Identifying lineage effects when controlling for population structure improves power in bacterial association studies

Sarah G. Earle†, Chieh-Hsi Wu†, Jane Charlesworth†, Nicole Stoesser†, N. Claire Gordon†, Timothy M. Walker†, Chris C. A. Spencer‡, Zamin Iqbal‡, David A. Clifton‡, Katie L. Hopkins§, Neil Woodford§, E. Grace Smith§, Nazir Ismail§–7, Martin J. Llewelyn§, Tim E. Peto†, Derrick W. Crook†, Gil McVean†,9, A. Sarah Walker† and Daniel J. Wilson†,2,*

Bacteria pose unique challenges for genome-wide association studies because of strong structuring into distinct strains and substantial linkage disequilibrium across the genome1–3. Although methods developed for human studies can correct for strain structure4–6, this risks considerable loss-of-power because genetic differences between strains often contribute substantial phenotypic variability5. Here, we propose a new method that captures lineage-level associations even when locus-specific associations cannot be fine-mapped. We demonstrate its ability to detect genes and genetic variants underlying resistance to 17 antimicrobials in 3,144 isolates from four taxonomically diverse clonal and recombining bacteria: Mycobacterium tuberculosis, Staphylococcus aureus, Escherichia coli and Klebsiella pneumoniae. Strong selection, recombination and penetrance confer high power to recover known antimicrobial resistance mechanisms and reveal a candidate association between the outer membrane protein nmpC and cefazolin resistance in E. coli. Hence, our method pinpoints locus-specific effects where possible and boosts power by detecting lineage-level differences when fine-mapping is intractable.

Mapping genetic variants underlying bacterial phenotypic variability is of great interest owing to the fundamental role of bacteria ecologically, industrially and in the global burden of disease6–8. Hospital-associated infections including Escherichia coli and Staphylococcus aureus are major global health challenges9–11. Treatment options continue to be eroded by the spread of antimicrobial resistance, with some strains resistant even to antibiotics of last resort12. Genome-wide association studies (GWASs) offer new opportunities to map bacterial phenotypes through inexpensive sequencing of entire genomes, enabling direct analysis of causal loci and functional validation via well-developed molecular approaches12–21. However, bacterial populations typically exhibit genome-wide linkage disequilibrium and strong structuring into geographically widespread genetic lineages or strains that are probably maintained by natural selection13. Approaches to controlling for this population structure have allowed for systematic phenotypic differences based on cluster membership15,16 or, in clonal species, phylogenetic history17–19. However, these and other approaches common in human GWASs20,21 risk masking causal variants because differences between strains account for large proportions of both phenotypic and genetic variability.

Here, we describe a new approach for controlling bacterial population structure that boosts power by recovering signals of lineage-level associations when associations cannot be pinpointed to individual loci because of strong population structure, strong linkage disequilibrium and a lack of homoplasy. We base our approach on linear mixed models (LMMs), which can control for close relatedness within samples by capturing the fine structure of populations more faithfully than other approaches22 and enjoy greater applicability than phylogenetic methods because recombination is evident in most bacteria23,24. Our approach offers biological insights into strain-level differences and identifies groups of loci that are collectively significant, even when individually insignificant, without sacrificing the power to detect locus-specific associations.

Controlling for population structure aims to avoid spurious associations arising from (1) linkage disequilibrium with genuine causal variants that are population-stratified, (2) uncontrolled environmental variables that are population-stratified and (3) population-stratified differences in sampling2. In the four species we investigated, we observed genome-wide linkage disequilibrium and strong population structure, with the first ten principal components (PCs)25 explaining 70–93% of genetic variation, compared with 27% in human chromosome 1 (Supplementary Fig. 1). Controlling artefacts arising from population structure therefore risks a loss of power to detect genuine associations in this large proportion of population-stratified loci.

For example, we investigated associations between fusidic acid resistance and the presence or absence of short 31 bp haplotypes or ‘kmers’ in S. aureus (see Methods and Supplementary Fig. 2). The kmer approach aims to capture resistance encoded by substitutions in the core genome, the presence of mobile accessory genes, or both13. Kmers linked to the presence of fusC, a mobile-element-associated resistance-conferring gene whose product prevents fusidic acid interacting with its target EF-G (ref. 27), showed the strongest genome-wide association by \( \chi^2 \) test (\( P = 10^{-122} \)).
**Figure 1 | Controlling for population structure in bacterial GWASs for fusidic acid resistance in S. aureus.**

