Creating a 7,000-strain *E. coli* genotype dataset with antimicrobial resistance phenotypes

We assembled a comprehensive *E. coli* antimicrobial resistance phenotype-genotype resource. This dataset will aid large-scale genetic studies on anti-microbial resistance and support research in phylogenetics and other fields.

Version 3, published Mar 24, 2025. Originally published Aug 22, 2024.

Arcadia Science

DOI: 10.57844/arcadia-d2cf-ebe5

Purpose

Mapping the connections between genotypes and phenotypes is central to understanding the principles governing biology and its innovations. One of Arcadia's long-term goals is to improve methods and datasets for these types of analyses. The increasing availability of genetic and phenotypic data from many different *Escherichia coli* infections presents an opportunity to investigate genotype–phenotype relationships with particular relevance to anti-microbial resistance (AMR). Additionally, the genetic diversity observed within this species is almost comparable to the diversity found between species, offering a unique opportunity to understand how genetic variation drives phenotypic diversity across different levels of biological organization. Therefore, using existing data, we produced a large-scale dataset to serve as a testbed for genotype–phenotype prediction models, exploring gene-gene interactions and establishing frameworks for bridging genetic analyses within and between species.

We constructed our dataset using the wealth of sequenced microbial genomes and associated phenotypes, including AMR. Specifically, from 72 well-studied and heavily sequenced *E. coli* strains, we assembled a pangenome, capturing the total genome content from these species, including the presence or absence of genetic material. We then aligned sequence data from ~7,000 *E. coli* strains to this pangenome and performed variant calling to identify genetic variation across strains. We found significant genetic diversity, identifying 2.4 million variants. To validate this dataset, we measured the association of variants in genes known to impact AMR with AMR phenotypes within the dataset. We successfully correlated genetic variation in known AMR-related genes to AMR phenotypes.

This work will be of particular interest to geneticists and evolutionary biologists. It may also be valuable for microbiologists and epidemiologists who are trying to develop more effective treatment and prevention strategies for bacterial infections.

- This pub is part of the **platform effort**, "Genetics: Decoding evolutionary drivers across biology." Visit the platform narrative for more background and context.
- You can find our **code**, including the code to generate the *E. coli* pangenome, the main pipeline to perform variant-calling, and the notebooks to further analyze the dataset, in this GitHub repository.
- The **data**, including the list of genome accession numbers for the pangenome, the SRA accession numbers of the strains in the dataset, and the output of the variant-calling process, are on Zenodo.

Background and goals

For over 100 years, mapping genetic variation to phenotypes has been a focus in fields such as agriculture, human health, and evolutionary biology. The development of high-throughput sequencing has significantly increased the amount of genetic information available, including data from entire populations and thousands of individuals. This expansion has accelerated efforts to map genetic variation with phenotypes, uncovering complex relationships between specific genetic changes and their associated traits. It has also facilitated the development of more accurate genotype–phenotype predictions in both eukaryotes and prokaryotes, enabling researchers to identify critical genetic markers linked to various biological functions and disease states [1][2].

Genotype-phenotype studies rely on managing complex, heterogeneous, and multidimensional data to connect genotype information to phenotypes. Simply assembling an appropriate dataset can be difficult and is complicated by an organism's genetic complexity and incomplete genetic or phenotypic information. In this work, we aimed to build a high-quality dataset as a resource for investigators developing analytical frameworks in the genotype–phenotype space, including genotype–phenotype predictions, identification of epistasis, and new approaches bridging within and between species analysis.

With comparatively simple genome structure and frequently complete genotype and phenotype information, bacteria present an opportunity to build a large, well-standardized dataset. Research deciphering the genetic underpinnings of traits like antibiotic resistance, pathogenicity, and metabolism has resulted in well-documented microbial genomes and phenotypes. Specifically, *E. coli* is an extensively studied, occasionally pathogenic bacterium, with some variants posing significant healthcare challenges due to the rise of anti-microbial resistance (AMR). As a result, clinical studies have tracked infection outbreaks, sequenced strain genomes, and documented measured or predicted antimicrobial resistance phenotypes for large populations.

Extensive *E. coli* research has demonstrated remarkable genetic diversity. Only 20 to 40% of the genome is present across strains (the species' core genome), while the presence of the remainder varies **[3]** **[4]**. This diversity means that, in a large population of *E. coli* strains, we can study genetic variation at both the nucleotide level (typical of within-species analyses) and the larger scale of gene presence-absence, frequently used when comparing species. Applying both types of analyses could lead to a more comprehensive understanding of how genetic variation drives phenotypic diversity within and between species.

In this work, we aimed to consolidate genotype–phenotype information of *E. coli* from a public database, mapping genetic diversity and variations to AMR phenotypes. We conducted variant calling on nearly 7,000 sequenced strains using a custom *E. coli* pangenome and correlated this data with AMR profiles. We've shared the dataset here. We hope it'll serve as a resource for large-scale genotype–phenotype studies and provide a benchmark dataset for developing new methods. While it may allow researchers to understand AMR mechanisms better and investigate genetic interactions, the dataset should also enable novel phenotype–phenotype prediction and phylogenetic research analyses.

The approach

Our primary objective was to create a dataset integrating available genotypic and phenotypic *E. coli* data to allow mapping between genetic variation and known antimicrobial resistance (AMR) phenotypes. We first identified an *E. coli* cohort with documented AMR phenotypes and available genomes. We then characterized the phenotype distribution within our cohort. Next, given the genetic diversity of *E. coli*, we constructed a pangenome from 72 especially well-studied strains and used it to conduct variant calling across all 7,057 strains, including these initial 72. Finally, we correlated the identified genetic variations with the known AMR phenotypes.

Generating the reference genome

Selecting an appropriate reference genome is essential for genotype–phenotype analyses that leverage precise genomic locations of genetic variants. The genome is the shared reference for genetic variation across strains and, thus, must provide comprehensive coverage across strains and accurately represent the genetic diversity of the cohort [5].

E. coli exhibits high genetic diversity among its strains in terms of single-nucleotide variation and, different from many eukaryotic species, the presence or absence of large portions of the genome [3]. We, therefore, need a reference genome encompassing this global diversity. We decided to generate a pangenome using the genomes of the ECOR collection [6], which consists of 72 *E. coli* strains isolated from a wide variety of hosts and geographical locations, including strains from different phylogenetic groups. This collection offers a broad representation of the natural diversity of the species.

Much pangenome analysis has focused on coding genes and excluded intergenic regions (IGRs). However, IGRs are essential for gene regulation and mediating gene-gene interactions. Therefore, we included IGRs in our pangenome assembly. To construct the pangenome for this study, we first obtained the sequenced genome for all 72 strains in the ECOR collection from [7] using the Batch Entrez API and the strains' GenBank accession numbers. Using the default parameters, we annotated the genomes with Prokka (version 1.14.6) [8]. Next, we generated the pangenome of coding genes with Roary (version 3.13.0) [9] using a 90% identity threshold to cluster the protein sequences and performing within-cluster alignments to identify the reference sequence in each cluster. Finally, we generated the pangenome of IGRs with Piggy (version 1.5) [10], using the program defaults parameters, and we combined both pangenomes into a single FASTA file (whole_pangenome.FASTA on Zenodo).

