pathol Lab Med. 2004 Dec;128(12):1337–1345.

49. Chewapreecha C, Marttinen P, Croucher NJ, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014 Aug;10(8):e1004547.

50. Sommer MOA, Munck C, Toft-Kehler RV, Andersson DI. Prediction of antibiotic resistance: time for a new preclinical paradigm? Nat Rev Microbiol. 2017 Jul;.

51. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012 Sep;13(9):601–612.

52. Guinane CM, Ben Zakour NL, Tormo-Mas MA, et al. Evolutionary genomics of Staphylococcus aureus reveals insights into the origin and molecular basis of ruminant host adaptation. Genome Biol Evol. 2010 Jul;2:454–466.

53. Mather AE, Reid SWJ, Maskell DJ, et al. Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. Science. 2013 Sep;341(6153):1514–1517.

54. Hawkey PM. Multidrug-resistant Gram-negative bacteria: a product of globalization. J Hosp Infect. 2015 Apr;89(4):241–247.

55. Al-Tawfiq JA, Memish ZA. Potential risk for drug resistance globalization at the Hajj. Clin Microbiol Infect. 2015 Feb;21(2):109–114.

56. Bennett GM, Moran NA. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. Genome Biol Evol. 2013;5(9):1675–1688.

57. Han K, Li ZF, Peng R, et al. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. Sci Rep. 2013;3:2101.

58. Ochman H, Jones IB. Evolutionary dynamics of full genome content in Escherichia coli. EMBO J. 2000 Dec;19(24):6637–6643.

59. Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. Annu Rev Genet. 2004;38:771–792.

60. Sezonov G, Joseleau-Petit D, D'Ari R. Escherichia coli physiology in Luria-Bertani broth. J Bacteriol. 2007 Dec;189(23):8746–8749.

61. Tobiason DM, Seifert HS. Genomic Content of Neisseria Species. J Bacteriol. 2010 Apr;192(8):2160–2168.

62. Renzoni A, Andrey DO, Jousselin A, et al. Whole genome sequencing and complete genetic analysis reveals novel pathways to glycopeptide resistance in Staphylococcus aureus. PLoS One. 2011 Jun;6(6):e21577.

63. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000 May;405(6784):299–304.

64. Hyman P, Abedon ST. Bacteriophage host range and bacterial resistance. Adv Appl Microbiol. 2010 Mar;70:217–248.

65. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005 Dec;15(6):589–594.

66. Sundin GW. Genomic insights into the contribution of phytopathogenic bacterial plasmids to the evolutionary history of their hosts. Annu Rev Phytopathol. 2007;45:129–151.

67. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic elements, and why? Heredity. 2011 Jan;106(1):1–10.

68. Jackson RW, Vinatzer B, Arnold DL, Dorus S, Murillo J. The influence of the accessory genome on bacterial pathogen evolution. Mob Genet Elements. 2011 May;1(1):55–65.

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

70. Prudhomme M, Attaiech L, Sanchez G, Martin B, Claverys JP. Antibiotic stress induces genetic transformability in the human pathogen Streptococcus pneumoniae. Science. 2006 Jul;313(5783):89–92.

71. Charpentier X, Kay E, Schneider D, Shuman HA. Antibiotics and UV radiation induce competence for natural transformation in Legionella pneumophila. J Bacteriol. 2011 Mar;193(5):1114–1121.

72. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends Microbiol. 2010 Jul;18(7):315–322.

73. Feil EJ, Spratt BG. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol. 2001;55:561–590.

74. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2008 Oct;3(2):199–208.

75. Reeves P. Evolution of Salmonella O antigen variation by interspecific gene transfer on a large scale. Trends Genet. 1993 Jan;9(1):17–22.

76. Croucher NJ, Harris SR, Fraser C, et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011 Jan;331(6016):430–434.

77. Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res. 2012 Apr;22(4):721–734.

78. van Tonder AJ, Mistry S, Bray JE, et al. Defining the estimated core genome of bacterial populations using a Bayesian decision model. PLoS Comput Biol. 2014 Aug;10(8):e1003788.

79. de Been M, van Schaik W, Cheng L, Corander J, Willems RJ. Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium. Genome Biol Evol. 2013;5(8):1524–1535.

80. Howell KJ, Weinert LA, Chaudhuri RR, et al. The use of genome wide association methods to investigate pathogenicity, population structure and serovar in Haemophilus parasuis. BMC Genomics. 2014 Dec;15:1179.

81. Croucher NJ, Finkelstein JA, Pelton SI, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013 Jun;45(6):656–663.

82. Bessen DE, Kumar N, Hall GS, et al. Whole-genome association study on tissue tropism phenotypes in group A Streptococcus. J Bacteriol. 2011 Dec;193(23):6651–6663.

83. Hacker J, Carniel E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. EMBO Rep. 2001 May;2(5):376–381.

84. Wassenaar TM, Gaastra W. Bacterial virulence: can we draw the line? FEMS Microbiol Lett. 2001 Jul;201(1):1–7.

85. Li R, Wang G, Shen B, et al. Random transposon vectors pUTTns for the markerless integration of exogenous genes into gram-negative eubacteria chromosomes. J Microbiol Methods. 2009 Nov;79(2):220–226.

86. Martínez-García E, de Lorenzo V. Transposon-Based and Plasmid-Based Genetic Tools for Editing Genomes of Gram-Negative Bacteria. In: Weber W, Fussenegger

M, editors. Synthetic Gene Networks: Methods and Protocols. Totowa, NJ: Humana Press; 2012. p. 267–283.

87. Kokes M, Dunn JD, Granek JA, et al. Integrating chemical mutagenesis and whole-genome sequencing as a platform for forward and reverse genetic analysis of Chlamydia. Cell Host Microbe. 2015 May;17(5):716–725.

88. Kaniga K, Delor I, Cornelis GR. A wide-host-range suicide vector for improving reverse genetics in gram-negative bacteria: inactivation of the blaA gene of Yersinia enterocolitica. Gene. 1991 Dec;109(1):137–141.

