Published on Mar 04, 2025 by Arcadia Science DOI: 10.57844/arcadia-8391-465f

Predicting antimicrobial resistance phenotypes across 7,000 *E. coli* genomes

We explored the genetic basis of antimicrobial resistance (AMR) phenotypes among 7,000 globally distributed strains of *E. coli*. AMR is associated with various genetic architectures that span multiple evolutionary scales.

Contributors (A-Z)

Audrey Bell, Erin McGeever, Austin H. Patton, George Sandler, Ryan York

Version 2 · Mar 31, 2025

Purpose

At Arcadia, we're interested in mapping genotype-phenotype relationships at broader evolutionary scales than previously attempted. To achieve this, we're developing models to capture genetic relationships — both linear and nonlinear — that may be inaccessible to conventional approaches. To further our development, we need a rare commodity: large-scale data from evolutionarily, genetically, *and* phenotypically diverse populations.

In this pub, we characterize a candidate dataset for model development composed of globally distributed samples of *E. coli*. Using exploratory population genomic and phylogenomic analyses of a previously published dataset of 7,000 *E. coli* genomes **[1]**, we describe extensive genetic diversity at both the strain level and among deeply branching phylogroups. We also verify that this dataset contains high-quality genotypic information that can be leveraged for model development. Finally, we show we can uncover the genetic basis of three antimicrobial resistance (AMR) phenotypes using conventional genomic prediction methods. These analyses expand our understanding of the evolution of these AMR phenotypes and set the baseline for future non-linear model development.

This work will interest anyone studying the evolution of antimicrobial resistance or the links between genotype and phenotype in evolutionary biology, breeding/agriculture, or genetics.

- **Data** from this pub is available on Zenodo.
- All associated **code** is available in this GitHub repository.
- This pub builds on a **dataset pub** we previously released, "Creating a 7,000-strain *E. coli* genotype dataset with antimicrobial resistance phenotypes."

Background and goals

Learning the genotype–phenotype map (i.e., how genotypic diversity translates to phenotypic diversity) is a fundamental goal in biology with many direct applications. As the availability of sequencing data has exploded over the past decade, many believed that genotype–phenotype maps would be resolved across the tree of life. However, while progress has been made in some cases — e.g., linking large-effect loci with specific traits — the genetic architecture underlying most phenotypic variation remains unresolved.

Why is this so? The methods used provide (at least) part of the answer. Most genotype–phenotype mapping approaches (e.g., genome-wide association studies; GWAS) are built on a strong assumption: we can explain the breadth of phenotypic variation by simply adding up the contributions of individual genomic loci **[2]**. However, it has been known for a long time that additive effects constitute just a portion of possible genomic interactions. It has also been known that by not capturing nonlinear effects like epistasis, linear models generally can't generalize across populations **[3]**. In these models, nonlinear effects are averaged over and effectively washed out when estimating a genetic variant's phenotypic contribution **[2]**. Additive models, therefore, can't capture the breadth of genetic patterns *within* populations and are unlikely to generalize *across* populations on broader evolutionary scales (where nonlinear interactions become increasingly prevalent). Given this, decoding genotype–phenotype maps will remain hard so long as exclusively linear models are used.

Models that capture linear *and* nonlinear interactions (e.g., autoencoders, transformers, graph neural networks, etc.) are appealing alternatives. Nonlinear models may have the flexibility to detect features like epistasis and additive interactions. In previous work, we found that nonlinear models can predict complex sets of phenotypes **[4]** and generalize across quantitative genetic applications **[5]**. While promising, these efforts were largely theoretical and/or relied on simulated data. Empirical datasets capturing natural genetic, phenotypic, and evolutionary complexity are needed to explore the full utility of nonlinear genotype–phenotype models.

We recently published a dataset of 7,000 globally distributed *E. coli* strains that, for several reasons, may fit this bill (referred to henceforth here as "*E. coli* 7k") **[1]**. First, *E. coli* has independently evolved antimicrobial resistance (AMR) phenotypes multiple times **[6]**. Second, *E. coli* displays substantial genetic variation, including both regular polymorphisms such as SNPs/short indels, as well as gene presence–absence variation **[7]**. Third, the global population of *E. coli* shows species, population, strain, and individual-level diversification **[8]**. These features suggest that this may be a good "model dataset" for testing nonlinear models that span multiple evolutionary scales.

In this pub, we characterize the extent to which these global features of *E. coli* diversity are present in this dataset. We then assess how well linear models predict variation in AMR phenotypes with diverse evolutionary histories. These analyses flesh out the utility of this dataset for model development and set a baseline for genomic prediction accuracy, highlighting diverse opportunities for future development.

> **SHOW ME THE DATA**: Raw genotype and phenotype data used in all analyses is available on Zenodo (DOI: 10.5281/zenodo.14364732).

The approach

Our goal was to assess the suitability of the previously published *E. coli* 7k dataset for genotype-to-phenotype mapping applications. To do this, we first performed some exploratory population genomic and phylogenetic analyses as both a sanity check on the genotyping calls and to assess the phylogenetic scope of the dataset. We then applied standard genomic prediction methods to three AMR phenotypes and analyzed the results in the context of previously published functional genetic data.

Data preparation and filtering

One caveat of the previously released *E. coli* 7k dataset is that pangenome reference genotypes were encoded as missing during genotype calling, meaning we can't differentiate between missing data and reference allele calls. The vast majority of the time, missing data should correspond to reference genotypes, so as an approximation, we assigned all samples with missing genotype calls as reference genotypes. This approach should be appropriate for most sites but will likely lead to reference-biased genotyping error in our analyses.

Our dataset is quite large, containing ~2.4 million genetic variants across our 7,000 strains. Although highly information-rich, the size of this dataset can be prohibitive for exploratory analyses. Furthermore, not all genotypic variants are independent of one another, whether due to physical proximity/linkage or due to evolutionary non-independence through patterns of shared ancestry. Thus, to pare down our genotypic data to a subset of sites suitable for downstream exploratory analyses, we applied several filtering criteria, constructing two analysis-specific datasets from this subset.

Dataset 1: Population genomic and phylogenetic analyses

Bacterial genomes can be divided broadly into a “core” genome, shared by all individuals, and an “accessory” genome that captures presence–absence variation **[9]**.

Polymorphism found within the core genome of bacterial species is likely to be of high significance, as these sites are transmitted vertically from generation to generation and thus reflect phylogenetic and population genomic signals considerably more than accessory gene content, which can be transmitted horizontally both between and within bacterial species **[9]**. Consequently, we first sought to identify which contigs in our reference pangenome correspond to the core *E. coli* genome. Looking across the 72 ECOR strains **[1][10]** used to construct our reference pangenome, we considered contigs that were present in all samples (i.e., weren't missing in any ECOR strain) to belong to the core *E. coli* genome.

