

UC San Diego

UC San Diego Previously Published Works

Title

Decomposition of the pangenome matrix reveals a structure in gene distribution in the Escherichia coli species.

Permalink

<https://escholarship.org/uc/item/1xm413w8>

Journal

Clinical and Vaccine Immunology, 10(1)

Authors

Chauhan, Siddharth

Ardalani, Omid

Hyun, Jason

et al.

Publication Date

2025-01-28

DOI

10.1128/msphere.00532-24

Peer reviewed

Decomposition of the pangenome matrix reveals a structure in gene distribution in the *Escherichia coli* species

Siddharth M. Chauhan,¹ Omid Ardalani,² Jason C. Hyun,¹ Jonathan M. Monk,¹ Patrick V. Phaneuf,² Bernhard O. Palsson^{1,2,3,4}

AUTHOR AFFILIATIONS See affiliation list on p. 15.

ABSTRACT Thousands of complete genome sequences for strains of a species that are now available enable the advancement of pangenome analytics to a new level of sophistication. We collected 2,377 publicly available complete genomes of *Escherichia coli* for detailed pangenome analysis. The core genome and accessory genomes consisted of 2,398 and 5,182 genes, respectively. We developed a machine learning approach to define the accessory genes characterizing the major phylogroups of *E. coli* plus *Shigella*: A, B1, B2, C, D, E, F, G, and *Shigella*. The analysis resulted in a detailed structure of the genetic basis of the phylogroups' differential traits. This pangenome structure was largely consistent with a housekeeping-gene-based MLST distribution, sequence-based Mash distance, and the Clermont quadruplex classification. The rare genome (consisting of genes found in <6.8% of all strains) consisted of 163,619 genes, about 79% of which represented variations of 315 underlying transposon elements. This analysis generated a mathematical definition of the genetic basis for a species.

IMPORTANCE The comprehensive analysis of the pangenome of *Escherichia coli* presented in this study marks a significant advancement in understanding bacterial genetic diversity. By employing machine learning techniques to analyze 2,377 complete *E. coli* genomes, the study provides a detailed mapping of core, accessory, and rare genes. This approach reveals the genetic basis for differential traits across phylogroups, offering insights into pathogenicity, antibiotic resistance, and evolutionary adaptations. The findings enhance the potential for genome-based diagnostics and pave the way for future studies aimed at achieving a global genetic definition of bacterial phylogeny.

KEYWORDS *Shigella*, *Escherichia coli*, genomics, typing, computational biology, genome analysis

The first complete bacterial genome sequence appeared in 1995 (1). Shortly thereafter, the genome sequence of the model *Escherichia coli* K-12 MG1655 strain appeared (2). The genome sequence of a second *E. coli* strain, the enterohemorrhagic O157:H7 strain, appeared in 2001 (3). It had about a 1 Mbp longer genomic sequence than MG1655, encoding about 1,000 additional genes representing different traits than those found in MG1655. Following the massive drop in DNA sequencing costs in the late 2000s (4), a large number of *E. coli* strain sequences became available (5, 6). These data form the basis for pangenome analysis of the *E. coli* species (7, 8). In 2013, a study analyzed 55 *E. coli* genome sequences (9). Using metabolic reconstructions and computational systems biology, auxotrophies and colonization sites could be predicted from these genome sequences. As the number of available genome sequences grew, subsequent studies showed that differential traits between phylogroups could be delineated from sequence and specific pathogenic properties could be deciphered (10).

With the availability of low-cost genomic sequencing, strain taxonomic classifications thus moved from phenotypes to genotypes. This started with the creation of the original

Editor Raffaele Zarrilli, University of Napoli Federico II, Naples, Italy

Address correspondence to Bernhard O. Palsson, bpalsson@ucsd.edu.

The authors declare no conflict of interest.

Received 9 July 2024

Accepted 8 November 2024

Published 31 December 2024

Copyright © 2024 Chauhan et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Achtman multilocus sequence typing (MLST) schema (11). Following this development, the Clermont triplex (12) and subsequently the quadruplex (13) appeared that deployed PCR assays for discriminating alleles to perform sequence-based phylogrouping. More recently, the whole genome-based Mash distance has been utilized to successfully phylogroup *E. coli* strains, moving the definition of phylogroups to the genome scale (14). Today, the number of *E. coli* sequences in the public domain has reached the 10^5 scale (15). These sequences contain the full gene complement of these strains. This data availability demands development of novel big data analytic methods that characterize the strains' genomes based on their full genome-wide gene content.

We can now call the presence/absence of genes across thousands of genomes. The results enable us to form the pangenome matrix (10, 16, 17) for the *E. coli* species. Once formed, this matrix allows us to develop machine learning methods to classify the entire gene complement of these strain sequences into phylogroups. Meaningful classification of strains would allow us to precisely define the genetic basis for differential traits observed between the phylogroups. If phylogroup- and strain-specific traits can be derived straight from sequence, it would reduce the need for strain cultivation in clinical settings and allow for accelerated diagnosis. The full phylogroup definition of the *E. coli* species thus has fundamental and applied implications.

RESULTS

Forming the pangenome matrix

We downloaded all available *E. coli* genomes from two public databases, BV-BRC and NCBI RefSeq. This sequencing data were subjected to quality controls and admissions criteria from pangenomic studies (Fig. 1a, Methods). The result was a collection of over 10,000 high-quality genome sequences, of which 2,377 were high-quality complete sequences that were used for pangenome analysis. These sequences were collected from a wide variety of isolation sources, including humans, land animals, and various species of birds (Fig. 1b). Most of the strains did not contain any plasmids, with notable exceptions (*e.g.*, a Phylogroup G strain containing seven plasmids) (Fig. 1c). We call this curated collection of sequences and its resulting pangenome a Genome Encyclopedia of Notable Observed Microorganisms Curated for Universal Study (GENOMiCUS).

Genomes can be classified using sequence characteristics. The Mash distance between genome sequences has been shown to quantify their differences (Fig. 1d) (14, 19). One can now cluster a series of genome sequences based on global sequence similarity. A heatmap classification of the sequences used in this study shows that Mash distances lead to phylogroup classification, consistent with a previous study (14). In addition, phylogroup designation based on the Clermont quadruplex standard can be computed from the sequences (summarized Fig. 1e) (13, 18). It shows that phylogroups A, B1, and B2 had the highest number of strains in the collection analyzed. In contrast, the recently defined phylogroup G (21, 22) had relatively few complete strain sequences available for analysis.

