RESEARCH ARTICLE

A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination

Caitlin Collins, Xavier Didelot*

Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom

* x.didelot@imperial.ac.uk

Abstract

Genome-Wide Association Studies (GWAS) in microbial organisms have the potential to vastly improve the way we understand, manage, and treat infectious diseases. Yet, microbial GWAS methods established thus far remain insufficiently able to capitalise on the growing wealth of bacterial and viral genetic sequence data. Facing clonal population structure and homologous recombination, existing GWAS methods struggle to achieve both the precision necessary to reject spurious findings and the power required to detect associations in microbes. In this paper, we introduce a novel phylogenetic approach that has been tailor-made for microbial GWAS, which is applicable to organisms ranging from purely clonal to frequently recombining, and to both binary and continuous phenotypes. Our approach is robust to the confounding effects of both population structure and recombination, while maintaining high statistical power to detect associations. Thorough testing via application to simulated data provides strong support for the power and specificity of our approach and demonstrates the advantages offered over alternative cluster-based and dimension-reduction methods. Two applications to *Neisseria meningitidis* illustrate the versatility and potential of our method, confirming previously-identified penicillin resistance loci and resulting in the identification of both well-characterised and novel drivers of invasive disease. Our method is implemented as an open-source R package called treeWAS which is freely available at https://github.com/caitiecollins/treeWAS.

Author summary

Measurable differences often exist within a microbial population, with important ecological or epidemiological consequences. Examples include differences in growth rates, host range, transmissibility, antimicrobial resistance, virulence, etc. Understanding the genetic factors involved in these phenotypic properties is a crucial aim in microbial genomics. A fundamental approach for doing so is to perform a Genome-Wide Association Study (GWAS), where genomes are compared to search for genetic markers systematically correlated with the property of interest. If this strategy were implemented naively in

OPEN ACCESS

Citation: Collins C, Didelot X (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol 14(2): e1005958. https://doi.org/10.1371/journal.pcbi.1005958

Editor: Alice Carolyn McHardy, Helmholtz-Zentrum fur Infektionsforschung GmbH, GERMANY

Received: June 6, 2017

Accepted: December 30, 2017

Published: February 5, 2018

Data Availability Statement: All data are contained within the Supporting Information files.

Funding: This work was support by BBSRC grant BB/L023458/1 and MRC grant MR/N010760/1. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

microbes, it could lead to spurious results due to the confounding effects of population structure and recombination. Here we present treeWAS, a new phylogenetic method to perform microbial GWAS that avoids these pitfalls. We show, using simulated datasets, that treeWAS is able to distinguish between genetic markers that are truly associated with the property of interest and those that are not. Furthermore, we demonstrate that treeWAS offers advantages in both sensitivity and specificity over alternative cluster-based and dimension-reduction techniques. We also showcase treeWAS in two applications to real datasets from *N. meningitidis*. We have developed an easy-to-use implementation of treeWAS in the R environment, which should be useful to a wide range of researchers in microbial genomics.

This is a PLoS Computational Biology Methods paper.

Introduction

Owing to rapid progress in sequencing technologies, the accumulation of microbial genome sequences has begun to outpace the development of statistical and computational tools for their analysis. As a result, opportunities to reduce the global burden of infectious disease are missed. Meanwhile, infectious diseases remain accountable for 15% of worldwide annual mortality [1]. Moreover, as globalisation continues to increase the rate and scope of human interaction, with each other and with animals, this process will likely be accompanied by parallel change in the spread and evolution of infectious pathogens [2–5]. Discovering the genetic basis of microbial traits would offer key insights into the biological mechanisms underlying infectious diseases, and would improve our ability to develop drugs and vaccines, target treatments, build predictive tools, benefit from surveillance, and enhance public health.

Genome-wide association studies (GWAS) can be used to make these inferences, linking genotype to phenotype by testing for statistical associations between the two. GWAS have become a tool of choice in human genetics, since the publication of the first such studies in the early 2000s [6–9], leading to the identification of over 11,000 trait-associated single nucleotide polymorphisms (SNPs) in humans [10]. It has been anticipated that by applying GWAS methods to microbes, similar discoveries could be made [11]. Indeed, although the advent of GWAS in microbes has been relatively recent, promising results can already be seen in the literature to date [12–14]. By contrast to GWAS in humans, however, microbial association mapping remains a technical challenge in search of an optimal methodological approach.

The purpose of GWAS is to identify statistically significant associations that may indicate the presence of a causal relationship between genotype and phenotype while rejecting spurious associations arising from confounding factors. In microbes, smaller genome sizes and the ability to manipulate these genomes in the laboratory may improve the power and computational ease of GWAS and facilitate the confirmation of candidate loci [11]. On the other hand, microbial association studies must overcome a multiplicity of confounding factors, such as, the stronger population structure that results from clonal reproduction [15], widespread linkage disequilibrium interrupted unpredictably by homologous recombination [16], diversity in genetic content [17], and variability in the phenotypic probability distribution for a given genotype [18].

Most microbial GWAS analyses to date have made an effort to control for the confounding potential of population structure. The strength of this confounding effect increases both with the degree to which allele frequencies differ between subpopulations in a sample and the extent to which phenotypic states cluster within these lineages or clades [14, 19]. Cluster-based methods [20] and dimension reduction techniques [21, 22] have been adopted to account for population structure in microbial association studies [13, 22–26], and recent refinements of these methods have been proposed to increase their statistical power [27, 28]. Nevertheless, like methods that rearrange the phenotype to assess significance [24, 26, 29], these approaches cannot appropriately evaluate the probability that population substructure will give rise to spurious associations because they do not factor the degree of phenotypic clustering into the analysis. Pairwise methods account for fine scale genetic differences and phenotypic clustering, but discard large volumes of valuable data [29, 30]. Clonal relatedness evidently remains a challenge for microbial GWAS based on these strategies.

Fortunately, clonality also enables the adoption of a phylogenetic solution [31–33]. 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. Adopting a phylogenetic approach does not require evolution to be treated as purely clonal, nor that recombination must be ignored, since the effect of recombination events can be considered within a phylogenetic framework [15, 34]. Nor do they require any loss of information, provided pairwise techniques are not used. 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.

Here we propose a new phylogenetic approach to GWAS called treeWAS that is able to overcome many of the limitations of existing microbial GWAS approaches. Within our analytical pipeline, data simulation based on parameters of the empirical dataset under analysis allows us to account for the composition of the genetic dataset, the population structure of the sample, and the confounding effects of recombination. We apply multiple complementary scores of association to enhance statistical power and improve detection of associations underlying subtle and complex phenotypes, such as host association or invasiveness. Below, we present the results of rigorous testing on simulated datasets, and compare performance with alternative approaches, including cluster-based and dimension-reduction methods. We also demonstrate how treeWAS responds to varying levels of recombination and contrast this to previous methods. We show that treeWAS provides both specificity and power in a wide range of settings, and consistently offers the best overall performance. Finally, we present two applications to real data from *Neisseria meningitidis*. First, we investigate penicillin resistance and demonstrate that our approach can confirm known resistance loci. Second, we examine invasive disease, which reveals both previously characterised and novel invasiveness factors and illustrates the ability of our methodology to identify associations when applied to complex phenotypes.

Materials and methods

Overview of the treeWAS method

Our central aim is to delineate true signals of association from a noisy background of spurious associations. To accomplish this, our method uses the simulation of a null genetic dataset to establish whether high association score values in the empirical dataset under analysis are likely to be truly significant or may, in fact, arise by chance as a result of confounding factors found in the empirical dataset. Outside of the GWAS literature, similar approaches comparing null and empirical distributions have been used to determine whether inferred ancestry supports the functional linkage of genes [35, 36]. In treeWAS, we characterise the evolutionary parameters of the empirical dataset and use these to generate the simulated genetic dataset, which represents the null hypothesis of no association. This null dataset resembles the empirical dataset in both genetic composition and population structure, but does not have any true association with the phenotype. By comparing associations in the empirical and simulated datasets, we are able to determine which signals of association have sufficient statistical and evolutionary support. This approach makes use of all information contained in the dataset, as well as that inferred in phylogenetic and ancestral state reconstructions. We aim to maintain strict control over the number of false positive findings. This makes possible the application of multiple complementary tests of association, which increases the power to detect associations. The entire treeWAS pipeline is typically completed in a matter of minutes or seconds, depending on dataset size (see S1 Appendix).

Implementation

The treeWAS approach is implemented in the following steps:

1. **Phylogenetic reconstruction** can be performed within treeWAS by distance-based [37–40] or maximum-likelihood (ML) [41] methods. However, where recombination is expected to distort the clonal genealogy, it is recommended that users provide a tree previously reconstructed by a recombination-aware approach [34, 42, 43]. Tools are provided for integration with ClonalFrameML [42].

2. **Computation of the homoplasy distribution**, containing site-specific numbers of substitutions drawn from the empirical dataset, is performed with the Fitch parsimony algorithm [44].

