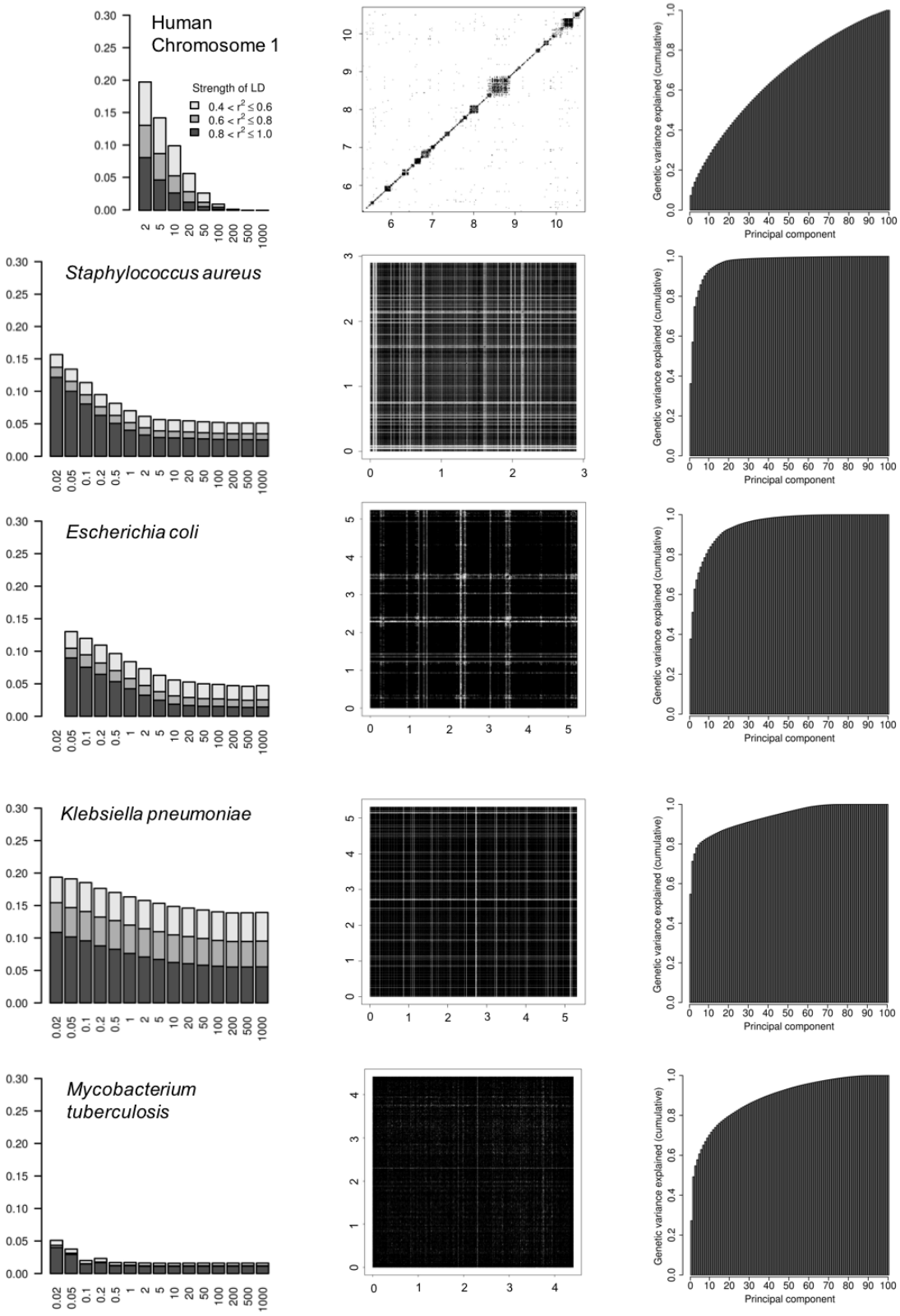
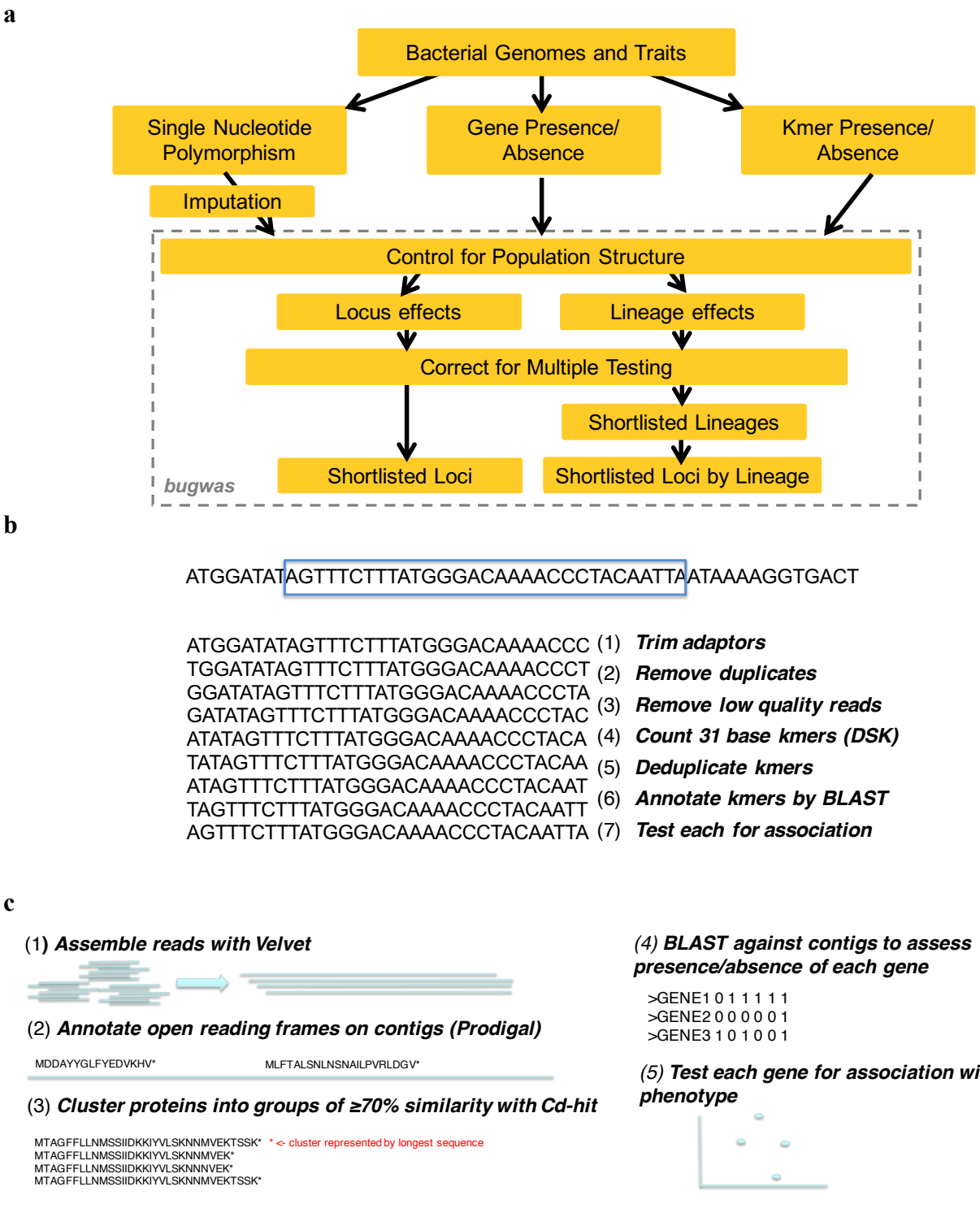
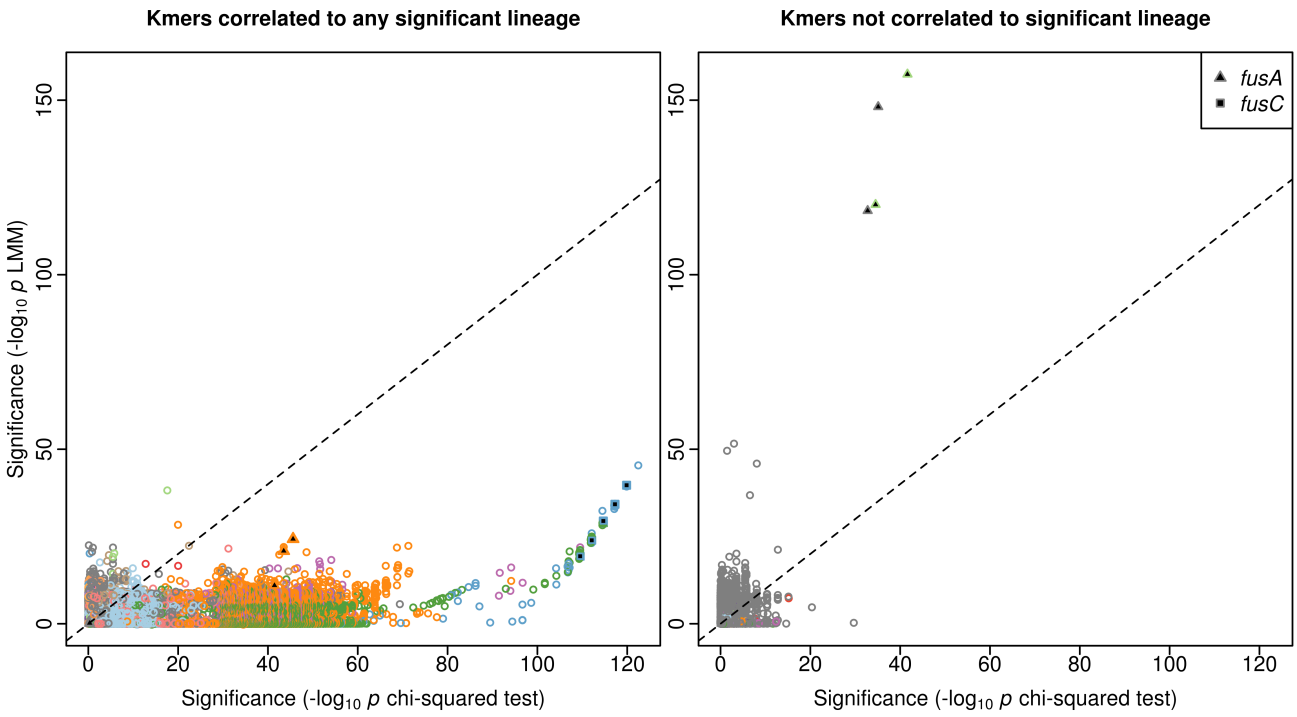
Supplementary Information
Earle, Wu, Charlesworth, *et al* (2016)
Identifying lineage effects when controlling for population structure improves power in bacterial association studies



Supplementary Figure 1. Linkage disequilibrium and population structure in humans vs bacteria is strikingly different. Left panels: LD (r^2) decays rapidly with physical distance (in kb) in bacteria, as in humans, but plateaus to residual levels resulting in genome-wide LD. Middle panels: Points show the position of sites (in Mb) in high LD ($r^2 > 0.7$). In contrast to the block-like structure of LD in humans, in bacteria the genome comprises one large LD block because of strong population structure and limited homologous recombination. Right panels: the cumulative proportion of genetic variability explained by leading principal components.