This final file contains the collective genetic content of the ECOR collection and served as the “reference genome” for this study. It contains 18,494 coding gene sequences, including 2,652 core genes found in 99% of the ECOR collection species and 13,947 IGRs. Altogether, the genome comprises 32,441 sequences, which we call contigs (or pangenome contigs). Information about the presence or absence of contigs in strains is in the file whole_pan_ecor_presence_absence.csv on Zenodo.

Finally, we conducted functional annotation of the pangenome's coding sequences using the eggNOG-mapper (version 2.1.12) web interface [11].

Selecting the working dataset

We used the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) to identify the study cohort and compile our dataset of available *E. coli* genomes and associated antimicrobial resistance phenotypes. We conducted two separate searches: first, we identified strains with genomic sequence data available on the Sequence Read Archive (SRA), and second, we collected AMR phenotype data from BV-BRC. By intersecting the results of these searches, we identified 6,985 *E. coli* strains for which genomic sequencing data were available from the SRA, and phenotypic information was available for at least one antibiotic.

Identifying genetic variants

To enable genotype–phenotype mapping, we needed to identify genetic variation across the cohort of *E. coli* strains. This procedure included multiple steps. First, we created a reference genome (see above) that we could use to identify single-nucleotide polymorphisms and the presence or absence of large portions of genetic material. Then, after downloading the available genome sequence data, we determined the allelic state for each strain at each genomic location that varies between strains. Finally, we created a genotype matrix containing each strain's predicted allelic state at each variant location.

Downloading sequencing reads from the Sequence Read Archive

We obtained the SRA accession numbers for the sequencing files of the 7,057 selected strains (6,983 strains from the original cohort and 72 strains from the ECOR collection). Using the GNU parallel shell tool [12] and the `faster-dump` tool from the sra-toolkit [13], we downloaded the FASTQ files from paired-end sequencing for each strain.

Variant calling in 7,057 samples

Variant calling identifies genetic differences between a strain and a reference genome by aligning genomic sequencing reads from the strains against the reference and identifying where they differ. These differences are then filtered based on the likelihood that they're true variants versus sequencing errors and compiled into a variant call format (VCF) file.

We performed variant calling independently for each strain from our cohort (6,985 strains) and the 72 ECOR strains. Our workflow proceeds as follows: First, we used fastp [14] to perform quality control, remove sequences corresponding to sequencing primers and remove reads shorter than 30 nucleotides or quality scores below a threshold (Phred score below 30 across a sliding window of four bases). Subsequently, we concatenated the resulting FASTQ files (one for read one and one for read two; all samples were “paired-end” read data) into a single file, which we then aligned against the reference pangenome (`BWA mem` [15]). We sorted the resulting alignment file and marked duplicates using SAMtools (version 1.20) [16]. Using Picard (version 2.27.5) [17], we added read group tags (RG tags), incorporating the strain name to ensure precise identification and traceability of each sample in the downstream process. Next, we indexed the alignment file using SAMtools [16] and generated the associated MPILEUP file using BCFtools [16]. An MPILEUP file is a text-based format that provides a per-base accumulation of sequencing reads against the reference sequence, detailing coverage and variant information. Finally, we called variants from the MPILEUP file using BCFtools [16], generating the corresponding VCF file.

We incorporated this workflow into a Snakefile, allowing for efficient parallel processing of multiple samples with Snakemake [18].

Merging VCF files

The output of this workflow was a single VCF file for each strain. To facilitate the analysis of genetic variation at the cohort level, we wanted to merge these VCF files. The procedure to merge VCF files (the `merge` function in the software package BCFtools) didn't run on more than 1,000 VCF files at a time. Therefore, we merged batches of 1,000 files. We re-indexed these files before conducting a final comprehensive merge (merged_output_all.vcf.gz on Zenodo).

Ultimately, we identified 3,119,517 variants in the cohort, including single-nucleotide polymorphism (SNP) and insertion-deletion (indel) variants.

Filtering variants

While many of these variants are likely to be true genetic variations across this set of strains, some may result from sequencing or alignment errors. We designed a filtering strategy to mitigate the risk of false positives while aiming to preserve as many true positives as possible. We used BCFtools [16] to refine our variant data and filtered out any variants with a QUAL score below 30 (a threshold for the probability of error) and a DP (depth of coverage) below 19.28 (a threshold for read depth).

Using a quality threshold of 30 represents a 99.9% probability that the variant is correctly identified. We established the read coverage threshold based on the coverage data from the 72 ECOR strains we used to generate the pangenome and our understanding of the presence or absence of each contig of the pangenome in these strains. We expand upon this determination process in the next section.

Applying these filters excluded 668,333 variants (21% of the initial dataset), leaving 2,451,184 variants in the dataset (filtered_output.vcf.gz on Zenodo). This filtration is conservative, and we should note that these excluded variants (~2.8% of all variants) may be true, potentially phenotypically impactful variation.

Defining the read depth filtering threshold

We defined the read depth threshold as the minimum number of reads required to confidently assert the presence of a nucleotide (and, by extension, the contig) in a strain. We analyzed the presence-absence data for the 72 ECOR strains to calculate this threshold, incorporating the coverage depth observed for each nucleotide that mapped against the pangenome.

Using custom Python scripts (extract_mpileup_info.py and numpy_merge_ecor.py), we first extracted the contig, position, and depth information for each nucleotide from the MPILEUP files of each of the 72 ECOR strains into a unified matrix (ecor72_array.txt on Zenodo). Then, a custom R notebook (Ecor72_averaging_contigDP.ipynb) calculated the average read depth per nucleotide at each contig for each strain as the total reads per nucleotide divided by the contig length.

Next, by integrating read depth and presence-absence data for each contig across all strains, we assessed coverage and read depth patterns in relation to the contig's presence-absence status (ECOR72_and_DP_threshold_analysis.Rmd). We observed that absent contigs are associated with significantly lower read depth (mean: 3 ± 10.85 reads/nucleotides) than present contigs (mean: 51 ± 28 reads/nucleotides). Some present contigs displayed low coverage and read depth, likely due to challenges in sequencing regions with high repeat content, high GC content, or complex secondary DNA structures. Conversely, some absent contigs exhibited full coverage and substantial read depth, potentially due to contamination, genome assembly errors, index hopping during sequencing, alignment artifacts, or highly repetitive sequences.

Ultimately, we established the DP threshold based on the distribution of read depth for absent contigs:

- DP threshold = mean read depth of absent contigs + 1.5 × standard deviation
- DP threshold = 3 + 1.5 × 10.85 = 19.28

This threshold indicates that we confidently consider any nucleotide or contig with a read depth exceeding 19.28 to be present.

Annotating variants

We performed automated variant annotation to help interpret the biological importance of variants in coding regions. Variant annotation uses information such as gene sequence and annotation to predict whether variants will have minimal or significant biological effects (e.g., categorizing variants into categories such as silent, missense, or nonsense mutations).

We used SnpEff (version 5.2c) [19] to annotate variants within the pangenome's coding sequences, excluding intergenic regions (IGRs). SnpEff analyzes input variants from a VCF file by annotating them based on a predefined database that includes gene annotations and gene sequence information. Since we used a custom pangenome, we first needed to construct a corresponding custom database before annotating.

