Published online 22 February 2022

NAR Genomics and Bioinformatics, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/nargab/lqac011

Genome-wide association, prediction and heritability in bacteria with application to *Streptococcus pneumoniae*

Sudaraka Mallawaarachchi[1,*], Gerry Tonkin-Hill[2], Nicholas J. Croucher[3], Paul Turner[4,5], Doug Speed[6,7,8], Jukka Corander[2,9,10] and David Balding[1,8,11,*]

[1]Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia, [2]Parasites and Microbes, Wellcome Sanger Institute, Cambridge CB10 1SA, UK, [3]Faculty of Medicine, School of Public Health, Imperial College, London SW7 2AZ, UK, [4]Cambodia-Oxford Medical Research Unit, Angkor Hospital for Children, Siem Reap 1710, Cambodia, [5]Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LG, UK, [6]Aarhus Institute of Advanced Studies (AIAS), Aarhus University, 8000 Aarhus, Denmark, [7]Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark, [8]UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom, [9]Department of Biostatistics, Faculty of Medicine, University of Oslo, 0372 Oslo, Norway, [10]Helsinki Institute of Information Technology, Department of Mathematics and Statistics, University of Helsinki, Helsinki 00014, Finland and [11]School of BioSciences, University of Melbourne, VIC 3010, Australia

Received November 16, 2021; Revised January 06, 2022; Editorial Decision January 30, 2022; Accepted February 01, 2022

ABSTRACT

Whole-genome sequencing has facilitated genome-wide analyses of association, prediction and heritability in many organisms. However, such analyses in bacteria are still in their infancy, being limited by difficulties including genome plasticity and strong population structure. Here we propose a suite of methods including linear mixed models, elastic net and LD-score regression, adapted to bacterial traits using innovations such as frequency-based allele coding, both insertion/deletion and nucleotide testing and heritability partitioning. We compare and validate our methods against the current state-of-art using simulations, and analyse three phenotypes of the major human pathogen *Streptococcus pneumoniae*, including the first analyses of minimum inhibitory concentrations (MIC) for penicillin and ceftriaxone. We show that the MIC traits are highly heritable with high prediction accuracy, explained by many genetic associations under good population structure control. In ceftriaxone MIC, this is surprising because none of the isolates are resistant as per the inhibition zone criteria. We estimate that half of the heritability of penicillin MIC is explained by a known drug-resistance region, which also contributes a quarter of the ceftriaxone MIC heritability. For the within-host carriage duration phenotype, no associations were observed, but the moderate heritability and prediction accuracy indicate a moderately polygenic trait.

INTRODUCTION

The ability to perform genome-wide analyses of DNA variations has enabled detailed investigations of the genetic architecture of traits in many organisms. In human genetics, the study of association, prediction and heritability across the genome has received considerable attention and the main statistical challenges related to problems such as the robust estimation of SNP (single-nucleotide polymorphism) heritability are being overcome (1,2). Similar studies in bacteria are emerging (3,4); however, the field is still in its infancy, and the pros and cons of many proposed methods have not yet been extensively evaluated using bacterial datasets.

To address this shortcoming, we present a suite of analyses that take into account the challenges of bacterial genetics such as genome-wide linkage disequilibrium (LD) and genome plasticity. Our methods are based on popular methods in human and bacterial genetics, but these are coupled with innovations to better adapt them to bacterial datasets. Our suite of methods uses linear mixed models (LMMs) and linkage disequilibrium score regression (LDSC) to investigate genome-wide association, heritability and heritability partitioning, along with elastic-net regression for trait prediction. We use simulation studies to validate our suite

*To whom correspondence should be addressed. Tel: +61 452589316; Email: smallawaarachchi@gmail.com
Correspondence may also be addressed to David Balding. Email: dbalding@unimelb.edu.au

of methods and demonstrate its capabilities in comparison with current state-of-art methods. We use the methods to analyse three traits, two of them previously unstudied, in *Streptococcus pneumoniae*.

Streptococcus pneumoniae, or the pneumococcus, is a Gram-positive human pathogen that can cause several invasive diseases such as pneumonia, meningitis and sepsis, as well as milder diseases such as acute otitis media and tonsillitis. Typically, pneumococci colonize the nasopharynx of a host asymptomatically and transmit effectively between young children, who frequently carry the bacterium until they develop broad natural immunity. This may be supplemented by vaccination with any of the polysaccharide conjugate vaccines (PCVs), which induce effective protection against some common virulent serotypes.

Several population genomic studies have characterized epidemiological traits of the pneumococcus. In a pioneering study, Lees *et al.* (3), found high heritability of the duration of carriage of *S. pneumoniae* in human hosts. Additionally, the strong genetic control of the binary trait antimicrobial resistance (AMR) is also well established from genome-wide association studies (GWAS) (5–8). However, the quantitative trait minimum inhibitory concentration (MIC) has previously been studied in *Mycobacterium tuberculosis* (9) but not in *S. pneumoniae*. For the two MIC traits, we find high heritability and predictive accuracy, explained by many associations. We also confirm that carriage duration (CD) is a polygenic trait with moderate heritability and predictive accuracy.

Given the increasing availability of large-scale bacterial genetic datasets, the developments presented here will provide a valuable guide to future studies.

MATERIALS AND METHODS

Source of data

The present study is based on nasopharyngeal swab data collected monthly from infants and their mothers in the Maela refugee camp in Thailand between 2007 and 2010 (10). Overall, 23 910 swabs were collected during the original cohort study, from which 19 359 swabs from 737 infants and 952 mothers were processed according to World Health Organization (WHO) pneumococcal carriage detection protocols (11) and/or the latex sweep method (12).

Penicillin and ceftriaxone susceptibilities were assessed using 1 μg oxacillin disks in accordance with the 2007 CLSI guidelines (13). Only isolates with an oxacillin zone diameter of <20 mm were subject to benzyl penicillin and ceftriaxone MIC measurements; other isolates were classified as susceptible.

Preparation of phenotypes

A carriage episode corresponds to one or more consecutive swabs in which a host carries the same *S. pneumoniae* strain. To allow for occasional false negatives in strain identification, we followed (3) and implemented a hidden Markov model, using the R package msm (14), to obtain maximum-likelihood estimates of CD values. Due to differences in immune response to bacterial infections between adults and infants (15), only data from infants were used for CD analyses, but we analysed all MIC values regardless of the host. To obtain approximate normal distributions, we log-transformed all three phenotypes (see Supplementary Figure S1 for histograms).

Preparation of genetic data

We used a published dataset (5) of high quality genome sequences from 2663 isolates, manually selected and aligned to the ATCC700669 reference genome using the snippy pipeline version 4.4.0 (16), with minimum coverage set at the default 10 reads. Of these, 1612 isolates were sampled during 1047 *S. pneumoniae* carriage episodes (mean 1.5, SD 1.0 isolates per episode) in 370 host infants (mean 2.8, SD 1.9 episodes per host). The median CD was 64 days (mean 110, SD 102).

By definition, the sequences from different isolates within the same carriage episode are of the same strain, but there can be sequence variation. For the 337 episodes represented by >1 genome sequence, we used the sequence from the last isolate sampled, which we expect to be the most representative sequence as it may incorporate some effects of host–pathogen interaction that increased CD. However, as within-strain sequence variation is low this choice has little impact, which we checked by repeating analyses using the sequence from a randomly chosen isolate from each of the 337 episodes, finding only negligible variation from the results reported here.