We further filtered to retain only bi-allelic sites annotated as synonymous, missense, or loss of function (LOF) and visualized their respective site frequency spectra (SFS). These spectra revealed an excess of quadrupleton synonymous sites, suggesting the persistence of bioinformatic artifacts in our data. More careful investigation revealed this pattern was driven by 14 samples that possessed an over-enrichment of such quadrupletons (> 500); we thus removed these samples from our data, resulting in a more reasonable-looking SFS (Figure 1, A). For all downstream analyses, we focused exclusively on synonymous sites that had passed all filtering criteria thus far. To improve the computational efficiency of downstream analyses, we randomly retained 10% of genotypes with a minimum derived minor allele count (hereafter MAC) of 10, leaving us with a total of 13,352 synonymous sites for population genomic and phylogenetic analyses.

Population genomics

To explore broad patterns of genetic similarity in our dataset, we conducted a principal component analysis (PCA) using our filtered genotypic data as implemented in the R package SNPRelate (v1.36) **[11]**. We explored how various sample metadata features such as country of isolation, year of isolation, and multilocus sequence type (MLST) mapped onto the first five PC axes using multinomial logistic regression using the R package nnet (v7.3-19) **[12]**.

Phylogenetic inference

We inferred a strain-level phylogeny using IQ-TREE 2 (v2.3.5) **[13]**, using a general time reversible substitution model with unequal rates and base frequencies (GTR). We also applied an ascertainment bias correction (+ASC **[14]**) to adjust branch length estimates from SNP data. Motivated by the findings of our genomic prediction analyses (see below), we applied the same procedure for constructing a gyrA gene tree in downstream analyses. For this analysis, we simply restricted the inference to all polymorphic sites found on the contig mapping to this gene.

Dataset 2: Genomic prediction

Next, we sought to construct a genotypic dataset suited to the task of genomic prediction, starting from the complete genome. We removed excessively rare variants for these analyses as most models will have very little statistical power to estimate the effects for such polymorphism. We thus first filtered to retain sites with non-reference alleles that occurred at appreciable frequency (derived MAC ≥ 250), leading to the retention of 326,625 markers for genomic prediction in 7,043 samples (we again excluded the 14 individuals identified as outliers earlier). We chose not to prune our sites for linkage disequilibrium (LD, the statistical association between alleles in a population). Instead, we used models that induce sparsity in marker effect estimates through regularization using a Bayesian prior for marker effect sizes, thus allowing genotype-to-phenotype associations to drive marker selection rather than random sampling.

We also chose not to focus our analyses solely on bi-allelic markers as this can lead to the exclusion of important multi-allelic polymorphisms in the genome, especially in large datasets such as ours **[15]**. Furthermore, previous work has shown that presence/absence variation in the accessory genome can also play important roles in AMR evolution through genetic mechanisms such as plasmid exchange or transposon-mediated resistance **[16]**. Consequently, we also included a set of markers tracking presence/absence status for all 32,441 pangenome contigs and all 7,043 individuals in our analysis.

To genotype individuals for contig presence/absence status, we used SAMtools (v1.20) `idxtstats` **[17]** to calculate the number of reads mapping to each contig in each sample. We then normalized these counts by the contig length and total number of reads mapped to the pangenome in each sample. We then chose 12 random ECOR strains and visualized their distributions of normalized coverage. Based on these plots, we chose a normalized coverage cutoff of $1e^{-10}$ that separated the two distinct distributions of coverage (one for present contigs and one for absent contigs) that were apparent. We combined SNP/indel marker data and presence/absence pseudo-marker data using PLINK (v1.90b6.21) **[18]**, creating a merged output file we used in downstream genomic prediction analyses.

Genomic prediction

We used Bayesian sparse linear mixed models (BSLMM) **[19]** to perform genomic prediction as implemented in GEMMA (v0.98) **[20]**. We selected the BSLMM model as it allows markers to draw effect sizes from two distributions: a distribution for minor effects (such as those often associated with quantitative traits) and a distribution of much rarer major effect markers. We assumed *a priori* that this combination of distributions should be appropriate for the AMR phenotypes we're modeling, given that individual substitutions or alleles are often major determinants of resistance **[16]**.

To fit the model, we first calculated a centered relatedness matrix based on our set of markers. Then, independently, we ran a probit version of the BSLMM model for each AMR phenotype. Strictly speaking, our phenotypes are encoded in three levels (susceptible, intermediate, and resistant). Susceptible was the most common state for almost all phenotypes, with a subset of individuals being resistant and a very small fraction labeled intermediate. Given the rarity of susceptible samples and that resistant/intermediate phenotypes likely share genetic features that lead to any level of resistance, we coerced our phenotypes to a binary state as required for a probit model, encoding all intermediate individuals as resistant (coding 0=susceptible, 1=intermediate/resistant). As in previous analyses, we assumed that all missing genotypes were, in fact, masked reference alleles. We left all other parameters as their defaults. Fitted BSLMMs return posterior mean estimates for marker effect size parameters, including alpha (corresponding to a minor effect estimate), beta (a major effect estimate), and gamma (a parameter estimating the probability that beta is non-null). We estimated overall marker effects as $\alpha + \beta * \gamma$, as suggested in the GEMMA manual, sorting markers based on the absolute value of this estimate of total effect.

Additional analyses

To look for evidence of physical linkage between candidate resistance markers, we calculated a measure of covariance in allelic state, LD. To do this, we used PLINK (v1.90b6.21) **[18]** and calculated all vs. all LD between our top ten largest-effect markers for all phenotypes (separately for each phenotype), using options `--r2 --allow-extra-chr --ld-window-r2 0` to calculate all vs. all inter-chromosome r2.

Our genomic prediction analyses revealed that ciprofloxacin resistance was chiefly driven by three tightly linked resistance mutations, which likely evolved independently several times across our phylogeny. We, therefore, also used corHMM (v2.8) **[21]** to try and tease apart the most likely/frequent order in which these three mutations arose during ciprofloxacin resistance evolution in our dataset. Using hidden Markov models to infer transition rates between discrete states (here, genotypic state at three resistance sites) along branches of a phylogeny, corHMM constructs models wherein unsampled (“hidden”) states are codistributed alongside the sampled state. These hidden states allow the model to capture more biologically realistic rate variation across the phylogeny.

We first time-calibrated the phylogeny we inferred from genome-wide SNP data using the least squares dating method for tip-dating **[22]** implemented in IQ-TREE using year of isolation metadata associated with our samples as time reference points. To estimate hidden state transition rates, we modeled the reference and alternate alleles as a binary phenotype for all three mutational positions (gyrA248, gyrA259, and parC239). We fit a model using default corHMM options with asymmetrical transition rates and a single hidden rate category (i.e., each transition is modeled using a single rate). We limit our interpretations of resistance evolution to forward single-step mutations estimated by corHMM as this is the most biologically plausible path of resistance evolution, barring rare double mutation events for which our dataset likely lacks the necessary resolution to capture.