3. **Simulation of null genetic data** enables the delineation of true associations from spurious associations. We compare the relationships between genotype and phenotype at all loci in the real data to those in a simulated dataset that embodies only potentially confounding factors. Simulation of this “null” genetic dataset is guided by three parameters: (i) the phylogenetic tree, (ii) the homoplasy distribution, (iii) the number of loci to be simulated, N_{sim}, which is recommended to be at least ten times the number of biallelic sites in the empirical dataset. 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 original phenotype is maintained across the leaves. By retaining the phylogenetic tree and the distribution of phenotypic states, but reassigning substitutions to new branches, we are able to produce a simulated dataset that resembles the empirical dataset in population structure and genetic composition, including the effects of mutation and recombination. Associations due to confounding factors, but no true associations with the phenotype, are thus recreated by the simulation process.

4. **Ancestral character estimation** is required prior to association testing. A marginal reconstruction of the ancestral states of both genotype and phenotype must be performed via parsimony [45] or ML [46, 47] (see S2 Appendix).

5. **Association testing** is performed by applying three independent tests of association to all loci (see below). Associations between simulated loci and the empirical phenotype are first measured, allowing for the identification of a null distribution of association score statistics under the null hypothesis of no association. Associations between empirical loci and the phenotype are then measured and evaluated with reference to this null distribution.

6. **Identification of significance threshold and associations** proceeds by drawing a threshold in the upper tail of the null distribution, at the value corresponding to a base p-value (e.g., $p = 0.01$) that has been corrected for multiple testing to account for both the number of genetic loci and the three association tests (via Bonferroni correction by default, though False Discovery Rate is also implemented). 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.

Tests of association

The design of treeWAS, particularly its use of the null distribution, enables strict control over the false positive rate. This presents the opportunity for power to be augmented by applying multiple independent tests of association. We therefore measure the association between each genetic locus and the phenotype with three separate tests, described below and illustrated in Fig 1. All three tests are applicable to any form of binary genetic data, including SNPs, indels, and gene presence or absence matrices, and can be used on binary phenotypes, discrete interval variables, and continuous phenotypic data. The following notation is used to describe the three association scores. p_i^{anc} and p_i^{des} denote the phenotypic state at ancestral and descendant node of branch i, respectively. g_i^{anc} and g_i^{des} denote respectively the genotypic state at the ancestral and descendant node of branch i. n and n_b denote respectively the number of leaves and branches on the phylogenetic tree.

Score 1, the "Terminal Score", measures sample-wide association across the leaves of the phylogenetic tree. For a binary phenotype, this score is equivalent to counting the four terminal state combinations, with and without the phenotype, and with and without the genotype, as previously proposed [48]. Generalizing to continuous phenotypes gives:

$$\textbf{Score 1} = \left| \sum_{i=1}^{n} \frac{1}{n} (p_i^{des} g_i^{des} + (1 - p_i^{des})(1 - g_i^{des}) - (1 - p_i^{des}) g_i^{des} - p_i^{des}(1 - g_i^{des})) \right| \qquad (1)$$

Eq 1 determines whether a given allele is over-represented among individuals of a particular phenotypic state. This allows Score 1 to detect associations that are upheld across a relatively large proportion of terminal nodes. Uniform association in all but six of the terminal nodes in Fig 1A, for example, leads to a high Score 1 value. In addition, because Score 1 is blind to ancestral information, it can detect terminal association even in the absence of strong ancestral indicators of association. Furthermore, its inferences remain robust to any incorrect estimates of the phylogenetic tree or ancestral state reconstructions.

Score 2, the "Simultaneous Score", measures the degree of parallel change in the phenotype and genotype across branches of the tree. For a binary phenotype with a parsimonious ancestral state reconstruction, as in Fig 1B, this means counting the number of branches containing a simultaneous substitution in genotype and phenotype [31]. A more general definition is given by:

$$\textbf{Score 2} = \left| \sum_{i=1}^{n_b} (p_i^{anc} - p_i^{des})(g_i^{anc} - g_i^{des}) \right| \qquad (2)$$

Unlike Score 1, Score 2 is able to use information contained in the tree structure and set of ancestral character states towards the detection of significant associations. In Fig 1B, for example, Score 2 reveals the presence of strong signals of association by detecting five instances of simultaneous substitution. Moreover, because Eq 2 imparts a cumulative character (simultaneous

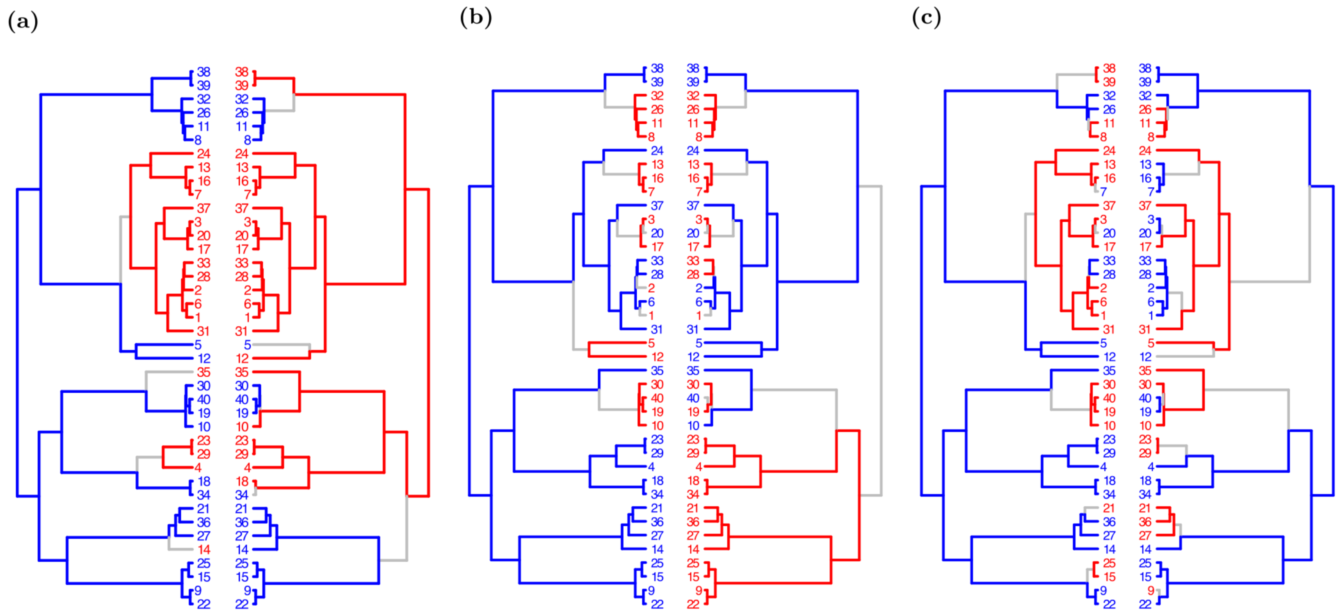

Fig 1. Evolutionary scenarios detected by treeWAS scores. The three complementary tests of association in treeWAS assign high scores to different patterns of association, examples of which are illustrated above. Each panel displays the phenotype (left) and the genotype of one associated locus (right), with binary states plotted along the tips of the phylogenetic tree (N = 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 to this scenario.

https://doi.org/10.1371/journal.pcbi.1005958.g001

substitutions increase the score, but branches where one or no variable changes do not decrease it), significance by Score 2 does not require sample-wide association. Score 2 may therefore be able to detect loci giving rise to the phenotype through complementary pathways, in addition to identifying loci whose associations with the phenotype persist across the tree.

Score 3, the "Subsequent Score", measures the proportion of the tree in which the genotype and phenotype co-exist. It is the mathematical solution to the integral of an association score along all points of the phylogenetic tree (see S3 Appendix):

$$\textbf{Score 3} = \left| \sum_{i=1}^{n_b} \frac{4}{3} p_i^{anc} g_i^{anc} + \frac{2}{3} p_i^{anc} g_i^{des} + \frac{2}{3} p_i^{des} g_i^{anc} + \frac{4}{3} p_i^{des} g_i^{des} - p_i^{anc} - p_i^{des} - g_i^{anc} - g_i^{des} + 1 \right| \quad (3)$$

Score 3 considers the maintenance of allelic enrichment in a given phenotypic state, as well as the change in both genotype and phenotype. Should association arise by a substitution in one variable being followed on a subsequent branch by a substitution in the other, as in Fig 1C, 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 host association, for example, where 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, Score 3 may be most effective [48]. Overall, this score should be sensitive to subtler and more probabilistic patterns of association.

Pooling the scores allows treeWAS to obtain a comprehensive picture of the association present in a dataset. As Fig 1 illustrates, the three scores complement each other and contribute distinct, if overlapping, patterns of association to the output of treeWAS. Score 1 recognises widespread terminal association in Fig 1A and returns a high score, while a Score 2 of zero and a low Score 3 result from their focus on ancestral co-evolution. By contrast, in Fig 1B, where one major clade and 50% of terminal nodes are out of association, we see a Score 1 of zero and a low Score 3, meanwhile the repeated identification of simultaneous substitutions results in a high Score 2. The lack of terminal association and absence of simultaneous substitutions in Fig 1C cause both Scores 1 and 2 to amount to zero, but relaxing both of their requirements and allowing substitutions to occur on subsequent branches results in a high value for Score 3.