A gene was considered a part of the core genome if it was observed in ≥ 95% of isolates, otherwise it was labelled as *accessory*. Pangenome data were extracted by assembling and annotating the read sequences using Prokka version 1.14.6 (17). Orthologous and paralogous gene clusters were then inferred using the Panaroo pangenome pipeline version 1.2.4, generating a gene presence/absence matrix (18). While the core genome was analysed at each variant site, the accessory genome was analysed at the level of genes, using standardized gene counts. The numbers of accessory genes showing variation in the CD and MIC datasets, respectively, were 2 310 and 2 242.

Association analyses

Testing gap and SNP effects. Five alleles are possible at each variant site, the four nucleotides and gap. Gaps are observed at approximately 71% of variant sites (see Figure 1 for the gap frequency distribution), while two, three and four nucleotide alleles are observed at 71%, 7% and 0.4% of variant sites, respectively. In human genetics, multiallelic SNPs and gaps are both rare and SNP alleles are usually coded as binary, leading to three diploid genotypes that can be coded using two degrees of freedom (df) or 1 df under an additive model. For haploid bacteria, a general coding would require up to 4 df per SNP. The usual approach in previous analyses is a 1 df binary coding indicating presence/absence of the major allele. This coding loses information if the minor alleles have different effects. In particular, gap and SNP effects can differ, due in part to different local-dependence effects of insertion/deletion lengths and recombination.

In previous bacterial GWAS analyses, variant sites with many gaps have often been removed. Reasons include that

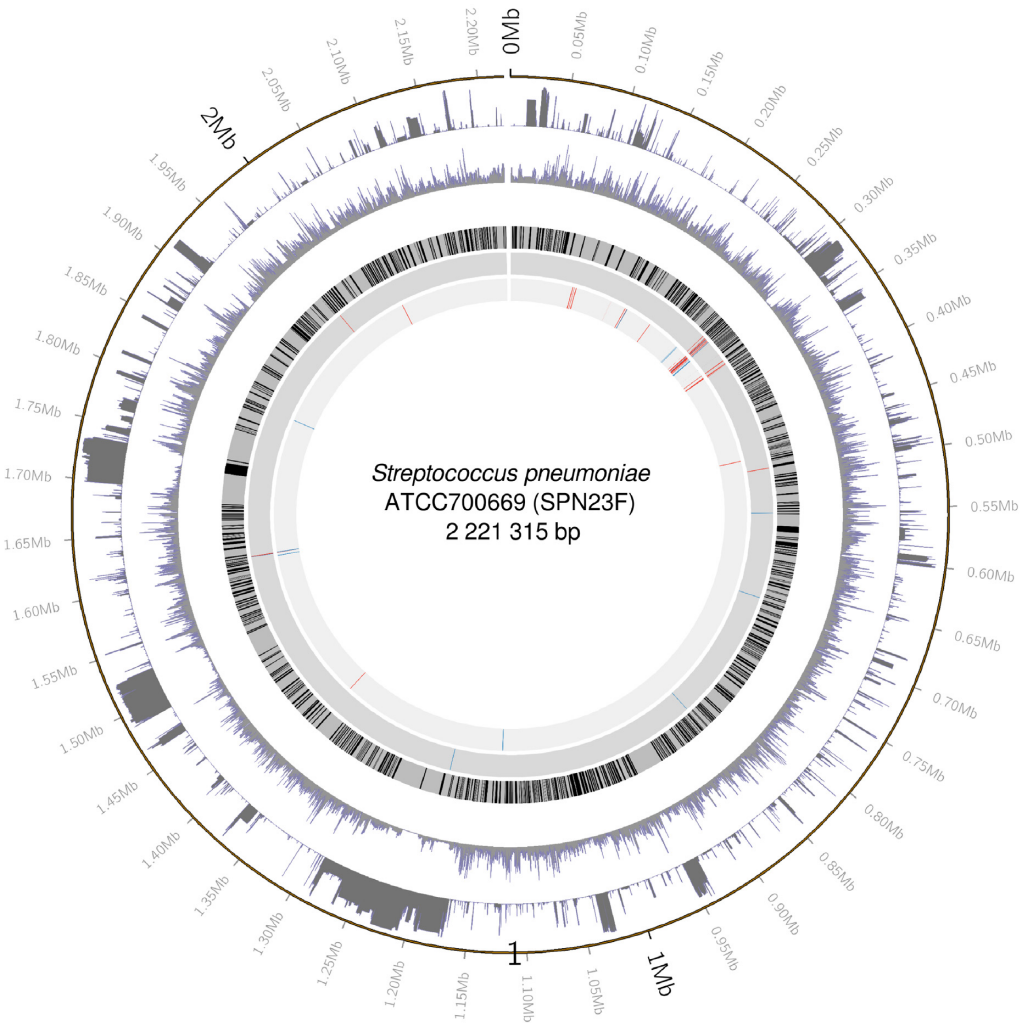

Figure 1. Mapping of association hits to the ATCC700669 reference genome. Working inwards from the outer circle showing basepair positions along the genome, the subsequent circles show the distributions of gap and minor allele frequencies in the MIC dataset, annotated core genes (in black), and SNPs associated with ceftriaxone MIC and penicillin MIC according to the gap test (blue) and SNP test (red). Figure prepared using circos (19).

a gap coding can reflect data quality issues other than a true insertion/deletion sequence state, and that the effects of large insertions or deletions cannot be localized to specific sites. However, insertions and deletions that generate gaps can affect phenotypes, and it is of interest to identify them, while recognizing that the ultimate cause of the association signal may be difficult to decipher. For the core genome variants, we first used a binary gap/non-gap coding to compute a gap test statistic at sites with $\geq$10 of both gap and non-gap sequences. The statistic at the jth variant was the squared standardized effect size: $b_j^2/\mathrm{Var}(b_j)$. Next we computed a 'SNP test' statistic, omitting gap sequences, at sites with $\geq$10 copies of at least two nucleotides. We used a 1 df allele coding equal to the sample frequency of the allele, which assumes that effect sizes vary linearly with allele frequency. For sites with both gap and SNP statistics available, the larger one was used ('max' statistic). In the simulation study we also combined the two statistics using Stouffer's method (divide their sum by $\sqrt{2}$), which we refer to as the 'combi' statistic.

To ensure a family-wise error rate (FWER) of 0.05, we performed 500 permutations of the ceftriaxone MIC phenotype, each time re-running the association analysis pipeline and recording the largest test statistic. From the resulting 500 values, we set the significance threshold for the real-data analyses to be the 25th largest (= 24.8). In comparison, the corresponding Bonferroni threshold based on 133K tests and a χ_1^2 null distribution is 25.8. Therefore, while taking the max of gap and SNP test statistics tends to inflate the null distribution, Bonferroni correction would still be conservative because it ignores the correlations among the statistics. Because of the similarity of the phenotype distributions (Supplementary Figure S1), for penicillin MIC we used the permutation threshold derived for ceftriaxone MIC.

For comparison, we also employed a 1 df association test based on presence/absence of the major allele at each variant, whether gap or a nucleotide, using the Bonferroni threshold. While this test allows some gap effects to be detected, if gap is not the major allele it assumes that the gap and minor nucleotide effects are the same. If gap is the ma-

jor allele then all nucleotide effects are assumed to be the same.