Supplementary Figure 2. (**a**) Overview of our approach. We identify significant loci and lineages, and the most significant loci within significant lineages. (**b**) Kmer analysis of hard-to-reach diversity. Some diversity (e.g. indels, repeats) is difficult to capture using standard variant calling tools. We directly analyse presence/absence of short haplotypes (*kmers*) to make sure we don't miss any associations. A kmer is a sliding window (in our case 31 bases long) of contiguous sequence. (**c**) Capturing the accessory genome. Differential presence or absence of genes or entire mobile elements is an important source of diversity in bacterial genomes. We test for associations with gene presence/absence by defining the accessory genome using Cd-hit and profiling each bug using BLAST.

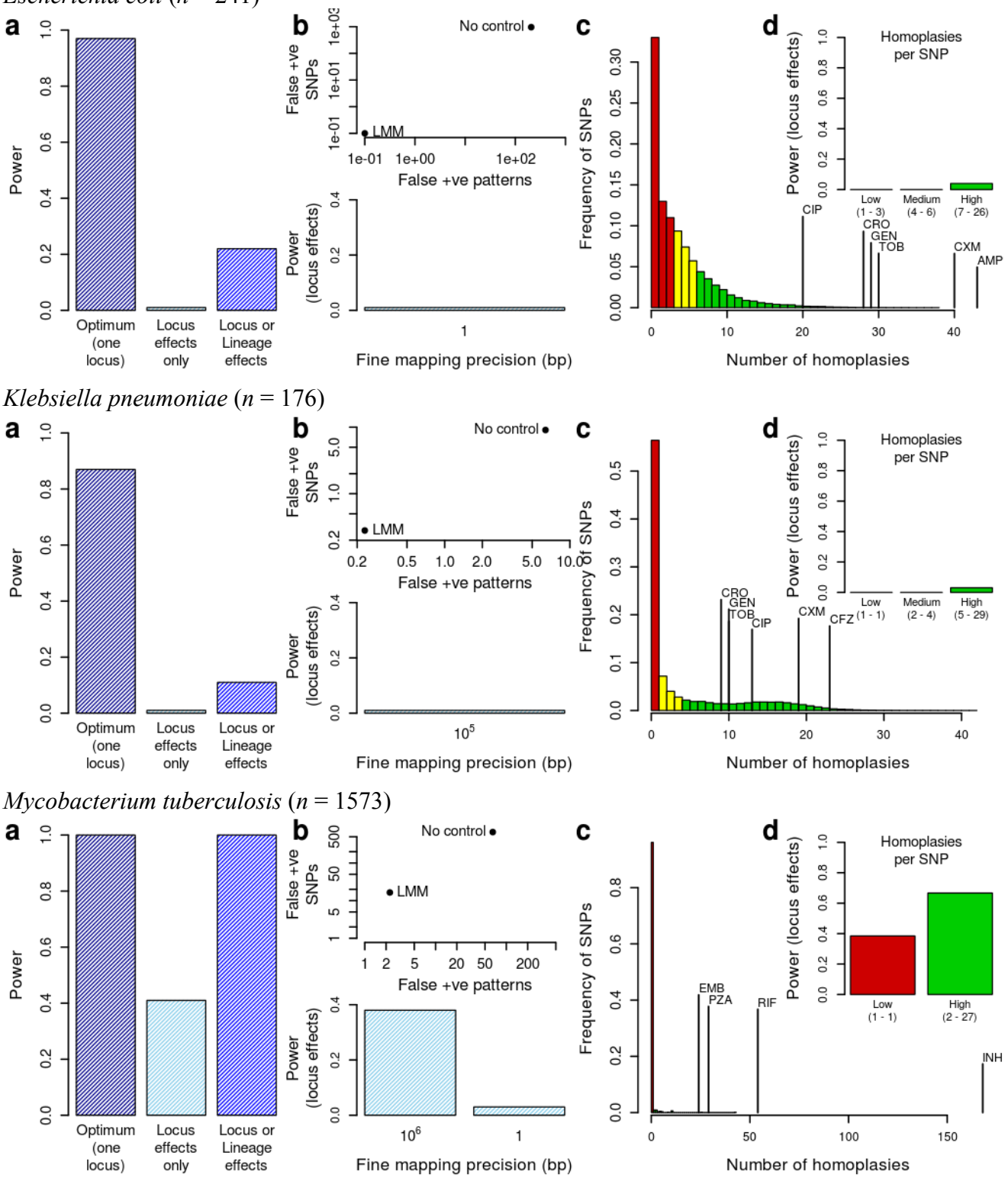


Supplementary Figure 3. Differential effect of controlling for population structure using LMM on significance of the association between fusidic acid resistance and kmers that are correlated vs. uncorrelated with any significant lineage in *S. aureus*. Kmers that are population-stratified by any lineage suffer greater loss of power when controlling for population structure. This can be visualized in the specific example of fusidic acid resistance by noting that kmers significantly associated with any lineage (left) suffered greater loss of significance than other kmers (right) after controlling for population structure using LMM (vertical axes) compared to before controlling for population structure (horizontal axes). 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 PC to which it is most strongly correlated, grey if it is not most strongly correlated to one of the 20 most significant PCs. Left: kmers with absolute correlation > 0.25 to any PC showing significant association to fusidic acid resistance. Right: kmers with no absolute correlation > 0.25 to any PC showing significant association to fusidic acid resistance.

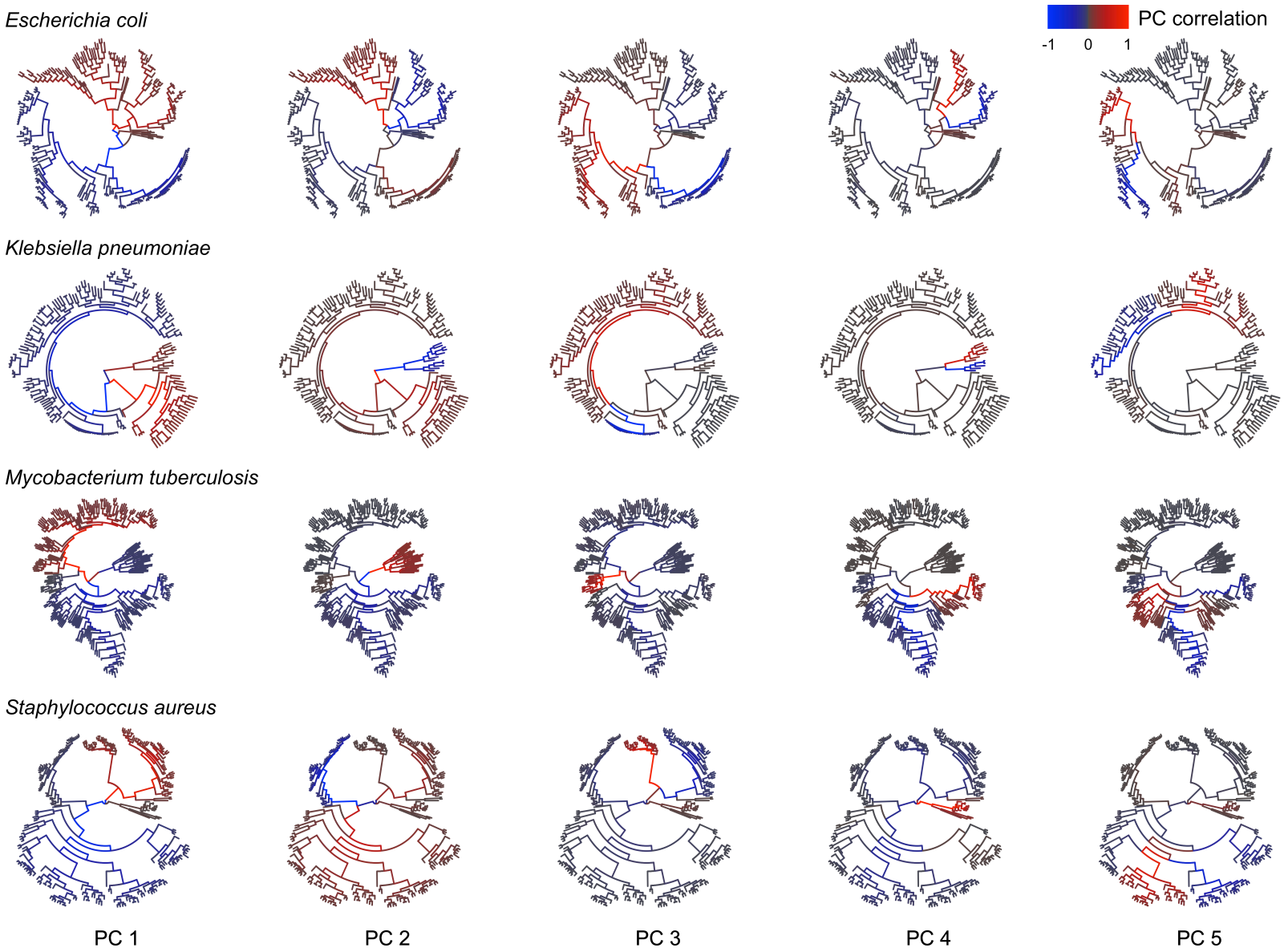
Escherichia coli (*n* = 241)