**a,** Effect of controlling for population structure using LMM on the significance of the presence or absence of 31 bp kmers. The 200,000 most-significant kmers prior to control for population structure and a random 200,000 are plotted. Each kmer is colour-coded according to the principal component to which it is most strongly correlated, and grey if it is not most strongly correlated to one of the 20 most significant principal components. **b,** Principal components correspond to lineages in the clonal genealogy. Branches are colour-coded by one of the 20 most significant principal components to which they are most correlated. Individual genomes are colour-coded with black or grey lines to indicate fusidic acid resistance and susceptibility, respectively. The circle passing through the line is colour-coded to indicate the phenotype predicted by the LMM. **c,** Wald tests of significance of lineage-specific associations. Some principal components, for example, PC-9, are hashed to indicate that no branch in the clonal genealogy was most strongly correlated with it. Asterisks above the bars, for example PC-25, indicate evidence for lineages associated with particular genomic regions. **d,** Manhattan plot showing significance of unique variants after controlling for population structure, with variants clustered by principal component. The horizontal ordering is randomized. This allows identification of the variants corresponding to the most significant lineage-specific associations.
However, *fusC*-encoded resistance was observed exclusively within strains ST-1 and ST-8. Thus, controlling for population structure using LMM28 reduced the significance to $P = 10^{-39}$, below other loci (Fig. 1a and Supplementary Fig. 3). kmers capturing resistance-conferring substitutions in *fusA,* which encodes EF-G, were propelled to greater significance, because these low-frequency variants were unstratified and LMM improves power in the presence of polygenic effects29 ($P = 10^{-11}$ by χ² test, $P = 10^{-157}$ by LMM). However, *fusA* variants explain only half as much resistance as *fusC* overall.

Although kmers linked to *fusC* did not suffer an outright loss of significance, as penetrance (proportion of *fusC* carriers expressing resistance) was very high, simulations show that for phenotypes with modest effect sizes (for example, odds ratios of 3), controlling for population structure risks loss of genome-wide significance at 59, 75, 99 and 99% of high-frequency causal variants in *M. tuberculosis* ($n = 1,573$), *S. aureus* ($n = 992$), *E. coli* ($n = 241$) and *K. pneumoniae* ($n = 176$) simulations, respectively, with the power loss being greatest when the sample size is low and the number of variants is high (Fig. 2a and Supplementary Fig. 4a).

Methods to limit loss of power such as ‘leave-one-chromosome-out’29 are impractical in bacteria, which typically have one chromosome. Instead, we developed a method to recover information discarded when controlling for population structure. In cases where population stratification reduces the power to detect locus-specific associations, our method infers lineage-specific associations, similar to a phylogenetic regression30,31, without sacrificing the power to detect locus-specific associations when possible.

Figure 2 | Power, false positives, fine mapping and homoplasy in *S. aureus*. Simulation results. a, Controlling for population structure and multiple testing lead to a drastic reduction in power to detect locus effects, compared with the theoretical optimum power for a single locus. The Wald test improves the power several-fold by detecting lineage-specific effects. b, Top: mean numbers of false-positive SNPs and patterns (that is, unique distributions of SNP alleles among individuals) are drastically reduced by controlling population structure with LMM. Bottom: fine mapping precision is very coarse owing to genome-wide linkage disequilibrium. Interpreting lineage effects is useful when the locus-specific signal cannot be fine-mapped. c, Number of times that common SNPs (minor allele frequency (MAF) > 20%) and antibiotic resistance phenotypes have emerged on the phylogeny. d, When homoplasy is high, the power to detect lineage-level associations is much improved, explaining the good power to map antibiotic resistance phenotypes. In the simulations, causal loci were selected at random from high-frequency SNPs (MAF > 20%) in the $n = 992$ isolates and phenotypes simulated per genome with case probabilities of 0.25 and 0.5 for the common and rare alleles, respectively (odds ratio of 3). Genome-wide significance (to detect locus effects) was based on a Bonferroni-corrected $P$ value threshold of $α$, equal to 0.05 divided by the number of SNP patterns.

We observed that leading principal components tend to correspond to major lineages in bacterial genealogies (or ‘clonal frames’32) despite substantial differences in recombination rates (Fig. 1b and Supplementary Fig. 5), reflecting an underlying relationship between genealogical history and principal component analysis33. Principal components are commonly used to control for population structure by including leading principal components as fixed effects in a regression34. The regression coefficients estimated for principal components could therefore be interpreted as capturing lineage-level phenotypic differences, and each principal component tested for an effect on the phenotype. Because principal components are guaranteed to be uncorrelated, defining lineages in terms of principal components, rather than as phylogenetic branches or genetic clusters, minimizes the loss of power to detect lineage-level associations caused by correlations between lineages.