Population structure, phylogeny and clustering. Levels of recombination vary over bacterial species, but in general asexual reproduction leads to strong population structure, which is challenging for association analyses (20,21). Population structure refers to groups of individuals (sub-populations) with greater genetic similarity among them than with other individuals, which causes genome-wide genetic correlations that can confound association signals. Sub-populations may also differ in environmental exposures, which can compound the problem.

There is no complete solution to the problems caused by population structure, and attempts to address them risk discarding true as well as spurious signal. Most approaches introduce either covariates or a genetic random effect into association models to absorb signals that can be explained by population structure, which then do not contribute to association statistics. The variance-covariance matrix **G** of a genetic random effect is assumed known *a priori* based on measures of similarity between pairs of sequences.

Sequence clusters can be used to define either **G**, via cluster distances, or population structure covariates via indicators of cluster membership. Clustering can proceed by constructing a phylogenetic tree that models the evolutionary history of the sequences (22), with nodes of the tree used as cluster identifiers and branch lengths used to define cluster distances. We inferred maximum-likelihood phylogenies of both CD and MIC datasets using IQTree version 2.0.6 (23) under the general time reversible model, with discrete Gamma (+G option) base substitution rates across sites (Figure 2). The model assumes no recombination, which is false for *S. pneumoniae*, and consequently the usefulness of the resulting phylogeny has been questioned (24).

FastBAPS, which extends hierBAPS, (28–30) was also used to cluster the isolates, without reference to a phylogeny. This approach generates an initial clustering using between-variant pairwise distances based on Ward's method (31), then an optimal set of clusters is identified using Bayesian hierarchical clustering (32).

In human studies, **G** was in the past computed from known pedigrees (33) and now usually as a genome-wide average allelic correlation (34). For bacteria, **G** can be defined using allelic correlations under any 1 df allele coding. Despite the success of this approach in human studies, our preliminary analyses could not identify an allele coding that led to good control of population structure effects, although using the gap presence/absence binary indicator gave the best results among those we tried. Conversely, despite the questionable validity of the phylogeny due to it ignoring recombination, defining **G** in terms of lengths of shared phylogenetic branches (35) led to good control of population structure, as evidenced by QQ plots.

Linear mixed model (LMM) analyses. We wish to test $b_j = 0$ within the LMM (36):

$$\mathbf{y} = b_j \mathbf{x}_j + \mathbf{u} + \epsilon, \quad \mathbf{u} \sim \mathcal{N}(0, \sigma_g^2 \mathbf{G}), \quad \epsilon \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}), \quad (1)$$

where $\mathbf{y}$ is a length-n phenotype vector, $\mathbf{x}_j$ is the vector encoding alleles at the jth variant, and $\mathbf{u}$ and ϵ are random vectors of genetic and environmental effects, with $\mathbf{I}$ the $n \times n$ identity matrix.

Pyseer (37) has recently been widely used in bacterial GWAS, and an extensive summary of its models with performance benchmarking is available (38). The Pyseer implementation of (1) is based on FaST-LMM (39) and includes likelihood ratio testing of $b_j = 0$. It requires binary coding of genetic variants, and so can be used for the gap and major-allele tests, but it cannot accommodate the frequency-coding or omission of the gap sequences at each SNP test. To overcome this problem, we used a two-stage LMM/GLS pipeline for the SNP test, similar to EMMAX (40), in which the phenotype for association testing was the residual from fitting (1) with $b_j = 0$. This LMM stage was performed using lme4qtl (33). The b_j were then estimated in a second stage using generalized least squares regression (GLS). In the CD analyses for the SNP test, we were able to incorporate an extra random effect to model shared host in the LMM/GLS pipeline, but for the gap and major-allele tests performed using Pyseer-LMM, this was replaced by a binary covariate indicating previous carriage.

Accessory genome genes were tested using the LMM/GLS pipeline, with a single test based on standardized gene counts.

Phylogenetic method treeWAS. For comparison, we also implemented the phylogeny-based treeWAS (41) using the major-allele coding. Use of a single phylogeny in treeWAS corresponds to an assumption of negligible recombination. As recommended for recombinant species such as *S. pneumoniae* (41), we first implemented the ClonalFrameML pipeline (Supplementary Figure S2) (42). Then treeWAS infers the ancestral phenotype and genotype states at each internal node of the phylogeny, before computing three association test statistics:

(1) **Terminal Score**: It measures sample-wide phenotype-genotype associations between leaves of the phylogeny.
(2) **Simultaneous Score**: It measures parallel changes in both phenotype and genotype on phylogeny branches.
(3) **Subsequent Score**: It measures the proportion of the tree within which genotype and phenotype 'co-exist'. It is equivalent to integrating association scores over all tree nodes.

For each sore, a significance threshold was estimated from null simulations of genetic data at 10 times as many sites as the observed dataset.

Phenotype prediction: whole genome elastic net (wg-enet)

We set up the Pyseer wg-enet model in glmnet (43) in order to use a frequency-based allele coding as in the SNP test except that gaps were counted as an allele. Following Pyseer guidelines (44), we omitted 25% of variants with the largest association P-values, and then removed highly correlated variants at a 0.75 threshold. We verified the finding of (44) that prediction accuracy is improved using weight w_i for the ith isolate, where w_i is proportional to the inverse of the size of the cluster that includes the isolate, and $\sum_i w_i = n$. After centering the phenotype values to have mean zero,

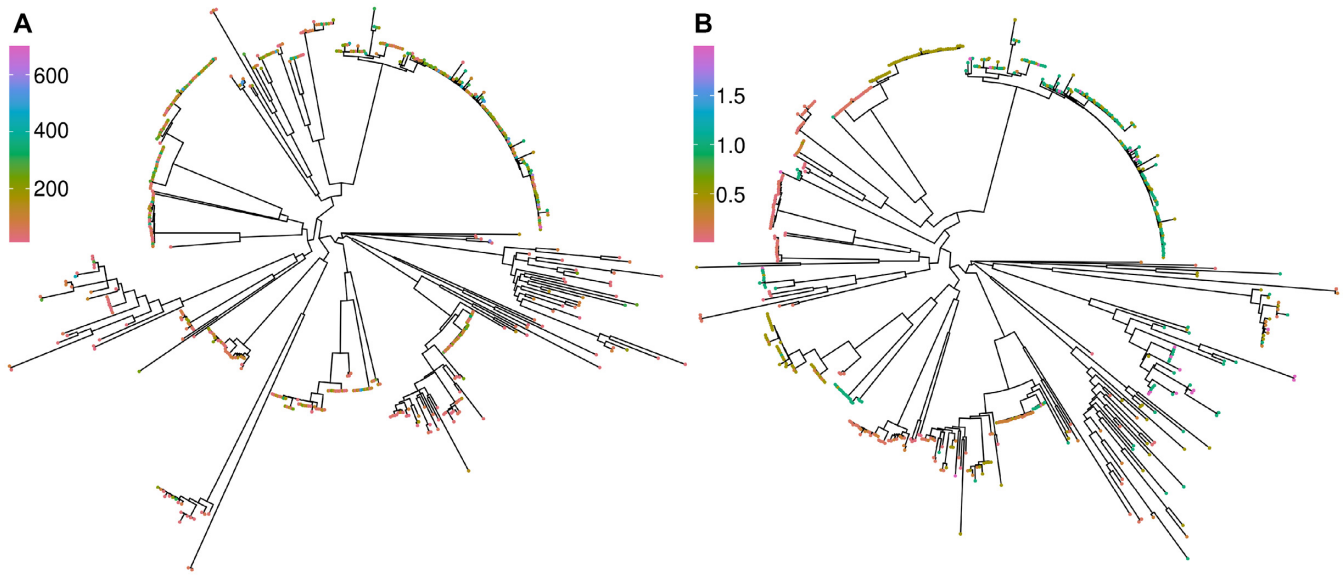

Figure 2. Phylogenies inferred using IQtree2 (**A**) 1047 isolates with a carriage duration (CD) phenotype, indicated by tip colour (in days). (**B**) 1332 isolates with MIC phenotypes, with the penicillin phenotype indicated by tip colour (in μg ml^{-1}). Plots generated after midpoint rooting using R packages ape (25), phytools (26) and ggtree (27).