Klebsiella pneumoniae (*n* = 176)

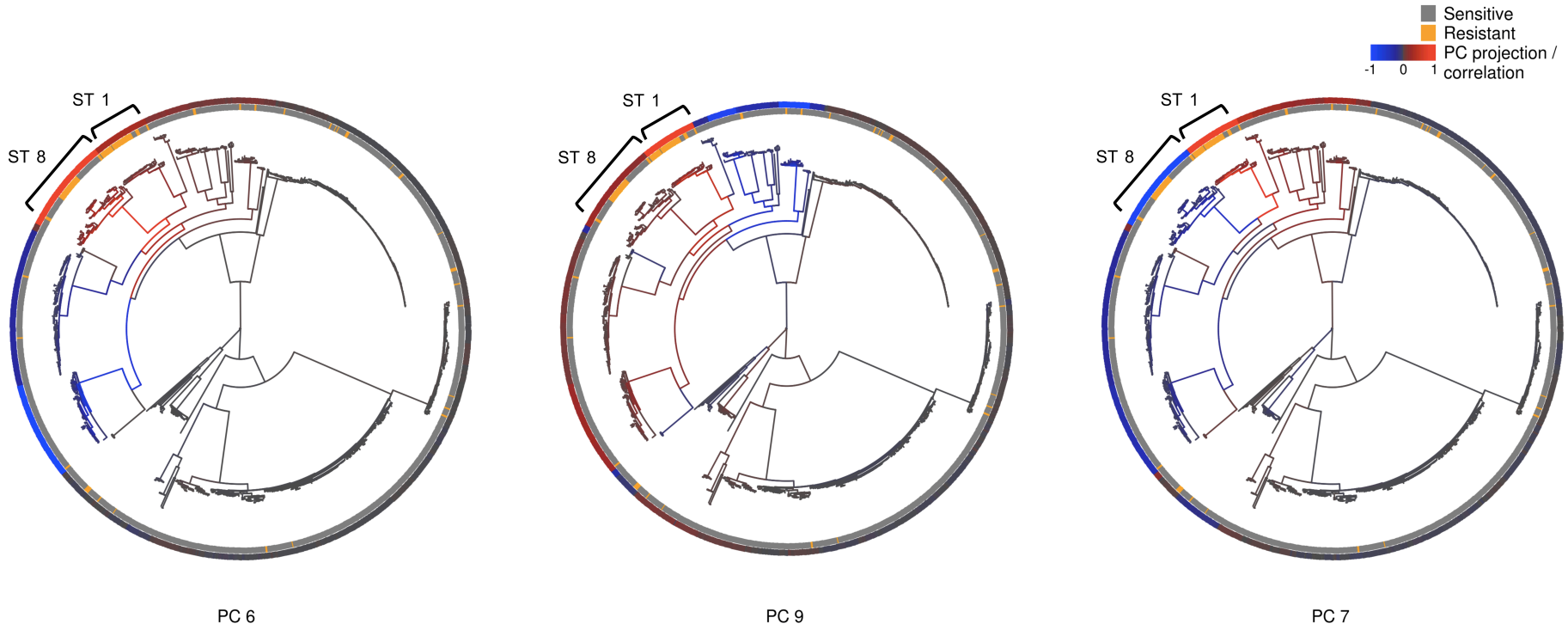
Mycobacterium tuberculosis (*n* = 1573)



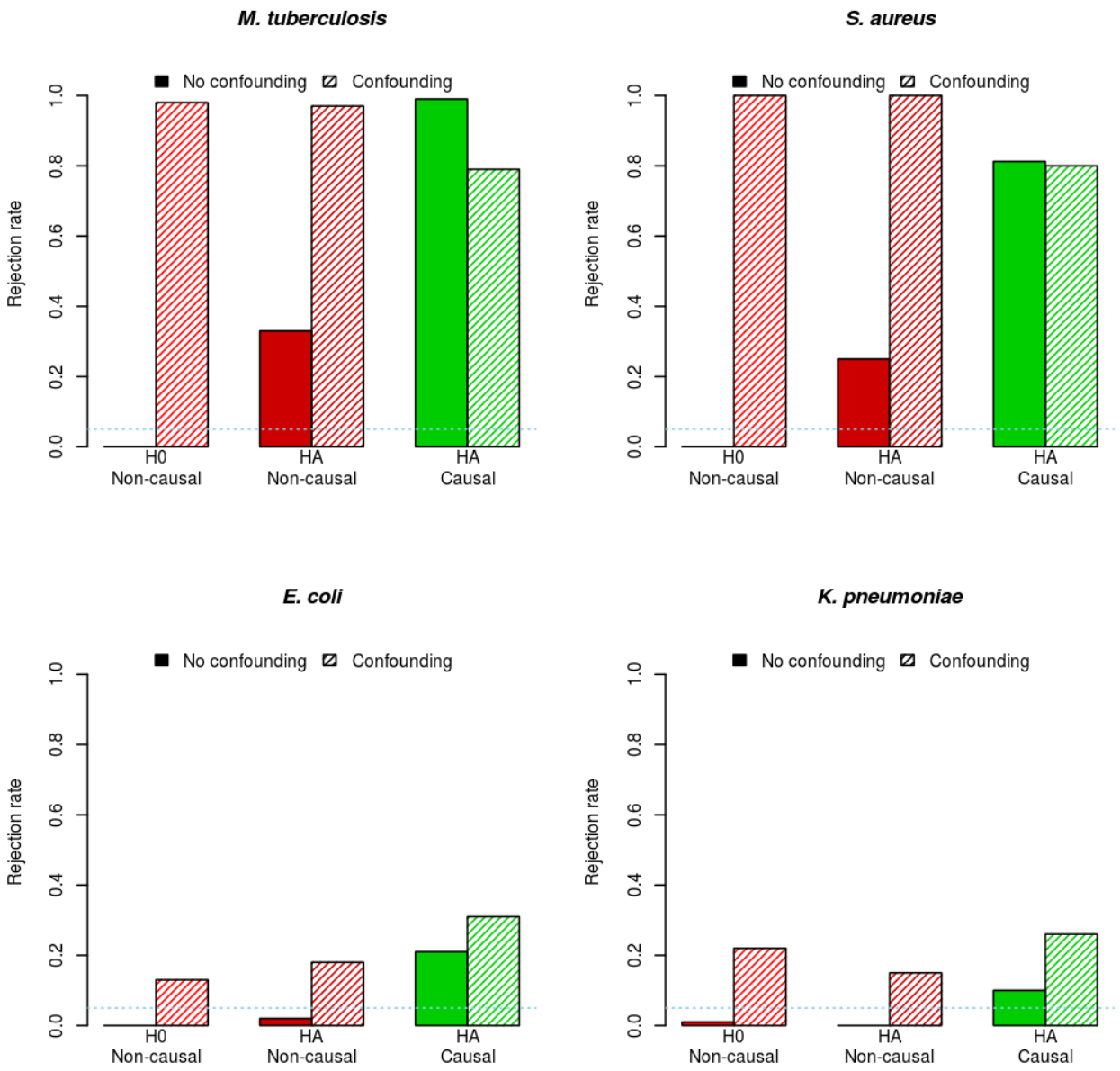
Supplementary Figure 4. Power, false positives, fine mapping and homoplasy in *E. coli, K.pneumoniae* and *M. tuberculosis*. Simulation results. (**a**) Controlling for population structure and multiple testing lead to a drastic reduction in power to detect locus effects, compared to the theoretical optimum power for a single locus. The Wald test improves power several-fold by detecting lineage-specific effects. (**b**) Top: Mean numbers of false positive SNPs and patterns (i.e. 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 LD. Interpreting lineage effects is useful when the locus-specific signal cannot be fine-mapped. (**c**) The number of times common SNPs (MAF>20%) and antibiotic resistance phenotypes have emerged on the phylogeny. (**d**) When homoplasy is high, power to detect locus effects 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* 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.



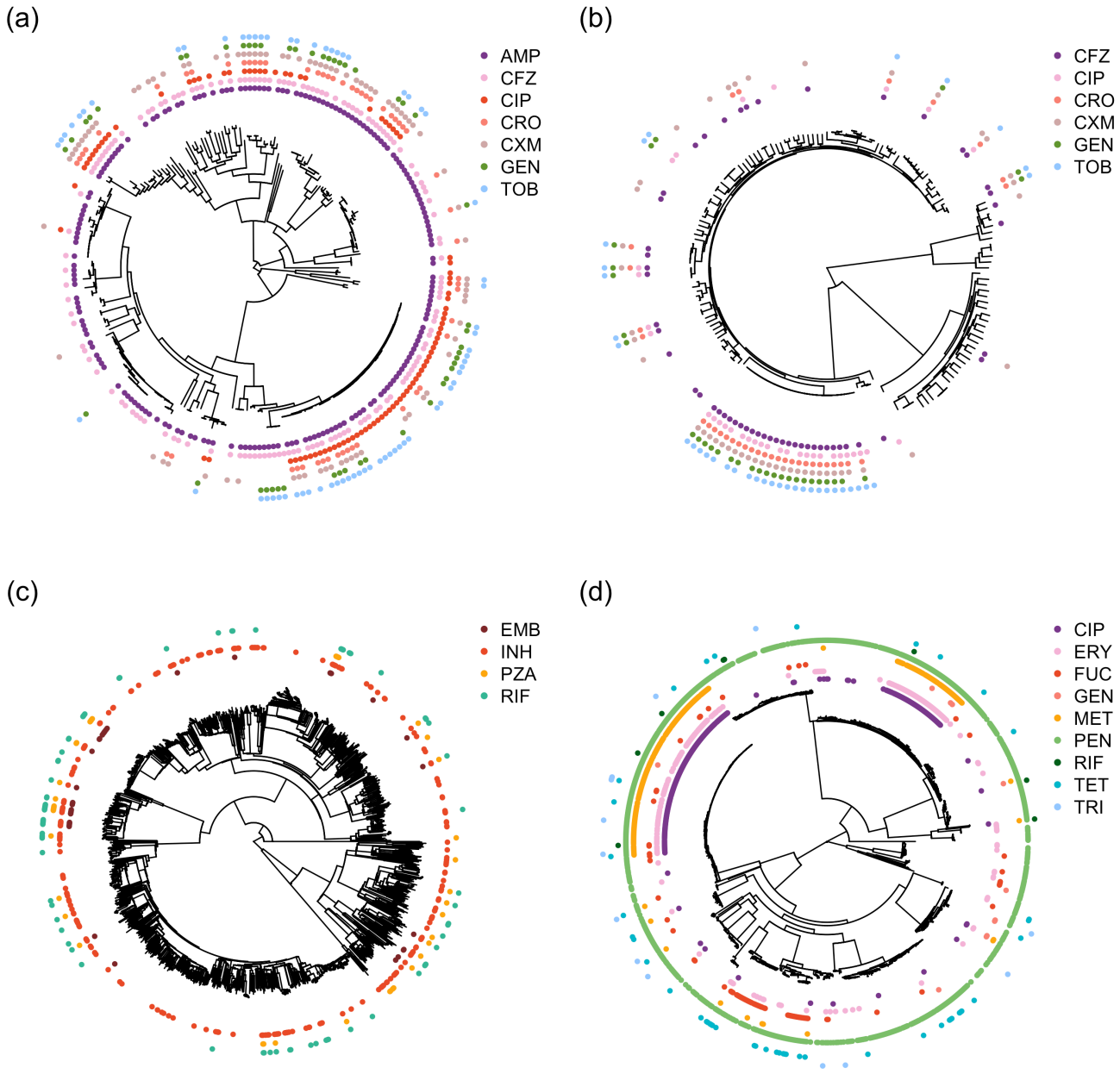
Supplementary Figure 5. Leading principal components correspond to major lineages in bacterial genealogies. For each species, branches of the maximum likelihood phylogeny are coloured by their correlation with each of PCs 1-5. Positive vs. negative correlations are shown in red vs. blue, brighter colours indicating stronger correlation. A square root transform was applied to branch lengths to visualize fine-scale structure.