Stratifying the pangenome into three categories of genes

A pangenome can be stratified into three main categories of genes:

- The **core genome** consists of the genes found in all, or nearly all, of the strains. These genes, therefore, can be taken to define the species. For the collection of strain sequences analyzed here, the core genome consists of 2,398 genes, 80% of which have known functions.
- The **accessory genome** is composed of 5,182 genes. These are genes that are found in many, but not all strains. The accessory genes, being variably present, can be used to define the gene portfolio of the phylogroups, as described below.

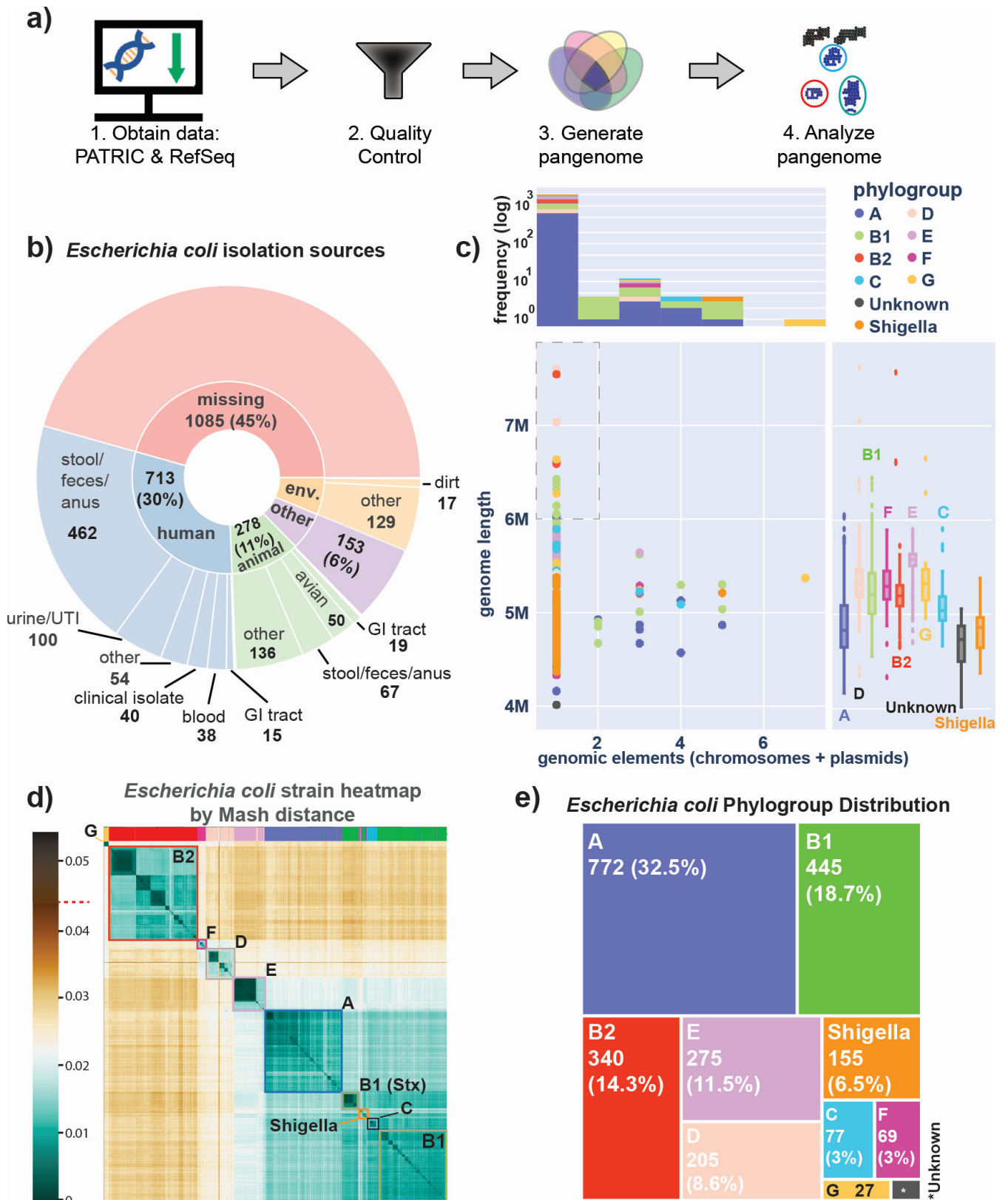


FIG 1 Processing and classification of a 2,377 complete *Escherichia coli* genome compendium (GENOMICUS). (a) The workflow used in this study. Genomes were downloaded from PATRIC (now BV-BRC) and RefSeq, after which they were deduplicated and filtered based on their quality metrics (see **Methods**). The resulting 2,377 complete genomes form a high-quality compendium of strains for detailed pangenome analysis. We call this compendium the Genome Encyclopedia of (Continued on next page)

Fig 1 (Continued)

Notable Observed Micro-organisms Curated for Universal Study (GENOMICUS). (b) A sunburst plot showcasing the different isolation sources for the bacteria in this compendium. While most of the 1,332 isolation site-annotated strains come from humans (713), there are many strains isolated from animals (278) and various other environmental niches (146). (c) Scatterplot summarizing properties of the genomes by genome length (y-axis) vs number of genomic elements (chromosomes + plasmids) (x-axis), colored by phylogroup as calculated *in silico* by the ClermonTyping github package (18). Note that many *Shigella* strains were incorrectly classified by ClermonTyping as belonging to Phylogroup A, and so any strains which were known to be *Shigella* were manually separated into a separate class for better identification. Nineteen strains were found to have a genome size greater than 6 Mb. Sixteen of those 19 strains were clinical isolates from ICDDR,B from patients who had diarrheal disorders. Above the scatterplot is a histogram showcasing the genomic element distribution within the strains of the pangenome, also colored by phylogroup. Note: in this context, a “genomic element” refers to both the main chromosome and any additional plasmids found in the organism. To the right of the scatterplot are phylogroup-specific boxplots describing the distribution of genome lengths per phylogroup. (d) A heatmap of the pairwise Mash distances for all 2,377 *E. coli* strains of GENOMICUS based on sequence analysis. Distances range from 0 to 0.04, and the highest Mash value (0.044) is denoted with a red dash on the color bar. Note that a pairwise Mash distance of 0.05 equates to an average nucleotide identity (ANI) of 95%, both of which correspond to a 70% DNA–DNA reassociation value, the historical definition of a bacterial species (19, 20). The highlighted bars at the top of the heatmap identify the Mash-based clusters of this compendium. Phylogroups are annotated on the heatmap, showing the correspondence between these phylogroups and the Mash-based clusters. (e) Treemap illustrating the distribution of *E. coli* strains by phylogroup as calculated *in silico* by the ClermonTyping github package (18).