the ith phenotype value is predicted by $\hat{\mathbf{b}}^T\mathbf{x}_i$, where $\mathbf{x}_i$ is the vector of allele indicators for the ith sequence, and

$$\hat{\mathbf{b}} = \underset{\mathbf{b}}{\operatorname{argmin}}\ \lambda\left[\frac{1-\alpha}{2}\|\mathbf{b}\|_2^2 + \alpha\|\mathbf{b}\|_1\right] + \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - \mathbf{b}^T\mathbf{x}_i)^2. \quad (2)$$

We use cross-validation (CV) to optimise the penalty parameter λ. When $\lambda = 0$ we have weighted least-squares regression, while increasing λ reduces overfitting but introduces bias. By default, both Pyseer and our pipeline set $\alpha = 0.01$. Although this value is close to that for ridge regression ($\alpha = 0$), which retains all predictors in the model, it is large enough that only about 10% of $\hat{\mathbf{b}}$ entries are non-zero.

Ten-fold (10F) and leave-one-strain-out (LOSO) (44) CV were used to assess prediction accuracy. Whereas 10F selects the training sets randomly, which can lead to instances of high similarity between test and training sequences, LOSO is a more challenging prediction task where an entire strain (= FastBAPS cluster) is predicted after training on the other strains.

Estimation of heritability

Genetic effects at different genome sites can interact (epistasis), but we restrict attention to the narrow-sense heritability h^2, with σ_g^2 assumed to be a sum of contributions from individual sites. The LMM estimates $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ (37). For the wg-enet heritability estimation, we used $\hat{h}^2 = R^2$, the proportion of phenotype variance explained by the model with $\alpha = 0$ (ridge regression) (44).

We also estimate h^2 using a modification of LDSC (45):

$$\mathbb{E}[S_j] \approx A + \frac{n-1}{m}h_g^2 l_j \ \text{ where } \ l_j = \sum_{k=1}^{m}\frac{(n-1)r_{jk}^2 - 1}{n-2}. \quad (3)$$

Here, S_j is the association test statistic at variant j, and r_{jk} is the sample correlation of frequency-based allele codes at variants j and k (or gene counts for the accessory genome). Following (46), prior to computing pairwise Pearson correlation coefficients we further transformed the allele codes using Gaussian quantile normalization.

The score l_j involves a sum over the whole genome. In human genetics applications only a neighbourhood of j is included, but the presence of genome-wide LD in *S. pneumoniae* makes it difficult to define a suitable neighbourhood. The definition of l_j also incorporates a bias adjustment (45) that can lead to $l_j < 0$, but typically $l_j \gg 1$. To account for heteroskedasticity and correlations among the S_j, the least-squares estimation of A and h_g^2 in (3) used weights $1/\max(1, l_j)$.

When choosing the testing method to generate the S_j for LDSC, we found that the very strong population structure effects distort the LDSC regression relationship in the absence of any adjustment, yet a fully effective adjustment for population structure was also unsatisfactory because it removed informative signal. The best compromise that we could identify between inadequate control for population structure effects and loss of association signal with effective control, was to compute the major-allele test statistic S_j in the fixed effect model (FEM):

$$\mathbf{y} = \mathbf{v}a + \mathbf{x_j}b_j + \epsilon, \quad (4)$$

where $\mathbf{v}$ is the first principal component (PC) of the sequence distances (explaining a large proportion of genetic variation) and a is the corresponding effect size. For the CD analyses, we also included the previous carriage covariate

in (4). We note again that $\mathbf{v}$ does not remove all population structure effects and the S_j tend to be inflated, but this is not important for LDSC estimation of h^2_g which uses the slope of the relationship of l_j with S_j. Because of inadequate control of population structure using all approaches that we attempted, which included FastBAPS cluster membership indicators and additional principal components (PC), we do not report association results based on this FEM and only use the S_j obtained under this model within LDSC.

As well as estimating genome-wide h^2_g, LDSC is useful for estimating the contributions to h^2_g from specified genome regions. This is challenging because simply omitting variants from a heritability analysis may not exclude their effects due to strong and long-range LD. For the MIC phenotypes, we computed $\hat{h}^2$ in (3) omitting effects from a known drug resistance genome region that includes the important penicillin-binding genes *pbp1a* and *pbp2x*. We first identified a set of large effect-size variants with basepair positions between 285 000 and 340 000 by clumping the frequency-coded variants using correlation threshold 0.85. These variants were used as fixed covariates when re-calculating the S_j for this analysis, which prevents tagging of effects from the omitted region.

Simulation-based validation of analyses

Association testing. Based on the CD dataset (1047 isolates, 134 383 variants), continuous traits were simulated under an additive model with $h^2 \in \{0.1, 0.2, ..., 0.5\}$. In each simulation, 5, 10, 15, 20 or 25 causal variants were randomly selected such that MAF > 0.05 and $r^2 < 0.2$ for all pairs of causal variants. Four replicates were performed for each of the 25 combinations of causal loci and h^2, and the resulting 100 simulated datasets included a total of 1500 causal loci ($\approx 0.011\%$ positives). Association testing was performed using gap/SNP (with both max and combi statistics), major-allele and treeWAS tests.

Heritability estimation. We used BacGWASim (38) to simulate 1000 bacterial genomes of length 250 kb under each of two LD scenarios: lateral gene transfer rate (lgtRate) $=$ 0.2 (Low-LD) and $=$ 0.1 (High-LD). For each scenario and each $h^2 \in \{0.1, 0.2, ..., 0.9\}$, we simulated 100 continuous traits using 10 randomly selected causal variants with MAF > 0.05 and $r^2 < 0.2$. We then computed $\hat{h}^2$ for each of the 1800 traits using LMM, wg-enet and LDSC.

RESULTS

Simulation analyses

The gap/SNP test with max statistic (used in the real-data analyses below) performed better than the alternatives we considered (Figure 3, see AUC values in legend box). At a Bonferroni corrected threshold of 0.05, the sensitivity and specificity were 0.433 and 0.986 for gap/SNP-max, 0.374 and 0.989 for gap/SNP-combi, 0.334 and 0.988 for major allele and 0.238 and 0.996 for treeWAS.