89. Muhl D, Filloux A. Site-directed mutagenesis and gene deletion using reverse genetics. Methods Mol Biol. 2014;1149:521–539.

90. Rappé MS, Giovannoni SJ. The uncultured microbial majority. Annu Rev Microbiol. 2003;57:369–394.

91. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006 Aug;14(8):353–355.

92. Sun YH, Bakshi S, Chalmers R, Tang CM. Functional genomics of Neisseria meningitidis pathogenesis. Nat Med. 2000 Nov;6(11):1269–1273.

93. Yi K, Stephens DS, Stojiljkovic I. Development and evaluation of an improved mouse model of meningococcal colonization. Infect Immun. 2003 Apr;71(4):1849–1855.

94. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015 Dec;13(12):787–794.

95. Genome Information by Organism; 2017. Accessed: 2017-8-15. https://www.ncbi.nlm.nih.gov/genome/browse/?report=2.

96. Grad YH, Kirkcaldy RD, Trees D, et al. Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime in the USA: a retrospective observational study. Lancet Infect Dis. 2014 Mar;14(3):220–226.

97. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017 Jan;18(1):41–50.

98. Howard SL, Gaunt MW, Hinds J, et al. Application of comparative phylogenomics to study the evolution of Yersinia enterocolitica and to identify genetic differences relating to pathogenicity. J Bacteriol. 2006 May;188(10):3645–3653.

99. Zhou X, Tan FK, Wang N, et al. Genome-wide association study for regions of systemic sclerosis susceptibility in a Choctaw Indian population with high disease prevalence. Arthritis Rheum. 2003 Sep;48(9):2585–2592.

100. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004 May;36(5):512–517.

101. Weiss LA, Veenstra-Vanderweele J, Newman DL, et al. Genome-wide association study identifies ITGB3 as a QTL for whole blood serotonin. Eur J Hum Genet. 2004 Nov;12(11):949–954.

102. Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science. 2005 Apr;308(5720):419–421.

103. Burdett T, Hastings E, Welter D, et al.. GWAS Catalog; 2017. Accessed: 2017-8-16. `https://www.ebi.ac.uk/gwas/home`.

104. Falush D. Toward the use of genomics to study microevolutionary change in bacteria. PLoS Genet. 2009 Oct;5(10):e1000627.

105. Paschou P, Ziv E, Burchard EG, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007 Sep;3(9):1672–1686.

106. Farhat M, Shapiro B, Sheppard S, Colijn C, Murray M. A phylogeny-based sampling strategy and power calculator informs genome-wide associations study design for microbial pathogens. Genome Med. 2014;6(11):101.

107. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–909.

108. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999 Dec;55(4):997–1004.

109. Kraft P, Cox DG. Study designs for genome-wide association studies. Adv Genet. 2008;60:465–504.

110. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6. 1901;2(11):559–572.

111. Bouaziz M, Ambroise C, Guedj M. Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. PLoS One. 2011 Dec;6(12):e28845.

112. McNally A, Cheng L, Harris SR, Corander J. The evolutionary path to extraintestinal pathogenic, drug-resistant Escherichia coli is marked by drastic reduction in detectable recombination within the core genome. Genome Biol Evol. 2013;5(4):699–710.

113. Sheppard SK, Cheng L, Méric G, et al. Cryptic ecology among host generalist Campylobacter jejuni in domestic animals. Mol Ecol. 2014 May;23(10):2442–2451.

114. Farhat MR, Shapiro BJ, Kieser KJ, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013 Oct;45(10):1183–1189.

115. Ledda A, Price JR, Cole K, et al. Re-emergence of methicillin susceptibility in a resistant lineage of Staphylococcus aureus. J Antimicrob Chemother. 2017 May;72(5):1285–1288.

116. Power RA, Davaniah S, Derache A, et al. Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. PLoS One. 2016 Sep;11(9):e0163746.

117. Read T, Massey R. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. 2014;6(11):109.

118. Sheppard SK, Didelot X, Jolley KA, et al. Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol. 2013 Feb;22(4):1051–1064.

119. Bartha I, Carlson JM, Brumme CJ, et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. Elife. 2013 Oct;2:e01123.

120. Laabei M, Recker M, Rudkin JK, et al. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014 May;24(5):839–849.

121. Hall BG. SNP-associations and phenotype predictions from hundreds of microbial genomes without genome alignments. PLoS One. 2014 Feb;9(2):e90490.

122. Alam MT, Petit RA 3rd, Crispell EK, et al. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. Genome Biol Evol. 2014 May;6(5):1174–1185.

123. Wozniak M, Tiuryn J, Wong L. GWAMAR: genome-wide assessment of mutations associated with drug resistance in bacteria. BMC Genomics. 2014 Dec;15 Suppl 10:S10.

124. Weinert LA, Chaudhuri RR, Wang J, et al. Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis. Nat Commun. 2015 Mar;6:6740.

125. Salipante SJ, Roach DJ, Kitzman JO, et al. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Genome Res. 2015 Jan;25(1):119–128.

126. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015 Mar;25:17–24.

127. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016 Nov;17(1):238.

128. Earle SG, Wu CH, Charlesworth J, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016 Apr;1:16041.

129. Lees JA, Vehkala M, Välimäki N, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016 Sep;7:12797.

130. Desjardins CA, Cohen KA, Munsamy V, et al. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. Nat Genet. 2016 May;48(5):544–551.

131. Nebenzahl-Guimaraes H, van Laarhoven A, Farhat MR, et al. Transmissible Mycobacterium tuberculosis Strains Share Genetic Markers and Immune Phenotypes. Am J Respir Crit Care Med. 2017 Jun;195(11):1519–1527.