Independently, each association test identifies a set of statistically significant genetic loci, and each of these findings constitutes a suitable candidate for further investigation. Although it may provide further support for a finding, identification by a second or third association test is not required for overall significance. Once identified, the three sets of significant genetic loci are pooled together and returned as the set of findings identified by treeWAS. Instead of merging the three well-defined association scores into an uninformative aggregate score, we report the significant association scores and p-values for each test separately. These measures enhance the interpretability of the output of treeWAS, by providing insight into the nature of the associations detected, alongside the list of significant findings.

Assessment of performance on simulated data

To evaluate the performance of treeWAS, we applied it and six alternative methods to 400 datasets simulated via three separate approaches. Approaches differed only in the nature of the simulated associations between genotype and phenotype. We present one approach below and the other two in S4 and S5 Appendices. Additional results based on the approach below are presented in S6 Appendix.

Each simulated dataset contained 100 individuals and 10,000 binary loci, of which ten loci were associated with a binary phenotype. Variation in the background recombination rate, the ancestral relationships between genomes, the degree of phenotypic clustering among related isolates, and the effect size of associations added complexity and noise to the simulated data and increased the challenge of association testing for treeWAS and all comparator methods. The non-associated loci were simulated using homoplasy distributions corresponding to four recombination rates (0, 0.01, 0.05, 0.1) (see S7 Appendix). Genetic data was simulated along randomly generated coalescent trees. For each simulated non-associated locus, the number of substitutions was drawn from the homoplasy distribution and assigned to branches of the phylogenetic tree with probabilities proportional to branch lengths. The ten associated loci were generated together with the phenotype according to an instantaneous transition rate matrix, Q, which controls the rates of transition between all four possible combinations of a binary genetic locus, G, and the binary phenotype, P, (i.e., $G_0\ P_0$, $G_0\ P_1$, $G_1\ P_0$, $G_1\ P_1$) between an ancestral node (in the rows) and a descendant node (in the columns):

$$Q = \begin{array}{c} \\ \mathrm{G_0P_0} \\ \mathrm{G_0P_1} \\ \mathrm{G_1P_0} \\ \mathrm{G_1P_1} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{G_0P_0} & \mathrm{G_0P_1} & \mathrm{G_1P_0} & \mathrm{G_1P_1} \end{array} \\ \begin{pmatrix} -2s & s & s & 0 \\ sa & -2sa & 0 & sa \\ sa & 0 & -2sa & sa \\ 0 & s & s & -2s \end{pmatrix} \end{array} \quad (4)$$

The Q matrix is parameterised by s, which controls the baseline substitution rate and applies to all columns, and a, an association factor that establishes the preference for one form of association ($G_0\ P_0$, $G_1\ P_1$) over the opposite ($G_0\ P_1$, $G_1\ P_0$). The parameter s is divided by the sum of the branch lengths before building Q. In all simulations, initial parameters were set to $s = 20$ and $a = 10$. To identify the probabilities of transition for a branch of length l, the instantaneous transition rate matrix, Q, is converted into a matrix of probabilities, $P = exp(Ql)$, via matrix exponentiation, which takes into account the length l of the branch in question.

Comparison with other GWAS methods

To each of the simulated datasets, in addition to treeWAS, we applied six alternative GWAS methods. The Fisher's exact test, and the χ^2 test available in PLINK version 1.07 [49] were used as benchmarks to demonstrate what results would be found by two standard tests of association without population structure control. The PLINK χ^2 test with Genomic Control (GC), has been used in bacterial GWAS [13] and provided a simple solution to population structure. Principal Components Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) represent more advanced and popular approaches to correcting for population structure [21, 22]. PCA is the "gold standard" method used in human GWAS [19, 50] and DAPC has been proposed as a potential improvement on PCA [22]. Both have been used in microbial GWAS [22, 25–28]. We followed the standard protocol used in human genetics and corrected for ancestry by regressing along the significant Principal Components (PCs) of PCA or DAPC (see S8 Appendix), and identified significant associations via χ^2 test [19]. The Cochran-Mantel-Haenszel (CMH, [51]) provided an alternative, cluster-based approach. The CMH test works directly with K population clusters by adopting a stratified 2x2xK design and has been used in bacterial GWAS [13, 23, 24].

Results/Discussion

Assessment of treeWAS performance on simulated data

The performance on simulated datasets was evaluated along four metrics: the False Positive Rate (FPR), Sensitivity, Positive Predictive Value (PPV), the proportion of results that are true positives, and the F1 Score, which is the harmonic mean of Sensitivity and PPV [52]. Our approach performed well along all four metrics. Fig 2A shows that treeWAS was able to consistently achieve a FPR of zero, indicating tight control over multiple confounding factors. In Fig 2B, we see that the sensitivity of treeWAS varied between zero and one, taking on 11 different values representing how many of the 10 associated loci were correctly identified. While our conservative approach produces moderate sensitivities in the three individual association scores within treeWAS, the contribution of true positive findings by each score gives treeWAS a very high sensitivity overall. Notably, although Score 2 generally achieved higher sensitivity than Scores 1 and 3, it was not always the leading contributor to the cumulative sensitivity of treeWAS in the analysis of these simulated datasets. This highlights the value of using multiple complementary measures to identify associations. Importantly, cumulative benefits in sensitivity were not undermined by cumulative reductions in PPV. Fig 2C reflects the fact that in most cases the total number of false positives found by treeWAS was zero or one. Overall, the high performance of our approach, as indicated by the composite F1 score in Fig 2D, provides strong support for the strategy adopted by treeWAS. This remains true when the number of individuals or loci simulated is varied (see S9 Appendix).

In addition to providing a thorough control over population structure, treeWAS was able to account explicitly for the varying confounding effect of recombination. In the simulated datasets analysed above, the phenotype and associated loci underwent between four and 23 binary state substitutions as a result of the probabilistic simulation process, with an average of 14 substitutions per tree. As such, the probability of chance correlation with the resulting pattern of phenotypic clustering increased as the number of substitutions among non-associated loci was elevated to similar levels by recombination (see S7 Appendix). Fig 3A shows that the FPR of treeWAS remained consistently low as the recombination rate was increased. Because