Additional methods

We used Grammarly Business to reorganize text using a template, reformat text according to a style guide, and help clarify and streamline the text that we wrote.

All **code** generated and used for the pub is available in this GitHub repository (DOI: 10.5281/zenodo.14941875).

The results

SHOW ME THE DATA: Raw genotype and phenotype data used in all analyses is available on Zenodo.

Population genomic patterns

The *E. coli* 7k dataset is genomically and phenotypically diverse. To facilitate downstream model development, we were interested in characterizing any technical (e.g., sequencing noise or bioinformatic errors) or biological (e.g., sampling imbalances or genetic complexity) factors that might limit its utility. To this end, we performed a series of population genomic and phylogenetic analyses to determine which patterns in the data reflected expected biological processes and which were the product of technical and/or biological artifacts.

First, we looked for evidence of genome-wide purifying selection, an expected population genetic phenomenon in real-world populations. Purifying selection removes deleterious mutations from a population, meaning that more deleterious mutations should, on average, segregate at lower allele frequencies than less deleterious ones. This results in a left-shifted and rapidly decaying site frequency spectrum (SFS). Deviations from this pattern can indicate technical/bioinformatic error. Population genomic datasets lacking signals of purifying selection may be corrupted by technical artifacts (e.g., sequencing or bioinformatic issues).

We compared the (unfolded) SFS of three types of mutations: 1) synonymous (low to no effect), 2) missense (moderate effect), and 3) loss-of-function (LOF; large effect). As expected, rare allele frequency increased with mutational effect, indicating that purifying selection has acted to remove more common deleterious mutations in *E. coli* (Figure 1, A). Moreover, the SFS of all types of mutations decays roughly monotonically with allele frequency, another expected biological signal. Most mutations in a population aren't expected to reach high frequencies due to random sampling. These results, therefore, rule out allele frequency imbalances due to technical errors.

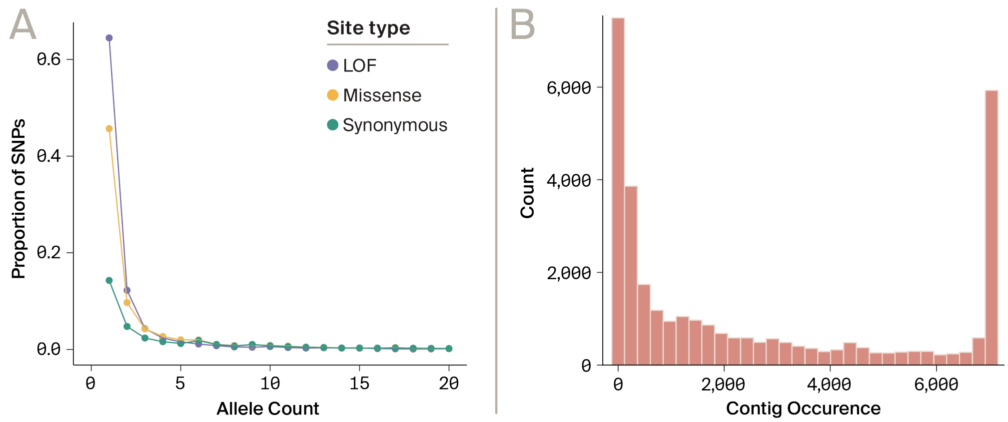

Figure 1

Patterns of SNP and presence–absence diversity in *E. coli*.

(A) Site frequency spectra of loss-of-function (LOF; purple), missense (yellow), and synonymous (green) mutations. Allele count represents the number of non-reference alleles present in the dataset for each mutation type.

(B) Distribution of contig presence–absence variation. The x-axis captures how many of 7,043 samples had a contig labeled as 'present' based on a normalized coverage cutoff threshold.

Next, we turned our attention to patterns of sequencing coverage in our dataset, which can also be informative about technical artifacts. *E. coli* strains possess a 'core genome' (a set of genes broadly shared by all taxa) and an "accessory genome" (genes that vary across taxa, strain/phylogroup/etc.), the structure of which can vary broadly. By analyzing the relative coverage of DNA segments (contigs) across the 7,000 genomes, we can reconstruct the presence of the core and accessory genomes in the dataset and compare these patterns to previously published work on the structure of the *E. coli* genome. Of 32,441 contigs, we found 2,847 were shared by all 7,043 samples, 687 were shared by all but one, and 287 were shared by all but two samples, thus reflecting a broadly shared core genome. On the other hand, many contigs were shared by a small subset of samples in the dataset (Figure 1, B), highlighting the accessory genome. These patterns align with previous estimates of the *E. coli* core genome size and presence–absence frequency. These observations further indicate that the *E. coli* 7k data and reference pan-genome are of high quality **[7][23]**.

Finally, we examined genome-wide patterns in the dataset to see if we identified expected *E. coli* population-level differences. A genomic PCA showed that genomes were broadly differentiated along axes defined by the 72 ECOR reference strains used to assemble the pangenome (Figure 2, A, Supplemental Table 1). Notably, there was little association between genomic diversity (represented by PC axes 1–5) and year of collection (pseudo r^2 = 0.05; multinomial regression; Figure 2, B). There was a similarly weak relationship between these genomic principal components and country of isolation (pseudo r^2 = 0.25; Figure 2, C). This tracks with previous work that has found phylogenetically distinct strains of *E. coli* co-localizing globally **[24][25]**, a pattern which likely obscures more subtle within lineage isolation by distance patterns.

However, multilocus sequencing type (MLST) — a commonly used taxonomic identifier of *E. coli* lineages — and broader phylogroup labels were strongly associated with genomic diversity (pseudo r^2 = 0.86; Figure 2, C). Furthermore, combining MLST/phylogroup with genomic diversity (PCs 1–5) strongly predicted the country of isolation (linear regression; pseudo r^2 = 0.89). This indicates that broad-scale population differences (i.e., MLST/phylogroup) co-occur with more recent, finer-scale geographic variation (i.e., country of origin). These results indicate that the *E. coli* 7k dataset represents various evolutionary scales and patterns.