132. Lees JA, Croucher NJ, Goldblatt D, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. Elife. 2017 Jul;6.

133. Maury MM, Tsai YH, Charlier C, et al. Uncovering Listeria monocytogenes hypervirulence by harnessing its biodiversity. Nat Genet. 2016 Mar;48(3):308–313.

134. Phelan J, Coll F, McNerney R, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC Med. 2016 Mar;14:31.

135. Coll F, Phelan J, Hill-Cawthorne GA, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018 Jan;.

136. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–575.

137. R Core Development Team. The R Project for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2013. Accessed: 2015-2-1. http://www.r-project.org/.

138. Lippert C, Listgarten J, Liu Y, et al. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011 Sep;8(10):833–835.

139. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012 Jun;44(7):821–824.

140. Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 2010 Feb;86(2):172–184.

141. Kang HM, Zaitlen NA, Wade CM, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008 Mar;178(3):1709–1723.

142. Pagel M. BayesTraits; 2017. Accessed: 2018-1-10. http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/BayesTraitsV3.0.1.html.

143. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2007 Mar;175(3):1251–1266.

144. Mantel N. Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. J Am Stat Assoc. 1963;58(303):690–700.

145. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010 Oct;11:94.

146. Thornton T, McPeek MS. ROADTRIPS 2.0 Software Documentation (Beta Version). J Hum Genet. 2012;86:172–184.

147. Sheppard SK, Didelot X, Meric G, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci U S A. 2013 Jul;110(29):11923–11927.

148. Maddison WP, FitzJohn RG. The unsolved challenge to phylogenetic correlation tests for categorical characters. Syst Biol. 2015 Jan;64(1):127–136.

149. Li M, Reilly MP, Rader DJ, Wang LS. Correcting population stratification in genetic association studies using a phylogenetic approach. Bioinformatics. 2010 Mar;26(6):798–806.

150. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):459–463.

151. Recker M, Laabei M, Toleman MS, et al. Clonal differences in Staphylococcus aureus bacteraemia-associated mortality. Nat Microbiol. 2017 Oct;2(10):1381–1388.

152. Didelot X, Meric G, Falush D, Darling A. Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli. BMC Genomics. 2012;13(1):256.

153. Harris SR, Clarke IN, Seth-Smith HMB, et al. Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet. 2012 Apr;44(4):413–9, S1.

154. Didelot X, Eyre DW, Cule M, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol. 2012 Dec;13(12):R118.

155. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018 Feb;14(2):e1005958.

156. Yahara K, Furuta Y, Oshima K, et al. Chromosome painting in silico in a bacterial species reveals fine population structure. Mol Biol Evol. 2013 Jun;30(6):1454–1464.

157. Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. Convergence of Campylobacter species: implications for bacterial evolution. Science. 2008 Apr;320(5873):237–239.

158. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009 Oct;5(10):e1000686.

159. Frandsen PB, Calcott B, Mayer C, Lanfear R. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol Biol. 2015 Feb;15:13.

160. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000 Jul;67(1):170–181.

161. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003 Aug;164(4):1567–1587.

162. Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. Springerplus. 2014 Aug;3:431.

163. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 2008 Dec;9:539.

164. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep;19(9):1655–1664.

165. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012 Jan;8(1):e1002453.

166. Marttinen P, Hanage WP, Croucher NJ, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012 Jan;40(1):e6.

167. Corander J, Waldmann P, Marttinen P, Sillanpää MJ. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics. 2004 Oct;20(15):2363–2369.

168. Bille E, Zahar JR, Perrin A, et al. A chromosomally integrated bacteriophage in invasive meningococci. J Exp Med. 2005 Jun;201(12):1905–1913.

169. Chewapreecha C, Harris SR, Croucher NJ, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014 Feb;46(3):305–309.

170. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Series B Stat Methodol. 2001 May;63(2):411–423.

171. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. J R Stat Soc Ser C Appl Stat. 1979;28(1):100–108.

172. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Cao J, Mao K, Cambria E, Man Z, Toh KA, editors. Proceedings of ELM-2014 Volume 1: Algorithms and Theories. Proceedings in Adaptation, Learning and Optimization. Springer International Publishing; 1967. p. 281–297.

173. Cavalli-Sforza LL. Population structure and human evolution. Proc R Soc Lond B Biol Sci. 1966;164:362–379.

174. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006 Dec;2(12):e190.

175. Jombart T, Pontier D, Dufour AB. Genetic markers in the playground of multivariate analysis. Heredity. 2009 Apr;102(4):330–341.

176. Sanchez-Mazas A, Langaney A. Common genetic pools between human populations. Hum Genet. 1988 Feb;78(2):161–166.

177. Smouse PE, Spielman RS, Park MH. Multiple-Locus Allocation of Individuals to Groups as a Function of the Genetic Variation Within and Differences Among Human Populations. Am Nat. 1982 Apr;119(4):445–463.

178. Baik J, Ben Arous G, Péché S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. Ann Probab. 2005 Sep;33(5):1643–1697.

179. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun;447(7145):661–678.

180. Clayton DG, Walker NM, Smyth DJ, et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet. 2005 Nov;37(11):1243–1246.

181. Jolliffe IT. Principal Component Analysis. 2nd ed. Statistics. Springer; 2002.

182. Kang HM, Sul JH, Service SK, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010 Mar;42:348.

183. Lees J. The background of bacterial GWAS. 2017;.

184. Peres-Neto PR, Jackson DA, Somers KM. How many principal components? stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal. 2005 Jun;49(4):974–997.

185. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015 Feb;43(3):e15.

186. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 Jul;4(4):406–425.

187. Legendre P, Legendre LFJ. Numerical Ecology. vol. 24. 2nd ed. Elsevier; 1983.

188. Sørensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter. 1948;5:1–34.

189. Wang LS, Warnow T, Moret BME, Jansen RK, Raubeson LA. Distance-based genome rearrangement phylogeny. J Mol Evol. 2006 Oct;63(4):473–483.