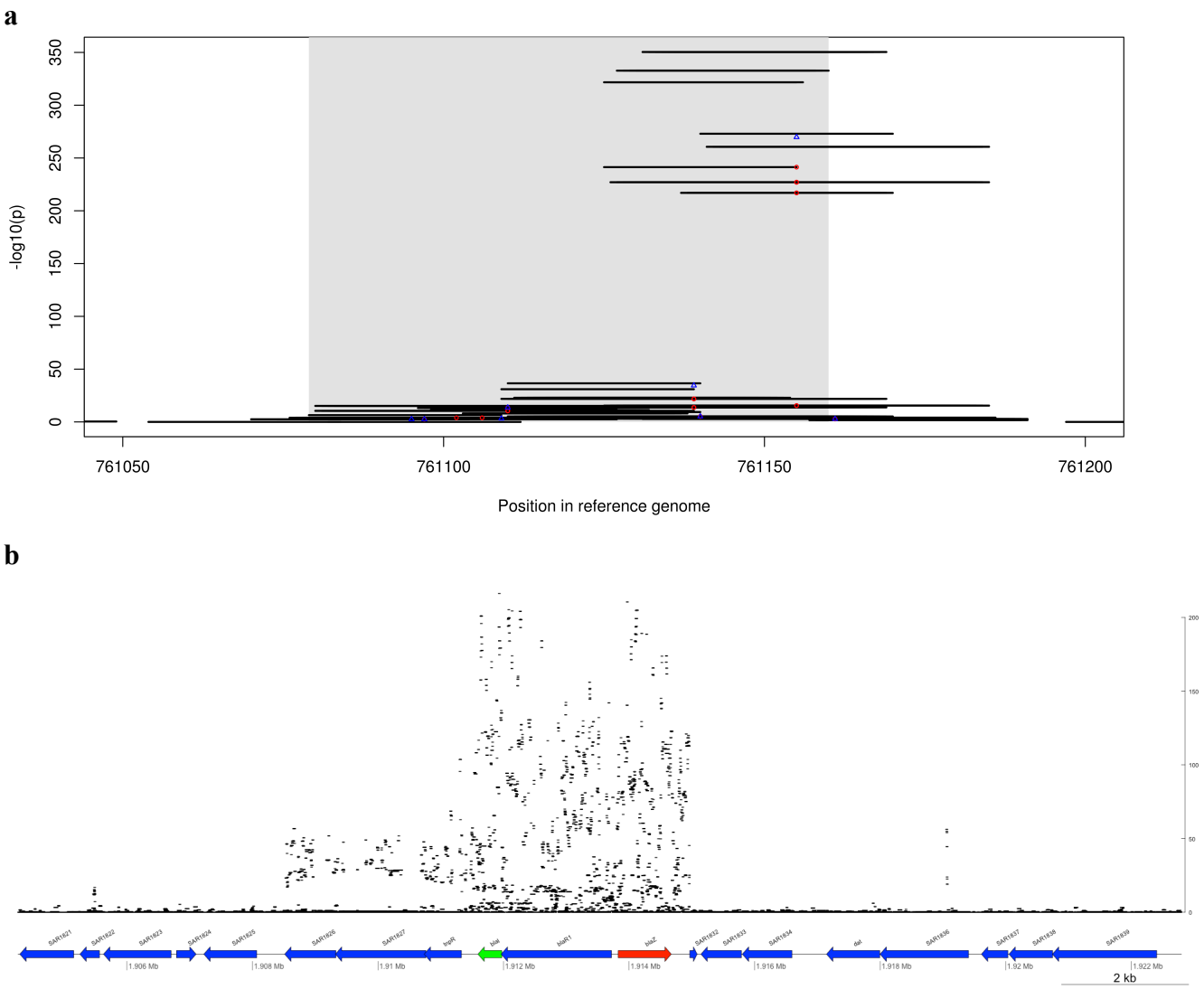
Supplementary Figure 6. The *S. aureus* phylogeny with the tips annotated by the fusidic acid resistance phenotype, orange for resistant and grey for sensitive. Branches of the tree are coloured according to their correlation with PC-6 (left), PC-9 (middle) and PC-7 (right). Positive correlations are coloured red and negative correlations blue, with brighter colouring indicating stronger correlation. Each PC appears to trace a path through the phylogeny, contrasting one lineage (red branches) with another (blue branches). The outer ring is annotated according to the projection of the individuals onto PC-6 (left), PC-9 (middle) and PC-7 (right), rescaled between -1 and 1 (from blue to red).



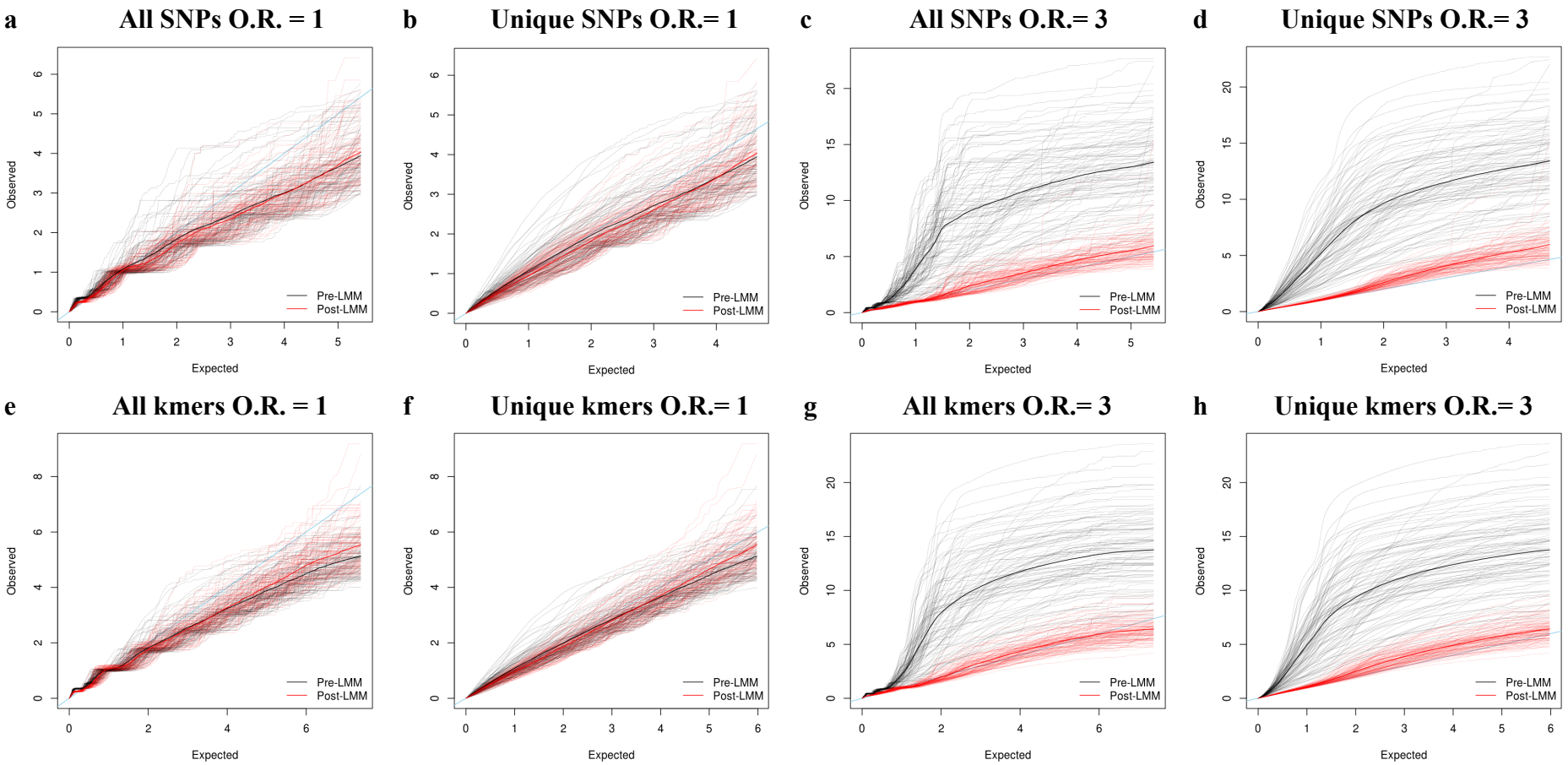
Supplementary Figure 7. Lineage effects are susceptible to confounding. For each of the four species, we conducted 100 simulations in the presence (HA) and absence (H0) of a causal SNP, in the presence and absence of confounding. The null hypothesis of no lineage effect was rejected when the PC-specific *p*-value from the Wald test exceeded the Bonferroni-corrected nominal rate of 0.05. A true rejection (green bars) was recorded for the PC most strongly correlated to the causal SNP (under HA); otherwise a false rejection (red bars) was recorded. In the absence of confounding (solid bars), the false rejection rate was well below the nominal rate under H0. Under HA, it was elevated in *M. tuberculosis* and *S. aureus* but it remained far below the true positive rate – this reflects the fact that multiple PCs are correlated to the causal SNP, therefore share some true signal. In the presence of confounding (hashed bars), the false rejection rate was inflated under both H0 and HA, demonstrating the susceptibility of lineage effects to capturing the effects of confounders in addition to genuine lineage-level differences. Baseline case probability was 0.25 and the rare allele at causal SNPs, chosen at random among those with frequency > 20% were assigned an odds ratio of 3. Confounding was simulated by generating phenotypic variability using the relatedness matrix, corresponding to mean heritability of 0.5 (*S. aureus*) 0.25 (*M. tuberculosis*) 0.1 (*E. coli*) and 0.1 (*K. pneumoniae*).



