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

Sarah G. Earle1, Chieh-Hsi Wu1, Jane Charlesworth1, Nicole Stoesser1, N. Claire Gordon1, Timothy M. Walker1, Chris C. A. Spencer2, Zamin Iqbal2, David A. Clifton3, Katie L. Hopkins4, Neil Woodford4, E. Grace Smith5, Nazir Ismail6,7, Martin J. Llewelyn8, Tim E. Peto1, Derrick W. Crook1, Gill McVean2,9, A. Sarah Walker1 and Daniel J. Wilson1,2*

Bacteria pose unique challenges for genome-wide association studies because of strong structuring into distinct strains and substantial linkage disequilibrium across the genome. Although methods developed for human studies can correct for strain structure, this risks considerable loss-of-power because genetic differences between strains often contribute substantial phenotypic variability. 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 recombinating 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 porin 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 disease. Hospital-associated infections including Staphylococcus aureus, Escherichia coli and Klebsiella pneumoniae represent a serious threat to the safe provision of healthcare, while the Mycobacterium tuberculosis pandemic remains a major global health challenge. Treatment options continue to be eroded by the spread of antimicrobial resistance, with some strains resistant even to antimicrobials of last resort.

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 approaches. 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 selection. Approaches to controlling for this population structure have allowed for systematic phenotypic differences based on cluster membership, or, in clonal species, phylogenetic history. However, these and other approaches common in human GWASs 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 approaches and enjoy greater replicability than phylogenetic methods because recombination is evident in most bacteria. 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 sampling. In the four species we investigated, we observed genome-wide linkage disequilibrium and strong population structure, with the first ten principal components (PCs) 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 both. 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 ($P=10^{-122}$).

1Nuffield Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DU, UK. 2Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. 3Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7QD, UK. 4Antimicrobial Resistance and Healthcare Associated Infections Reference Unit, Public Health England, London NW9 5EQ, UK. 5Public Health England, West Midlands Public Health Laboratory, Heartlands Hospital, Birmingham B9 5SS, UK. 6Centre for Tuberculosis, National Institute for Communicable Diseases, Johannesburg 2131 South Africa. 7Department of Medical Microbiology, University of Pretoria, Pretoria, South Africa. 8Department of Infectious Diseases and Microbiology, Royal Sussex County Hospital, Brighton BN2 5BE, UK. 9The Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7FZ, UK. *These authors contributed equally to this work. *e-mail: daniel.wilson@ndm.ox.ac.uk
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.
variants is high (Fig. 2a and Supplementary Fig. 4a). Being greatest when the sample size is low and the number of principal tested for an effect on the phenotype. Because principal components are commonly used to control for relationship between genealogical history and principal component, each principal component estimated is the difference between the mean and the observed value of the phenotype. Thus, identifying loci contributing to the most significant associations caused by correlations between lineages.

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 reduces the power to a level below the theoretical optimum power for a single locus. The Wald test improves the power several-fold by detecting lineage-specific effects. The 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. Because LMM improves precision in the presence of polygenic effects, we thus decomposed the random effects estimated by controlling for population structure and multiple testing (Fig. 1b and Supplementary Fig. 3). Kmers capturing resistance-conferring substitutions in fusA, which encodes EF-G, were to greater significance, because these low-frequency variants were unstratified and LMM improves power in the presence of polygenic effects (P = 10^{-11} by \chi^2 test, P = 10^{-15} by LMM). However, fusA variants explain only half as much resistance as fusC overall.

However, fusC-encoded resistance was observed exclusively within strains ST-1 and ST-8. Thus, controlling for population structure using LMM 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 to greater significance, because these low-frequency variants were unstratified and LMM improves power in the presence of polygenic effects (P = 10^{-11} by \chi^2 test, P = 10^{-15} by LMM). However, fusA variants explain only half as much resistance as fusC.