190. Gascuel O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 1997 Jul;14(7):685–695.

191. Gascuel O, Steel M. Neighbor-joining revealed. Mol Biol Evol. 2006 Nov;23(11):1997–2000.

192. Zuckerland E, Pauling LB. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in Biochemistry. Academic Press; 1962. p. 189–225.

193. Swofford DL, Maddison WP. Reconstructing ancestral character states under Wagner parsimony. Math Biosci. 1987;87(2):199–229.

194. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989;5:164–166.

195. Merker M, Blin C, Mona S, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet. 2015 Mar;47(3):242–249.

196. Morelli G, Song Y, Mazzoni CJ, et al. Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet. 2010 Dec;42(12):1140–1143.

197. Cui Y, Yu C, Yan Y, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc Natl Acad Sci U S A. 2013 Jan;110(2):577–582.

198. Zhou Z, McCann A, Litrup E, et al. Neutral genomic microevolution of a recently emerged pathogen, Salmonella enterica serovar Agona. PLoS Genet. 2013 Apr;9(4):e1003471.

199. Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet. 2003 Apr;4(4):275–284.

200. Guindon S, Dufayard JF, Lefort V, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010 May;59(3):307–321.

201. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006 Nov;22(21):2688–2690.

202. Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph. D. dissertation, The University of Texas at Austin.; 2006.

203. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009 Jul;26(7):1641–1650.

204. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010 Mar;5(3):e9490.

205. Mutreja A, Kim DW, Thomson NR, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011 Sep;477(7365):462–465.

206. Harris SR, Feil EJ, Holden MTG, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010 Jan;327(5964):469–474.

207. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. J Chem Phys. 1953 Jun;21(6):1087–1092.

208. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970 Apr;57(1):97–109.

209. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012 May;61(3):539–542.

210. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 2007 Nov;7:214.

211. Bouckaert R, Heled J, Kühnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):e1003537.

212. Felsenstein J. Inferring Phylogenies. Sinauer Assoc.; 2004.

213. Schierup MH, Hein J. Recombination and the molecular clock. Mol Biol Evol. 2000 Oct;17(10):1578–1579.

214. Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. Genetics. 2000 Oct;156(2):879–891.

215. Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol. 2002 Mar;54(3):396–402.

216. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. MBio. 2014 Nov;5(6):e02158.

217. Rannala B, Yang Z. Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet. 2008;9:217–231.

218. Milkman R, Bridges MM. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. Genetics. 1990 Nov;126(3):505–517.

219. Dress AWM, Flamm C, Fritzsch G, et al. Noisy: identification of problematic columns in multiple sequence alignments. Algorithms Mol Biol. 2008 Jun;3:7.

220. Hornstra HM, Priestley RA, Georgia SM, et al. Rapid typing of Coxiella burnetii. PLoS One. 2011 Nov;6(11):e26201.

221. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015 Feb;11(2):e1004041.

222. Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD. Interplay of recombination and selection in the genomes of Chlamydia trachomatis. Biol Direct. 2011 May;6:28.

223. Joseph SJ, Didelot X, Rothschild J, et al. Population genomics of Chlamydia trachomatis: insights on drift, selection, recombination, and population structure. Mol Biol Evol. 2012 Dec;29(12):3933–3946.

224. Walker TM, Kohl TA, Omar SV, et al. Whole-genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect Dis. 2015 Oct;15(10):1193–1202.

225. Dearlove BL, Cody AJ, Pascoe B, et al. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. ISME J. 2015 Aug;.

226. van Tonder AJ, Bray JE, Roalfe L, et al. Genomics Reveals the Worldwide Distribution of Multidrug-Resistant Serotype 6E Pneumococci. J Clin Microbiol. 2015 Jul;53(7):2271–2285.

227. Ashkenazy H, Penn O, Doron-Faigenboim A, et al. FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W580–4.

228. Croucher NJ, Finkelstein JA, Pelton SI, et al. Population genomic datasets describing the post-vaccine evolutionary epidemiology of Streptococcus pneumoniae. Sci Data. 2015 Oct;2:150058.

229. Cornick JE, Chaguza C, Harris SR, et al. Region-specific diversification of the highly virulent serotype 1 Streptococcus pneumoniae. Microbial Genomics. 2015;1(2).

230. Kamng'ona AW, Hinds J, Bar-Zeev N, et al. High multiple carriage and emergence of Streptococcus pneumoniae vaccine serotype variants in Malawian children. BMC Infect Dis. 2015 Jun;15:234.

231. Turner CE, Abbott J, Lamagni T, et al. Emergence of a New Highly Successful Acapsular Group A Streptococcus Clade of Genotype emm89 in the United Kingdom. MBio. 2015 Jul;6(4):e00622.

232. Stasiewicz MJ, Oliver HF, Wiedmann M, den Bakker HC. Whole-Genome Sequencing Allows for Improved Identification of Persistent Listeria monocytogenes in Food-Associated Environments. Appl Environ Microbiol. 2015 Sep;81(17):6024–6037.

233. Didelot X, Lawson D, Darling A, Falush D. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics. 2010 Dec;186(4):1435–1449.

234. Read AF, Nee S. Inference from binary comparative data. J Theor Biol. 1995 Mar;173(1):99–108.

235. Weimann A, Mooren K, Frank J, et al. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. mSystems. 2016 Nov;1(6).

236. Pagel M. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. Proceedings of the Royal Society of London B: Biological Sciences. 1994 Jan;255(1342):37–45.

237. Handelman SK, Aaronson JM, Seweryn M, et al. Cladograms with Path to Event (ClaPTE): a novel algorithm to detect associations between genotypes or phenotypes using phylogenies. Comput Biol Med. 2015 Mar;58:1–13.

238. Martins EP, Garland T Jr. Phylogenetic Analyses of the Correlated Evolution of Continuous Characters: a Simulation Study. Evolution. 1991 May;45(3):534–557.