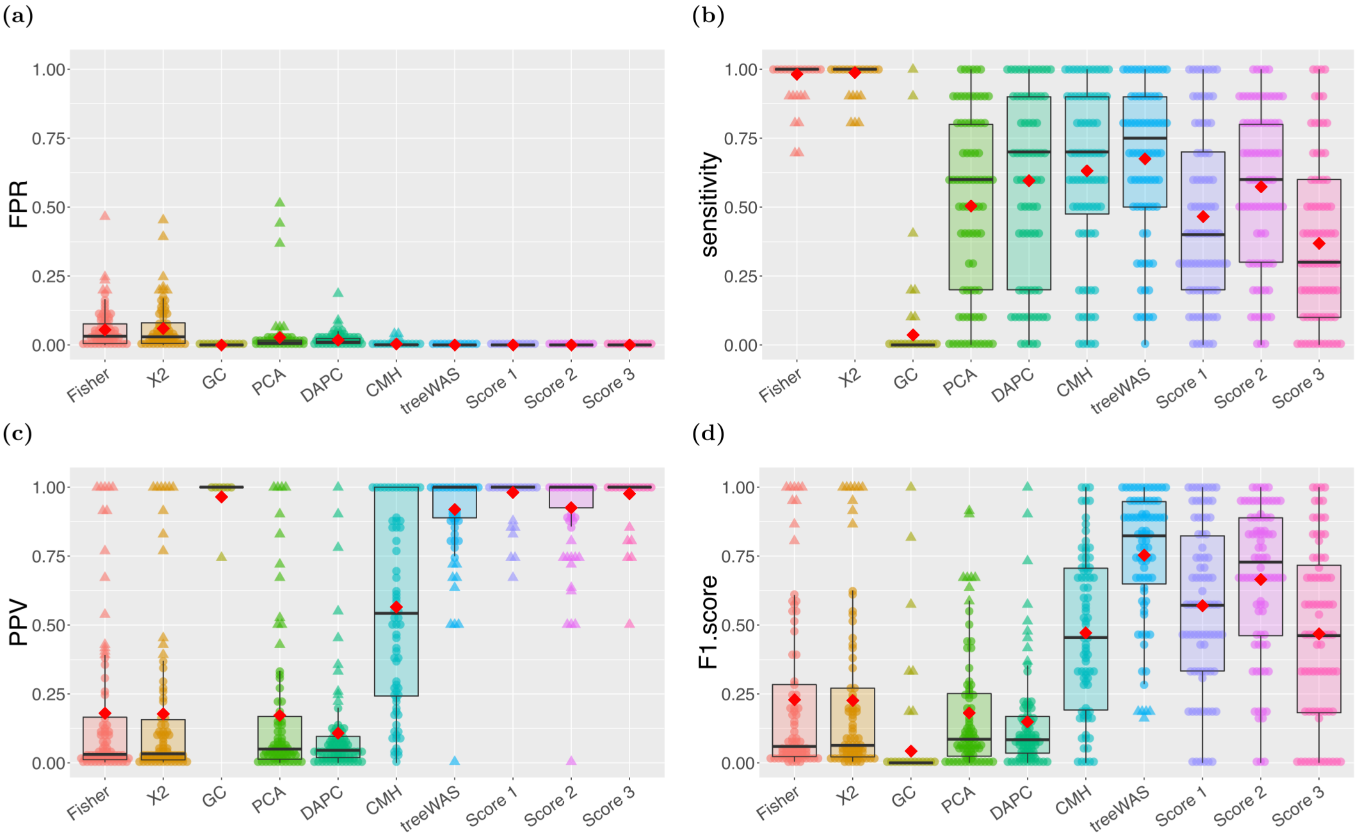

Fig 2. Performance by association test. The performance on simulated datasets for the six comparator GWAS methods and treeWAS, alongside its three association tests individually, is summarised along the four metrics of evaluation. Box plots display the median and interquartile range, red diamonds indicate 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.

https://doi.org/10.1371/journal.pcbi.1005958.g002

data simulation within treeWAS is guided by the empirical homoplasy distribution, the elevated risk of chance association due to recombination was accounted for in the null distribution. Fig 3B illustrates a second implication of this feature: in this analysis, sensitivity decreased with increasing recombination, as treeWAS could no longer attribute significance to some more weakly associated loci when similar patterns of association were likely to occur by chance. Taking into account the parameters of the data causes the impact of recombination on the sensitivity of treeWAS to vary by context (see S4 and S5 Appendices). This data-dependent behaviour is necessary to keep FPR at a minimum (Fig 3A). We do nevertheless see in Fig 3C a slight decline in the PPV of treeWAS, indicating a shift from an average of zero to one false positive findings with increasing recombination. No further PPV losses are observed, however, even in S9 Appendix where frequent gene gain and loss typical of the accessory genome is simulated. Ultimately, as the F1 score in Fig 3D demonstrates, the approach adopted by treeWAS not only produces good overall performance, but by accounting for recombination, it is able to maintain good performance across a range of backgrounds, from purely clonal to frequently recombining.

Comparison with other GWAS methods

Having described the performance of treeWAS on simulated data, we now compare it with the performance of alternative methods. Fig 2A reveals that only treeWAS and the conservative

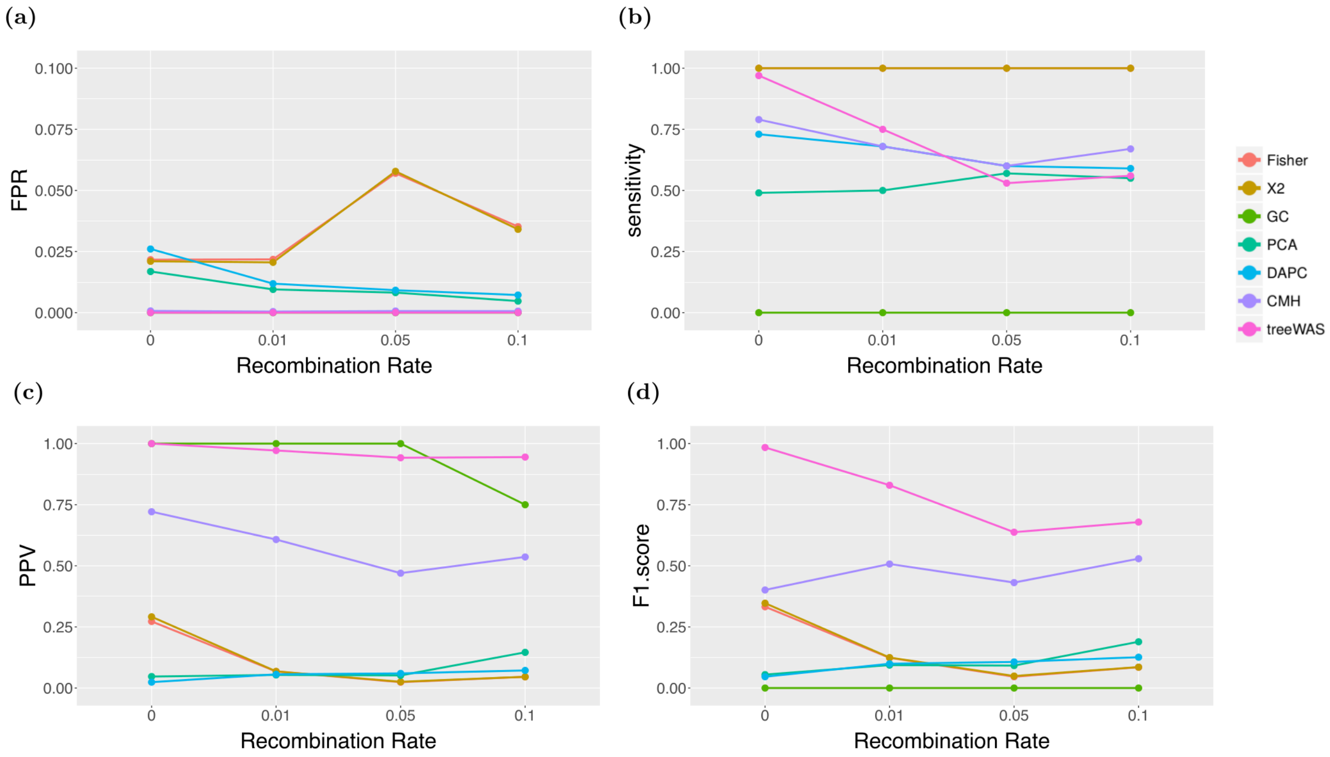

Fig 3. Performance by recombination rate. Interquartile mean performance by GWAS method and recombination rate is plotted along four statistics. **A:** False Positive Rate. **B:** Sensitivity. **C:** Positive Predictive Value. **D:** F1 Score.

https://doi.org/10.1371/journal.pcbi.1005958.g003

GC approach consistently rejected all false positive findings. PCA, DAPC, and the CMH test reduced FPR below the level incurred with no correction for population structure, but still returned undesirably high volumes of false positives. Fig 2A thus suggests that even the most popular dimension-reduction and cluster-based methods do not sufficiently correct for population structure in microbial contexts.

All of the non-phylogenetic approaches to correcting for population structure inherently simplify the extensive genetic relationships between isolates. The k-means clusters used in the CMH test, and the principal components of PCA and DAPC have been previously shown to correspond to the major clades and genealogical divisions of a phylogenetic tree [25, 28, 53]. In PCA, DAPC, and the CMH test, therefore, the user must make a conceptual delineation, at a given height on this tree, between what will and will not be considered parts of the population structure. Our approach, by contrast, works directly with the whole phylogenetic tree, retaining the information it provides at all levels of the clonal population structure. In addition, because the distribution of the phenotype is observed along the tips of the phylogeny but less clear in clusters and PCs, treeWAS is able to account directly for the degree of correlation observed between population structuring alleles and phenotypic clusters. The design of treeWAS therefore allows it to determine what degree of association is unlikely to have arisen by chance, given the evolutionary history inferred. For these reasons, our phylogenetic approach provides a more natural and complete solution to the problem posed by population structure, which drives the FPR gap in Fig 2A between treeWAS and its competitors.

Building on the foundation of low FPR, treeWAS is able to enhance power by drawing on the cumulative findings of multiple tests of association. The non-phylogenetic alternatives, by contrast, face an inherent trade-off between sensitivity and specificity. Fig 2B and 2C show that, consequently, none of these approaches offered simultaneously a high PPV and high sensitivity. The only non-phylogenetic method with an acceptably high PPV was GC (Fig 2C), but it also had the lowest sensitivity (Fig 2B) meaning that in our simulations it almost never found any association, correct or incorrect. PCA and DAPC, despite being too permissive of false positives, achieved only moderate sensitivities, considerably lower than treeWAS (Fig 2B). These methods undermine sensitivity by discounting higher-order lineage effects and proceed with the assumption that remaining genetic variation is ancestrally homogenous [28, 54]. Because these methods reduce power by eliminating variation with every additional PC, increasing the number of selected PCs in an effort to lower FPR in this study 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.