Methods to limit loss of power such as ‘leave-one-chromosome-out’ 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 regression, without sacrificing the power to detect locus-specific associations when possible.

We observed that leading principal components tend to correspond to major lineages in bacterial genealogies (or ‘clonal frames’) despite substantial differences in recombination rates (Fig. 1b and Supplementary Fig. 5), reflecting an underlying relationship between genealogical history and principal component analysis. Principal components are commonly used to control for population structure by including leading principal components as fixed effects in a regression. 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. 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 test to assess the significance of the association between each lineage and the phenotype.

Our method, implemented in the R package bugwars, revealed strong signals of association between fusC and fusC variants in linkage disequilibrium with fusC accounted for the strongest signals within PC-6 and PC-9 (P = 10^{-70}), comparable in significance to the low-frequency variants at fusC (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 staphylococcal 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
Table 1 | Number of resistant and sensitive isolates by species and antibiotics, known mechanisms of resistance and main results.

| Antibiotic   | R   | S   | Resistance mechanism | SNP/gene rank | SNP/gene LMM rank | Kmer rank | Kmer LMM rank |
|--------------|-----|-----|----------------------|---------------|-------------------|-----------|---------------|
| **E. coli**  |     |     |                      |               |                   |           |               |
| Ampicillin   | 189 | 52  | β-lactamase genes    | Gene presence | 1                 | 6 (tnp)*   | 6 (tnp)*  
| Cefazolin    | 139 | 102 | β-lactamase genes    | Gene presence | 2 (nmpC)**       | 121,710 (nmpC)** | 3,690 (nmpC)**  
| Cefuroxime   | 81  | 160 | β-lactamase genes    | Gene presence | 1                 | 1,598 (162-192 upstream blacMX-M)* | 470 (162-192 upstream blacMX-M)*  
| Ceftriaxone  | 55  | 186 | β-lactamase genes    | Gene presence | 1                 | 1,403 (tnp)*  
| Ciprofloxacin| 91  | 150 | SNPs in gyrA, gyrB, parC or parE or presence of PMQR (qraA,B,M, qraA,B,M)# | Gene presence or SNPs, or both | 1#  
| Gentamicin   | 48  | 193 | aac (aac(3)-I), ant, aph or rRNA methylase | Gene presence | 1                 | 1         | 1  
| Tobramycin   | 67  | 174 | aac (aac(3)-I), ant or rRNA methylase | Gene presence | 1                 | 1         | 1  
| **K. pneumonia** |     |     |                      |               |                   |           |               |
| Cefazolin    | 53  | 123 | β-lactamase genes    | Gene presence | 1 + HP + wbcC   | 762 (tnp)* | 837 (tnp)*  
| Cefuroxime   | 46  | 130 | β-lactamase genes    | Gene presence | 1 + HP + wbcC 1  
| Ceftriaxone  | 35  | 141 | β-lactamase genes    | Gene presence | 1 + HP + wbcC 1+ HP + wbcC | 762 (tnp)* | 1,480 (tnp)*  
| Ciprofloxacin| 34  | 142 | SNPs in gyrA, gyrB, parC or parE or presence of PMQR (qraA,B,M, qraA,B,M)# | Gene presence or SNPs, or both | 2# (tnp)*  
| Gentamicin   | 31  | 145 | aac (aac(3)-I), ant, aph or rRNA methylase | Gene presence | 1                 | 1         | 1  
| Tobramycin   | 36  | 140 | aac (aac(3)-I), ant or rRNA methylase | Gene presence | 1                 | 1         | 1  
| **M. tuberculosis** |     |     |                      |               |                   |           |               |
| Ethambutol   | 41  | 1589| embB                 | SNPs          | 2 (rpoB)**      | 1          | 1  
| Isoniazid    | 239 | 1,470| katG, fabG1          | SNPs          | 1                 | 1         | 1  
| Pyrazinamide | 45  | 1,662| pncA                 | SNPs          | 142 (rpoB)**     | 126 (rpoB)** | 1  
| Rifampicin   | 86  | 1,487| rpoB                 | SNPs          | 1                 | 1         | 1  
| S. aureus    |     |     |                      |               |                   |           |               |
| Ciprofloxacin| 242 | 750 | gyrA or gyrA         | SNPs          | 1                 | 1         | 1  
| Erythromycin | 216 | 776 | ermA, ermC, ermT or msrA | Gene presence | 1                 | 1         | 1  
| Fusidic acid | 84  | 908 | SNPs in fusA or presence of fusA or fusC | Gene presence or SNPs, or both | 4# (SAS0037)* | 1# | 75# (SAS0040)* | 1# |
| Gentamicin   | 11  | 981 | aacA/aphD            | Gene presence | 1 + GNAT acetyltransferase | 1 + GNAT acetyltransferase | 1 + 415 bases upstream to 100 bases downstream | 1 + 415 bases upstream to 100 bases downstream | 2 (blaI)* | 1 + SCCmec genes | 1 + SCCmec genes | 1 + plasmid genes | 1 + plasmid genes |
| Penicillin   | 824 | 168 | blaZ                 | Gene presence | 1                 | 1          | 1  
| Methicillin  | 216 | 776 | mecA                 | Gene presence | 1 + mecR1 | 2 (blaI)*  
| Tetracycline | 46  | 946 | tetK, tetL or tetM   | Gene presence | 2 (repC)*  
| Trimethoprim | 15  | 308 | SNPs in dfrB, presence of dfrG or dfrA | Gene presence or SNPs, or both | 1         | 1         | 1  
| Rifampicin   | 8   | 984 | SNPs                 | Gene presence | 1                 | 1         | 1  