To identify lineage effects we exploited a connection between principal components and LMMs. In an LMM, every locus is included as a random effect in a regression. This is equivalent to including every principal component in the regression as a random effect, increasing the power 2.5-fold. However, identifying loci contributing to the most significant association at individual loci was lost by controlling for population structure, increasing the power 2.5-fold. We thus decomposed the random effects estimated by the LMM to obtain coefficients and standard errors for every principal component (see Methods). We then used a Wald test35 to assess the significance of the association between each lineage and the phenotype.

Our method, implemented in the R package *bugwas,* revealed strong signals of association between fusidic acid resistance and lineages including PC-6 and PC-9 ($P = 10^{-70}$), comparable in significance to the low-frequency variants at *fusA* (Fig. 1c and Supplementary Fig. 6). We next reassessed locus-specific effects by assigning variants to lineages according to the principal component to which they were most correlated, then comparing the significance of variants within lineages. This showed that *fusC* and variants in linkage disequilibrium with *fusC* accounted for the strongest signals within PC-6 and PC-9 ($P = 10^{-34}$ and $10^{-45}$ respectively, Fig. 1d), with the strongest locus-specific associations localized to a 20 kb region containing the staphyllococcal cassette chromosome (SCC), the most significant hit mapping to the gene adjacent to *fusC*. Thus, identifying loci contributing to the most significant lineages provides an alternative to prioritizing variants for follow-up based solely on locus-specific significance.

In simulations, our method was able to recover signals of lineage-level associations in cases where significance at individual loci was lost by controlling for population structure, increasing the power 2.5-fold (*M. tuberculosis*) to 22.0-fold (*E. coli*) (Fig. 2a and Supplementary Fig. 4a). LMM reduced the number of falsely detected single nucleotide polymorphisms (SNPs) by 30-fold (*K. pneumoniae*) to 3,600-fold (*S. aureus*). However, fine-mapping of causal variants to specific chromosomal regions frequently suffered from genome-wide linkage disequilibrium, because linkage disequilibrium is not
that was most significant was not the expected mechanism, unless indicated by *(most significant variant was in physical linkage (PL) with the expected mechanism) or **(most significant variant was in physical linkage (PL) with the expected mechanism). The rank of the most significant result for an expected causal mechanism for each GWAS is reported, plus, in brackets, the gene or mechanism causes resistance, the variant we found is underlined, or referred to by # and ##. R, resistant; S, sensitive; HP, hypothetical protein; trp, transposase; PMQR, plasmid mediated quinoline resistance. See Supplementary Tables 3-6 for more detail.

Table 1 | Number of resistant and sensitive isolates by species and antibiotics, known mechanisms of resistance and main results.