- The **rare genome** consists of genes unique to a strain or found in a relatively small number of strains. The exact number defining this cutoff is determined using the protocol by Hyun et al. (23).

These three categories of genes are deciphered from the frequency of gene occurrence in the collection of strain sequences (Fig. 2a). This gene frequency histogram shows the number of genomes containing a particular gene. Taking the cumulative sum of the gene frequency, we get a cumulative gene distribution that is used to formally determine the boundaries for the core, accessory, and rare genomes (Fig. 2b) (23)

Defining the genes in the core, accessory, and rare genomes

The boundary between the core and accessory genome separates (near-)omnipresent genes from variably present accessory genes. Defining the boundary between the rare and the accessory genome is more subjective. The definition of these boundaries in this study is described in **Methods**, and they lead to the identification of 2,398 core genes, 5,182 accessory genes, and 163,619 rare genes. The exact definition of these boundaries does not affect the major conclusion of this study (see S1).

The number of genes classified into the core genome can be plotted with the number of strains considered (Heaps' Plot, Fig. S1). This curve levels off fairly quickly with the number of genomes considered, and stays flat at 2,398 genes, defining a closed core genome. The number of genes classified as accessory genes similarly levels off at 5,182 genes. This observation shows that the accessory genome is also closed. The closed nature of the accessory genome makes it possible to analyze the phylogroup gene content in novel and mathematically rigorous ways, as shown in the next section.

Thus, after a certain number of strains, the discovery of novel genes in the pangenome is driven by occurrence of rare genes; the median number of such genes per strain is 270, and 163,619 total among the 2,377 genomes studied. These rare genes will confer unique characteristics onto the strain in which they reside.

Traits in the core genome

The genes common to the 2,377 strains represent the core genome. Thus, there is a uniform genetic basis for certain traits. For instance, the core genome contains 18 of 29 two-component systems, consistent with previous findings (24). One of 68 biosynthetic gene clusters, two of 130 AMR genes (*Int*, an apolipoprotein N-acyltransferase and *narP*, a nitrate/nitrite response regulator), 382 of 127,223 transposable elements, and 21 of 7,925 motility genes (pili, fimbriae, flagella, and supporting proteins) are in the core genome. There are 1,006 metabolic genes in the core genome, which is slightly higher than the