The CMH test more effectively managed the sensitivity-specificity trade-off by applying a more conservative stratified test of association to the genetic data matrix without regressing out any relevant information. In fact, we see from Fig 2B that the sensitivity of the CMH test was only slightly lower than that of treeWAS. The CMH test also had better PPV than PCA and DAPC (Fig 2C), although it fell well below that of treeWAS. Indeed, with an mean PPV of 0.56, almost half of the results identified by the CMH test were false positives, whereas the PPV of treeWAS indicates that 92% of our findings were correct. Overall, the F1 score (Fig 2D) of CMH was similar to that of the lowest-performing individual association test within treeWAS. Yet, the CMH test achieves high F1 score values by adopting the less stringent approach of favouring high sensitivity over high PPV. In practice, it may be preferable to incur modest sensitivity losses so that the number of false positive findings may be kept at a minimum.

Additionally, Fig 3B and 3C show that because the CMH test is naive to recombination, its behaviour differed markedly from that of treeWAS. In response to increasing recombination, the CMH test maintained relatively stable sensitivity, but experienced a decrease in PPV as the number of false positive findings increased, demonstrating a lack of control for recombination. Furthermore, although the composite F1 scores in Fig 3D appear to narrow with increasing recombination, it is important to note the practical implications of the trade-offs being made by treeWAS and the CMH test. Even at the highest recombination rate examined in Fig 3, the CMH test identified only one more true positive than treeWAS. On the other hand, treeWAS found less than one false positive on average, while the CMH test found as many false positives as true positives (Fig 3C). Increasing recombination in other simulations caused the F1 score gap between the CMH test and treeWAS to increase or remain unchanged (S4 and S5 appendices).

Overall, the comparison of methods in Figs 2 and 3 indicates that the performance of the non-phylogenetic methods was limited by multiple factors: the focus on higher level population structure, the inability to control for its confounding effects in sufficient detail, the necessary trade-off between sensitivity and PPV, and the poor response to varying rates of recombination. By deliberately avoiding all of these pitfalls, the design of treeWAS achieved stronger performance on these simulated datasets.

Application to *Neisseria meningitidis*: Identifying penicillin resistance factors

To determine whether our approach could confirm previously-identified associations, and illustrate its applicability to both binary and continuous phenotypes, we applied treeWAS to a dataset of *N. meningitidis* isolates with a penicillin resistance phenotype. We used the *Neisseria* Bacterial Isolate Genome Sequence Database (BIGSdb accessible at https://pubmlst.org/neisseria/, [55]) to download 171 *N. meningitidis* sequences from serogroup B (see S1 File), extracting both the core SNPs (166,848 SNPs) and an accessory gene presence or absence matrix (2,808 genes). We reconstructed the phylogenetic tree from whole-genome sequences with ClonalFrameML to account for recombination [42]. *N. meningitidis* has a high recombination rate, though recombination in *N. meningitidis* is not so rampant as to entirely obscure the clonal genealogy, as would be the case for example in *Helicobacter pylori* [56]. Because treeWAS accounts explicitly for the confounding effects of recombination, our approach was appropriate for this context, provided recombination-aware phylogenetic methods were used [57]. treeWAS completed the analysis of the accessory genome in 23 seconds and analysed the SNPs matrix in 24 minutes.

The penicillin resistance phenotype was analysed in two ways: as a binary and as a continuous variable. The binary phenotype was categorised according to the penicillin minimum inhibitory concentration (MIC), defining susceptible as MIC ≤ 0.06 and resistant as MIC > 0.06. The continuous phenotype was defined as the ranks of the MIC values, rather than the MIC values themselves, whose distribution was highly skewed and which were less informative than the relative MIC values.

Analysis of the accessory gene presence or absence data did not result in the identification of any gene significantly associated with either the binary or continuous penicillin resistance phenotype. However, application of treeWAS to the set of core SNPs led to the identification of many significant loci. Analysis of the binary penicillin resistance phenotype resulted in the identification of 162 significant SNPs, all of which were located in the well-characterised NEIS1753 (*penA*) gene, encoding penicillin-binding protein 2 (see S2 File and S1 Fig). This finding is consistent with the literature, which indicates that penicillin resistance in *N. meningitidis* occurs when altered forms of a penicillin-binding protein (PBP) are produced [58]. Previous work also indicates that the resistance phenotype, and the mosaic structure of *penA*, arise via homologous recombination [59, 60]. It is therefore natural that our analysis uncovered significant associations among SNPs in this gene rather than in other core or accessory genes. Indeed, the alignment displayed in S1A Fig is consistent with previous accounts in the literature describing uniformity in *penA* sequences among susceptible isolates [61] and considerable diversity among those in resistant isolates [62].

Analysis of the continuous penicillin MIC phenotype returned 30 significant SNPs (see S3 File and S2 Fig). The majority of these were also located in the *penA* gene, although SNPs were also identified in three additional genes. In the presence of antibiotics, many loci not essential to the resistance phenotype may confer a slight selective advantage [12]. For example the UDP-N-acetylmuramoylalanyl-D-glutamate–2, 6-diaminopimelate ligase is involved in cell wall formation via peptidoglycan synthesis, the process targeted by penicillin [60]. The two additional genes in which significant SNPs were identified have roles in stress response and DNA damage repair, which may not be directly related to penicillin resistance but which may instead confer a minor fitness advantage that would slightly increase MIC values. We recommend that future laboratory analyses be undertaken to determine whether the statistically significant associations between MIC and these novel candidate loci represent true causal links, via the proposed mechanisms or otherwise. It should be noted that although the classification scheme we adopted designated MIC > 0.06 as "resistant", these isolates would in fact usually be classified as of "intermediate resistance", as they did not exceed the standard resistance threshold of MIC > 1 (see S1 File). In light of the narrow range of MIC values in the sample, it is remarkable that treeWAS was nonetheless able to identify significant associations. By analysing resistance as a continuous variable, our approach was able to retain all of the phenotypic information available. Hence, in spite of the relatively small phenotypic effect observed, treeWAS not only retained the power to detect the central PBP gene, *penA*, but it gained sensitivity to significant SNPs in three additional genes.

Application to *Neisseria meningitidis*: Identifying drivers of invasive disease

The overall design of treeWAS, in particular the implementation of the three association scores in Eqs 1–3, was developed with the aim of detecting genetic loci associated with subtle and complex phenotypes which may not be entirely determined by genetic factors [18]. To

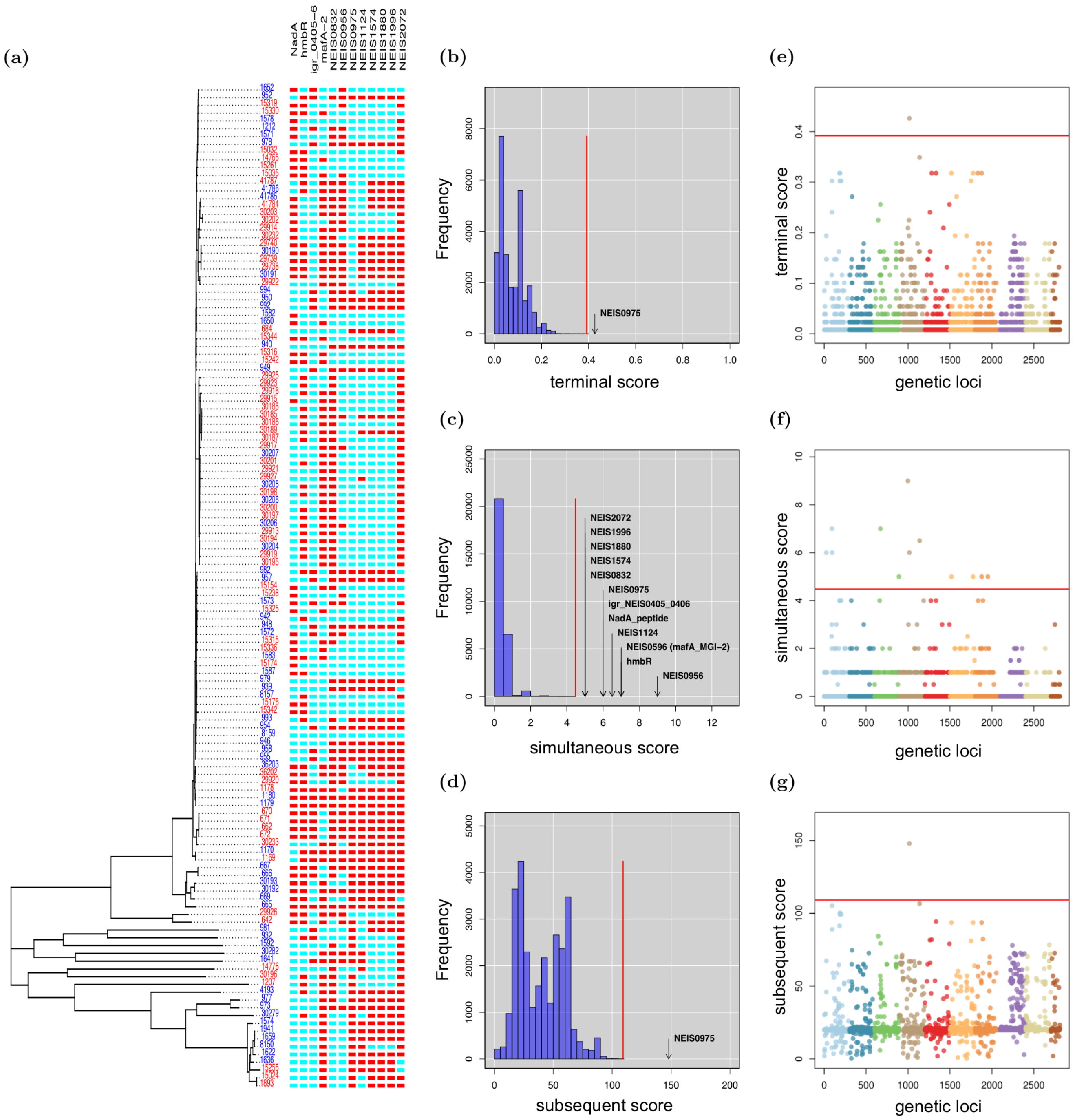

Fig 4. Invasive disease in the *N. meningitidis* accessory genome. treeWAS identified 12 genes associated with invasive disease. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = carrier, red = invasive). At right, an alignment of the 12 significant 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), 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), above which points indicate significant associations.