<table>
<thead>
<tr>
<th>Antibiotic</th>
<th>R</th>
<th>S</th>
<th>Resistance mechanism</th>
<th>Resistance determined by</th>
<th>SNP/gene rank</th>
<th>SNP/gene LMM rank</th>
<th>Kmer rank</th>
<th>Kmer LMM rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>189</td>
<td>142</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>6 (tnp)*</td>
<td>6 (tnp)*</td>
</tr>
<tr>
<td>Ampicillin</td>
<td>139</td>
<td>102</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>2 (tmpC)**</td>
<td>3 (tmpC)**</td>
<td>1,403 (tnp)*</td>
<td>470 (tnp)*</td>
</tr>
<tr>
<td>Cefuroxime</td>
<td>55</td>
<td>186</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>1,598 (162-192 upstream)</td>
<td>470 (162-192 upstream)</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>91</td>
<td>150</td>
<td>SNPs in gyrA, gyrB, parC or parE or presence of PMQR</td>
<td>Gene presence or SNPs, or both</td>
<td>1#</td>
<td>1##</td>
<td>1,403 (tnp)*</td>
<td>470 (tnp)*</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>48</td>
<td>193</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gentamicin</td>
<td>67</td>
<td>174</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tobramycin</td>
<td>53</td>
<td>123</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>762 (tnp)*</td>
<td>837 (tnp)*</td>
</tr>
<tr>
<td>Cefuroxime</td>
<td>46</td>
<td>130</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>762 (tnp)*</td>
<td>1,480 (tnp)*</td>
</tr>
<tr>
<td>Ceftriaxone</td>
<td>35</td>
<td>141</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>771 (tnp)*</td>
<td>812 (tnp)*</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>34</td>
<td>142</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>2#</td>
<td>2#</td>
<td>1,853##</td>
<td>4,427##</td>
</tr>
<tr>
<td>Gentamicin</td>
<td>31</td>
<td>145</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tobramycin</td>
<td>36</td>
<td>140</td>
<td>(\beta)-lactamases</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>41</td>
<td>1,589</td>
<td>embB</td>
<td>SNPs</td>
<td>2 (rpoB)**</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ethambutol</td>
<td>239</td>
<td>1,470</td>
<td>katG, fabG1</td>
<td>SNPs</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Isoniazid</td>
<td>45</td>
<td>1,662</td>
<td>pncA</td>
<td>SNPs</td>
<td>142 (rpoB)**</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Rifampicin</td>
<td>86</td>
<td>1,487</td>
<td>rpoB</td>
<td>SNPs</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S. aureus</td>
<td>242</td>
<td>750</td>
<td>grlA or gyrA</td>
<td>SNPs</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>216</td>
<td>776</td>
<td>ermA, ermC, ermT</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Erythromycin</td>
<td>84</td>
<td>908</td>
<td>fusA or presence of fusB or fusC</td>
<td>Gene presence or SNPs, or both</td>
<td>4# (SAS0037)*</td>
<td>1#</td>
<td>75# (SAS0040)*</td>
<td>1#</td>
</tr>
<tr>
<td>Fusidic acid</td>
<td>11</td>
<td>981</td>
<td>aac(6)-I</td>
<td>Gene presence</td>
<td>1+ GNAT</td>
<td>1+ GNAT</td>
<td>1+415 bases upstream to 100 bases downstream</td>
<td>1+415 bases upstream to 100 bases downstream</td>
</tr>
<tr>
<td>Gentamicin</td>
<td>824</td>
<td>168</td>
<td>blaz</td>
<td>Gene presence</td>
<td>1</td>
<td>1</td>
<td>2 (blaI)*</td>
<td>1+ SCCmec genes</td>
</tr>
<tr>
<td>Penicillin</td>
<td>216</td>
<td>776</td>
<td>mecA</td>
<td>Gene presence</td>
<td>1+ mecRl</td>
<td>1+ SCCmec genes</td>
<td>1+ plasmid genes</td>
<td></td>
</tr>
<tr>
<td>Methicillin</td>
<td>46</td>
<td>946</td>
<td>tetK, tetL or tetM</td>
<td>Gene presence</td>
<td>2 (repC)*</td>
<td>2 (repC)*</td>
<td>1+ plasmid genes</td>
<td></td>
</tr>
<tr>
<td>Tetracycline</td>
<td>15</td>
<td>308</td>
<td>SNPs in dfrB, presence of dfrG or dfrA</td>
<td>Gene presence or SNPs, or both</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Trimethoprim</td>
<td>8</td>
<td>984</td>
<td>rpoB</td>
<td>SNPs</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
generally organized into physically linked blocks along the chromosome (Fig. 2b and Supplementary Fig. 4b), underlining the importance of recovering power by interpreting lineage effects.

We noted a trade-off to interpreting lineage effects, because they are susceptible to confounding with population-stratified differences in environment or sampling (Supplementary Fig. 7). Therefore, non-random associations between lineages and uncontrolled variables that influence phenotype risk false detection of lineage-level differences.

Confronted with a strong population structure and genome-wide linkage disequilibrium in bacteria, we wished to test empirically the ability of GWASs to pinpoint genuine causal variants more generally. We therefore conducted 26 GWASs for resistance to 17 antimicrobials in 3,144 isolates across the major pathogens _M. tuberculosis_35, _S. aureus_37, _E. coli_ and _K. pneumoniae_38 (Supplementary Fig. 8).

We supplemented the kmer approach by surveying the variation in SNPs and gene presence or absence. We imputed missing SNP calls by reconstructing the clonal frame followed by ancestral state reconstruction, an approach that generally outperformed imputation using Beagle (Supplementary Table 1, see Methods).

Correlated phenotypes caused by the presence of multi-drug-resistant isolates led to significant results in unexpected loci or regions in some analyses. A combination of first-line drug regimens contributes to multi-drug resistance co-occurrence in _M. tuberculosis_, which led to spurious associations as the top hit before controlling for population structure between ethambutol and pyrazinamide resistance and SNPs in rifampicin resistance-conferring _ropB_. Even after controlling for population structure, these associations remained genome-wide significant at _P_ = 10^−45 and _P_ = 10^−54.

Antimicrobial resistance has arisen over 20 times per drug in the _M. tuberculosis_ tree, through frequent convergent evolution (Supplementary Fig. 4c and Supplementary Fig. 8). Within a single gene, such as _ropB_, there are multiple targets for selection. Both SNP and kmer-based approaches correctly identified variants in known resistance-causing codons, but greater significance was attained in the latter because the targets for selection were typically within 31 bp (Supplementary Fig. 9a). In these cases, absence of the wild-type allele was found to confer resistance, with power gained by pooling over the alternative mutant alleles.