Creating a SnpEff database and annotating variants

Creating a genome-specific database for variant annotation using SnpEff requires genome information in FASTA format and genome annotation in either GTF or GFF format. To this end, we annotated the pangenome coding sequences (pangenome_cds.fa on Zenodo) using Prokka (version 1.14.6) [8], and we then used the output (genes.gff on Zenodo) for SnpEff database creation.

Following the steps outlined in the SnpEff guidelines, we created the custom database with `-noCheckCds -noCheckProtein` to bypass the use of transcript and protein sequence information, which wasn't available.

The final database comprises 16,736 coding sequences out of the 18,494 initially present in the pangenome. Prokka uses Prodigal [20] to identify open reading frames (ORFs). Prodigal failed to identify 1,758 ORFs, possibly because Prodigal isn't designed to use pangenomes. Our pangenome contains many more contigs (thousands) than a typical genome (10s of contigs) expected by Prodigal. This could impact Prodigal in the following ways: first, during the training phase, Prodigal analyzes the genome to understand its characteristics, such as codon usage patterns and nucleotide composition, and adapts to the specific features of the input genome. The diverse and fragmented nature of a pangenome may hinder Prodigal's ability to accurately train this model. Second, during the gene prediction phase, Prodigal optimizes the selection of ORFs to predict the most likely set of genes in the input while ensuring that ORFs don't overlap improperly. Prodigal assigns scores to each potential ORF and, during the optimization step, ensures that the final set of predicted ORFs isn't redundant during the optimization step. Thus, if the same ORF (or substantially overlapping ORFs) is detected multiple times, only the highest-scoring version is retained. Despite Roary identifying the ORFs as 90% divergent at the protein level in our pangenome, Prodigal may still consider them overlapping due to their nucleotide-level similarity. This could lead to the exclusion of certain ORFs from the final predicted set, contributing to the observed discrepancy in ORF identification.

We then used this custom database to annotate the filtered VCF file and generated a new annotated VCF file (annotated_output.vcf.gz on Zenodo) and the corresponding SnpEff report (annot_summary_filtered.html on Zenodo).

Filtering silent mutations

We used SnpSift [21], a component of the SnpEff suite, to remove silent mutations from our set of annotated variants. The remaining variants (output.non_silent.vcf.gz on Zenodo) are likely to impact biology and result in missense, nonsense, and frameshift mutations.

Analyzing antimicrobial resistance-associated variants

To assess the veracity of our dataset, we tested whether we could associate AMR phenotypes with variants in genes known to be involved in AMR. We leveraged extensive previous work identifying genetic markers linked to AMR in *E. coli&nbsp;*[22]. Specifically, we focused on five genes known to confer resistance that we also identified in the pangenome:

- *tetA_1* (contig: LMHPMMMF_04732), *tetA_2* (contig: APHKLHJA_00520), and *tet_3* (contig: FCDKFLAE_04147), associated with tetracycline resistance
- *dfrD* (contig: NGHFEPFE_01999), associated with trimethoprim resistance
- *catA1* (contig: DHJNCGMO_04398), associated with chloramphenicol resistance

We extracted non-silent variant information for these contigs (resistance_output.non_silent.vcf.gz on Zenodo) and analyzed the distribution of these variants across the cohort. We then correlated the presence or absence of these variants with AMR phenotypes. These analyses are documented in an R notebook (Antimicrobial_resistance_investigation.Rmd).

Additional methods

We used ChatGPT to help write code and suggest wording ideas, which we then chose small phrases or sentence structure ideas to use. We used Grammarly Premium to help copy-edit draft text to match Arcadia's style and to clarify and streamline text that we wrote.

The dataset

We initiated our effort to build a large genotype–phenotype dataset by querying the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) [23] to identify a large cohort of *E. coli* strains with available genome sequence data and anti-microbial resistance (AMR) phenotypes for at least one antibiotic.

We chose *E. coli* for our study primarily due to the availability of thousands of genomes and extensively documented AMR phenotypes. Although similar data exist for other global pathogenic bacterial species, the remarkable genetic diversity within *E. coli* — with strains sharing only 20% to 40% of a core genome and possessing a wide array of accessory genes [3][4] — makes it uniquely suited for our research. This diversity allows us to interrogate this dataset using methods typically applied to interspecies comparisons and traditional intraspecies studies.

Our cohort contains a wide diversity of strains sampled over many years and countries

We identified a set of 6,983 *E. coli* strains with documented resistance or susceptibility to 50 antibiotics or combinations of two antibiotics (for instance, a treatment that includes both ampicillin and clavulanic acid). Genome sizes within this cohort ranged from 4.07 Mb to 5.98 Mb, with a median of 5.09 Mb and a mean of 5.07 Mb (Figure 1, A). Correspondingly, the number of coding sequences (CDS) varied from 3,992 to 6,135, with a mean of 5,069 CDS. These metrics reflect this species' typical genome size and genetic diversity [24].

In addition to genetic data, we retrieved metadata information from the database, including isolation country, collection year, and original host. Isolation country data was unavailable for only four strains; the remaining strains came from 14 countries (Figure 1, B), with the majority originating from the United Kingdom and Norway. Collection year information was available for 4,910 strains collected between 2001 and 2017. Host information was available for 3,922 strains, primarily isolated from humans and sourced from five other hosts (cow, dog, pig, cat, and chicken).

The diversity of location, time of collection, and host species present in our cohort demonstrates that our dataset isn't limited to a specific outbreak, environment, or timeframe. Thus, the genetic diversity within our cohort may be more analogous to species–species differences than the diversity observed in more closely related populations.

Our dataset includes susceptibility and resistance data for many antibiotics

We evaluated the diversity and distribution of AMR phenotypes to gain further insights into their patterns and prevalence within our cohort. This included organizing and analyzing documented AMR phenotypes across strains and antibiotics (or antibiotic classes) and identifying potential multidrug-resistant strains.

First, we assessed the number of antibiotics or antibiotic combinations for which strains had documented AMR phenotypes. The majority of strains had known phenotypes (either “susceptible,” “intermediate,” or “resistant”) for between eight and 11 antibiotics, and six strains had known phenotypes for only one antibiotic (Figure 2, A). Notably, one strain had phenotype information for 33 antibiotics. We analyzed the distribution of phenotypes for each antibiotic. Gentamicin had the highest number of documented