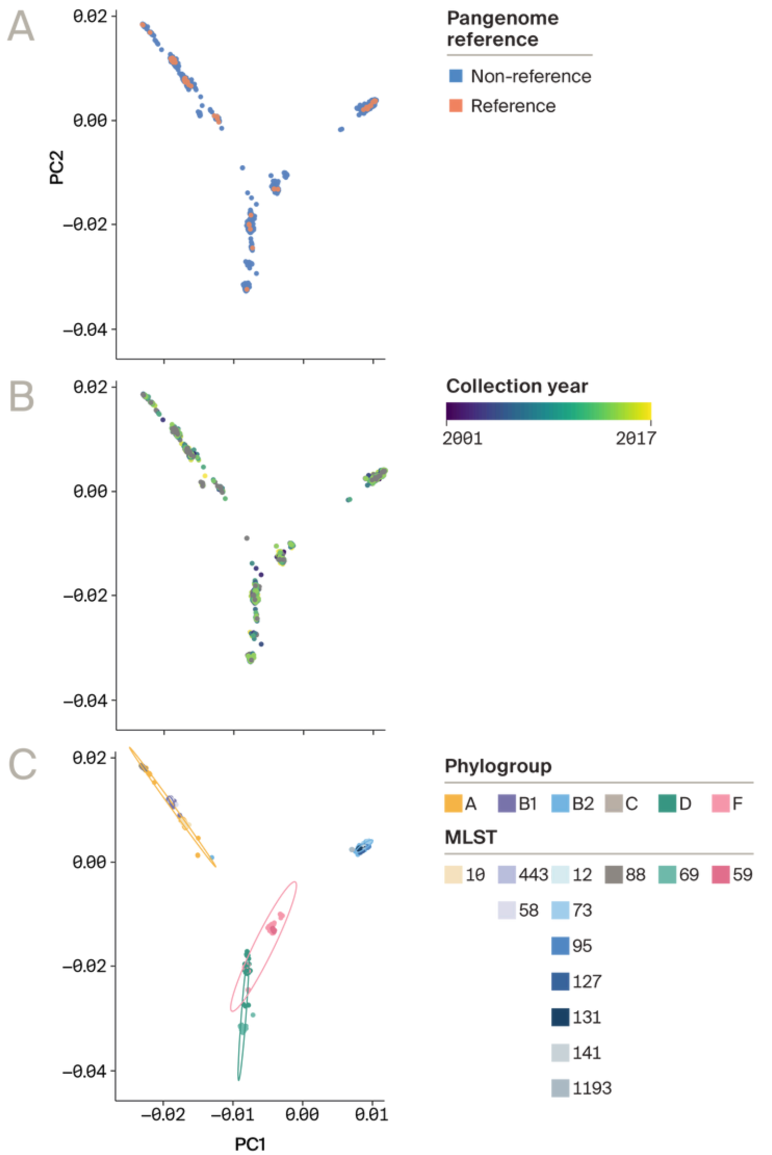

Figure 2

PCA of 7,000 *E. coli* genotypes with visualization of various metadata features.

(A) Labeling of samples included in the construction of the pan-genome reference.

(B) Labeling of samples by year of isolation.

(C) Labeling of samples by either broader phylogroup or multilocus sequence type (MLST). MLST labels are nested under the broader phylogroup to which they belong.

Phylogenetic analysis

To further interrogate these patterns, we inferred a core-genome phylogeny and analyzed the distribution of major phylogroups and MLSTs (Figure 3, A). As expected, MLSTs formed clades nested within deeply branching phylogroups (Figure 3, A). For example, MLST 131 is a sub-lineage of phylogroup B2 and correctly appears as such in our phylogeny **[26]**. We did note that some samples labeled as phylogroup F appeared within phylogroup D (Figure 3, A). These samples lack signals of phylogenetic (e.g., excessively long branch lengths) or bioinformatic error (unexpected SFS patterns), suggesting that they're incorrectly labeled phylogroup F in the metadata. Overall, the phylogenetic patterns match previous work and recover expected relationships both within strains and among major phylogroups **[8]**.

We next used the phylogeny to infer how the three AMR phenotypes (ampicillin, trimethoprim/sulfamethoxazole, ciprofloxacin) have diversified over time (Figure 3, B). We wanted to know if these phenotypes have followed the same pattern or if they represent different diversification modes. The answer would help us gauge how much power the *E. coli* 7k dataset contains for modeling genotype–phenotype relationships over different evolutionary scales.

Ampicillin and trimethoprim/sulfamethoxazole resistance were fairly evenly distributed across the tree (Figure 3, B). On the other hand, ciprofloxacin was restricted to specific clades, potentially representing multiple independent transitions between antibiotic susceptibility and resistance (Figure 3, B). Of the 433 unique clade labels (MLST or broader phylogroups) in our dataset, only 70 (16%) contained at least one resistant ciprofloxacin observation. Ampicillin and trimethoprim/sulfamethoxazole were more broadly distributed: 165/112 (38%/26%) clades contained at least one observation, respectively. These observations suggest two types of phenotypic distribution in the dataset: broad (ampicillin, trimethoprim/sulfamethoxazole) and clade-restricted (ciprofloxacin).

The two diversification types may be associated with different genetic architectures. For example, closely related strains exhibit different AMR phenotypes in the broad distribution, suggesting that just a few mutations may be needed to evolve resistance. Encouragingly, this type of recurrent diversification helps control for population differences, allowing causal loci to be decoupled from the genomic background. On the other hand, the putatively independent, repeated evolutionary origins of ciprofloxacin resistance may point to a common mutational pathway through which this trait has evolved among distinct *E. coli* strains. We thus sought to identify which loci contribute to AMR's evolution and the extent to which these contributions persist across evolutionary scales.

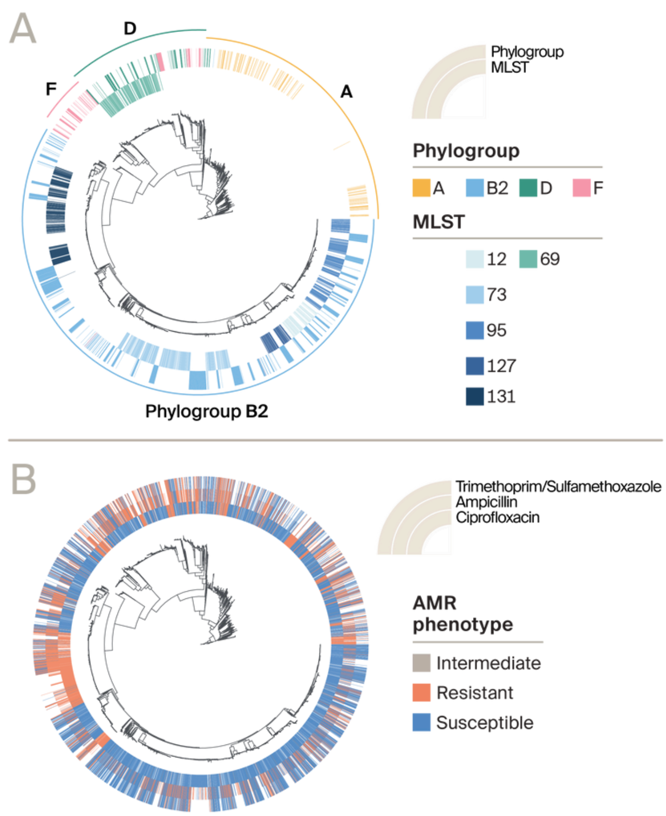

Figure 3

Species tree constructed from sub-sampled synonymous sites (minor-allele count: MAC > 10) in the core genome.