For each drug and species, we evaluated whether the most significant hit identified by GWAS matched a known causal variant36–38 (Supplementary Table 2). By this measure, the performance of GWASs across species was very good, identifying genuine causal loci or regions in physical linkage with those loci for antimicrobial resistance in 25/26 cases for the SNP and gene approach and the kmer approach after controlling for population structure (Table 1 and Supplementary Table 3). For accessory genes such as _β_-lactams, in particular, mobile element-associated regions of linkage disequilibrium were often detected together with the causal locus (Supplementary Fig. 9b).

Genuine resistance-conferring variants were detected in all but one study, demonstrating that the high accuracy attained in predicting antimicrobial resistance phenotypes from genotypes known from the literature37,39 is mirrored by good power to map the genotypes that confer antimicrobial resistance phenotypes using GWASs. However, these results also reflect the extraordinary selection pressures exerted by antimicrobials. High homoplasy at resistance-conferring loci caused by repeat mutation and recombination breaks down linkage disequilibrium, assisting mapping (Fig. 2c and Supplementary Fig. 4c).

For one drug, cefazolin, in _E. coli_, we identified a variation in the presence of an unexpected gene as the most strongly associated with resistance, _nmrc_ ( _P_ = 10^−125). This gene encodes an outer membrane porin over-represented in susceptible individuals. Permeability in the _Salmonella typhimurium_ homologue mediates resistance to other cephalosporin _β_-lactams40, making this a strong candidate for a novel resistance-conferring mechanism discovered in _E. coli_.

Population structure presents the greatest challenge for GWASs in bacteria, because of the inherent trade-off between the power to detect genuine associations of population-stratified variants and robustness to unmeasured, population-stratified confounders. By introducing a test for lineage-specific associations, we allow these signals to be recovered even in the absence of homoplasy, while acknowledging the increased risk of confounding. Detecting lineage effects is valuable, because characterizing phenotypic variability in terms of strain-level differences is helpful for biological understanding and it permits the prediction of traits, including clinically actionable phenotypes, from strain designation.

Identifying loci that contribute to the most significant lineage-level associations offers flexibility in the interpretation of bacterial GWASs, where it will often be difficult to pinpoint significance to individual locus effects and where linkage disequilibrium can make the fine-mapping of causal loci a genome-wide problem. Loci can be prioritized for follow-up by identifying groups of lineage-associated variants that collectively show a strong signal of phenotypic association, but which cannot be distinguished statistically. This strategy provides an alternative to prioritizing variants based solely on locus-specific significance, but it carries risks, because lineage-associated effects are more susceptible to confounding with population-stratified differences in environment or sampling. This trade-off between power and robustness underlines the importance of functional validation for bacterial GWASs going forward.

Methods

Linear mixed model. In the LMM40–45, the phenotype is modelled as depending on the fixed effects of covariates including an intercept, the ‘foreground’ fixed effect of the locus whose individual contribution is to be tested, the ‘background’ random effects of all the loci whose cumulative contribution to phenotype variability we will decompose into lineage-level effects, and the random effect of the environment:

\[ \text{phenotype} = \text{covariates} + \text{foreground locus} + \text{background loci} + \text{environment} \]

Formally,

\[ y_i = \beta_0 + \sum a_i \gamma_i + X_i \beta_1 + \sum X_i \gamma_i + \epsilon_i \]

where there are _n_ individuals, _c_ covariates, _l_ loci, _j_ is the foreground locus, _y_i_ is the phenotype in individual _i_, _W_j_ is covariate _j_ in individual _i_, _α_i_ is the effect of covariate _j_ on individual _i_, _β_j_ is the foreground effect of locus _j_, _γ_j_ is the background effect of locus _j_ and _ε_i_ is the effect of the environment (or error) on individual _i_. Biallelic genotypes are numerically encoded as _−1_ (common allele) or _1_ (rare allele), where _f_j_ is the frequency of the rare allele at locus _j_. This convention ensures that the mean value of _X_i_ over individuals _i_ is zero for any locus _j_. Because triallelic and tetraallelic loci are rare, we use only biallelic loci to model background effects. When the foreground locus is triallelic ( _K_ = 3) or tetraallelic ( _K_ = 4), the genotype in individual _i_ is encoded as a vector indicating the presence (1) or absence (0) of the first ( _K_ − 1) alleles and _β_j_ becomes a vector of length ( _K_ − 1).