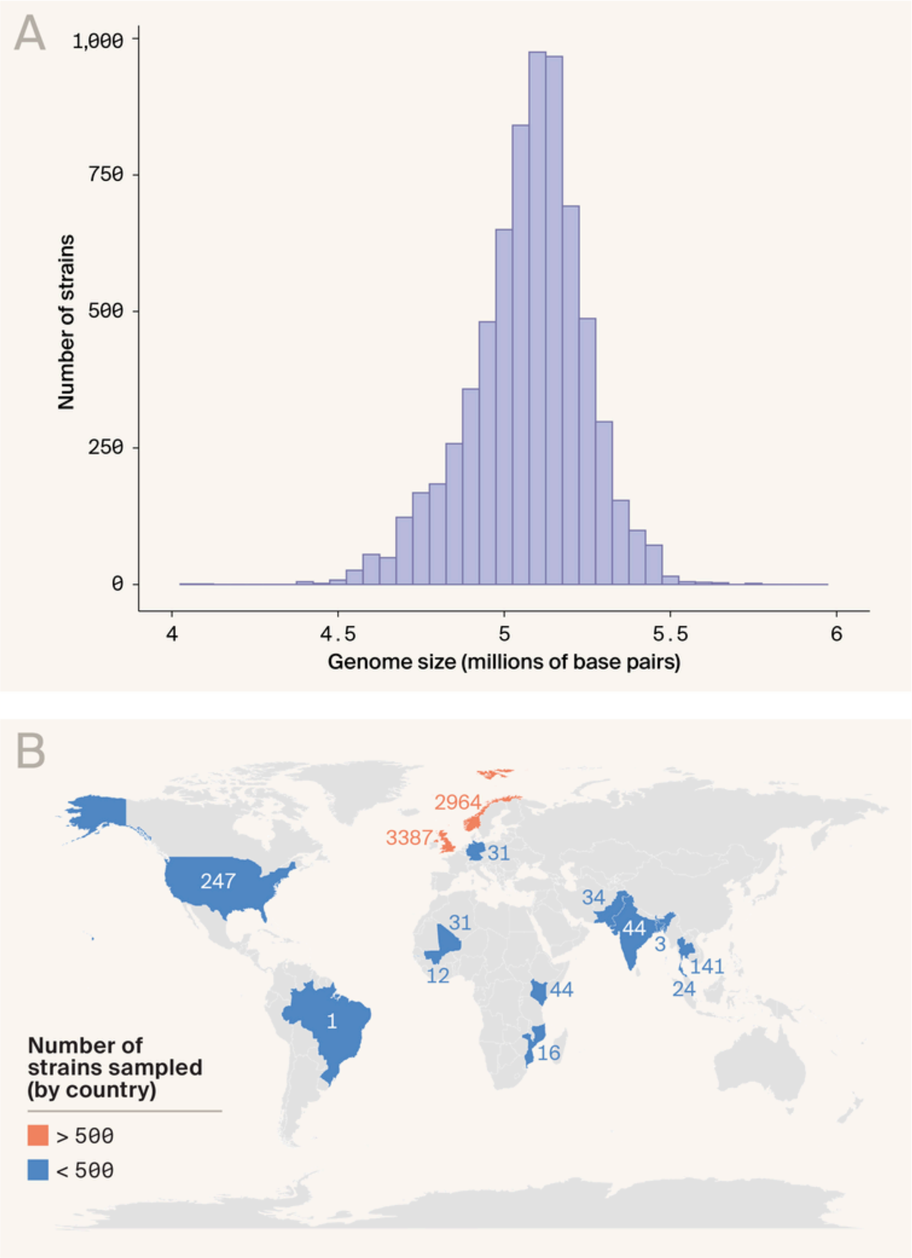

Figure 1. **The worldwide isolated strains in our cohort span the expected genome size range for** *E. coli*.

(A) Genome size distribution (number of base pairs).

(B) Map of the countries from which strains have been isolated.

phenotypes (6,043 strains), and 20 other antibiotics had phenotypes for more than 500 strains. We further focused on these antibiotics to evaluate the distribution of phenotypes (Figure 2, B). For most antibiotics, strains predominantly exhibited a “susceptible” phenotype, with some exceptions, such as ampicillin.

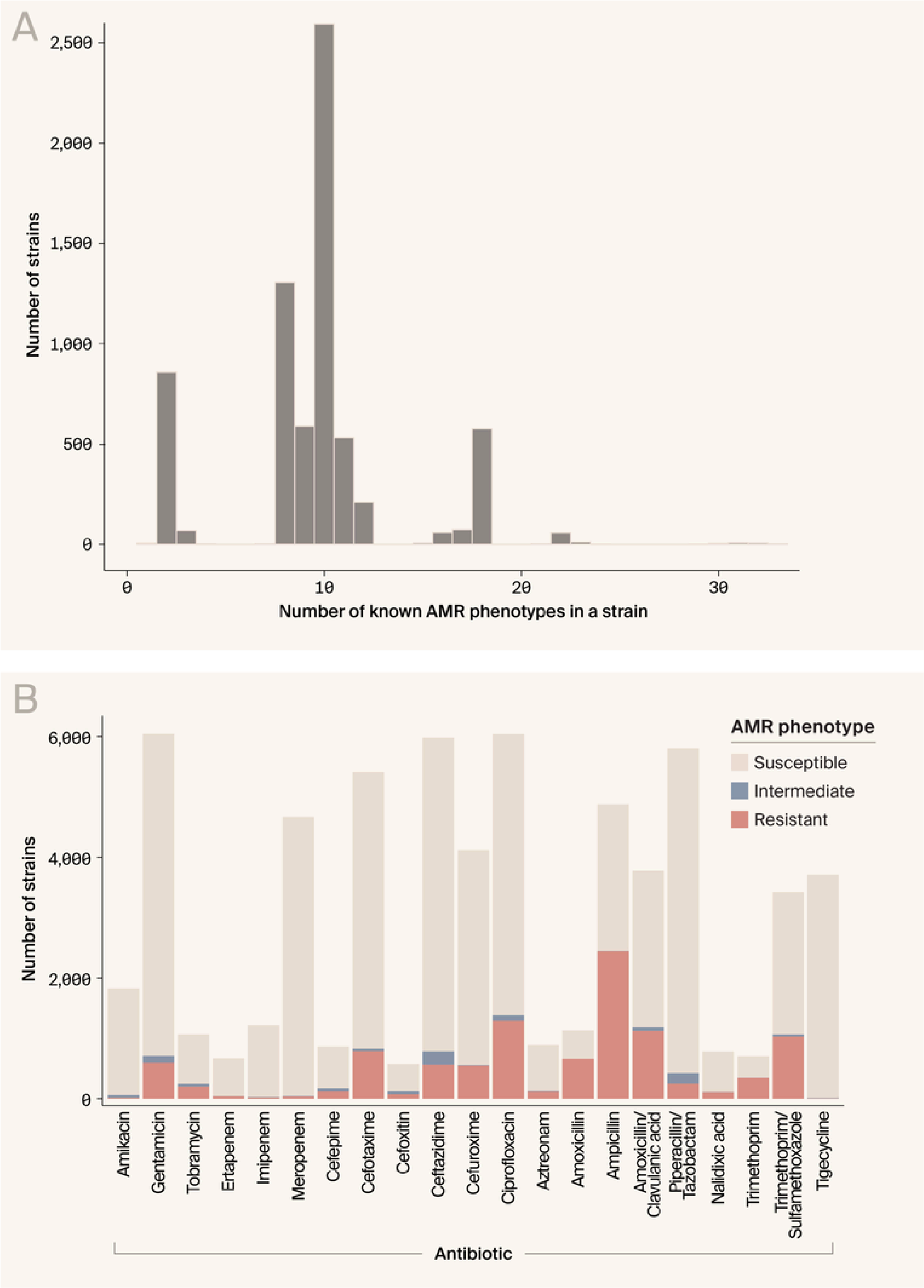

Figure 2. **Our cohort exhibits diverse AMR phenotypes across strains and antibiotics**.