239. Garland T Jr, Bennett AF, Rezende EL. Phylogenetic approaches in comparative physiology. J Exp Biol. 2005 Aug;208(Pt 16):3015–3035.

240. Pei YF, Zhang L, Li J, Deng HW. Analyses and comparison of imputation-based association methods. PLoS One. 2010 May;5(5):e10827.

241. Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. Nat Rev Genet. 2004 Feb;5(2):89–100.

242. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol. 2001 Jan;20(1):4–16.

243. Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999 Jun;284(5423):2124–2129.

244. Bloomquist EW, Suchard MA. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst Biol. 2010 Jan;59(1):27–41.

245. Tibayrenc M, Ayala FJ. Chapter Six - Is Predominant Clonal Evolution a Common Evolutionary Adaptation to Parasitism in Pathogenic Parasitic Protozoa, Fungi, Bacteria, and Viruses? In: Rollinson D, Stothard JR, editors. Advances in Parasitology. vol. 97. Academic Press; 2017. p. 243–325.

246. Shapiro BJ, Friedman J, Cordero OX, et al. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012 Apr;336(6077):48–51.

247. Spratt BG, Bowler LD, Zhang QY, Zhou J, Smith JM. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species. J Mol Evol. 1992 Feb;34(2):115–125.

248. Wirth T, Morelli G, Kusecek B, et al. The rise and spread of a new pathogen: seroresistant Moraxella catarrhalis. Genome Res. 2007 Nov;17(11):1647–1656.

249. Smith JM, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? Proc Natl Acad Sci U S A. 1993 May;90(10):4384–4388.

250. Dykhuizen DE, Green L. Recombination in Escherichia coli and the definition of biological species. J Bacteriol. 1991 Nov;173(22):7257–7268.

251. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics. 1985 Sep;111(1):147–164.

252. Lewontin RC. THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS. Genetics. 1964 Jan;49(1):49–67.

253. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theor Appl Genet. 1968 Jun;38(6):226–231.

254. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D. Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol. 2014 Jun;31(6):1593–1605.

255. Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Syst Biol. 1971 Dec;20(4):406–416.

256. Joy JB, Liang RH, Mccloskey RM. Ancestral Reconstruction. PLoS Computational Biology. 2016;.

257. Sankoff D. Minimal Mutation Trees of Sequences. SIAM J Appl Math. 1975 Jan;28(1):35–42.

258. Schluter D, Price T, Mooers AØ, Ludwig D. Likelihood of Ancestor States in Adaptive Radiation. Evolution. 1997;51(6):1699–1711.

259. Rogers JS. Deriving Phylogenetic Trees from Allele Frequencies. Syst Biol. 1984 Mar;33(1):52–63.

260. Platt A, Vilhjálmsson BJ, Nordborg M. Conditions under which genome-wide association studies will be positively misleading. Genetics. 2010 Nov;186(3):1045–1052.

261. Anderson TJC, Williams JT, Nair S, et al. Inferred relatedness and heritability in malaria parasites. Proc Biol Sci. 2010 Aug;277(1693):2531–2540.

262. Blangero J, Williams JT, Almasy L. Variance component methods for detecting complex trait loci. Adv Genet. 2001;42:151–181.

263. Walsh M, Lynch B. Genetics and Analysis of Quantitative Traits. 1998;.

264. Fraser C, Lythgoe K, Leventhal GE, et al. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. Science. 2014 Mar;343(6177):1243727.

265. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999 Oct;401(6756):877–884.

266. Housworth EA, Martins EP, Lynch M. The phylogenetic mixed model. Am Nat. 2004 Jan;163(1):84–96.

267. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May;18(5):821–829.

268. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res. 2008 Feb;18(2):324–330.

269. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455–477.

270. McCarroll SA. Extending genome-wide association studies to copy-number variation. Hum Mol Genet. 2008 Oct;17(R2):R135–42.

271. Xu L, Cole JB, Bickhart DM, et al. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. BMC Genomics. 2014 Aug;15:683.

272. Kryazhimskiy S, Plotkin JB. The Population Genetics of dN/dS. PLoS Genet. 2008 Dec;4(12):e1000304.

273. Felsenstein J. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 1973 Sep;25(5):471–492.

274. Remmele CW, Xian Y, Albrecht M, et al. Transcriptional landscape and essential genes of Neisseria gonorrhoeae. Nucleic Acids Res. 2014 Aug;42(16):10579–10595.

275. Wang Y, Qiu C, Cui Q. A Large-Scale Analysis of the Relationship of Synonymous SNPs Changing MicroRNA Regulation with Functionality and Disease. Int J Mol Sci. 2015 Sep;16(10):23545–23555.

276. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics. 2010 Oct;11:523.

277. Nixon KC. The Parsimony Ratchet, a New Method for Rapid Parsimony Analysis. Cladistics. 1999;15:407–414.

278. Jukes TH, Cantor CR, Others. Evolution of protein molecules. Mammalian protein metabolism. 1969;3(21):132.

279. Von Neumann J. Various techniques used in connection with random digits. Appl Math Ser. 1951;12(36-38):3.

280. Kim SY, Lohmueller KE, Albrechtsen A, et al. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics. 2011 Jun;12:231.

281. Kruskal WH. Ordinal Measures of Association. J Am Stat Assoc. 1958;53(284):814–861.

282. Dunn OJ. Multiple Comparisons Among Means. J Am Stat Assoc. 1961;56(293):52–64.

283. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300.

284. Silverman BW. Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability, London: Chapman and Hall, 1986. 1986;.

285. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010 Jun;5(6):e11147.

286. Fu S, Octavia S, Tanaka MM, Sintchenko V, Lan R. Defining the Core Genome of Salmonella enterica Serovar Typhimurium for Genomic Surveillance and Epidemiological Typing. J Clin Microbiol. 2015 Aug;53(8):2530–2538.

287. McNally A, Oren Y, Kelly D, et al. Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. PLoS Genet. 2016 Sep;12(9):e1006280.