In heritability estimation, LDSC is the best-performing method, although it tends to slightly under-estimate, particularly in the high-LD scenario and for higher h^2 (Figure 4). LMM greatly over-estimates, particularly in the range 0.2

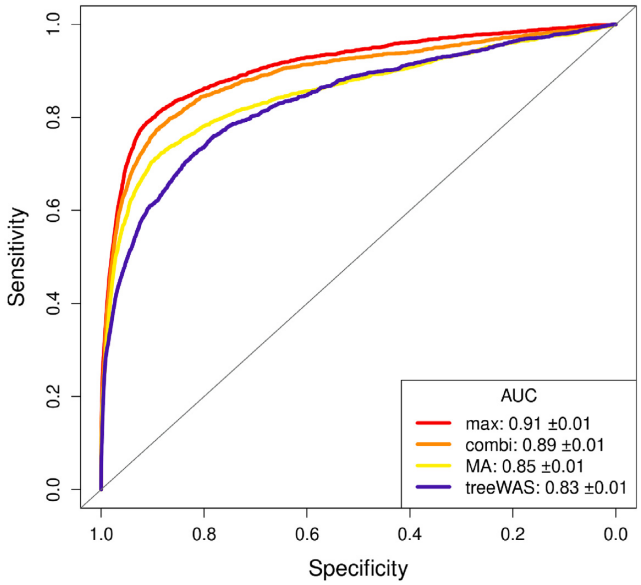

Figure 3. ROC curves for association tests. Based on traits simulated from CD dataset sequences. In the legend box, 'max' and 'combi' are alternative methods for combining gap and SNP test statistics in the gap/SNP test. Only max is used elsewhere in this paper. 'MA' is the major-allele test. For treeWAS, curves were obtained for each of the three scores and the pointwise maximum is shown.

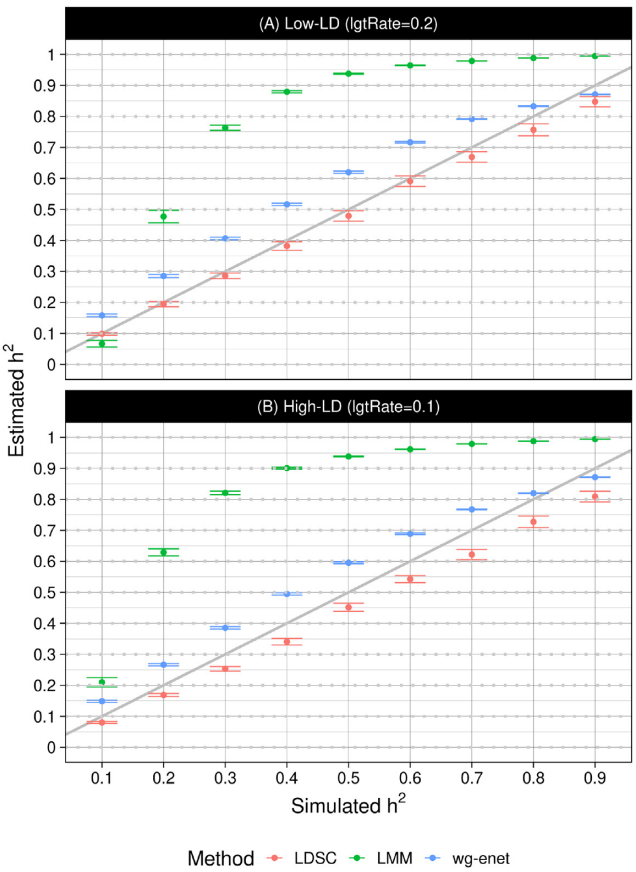

Figure 4. Estimating the heritability of simulated bacterial phenotypes. In the (**A**) Low-LD genome simulation, average relative errors for LMM (green), wg-enet (blue) and LDSC (red) are $28.3 \pm 0.6\%$, $7.7 \pm 0.2\%$ and $-2.1 \pm 0.4\%$. In the (**B**) High-LD genome simulation, average errors for LMM (green), wg-enet (blue) and LDSC (red) $32.4 \pm 0.6\%$, $6.0 \pm 0.2\%$ and $-5.6 \pm 0.4\%$. The error bars show estimated standard error of the mean.

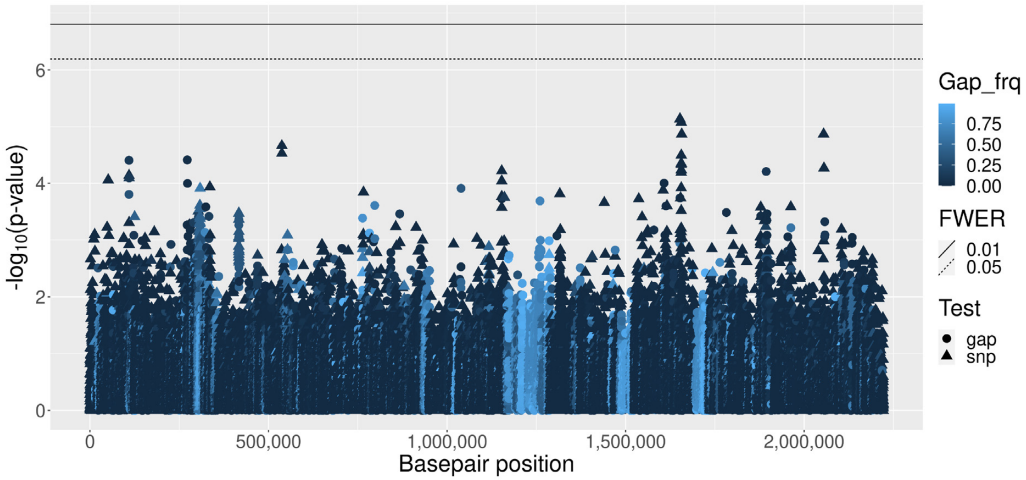

Figure 5. Carriage duration (CD) Manhattan plot for core genome variants. Accessory genes are not shown. See legend for shading that indicates gap frequency and symbol shape indicating gap or SNP test. Basepair positions are obtained from the ATCC700669 reference genome alignment.

$< h^2 < 0.6$. Wg-enet also tends to over-estimate, but it performs slightly better than LDSC when $h^2 > 0.8$. Both LMM and wg-enet estimates are more precise than LDSC but less accurate.

Carriage duration (CD)

None of the 2 310 tested accessory genes were associated with CD. Similarly there were no genome-wide significant results among the 44 097 gap and 91 822 SNP tests at core genome variants (Figure 5). The shared-host random effect explained 1.4% of variance for CD, and $R^2 = 0.0022$ for the previous carriage fixed effect ($\beta = -0.097$, SE =0.026). The QQ-plot (Supplementary Figure S3) indicates some inflation of test statistics suggestive of population structure effects (genome inflation factor, GIF = 1.44). The major-allele test also identified no associations (GIF = 1.22, Supplementary Figure S4) and treeWAS identified 3 hits in 2 genes: *purF* and *polA* (Supplementary Figure S5).

Despite the lack of associations for CD, prediction accuracy (Table 1) and heritability estimates (Table 2) are significantly above zero, suggesting a polygenic trait. As expected, LOSO prediction is less accurate than 10F CV. Pangenome estimates from wg-enet, LMM and LDSC are similar ($0.32 \leq \hat{h}^2 \leq 0.34$) with all methods also agreeing on a negligible contribution to h^2 from the accessory genome. LDSC analyses also confirmed only a small contribution to h^2 from the known drug-resistance region (see Supplementary Figure S6 for LDSC plots). Furthermore, phenotype prediction with allele frequency-based coding of variants slightly outperformed major-allele coding (Supplementary Appendix S2 and Supplementary Figure S7).