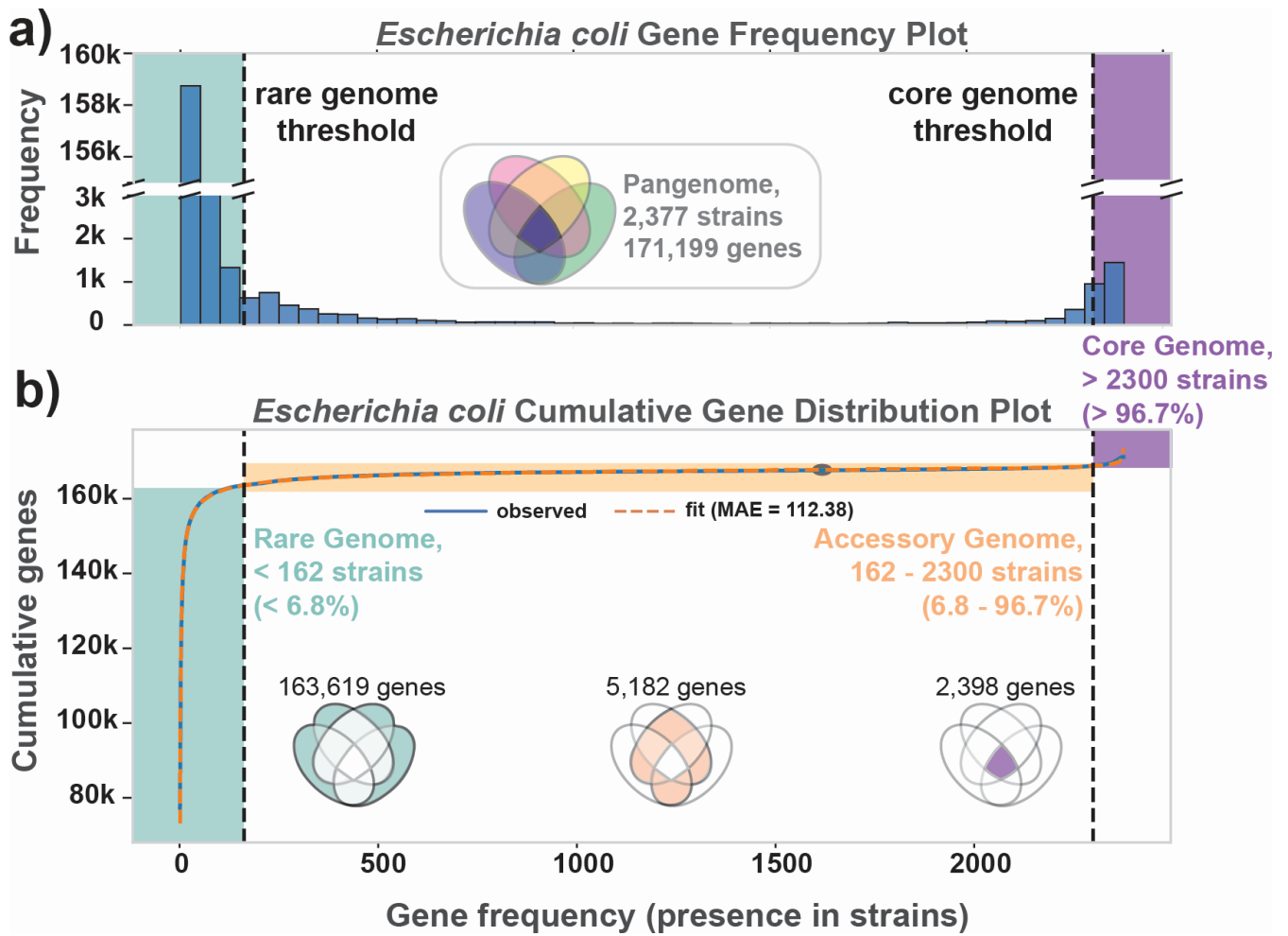


FIG 2 Global distributions of gene frequencies and functions in the *Escherichia coli* pangenome. (a) Gene frequency distribution across the 2,377 curated genomes in GENOMICUS. Genes present in all 2,377 strains appear at the histogram’s right end. Progressing leftward, subsequent bars show genes found in nearly all strains, decreasing in frequency, until reaching genes unique to just one strain at the extreme left. (b) The cumulative gene distribution function (23). The gene frequency distribution was fitted to a double-exponential form (with median absolute error or MAE = 176.31) and the inflection points determined. Based on these inflection points, the genes in the pangenome were divided into the core (comprising 2,398 genes), accessory (comprising 5,182 genes), and rare (comprising 163,619 genes) genomes (See Methods).

976 genes previously reported (25) (Fig. S2). There are still 462 genes of unknown function (y-genes) in the core genome, comprising 19% of all core genes.

The accessory genome has a clear mathematical structure

The accessory genome is effectively closed (Fig. S1), enabling a comprehensive analysis of gene-level diversity among the 2,377 strains. To do so, we form the **P** matrix (genes × strains) for just the accessory genome. This **P** matrix can be decomposed using non-negative matrix factorization (NMF) (26, 27) to define the genes that belong to strains of a particular phylogroup. NMF factors **P** into two matrices:

$$P = LA$$

L, indicate columns consisting of gene weightings that define a *phylon* (the genes common among similar strains, often belonging to the same Clermont phylogroup and/or MLST cluster), and **A**, indicate rows giving a strain’s *affinity* (or closeness) of a genome to a phylon (Fig. 3a). The column space of **L** is a convex cone, as all its values

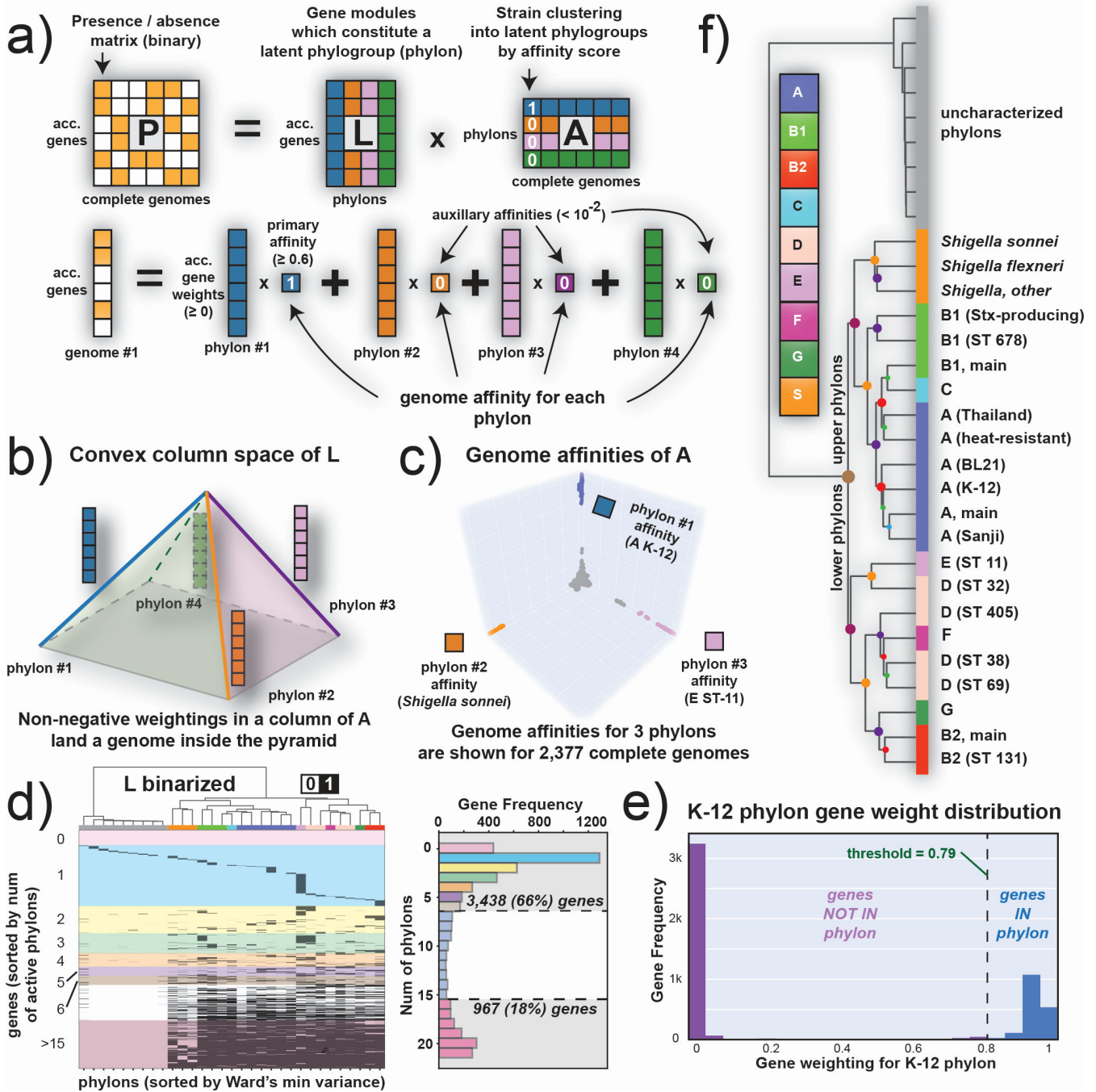


FIG 3 The fundamental mathematical structure of the *E. coli* accessory genome. Characteristics of the NMF decomposition of the pangenome matrix P . (a) A column of P (i.e., genome #1) is a linear combination of the phylon vectors as determined by the weights in the corresponding column of A . (b) Since the phylon vectors are non-negative, they span a polygon as its edge vectors. A positive linear combination of the L_i vectors lands inside the polygon. (c) Since there is typically only one dominant value in a column of A , the reconstruction of a column in P (i.e., one genome) lies close to a phylon vector (i.e., the edges of the polygon) as is evident for the 2,377 sequenced strains used. (d) A clustermap of the binarized L matrix. Colors on top correspond with classically defined phylogroups as determined by ClermonTyping. Columns are clustered using Ward's minimum variance method, and rows are sorted by gene frequency in each phylon (i.e., genes in zero phylons are at the top, genes in 22 phylons are at the bottom). The dendrogram at the top of L , showing the clustering of its columns, is the same as that used in panel (f). In this graphical representation the black elements designate that the gene responding to that row is found in the phylon that the column represents. White elements mean that the corresponding gene is not found in the phylon. The histogram to the right of the clustered L matrix showcases the gene frequency across multiple phylons (i.e., how many phylons a gene is present in). The colors in L -binarized correspond to the colors on this histogram and showcase the distribution of genes by their number of active phylons; 3,438 (66%) of the 5,182 accessory genes are found in six or fewer phylons, (Continued on next page)