https://doi.org/10.1371/journal.pcbi.1005958.g004

illustrate this, we applied treeWAS to a separate *N. meningitidis* dataset, with the more challenging phenotype of invasive disease versus carriage. Invasiveness is determined more probabilistically than penicillin resistance, on the basis of both pathogen genetics and external factors, such as host immunity [63].

From the *Neisseria* BIGSdb database, we downloaded 129 European *N. meningitidis* sequences from serogroup C (see S4 File), including both core SNPs (115,386 SNPs) and accessory gene presence or absence data (2,809 genes). ClonalFrameML was used to reconstruct the phylogenetic tree from whole-genome sequences while accounting for recombination [42]. Analyses by treeWAS of the accessory genome and core SNPs datasets were completed in 17 seconds and 7 minutes, respectively.

In the analysis of the accessory gene presence or absence data, treeWAS identified 12 genes associated with carriage or invasiveness (Fig 4, Table 1). Three genes were found to be associated with invasive disease, and the role of each was confirmed by the literature. *NadA* (Neisserial adhesin A) has well-characterised roles in virulence, enabling adhesion, colonisation, and invasion of mucosal cells [64, 65]. *MafA2*, another adhesin, plays a similar role in pathogenic *Neisseria* [66, 67]. Epidemiological evidence and rat models have also linked the haemoglobin receptor protein, *hmbR*, to invasive disease in *N. meningitidis* [68, 69]. Moreover, as this gene is highly conserved, *hmbR* may be a good target for vaccine development [70].

We also identified nine genes whose presence was associated with Neisserial carriage. These included the cell-surface protein encoded by NEIS0956 and the DNA transport competence proteins encoded by NEIS1574, NEIS1880, and NEIS1996, which enable genetic transformation [71, 72]. These genes may confer an adaptive advantage to *N. meningitidis* that enables immune evasion via surface modulation, and favours colonisation and survival in the nasopharyngeal niche [73]. This relationship is not entirely clear, however, as non-pathogenic carriage remains incompletely characterised at a molecular level, despite being a fundamental element of the Neisserial life cycle [74].

In the analysis of core SNPs, treeWAS identified seven associated loci (Fig 5, Table 2). Among these, the *porA* gene is well known for encoding a surface protein that drives hyperinvasivity in *N. meningitidis* [75–78]. Likewise, *gapA-2* may facilitate the adhesion to and invasion of host tissues [79]. As the genetic basis of invasiveness in *N. meningitidis* is not yet fully understood, we anticipate that future work will elucidate the roles of other loci in Table 2.

Table 1. Genes associated with invasive disease in *N. meningitidis*. These 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.

Gene	Gene product
NadA_peptide	*NadA* peptide
hmbR	Haemoglobin receptor protein
igr_NEIS0405_0406	intergenic region between NEIS0405 and NEIS0406
NEIS0596 (mafA_MGI-2)	*MafA-2* adhesin
NEIS0832	hypothetical protein
NEIS0956	cell-surface protein
NEIS0975	hypothetical protein
NEIS1124	hypothetical protein
NEIS1574	DNA transport competence protein
NEIS1880	DNA transport competence protein
NEIS1996	DNA transport competence protein
NEIS2072	putative periplasmic protein

https://doi.org/10.1371/journal.pcbi.1005958.t001

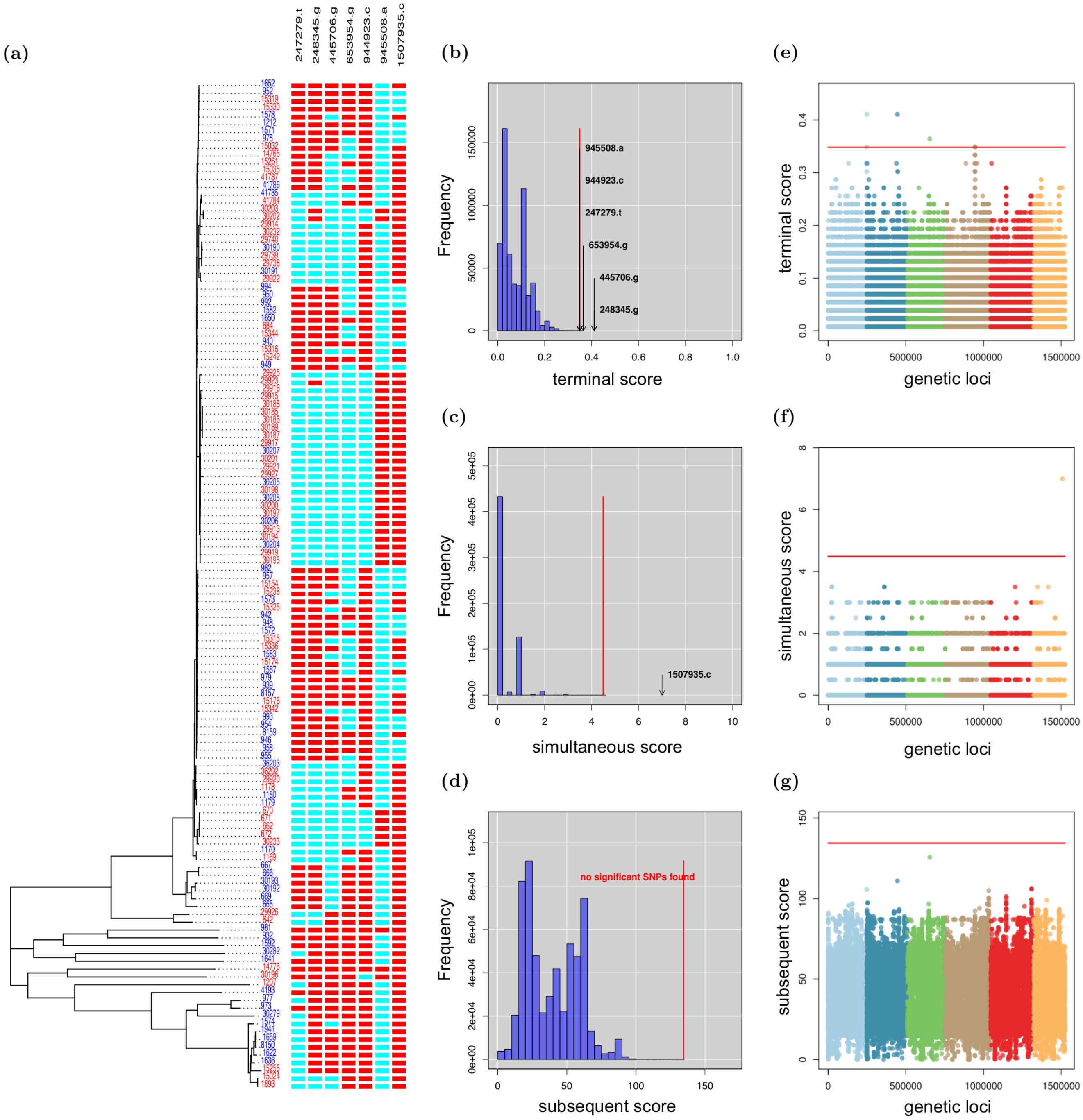

Fig 5. Invasive disease in *N. meningitidis* core SNPs. treeWAS identified 7 SNPs associated with invasive disease. **A:** At left, the clonal genealogy reconstructed with ClonalFrameML, and terminal phenotype (blue = carrier, red = invasive). At right, an alignment of the 7 significant SNPs (blue = 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), 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), above which points indicate significant associations.

https://doi.org/10.1371/journal.pcbi.1005958.g005

Overall, treeWAS was able to identify both previously-known and putatively novel genes and SNPs in significant association with the commensal or invasive phenotype in *N. meningitidis*. Subsequent analyses in the laboratory would be required to confirm that a true biological or causal relationship accompanies this statistical significance.