Treating the background effects of the loci as random effects means the precise values of coefficients _γ_j_ are averaged. The _γ_j_ are assumed to follow independent normal distributions with common mean 0 and variance _σ_j^2_ to be estimated. As most loci are expected to have little or no effect on a particular phenotype, this tends to constrain the magnitude of the background effect sizes to be small. The environmental effects are also treated as random effects assumed to follow independent normal distributions with mean 0 and variance _γ_j^2_. The model can be rewritten in matrix form as

\[ y = X \beta + X_1 \gamma_1 + \epsilon \]

with

\[ u = X_1 \gamma_1 + \epsilon \]

\[ \mu \sim \text{MVN}(0, \sigma_j^2 \mathbf{I}) \]

\[ \epsilon \sim \text{MVN}(0, \gamma_j^2 \mathbf{I}) \]

where _u_ represents the cumulative background effects of the loci, _MVN_ denotes the multivariate normal distribution, _I_j_ is an _n_ × _n_ identity matrix, and _K_ is an _n_ × _n_ relatedness matrix defined as _K_ = _XX^T_, which captures the genetic covariance between individuals.
Testing for locus effects. To assess the significance of the effect of an individual locus on the phenotype, controlling for population structure and background genetic effects, the parameters of the linear mixed model \( \alpha_i, \beta_i, \lambda_i \) and \( \tau \) were estimated by maximum likelihood, and a likelihood ratio test with \( (k-1) \) degrees of freedom was performed against the null hypothesis that \( \beta_i = 0 \) using the software GEMMA.

Testing for lineage effects. Because controlling for population structure drastically reduces the power at population-stratified variants, and because a large proportion of variants are typically population-stratified in bacteria, we recovered information from the LMM regarding lineage-level differences in phenotypy. We defined lineages using principal components because we observed that principal components tend to trace paths through the clonal frame genealogy corresponding to recognizable lineages (as seen by the branch colouring in Fig. 1b and Supplementary Fig. 5) and because principal components are mutually uncorrelated, minimizing loss of power to detect differences between lineages due to correlations. Principal components were computed based on biallelic SNPs using the R function `prcomp()`, producing an \( L \) by \( n \) loading matrix \( D \) and an \( n \) by \( n \) score matrix \( T \) where \( T = XD \).\( D \) records the contribution of biallelic SNP \( t \) to the definition of principal component \( j \).

Point estimates and standard errors for the background locus effects are usually overestimated because the assumed normal distribution with common mean \( 0 \) and variance \( \lambda^2 \) tends to cause them to be small in magnitude and not significantly different from zero. However, cumulatively, the background locus effects can capture systemic phenotypic differences between lineages. We therefore recovered the post-data distribution (equivalent to an empirical Bayes posterior distribution) of the background locus random effects, \( \gamma \), from the LMM, and reinterpreted it in terms of lineage-level differences in phenotypy.

Empirically, we found that the post-data distribution of the background random effects was generally insensitive to the identity of the foreground locus and comparable under the null hypothesis (\( \beta_i = 0 \)). We therefore calculated the mean and variance–covariance matrix of the multivariate normal post-data distribution of \( \gamma \) in the LMM null model. These are equivalent to those of a ridge regression\(^{48} \) and were computed as

\[
\mu = (X^TX + 1/\lambda^2I)^{-1}y + \sum_i (rX^Ti + 1/\lambda^2I)^{-1}S_i^j
\]

respectively. Both \( \lambda \) and \( \tau \) were estimated by GEMMA under the LMM null model. Using the inverse transformation of the biallelic variants from PCA, \( X = TD \), the background random effects can be rewritten in terms of the contribution of the \( n \) principal

\[
u = X_gy_1 + \ldots + X_gy_n
\]

where \( g = D^{-1}y \) and \( g_i \) is the background effect of principal component \( j \) on the phenotype. We computed the mean and variance of the post-data distribution of \( g \) as

\[
m = D^{-1}S \quad \text{and} \quad \text{var}(g) = (D^{-1}S D)^{-1}
\]

Restricting analysis to sites called in all genomes is undesirable while ignoring uncalled sites by removing individuals with missing data at individual sites generates \( P \) values that cannot be validly compared between sites because they are calculated using data from differing sets of isolates.

SNP imputation. Because Illumina sequencing is inherently more error-prone than Sanger sequencing, strict filtering is required for reliable mapping-based SNP calling, contributing to a small but appreciable frequency of uncalled bases in the genome due to ambiguity or deletion. Restricting analysis to sites called in all genomes is undesirable while ignoring uncalled sites by removing individuals with missing data at individual sites generates \( P \) values that cannot be validly compared between sites because they are calculated using data from differing sets of isolates.

SNP imputation is therefore generally considered necessary for GWAS\(^{46} \). We imputed missing base calls using two approaches, ClonalFrameML\(^{27} \) and Beagle\(^{48} \). Imputation using ClonalFrameML\(^{27} \) involves estimating the clonal frame by maximum likelihood\(^{48} \), then jointly reconstructing ancestral states and missing base calls by maximum likelihood utilizing the phylogeny reconstructed earlier\(^{48} \). To use Beagle, the mapped data were coded as haploid (one column per individual) and input as phased data\(^{40} \).