Fig 3 (Continued)

with the plurality being genes active in only one phylon (1,289 single-phylon genes, 25% of all 5,182 accessory genes). (e) A gene weight distribution for one particular phylon consisting of K-12 strains in the L matrix. Most genes have a weighting close to zero, with a notable cluster having weightings between 0.8 and 1. The genes with low weightings (below the threshold indicated by the dashed line) are binarized to zero and considered not to be part of the phylon, while genes with high weightings are binarized to one and considered to be constituents of this phylon. The threshold for binarization is determined for each phylon using k -means clustering (see **Methods**). (f) A dendrogram of all 31 phylons based on clustering the binarized L matrix shown in panel (d). The uncharacterized phylons are separated, mainly consisting of phage genes and other mobile elements.

must be non-negative. Each column of L (a phylon vector) represents an edge of a polygon (Fig. 3b).

NMF gives a clear mathematical description of the gene portfolio of a phylogroup found in the *E. coli* pangenome. The gene list found in all strains of a phylogroup is given by a phylon, or a column in L . Few strains will correspond perfectly to a phylon as its gene list may differ slightly from that given by the columns in L . The affinity matrix, A , shows how close a strain is to a phylon as the elements in a column in A give the phylon composition of a particular strain. This feature is demonstrated with the color coding of the matrices in Fig. 3a. The 3D image of the location of all strains relative to three of the columns of L is shown in Fig. 3c for all 2,377 genomes in this study. Strains of a phylogroup are close to one of the phylon vectors shown (*i.e.*, high affinity for the phylon), while the rest of the strains that are not in these three phylogroups are close to the origin (*i.e.*, low affinity for these phylons).

Almost all strains have a dominant phylon as they lie close to one edge of the column space. Thus, the affinity scores in the column of the A matrix that corresponds to a particular strain places each genome inside the convex solution space. Most of these affinities are small and close to zero, typically with only one dominant affinity per genome, revealing that most strains reside close to the edges of the convex space. An image of the binarized form of A is shown in (Fig. S4)

Biological meaning of the pangenome's mathematical structure

The columns of L show that NMF breaks P into the eight classically defined phylogroups (plus *Shigella*, see **Methods**) and sub-phylogroups thereof (Fig. 3d). There are 22 of these columns, and then an additional nine unclassified columns of L that represent mobile elements (see below). The L matrix shown in Fig. 3d is binarized. The weightings are close to unity (gene in phylon) or zero (gene not in phylon), as shown in Fig. 3e.

Thus, the NMF decomposition of P for the accessory genome reveals phylons defined by their list of genes. It also shows how each strain's gene set maps onto these phylons. NMF segregates genes to a phylon concordant with previous phylo-grouping methods of strains in *E. coli* (Fig. 3f): the Mash distances (Fig. 1d), the Clermont quadruplex, and the MLST typing. NMF allows us to go from differential traits between phylogroups to their genetic basis.

Utilizing the binarized L and A matrices, we can multiply them to generate a reconstructed P matrix of gene presence/absence that we can compare with the original table (which serves as our ground truth). From this comparison, we find that this reconstructed P matrix has an accuracy of 87% (Table S1). This showcases how well NMF-derived phylons approximate the actual structure of the pangenome. With few false positives (false positive rate of 0.04), phylon membership is a conserved estimate of all co-occurrent gene groupings.

The *E. coli* pangenome consists of two distinct groups of phylons

The phylons are divided into two major groups (Fig. 3f). One group, which we collectively call the lower phylons – G, B2, D, F, and E – is genetically dissimilar to the strains found in the upper phylons that correspond to phylogroups A, B1, C, and *Shigella*. This is the first split in the hierarchical phylon clustering tree. Of the 5,182 genes found in the accessory genome, 765 are found exclusively in strains of the upper phylons, with 1,244