288. Thorpe HA, Bayliss SC, Hurst LD, Feil EJ. Comparative Analyses of Selection Operating on Nontranslated Intergenic Regions of Diverse Bacterial Species. Genetics. 2017 May;206(1):363–376.

289. Brown T, Didelot X, Wilson DJ, De Maio N. SimBac: simulation of whole bacterial genomes with homologous recombination. Microb Genom. 2016 Jan;2(1).

290. Comas I, Coscolla M, Luo T, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013 Oct;45(10):1176–1182.

291. Rödel E. Fisher, R. A.: Statistical Methods for Research Workers, 14. Aufl., Oliver & Boyd, Edinburgh, London 1970. XIII, 362 S., 12 Abb., 74 Tab., 40 s. Biom J. 1970;13(6):429–430.

292. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008 Jun;24(11):1403–1405.

293. Lee C, Abdool A, Huang CH. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics. 2009 Jan;10 Suppl 1:S73.

294. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011 Nov;27(21):3070–3071.

295. Collins C. treeWAS: A Phylogenetic Tree-Based Tool for Genome-Wide Association Studies in Microbes; 2016. Accessed: 2016-12-12. `https://github.com/caitiecollins/treeWAS`.

296. Rijsbergen CJV. Information Retrieval. 2nd ed. Newton, MA, USA: Butterworth-Heinemann; 1979.

297. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet. 2008 Oct;17(R2):R143–50.

298. Tibayrenc M, Ayala FJ. Reproductive clonality of pathogens: a perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. Proc Natl Acad Sci U S A. 2012 Nov;109(48):E3305–13.

299. Zapun A, Morlot C, Taha MK. Resistance to $\beta$-Lactams in Neisseria ssp Due to Chromosomally Encoded Penicillin-Binding Proteins. Antibiotics (Basel). 2016 Sep;5(4).

300. Unemo M, Golparian D, Nicholas R, et al. High-level cefixime- and ceftriaxone-resistant Neisseria gonorrhoeae in France: novel penA mosaic allele in a successful international clone causes treatment failure. Antimicrob Agents Chemother. 2012 Mar;56(3):1273–1280.

301. Whittles LK, White PJ, Didelot X. Quantifying The Fitness Benefit And Cost Of Cefixime Resistance In Neisseria gonorrhoeae To Inform Prescription Policy; 2017.

302. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11(1):595.

303. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403–410.

304. Jomantiene R, Davis RE. Clusters of diverse genes existing as multiple, sequence-variable mosaics in a phytoplasma genome. FEMS Microbiol Lett. 2006 Feb;255(1):59–65.

305. Maiden MC. Population genomics: diversity and virulence in the Neisseria. Curr Opin Microbiol. 2008 Oct;11(5):467–471.

306. Harrison OB, Clemence M, Dillard JP, et al. Genomic analyses of Neisseria gonorrhoeae reveal an association of the gonococcal genetic island with antimicrobial resistance. J Infect. 2016 Dec;73(6):578–587.

307. Gardete S, Ludovice AM, Sobral RG, et al. Role of murE in the Expression of beta-lactam antibiotic resistance in Staphylococcus aureus. J Bacteriol. 2004 Mar;186(6):1705–1713.

308. Ohnishi M, Watanabe Y, Ono E, et al. Spread of a chromosomal cefixime-resistant penA gene among different Neisseria gonorrhoeae lineages. Antimicrob Agents Chemother. 2010 Mar;54(3):1060–1067.

309. Schubert B, Maddamsetti R, Nyman J, Farhat MR, others. Genome-wide discovery of epistatic loci affecting antibiotic resistance using evolutionary couplings. bioRxiv. 2018;.

310. Maiden MC. Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. Clin Infect Dis. 1998 Aug;27 Suppl 1:S12–20.

311. Gong Z, Lai W, Liu M, et al. Novel Genes Related to Ceftriaxone Resistance Found among Ceftriaxone-Resistant Neisseria gonorrhoeae Strains Selected In Vitro. Antimicrob Agents Chemother. 2016 Apr;60(4):2043–2051.

312. Tomberg J, Unemo M, Davies C, Nicholas RA. Molecular and structural analysis of mosaic variants of penicillin-binding protein 2 conferring decreased susceptibility to expanded-spectrum cephalosporins in Neisseria gonorrhoeae: role of epistatic mutations. Biochemistry. 2010 Sep;49(37):8062–8070.

313. Takahata S, Senju N, Osaki Y, Yoshida T, Ida T. Amino acid substitutions in mosaic penicillin-binding protein 2 associated with reduced susceptibility to cefixime in clinical isolates of Neisseria gonorrhoeae. Antimicrob Agents Chemother. 2006 Nov;50(11):3638–3645.

314. Arnold BJ, Gutmann MU, Grad YH, et al. Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics. 2018 Mar;208(3):1247–1260.

315. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet. 2015 Oct;24(R1):R111–9.

316. Mobegi FM, Cremers AJH, de Jonge MI, et al. Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data. Sci Rep. 2017 Feb;7:42808.

317. Unemo M, Shafer WM. Antimicrobial resistance in Neisseria gonorrhoeae in the 21st century: past, evolution, and future. Clin Microbiol Rev. 2014 Jul;27(3):587–613.

318. Torok E, Moran E, Cooke F. Oxford Handbook of Infectious Diseases and Microbiology. OUP Oxford; 2009.

319. Eyre DW, De Silva D, Cole K, et al. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. J Antimicrob Chemother. 2017 Mar;.

320. Pérez-Losada M, Browne EB, Madsen A, et al. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol. 2006 Mar;6(2):97–112.

321. Bowler LD, Zhang QY, Riou JY, Spratt BG. Interspecies recombination between the penA genes of Neisseria meningitidis and commensal Neisseria species during the emergence of penicillin resistance in N. meningitidis: natural events and laboratory simulation. J Bacteriol. 1994 Jan;176(2):333–337.