(A) Species tree constructed from sub-sampled synonymous sites (minor-allele count: MAC > 10) with Phylogroup/MLST labels, only groups with at least 100 samples shown.

(B) Species tree constructed from sub-sampled synonymous sites (MAC > 10) with three focal AMR phenotypes and their state distributions.

Genomic prediction overview

In the previous sections, we found that the *E. coli* 7k dataset has several desirable features for downstream model development. The dataset lacks clear evidence of technical and biological noise, encompasses a substantial portion of *E. coli*'s global diversity, and contains genomic and phenotypic diversity spanning multiple evolutionary scales. Given these positive signs, we were next interested in probing the genetic architectures of three AMR phenotypes.

Using linear genomic prediction methods, we inferred the size, structure, magnitude, and heritability of AMR-associated genetic markers (Figure 4). As previously discussed, these methods often fail to capture nonlinear processes like epistasis. However, by estimating how much phenotypic variation can be explained solely by additive interactions, linear models can be useful for establishing a predictive baseline. For example, all three phenotypes had very high estimates of narrow sense heritability (proportion of variance explained: 0.99, 0.95, 0.94 for ciprofloxacin, ampicillin, and trimethoprim/sulfamethoxazole resistance, respectively), indicating that these traits have relatively simple genetic architectures that might allow genomic prediction models to capture a majority of their phenotypic variance.

This section provides an in-depth exploration of these genomic prediction results. Please jump to the Key takeaways and Next steps for a quicker overview.

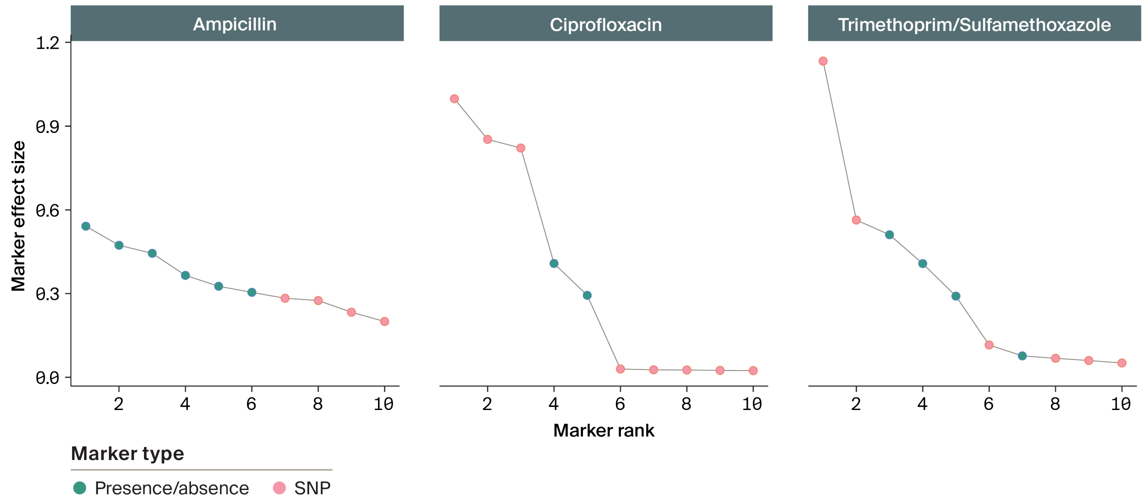

Figure 4

Effect size distribution for the ten largest-effect markers from a genomic prediction analysis on three AMR phenotypes.

Effect size is measured as the sum of minor and major effects estimated by GEMMA's BSLMM model. Markers are labeled as either SNP or presence–absence variation.

Ciprofloxacin resistance

The three markers with the largest effect sizes strongly predicted ciprofloxacin resistance. These markers were two missense SNPs in the gyrA gene (gyrA248, gyrA259 leading to substitutions Ser83Leu, Asp87Asn/His/Tyr respectively), and one missense SNP in the topoisomerase IV subunit A (ParC239, substitution Ser80Ile) (Figure 4, Supplemental Table 2). All three substitutions have been previously identified as resistance mutations in lab studies **[27]** and likely underlie the vast majority of ciprofloxacin resistance in our dataset. We constructed a gene tree based on gyrA and mapped the occurrence of all three resistance markers onto it to check if these substitutions have evolved independently more than once. The distribution of the resistance phenotype on this gene tree suggests a dynamic evolutionary history, with resistance to ciprofloxacin being repeatedly gained and/or lost through mutation in the core genome (Figure 5, A).

As we're ultimately interested in nonlinear genotype–phenotype modeling, we looked further into possible interactions between the three key ciprofloxacin resistance substitutions, given their multiple independent origins and abundance in our dataset. To test for first-order interactions between the three substitutions, we fit a logistic regression predicting resistance phenotype using individual marker-state and all pairwise (N = 9) marker interaction terms possible (Supplemental Table 3). For the singular marker terms, the logistic regression recapitulated the effect size ranking of our initial analysis [gyrA248 (effect size = 3.24) > gyrA259 (2.64) > parC239 (2.11)]. None of the interaction terms significantly differed from 0 in the fitted model, likely due to the low number of observations of intermediate genotypes. However, the model did predict a positive (albeit insignificant) interaction between gyrA248 and gyrA259(p = 0.12, effect size = 2.29), a finding that's supported by observations of ciprofloxacin resistance in the lab **[27]**, and by plotting resistance phenotype distributions by genotype (Figure 5, B).

While the tight associations between the resistance markers for ciprofloxacin hamper model fitting, they are in themselves informative since an overabundance of certain combinations (i.e., LD) implies they're more fit than others. Consequently, we expect that mutational trajectories from non-resistant wild-type genotypes to antibiotic-resistant genotypes should disproportionately pass through these favorable genotypic combinations, avoiding unfit genotypes. To test these predictions, we fit models of discrete trait evolution to infer transition rates between different genotypes, thus obtaining estimates of the relative probabilities of different mutational trajectories between ciprofloxacin susceptible/resistant strains. We accomplished this using corHMM(v2.8) **[21]** to estimate all possible single-step transition rates between the presumed ancestral state (the reference allele at all three positions) to the final full mutation stack resistance phenotype.

This analysis suggested high reverse mutation rates, particularly in mutational states involving one or two resistance mutations. This likely occurs because such genotypes are generally rare in our dataset (and difficult to sample in any dataset of this size) and distributed within clades with very short internal branch lengths, where the true fine-scale phylogenetic relationships are hard to estimate. As a result, we limit our focus on biologically plausible forward, single-step mutational rates estimated in corHMM.