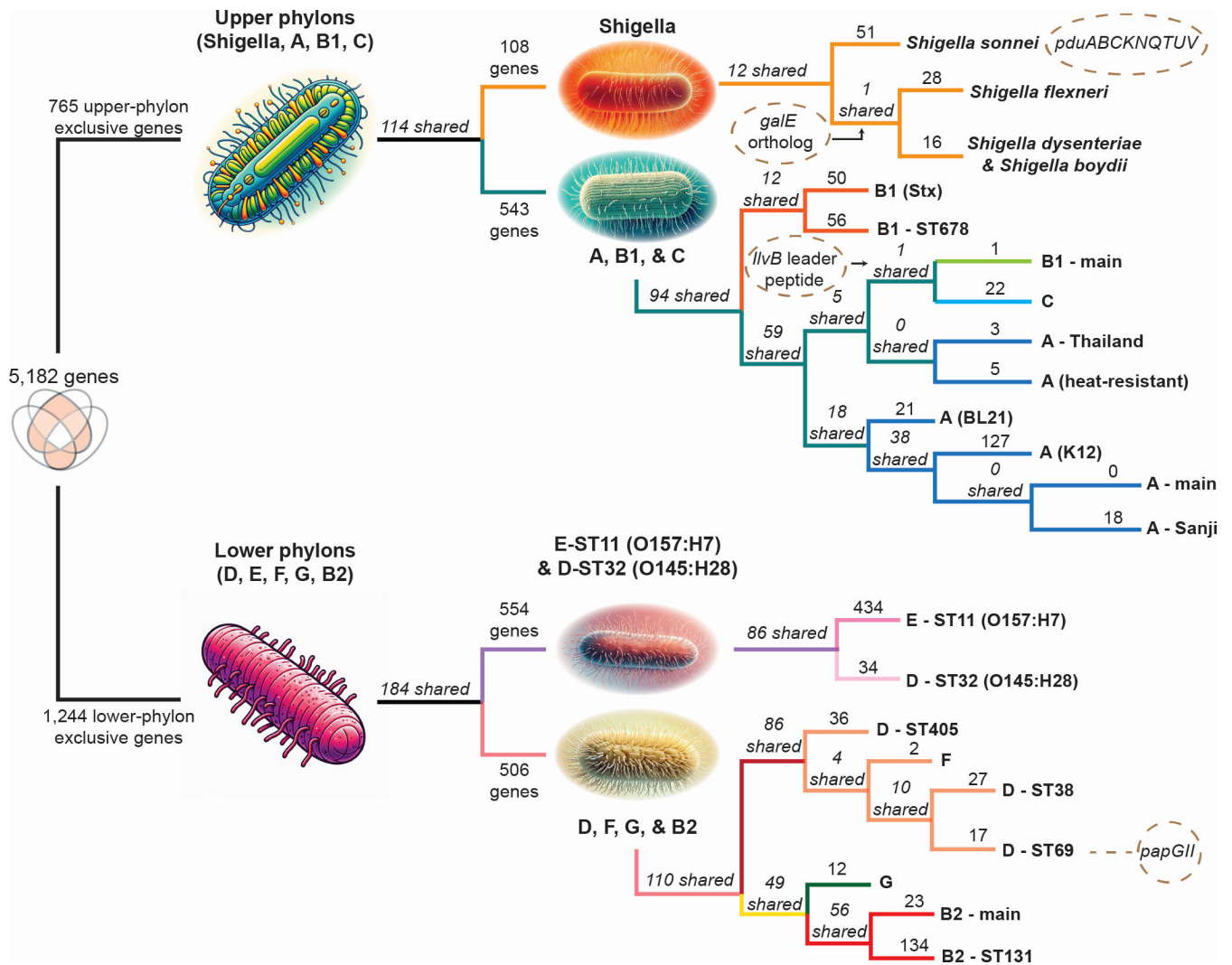


FIG 4 A clustering diagram of the phylons (see Fig. 3f) that highlights the groups of exclusive genes that follow one branch and not the other at each branch point. The numbers above the line leading to a split indicate exclusive genes (*i.e.*, genes found in one group of phylons but absent in the other). Numbers in italics specifically indicate shared genes that are found across all groups. The function and identity of special genes of interest are discussed in the main text, and detailed in Table S2. Four specific genetic traits of interest are highlighted in dashed ovals, such as the *papGII* operon to phylon D (ST69). This sequence variant of *papG* in this operon is associated with UTIs that can become bacteremic (28, 29).

found exclusively in strains of the lower phylons (Fig. 4). While some of these genes have no known function/ortholog, many do. In fact, 98 of the 765 upper phylon-exclusive genes have a known metabolic function, with an additional 34 having a known motility function. Similarly, 213 of the 1,244 lower phylon-exclusive genes are metabolic in nature, with 73 having motility functions. The distribution of genes in many other functional categories (such as transcription factors, metabolic functions, pili, motility, membrane-, and phage-genes) are highlighted in Table S2. The columns of L thus gives us detailed information about the differential gene contents of the phylons, and thus gives the basis for finding the genetic basis for differential traits between phylogroups.

A further examination of the metabolic gene content between these two distinct groups of phylons reveals that the upper phylon-exclusive genes code for glycosyltransferases, exo-alpha-sialidases (glycoconjugates), beta-glucosidases, beta-glucuronidases, hyaluronidase, and monogalactosyldiacylglycerol lipases (monogalactosyl-diacylglycerols), among others. These enzymes are all involved in processing and digesting glycans in a host, including sialic acids. Sialic acids and their derivatives often form the ends

of glycans on various glycoproteins and glycolipids that coat the surfaces of most vertebrate and bacterial cells. They are known to behave as a signal to specific bacteria upon reaching a vertebrate environment suitable for colonization (30, 31). The presence of these enzymes in the upper phylons suggests that *Shigella*, Shiga toxin-producing *E. coli* (STEC), and other pathogenic strains in phylogroups A, B1, and C have the ability to identify and process these sugar moieties, which may enable them to better colonize their hosts.

A parallel inspection into the lower phylon-exclusive metabolic genes reveals that they code for various bacterial capsule formation proteins, Amadori product degradation (specifically fructoselysine/psicoselysine degradation), as well as three distinct fructose-bisphosphate aldolases, one of which is GatY (10). All of these are known proteins found in tagatose-competent *E. coli* strains (primarily found in phylogroups B2 and D), which can utilize these enzymes to digest the glycans in the mucus within the human GI tract (10).

The secondary splits in the phylon clustering tree allow further tracing of genes and thus segregation of the genetic basis for traits. Three further splits are discussed here. A full description of the classification tree requires a comprehensive study that will result in the full genetic definition of *E. coli* and all its phylogroups.

***Shigella* strains exhibit gene gains as well as gene losses**

Continuing the segregation of genes down the tree of phylons, as defined by NMF (Fig. 3f and 4), we see that the upper phylons split by gene content into *Shigella* strains and those that belong to the classically defined A, B1, and C phylogroups. A closer look at this split reveals that the strains in the *Shigella* phylons contain 108 exclusive genes not found in the A, B1, and C strains, in addition to not containing 543 genes found in these three classically defined phylogroups. This suggests that *Shigella* has undergone both gene gain and loss during its restriction to human hosts and adaptation to the human intestinal mucosa (32). For example, in the *nadA* and/or *nadB* genes encoding the enzyme complex that converts L-aspartate to quinolinate, a precursor to NAD resulting in nicotinic acid auxotrophy is lost in these strains (33). However, other genes are gained, including nine genes that form the propanediol utilization (*pdu*) operon. Propanediol is produced when fucose (a component of mucin) is metabolized under anaerobic conditions (34). Of particular note is the *pduC* gene in this operon, which is enriched in adherent-invasive *E. coli* found in the microbiome of Crohn's disease patients (35). Note that *Shigella sonnei* is separated from the rest of the *Shigella* strains, which all contain a *galE* ortholog that *S. sonnei* itself lacks (Fig. 4).