We also performed association testing on all 1612 isolates linked to a carriage episode. This analysis identified four sites at basepair positions 1 522 542–1 522 896, near the previously-reported phage hit based on *k*-mer analysis (44). However, our 4 hits are due to the same 15 isolates, of which 6 are from the same long (517 day) episode (see detailed results in Supplementary Appendix S1). Furthermore, when the all-isolates dataset was analysed using treeWAS, 9 associations were identified (Supplementary Appendix S3), but these did not include *purF* and *polA* (reported above) nor the region identified in our LMM analyses. We conclude that we are unable to reliably identify individual associations for CD, but there is good evidence for it being a moderately-heritable polygenic trait.

Table 1. Phenotype prediction. Mean squared error (MSE) and the correlation between observed and predicted test values using 10-fold (10F) and leave-one-strain-out (LOSO) cross validation (CV). Predictions were performed using a wg-enet model ($\alpha = 0.01$) in glmnet, with frequency-based allele coding (all five alleles coded according to their frequency). Approximately 2% of available predictors were used for CD and 1% were used for the two MIC phenotypes. For corresponding results from major-allele coded variants, see Supplementary Appendix S2

Phenotype	10F CV		LOSO CV	
(log scale)	MSE (SE)	Cor (SE)	MSE (SE)	Cor (SE)
CD	0.10 (0.004)	0.55 (0.022)	0.12 (0.005)	0.44 (0.025)
Ceftriaxone MIC	0.03 (0.002)	0.91 (0.005)	0.08 (0.003)	0.77 (0.005)
Penicillin MIC	0.04 (0.003)	0.91 (0.005)	0.13 (0.051)	0.69 (0.014)

Table 2. Heritability estimates ($\hat{h}^2$). The upper and lower values in each cell are for core genome and pangenome (= core genome plus accessory genes). Under 'w/o DR' are results from analyses that omit effects from a genome region that is known to be associated with drug resistance.

Phenotype	LDSC			
	wg	w/o DR	enet	LMM
CD	0.34	0.30	0.34	0.32
with accessory genes	0.34	0.31	0.34	0.32
Ceftriaxone MIC	0.86	0.22	0.92	0.98
with accessory genes	0.87	0.22	0.93	0.98
Penicillin MIC	0.72	0.40	0.94	0.98
with accessory genes	0.72	0.41	0.94	0.98

Minimum inhibitory concentration (MIC) phenotypes

For both MIC phenotypes, from the 2242 accessory genes tested, one (with Panaroo label group_102) showed genome-wide significant association. Gap and SNP tests were per-

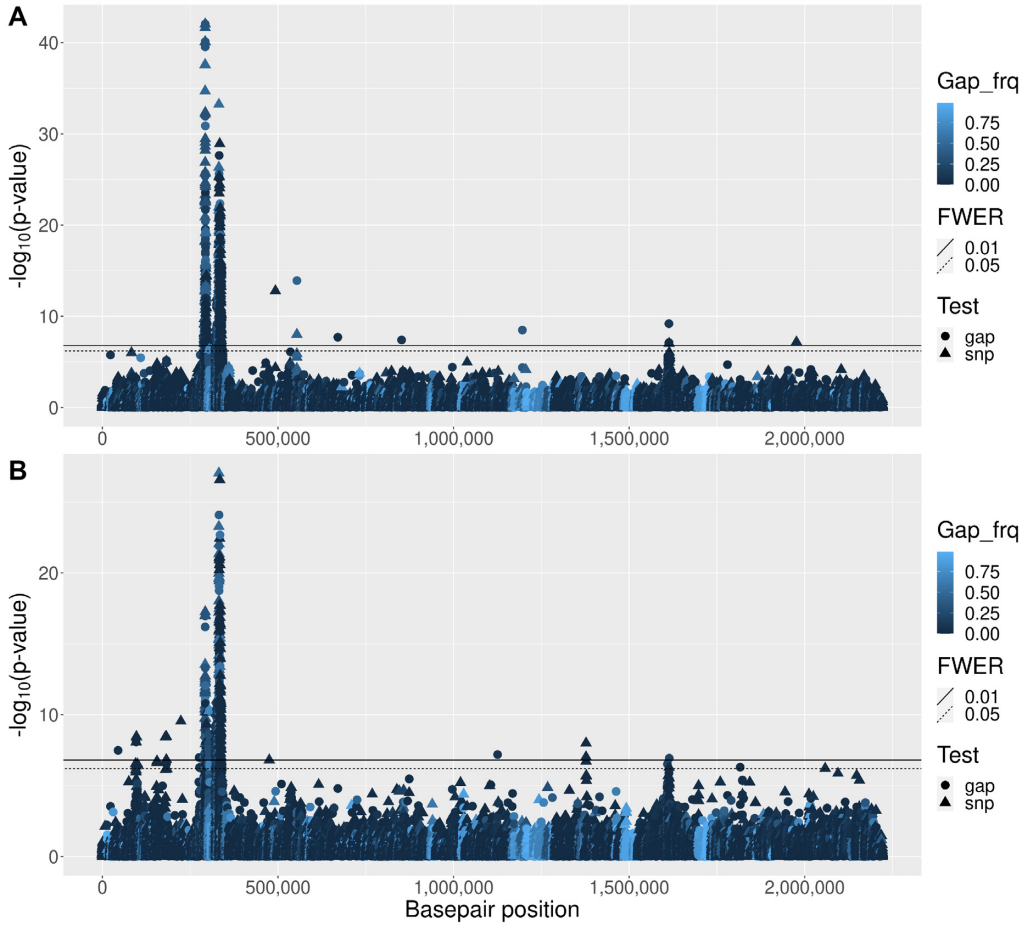

Figure 6. Manhattan plots for (**A**) Ceftriaxone MIC and (**B**) Penicillin MIC. The shading and symbol shapes (see legend) are the same as for Figure 5.

Table 3. Genes showing significant association with MIC phenotypes.

Phenotype (log)	Core genes	Acc. gene
Ceftriaxone only	mraW, clpL, csrR, rplK, aliB, plr, valS	
Both	pbp1a, aliA, pbp2x, mraY, recU, gnd, dexB, luxS, wzg, pbp2b	group_102
Penicillin only	aliB, clpL, wzd, wzh, blpY, galK, hasC, leuB, leuS, murF, recO	

formed at 36 020 and 97 224 core genome sites, respectively. For ceftriaxone MIC and penicillin MIC, respectively, 998 and 833 variants showed genome-wide significance (Figure 6), and 688 and 504 of these were within annotated gene regions of the ATCC700669 reference genome (47) (Table 3). Approximately 35% of hits were from the gap test, associations that have largely been ignored in previous analyses. For ceftriaxone MIC and penicillin MIC, GIF = 1.14 and 1.28 respectively, but the QQ plots (Supplementary Figure S9) suggest that, rather than genome-wide inflation caused by population structure, GIF >1 is due to a large fraction of the genome showing causal association with these highly heritable, polygenic traits.

For ceftriaxone MIC, the largest statistics are of similar magnitude for gap and SNP tests (Figure 7), but for low allele frequencies there are few large gap statistics and many large SNP statistics, suggesting that there are few rare deletions, but many rare nucleotides of large effect. There are also few large gap statistics with frequency >0.6, suggesting few sequence insertions of large effect. Many large SNP statistics with frequency above 0.4 were not recorded as significant under the major-allele test, which may reflect a benefit of frequency-based allele coding. Here, the 7th order regression fit for the 90th percentile was generated using the R package quantreg (48).