Supplementary Figure 8. Evolutionary relationships and distribution of antibiotic resistance in the four species: (**a**) *Escherichia coli* (n = 241), (**b**) *Klebsiella pneumoniae* (n = 176), (**c**) *Mycobacterium tuberculosis* (n = 1735) and (**d**) *Staphylococcus aureus* (n = 992). In each case the midpoint-rooted maximum likelihood clonal frame tree is shown.
AMP = Ampicillin, CFZ = Cefazolin, CIP = Ciprofloxacin, CRO = Ceftriaxone, CXM = Cefuroxime, GEN = Gentamicin, TOB = Tobramycin, EMB = Ethambutol, INH = Isoniazid, PZA = Pyrazinamide, RIF = Rifampicin, ERY = Erythromycin, FUC = Fusidic acid, GEN = Gentamicin, MET = Methicillin, PEN = Penicillin, RIF = Rifampicin, TET = Tetracycline, TRI = Trimethoprim.



Supplementary Figure 9. (a) A close up of the *rpoB* gene from the *M. tuberculosis* rifampicin kmer GWAS after controlling for population structure. The shaded region represents the codon positions within *rpoB* where amino acid substitutions alter resistance. Kmers were plotted as black lines, with their significance of association with rifampicin resistance after controlling for population structure on the y-axis. Blue triangles illustrate non-synonymous, nonsense and read-through SNP results after controlling for population structure. Kmers were annotated by SNPs they contained, with red circles representing non-synonymous, nonsense or read-through SNPs within the shaded region. The most significant SNP was a non-synonymous SNP causal of resistance, and the most significant kmer coincided with this region. However, the kmer was more significant, indicating an increase in power by pooling over multiple variants. This is shown by the 'wild-type' kmer containing no SNPs, thus it was found to be protective. **(b)** The region containing the penicillin resistance-conferring *blaZ* gene showing the kmer results for penicillin resistance in *S. aureus* after controlling for population structure. The gene containing the most significant kmers (the *blaZ* repressor *blaI*) is in green and *blaZ* is in red. The mobile element-associated region of LD was detected along with the causal mechanism.



Supplementary Figure 10. Bonferroni correction for the number of unique SNP or kmer patterns is better calibrated than using the total number of SNPs or kmers. We conducted 100 simulations in *S. aureus* (results were similar for other species) in the absence of confounding in the absence (**O.R.=1**) and presence (**O.R.=3**) of a causal variant with minor allele frequency > 20%, baseline case probability of 0.25 and odds ratio 3. For each simulation, we obtained *p*-values for the locus effect of every variant using logistic regression (black lines) and LMM (red lines). The thick lines show the mean across simulations. For each simulation, the line shows a qq-plot of the expected versus the observed significance ($-\log_{10}$ *p*-value), after ordering variants by significance. In (**a,e**) the distribution of observed *p*-values was substantially deflated relative to that expected (blue line), indicating they are conservative. However, strong population structure leads many SNPs to be indistinguishable phylogenetically. In (**b,d**) this removes most of the deflation, indicating that Bonferroni correction based on the number of unique patterns produces better-calibrated, less conservative *p*-values in the absence of a causal variant. Simulations in the presence of a causal variant (**c**,**d,g,h**) show that failure to control population structure in the logistic regression leads to massive inflation of significance, a problem that is largely alleviated by the LMM. However, in correcting for population structure, the LMM often loses the significance of the causal variant because it is population-stratified, emphasizing the desirability of recovering this signal through our test for lineage effects.

Species	% Ns to impute at the site	Sites originally imputed by CFML		Sites originally called		Allele frequency at the site	Sites originally imputed by CFML		Sites originally called	
		% Correct imputation CFML	% Correct imputation Beagle	% Correct imputation CFML	% Correct imputation Beagle		% Correct imputation CFML	% Correct imputation Beagle	% Correct imputation CFML	% Correct imputation Beagle
Escherichia coli										
	0-10%	97.6	98.0	96.5	97.6	0.00-0.01	92.0	89.3	93.2	91.9
	10-20%	96.0	96.5	94.9	95.6	0.01-0.05	90.6	86.9	90.9	88.1
	20-30%	94.8	95.1	94.0	94.3	0.05-0.10	91.4	90.5	91.8	90.5
	30-40%	94.6	94.7	93.7	93.9	0.10-0.15	92.6	92.4	91.9	91.6
	40-50%	92.3	91.1	91.9	90.9	0.15-0.20	93.8	93.5	92.9	92.8
	50-60%	93.4	92.8	93.0	92.1	0.20-0.25	93.0	93.3	91.9	92.1
	60-70%	92.3	89.6	91.7	89.0	0.25-0.30	93.8	94.5	92.3	93.1
	70-80%	91.0	89.0	90.1	87.7	0.30-0.35	94.9	95.0	93.3	93.4
	80-90%	88.2	83.5	87.1	82.8	0.35-0.40	96.4	96.3	94.6	94.8
	90-100%	82.9	74.3	81.7	73.4	0.40-0.45	98.0	97.9	96.5	96.9
						0.45-0.50	97.3	97.7	96.1	97.0
Klebsiella pneumoniae										
	0-10%	95.8	93.7	95.8	94.3	0.00-0.01	90.1	86.5	92.1	88.6
	10-20%	92.9	91.2	92.0	90.1	0.01-0.05	90.9	88.5	92.3	89.8
	20-30%	90.4	88.4	91.3	89.2	0.05-0.10	91.8	91.2	92.3	91.2
	30-40%	94.8	94.1	95.6	94.6	0.10-0.15	93.7	93.5	94.9	94.3
	40-50%	94.6	94.4	95.8	95.1	0.15-0.20	93.1	92.5	94.3	92.9
	50-60%	95.6	95.3	96.6	96.0	0.20-0.25	91.7	91.1	96.0	94.5
	60-70%	94.6	93.8	95.6	94.3	0.25-0.30	84.6	85.8	84.5	87.2
	70-80%	92.8	90.7	94.0	90.6	0.30-0.35	85.0	85.0	82.1	85.7
	80-90%	80.7	78.1	79.7	76.1	0.35-0.40	85.4	85.2	82.2	85.9
	90-100%	82.5	72.7	83.9	71.6	0.40-0.45	90.2	90.5	86.0	89.3
						0.45-0.50	94.2	93.2	88.4	91.7
Mycobacterium tuberculosis										
	0-10%	99.8	98.0	99.8	99.0	0.00-0.01	93.8	86.4	96.8	94.9
	10-20%	99.9	98.7	99.8	99.0	0.01-0.05	98.4	81.9	98.3	90.2
	20-30%	99.7	98.7	99.8	98.8	0.05-0.10	99.5	91.8	99.8	93.8
	30-40%	99.5	97.8	99.7	98.3	0.10-0.15	98.4	88.3	99.6	94.3
	40-50%	99.9	98.4	99.7	98.4	0.15-0.20	100	97.1	100.0	96.7
	50-60%	99.4	93.6	99.6	98.4	0.20-0.25	100	96.9	99.7	96.2
	60-70%	95.7	94.2	99.0	96.2	0.25-0.30	99.6	93.5	99.9	96.3
	70-80%	98.8	92.3	98.4	95.3	0.30-0.35	100	95.9	100	96.4
	80-90%	98.4	88.4	97.8	91.5	0.35-0.40	100	97.2	100.0	96.0
	90-100%	89.5	72.0	88.3	78.1	0.40-0.45	100	75	100	98.5
						0.45-0.50	NA	NA	NA	NA
Staphylococcus aureus										
	0-10%	98.1	97.7	98.0	97.9	0.00-0.01	84.3	79.8	92.5	89.8
	10-20%	97.0	96.5	97.7	97.1	0.01-0.05	87.5	80.7	91.4	85.9
	20-30%	93.9	92.8	94.9	94.4	0.05-0.10	88.4	80.8	92.7	88.8
	30-40%	89.2	85.2	91.7	88.2	0.10-0.15	90.2	86.4	91.5	87.8
	40-50%	91.7	87.2	93.7	89.6	0.15-0.20	91.8	90.1	92.9	90.5
	50-60%	87.5	79.0	90.7	84.4	0.20-0.25	91.0	89.1	92.9	90.8
	60-70%	85.9	77.5	89.0	82.3	0.25-0.30	93.9	91.0	94.3	91.4
	70-80%	88.7	81.2	91.1	84.6	0.30-0.35	95.7	93.5	95.3	93.6
	80-90%	85.6	79.3	88.4	82.9	0.35-0.40	95.9	95.9	95.2	95.4
	90-100%	73.8	67.4	77.5	69.1	0.40-0.45	96.7	96.8	95.9	96.2
						0.45-0.50	97.3	96.8	95.0	96.4

Supplementary Table 1. Imputation accuracy for ClonalFrameML and Beagle per species. Results have been stratified into accuracy for sites which were originally imputed using ClonalFrameML, and accuracy for sites which were originally all called, so that the results are not biased by the original imputation by ClonalFrameML.