322. Antignac A, Kriz P, Tzanakaki G, Alonso JM, Taha MK. Polymorphism of Neisseria meningitidis penA gene associated with reduced susceptibility to penicillin. J Antimicrob Chemother. 2001 Mar;47(3):285–296.

323. Taha MK, Vázquez JA, Hong E, et al. Target gene sequencing to characterize the penicillin G susceptibility of Neisseria meningitidis. Antimicrob Agents Chemother. 2007 Aug;51(8):2784–2792.

324. Antignac A, Rousselle JC, Namane A, et al. Detailed structural analysis of the peptidoglycan of the human pathogen Neisseria meningitidis. J Biol Chem. 2003 Aug;278(34):31521–31528.

325. Neri A, Mignogna G, Fazio C, et al. Neisseria meningitidis rifampicin resistant strains: analysis of protein differentially expressed. BMC Microbiol. 2010 Sep;10:246.

326. Jin DJ, Cashel M, Friedman DI, et al. Effects of rifampicin resistant rpoB mutations on antitermination and interaction with nusA in Escherichia coli. J Mol Biol. 1988 Nov;204(2):247–261.

327. Renner-Schneck M, Hinderberger I, Gisin J, et al. Crystal Structure of the N-Acetylmuramic Acid $\alpha$-1-Phosphate (MurNAc-$\alpha$1-P) Uridylyltransferase MurU, a Minimal Sugar Nucleotidyltransferase and Potential Drug Target Enzyme in Gram-negative Pathogens. J Biol Chem. 2015 Apr;290(17):10804–10813.

328. Pizza M, Rappuoli R. Neisseria meningitidis: pathogenesis and immunity. Curr Opin Microbiol. 2015 Feb;23:68–72.

329. Taha MK, Deghmane AE, Antignac A, et al. The duality of virulence and transmissibility in Neisseria meningitidis. Trends Microbiol. 2002 Aug;10(8):376–382.

330. Caugant DA. Genetics and evolution of Neisseria meningitidis: importance for the epidemiology of meningococcal disease. Infect Genet Evol. 2008 Sep;8(5):558–565.

331. Bauer FJ, Rudel T, Stein M, Meyer TF. Mutagenesis of the Neisseria gonorrhoeae porin reduces invasion in epithelial cells and enhances phagocyte responsiveness. Mol Microbiol. 1999 Feb;31(3):903–913.

332. Urwin R, Russell JE, Thompson EAL, et al. Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. Infect Immun. 2004 Oct;72(10):5955–5962.

333. Brehony C, Rodrigues CMC, Borrow R, et al. Distribution of Bexsero® Antigen Sequence Types (BASTs) in invasive meningococcal disease isolates: Implications for immunisation. Vaccine. 2016 Sep;34(39):4690–4697.

334. Dietrich G, Kurz S, Hübner C, et al. Transcriptome analysis of Neisseria meningitidis during infection. J Bacteriol. 2003 Jan;185(1):155–164.

335. Merighi M, Septer AN, Carroll-Portillo A, et al. Genome-wide analysis of the PreA/PreB (QseB/QseC) regulon of Salmonella enterica serovar Typhimurium. BMC Microbiol. 2009 Feb;9(1):42.

336. Beyene GT, Kalayou S, Riaz T, Tonjum T. Comparative proteomic analysis of Neisseria meningitidis wildtype and dprA null mutant strains links DNA processing to pilus biogenesis. BMC Microbiol. 2017 Apr;17(1):96.

337. Snyder LAS, Cole JA, Pallen MJ. Comparative analysis of two Neisseria gonorrhoeae genome sequences reveals evidence of mobilization of Correia Repeat Enclosed Elements and their role in regulation. BMC Genomics. 2009 Feb;10:70.

338. Baart GJE, Zomer B, de Haan A, et al. Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes. Genome Biol. 2007 Jul;8(7):R136.

339. Ieva R, Alaimo C, Delany I, et al. CrgA Is an Inducible LysR-Type Regulator of Neisseria meningitidis, Acting both as a Repressor and as an Activator of Gene Transcription. J Bacteriol. 2005 May;187(10):3421–3430.

340. Grandi G. Rational antibacterial vaccine design through genomic technologies. Int J Parasitol. 2003 May;33(5-6):615–620.

341. Hey A, Li MS, Hudson MJ, Langford PR, Kroll JS. Transcriptional profiling of Neisseria meningitidis interacting with human epithelial cells in a long-term in vitro colonization model. Infect Immun. 2013 Nov;81(11):4149–4159.

342. ; Gene Expression During Meningococcus Adhesion. 02448284; 2002.

343. Grifantini R, Bartolini E, Muzzi A, et al. Gene expression profile in Neisseria meningitidis and Neisseria lactamica upon host-cell contact: from basic research to vaccine development. Ann N Y Acad Sci. 2002 Dec;975:202–216.

344. Capecchi B, Adu-Bobie J, Di Marcello F, et al. Neisseria meningitidis NadA is a new invasin which promotes bacterial adhesion to and penetration into human epithelial cells. Mol Microbiol. 2005 Feb;55(3):687–698.

345. Comanducci M, Bambini S, Brunelli B, et al. NadA, a novel vaccine candidate of Neisseria meningitidis. J Exp Med. 2002 Jun;195(11):1445–1454.

346. Bentley SD, Vernikos GS, Snyder LAS, et al. Meningococcal Genetic Variation Mechanisms Viewed through Comparative Analysis of Serogroup C Strain FAM18. PLoS Genet. 2007 Feb;3(2):e23.

347. Fagnocchi L, Pigozzi E, Scarlato V, Delany I. In the NadR regulon, adhesins and diverse meningococcal functions are regulated in response to signals in human saliva. J Bacteriol. 2012 Jan;194(2):460–474.

348. Harrison OB, Evans NJ, Blair JM, et al. Epidemiological evidence for the role of the hemoglobin receptor, hmbR, in meningococcal virulence. J Infect Dis. 2009 Jul;200(1):94–98.