These rates suggest that the most likely single-step mutational pathway toward resistance first requires a mutation at gyrA248, followed by rapid mutation at either of the two following sites (Figure 5, C, Supplemental Table 4). This lines up well with the phenotypic data, as gyrA248 is the only mutation that appears to show some level of resistance when found alone (Figure 5, B) and is the only mutation that's found at

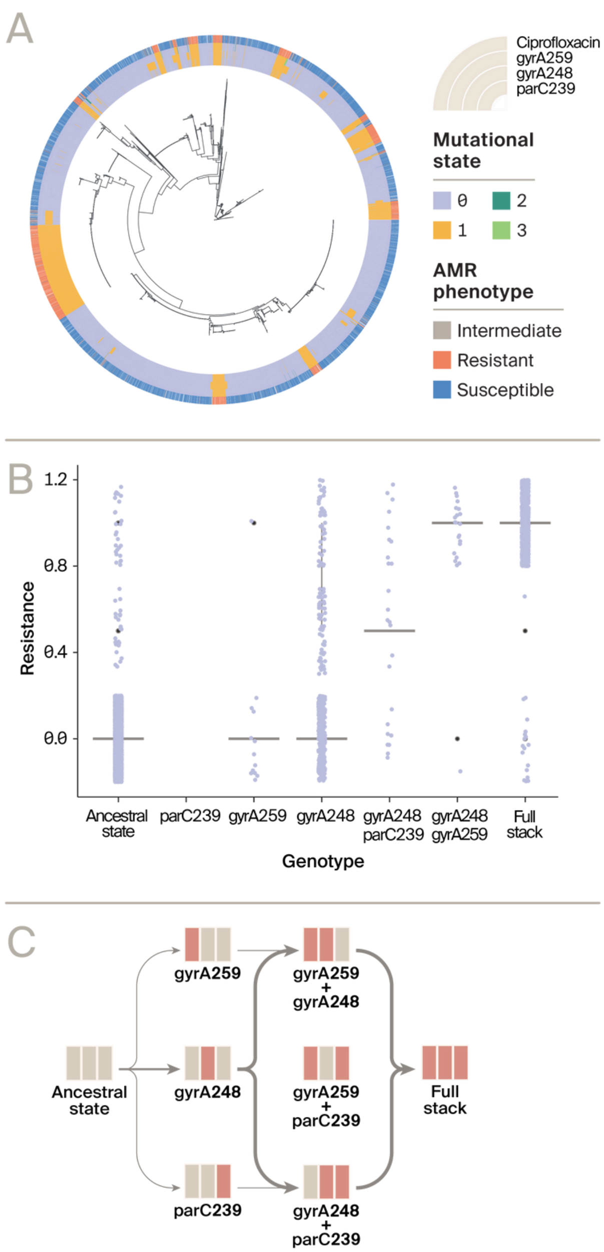

Figure 5

Evolutionary history of three major ciprofloxacin resistance mutations.

(A) Gene tree of gyrA locus with mutational state at three resistance SNPs (inner 3 rings: gyrA248, gyrA259, parC239), as well as ciprofloxacin resistance phenotype state (outer ring). Mutational state of 0 denotes ancestral allele, and mutational state of 1,2,3

appreciable frequency by itself, further suggesting that it's the most likely initial mutational step. Once this mutation arises, it seems to potentiate ciprofloxacin resistance evolution as further accumulation of the next two mutations achieves full resistance (Figure 5, B and C).

Overall, our results for genomic prediction in ciprofloxacin point to a fairly simple genetic architecture dominated by a series of three ordered mutations but reveal more subtle signatures of epistasis that are easily missed in our initial linear genomic prediction analysis. This provides a nice baseline for future nonlinear modeling work on this phenotype. Simultaneously, however, these results point to the disadvantage of using natural datasets. Epistasis naturally quickly creates LD between mutations, thereby purging unfavorable genotypes from a population **[28]**. However, this hampers model fitting, as unfavorable genotypes (i.e., “true negatives”) are needed to train models aiming to connect phenotype to genotypes. This highlights the advantage of studies with controlled crosses where both fit and unfit genotypes can be observed and phenotyped.

denote various derived resistance alleles.

(B) Distribution of ciprofloxacin resistance phenotypes as a function of various combinations of ancestral/derived mutations at three resistance sites. For simplicity, all derived alleles are considered equivalent. Resistance values were encoded as 0 = susceptible, 0.5 = intermediate, and 1 = resistant.

(C) Graph of transition rate estimates between genotypes going from ancestral state of no resistance mutations, to full resistance across three resistance sites. Rates were estimated using corHMM; only single-step forward transition rates are illustrated for interpretability. Width of edges corresponds to transition rate from the source (left) to target (right) genotype.

Ampicillin resistance

Unlike ciprofloxacin, ampicillin resistance was characterized by a gentler decay of marker effect size (Figure 4). These resistance markers mapped to an assortment of presence–absence loci corresponding to putative plasmid and transposon fragments in the pan-genome (Supplemental Table 2). The marker with the largest effect size in our results is a class A beta-lactamase (TEM-1) fragment based on BLASTx hits (e.g., 74% query cover and 93.3% sequence identity with an *Enterobacter hormaechei* class A beta-lactamase), a reasonable resistance locus for ampicillin resistance **[29]**. The remaining markers were enriched for Tn3 transposon family components (Supplemental Table 2). Such transposons often harbor beta-lactam genes associated with resistance evolution **[16]**.

Most markers associated with ampicillin resistance in our dataset appear to map to genomic components that are likely physically linked to causal resistance loci (such as beta-lactams) rather than resistance loci themselves, a consequence of the fragmented pan-genome we're using. To overcome this limitation, we looked at patterns of LD between our ten largest-effect markers to see if we could find evidence of linkage among them. We found clear signals of enriched LD among the six markers with the largest effect size, especially the largest-effect marker, a beta-lactam gene, and the marker with the third-largest effect size, a Tn3 transposon fragment (Figure 6). The observation of moderate/high but not near perfect (> 0.9 r^2) LD among these markers likely reflects complex patterns of physical linkage between them whereby they're likely linked in some parts of the phylogeny but not in others.

Our results point to a distinct genetic architecture for ampicillin resistance that involves the acquisition of any of a variety of accessory genome components rather than specific core genome mutations, as in the case of ciprofloxacin resistance. While our analysis only captures one previously validated causal resistance locus, we can still conclude that ampicillin resistance generally arises via plasmid and transposon resistance locus acquisition in our dataset, a finding corroborated by previous research on the occurrence of resistant beta-lactam genes **[30]**.