(A) Distribution of the number of known AMR phenotypes per strain.

(B) Distribution of AMR phenotypes for antibiotics with AMR information available in 500 or more.

(C) Word cloud of antibiotic classes associated with the most resistant phenotypes.

Finally, we focused on the “resistant” phenotype, identifying 1,319 strains resistant to at least one antibiotic, including 139 strains resistant to 10 or more. These resistant phenotypes span 15 antibiotic classes; the most represented classes were penicillin, cephalosporin, fluoroquinolone, penicillin/beta-lactamase inhibitors, and sulfonamide. We identified 1,425 strains resistant to at least one antibiotic from three different classes, indicating multidrug resistance.

Our study encompasses a broad spectrum of antibiotic phenotypes, including susceptibility and resistance to a large number of differing antibiotics and many multi-drug-resistant strains. While not developed in this work, we believe that this dataset includes a diversity of phenotypic information that would allow us to identify correlations of AMR phenotypes within and between antibiotic classes, thus simplifying and improving predictions of AMR phenotypes.

We found broad genetic diversity across this cohort consistent with the diversity found in other *E. coli* populations

We identified genomic locations that vary across our cohort by first creating a pangenomic reference, aligning sequence data from all strains to this reference, and identifying locations that varied relative to the reference (variant calling). Given the broad genetic diversity commonly found in *E. coli*, selecting a representative reference genome was crucial. To this end, we constructed a pangenome from the 72 strains of the ECOR collection, which encompasses the species’ natural diversity [6]. The pangenome included intergenic regions (IGRs), which can play a critical regulatory role, and coding sequences (CDS). The final pangenome comprised 32,441 genetic sequences: 18,494 CDS and 13,947 IGRs. We further refer to these sequences as contigs.

By mapping all the strains against this pangenome and conducting variant calling, we identified 3,119,517 variants. This number was reduced to 2,451,184 after we filtered to control for false positives (see “The approach” for details on our filtering strategy). 2,407,385 of these variants were single-nucleotide polymorphisms (SNPs), and 43,799 were insertions or deletions (indels). Furthermore, we identified 376,038 multi-allelic variants, with the majority featuring two or three alternative alleles.

We further characterized the variants by their contig type (CDS versus IGR), assessed the variant rate per contig, and analyzed variant distribution within the cohort to identify likely rare variants and common variants. We detected variants in 85% of all contigs (27,637 of 32,441), with a higher tendency for variation in CDS contigs (16,715 out of 18,494; 90%) compared to IGR contigs (10,922 out of 13,947; 78%). The variant rate, the ratio of the contig length to the number of variants in that contig, characterizes a contig’s disposition to variation. For example, a variant rate of three suggests that variants occur every three nucleotides, on average. While variant rates varied widely, indicating different propensities for variation among contigs, the median variation rates were eight for CDS and 10 for IGRs, indicating slightly higher variability in coding sequences.

Finally, we evaluated the prevalence of each allelic variant across the strain cohort, measuring how often alternative alleles appeared. Notably, 19% of the variants (527,686 variants: 452,357 in CDS, 75,329 in IGR) were found in only one strain, indicating their rarity. Conversely, 6,167 variants (5,515 in CDS and 652 in IGR) appeared in 6,350 strains or more, suggesting these represent the more typical genetic composition of these contigs rather than true variants as they're found in at least 90% of the strains.

Altogether, our analysis of the variants in the *E. coli* cohort highlights the extensive genetic diversity within the species.

On average, around 40% of the variations in CDS are non-silent

The diversity of genetic variants identified in our study cohort is extensive, and we detected many variants within coding sequences (CDS contigs). However, we expect only some of these to be impactful. Genetic variation in coding regions can lead to silent mutations that don't affect the protein sequence, or non-silent mutations can alter protein sequence. To better understand the effects of these variants, we performed variant annotation to distinguish between silent and non-silent mutations within CDS contigs (see “The approach” for details).

We were able to annotate variants in 16,736 of the original 18,494 CDS contigs of the pangenome. The incompleteness of our annotation is likely due to limitations in using a pangenome in genome-based annotation algorithms and the exclusion of CDS that are considered redundant with already annotated CDS. The missing CDSs represent 9% of the CDS of the pangenome, and the absence of annotations for these variants is one limitation of this analysis.

Non-silent mutations alter protein sequence and are more likely to impact phenotypes. We analyzed the 783,436 variants classified as non-silent, which included frameshift, nonsense, and missense variants. Among these, 765,536 were SNPs, 17,900 were INDELs, and 95,581 were multiallelic variants. Notably, 33% of these variants (258,194 variants) were rare and found in only one strain. Conversely, 325 variants were prevalent in at least 90% of strains, suggesting these may be present in the majority of *E. coli* strains and that the allelic states in the ECOR collection (used to create the pangenome) are of lower prevalence.

Finally, we assessed the non-silent variant rate within each CDS contig to explore these sequences’ functional and evolutionary dynamics. Contigs with many non-silent mutations may indicate positive selection for these variants and suggest that CDS may impact phenotypes that enhance survival in some settings. In contrast, contigs predominantly containing silent variants are likely essential for cellular functions, and non-silent mutations would be deleterious. In the average CDS contig, ~42% of variants are non-silent (Figure 3, A). Furthermore, we found 316 contigs where non-silent variants accounted for at least 90% of the variants and 117 contigs where non-silent variants were less than 10% of the variants.

We used the Clusters of Orthologous Groups (COGs) database [25] to assign likely functions to each of the CDS contigs (Figure 3, B). For both groups — those with high and low non-silent mutation rates — most contigs were poorly characterized in this database (156 contigs in the high non-silent mutation rates group and 45 in the low rates group). However, for contigs associated with high rates of non-silent variants (> 90% of variants non-silent), the most represented COG categories were “Replication and Repair” (COG L: 8.1% of COG annotations) and “Cell Wall/Membrane/Envelope Biogenesis” (COG M: 4.2% of the COG annotations). Notably, these categories are some of the essential mechanisms of antibiotic resistance [26][27][28]. Conversely, the COG categories that are enriched in contigs associated with low non-silent variant rates (< 10% of variants non-silent) were Intracellular “Trafficking and Secretion” (COG U: 9.3%) and “Transcription” (COG K: 6.2%), essential cellular functions that, when altered, can lead to significant detrimental effects and tend to be conserved [29].

Within this cohort, there are many non-silent mutations in coding regions. Contigs associated with DNA replication, DNA repair, and cell wall biogenesis, functions associated with biological processes that adapt to evade antibiotics, tend to have higher rates of non-silent variants, consistent with *E. coli*’s reported adaptability.