Species	Antimicrobial(s)	Gene	Product	SNP	Gene	Amino acid substitutions/region
Escherichia coli, Klebsiella pneumonia						
	Penicillins: Ampicillin	*bla*$_{TEM}$	beta-lactamase			
		bla$_{SHV}$	beta-lactamase			
		bla$_{CBL}$	beta-lactamase			
		bla$_{BRO}$	beta-lactamase			
		bla$_{HERA}$	beta-lactamase			
		bla$_{LEN}$	beta-lactamase			
		bla$_{M}$	beta-lactamase			
		bla$_{P}$	beta-lactamase			
		bla$_{ROB}$	beta-lactamase			
		bla$_{VHH}$	beta-lactamase			
		bla$_{VHW}$	beta-lactamase			
		ampC promoter	mutations in *ampC* promoter			
		bla$_{TEM}$ *promoter*	mutations in *bla*$_{TEM}$ promoter			
	Cephalosporins: Cefazolin, Cefuroxime, Ceftriaxone	*bla*$_{TEM}$	beta-lactamase			
		bla$_{SHV}$	beta-lactamase			
		bla$_{CTX-M}$	beta-lactamase			
		bla$_{ACC}$	beta-lactamase			
		bla$_{ACI}$	beta-lactamase			
		bla$_{ACT}$	beta-lactamase			
		bla$_{AST}$	beta-lactamase			
		bla$_{BES}$	beta-lactamase			
		bla$_{BIC}$	beta-lactamase			
		bla$_{BIL}$	beta-lactamase			
		bla$_{CBL}$	beta-lactamase			
		bla$_{CEPA}$	beta-lactamase			
		bla$_{CGA}$	beta-lactamase			
		bla$_{CKO}$	beta-lactamase			
		bla$_{CIA}$	beta-lactamase			
		bla$_{CME}$	beta-lactamase			
		bla$_{CSP}$	beta-lactamase			
		bla$_{CMY}$	beta-lactamase			
		bla$_{CFE}$	beta-lactamase			
		bla$_{DES}$	beta-lactamase			
		bla$_{DHA}$	beta-lactamase			
		bla$_{ERP}$	beta-lactamase			
		bla$_{FONA}$	beta-lactamase			
		bla$_{FOX}$	beta-lactamase			
		bla$_{GES}$	beta-lactamase			
		bla$_{IBC}$	beta-lactamase			
		bla$_{LAP}$	beta-lactamase			
		bla$_{LAT}$	beta-lactamase			
		bla$_{LUT}$	beta-lactamase			
		bla$_{MIR}$	beta-lactamase			
		bla$_{MOR}$	beta-lactamase			
		bla$_{MOX}$	beta-lactamase			

	bla$_{OCH}$	beta-lactamase	
	bla$_{OKP}$	beta-lactamase	
	bla$_{ORN}$	beta-lactamase	
	bla$_{OXA}$	beta-lactamase	
	bla$_{OXY}$	beta-lactamase	
	bla$_{PER}$	beta-lactamase	
	bla$_{PLA}$	beta-lactamase	
	bla$_{PME}$	beta-lactamase	
	bla$_{RAHN}$	beta-lactamase	
	bla$_{SED}$	beta-lactamase	
	bla$_{SFO}$	beta-lactamase	
	bla$_{SPU}$	beta-lactamase	
	bla$_{SRT}$	beta-lactamase	
	bla$_{SST}$	beta-lactamase	
	bla$_{TER}$	beta-lactamase	
	bla$_{TLA}$	beta-lactamase	
	bla$_{TRU}$	beta-lactamase	
	bla$_{VEB}$	beta-lactamase	
	bla$_{ZEG}$	beta-lactamase	
	cfxA	beta-lactamase precursor	
	mecA	low-affinity PBP2	
Ciprofloxacin	*gyrA*	DNA topoisomerase (ATP-hydrolyzing) subunit A	quinolone-resistance determining region: amino acids 67-106
	gyrB	DNA gyrase, subunit B	quinolone-resistance determining region: amino acids 426-464
	parC	DNA topoisomerase IV subunit A	quinolone-resistance determining region: amino acids 47-133
	parE	DNA topoisomerase IV, subunit B	quinolone-resistance determining region: amino acids 420-458
	aac(6')-II	aminoglycoside N-acetyltransferase	
	qnr	quinolone resistance protein	
	qepA	quinolone efflux pump	
	oqxA	multidrug efflux membrane fusion protein	
	oqxB	multidrug efflux membrane fusion protein	
Gentamicin	*aac(3)-I*	aminoglycoside N-acetyltransferase	
	aac(3)-II	aminoglycoside N-acetyltransferase	
	aac(3)-III	aminoglycoside N-acetyltransferase	
	aac(3)-IV	aminoglycoside N-acetyltransferase	
	aac(3)-VI	aminoglycoside N-acetyltransferase	
	aac(3)-VII	aminoglycoside N-acetyltransferase	
	aac(6')-I	aminoglycoside N-acetyltransferase	
	aac(6')-II	aminoglycoside N-acetyltransferase	
	aac fusion variants	aminoglycoside N-acetyltransferase	
	ant(2'')-I	aminoglycoside O-nucleotidyltransferase	
	aph(2'')-I	aminoglycoside O-phosphotransferase	
	aph(2'')-II	aminoglycoside O-phosphotransferase	
	aph(2'')-III	aminoglycoside O-phosphotransferase	
	aph(2'')-IV	aminoglycoside O-phosphotransferase	
	apmA	rRNA methylase	
	npmA	rRNA methylase	
	armA	rRNA methylase	
	rmt	rRNA methylase	
Tobramycin	*aac(3)-II*	aminoglycoside N-acetyltransferase	
	aac(3)-III	aminoglycoside N-acetyltransferase	

	aac(3)-IV	aminoglycoside N-acetyltransferase	
	aac(2')-I	aminoglycoside N-acetyltransferase	
	aac(6')-I	aminoglycoside N-acetyltransferase	
	aac(6')-Ib-cr	aminoglycoside N-acetyltransferase	
	aac(6')-II	aminoglycoside N-acetyltransferase	
	aacA4	aminoglycoside N-acetyltransferase	
	aac fusion variants	aminoglycoside N-acetyltransferase	
	ant(2'')-I	aminoglycoside O-nucleotidyltransferase	
	ant(4)-I	aminoglycoside O-nucleotidyltransferase	
	ant(4)-II	aminoglycoside O-nucleotidyltransferase	
	apmA	rRNA methylase	
	npmA	rRNA methylase	
	armA	rRNA methylase	
	rmt	rRNA methylase	
Staphylococcus aureus			
Penicillin	*blaZ*	class A beta-lactamase	
Methicillin	*mecA*	low-affinity PBP2	
Erythromycin	*msrA*	erythromycin resistance protein	
	ermA	rRNA adenine N-6-methyltransferase	
	ermB	rRNA adenine N-6-methyltransferase	
	ermC	rRNA adenine N-6-methyltransferase	
	ermT	23S rRNA methylase	
Tetracycline	*tetK*	MFS tetracycline effux pump	
	tetL	MFS tetracycline effux pump	
	tetM	ribosomal protection protein	
Fusidic acid	*fusA*	fusidic acid resistance protein; elongation factor G (EF-G)	A160V*, A376V, A655E, A655P*, A655V*, A67T*, A70V*, A71V*, B434N, C473S*, D189G*, D189V*, D373N*, D463G*, E233Q*, E444K, E444V*, E449K*, F441Y, F652S*, G451V, G452C, G452S, G556S, G617D, G664S, H438N, H457Q, H457Y, L430S*, L456F, L461K, L461S, M161I*, M453I, M651I, P114H, P404L, P404Q, P406L, P478S, Q115L, R464C, R464H, R464S, R659C, R659H, R659L, R659S, R76C*, S416F*, T385N, T387I*, T436I, T656K, V607I, V90A, V90I, Y654N*
	fusB	fusidic acid resistance protein; EF-G-binding protein	
	fusC	fusidic acid resistance protein; EF-G-binding protein	
Trimethoprim	*dfrA*	trimethoprim resistance protein	
	dfrB	insensitive dihydrofolate reductase	F99Y, F99S, F99I, H31N, L41F, H150R, L21V*, N60I*
	dfrG	insensitive dihydrofolate reductase	
Gentamicin	*aacA-aphD*	6'-aminoglycoside N-acetyltransferase/ 2"-aminoglycoside phosphotransferase	
Ciprofloxacin	*gyrA*	DNA topoisomerase (ATP-hydrolyzing) subunit A	S84L, E88K, G106D, S85P, E88G, E88L
	grlA	DNA topoisomerase IV subunit A	S80F, S80Y, E84K, E84G, E84V, D432G, Y83N, A116E, I45M, A48T, D79V, V41G, S108N
	grlB	DNA topoisomerase IV subunit B	R470D*, E422D*, P451S*, P585S*, D443E*, R444S
Rifampicin	*rpoB*	DNA-directed RNA polymerase subunit beta	A473T*, A477D, A477T*, A477V, D471G*, D471Y, D550G, H481D, H481N, H481Y, I527F, I527L*, I527 M*, ins 475H, ins G475*, L466S*, M470T*, N474K*, Q456K, Q468K, Q468L, Q468R, Q565R*, R484H, S463P, S464P, S486L, S529L*
Mycobacterium tuberculosis			
Ethambutol	*embA*	arabinosyltransferase A	S49R, M153T
	embB	arabinosyltransferase B	*306*, G406D, G406S, Q1002R, Q497R, T506N, W332R, E378A, K561R, S565G

	embC	arabinosyltransferase C	T270I, V981L, R738Q
Isoniazid	*ahpC*	alkyl hydroperoxide reductase subunit	C-39T, G-46A, G-88A
	fabG1	3-oxoacyl-ACP reductase	*-15*, *-16*, *-8*
	katG	catalase-peroxidase	G279D, T180K, T302R, V473F, Y300C, R463L
Pyrazinamide	*pncA*	pyrazinamidase/nicotinamidase	A11G, A102V, A134V, A146E, A161P, A171E, A46V, C138R, C138Y, C14R, C72R, D12A, D63G, D8G, G162D, G17D, G78C, G97D, G97S, H137R, H51P, H57D, H57Y, H71R, H82R, K96N, K96T, L116R, L159P, L172E, L172P, L19P, L27P, L35R, L4S, L85P, P54T, Q10P, Q141P, R121P, R140S, S104R, S185T, S66P, S67P, T-12C, T114P, T142K, T160P, T76P, V125G, V128G, V130G, V139L, V155G, V21G, W68C, W68G, W68R, Y34S
Rifampicin	*rpoB*	DNA-directed RNA polymerase subunit beta	*425*, *426*, *427*, *428*, *429*, *430*, *431*, *432*, *433*, *434*, *435*, *436*, *437*, *438*, *439*, *440*, *441*, *442*, *443*, *444*, *445*, *446*, *447*, *448*, *449*, *450*, *451*, *452*

Supplementary Table 2. Summary of known antimicrobial resistance-conferring mechanisms. Compiled from Stoesser, N. *et al. J. Antimicrob. Chemother.* 68, 2234-2244 (2013), Gordon, N.C. *et al. J. Clin. Microbiol.* 52, 1182-1191 (2014) and Walker, T.M. *et al. Lancet Infect. Dis.* 15, 1193-1202 (2015).