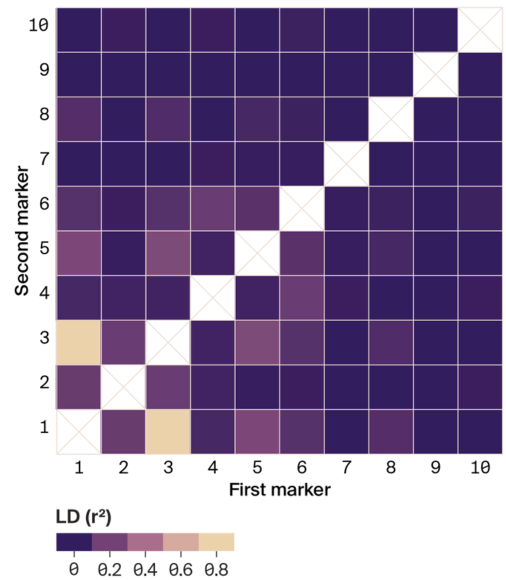

Figure 6

Pairwise linkage disequilibrium (r^2) among the ten largest-effect ampicillin resistance markers.

Trimethoprim/sulfamethoxazole resistance

Markers of varying effect sizes characterized trimethoprim/sulfamethoxazole resistance (Figure 4). Similar to the case of ciprofloxacin, one marker had a particularly large effect size (the largest in any of our genomic-prediction analyses). The top ten largest-effect markers were a mixture of SNPs and presence–absence markers (six and four markers, respectively, a pattern that is intermediate to the results for ciprofloxacin and ampicillin (Supplemental Table 2).

Our list of the top ten largest-effect markers seemed to be enriched for AadA (aminoglycoside adenylyltransferase) and GNAT (GCN5-related N-acetyltransferases) family proteins. The marker with the largest effect size is a SNP located on a short contig (160 bp), which has multiple significant BLASTx hits to AadA family proteins (e.g., 99% query cover and 100% sequence identity to an AadA in *Pseudomonas gessardii*) in addition to hits to partial nucleotidyltransferase domain-containing proteins. This wasn't the only association with AadA proteins that we found. The marker with the second-largest effect size (a SNP) also had significant BLASTx hits to AadA1/ANT(3”) (99% query cover and 100% sequence identity to AadA1 in *E. coli*). Finally, the marker with the fourth-largest effect size (a presence–absence marker) had significant BLASTx hits to a GNAT family protein (80% query cover and 86% sequence identity to a *Klebsiella pneumoniae* putative aminoglycoside N(6')-acetyltransferase (AAC(6”)).

Functionally, this enrichment for AadA and GNAT family proteins is perplexing. While these protein families do indeed play crucial roles in antibiotic resistance, their mechanism of action is the enzymatic modification of aminoglycoside family antibiotics, which doesn't include either trimethoprim or sulfamethoxazole **[31]**. The association between trimethoprim/sulfamethoxazole resistance and the largest-effect marker was particularly strong and statistically robust. We considered the possibility that the phenotypic data was mislabelled, checking both the source of the data (BV-BRC) and the underlying studies **[32]**, but we found no evidence of phenotype data errors. We also didn’t find a strong phenotypic correlation between trimethoprim/sulfamethoxazole and aminoglycoside antibiotics like gentamycin (r^2 = 0.20) in our dataset.

While our genomic prediction analyses did a good job predicting trimethoprim/sulfamethoxazole resistance, we couldn't link the largest-effect markers to putative resistance loci. We suspect this is chiefly driven by the fragmented nature of the pan-genome we use for this dataset. As resistance is often acquired through plasmids/transposons **[16][33]**, and such contigs are poorly assembled in our pangenome, we might struggle to find candidate resistance loci among these fragments. It could well be that an unassembled causal resistant *dfrA* locus **[34]** is linked to the largest effect *aadA* gene we observe as predictive of trimethoprim/sulfamethoxazole resistance, leading to our confusing results. This isn't an unlikely hypothesis given the fact that resistance genes for multiple different types of antibiotics are known to be stacked within single plasmids/transposon **[33]**. This result implies that while we likely can predict most AMR phenotypes very well with both linear and nonlinear models in this dataset, the interpretability of findings may be challenging in some instances.

Key takeaways

- We performed exploratory analyses in the recently published *E. coli* 7k dataset

- The *E. coli* 7k dataset captures the global genomic diversity of *E. coli* and captures both fine- and broad-scale diversity across evolutionary scales
- Genomic prediction analyses identified expected causal AMR loci for ciprofloxacin and ampicillin but no interpretable genomic resistance targets for trimethoprim/sulfamethoxazole
- Follow-up analyses demonstrate that while the overall genetic architecture of resistance is often simple, it nonetheless can be dependent on more subtle epistatic interactions

Next steps

Dataset availability currently limits the creation of realistic genetic models that can account for linear *and* nonlinear phenomena. In this pub, we stress-tested the *E. coli* 7k dataset for such modeling applications.

The *E. coli* 7k dataset lacks common biological and technical error signals. For example, we detected purifying selection, a stable core genome, and isolation by distance in genetic similarity among samples. Phylogenetic analyses indicated the presence of diverse evolutionary scales: deeply branching phylogroups and rapidly diversifying strains. Antibiotic resistance has evolved in multiple ways across this tree, displaying both broad and clade-restricted distributions. Together, these results provide confidence in the quality of the *E. coli* 7k dataset and confirm its suitability for model development.

Linear genomic prediction methods were able to confidently predict AMR phenotypes. However, the interpretability of these results varied. For example, we identified three epistatic mutations underlying ciprofloxacin resistance in follow-up analyses. Ampicillin resistance was also associated with interpretable loci, particularly plasmid and transposon components likely linked to resistance genes. On the other hand, we found no obvious link between loci and potential resistance mechanisms for trimethoprim/sulfamethoxazole. These results provide a suitable baseline for comparison as more complex models are developed.

We note some outstanding issues. First, the *E. coli* 7k dataset is centered on a fragmented pan-genome, which makes it challenging to link genetic markers with AMR phenotypes functionally. Second, it’s possible that the genomic prediction methods were underpowered because this is a natural population; selection will have eroded unfit but informationally rich genotypes that could be uncovered in other contexts.

Overall, our findings set the stage for us to exploit this dataset to guide the use of more complex nonlinear genomic prediction models. However, it's useful to consider when and where we expect nonlinear models to provide an edge over linear genomic prediction models. We hypothesize likely candidates for nonlinear models are populations in which epistasis or gene-by-environment interactions are prevalent. For example, the simplest form of epistasis — two-locus interactions — will be most powerful in populations with intermediate (~0.5) allele frequencies **[2]**. This requirement is most likely to be met in highly structured populations — especially products of artificial crosses or selection — or where mutations are fixed between evolutionarily diverged lineages **[35][36]**. Gene-by-environment interactions are more complicated to predict and are likely implicated in complex, highly polygenic traits sensitive to environmental conditions (e.g., agronomic yield **[37]**). Datasets in which phenotypes were measured in the field are likely candidates for approaching gene-by-environment signals. It'll be useful to continue developing intuitions for which model architectures will best capture these various processes of diversification, pushing our ability to extract inference in contexts where our knowledge of the genotype–phenotype map is much less understood.