Table 2. SNPs associated with invasive disease in *N. meningitidis*. These SNPs were identified as significantly associated with invasive disease when treeWAS was applied to 129 whole-genome sequences from *N. meningitidis* serogroup C.

Locus	Gene	Gene product
247279.t	NEIS0343	N-acetylglutamate synthase
248345.g	NEIS0344	hypothetical protein
445706.g	NEIS0614	DNA ligase
653954.g	NEIS0361	hypothetical protein
934483.c	NEIS1348	hypothetical protein
945508.a	NEIS1364 (*porA*)	*PorA*, porin, class 1 outer membrane protein
1507935.c	NEIS2137 (*gapA2*)	glyceraldehyde 3-phosphate dehydrogenase C

https://doi.org/10.1371/journal.pcbi.1005958.t002

Conclusions

Microbial GWAS has the potential to reveal many important features of microbial genomes. Application has however been so far hampered by a lack of well founded and thoroughly tested methodology. Here we proposed a new phylogenetic approach to microbial GWAS that is able to control for the disruptive effects of both population structure and recombination, whilst still retaining a high statistical power to detect real associations. Application to both simulated and real datasets demonstrated that our method is accurate, efficient and versatile, being able to detect associations in both the core and pan-genome and for both categorical and continuous phenotypic measurements. We have implemented our approach in a user-friendly R package, treeWAS, which is freely available for public use at https://github.com/caitiecollins/treeWAS.

Supporting information

S1 Appendix. Computational time to run treeWAS.
(PDF)

S2 Appendix. Choosing the Method of Reconstruction.
(PDF)

S3 Appendix. Derivation of Score 3.
(PDF)

S4 Appendix. Simulation Set A (simple association).
(PDF)

S5 Appendix. Simulation Set B (complementary pathways).
(PDF)

S6 Appendix. Simulation Set C (main set).
(PDF)

S7 Appendix. Simulating homoplasy distributions by recombination rate with SimBac.
(PDF)

S8 Appendix. Significant PCs and clusters in PCA, DAPC, and CMH.
(PDF)

S9 Appendix. Simulation Set C (variable size).
(PDF)

S1 Fig. Application to penicillin resistance in *N. meningitidis* core SNPs. treeWAS identified 140 SNPs associated with penicillin resistance. **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 140 significant SNPs. **B-D:** Null distributions of simulated association scores for (B) Score 1, (C) Score 2, (D) Score 3, a significance threshold (red), 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), above which points indicate significant associations.
(PDF)

S2 Fig. Application to penicillin MIC in *N. meningitidis* core SNPs. treeWAS identified 30 SNPs associated with the ranked penicillin MIC values. **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), 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), above which points indicate significant associations.
(PDF)

S1 File. Metadata for *N. meningitidis* isolates analysed for associations with penicillin resistance and MIC.
(XLS)

S2 File. Loci associated with penicillin MIC in *N. meningitidis*. These loci were identified as significantly associated with the continuous ranked penicillin MIC phenotype, when treeWAS was applied to a dataset of 171 core SNPs extracted from serogroup B *Neisseria meningitidis* whole-genome sequences.
(XLS)

S3 File. Metadata for *N. meningitidis* isolates analysed for associations with invasive disease.
(XLS)

S4 File. Loci associated with penicillin resistance in *N. meningitidis*. These loci were identified as significantly associated with the binary penicillin resistance phenotype, resistant (MIC $< =$ 0.06) *versus* susceptible (MIC $>$ 0.06), when treeWAS was applied to a dataset of 171 core SNPs extracted from serogroup B *Neisseria meningitidis* whole-genome sequences.
(XLS)

Acknowledgments

We would like to thank David Aanensen, Christophe Fraser and Daniel Wilson for useful discussions.

Author Contributions

Conceptualization: Caitlin Collins, Xavier Didelot.

Formal analysis: Caitlin Collins.

Funding acquisition: Xavier Didelot.

Methodology: Caitlin Collins, Xavier Didelot.

Software: Caitlin Collins.

Writing – original draft: Caitlin Collins.

Writing – review & editing: Caitlin Collins, Xavier Didelot.

References

1. WHO. World Health Statistics. Global Health Indicators: Cause-specific mortality and morbidity. World Health Organisation. 2015;p. 72.
2. Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, et al. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. Proc Natl Acad Sci U S A. 2009 17 Nov; 106(46):19545–19550. https://doi.org/10.1073/pnas.0909285106 PMID: 19884497
3. Guinane CM, Ben Zakour NL, Tormo-Mas MA, Weinert LA, Lowder BV, Cartwright RA, et al. Evolutionary genomics of Staphylococcus aureus reveals insights into the origin and molecular basis of ruminant host adaptation. Genome Biol Evol. 2010 12 Jul; 2:454–466. https://doi.org/10.1093/gbe/evq031 PMID: 20624747
4. Kiechle FL, Zhang X, Holland-Staley CA. The -omics era and its impact. Arch Pathol Lab Med. 2004 Dec; 128(12):1337–1345. PMID: 15578876
5. Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, 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. https://doi.org/10.1101/gr.147710.112 PMID: 23299977
6. 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. https://doi.org/10.1038/ng1337 PMID: 15052271
7. Weiss LA, Veenstra-Vanderweele J, Newman DL, Kim SJ, Dytch H, McPeek MS, 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. https://doi.org/10.1038/sj.ejhg.5201239 PMID: 15292919
8. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement factor H variant increases the risk of age-related macular degeneration. Science. 2005 15 Apr; 308(5720):419–421. https://doi.org/10.1126/science.1110359 PMID: 15761120
9. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 15 Apr; 308(5720):385–389. https://doi.org/10.1126/science.1109557 PMID: 15761122
10. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan; 42(Database issue):D1001–6. https://doi.org/10.1093/nar/gkt1229 PMID: 24316577
11. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends Microbiol. 2006 Aug; 14 (8):353–355. PMID: 16782339
12. 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. https://doi.org/10.1186/s13073-014-0109-z PMID: 25593593
13. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015 25 Mar; 25:17–24. https://doi.org/10.1016/j.mib.2015.03.002 PMID: 25835153
14. 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. https://doi.org/10.1038/nrg.2016.132 PMID: 27840430
15. 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. https://doi.org/10.1534/genetics.110.120121 PMID: 20923983
16. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends Microbiol. 2010 Jul; 18 (7):315–322. https://doi.org/10.1016/j.tim.2010.04.002 PMID: 20452218
17. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. Curr Opin Microbiol. 2015; 23:148–154. https://doi.org/10.1016/j.mib.2014.11.016 PMID: 25483351

18. Ansari MA, Didelot X. Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree. Genetics. 2016 Sep; 204(1):89–98. https://doi.org/10.1534/genetics.116.190496 PMID: 27412711

19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug; 38(8):904–909. https://doi.org/10.1038/ng1847 PMID: 16862161

20. 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. https://doi.org/10.1080/01621459.1963.10500879

21. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6. 1901; 2(11):559–572. https://doi.org/10.1080/14786440109462720

22. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010 15 Oct; 11:94. https://doi.org/10.1186/1471-2156-11-94 PMID: 20950446

23. Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, et al. Genomic signatures of human and animal disease in the zoonotic pathogen Streptococcus suis. Nat Commun. 2015 31 Mar; 6:6740. https://doi.org/10.1038/ncomms7740 PMID: 25824154

24. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014 Aug; 10(8):e1004547. https://doi.org/10.1371/journal.pgen.1004547 PMID: 25101644

25. Howell KJ, Weinert LA, Chaudhuri RR, Luan SL, Peters SE, Corander J, et al. The use of genome wide association methods to investigate pathogenicity, population structure and serovar in Haemophilus parasuis. BMC Genomics. 2014 24 Dec; 15:1179. https://doi.org/10.1186/1471-2164-15-1179 PMID: 25539682

26. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, et al. Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. PLoS One. 2016 27 Sep; 11(9):e0163746. https://doi.org/10.1371/journal.pone.0163746 PMID: 27677172

27. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016 16 Sep; 7:12797. https://doi.org/10.1038/ncomms12797 PMID: 27633831

28. Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016 4 Apr; 1:16041. https://doi.org/10.1038/nmicrobiol.2016.41 PMID: 27572646

29. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016 25 Nov; 17(1):238. https://doi.org/10.1186/s13059-016-1108-8 PMID: 27887642

30. 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. https://doi.org/10.1186/s13073-014-0101-7 PMID: 25484920

31. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013 Oct; 45(10):1183–1189. https://doi.org/10.1038/ng.2747 PMID: 23995135

32. Sheppard SK, Didelot X, Jolley KA, Darling AE, Pascoe B, Meric G, et al. Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol. 2013 Feb; 22(4):1051–1064. https://doi.org/10.1111/mec.12162 PMID: 23279096

33. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC. From genomes to phenotypes: Traitar, the microbial trait analyzer. mSystems. 2016; 1(6):e00101–16. https://doi.org/10.1128/mSystems.00101-16 PMID: 28066816

34. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2007 Mar; 175(3):1251–1266. https://doi.org/10.1534/genetics.106.063305 PMID: 17151252

35. Barker D, Pagel M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol. 2005 Jun; 1(1):e3. https://doi.org/10.1371/journal.pcbi.0010003 PMID: 16103904

36. Cohen O, Ashkenazy H, Burstein D, Pupko T. Uncovering the co-evolutionary network among prokaryotic genes. Bioinformatics. 2012 15 Sep; 28(18):i389–i394. https://doi.org/10.1093/bioinformatics/bts396 PMID: 22962457

37. Sokal R, Michener C. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 1958; 38:1409–1438.

38. 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. https://doi.org/10.1093/oxfordjournals.molbev.a025808 PMID: 9254330

39. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987 1 Jul; 4(4):406–425. PMID: 3447015

40. Criscuolo A, Gascuel O. Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics. 2008 26 Mar; 9:166. https://doi.org/10.1186/1471-2105-9-166 PMID: 18366787

41. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981; 17(6):368–376. https://doi.org/10.1007/BF01734359 PMID: 7288891

42. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015 Feb; 11(2):e1004041. https://doi.org/10.1371/journal.pcbi.1004041 PMID: 25675341

43. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015 18 Feb; 43(3):e15. https://doi.org/10.1093/nar/gku1196 PMID: 25414349

44. Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. Syst Biol. 1971 1 Dec; 20(4):406–416. https://doi.org/10.1093/sysbio/20.4.406

45. Swofford DL, Maddison WP. Reconstructing ancestral character states under Wagner parsimony. Math Biosci. 1987; 87(2):199–229. https://doi.org/10.1016/0025-5564(87)90074-5

46. 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 22 Jan; 255(1342):37–45. https://doi.org/10.1098/rspb.1994.0006

47. Felsenstein J. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 1973 Sep; 25(5):471–492. PMID: 4741844

48. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, 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 16 Jul; 110(29):11923–11927. https://doi.org/10.1073/pnas.1305559110 PMID: 23818615

49. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, 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. https://doi.org/10.1086/519795 PMID: 17701901

50. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006 Dec; 2(12): e190. https://doi.org/10.1371/journal.pgen.0020190 PMID: 17194218

51. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst. 1959 Apr; 22(4):719–748. PMID: 13655060

52. Rijsbergen CJV. Information Retrieval. 2nd ed. Newton, MA, USA: Butterworth-Heinemann; 1979.

53. 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 10 Feb; 15:13. https://doi.org/10.1186/s12862-015-0283-7 PMID: 25887041

54. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet. 2008 15 Oct; 17(R2):R143–50. https://doi.org/10.1093/hmg/ddn268 PMID: 18852203

55. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010; 11(1):595. https://doi.org/10.1186/1471-2105-11-595 PMID: 21143983

56. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. ISME J. 2008 2 Oct; 3(2):199–208. https://doi.org/10.1038/ismej.2008.93 PMID: 18830278

57. Collins C, Didelot X. Reconstructing the Ancestral Relationships Between Bacterial Pathogen Genomes. Methods Mol Biol. 2017; 1535:109–137. https://doi.org/10.1007/978-1-4939-6673-8_8 PMID: 27914076

58. Oppenheim BA. Antibiotic resistance in Neisseria meningitidis. Clin Infect Dis. 1997 Jan; 24 Suppl 1: S98–101. https://doi.org/10.1093/clinids/24.Supplement_1.S98 PMID: 8994787

59. 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. https://doi.org/10.1128/jb.176.2.333-337.1994 PMID: 8288526

60. Maiden MC. Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. Clin Infect Dis. 1998 Aug; 27 Suppl 1:S12–20. https://doi.org/10.1086/514917 PMID: 9710667

61. Spratt BG, Zhang QY, Jones DM, Hutchison A, Brannigan JA, Dowson CG. Recruitment of a penicillin-binding protein gene from Neisseria flavescens during the emergence of penicillin resistance in Neisseria meningitidis. Proc Natl Acad Sci U S A. 1989 Nov; 86(22):8988–8992.

62. Zhang QY, Jones DM, Sáez Nieto JA, Pérez Trallero E, Spratt BG. Genetic diversity of penicillin-binding protein 2 genes of penicillin-resistant strains of Neisseria meningitidis revealed by fingerprinting of amplified DNA. Antimicrob Agents Chemother. 1990 Aug; 34(8):1523–1528. https://doi.org/10.1128/AAC.34.8.1523 PMID: 2121092

63. Pizza M, Rappuoli R. Neisseria meningitidis: pathogenesis and immunity. Curr Opin Microbiol. 2015 Feb; 23:68–72. https://doi.org/10.1016/j.mib.2014.11.006 PMID: 25461575

64. Capecchi B, Adu-Bobie J, Di Marcello F, Ciucchi L, Masignani V, Taddei A, 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. https://doi.org/10.1111/j.1365-2958.2004.04423.x PMID: 15660996

65. Comanducci M, Bambini S, Brunelli B, Adu-Bobie J, Aricò B, Capecchi B, et al. NadA, a novel vaccine candidate of Neisseria meningitidis. J Exp Med. 2002 3 Jun; 195(11):1445–1454. https://doi.org/10.1084/jem.20020407 PMID: 12045242

66. 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. https://doi.org/10.1128/JB.06161-11 PMID: 22081399

67. Bentley SD, Vernikos GS, Snyder LAS, Churcher C, Arrowsmith C, Chillingworth T, et al. Meningococcal Genetic Variation Mechanisms Viewed through Comparative Analysis of Serogroup C Strain FAM18. PLoS Genet. 2007 16 Feb; 3(2):e23. https://doi.org/10.1371/journal.pgen.0030023 PMID: 17305430

68. Harrison OB, Evans NJ, Blair JM, Grimes HS, Tinsley CR, Nassif X, et al. Epidemiological evidence for the role of the hemoglobin receptor, hmbR, in meningococcal virulence. J Infect Dis. 2009 1 Jul; 200 (1):94–98. https://doi.org/10.1086/599377 PMID: 19476432

69. Stojiljkovic I, Hwa V, de Saint Martin L, O'Gaora P, Nassif X, Heffron F, et al. The Neisseria meningitidis haemoglobin receptor: its role in iron utilization and virulence. Mol Microbiol. 1995 Feb; 15(3):531–541. https://doi.org/10.1111/j.1365-2958.1995.tb02266.x PMID: 7783623

70. 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. https://doi.org/10.1128/jb.178.15.4670-4678.1996 PMID: 8755899

71. Chen I, Gotschlich EC. ComE, a competence protein from Neisseria gonorrhoeae with DNA-binding activity. J Bacteriol. 2001 May; 183(10):3160–3168. https://doi.org/10.1128/JB.183.10.3160-3168.2001 PMID: 11325945

72. 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 9 Feb; 10:70. https://doi.org/10.1186/1471-2164-10-70 PMID: 19203353

73. Hill DJ, Griffiths NJ, Borodina E, Virji M. Cellular and molecular biology of Neisseria meningitidis colonization and invasive disease. Clin Sci. 2010 9 Feb; 118(9):547–564. https://doi.org/10.1042/CS20090513 PMID: 20132098

74. Capel E, Zomer AL, Nussbaumer T, Bole C, Izac B, Frapy E, et al. Comprehensive Identification of Meningococcal Genes and Small Noncoding RNAs Required for Host Cell Colonization. MBio. 2016 7 Sep; 7(4). https://doi.org/10.1128/mBio.01173-16 PMID: 27486197

75. Urwin R, Russell JE, Thompson EAL, Holmes EC, Feavers IM, Maiden MCJ. Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design. Infect Immun. 2004 Oct; 72(10):5955–5962. https://doi.org/10.1128/IAI.72.10.5955-5962.2004 PMID: 15385499

76. Russell JE, Jolley KA, Feavers IM, Maiden MCJ, Suker J. PorA variable regions of Neisseria meningitidis. Emerg Infect Dis. 2004 Apr; 10(4):674–678. https://doi.org/10.3201/eid1004.030247 PMID: 15200858

77. Derrick JP, Urwin R, Suker J, Feavers IM, Maiden MC. Structural and evolutionary inference from molecular variation in Neisseria porins. Infect Immun. 1999 May; 67(5):2406–2413. PMID: 10225902

78. Suker J, Feavers IM, Achtman M, Morelli G, Wang JF, Maiden MC. The porA gene in serogroup A meningococci: evolutionary stability and mechanism of genetic variation. Mol Microbiol. 1994 Apr; 12 (2):253–265. https://doi.org/10.1111/j.1365-2958.1994.tb01014.x PMID: 8057850

79. Tunio SA, Oldfield NJ, Ala'Aldeen DAA, Wooldridge KG, Turner DPJ. The role of glyceraldehyde 3-phosphate dehydrogenase (GapA-1) in Neisseria meningitidis adherence to human cells. BMC Microbiol. 2010 9 Nov; 10:280. https://doi.org/10.1186/1471-2180-10-280 PMID: 21062461