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log10(*p*)	Rank	-log10(*p*) LMM	Rank LMM
				Pre-LMM	LMM													
AMP	*β-lactamase genes*	SNP / gene	$bla_{TEM-208}$	NC_017659.1				51	1	-	59	130	-	112.4	20.1	1	19.4	1
		Kmer	$bla_{TEM-208}$	NC_017654.1 (1455 - 1485)				52	0	-	53	136	-	Inf	23.6	6	19.7	6
	Tn3-like transposase	Kmer	Linked to $bla_{OXA-181}$	KP400525.1 (51445-51475)				50	2	-	43	146	-	84.9	25.9	1	21.2	1
CIP	*gyrA*	SNP / gene	D87N, D87W	2626015		C, T, A	NS	147	2	1	5	86	0	1264.2, 0, 0	55.7	2	18.5	2
		Kmer	-	2626026 - 2626056		-	-	136	14	-	1	90	-	874.3	41.5	45	43.4	1
	gyrB	SNP / gene		4380590		A, C, G	S	107	35	8	7	39	45	17.0, 86.0, 5.0	26.0	43	21.4	8
		Kmer																
	parC	SNP / gene	S80I	3595065		G, A	NS	136	14	-	1	90	-	874.3	59.6	1	55.8	1
		Kmer		3595065 - 3595095				4	146	-	86	5		0.00159	45.6	1	38.7	28
	parE	SNP / gene	I529L	3610603		T, G	NS	142	8	-	46	45	-	17.4	15.2	6626	14.8	156
		Kmer		NC_013361.1 (627842 - 627812)														
CFZ	*β-lactamase genes*	SNP / gene	$bla_{CTX-M-15}$	U5SQ39				102	0	-	96	43	-	Inf	12.7	2	6	3
		Kmer	$bla_{CTX-M-15}$	DQ335219.1 (405 - 435)				102	0	-	107	32	-	Inf	6.71	121710	3.99	3690
	nmpC	SNP / gene		P21420				15	87	-	91	48	-	0.09	15.4	1	12.4	1
		Kmer	*nmpC*	1985557 – 1985587				16	86	-	91	48	-	0.1	13.8	1	9.6	1
CXM	*β-lactamase genes*	SNP / gene	$bla_{CTX-M-14}$	U5SQ39				159	1	-	39	42	-	171.2	23.2	1	18.99	1
		Kmer	$bla_{CTX-M-15}$	KP268826.1 (7 - 37)				160	0	-	50	31	-	Inf	16.3	1598	15.4	470
	Intergenic	Kmer	Linked to bla_{CMY-2} (31177 - 32322)	LC019731.1 (31015 - 31045)				160	0	-	38	43	-	Inf	25.6	1	20.0	1

CRO	β-*lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	NC_022648.1	185	1	-	13	42	-	597.7	34.5	1	48.2	1
		Kmer	*bla*$_{CTX-M-15}$	KP268826.1 (7 - 37)	186	0	-	24	31	-	Inf	27.3	1403	34.9	470
	tnpA – ISECP1	Kmer	Linked to bla$_{CTX-M-132}$ (8362 - 9237; complement)	KM207012.2 (9298 – 9328)	186	0	-	12	43	-	Inf	39.7	1	56.4	1
GEN	aac	SNP / gene	*aac(3)-II*	ESD46483.1	192	1	-	9	39	-	832.0	35.5	1	68.4	1
		Kmer	*aac(3)-II*	CP008735.1(7913-7943) CP008735.1(7913-7943)	193	0	-	9	39	-	Inf	41.9	1	74.0	1
TOB	aac	SNP / gene	*aac(3)-II*	ESD46483.1	174	0	-	27	40	-	Inf	28.6	1	30.5	1
		Kmer	*aac(3)-II*	KJ850481(134-164) KJ850481(134-164)	174	0		27	40	-	Inf	28.2	1	30.5	1

Supplementary Table 3a. *Escherichia coli* results. AMP = Ampicillin, CFZ = Cefazolin, CIP = Ciprofloxacin, CRO = Ceftriaxone, CXM = Cefuroxime, GEN = Gentamicin, TOB = Tobramycin. Case = phenotypically resistant, control = phenotypically sensitive, aac = Aminoglycoside N-acetyltransferase genes, ant = Aminoglycoside N-acetyltransferase genes, aph = Aminoglycoside O-phosphotransferase genes. Causal gene names are coloured according to their resistance causing mechanism, red if it's presence determines resistance, blue if substitutions within the gene causes resistance.

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log10(*p*)	Rank	-log10(*p*) LMM	Rank LMM
				Pre-LMM	LMM													
CFZ	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	A0A075VKM9				122	1	-	20	33	-	201.3	20.8	1	15.2	2
		Kmer	*bla*$_{CTX-M-15}$	DQ335219.1 (110-140)				122	1	-	20	33	-	201.3	20.6	762	15.2	837
	HP from ISEcp1	SNP / gene						122	1	-	20	33	-	201.3	20.8	1	15.2	2
	protein WbuC	SNP / gene		AIG86706.1				122	1	-	20	33	-	201.3	20.8	1	15.2	1
	ISEcp1 tnpA	Kmer	Linked to *bla*$_{CTX-M-15}$	EU418923.1 (10812 - 10842, reverse)				122	1	-	19	34	-	218.3	21.3	1	18.3	1
CXM	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	A0A075VKM9				129	1	-	13	33	-	327.5	24.2	1	23.4	1
		Kmer	*bla*$_{CTX-M-24}$	NC_022078.1 (127606 - 127633, reverse)				129	1	-	13	33	-	327.5	25.0	772	23.4	1480
	ISEcp1 tnpA	Kmer	Linked to *bla*$_{CTX-M-15}$	EU418923.1 (10812 - 10842, reverse)				129	1	-	12	34	-	365.5	25.9	1	26.6	1
	HP from ISEcp1	SNP / gene						129	1	-	13	33	-	327.5	24.2	1	23.4	1
	protein wbuC	SNP / gene		AIG86706.1				129	1	-	13	33	-	327.5	24.2	1	23.4	1
CRO	*β-lactamase genes*	SNP / gene	*bla*$_{CTX-M-15}$	AIG86707.1				140	1	-	2	33	-	2310	32.8	1	60.5	1
		Kmer	*bla*$_{CTX-M-24}$	NC_022078.1 (127606 - 127633, reverse)				140	1	-	2	33	-	2310	35.4	762	60.5	803
	HP from ISEcp1	SNP / gene						140	1	-	2	33	-	2310	32.8	1	60.5	1

	protein wbuC	SNP / gene		AIG86706.1	140	1	-	2	33	-	2310	32.8	1	60.5	1
	ISEcp1 tnpA	Kmer	Linked to $bla_{CTX\text{-}M\text{-}15}$	EU418923.1 (10812 - 10842, reverse)	140	1	-	1	34	-	4760	36.7	1	76.0	1
CIP	Plasmid-mediated quinolone resistance genes	SNP / gene	*aac(6')-Ib-c*	ACV60575.1	138	4	-	8	26	-	112.1	20	5	16.7	4
		SNP / gene	*qnr-B1*	A0A075VJL2	140	2	-	9	25	-	194.4	20.7	2	19.5	2
		Kmer	*qnr-B19*	JX298080.1 (481 – 511)	130	2	-	9	25	-	180.6	25.0	1846	19.5	4423
	tnpA (truncated)	Kmer	Linked to *qnr-B19*	JX298080.1 (1520 – 1550)	135	7	-	3	31	-	199.3	27.3	1	28.5	1
	tnpA	SNP / gene	Linked to aac	BAD08693.1	131	11	-	2	32	-	190.5	23.5	1	19.9	1
GEN	aac	SNP / gene	*aac(3)-II*	AHI38985.1 .	145	0	-	0	31	-	NA	36.8	1	>100	1
		Kmer	*aac(3)-II*	AJD77170.1(383 – 413)	145	0	-	0	31	-	Inf	39.4	1	15.6	397987
	ant	SNP / gene	*aac(6)*	AIG86041.1	141	4	-	5	26	-	183.3	22.1	6	18.6	7
	tmrB_2 (Tunicamycin resistance protein)	Kmer	Linked to aacA4 (314520 – 315119, complement)	CP011314.1 (310752 – 310782)	145	0	-	2	29	-	Inf	36.5	519	146.4	1

TOB	aac	SNP / gene	***aac(3)-II***	AIG86707.1	140	0	-	5	31	-	Inf	30.4	1	43.3	1
		Kmer	***aac(3)-II***	LK391770.1 (21132 - 21162)	140	0	-	5	31	-	Inf	33.0	1	43.3	1
	ant	SNP / gene	***aac(6)***	AIG86041.1	139	1	-	7	29	-	575.6	25.6	3	27	3

Supplementary Table 3b. *Klebsiella pneumoniae* results for drugs where resistance is determined by the presence of a gene. CFZ = Cefazolin, CIP = Ciprofloxacin, CRO = Ceftriaxone, CXM = Cefuroxime, GEN = Gentamicin, TOB = Tobramycin. Case = phenotypically resistant, control = phenotypically sensitive. aac = Aminoglycoside N-acetyltransferase genes, ant = Aminoglycoside N-acetyltransferase genes, aph = Aminoglycoside O-phosphotransferase genes. Causal gene names are coloured according to their resistance causing mechanism, red if it's presence determines resistance, blue if substitutions within the gene causes resistance.