Consistent with the simulation results, the gap/SNP test identified more associations than the major-allele and treeWAS tests (combined over the two MIC phenotypes: 1 831 versus 1 419 versus 206), and had lower GIF than the major-allele test (1.14 versus 1.20 and 1.28 versus 1.56; GIF not available for treeWAS). Further results for the major-allele test are in Supplementary Figures S10 and S11, and for treeWAS in Supplementary Figures S12 and S13. The lists of genes identified are in Supplementary Appendix S3.

As expected from the large number of associations, prediction accuracy for both MIC phenotypes is very high under 10F CV (Table 1), but less so for LOSO CV, with high SE for penicillin MIC indicating hard-to-predict clusters (Supplementary Figure S14).

The values of $\hat{h}^2$ also reflect the simulation studies, with LMM > wg-enet > LDSC for both MIC phenotypes (Table 2). Whereas LMM and wg-enet agreed closely across the two MIC phenotypes, the LDSC $\hat{h}^2$ differ consistent with

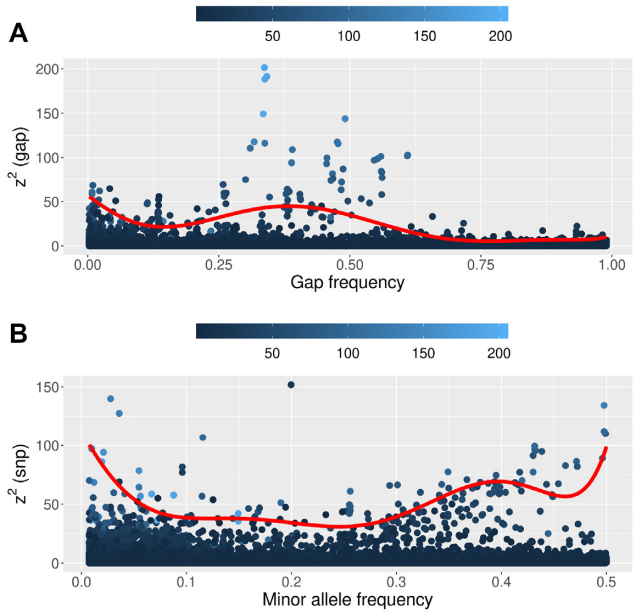

Figure 7. Association test statistics against variant frequency for ceftriaxone MIC. Each point shows the z^2 statistic from (**A**) gap and (**B**) SNP test at a core genome variant. The *x*-axis shows frequencies of (A) gap and (B) minor nucleotide as a fraction of all nucleotides. Points are shaded according to the major-allele test statistic and the red curve shows 7th order regression fit for the 90th percentile. See Supplementary Figure S8 for this analysis on the other two phenotypes.

the higher LOSO prediction accuracy and higher numbers and significance of associations for ceftriaxone MIC compared with penicillin MIC (see Supplementary Figure S6 for LDSC plots). Using LDSC we also estimate that around a quarter of h^2 for ceftriaxone MIC can be attributed to the known drug resistance region, which represents only 2.5% of the core genome. This fraction rises to over half of h^2 for penicillin MIC.

DISCUSSION

We present new and improved approaches for association, prediction and heritability analyses for quantitative bacterial traits. The superior performance of our proposed methods is verified through simulation studies and real-data analyses.

The innovations in our association analysis pipeline include separate testing of gap and SNP effects, with a permutation approach to control FWER and frequency-based allele coding. This approach performed better than the alternatives of major-allele and treeWAS tests, detecting more associations under good control of population structure effects.

Our phenotype prediction analysis used frequency-coded variants within a glmnet-based whole genome elastic net model. For heritability estimation, we have introduced modifications to allow LDSC to be used for bacterial traits, including for genome partitioning of heritability, and verified its advantages over existing approaches.

Using these innovations, we present the first genomic analyses of *S. pneumoniae* minimum inhibitory concentration (MIC) for the beta-lactam antibiotics ceftriaxone and penicillin, finding many associations and high heritability. Prediction of MIC traits was correspondingly accurate under 10F CV.

The genome regions identified as associated with the MIC phenotypes overlap those previousy reported for the binary AMR phenotypes, even in the case of ceftriaxone for which none of the tested isolates was resistant. Many of the associated genes are in the peptidoglycan biosysthesis pathway, including penicillin binding proteins (PBPs: *pbp1a*, *pbp2b*, *pbp2x*) and transferases required for cell wall biogenesis (*mraY* and *mraW* for ceftriaxone MIC). A single heat shock protein (*clpL*) and a gene from the recombination pathway (*recU*) were also identified as associated (6). When present, the group_102 accessory gene is located adjacent to *pbp1a*, which generates an enzyme involved in cell wall remodelling, which may contribute to the association signal for the MIC phenotypes. However, most of the genes identified for the MIC phenotypes are in tight linkage with the three PBPs and may not represent independent effects.

We found no reliable associations for *S. pneumoniae* carriage duration (CD), but strong evidence that it is a polygenic trait of moderate heritability that is predictable from the genome sequence (0.55 and 0.44 correlation between predicted and true phenotype under 10F and LOSO CV, respectively).

A previous analysis of CD using data from the same study (3), provided a lower-bound h^2 estimate of 0.45 using warped-lmm (49), concluding that CD is a highly heritable trait. Our estimates are lower ($\hat{h}^2 \approx 0.33$), which may be due to our decision to use only one isolate per carriage episode (Supplementary Appendix S1).

Penicillin AMR $\hat{h}^2$ in the Maela data set was recently reported in the range 0.67–0.83 (4). Our most reliable (LDSC) estimate for the quantitative penicillin MIC phenotype is within this range (0.72). For ceftriaxone MIC, $\hat{h}^2$ is even higher (0.87).

The attribution of over half of h^2 for penicillin MIC to known drug resistance genome regions in *S. pneumoniae* contrasts with results from *M. tuberculosis*, where the largest reduction in h^2 (measured using GEMMA (50)) was only 27% (9), which is close to our result for ceftriaxone MIC.

In summary, our results support the use of separate testing of gap and SNP effects, and wg-enet for prediction of quantitative traits. We find that LDSC performs best for heritability analyses. Further work is required to assess optimal strategies in a wider range of settings for population structure in bacterial genomes.

DATA AVAILABILITY

All codes, figures and accession details for the genetic data used in this analysis are available at https://github.com/Sudaraka88/bacterial-heritability.

SUPPLEMENTARY DATA

Supplementary data are available at NARGAB Online.

ACKNOWLEDGEMENTS

S.M. would like to acknowledge M. P. N. Perera and members at Melbourne Integrative Genomics for encouragement and insightful discussions.