For each antibiotic, the most significant variant was the expected mechanism, unless indicated by *(most significant variant was in physical linkage (PL) with the expected mechanism) or **(most significant variant was not the expected mechanism or in 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 that was most significant when it was not causal. Where more than one gene or mechanism causes resistance, the variant we found is underlined, or referred to by * or **. R, resistant; S, sensitive; HP, hypothetical protein; tnp, transposase; PMQR, plasmid mediated quinoline resistance. See Supplementary Tables 3-6 for more detail.
generally organized into physically linked blocks along the chromo-
some (Fig. 2b and Supplementary Fig. 4b), underlining the impor-
tance of recovering power by interpreting linkage effects.

We noted a trade-off in interpreting lineage effects, because they
are susceptible to confounding with population-stratified differ-
ences in environment or sampling (Supplementary Fig. 7).

Confronted with a strong population structure and genome-wide
linkage disequilibrium in bacteria, we wished to test empirically
lineage-level differences.

In one study, demonstrating that the high accuracy attained in predict-
ing antimicrobial resistance phenotypes from genotypes known
(Supplementary Fig. 9b).

For accessory genes such as
loci or regions in physical linkage with those loci for antimicrobial
GWASs across species was very good, identifying genuine causal

-loci, which led to spurious associations as the top hit before controlling for

differences. 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 confound-
ing with population-stratified differences in environment or
sampling. This trade-off between power and robustness underlines
the importance of functional validation for bacterial GWASs

\( \beta \) and Supplementary Table 3). For accessory genes such as
resistance in 25/26 cases for the SNP and gene approach and the
resistance (Supplementary Fig. 8).

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 \( M. \) \( \text{tuberculosis} \),
which led to spurious associations as the top hit before controlling for
controlling for population structure between ethambutol and pyrazinamide resist-
ance and SNPs in rifampicin resistance-conferring \( rpoB \). Even after
controlling for population structure, these associations remained
genome-wide significant at \( P = 10^{-45} \) and \( P = 10^{-34} \).