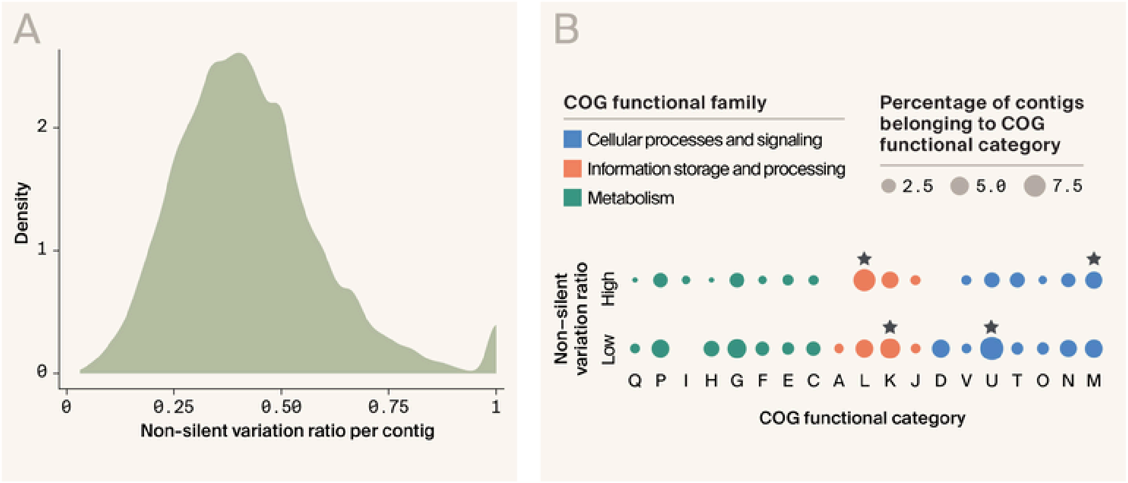

Figure 3. **Analysis of non-silent variant distribution**.

(A) Distribution of non-silent variant rate in the annotated CDS contigs. The ratio on the x-axis is the fraction of non-silent variants within each annotated CDS contig (per contig: non-silent variant / total number of variants).

(B) Distribution of COG functional categories among the contigs associated with high (> 90% non-silent mutations in a contig — 316 contigs) or low (< 10% of non-silent mutations in a contig — 117 contigs) non-silent variant rates. “Percentage” indicates the percentage of CDS in each dataset associated with the COG functional category. Stars identify the two COG functional categories (discussed in the text) that were the most represented among the contigs with high or low levels of non-silent polymorphisms. COG categories: A: RNA processing and modification; C: Energy production and conversion; D: Cell cycle control and mitosis; E: Amino acid metabolism and transport; F: Nucleotide metabolism and transport; G: Carbohydrate metabolism and transport; H: Coenzyme metabolism; I: Lipid metabolism; J: Translation; K: Transcription; L: Replication and repair; M: Cell wall/membrane/envelope biogenesis; N: Cell motility; O: Post-translational modification, protein turnover, chaperone functions; P: Inorganic ion transport and metabolism; Q: Secondary structure; S: Function unknown; T: Signal transduction; U: Intracellular trafficking and secretion; V: Defense mechanisms.

Assessing the veracity of the dataset: Identifying genetic variants associated with AMR

To evaluate the utility of our dataset in identifying polymorphisms associated with phenotypes, we correlated the presence of non-silent variants in coding sequences (CDS) previously associated with antimicrobial resistance (AMR).

We focused on resistance to three well-studied antibiotics: tetracycline, chloramphenicol, and trimethoprim, and five genes associated with resistance to these antibiotics. The genes *tetA_1*, *tetA_2*, and *tetA_3* code for tetracycline efflux pumps that expel tetracycline from the cell, thereby conferring resistance [30]. The gene *catA1*, which encodes chloramphenicol acetyltransferase, confers resistance to chloramphenicol by acetylating the antibiotic, thus preventing its binding to the bacterial ribosome [31]. The gene *dfrD* codes for a variant of the dihydrofolate reductase DhfR, the main target of the antibiotic trimethoprim. The DfrD protein can have a lower affinity for trimethoprim than DhfR and compensate for the DhfR function, conferring resistance to trimethoprim [32][33].

We identified 275 strains with non-silent variants in *tetA_1*, *tetA_2*, or *tetA_3* genes. Of these 275 strains, only 26 had a documented AMR phenotype for tetracycline, and all 26 were resistant. However, in the entire cohort, phenotype information for tetracycline was available for 393 strains (including the 26 strains just mentioned), of which 237 were resistant and 156 were susceptible (Figure 4). We performed a hypergeometric test to assess whether the set of 26 strains is significantly enriched for tetracycline resistance. This test calculates the probability of randomly selecting 26 resistant strains out of the 393 strains with available phenotype data. We found that the non-silent variants in *tetA_1*, *tetA_2*, or *tetA_3* are significantly associated with tetracycline resistance ($p < 0.001$). The resistance of the remaining 211 strains lacking variants in the *tetA* genes may be attributed to alterations in other tetracycline resistance genes, including additional efflux pumps (*tetB*, *tetC*, *tetD*, and *tetE*) or ribosomal protection proteins (*tetM* and *tetO*) [30].

When we investigated resistance to chloramphenicol, we identified 58 strains with non-silent variants in *catA1*, but only two of these 58 strains had documented chloramphenicol AMR phenotypes — both resistant. In the cohort, phenotype information for chloramphenicol was available for 253 other strains with no variant information, 72 of which were resistant (Figure 4). Given the small number of strains with variants in the *catA1* contig with an available AMR phenotype, we were unable to find a significant connection between variants in *catA1* and resistance to chloramphenicol ($p > 0.05$, hypergeometric test). It's possible that the resistant phenotypes of the strains lacking variations in *catA1* are associated with variations in other known chloramphenicol resistance genes, such as other *cat* genes or *cml* genes [34].

Last, we assessed the possible genetic basis of trimethoprim resistance. We found 571 strains with non-silent variants in *dfrD*, and AMR phenotype information was available for 40 of them (Figure 4). These 40 strains displayed both susceptible (19 strains) and resistant phenotypes (21 strains). In the whole cohort, trimethoprim resistance data was available for 704 strains, including the 40 aforementioned strains. Ultimately, we didn't find a significant association between the presence of trimethoprim resistance and non-silent mutations in *dfrD* ($p > 0.05$, hypergeometric test). Across the *dfrD* gene, there were a total of seven non-silent alleles. One possibility is that some of these non-silent alleles significantly impact the functioning of *dfrD* while others do not. This potentially motivates a more targeted analysis of these individual polymorphisms but is beyond the scope of this work.

Our analysis focuses on resistance to three antibiotics and five genes successfully correlated genes known to influence tetracycline resistance but failed to identify genes linked to chloramphenicol or trimethoprim. For the latter two, one non-trivial explanation could be statistical power as the number of strains with documented chloramphenicol resistance was small (two), and the number of non-silent loci in *dfrD* (known to influence trimethoprim resistance) was numerous, possibly reducing the statistical power of the test.