349. Stojiljkovic I, Hwa V, de Saint Martin L, et al. The Neisseria meningitidis haemoglobin receptor: its role in iron utilization and virulence. Mol Microbiol. 1995 Feb;15(3):531–541.

350. Stojiljkovic I, Larson J, Hwa V, Anic S, So M. HmbR outer membrane receptors of pathogenic Neisseria spp.: iron-regulated, hemoglobin-binding proteins with a high level of primary structure conservation. J Bacteriol. 1996 Aug;178(15):4670–4678.

351. Schoen C, Blom J, Claus H, et al. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis. Proc Natl Acad Sci U S A. 2008 Mar;105(9):3473–3478.

352. Croucher NJ, Mostowy R, Wymant C, et al. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. PLoS Biol. 2016 Mar;14(3):e1002394.

353. Marri PR, Paniscus M, Weyand NJ, et al. Genome sequencing reveals widespread virulence gene exchange among human Neisseria species. PLoS One. 2010 Jul;5(7):e11835.

354. Hill DJ, Griffiths NJ, Borodina E, Virji M. Cellular and molecular biology of Neisseria meningitidis colonization and invasive disease. Clin Sci. 2010 Feb;118(9):547–564.

355. Masignani V, Giuliani MM, Tettelin H, et al. Mu-Like Prophage in Serogroup B Neisseria meningitidis Coding for Surface-Exposed Antigens. Infect Immun. 2001 Apr;69(4):2580–2588.

356. Bennett JS, Bentley SD, Vernikos GS, et al. Independent evolution of the core and accessory gene sets in the genus Neisseria: insights gained from the genome of Neisseria lactamica isolate 020-06. BMC Genomics. 2010 Nov;11:652.

357. Snyder LAS, Saunders NJ. The majority of genes in the pathogenic Neisseria species are present in non-pathogenic Neisseria lactamica, including those designated as 'virulence genes'. BMC Genomics. 2006 May;7:128.

358. Dunning Hotopp JC, Grifantini R, Kumar N, et al. Comparative genomics of Neisseria meningitidis: core genome, islands of horizontal transfer and pathogen-specific genes. Microbiology. 2006 Dec;152(Pt 12):3733–3749.

359. Chen I, Dubnau D. DNA uptake during bacterial transformation. Nat Rev Microbiol. 2004 Mar;2(3):241–249.

360. Johnsborg O, Eldholm V, Håvarstein LS. Natural genetic transformation: prevalence, mechanisms and function. Res Microbiol. 2007 Dec;158(10):767–778.

361. Melnyk RA, Hossain SS, Haney CH. Convergent gain and loss of genomic islands drives lifestyle changes in plant-associated bacteria. bioRxiv. 2018;.

362. Farhat MR, Freschi L, Calderon R, et al. Genome wide association with quantitative resistance phenotypes in Mycobacterium tuberculosis reveals novel resistance genes and regulatory regions. bioRxiv. 2018;.

363. Tabor J. Investigating the Investigative Task: Testing for Skewness: An Investigation of Different Test Statistics and Their Power to Detect Skewness. J Stat Educ. 2010;18(2).

364. Aschard H, Vilhjálmsson BJ, Joshi AD, Price AL, Kraft P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am J Hum Genet. 2015 Feb;96(2):329–339.

365. O'Reilly PF, Hoggart CJ, Pomyen Y, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One. 2012 May;7(5):e34861.

366. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. 2017 Apr;13(4):e1006711.

367. Lees JA. Host and pathogen genetics associated with pneumococcal meningitis. 2017;.

368. Naret O, Chaturvedi N, Bartha I, Hammer C, Fellay J. Correcting for population stratification reduces false positive and false negative results in joint analyses of host and pathogen genomes; 2017.

369. Kimura M. The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet. 1991 Aug;66(4):367–386.

370. Hughes AL. The origin of adaptive phenotypes. Proc Natl Acad Sci U S A. 2008 Sep;105(36):13193–13194.

371. Wollenberg KR, Atchley WR. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. Proc Natl Acad Sci U S A. 2000 Mar;97(7):3288–3291.

372. Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Stat. 1979 Jan;7(1):1–26.

373. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Lancaster, Pa. :: Society for the Study of Evolution; 1985.

374. Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJP. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol. 2003 Feb;20(2):248–254.

375. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001 Aug;17(8):754–755.

376. Liberman U, Feldman M. On the evolution of epistasis III: the haploid case with mutation. Theor Popul Biol. 2008 Mar;73(2):307–316.

377. Bloom JD, Adami C. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. BMC Evol Biol. 2004 Jun;4:14.

378. Martin G, Elena SF, Lenormand T. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat Genet. 2007 Apr;39(4):555–560.

379. Codoñer FM, O'Dea S, Fares MA. Reducing the false positive rate in the non-parametric analysis of molecular coevolution. BMC Evol Biol. 2008 Apr;8:106.

380. Bergé M, Moscoso M, Prudhomme M, Martin B, Claverys JP. Uptake of transforming DNA in Gram-positive bacteria: a view from Streptococcus pneumoniae. Mol Microbiol. 2002 Jul;45(2):411–421.

381. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. Genes under positive selection in Escherichia coli. Genome Res. 2007 Sep;17(9):1336–1343.

382. Touchon M, Hoede C, Tenaillon O, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009 Jan;5(1):e1000344.

383. Milkman R, Jaeger E, McBride RD. Molecular evolution of the Escherichia coli chromosome. VI. Two regions of high effective recombination. Genetics. 2003 Feb;163(2):475–483.

384. Ganesh K. Molecular characterization of non-groupable Neisseria meningitidis causing invasive disease in South Africa; 2017.

385. Rouphael NG, Stephens DS. Neisseria meningitidis: biology, microbiology, and epidemiology. Methods Mol Biol. 2012;799:1–20.