Pathogenic A strains are more closely related to B1 and C strains than to commensal A strains

The upper phylons split into *Shigella* strains and strains in A, B1, and C phylogroups. The A, B1, and C strains further split with Shiga-toxin producing *E. coli* strains (B1-Stx and B1-ST678) forming their own subgroup. Interestingly, the next split between these strains in the other branch occurs between commensal A strains and B1, C, and pathogenic A strains. Specifically, these pathogenic A strains are those that came from foodborne illnesses in Thailand (A-Thailand) and those found to be heat-resistant in meat (A-heat-resistant). The commensal A strains are heavily used in laboratory work and in biomanufacturing (specifically the K-12 MG1655 strain).

Lower phylons have a subgroup containing *E. coli* O157:H7 and O145:H28 strains

The lower phylons split into two subgroups, with E-ST11 (O157:H7) and D-ST32 (O145:H28) strains separating from the other group of lower phylons (see Fig. 4). These two serovars (O157:H7 and O145:H28) are known to have shared a common evolutionary lineage (36). This shared lineage is directly reflected in their shared gene content, with 86 shared (accessory) genes between them. Of these genes, 33 have no known function, while 11

of them are metabolic, 10 are motility-related, and nine are transcription factors. The 11 metabolic genes primarily code for various transporters, oxidoreductases involved in glycolytic pathways, and a class-II fructose-bisphosphate aldolase. Furthermore, all motility genes code for fimbrial gene orthologs of the *yadCKLM-htrE-yadVN* operon. This operon is cryptic under normal laboratory conditions but when constitutive expression is induced, it promotes biofilm formation in minimal media on a variety of abiotic surfaces and produces surface fimbrial structures (37). Constitutive expression of this operon also results in increased adhesion of cells to xylose-rich glycans, increased adherence to intestinal epithelial cells, and can also modulate the inflammatory response of host cells (38).

D and F strains are very closely linked, as are G and B2 strains

The remaining lower phylons consist of D, E, F, and B2 strains. These strains cluster into two distinct groups: the first group consists of D and F strains, while the second group consists of G and B2 strains. The first cluster shares 86 genes, and D-ST405 strains appear to be the most distinct, even more so than F strains. This clustering suggests that the classically-defined D and F phylogroups of *E. coli* are more closely related to each other genetically than previously thought. This is similarly the case for G strains, which were already known to be more closely related to F and B2 strains than others (22).

Uncharacterized phylons contain mobile genetic elements

The same mobile elements can be found in strains of many phylogroups. Remarkably, NMF detects this characteristic and factors out these mobile genes into a set of nine “uncharacterized” phylons since they can be columns in the addition that forms the gene set in a strain (see Fig. 3b). These mobile elements are described in Table S2 and include sex pili, F-plasmid operons, and various phage genes, among others. The mobilome of *E. coli* is nine-dimensional.

Traits found in the rare genome

The rare genome consists of 163,619 genes (Fig. 2b), of which 127,223 (or 79%) are transposable elements (TEs) (Fig. S5a). These TEs fall into 315 unique categories of TEs. The number of the 40 most frequent TEs and the number of passenger genes (a.k.a cargo genes) they carry are shown in Fig. 5a. About 3% of the most frequent TEs carry 773 unique passenger genes, that with replicate occurrences give a total count of 3,631 rare genes. These 3,631 genes fall into 24 functional categories, of which the largest COG category is unknown function (S), followed by energy production and conversion (C) and transcription (K) (Fig. S5b). Thus, the genetic diversity of the rare genome is effectively much narrower than its raw gene count indicates.

In the Tn3 family, particularly within the *TnAs1* transposase family, a notably higher ratio of passenger-associated TEs (42%) was observed compared to all other TEs categories; interestingly, studies have identified the Tn3 family of transposons as a factor in the recent surge of carbapenem- and colistin-resistant Enterobacteriaceae (39–42). The richness (representing the number of distinct types of TEs found in a phylon) and entropy (indicating the degree of unevenness in the distribution of various transposable element types within a phylon, with higher values suggesting a less uniform population) of TEs are greater in uncharacterized phylons, indicating higher diversity and greater heterogeneity in the distribution of TEs (Fig. 5b and c). In three out of the four phylons related to phylogroup A, comparatively low levels of TE entropy, richness, and count were observed. This result suggests that reduced replicative activity of TEs is exhibited within these phylons. Furthermore, these phylons are characterized by the dominance of a specific TE type, namely *yhnl* for A-Thailand and A-BL21, and “*insH11*” for A-K12. Interestingly, similar patterns were displayed by all *Shigella* phylons, featuring low richness and moderate to high entropy. Notably, a common theme across these phylons was the prevalence of *insG* as the dominant TE (Fig. 5b).

Fig 5 (Continued)

cell in the heatmap represents the presence (black) or absence (white) of a specific group of TEs in a given phylon. The hierarchical clustering, employing the Ward method, is applied both horizontally and vertically, illustrating the grouping of similar TEs and phylons, respectively. The dendrograms adjacent to the rows and columns indicate the clustering relationships.

The remaining 21% of genes in the rare genome result from TE insertions that fracture coding regions into smaller coding regions (43) or horizontal gene transfer (HGT). As an example of HGT, a gene family found in the rare genome is *lapA*, encoding for a large adhesion pilus of over 6,000 amino acids in length, and commonly found in *Pseudomonas* species (see details in SI).

DISCUSSION

Taxonomy going back to Linnaeus' time was based on the form and phenotypic function of organisms. Then, classification based on genetics, such as certain alleles, emerged (*e.g.*, multi-locus sequence tags [44], and Clermont quadruplex [13]). With whole-genome sequences becoming available, the MASH distance (19) could be used to assess the relatedness of strains using genome-scale sequence similarity metrics. Now with genome sequences for a large number strains becoming available, we can annotate them and determine the species' pangenome.

The pangenome is formally represented by the pangenome matrix, P , whose columns represent genomes and whose rows represent genes. Every column is then filled in with 0 or 1, the absence or presence of the gene in that genome, respectively. Since the number of sequenced genomes is now large, mathematical and machine learning methods can be applied to formulate global and rigorous classification schemes for a strain's phylogeny based on the pangenome matrix. Such classification schemes are fundamental and will be at the root of bacterial taxonomy.

In this study, we developed a classification schema based on the pangenome matrix using methods of machine learning. Remarkably, this approach gives a very clear definition of the gene content that differentiates strains that closely follows classical phylogroup definitions. The variably present genes populate the accessory genome, whose gene distribution amongst the strains can be used to obtain a mathematical definition of phylons, which are lists of genes that are found in the majority of the strains of a phylogroup.