Antimicrobial resistance has arisen over 20 times per drug in the
\( M. \) \( \text{tuberculosis} \) tree, through frequent convergent evolution
(Supplementary Fig. 4c and Supplementary Fig. 8). Within a
single gene, such as \( rpoB \), 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 signifi-
cant hit identified by GWAS matched a known causal variant\(^{26-38} \)
(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 \( \beta \)-lacta-
mases, 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 predict-
ing antimicrobial resistance phenotypes from genotypes known
from the literature\(^{37,39} \) 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 recombi-
nation breaks down linkage disequilibrium, assisting mapping
(Fig. 2c and Supplementary Fig. 4c).

For one drug, cefazolin, in \( E. \) \( \text{coli} \), we identified a variation in
the presence of an unexpected gene as the most strongly associated
with resistance, \( rmpC \) (\( P = 10^{-12.5} \)). This gene encodes an outer mem-
brane porin over-represented in susceptible individuals. Permeability
in the \( \text{Salmonella typhimurium} \) homologue mediates resistance to
other cephalosporin \( \beta \)-lactams,\(^{40} \), making this a strong candidate for
a novel resistance-conferring mechanism discovered in \( E. \) \( \text{coli} \).

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
linkage effects is valuable, because characterizing phenotypic vari-
ability 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 statisti-
cally. 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 confound-
ing with population-stratified differences in environment or
sampling. This trade-off between power and robustness underlines
the importance of functional validation for bacterial GWASs

### Methods

#### Linear mixed model.

In the LMM,\(^{41-45} \), 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,

\[
y_i = w_i \alpha_i + \ldots + w_i \alpha_i + X_0 \beta_i + X_1 \gamma_1 + \ldots + X_l \gamma_l + \epsilon_i
\]

where there are \( n \) individuals, \( c \) covariates, \( l \) loci, \( j \) is the foreground locus, \( y_i \) is the
phenotype in individual \( i \), \( W_i \) is covariate \( j \) in individual \( i \), \( \gamma_j \) is the effect of covariate
\( j \), \( X_{0,j} \) is the genotype of locus \( j \) in individual \( i \), \( \beta_i \) is the foreground effect of locus \( j \), \( y_i \) is the
background effect of locus \( j \) and \( \epsilon_i \) is the effect of the environment (or error) on
individual \( i \). Biallelic genotypes are numerically encoded as \( -1/1 \) (common allele) or
\( -1/1 \) (rare allele), where \( f \) 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
triatlelic and tetralelic loci are rare, we use only biallelic loci to model background
effects. When the foreground locus is triallelic (\( K = 3 \)) or tetralelic (\( K = 4 \)), the
genotype in individual \( i \) is encoded as a vector indicating the presence (1) or
absence (0) of the first (\( K = 1 \) alleles and \( \beta_i \) becomes a vector of length (\( K – 1 \).

Treating the background effects of the loci as random effects means the precise values of coefficients \( \gamma_j \) are averaged. The \( \gamma_j \) are assumed to follow independent
normal distributions with common mean \( 0 \) and variance \( \tau^{-1} \) 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 \( \tau^{-1} \). The model can be
rewritten in matrix form as

\[
y = Wa + X \beta_i + u + \epsilon
\]

with

\[
u = X_1 \gamma_1 + \ldots + X_1 \gamma_l
\]

\[
\mu \sim \text{MVN}(0, \lambda \tau^{-1}K)
\]

\[
\epsilon \sim \text{MVN}(0, \tau^{-1}I)
\]

where \( u \) represents the cumulative background effects of the loci, \( \text{MVN} \) denotates
the multivariate normal distribution, \( I \) is an \( n \times n \) identity matrix, and \( K \) is an \( n \times n \)
relatedness matrix defined as \( K = XX' \), which captures the genetic covariance
between individuals.
Testing for locus effects. To assess the significance of the effect of an individual locus \( l \) on the phenotype, controlling for population structure and background genetic effects, the parameters of the linear mixed model \( \alpha_l, \alpha_r, \beta_l, \lambda \) 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 \( \beta_l = 0 \) using the software GEMMA\(^{35}\).