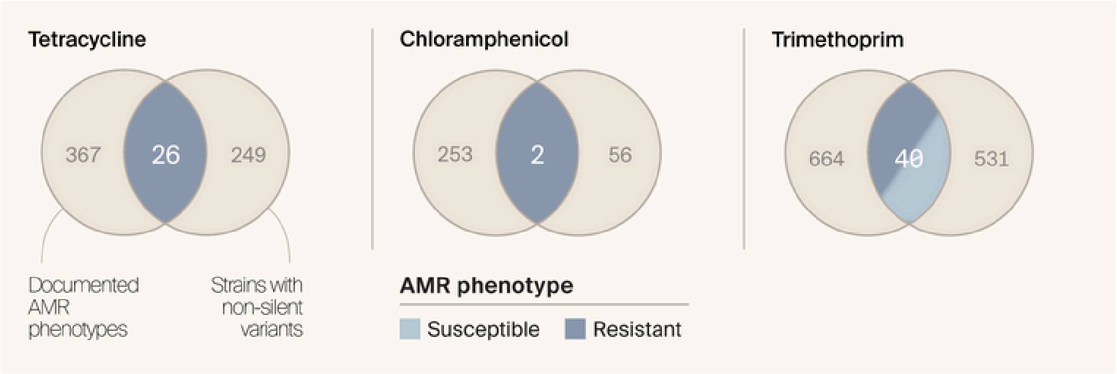

Figure 4. **Analysis of non-silent variants in known genes associated with** *E. coli* **antimicrobial resistance**.

Intersections of strains with non-silent variants in antibiotic resistance contigs and available AMR phenotypes. Each Venn diagram shows the intersections of strains that have identified non-silent variants in resistance genes of the indicated antibiotic (chloramphenicol, trimethoprim, or tetracycline) and corresponding AMR phenotypes. The color of the intersection indicates the AMR phenotypes observed for these strains (pale blue: susceptible, darker blue: resistant).

Key takeaways

In this work, we've compiled a genotype–phenotype dataset from public databases for around 7,000 *E. coli* strains, focusing on their anti-microbial resistance (AMR) phenotypes. After retrieving AMR phenotypes for 50 antibiotics or combinations, we performed variant calling for these strains against a custom pangenome generated from 72 different *E. coli* strains.

The cohort features significant genetic diversity, with 2.4 million variants in 85% of the pangenome's coding sequences and intergenic regions. We also identified non-silent variants that could affect protein integrity and function. Specifically, we focused on variants within genes reported to be associated with AMR, and we were able to find resistant variants in known AMR-related genes.

Altogether, we hope this dataset will be a versatile resource for microbiologists, geneticists, and evolutionary biologists who want to probe genotype–phenotype associations or delve deeper into the genetic basis of AMR.

Challenges and limitations

Compiling this dataset presented multiple challenges, which led us to make specific decisions that shaped the scope of our analysis. As a result, we may have excluded some AMR genetic markers, making our dataset less comprehensive for a complete analysis of antimicrobial resistance in *E. coli*.

We discuss these challenges, the choices we made in response, and the probable limitations of the dataset that result from those choices here.

Incomplete data

We compiled data for around 7,000 *E. coli* strains, but AMR phenotypes weren't available for all antibiotics and strains. For some antibiotics, like gentamicin or ciprofloxacin, AMR phenotypes were available for over 6,000 strains. Unfortunately, for most antibiotics, phenotypes were available for less than 1,000 strains, reducing the utility of this dataset.

Choice of phenotype

AMR is one of the best-documented phenotypes for microbes, so it made a great option for building a large-scale genotype–phenotype dataset. However, many AMR genes are found on plasmids, contributing significantly to the rapid spread of AMR resistance through processes like horizontal gene transfer [35][36]. Although our pangenome contains likely plasmid sequences, it doesn't capture the full diversity of plasmids in *E. coli*. Genome assembly from short reads doesn’t allow for the efficient recovery of plasmid sequences, and the 72 ECOR strains [6] are unlikely to cover the entire plasmid diversity within *E. coli*. Therefore, our dataset may only include variation for some AMR genes.

Working with a pangenome

While using a pangenome doesn't fully capture plasmid diversity, it allowed us to capture more of the genetic diversity in our *E. coli* cohort than using a single reference strain. However, this choice, too, brought inherent challenges and further analytical considerations worth discussing.

One major challenge is determining the number of species to include in the pangenome. This choice depends on multiple factors. While covering most of the diversity is desirable, adding more genomes increases the risk of inaccuracies due to sequencing, assembly, and annotation errors. We chose to work with the ECOR collection because it's been extensively studied, but even in this cohort, there are reported sequencing and assembly errors **[7]**. While it includes strains from different *E. coli* populations, it nonetheless leaves some diversity of the whole *E. coli* species unaccounted for, possibly missing important information regarding the genetic basis of AMR phenotypes.

Challenges also arise during the creation of the pangenome, including clustering and classification of orthologous and paralogous genes and errors in automated gene identification, which can introduce inaccuracies in the final set of sequences. Additionally, programs designed for genome or pre-assembled genome processing, such as Prokka **[8]** and Prodigal **[20]**, might not perform as well with a pangenome. For instance, CDS identification by Prodigal is informed by the genome structure and organization and is optimized based on what’s expected to be a complete, non-redundant set of ORFs in a prokaryotic genome. This reliance on a typical genome structure could have led to the lack of annotation of some CDS in our pangenome, resulting in potentially important missing information regarding variants and non-silent variants in CDS regions, including potential AMR contigs.

Ultimately, our pangenome resulted in a reference sequence four to five times the size of a regular bacterial genome, leading to a more significant computational load compared to typical microbial genomics studies.

Despite these challenges, the pangenome allowed us to intentionally generate a dataset that examines single-nucleotide variation in specific genomic locations and encompasses the loss and gain coding regions between strains. This common backbone was essential to investigate a cohort of *E. coli* strains, as *E. coli* is known for its dynamic genome, characterized by frequent DNA loss and gain [3]. As a result, individual strains in this dataset are somewhat similar to separate species where gene gain/loss is a critical differentiator, but also resemble actively interbreeding individuals where single-nucleotide polymorphisms are the primary form of genetic variation. This complexity allows for analysis from both a phylogenetic and a population genetic perspective. This data can serve as a testbed for methods that apply to the study of genetic variation within and between species, with the goal of integrating these two approaches more completely.

These challenges and considerations highlight the importance of accurately defining the scope and ambitions of a genotype–phenotype study to rationally decide whether a pangenome is more valuable than a single reference strain.

Computational resources and data limitations

Producing the entire dataset required significant computational power and runtime. For instance, generating all individual variant calling files took a week with 50 CPUs, and the data generated, including important intermediary files such as alignment files and MPILEUP files, represented 15 to 20 terabytes. Such requirements inevitably limit the number of genotype–phenotype datasets and studies for large populations.

This approach also limited the level of data and completeness of the information we could save or generate. To reduce the size of the MPILEUP files, we included information only for locations with aligned reads. Similarly, when identifying variants, we saved information in the vcf.gz file only for locations where we identified an allelic variant. This choice further limited our ability to call the reference allelic states in our final variant population. Consequently, our final dataset in the variant calling format (VCF) file is incomplete, as it only reports identified allelic variants. However, if further investigations are needed, the missing information can be retrieved from the associated BAM files.

Next steps

We have a few things in mind for using this dataset at Arcadia.

First, it’ll be an excellent candidate for testing and using our phenotype encoder [37] and for carrying out phenotype–phenotype predictions. It'll provide an opportunity to evaluate how the autoencoder from our previous work handles datasets with multiple missing phenotypes. Biologically, it’ll offer deeper insights into the distribution and correlation of AMR phenotypes by identifying interesting patterns of AMR phenotypes associations between antibiotics and antibiotic classes.