Testing imputation accuracy. To simulate data for testing imputation accuracy, 100 sequences were randomly sampled from each GWAS data set across the phylogeny. Maximum likelihood phylogenies were estimated for the 100 sequences of each species using RAxML\(^{47} \), as above. Any columns in the alignment corresponding to ambiguous positions in the reference genome were removed. One round of imputation was performed using ClonalFrameML to produce complete data sets with no ambiguous bases (Ns), which were then treated as the truth for the purpose of testing. The empirical distributions of Ns per site in the data sets of 100 sequences were determined, and these were sampled with replacement to reintroduce Ns to the variable sites in 100 simulated data sets. These sequences were then imputed again using ClonalFrameML and Beagle. Accuracy was summarized per site as a function of the frequency of Ns per site and the minor allele frequency. Overall, ClonalFrameML was more accurate than Beagle, so ClonalFrameML was used for all GWAS analyses (Supplementary Table 1).

Calculating association statistics before controlling population structure. We wished to compare the significance of associations before and after controlling for population structure. For the SNP and gene presence or absence data, an association between each SNP or gene and the phenotype was tested by logistic regression implemented in R. For the kmer analyses, an association between each SNP or gene and the presence or absence of each kmer was tested using a \( \chi^2 \) test implemented in C++. For each variant a \( P \) value was computed.

Correction for multiple testing. Multiple testing was accounted for by applying a Bonferroni correction\(^{47} \); the individual locus effect of a variant (SNP, gene or kmer) was considered significant if its \( P \) value was smaller than \( \alpha/n_{\text{var}} \), where we took \( \alpha = 0.05 \) to be the genome-wide false-positive rate and \( n_{\text{var}} \) to be the number of SNPs and genes, or kmers, with unique phylogenetic patterns, that is, unique partitions of individuals according to allele membership. Because the phenotypic contribution of multiple variants with identical phylogenetic patterns cannot be disentangled statistically, we found that pooling such variants improved the power by demanding methods\(^{48} \) employing Stempy\(^{48} \) to map reads to reference strains CPT073 (genbank accession no. AE014875.1), MGH 78778 (CP000064.1), H77177 (NC_000962.2) and MRSA252 (BX571856.1) for E. coli, K. pneumoniae, M. tuberculosis and S. aureus, respectively. The distributions of biallelic SNP frequencies are provided in Supplementary Table 4.
a less conservative Bonferroni correction than correcting for the total number of variants (Supplementary Fig. 10).

The genome-wide \(-\log_{10} P\) value threshold for SNPs and genes (or kmers) was 6.1 (7.3) for *S. aureus* ciprofloxacin, erythromycin, fusidic acid, gentamicin, penicillin, methicillin, tetracycline and rifampicin, 5.9 (6.7) for *S. aureus* trimethoprim, 6.5 (7.3) for all antimicrobials tested for *E. coli*, 6.6 (7.3) for all antimicrobials tested for *K. pneumoniae* and 5.0 (7.6) for all antimicrobials tested for *M. tuberculosis*. We also accounted for multiple testing of lineage effects by applying a Bonferroni correction for the number of principal components, which equals the sample size n.

Running GEMMA. For the analyses of SNPs, genes and kmers, we computed the relatedness matrix K from biallelic SNPs only. We tested for foreground effects at all biallelic, triallelic and tetraallelic SNPs, genes and kmers. GEMMA was run using a minor allele frequency of 0 to include all SNPs. GEMMA was modified to output the ML log-likelihood under the null, and alternative and \(-\log_{10} P\) values were calculated using R.

To perform LMM on tri- and tetra-allelic SNPs, each SNP was encoded as K = 1 binary columns corresponding to the first K – 1 alleles. For each column, an individual was encoded 1 if it contained that allele and 0 otherwise. The first column was input as the genotype, and the others as covariates into GEMMA. The log-likelihood under the alternative for each of the SNPs, was used to calculate the \(P\) value per SNP.

Due to the large number of kmers present within each data set, it was not feasible to run LMM on all kmers. We therefore applied the LMM to the top 200,000 most significant kmers from the logistic regression, plus 200,000 randomly selected kmers of those remaining. The randomly selected kmers were used to indicate whether some were becoming relatively more significant than the top 200,000, providing a warning in the case where large numbers of kmers became significant only after controlling for population structure.