FUNDING

S.M. was funded by Australian Research Council [DP190103188 to D.B. and D.S.]; ERC [742158 to J.C.]. Funding for open access charge: ERC [742158 to J.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Speed,D. and Balding,D.J. (2019) SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.*, **51**, 277–284.
2. Speed,D., Holmes,J. and Balding,D.J. (2020) Evaluating and improving heritability models using summary statistics. *Nat. Genet.*, **52**, 458–462.
3. Lees,J.A., Croucher,N.J., Goldblatt,D., Nosten,F., Parkhill,J., Turner,C., Turner,P. and Bentley,S.D. (2017) Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife*, **6**, e26255.
4. Mai,T.T., Turner,P. and Corander,J. (2021) Boosting heritability: estimating the genetic component of phenotypic variation with multiple sample splitting. *BMC Bioinform.*, **22**, 164.
5. Chewapreecha,C., Harris,S.R., Croucher,N.J., Turner,C., Marttinen,P., Cheng,L., Pessia,A., Aanensen,D.M., Mather,A.E., Page,A.J. *et al.* (2014) Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.*, **46**, 305–309.
6. Chewapreecha,C., Marttinen,P., Croucher,N.J., Salter,S.J., Harris,S.R., Mather,A.E., Hanage,W.P., Goldblatt,D., Nosten,F.H., Turner,C. *et al.* (2014) Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.*, **10**, e1004547.
7. Mobegi,F.M., Cremers,A.J., De Jonge,M.I., Bentley,S.D., Van Hijum,S.A. and Zomer,A. (2017) Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data. *Sci. Rep.-UK*, **7**, 42808.
8. Skwark,M.J., Croucher,N.J., Puranen,S., Chewapreecha,C., Pesonen,M., Xu,Y.Y., Turner,P., Harris,S.R., Beres,S.B., Musser,J.M. *et al.* (2017) Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis. *PLoS Genet.*, **13**, e1006508.
9. Farhat,M.R., Freschi,L., Calderon,R., Ioerger,T., Snyder,M., Meehan,C.J., de Jong,B., Rigouts,L., Sloutsky,A., Kaur,D. *et al.* (2019) GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nat. Commun.*, **10**, 2128.
10. Turner,P., Turner,C., Jankhot,A., Helen,N., Lee,S.J., Day,N.P., White,N.J., Nosten,F. and Goldblatt,D. (2012) A longitudinal study of Streptococcus pneumoniae carriage in a cohort of infants and their mothers on the Thailand-Myanmar border. *PLoS one*, **7**, e38271.
11. O'Brien,K.L., Nohynek,H. and World Health Organization Pneumococcal Vaccine Trials Carriage Working Group (2003) Report from a WHO Working Group: standard method for detecting upper respiratory carriage of Streptococcus pneumoniae. *Pediatr. Infect. Dis. J.*, **22**, e1–e11.
12. Turner,P., Turner,C., Jankhot,A., Phakaudom,K., Nosten,F. and Goldblatt,D. (2013) Field evaluation of culture plus latex sweep serotyping for detection of multiple pneumococcal serotype colonisation in infants and young children. *PLoS One*, **8**, e67933.
13. Clinical and Laboratory Standards Institute (2017) Performance standards for antimicrobial susceptibility testing. Clinical and Laboratory Standards Institute, PA, USA.
14. Jackson,C. (2011) Multi-state models for panel data: the msm package for R. *J. Stat. Soft.*, **38**, 1–28.
15. Maródi,L. (2006) Neonatal innate immunity to infectious agents. *Infect. Immun.*, **74**, 1999–2006.
16. Seemann,T. (2020) Snippy 4.6.0: Rapid haploid variant calling and core genome alignment [Internet]. https://github.com/tseemann/snippy.
17. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
18. Tonkin-Hill,G., MacAlasdair,N., Ruis,C., Weimann,A., Horesh,G., Lees,J.A., Gladstone,R.A., Lo,S., Beaudoin,C., Floto,R.A. *et al.* (2020) Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.*, **21**, 180.
19. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
20. Corander,J., Croucher,N.J., Harris,S.R., Lees,J.A. and Tonkin-Hill,G. (2019) In: *Bacterial Population Genomics chapter 36*. John Wiley & Sons, Ltd., Oxford, UK, pp. 997–1020
21. Chen,P.E. and Shapiro,B.J. (2015) The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.*, **25**, 17–24.
22. Earle,S.G., Wu,C.-H., Charlesworth,J., Stoesser,N., Gordon,N.C., Walker,T.M., Spencer,C.C., Iqbal,Z., Clifton,D.A., Hopkins,K.L. *et al.* (2016) Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.*, **1**, 16041.
23. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., Von Haeseler,A. and Lanfear,R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.
24. Sakoparnig,T., Field,C. and van Nimwegen,E. (2021) Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife*, **10**, e65366.
25. Paradis,E. and Schliep,K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
26. Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, **3**, 217–223.
27. Yu,G. (2020) Using ggtree to visualize data on tree-like structures. *Curr. Prot. Bioinform.*, **69**, e96.
28. Tonkin-Hill,G., Lees,J.A., Bentley,S.D., Frost,S.D. and Corander,J. (2019) Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.*, **47**, 5539–5549.
29. Cheng,L., Connor,T.R., Sirén,J., Aanensen,D.M. and Corander,J. (2013) Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.*, **30**, 1224–1228.
30. Corander,J., Waldmann,P., Marttinen,P. and Sillanpää,M.J. (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**, 2363–2369.
31. Murtagh,F. and Legendre,P. (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *J. Classi.*, **31**, 274–295.
32. Heller,K.A. and Ghahramani,Z. (2005) Bayesian hierarchical clustering. In: *Proceedings of the 22nd international conference on Machine learning*. Bonn Germany, pp. 297–304.
33. Ziyatdinov,A., Vázquez-Santiago,M., Brunel,H., Martinez-Perez,A., Aschard,H. and Soria,J.M. (2018) lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC Bioinformatics*, **19**, 68.
34. Astle,W. and Balding,D.J. (2009) Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.*, **24**, 451–471.
35. Pagel,M. (1997) Inferring evolutionary processes from phylogenies. *Zool. Scr.*, **26**, 331–348.
36. Yang,J., Zaitlen,N.A., Goddard,M.E., Visscher,P.M. and Price,A.L. (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.*, **46**, 100–106.
37. Lees,J.A., Galardini,M., Bentley,S.D., Weiser,J.N. and Corander,J. (2018) Pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*, **34**, 4310–4312.
38. Saber,M.M. and Shapiro,B.J. (2020) Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microbial. Genom.*, **6**, e000337.
39. Lippert,C., Listgarten,J., Liu,Y., Kadie,C.M., Davidson,R.I. and Heckerman,D. (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.

40. Kang,H.M., Sul,J.H., Service,S.K., Zaitlen,N.A., Kong,S.-y., Freimer,N.B., Sabatti,C. and Eskin,E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
41. Collins,C. and 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**, e1005958.
42. Didelot,X. and Wilson,D.J. (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.*, **11**, e1004041.
43. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.*, **33**, 1.
44. Lees,J.A., Mai,T.T., Galardini,M., Wheeler,N.E., Horsfield,S.T., Parkhill,J. and Corander,J. (2020) Improved prediction of bacterial genotype-phenotype associations using interpretable pangenome-spanning regressions. *Mbio*, **11**, e01344-20.
45. Bulik-Sullivan,B.K., Loh,P.-R., Finucane,H.K., Ripke,S., Yang,J., Patterson,N., Daly,M.J., Price,A.L., Neale,B.M., Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291.
46. Bishara,A.J. and Hittner,J.B. (2015) Reducing bias and error in the correlation coefficient due to nonnormality. *Educ. Psychol. Meas.*, **75**, 785–804.
47. Croucher,N.J., Walker,D., Romero,P., Lennard,N., Paterson,G.K., Bason,N.C., Mitchell,A.M., Quail,M.A., Andrew,P.W., Parkhill,J. *et al.* (2009) Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniae Spain23F ST81. *J. Bacteriol.*, **191**, 1480–1489.
48. Koenker,R. (2021) quantreg: Quantile Regression. R package version 5.86. https://cran.r-project.org/web/packages/quantreg/.
49. Fusi,N., Lippert,C., Lawrence,N.D. and Stegle,O. (2014) Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nat. Comm.*, **5**, 4890.
50. Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.