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log10(*p*)	Rank	-log10(*p*) LMM	Rank LMM
				Pre-LMM	LMM													
EMB	*embB*	SNP	M306L, M306V	4247429		A, G, C	NS	1586	2	1	24	16	1	528.7, 66.1, 0.1	25.6	2	82.8	1
		Kmer		4247429 - 4247459				31	1558	-	31	10	-	0.006	130.2	1	107.5	1
	rpoB	SNP	S450L, S450W	761155		C, T, G	NS	1563	23	3	17	23	1	91.9, 30.6 0.333	27.5	1	45.9	2
INH	*katG*	SNP	S315T	2155168		C, G	NS	1468	2	-	86	153	-	2475.8	151.1	1	169.4	1
		Kmer		2155145 - 2155175	2155145 - 2155175			1468	2	-	87	152	-	1282.4	220.9	1	172.4	1
PZA	*pncA*	SNP	V125G	2288868	2288868	A, C	NS	1662	0	-	41	4	-	Inf	7.2	142	60.0	1
		Kmer		2288847- 2288877	2288847- 2288877			1662	0	-	41	4	-	Inf	33.3	7890	60.0	1
		Kmer						1	1661	-	7	38	-	0.003	50.2	174	25.7	653
	rpoB	SNP	S450L, S450W	761155	761155	C, T, G	NS	1632	28	2	23	21	1	53.2, 35.5, 0.7	22.3	1	54.4	2
RIF	*rpoB*	SNP	S450L, S450W	761155	761155	C, T, G	NS	1486	0	1	34	49	3	Inf, 131.1, 0	73.2	1	269.8	1
		Kmer		761136 - 761166	761136 - 761166			7	1480	-	70	16	-	0.001	250.0	1	0	1
		Kmer		761126 - 761156	761126 - 761156			6	1481	-	66	20	-	0.001	237.2	14	321.7	1

Supplementary Table 3c. *Mycobacterium tuberculosis* results. EMB = Ethambutol, INH = Isoniazid, PZA = Pyrazinamide, RIF = Rifampicin. Case = phenotypically resistant, control = phenotypically sensitive. Causal gene names are coloured according to their resistance causing mechanism, red if it's presence determines resistance, blue if substitutions within the gene causes resistance.

Drug	Gene	Study	Variant	Genome position (in reference genome) or BLAST accession		Alleles	Type	Ctrl 1	Ctrl 2	Ctrl 3	Case 1	Case 2	Case 3	Odds ratio	-log10(*p*)	Rank	-log10(*p*) LMM	Rank LMM
				Pre-LMM	LMM													
CIP	*grlA*	SNP / gene	S80F, S80Y, F80Y	1419998		C, T, A	NS, NS, NS	745	5	0	14	226	2	2405.3, Inf, Inf	198.6	1	138.2	1
		Kmer		1419968-1420098				745	5	-	18	224	-	1854.2	190.5	1	99.1	9
		Kmer			1419995-1420025			22	728	-	228	14	-	0.002	177.2	46	113.6	1
	gyrA	SNP / gene	S84L	7255		C, T	NS	745	5	-	20	222	-	1653.9	190.3	2	81.6	2
		Kmer		7726 - 7756				6	744	-	223	19	-	0.0007	188.3	23	82.3	47
ERY	*ermC*	SNP / gene	Q2FDD1					774	2	-	103	113	-	424.6	85.0	1	193.5	1
		Kmer		NC_022228.1 (743-781)	NC_022228.1 (279-309)			770	6	-	102	114	-	143.4	94.7	1	192.4	12
	ermA	SNP / gene	P0A0H3					775	1	-	121	95	-	608.5	70.7	12	75.7	3
	ermT	SNP / gene					clusters with ermA in pan-genome											
	Intergenic, 30bp upstream to *ermC*	Kmer		AE002098.2 (75634-75664)				772	4	-	106	110	-	200.3	93.1	92	195.7	1
FUS	*fusA*	SNP / gene	L461-, L461S, -461S	601084		T, A, C	Nonsense, NS, read-through	908	0	-	83	1	-	Inf	11.0	9754	62.3	1
		kmer		601054-601084				1	907	-	18	66	-	0.004	41.6	115876	157.4	1
	fusC	SNP / gene	Q8GNY5					908	0	-	40	44	-	Inf	54.3	4	19.3	39
		kmer		NC_002953.3 (53216-53246)				908	0	-	36	48	-	Inf	119.9	75	39.7	58
	fusB	SNP / gene	Q8GNY5					908	0	-	76	8	-	Inf	9.6	18494	16.6	76
	SAS0040	Kmer						908	0	-	35	49	-	Inf	122.5	1	45.4	36
	SAS0037	SNP / gene						908	0	-	38	46	-	Inf	57.0	1	28.5	21
	SAS0040	SNP / gene						908	0	-	38	46	-	Inf	57.0	1	28.5	21

GEN	*aacA/aphD*	SNP / gene	P0A0C1				981	0	-	2	9	-	Inf	21.1	1	380.8	1
		Kmer	-	AY971367.1 (727-2683)			981	0	-	2	9	-	Inf	177.4	1	380.8	1
	GNAT acetyltransferase	SNP / gene	D2J631				981	0	-	2	9	-	Inf	21.1	1	380.8	1
MET	*mecA*	SNP / gene	P60185				773	4	-	3	212	-	13656.3	209.6	1	374.9	1
		Kmer		NC_022604.1 (78438- 78468)			772	4	-	3	213	-	13703	208.2	1	375.6	1
	HP in SCC-mec	SNP/gene					773	4	-	3	212	-	13656.3	209.6	1	374.9	1
PEN	*blaZ*	SNP / gene	P00807				145	23	-	28	796	-	179.2	118.6	1	140.1	1
		Kmer		NC_022604.1 (2824752-2824782)			143	25	-	9	815	-	518.0	166.4	2	210.5	2
	blaI			NC_022604.1 (2822414-2822444)			142	26	-	7	817	-	637.4	167.8	1	216.2	1
RIF	*rpoB*	SNP / gene	H481Y	592271	C, T	NS	983	1	-	3	5	-	1638.3	11.1	1	158.6	1
		kmer		592260-592290			2	982	-	6	2	-	0.0007	122.0	1	177.9	1
TET	*tetK*	SNP / gene	B0FYM6				945	1	-	9	37	-	3885	58.0	2	315.4	2
		Kmer		KM281803.1 (1198-1228)			945	1	-	4	42	-	9922.5	192.6	1	464.1	1
	tetL																
	tetM				cluster with *tetK* in the pan-genome												
	repC	SNP / gene	Q5701				944	2	-	6	40	-	3146.7	62.9	1	364.6	1
TRI	*dfrB*	SNP / gene	F99Y	1497290	A, T	NS	308	0	-	10	5	-	Inf	7.9	1	28.5	1
		Kmer		1497269- 1497299			308	0	-	10	5	-	Inf	23.8	1	28.5	1
	dfrG	SNP/gene					308	0	-	12	3		Inf	4.9	2	16.1	1
	dfrA				clusters with dfrG in pan-genome												
	orfu1	SNP/gene					308	0	-	12	3		Inf	4.9	2	16.1	1
	LPXTG surface protein	SNP/gene					301	7	-	9	6		28.7	5.6	1	13.5	2

Supplementary Table 3d. *Staphylococcus aureus* results. CIP = Ciprofloxacin, ERY = Erythromycin, FUS = Fusidic acid, GEN = Gentamicin, MET = Methicillin, PEN = Penicillin, RIF = Rifampicin, TET = Tetracycline, TRI = Trimethoprim. Case = phenotypically resistant, control = phenotypically sensitive. Causal gene names are coloured according to their resistance causing mechanism, red if it's presence determines resistance, blue if substitutions within the gene causes resistance.

	Minor allele frequency (%)					
Species	**0 – 1%**	**1 – 2%**	**2 – 5%**	**5 – 10%**	**10 – 20%**	**20 – 50%**
Escherichia coli	26.5	9.0	11.4	10.7	12.9	29.5
Klebsiella pneumoniae	21.2	11.3	17.1	17.1	16.5	16.8
Mycobacterium tuberculosis	93.9	1.8	1.6	1.0	1.1	0.7
Staphylococcus aureus	49.2	6.9	6.5	9.7	5.2	22.5

Supplementary Table 4. Distribution of minor allele frequencies for biallelic SNPs.

Species	GWAS studies	Reference genome for mapping	# Bialellic SNPs	# Triallelic SNPs	# Tetra-allelic SNPs	# Kmers	# Gene clusters
Staphylococcus aureus	Ciprofloxacin, Erythromycin, Fusidic acid, Gentamicin, Penicillin, Methicillin, Tetracycline, Rifampicin	*Staphylococcus aureus* MRSA252 (GenBank accession no. BX571856.1)	264604	14731	519	24154606	13881
	Trimethoprim	*Staphylococcus aureus* MRSA252 (GenBank accession no. BX571856.1)	196996	8712	269	15840354	10261
Escherichia coli	β-lactam: Ampicillin, Ceftazidime, Cefuroxime, Ceftriaxone; Quinolone: Ciprofloxacin; Amincoglycoside: Gentamicin, Tobramycin	*Escherichia coli* CFT073 (Genbank accession AE014075.1)	417645	25298	1287	39918870	23502
Klebsiella pneumoniae	β-lactam: Ampicillin, Ceftazidime, Cefuroxime, Ceftriaxone; Quinolone: Ciprofloxacin; Amincoglycoside: Gentamicin, Tobramycin	*Klebsiella pneumoniae* subsp. *pneumoniae* MGH 78578 (GenBank accession no. CP000647.1)	654425	63639	5029	53816250	21382
Mycobacterium tuberculosis	Ethambutol	*Mycobacterium tuberculosis* H37Rv (GenBank accession no. NC_000962.2)	107480	954	8	15680376	-
	Isoniazid	*Mycobacterium tuberculosis* H37Rv (GenBank accession no. NC_000962.2)	110400	1020	10	15941713	-
	Pyrazidamide	*Mycobacterium tuberculosis* H37Rv (GenBank accession no. NC_000962.2)	110162	1012	10	15963479	-
	Rifampicin	*Mycobacterium tuberculosis* H37Rv (GenBank accession no. NC_000962.2)	101968	864	8	15554437	-

Supplementary Table 5. Variant information for all GWAS studies.