References

1 Mets DG, Morin M. (2024). Creating a 7,000-strain E. coli genotype dataset with antimicrobial resistance phenotypes. https://doi.org/10.57844/ARCADIA-D2CF-EBE5

2 Mackay TFC. (2013). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. https://doi.org/10.1038/nrg3627

3 Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. https://doi.org/10.1038/s41467-019-11112-0

4 Avasthi P, Mets DG, York R. (2024). Harnessing genotype-phenotype nonlinearity to accelerate biological prediction. https://doi.org/10.57844/ARCADIA-5953-995F

5 Mets DG, York R. (2024). Applying information theory to genetics can better explain biological phenomena. https://doi.org/10.57844/ARCADIA-53F4-DA1A

6 Poirel L, Madec J-Y, Lupo A, Schink A-K, Kieffer N, Nordmann P, Schwarz S. (2018). Antimicrobial Resistance in*Escherichia coli*. https://doi.org/10.1128/microbiolspec.arba-0026-2017

7 McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, Vehkala M, Välimäki N, Prentice MB, Ashour A, Avram O, Pupko T, Dobrindt U, Literak I, Guenther S, Schaufler K, Wieler LH, Zhiyong Z, Sheppard SK, McInerney JO, Corander J. (2016). Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations. https://doi.org/10.1371/journal.pgen.1006280

8 Denamur E, Clermont O, Bonacorsi S, Gordon D. (2020). The population genetics of pathogenic Escherichia coli. https://doi.org/10.1038/s41579-020-0416-x

9 Zhou Z, Charlesworth J, Achtman M. (2020). Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. https://doi.org/10.1101/gr.260828.120

10 Ochman H, Selander RK. (1984). Standard reference strains of Escherichia coli from natural populations. https://doi.org/10.1128/jb.157.2.690-693.1984

11 Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. https://doi.org/10.1093/bioinformatics/bts606

12 Venables WN, Ripley BD. (2002). Modern Applied Statistics with S. https://doi.org/10.1007/978-0-387-21706-2

13 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. https://doi.org/10.1093/molbev/msaa015

14 Lewis PO. (2001). A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. https://doi.org/10.1080/106351501753462876

15 Kreiner JM, Sandler G, Stern AJ, Tranel PJ, Weigel D, Stinchcombe JR, Wright SI. (2022). Repeated origins, widespread gene flow, and allelic interactions of target-site herbicide resistance mutations. https://doi.org/10.7554/elife.70242

16 Partridge SR, Kwong SM, Firth N, Jensen SO. (2018). Mobile Genetic Elements Associated with Antimicrobial Resistance. https://doi.org/10.1128/cmr.00088-17

17 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. (2021). Twelve years of SAMtools and BCFtools. https://doi.org/10.1093/gigascience/giab008

18 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. https://doi.org/10.1086/519795

19 Zhou X, Carbonetto P, Stephens M. (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. https://doi.org/10.1371/journal.pgen.1003264

20 Zhou X, Stephens M. (2012). Genome-wide efficient mixed-model analysis for association studies. https://doi.org/10.1038/ng.2310

21 Beaulieu JM, O'Meara BC, Donoghue MJ. (2013). Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. https://doi.org/10.1093/sysbio/syt034

22 To T-H, Jung M, Lycett S, Gascuel O. (2015). Fast Dating Using Least-Squares Criteria and Algorithms. https://doi.org/10.1093/sysbio/syv068

23 Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT. (2011). Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. https://doi.org/10.1073/pnas.1015622108

24 Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EP. (2020). Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. https://doi.org/10.1371/journal.pgen.1008866

25 Stoppe N de C, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM, Torres TT. (2017). Worldwide Phylogenetic Group Patterns of Escherichia coli from Commensal Human and Wastewater Treatment Plant Isolates. https://doi.org/10.3389/fmicb.2017.02512

26 Nicolas-Chanoine M-H, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park Y-J, Lavigne J-P, Pitout J, Johnson JR. (2007). Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15. https://doi.org/10.1093/jac/dkm464

27 Shariati A, Arshadi M, Khosrojerdi MA, Abedinzadeh M, Ganjalishahi M, Maleki A, Heidary M, Khoshnood S. (2022). The resistance mechanisms of bacteria against ciprofloxacin and new approaches for enhancing the efficacy of this antibiotic. https://doi.org/10.3389/fpubh.2022.1025633

28 Kimura M, Maruyama T. (1966). THE MUTATIONAL LOAD WITH EPISTATIC GENE INTERACTIONS IN FITNESS. https://doi.org/10.1093/genetics/54.6.1337

29 Cooksey R, Swenson J, Clark N, Gay E, Thornsberry C. (1990). Patterns and mechanisms of beta-lactam resistance among isolates of Escherichia coli from hospitals in the United States. https://doi.org/10.1128/aac.34.5.739

30 Hussain HI, Aqib AI, Seleem MN, Shabbir MA, Hao H, Iqbal Z, Kulyar MF-A, Zaheer T, Li K. (2021). Genetic basis of molecular mechanisms in β-lactam resistant gram-negative bacteria. https://doi.org/10.1016/j.micpath.2021.105040

31 Doi Y, Wachino J, Arakawa Y. (2016). Aminoglycoside Resistance. https://doi.org/10.1016/j.idc.2016.02.011

32 Rafique M, Potter RF, Ferreiro A, Wallace MA, Rahim A, Ali Malik A, Siddique N, Abbas MA, D'Souza AW, Burnham C-AD, Ali N, Dantas G. (2020). Genomic Characterization of Antibiotic Resistant Escherichia coli Isolated From Domestic Chickens in Pakistan. https://doi.org/10.3389/fmicb.2019.03052

33 Nikaido H. (2009). Multidrug Resistance in Bacteria. https://doi.org/10.1146/annurev.biochem.78.082907.145923

34 Sánchez-Osuna M, Cortés P, Llagostera M, Barbé J, Erill I. (2020). Exploration into the origins and mobilization of di-hydrofolate reductase genes and the emergence of clinical resistance to trimethoprim. https://doi.org/10.1099/mgen.0.000440

35 Phillips PC. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. https://doi.org/10.1038/nrg2452

36 Hill WG, Goddard ME, Visscher PM. (2008). Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. https://doi.org/10.1371/journal.pgen.1000008

37 Malosetti M, Ribaut J-M, van Eeuwijk FA. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. https://doi.org/10.3389/fphys.2013.00044