Second, this dataset will be valuable for investigating gene–gene interactions and building models to reconstruct gene networks from genotype–phenotype information. This will enable a move from a single-gene definition of phenotypes to a better characterization of epistasis. Again, the existence of details regarding gene–gene interaction in antimicrobial resistance will allow us to assess the power and reliability of these models. This would also provide better insights into the gene networks underlying antimicrobial resistance.

Outside of Arcadia, we hope that microbiologists, population geneticists, epidemiologists, and evolutionary biologists will also find this dataset valuable. We're eager to hear from researchers using it to better understand antimicrobial resistance mechanisms in *E. coli*, predict the emergence of resistance, or conduct genotype–phenotype predictions in future outbreaks. Although AMR is the example phenotype for this work, the extensive variant matrix we generated, mapping out the diversity within this large *E. coli* cohort, isn't limited to studying AMR phenotypes. Researchers can use it to identify the genetic basis of many other phenotypes in *E. coli* (across all 7,000 strains or a subset). We hope researchers will use this genotype mine to investigate phenotypes like host association, metabolism, or stress response and provide feedback on other interesting phenotypes to explore.

Contributors (A–Z)

- **Audrey Bell**: Visualization
- **Megan L. Hochstrasser**: Editing
- **Elizabeth A. McDaniel**: Validation
- **David G. Mets**: Conceptualization, Methodology, Resources, Supervision
- **Manon Morin**: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing

References

1. Lehner B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. https://doi.org/10.1038/nrg3404
2. Benfey PN, Mitchell-Olds T. (2008). From Genotype to Phenotype: Systems Biology Meets Natural Variation. https://doi.org/10.1126/science.1153716

3. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EPC, Denamur E. (2009). Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. https://doi.org/10.1371/journal.pgen.1000344

4. Lukjancenko O, Wassenaar TM, Ussery DW. (2010). Comparison of 61 Sequenced Escherichia coli Genomes. https://doi.org/10.1007/s00248-010-9717-3

5. Tonkin-Hill G, Corander J, Parkhill J. (2023). Challenges in prokaryote pangenomics. https://doi.org/10.1099/mgen.0.001021

6. 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

7. Patel IR, Gangiredla J, Mammel MK, Lampel KA, Elkins CA, Lacher DW. (2018). Draft Genome Sequences of the Escherichia coli Reference (ECOR) Collection. https://doi.org/10.1128/mra.01133-18

8. Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. https://doi.org/10.1093/bioinformatics/btu153

9. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. https://doi.org/10.1093/bioinformatics/btv421

10. Thorpe HA, Bayliss SC, Sheppard SK, Feil EJ. (2018). Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. https://doi.org/10.1093/gigascience/giy015

11. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. https://doi.org/10.1093/molbev/msab293

12. Tange O. (2018). Gnu Parallel 2018. https://doi.org/10.5281/zenodo.1146014

13. https://github.com/ncbi/sra-tools/wiki/01.-downloading-sra-toolkit

14. Chen S. (2023). Ultrafast one pass FASTQ data preprocessing, quality control, and deduplication using fastp. https://doi.org/10.1002/imt2.107

15. Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arxiv.1303.3997

16. 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

17. https://github.com/broadinstitute/picard/releases/tag/2.27.5

18. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. (2021). Sustainable data analysis with Snakemake. https://doi.org/10.12688/f1000research.29032.1

19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. https://doi.org/10.4161/fly.19695

20. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. https://doi.org/10.1186/1471-2105-11-119

21. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. https://doi.org/10.3389/fgene.2012.00035

22. Leekitcharoenphon P, Johansson MHK, Munk P, Malorny B, Skarżyńska M, Wadepohl K, Moyano G, Hesp A, Veldman KT, Bossers A, EFFORT Consortium, Graveland H, van Essen A, Battisti A, Caprioli A, Blaha T, Hald T, Daskalov H, Saatkamp HW, Stärk KDC, Luiken REC, Van Gompel L, Hansen RB, Dewulf J, Duarte ASR, Zając M, Wasyl D, Sanders P, Gonzalez-Zorn B, Brouwer MSM, Wagenaar JA, Heederik DJJ, Mevius D, Aarestrup FM. (2021). Genomic evolution of antimicrobial resistance in Escherichia coli. https://doi.org/10.1038/s41598-021-93970-7

23. Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis J, Dempsey D, Dickerman A, Dietrich E, Kenyon R, Kuscuoglu M, Lefkowitz E, Lu J, Machi D, Macken C, Mao C, Niewiadomska A, Nguyen M, Olsen G,

Overbeek J, Parrello B, Parrello V, Porter J, Pusch G, Shukla M, Singh I, Stewart L, Tan G, Thomas C, VanOeffelen M, Vonstein V, Wallace Z, Warren A, Wattam A, Xia F, Yoo H, Zhang Y, Zmasek C, Scheuermann R, Stevens R. (2022). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. https://doi.org/10.1093/nar/gkac1003

24. Bergthorsson U, Ochman H. (1995). Heterogeneity of genome sizes among natural isolates of Escherichia coli. https://doi.org/10.1128/jb.177.20.5784-5789.1995

25. Tatusov RL. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. https://doi.org/10.1093/nar/28.1.33

26. Kohanski MA, Dwyer DJ, Collins JJ. (2010). How antibiotics kill bacteria: from targets to networks. https://doi.org/10.1038/nrmicro2333

27. Gauba A, Rahman KM. (2023). Evaluation of Antibiotic Resistance Mechanisms in Gram-Negative Bacteria. https://doi.org/10.3390/antibiotics12111590

28. Traverse CC, Ochman H. (2016). Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. https://doi.org/10.1073/pnas.1525329113

29. Chopra I, Roberts M. (2001). Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance. https://doi.org/10.1128/mmbr.65.2.232-260.2001

30. Shaw WV. (1983). Chloramphenicol Acetyltransferase: Enzymology and Molecular Biology. https://doi.org/10.3109/10409238309102789

31. Cammarata M, Thyer R, Lombardo M, Anderson A, Wright D, Ellington A, Brodbelt JS. (2017). Characterization of trimethoprim resistant E. coli dihydrofolate reductase mutants by mass spectrometry and inhibition by propargyl-linked antifolates. https://doi.org/10.1039/c6sc05235e

32. Huovinen P, Sundström L, Swedberg G, Sköld O. (1995). Trimethoprim and sulfonamide resistance. https://doi.org/10.1128/aac.39.2.279

33. Bissonnette L, Champetier S, Buisson JP, Roy PH. (1991). Characterization of the nonenzymatic chloramphenicol resistance (cmlA) gene of the In4 integron of Tn1696: similarity of the product to

transmembrane transport proteins. https://doi.org/10.1128/jb.173.14.4493-4502.1991

34. Bennett PM. (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. https://doi.org/10.1038/sj.bjp.0707607

35. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, Savelkoul PHM, Wolffs PFG. (2016). Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. https://doi.org/10.3389/fmicb.2016.00173

36. Avasthi P, Mets DG, York R. (2023). Harnessing genotype-phenotype nonlinearity to accelerate biological prediction. https://doi.org/10.57844/arcadia-5953-995f