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 \( R \) function `prcomp()`, producing an \( L \) by \( n \) loading matrix \( D \) and an \( n \) by \( n \) score matrix \( T \) where \( T = XD \). \( D_n \) records the contribution of biallelic SNP \( n \) to the definition of principal component \( j \), while \( T_{ij} \) represents the projection of individual \( i \) onto 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 \( \lambda \tau^2 \) tends to cause them to be small in magnitude and not significantly different from zero. However, cumulatively, the background locus effects can capture systematic phenotypic differences between lineages. We therefore recovered the post-data distribution (equivalent to an empirical Bayesian 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 (\( \beta_l = 0 \)). We therefore calculated the mean and variance–covariance matrix of the multivariate normal posterior distribution of \( \gamma \) in the LMM null model. These are equivalent to those of a ridge regression\(^{36}\) and were computed as

\[
\mu = (XX' + \lambda LL')^{-1}X'y \quad \text{and} \quad \Sigma = \lambda \tau^2 (XX' + \lambda LL')^{-1}
\]

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 components.

\[
u = X_1\gamma_1 + \ldots + X_n\gamma_n \quad \text{where} \quad \gamma = D^{-1}y \quad \text{and} \quad g_n \text{ is the background effect of principal component } j \text{ on the phenotype. We computed the mean and variance of the post-data distribution of } g_n \text{ as } m = D^{-1} \mu \quad \text{and} \quad S = D^{-1} \Sigma D^{-1}, \text{ respectively, using the affine transformation for a multivariate normal distribution. To test the null hypothesis of no background effect of principal component } j \text{ (that is, } g_n = 0 \text{), we used a Wald test with test statistic}
\]

\[
\chi^2 = \frac{(\gamma - \bar{\gamma})'(V_n^{-1}/n)(\gamma - \bar{\gamma})}{\text{variance distribution with degrees of freedom equal to the number of bins minus one to obtain a } P \text{ value.}}
\]

Although we identified and tested for lineage effects in the LMM setting, lineage effects could also be identified and tested for by interpreting the coefficients of leading principal components or genetic cluster membership included as fixed effects in a regression, both of which represent alternative methods for controlling for population structure.

Identifying non-genome-wide principal components. Some principal components capture variation localized to particular areas of the genome. We identified non-genome-wide principal components by testing for spatial heterogeneity of the loading matrix \( W \) for biallelic SNPs across the genome. SNPs were grouped into \( 20 \) contiguous bins (indexed by \( j \)) of nearly equal size \( N_c \) and the mean \( O_c \) and variance \( V_c \) in the absolute value of the SNP loadings for principal component \( i \) in bin \( j \) were calculated, as well as the mean absolute value \( E_i \) of the SNP loadings for principal component \( i \) across all SNPs. The null hypothesis of no heterogeneity was assessed by comparing the test statistic \( \chi^2 = \sum (O_c - E_i)^2/V_c/N_c \) to a \( \chi^2 \) distribution with degrees of freedom equal to the number of bins minus one to obtain a \( P \) value.

Antimicrobial resistance testing, genome sequencing and SNP calling. We investigated \( 241 \) \( E. \) coli and \( 176 \) \( K. \) pneumoniae UK clinical isolates newly reported here, together with \( 992 \) \( S. \) aureus and \( 1,735 \) \( M. \) tuberculosis isolates reported previously\(^{12-15}\). All isolates were tested for resistance to multiple antimicrobials based on drug susceptibility clinical laboratory protocols, and DNA was extracted and sequenced on Illumina platforms as previously described\(^{30-34}\) and \( \text{SNP calling.} \)
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, trisomic and tetrallelic 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, which 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 BLAST 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 kmer 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 BLAST searches of each cluster sequence against a database of curated protein sequences downloaded from UNIPROT.