Variant annotation. SNPs were annotated in R using the reference fasta and genbank files to determine SNP type (synonymous, non-synonymous, nonsense, read-through and intergenic), the codon and codon position, reference and non-reference amino acid, gene name and gene product.

Unlike the SNP approach, where we can easily refer to the reference genome to find what gene the SNP is in and the effect that it may have, annotation of the kmers is more difficult. We used BLASTSS to identify the kmers in databases of annotated sequences. Each kmer was first annotated against a BLAST database created of all refseq genomes of the relevant genus on NCBI. This enabled automatic annotation of all kmers that gave a small e-value against the genus-specific database. All kmers were also searched against the whole nucleotide NCBI database, first to compare and confirm the matches made against the first database and second to annotate the kmers that did not match anything in the within-genus database. Finally, when the resistance-determining mechanism was a SNP, the top 10,000 kmers were mapped to a relevant reference genome using Bowtie2 (ref. 63). This was used to determine whether the most significant kmers covered the position of the resistance-causing SNP or whether they were found elsewhere in the gene.

Genes were annotated for each CD-hit gene cluster by performing BLAST2 searches of each cluster sequence against a database of curated protein sequences downloaded from UNIPROTEK.

Testing power by simulating phenotypes. To assess the performance of the method for controlling population structure, we performed 100 simulations per species. In each simulation, a biallelic SNP was chosen randomly (from those SNPs with minor allele frequency above 20%) to be the causal SNP. Binary phenotypes (case or control) were then simulated for each genome with case probabilities of 0.25 and 0.5, allele frequency above 20%) to be the causal SNP. Binary phenotypes (case or control) were then simulated for each genome with case probabilities of 0.25 and 0.5.

References

Acknowledgements

The authors thank J.-B. Veyrieras, D. Charlesworth and B. Charlesworth for comments on the manuscript, X. Zhou and M. Stephens for helping adapt their software, S. Niemann for assisting with tuberculosis isolates and X. Didelot, D. Falush, R. Bowden, S. Myers, J. Marchini, J. Pickrell, P. Visscher, A. Price and P. Donnelly for discussions. This study was supported by the Oxford NIHR Biomedical Research Centre, a Mérieux Research Grant and the UKCRC Modernising Medical Microbiology Consortium, the latter funded under the UKCRC Translational Infection Research Initiative supported by the Medical Research Council, the Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the UK Department of Health (grant no. G0800778) and the Wellcome Trust (grant no. 087646/Z/08/Z). T.M.W. is an MRC research training fellow. C.C.A.S. was supported by a Wellcome Trust Career Development Fellowship (grant no. 097364/Z/11/Z). D.A.C. is funded by the Royal Academy of Engineering and an EPSRC Healthcare Technologies Challenge Award. T.E.P. and D.W.C. are NIHR Senior Investigators. G.M. is supported by a Wellcome Trust Investigator Award (grant no. 100956/Z/13/Z). D.J.W. and Z.I. are Sir Henry Dale Fellows, jointly funded by the Wellcome Trust and the Royal Society (grants nos. 101237/Z/13/Z and 102541/Z/13/Z).

Author contributions

S.G.E., C.-H.W., J.C. and D.J.W. designed the study, developed the methods, performed the analysis, interpreted the results and wrote the manuscript. Z.I. and D.A.C. assisted the manuscript, X. Zhou and M. Stephens for helping adapt their software, S. Niemann for assisting with tuberculosis isolates and X. Didelot, D. Falush, R. Bowden, S. Myers, J. Marchini, J. Pickrell, P. Visscher, A. Price and P. Donnelly for discussions. This study was supported by the Oxford NIHR Biomedical Research Centre, a Mérieux Research Grant and the UKCRC Modernising Medical Microbiology Consortium, the latter funded under the UKCRC Translational Infection Research Initiative supported by the Medical Research Council, the Biotechnology and Biological Sciences Research Council and the National Institute for Health Research on behalf of the UK Department of Health (grant no. G0800778) and the Wellcome Trust (grant no. 087646/Z/08/Z). T.M.W. is an MRC research training fellow. C.C.A.S. was supported by a Wellcome Trust Career Development Fellowship (grant no. 097364/Z/11/Z). D.A.C. is funded by the Royal Academy of Engineering and an EPSRC Healthcare Technologies Challenge Award. T.E.P. and D.W.C. are NIHR Senior Investigators. G.M. is supported by a Wellcome Trust Investigator Award (grant no. 100956/Z/13/Z). D.J.W. and Z.I. are Sir Henry Dale Fellows, jointly funded by the Wellcome Trust and the Royal Society (grants nos. 101237/Z/13/Z and 102541/Z/13/Z).

Additional information

Supplementary information is available online. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to D.J.W.

Competing interests

The authors declare no competing financial interests.