**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, respectively, in individuals with the common and rare allele at the causal SNP (analogous to both control and case samples). In order to control for stratiﬁcation, 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, respectively, in individuals with the common and rare allele at the causal SNP (analogous to both control and case samples).

**Accession codes.** All genomes were deposited in NCBI and EBI short read archives under BioProject accession nos. PRJNA306133 (*E. coli*) and PRJNA308279 (*M. tuberculosis*) and PRJNA308283 (*S. aureus*). Individual BioSample accession numbers and antimicrobial resistance phenotypes are detailed in Supplementary Data 1.

**References.**

1. Feil, E. J. & Spratt, B. G. Recombination and the structures of bacterial pathogens. *Ann. Rev. Microbiol.* 55, 561–590 (2001).
2. Falush, D. & Bowden, R. Genome-wide association mapping in bacteria? *Trends Microbiol.* 14, 353–355 (2006).
3. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Rev. Genet.* 10, 681–690 (2009).
4. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24 (2012).
5. Coroed, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Rev. Microbiol.* 12, 263–273 (2014).
6. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* 95, 6578–6583 (1998).
7. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth’s biogeochemical cycles. *Science* 320, 1034–1039 (2008).
8. World Health Organization. The Global Burden of Disease: 2004 Update (2008); http://www.who.int/healthinfo/global_burden_disease.
9. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* 74, 417–433 (2010).
10. European Centre for Disease Prevention and Control. Surveillance of Surgical- Site Infections in Europe, 2009–2009 (2012); http://ecdc.europa.eu/en/publications/Publications/120215_SUR_SSL_2008-2009.pdf.
11. World Health Organization. Global Tuberculosis Report 2014 (2014); http://apps.who.int/iris/bitstream/10665/173094/1/9789241546409_eng.pdf.
12. World Health Organization. Antimicrobial Resistance: A Global Report on Surveillance (2014); http://apps.who.int/iris/bitstream/10665/112642/1/9789241564748_eng.pdf.
13. Sheppard, S. K. et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in *Campylobacter*. *Proc. Natl Acad. Sci. USA* 110, 11923–11927 (2013).
14. Alam, M. T. et al. Dissecting vancomycin-intermediate resistance in *Staphylococcus aureus* using genome-wide association. *Genome Biol. Evol.* 6, 1174–1185 (2014).
15. Laabei, M. et al. Predicting the virulence of MRSA from its genome sequence. *Genome Res.* 24, 839–849 (2014).
16. Chewaprolcha, C. et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet.* 10, e1004547 (2014).
17. Salipante, S. J. et al. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res.* 25, 119–128 (2014).
18. Read, T. D. & Massey, R. C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med.* 6, 109 (2014).
19. Fahat, M. R., Shapiro, B. J., Sheppard, S. K., Colijn, C. & Murray, M. A phylogeny-based sampling strategy and power calculator informs genome-wide association studies design for microbial pathogens. *Genome Med.* 6, 101 (2014).
20. Hall, B. G. SNP-associations and phenotype predictions from hundreds of microbial genomes without genome alignment. *PLoS ONE* 9, e90490 (2014).
21. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* 25, 17–24 (2015).
22. Holt, K. E. et al. Genome analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl Acad. Sci. USA* 112, E3574–E3581 (2015).
23. Rice, A. L., Zatlken, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Rev. Genet.* 11, 459–463 (2010).
24. Perez-Losada, M. et al. Population genetics of microbial pathogens estimated from multifocus sequence typing (MLST) data. *Infect. Genet. Evol.* 6, 97–111 (2006).
25. Vas, M. & Didelez, X. A comparison of homologous recombination rates in bacteria and archaea. *IMBE* J. 3, 199–208 (2009).
26. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* 38, 904–909 (2006).
27. O’Neill et al. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli*. *Genome Med.* 51, 1737–1740 (2007).
28. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genet.* 44, 821–824 (2012).
29. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. Nature Genet. 46, 100–106 (2014).
30. Graeven, A. The phylogenetic regression. Phil. Trans. R. Soc. Lond. B 326, 119–157 (1989).
31. Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. Am. Nat. 149, 646–667 (1997).
32. Millikan, R. & Bridges, M. M. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. Genetics 126, 505–517 (1990).
33. McVean, G. A genealogical interpretation of principal components analysis. Nature Genetics 52, 1681–1685 (2010).
34. McVean, G. A genealogical interpretation of principal components analysis. PLoS Genet. 5, e1000686 (2009).
35. Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. 54, 426–482 (1943).
36. Walker, T. M. et al. Whole genome sequencing for prediction of Mycobacterium tuberculosis drug susceptibility and resistance: a retrospective cohort study. Lancet Infect. Dis. 15, 1193–1202 (2015).
37. Wong, O. et al. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. J. Clin. Microbiol. 52, 1182–1191 (2014).
38. Stoeßner, N. et al. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genome sequence data. J. Antimicrob. Chemother. 68, 2234–2244 (2013).
39. Bradley, P. et al. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nature Commun. 6, 10063 (2015).
40. Sun, S., Berg, O. G., Roth, J. R. & Andersson, D. I. Contribution of gene conversion to evolution of increased antibiotic resistance in Salmonella typhimurium. Genetics 182, 1183–1195 (2009).
41. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genet. 38, 203–208 (2006).
42. Kang, H. M. et al. Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723 (2008).
43. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genet. 42, 348–354 (2010).
44. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. Nature Methods 8, 833–835 (2011).
45. Listgarten, J. et al. Improved linear mixed models for genome-wide association studies. Nature Methods 9, 525–526 (2012).
46. O’Hagan, A. & Forster, J. in Kendall’s Advanced Theory of Statistics Volume 2B Bayesian Inference 2nd edn, Ch. 11 (Wiley-Blackwell, 2010).
47. Eyre, D. W. et al. A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open 2, e001124 (2012).
48. Everitt, R. G. et al. Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. Nature Commun. 5, 3956 (2014).
49. Lunter, G. & Goodson, M. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Res. 21, 936–939 (2011).
50. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829 (2008).
51. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119 (2010).
52. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659 (2006).
53. Rzik, G., Lavenier, D. & Chikhi, R. DSK: k-mer counting with very low memory usage. Bioinformatics 29, 652–653 (2013).
54. Bolger, A. M., Lohse, M. & Usadel, B. Trimomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).
56. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084–1097 (2007).
57. Dinkelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput. Biol. 11, e1004041 (2015).
58. Hedge, J. & Wilson, D. J. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. mBio 5, e02158–14 (2014).
59. Pupko, T., Pe'er, I., Shamir, R. & Graur, D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol. Biol. Evol. 17, 890–896 (2000).
60. Yahara, K., Didelot, X., Ansari, M., Sheppard, S. K. & Falush, D. Efficient inference of recombination hot regions in bacterial genomes. Mol. Biol. Evol. 31, 1593–1605 (2014).
61. Dunn, O. J. Estimation of the medians for dependent variables. Ann. Math. Stat. 30, 192–197 (1959).
62. Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 431 (2009).
63. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357–359 (2012).
64. UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–D212 (2015).

Acknowledgements
The authors thank J.-B. Veyrerès, D. Charlesworth and B. Charlesworth for comments on the manuscript, X. Zhou and M. Stephens for helping adapt their software, S. Nienmann 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. G0800787) 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.I.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.I.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 analysis and commented on the manuscript. N.S., N.C.G., T.E.P. and D.W.C. designed and implemented isolate collection, drug susceptibility testing and whole-genome sequencing, and assisted with interpretation. C.C.A.S., G.M. and A.S.W. assisted with methods development and writing of the manuscript.

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.I.W.

Competing interests
The authors declare no competing financial interests.