This study enables a detailed, genomewide analysis of the genetic basis for the differential traits of the strains in the defined phylons, and it gives a global multi-scale genetic structure of a species. This full exposé of the genetic composition of a bacterial species has many implications. With the availability of the alleleome (45), representing the global assessment of sequence variation of coding and intergenic regions, we can begin to understand the evolutionary history of a species and its phylons. The rare genome can keep track of horizontal gene transfer events and how they are assimilated into the species and can provide new insights for understanding unique traits of a particular strain. Additionally, the hypothesized fractal nature of the *E. coli* pangenome can be further studied (*e.g.*, the most distinct phylon consists of phylogroup E strains which themselves are known to be quite diverse) (46). The differentiation of phylons into classically defined phylogroups and MLST clusters versus the uncharacterized phylons, which are primarily full of plasmid and mobilome elements, is also another aspect that is worth investigating. The NMF's ability to cleanly separate these groups of phylons alludes to its ability to distinguish between classical (Darwinian) evolution in *E. coli* vs horizontal gene acquisition (Woesian evolution) (Fig. 3). Remarkably, these different evolutionary pathways, while convoluted with each other in the genome, are cleanly separated based on pure gene presence/absence when looking at the species's pangenome.

Detailing the phenotypic consequences based on the differential gene presence may take many years to fully resolve for all genetic traits of interest. This undertaking may have a fundamental effect on infectious disease. Reliable sequence-based rapid

classification of pathogens isolated from a patient can accelerate physicians' decision-making about pathogen identity and selection of treatment modalities.

One can anticipate that with a good coverage of genome sequences across the phylogenetic tree, we will be able to repeat the results of this study for larger and larger swaths of the tree. Currently GENOMICUS only contains *E. coli*, but perhaps, in the fullness of time, we will achieve a global genetic definition of the entire phylogenetic tree of bacteria.

MATERIALS AND METHODS

Gathering and processing of sequence data from BV-BRC and NCBI RefSeq

We first downloaded the metadata of all genomes available on BV-BRC at the onset of this project (2021) (5). Using this metadata file, we filtered out strains that were not *E. coli* and those which only contained plasmid sequences. All "complete" sequences were further filtered by their L50 score (must equal 1) and their N50 score (greater than 4,000,000). Fragmented genomes were first filtered by their contig count, which was capped at 355 using previously defined metrics (23). CheckM contamination (<3.1%) and completeness (>98.1%) scores were then used to filter fragmented genomes further (see code for more details on exact numerical thresholds chosen). The final collection was then downloaded from BV-BRC. This exact process was repeated for all *Shigella* strains and similarly for downloading *E. coli* strains from NCBI RefSeq. Genomes were then deduplicated and collated for further quality control (*i.e.*, Mash filtration). In the end, only "complete" sequences were selected for pangenome analysis to ensure the pangenome had a gene presence/absence matrix (P matrix) of the highest quality.

Genome annotation and pangenome generation

All downloaded genomes were re-annotated using PROKKA (47) for consistency in gene annotation when generating the pangenome. All re-annotated genomes were then screened by the *E. coli* PubMLST schema (44) through the mlst github package (48) to identify the sequence types for all strains in the pangenome. After this, the phylogroup for each strain was identified *in silico* using the ClermonTyping github package (18). Genomes were then collated to form a pangenome using CD-HIT (49, 50). Gene families were identified using a sequence similarity and alignment cutoffs of 80% for both, as used in previous pangenome studies (23). Once the pangenome was generated, all representative alleles that define a gene family as identified by CD-HIT were extracted and subjected to eggNOG gene annotation (51–53). Genomes were also annotated for AMR gene annotation using the Resistance Gene Identifier tool (54).

Mash filtration and analysis

All downloaded genomes were run to generate pairwise Mash distance values. They were then separated into six groups: *Escherichia coli*, *Shigella sonnei*, *Shigella boydii*, *Shigella dysenteriae*, *Shigella flexneri*, and other *Shigella* species. For each group, the 99th percentile was calculated relative to the reference strain for each group and used as the filtration limit (that is, the top 1% of genomes in terms of Mash distance were filtered out for each group). Then, the Mash distance values were converted into Pearson correlation coefficients, which in turn were converted into Pearson correlation distances for Mash clustering, as outlined in Abram *et al.* (14). A sensitivity analysis was performed to find the best threshold for clustering these values, which led to a value of 0.1. Specifically, the cutoff threshold for hierarchical clustering using seaborn's inbuilt clustermap function was set to various values until all major phylogroups of *E. coli* were represented. This value (0.1) agreed with known domain knowledge, as Phylogroup C strains did not form their own cluster for any threshold value above 0.13 (14). This led to a total of 31 clusters, which was used to inform the rank of NMF decomposition.

Defining the core, accessory, and rare genomes

The core, accessory, and rare genomes are defined using the cumulative gene distribution plot (Fig. 2b) using methods outlined earlier (23). Briefly, this gene plot forms an S-shaped curve and thus will always have an inflection point. The core genome is defined by taking the highest endpoint and traveling 90% of the distance from the inflection point to the endpoint. This corresponds with the elbow in the plot defining the core genes. A similar approach is used for defining the rare genome, except with the lowest endpoint instead of the highest one.

For the TE analysis, transposable elements initially identified by PROKKA annotation were filtered. The start and end locations of each transposable element were determined on genomes across the entire pangenome. Genes that were not classified as transposable elements but had start and end locations within a transposable element were designated as passenger genes.

TEs richness and evenness calculation

A systematic genome sampling approach was implemented to ensure a representative and manageable data set for this analysis. Phylons with fewer than 30 genomes were excluded from the analysis. For phylons having 30 or more genomes, we conducted a random sampling procedure to select 30 unique genomes.

This genome sampling strategy guaranteed that the data set used for richness and evenness calculations was not only representative but also possessed a more even distribution. Consequently, it allowed for meaningful insights into TE diversity across various phylogroups. Fig. S5c illustrates the sensitivity of TE richness, evenness, and count metrics to the random sampling of genomes.

Richness, representing the number of unique TEs within each phylon, was computed. Evenness, quantifying the distribution uniformity of TEs within a phylon, was calculated using the Shannon entropy formula.

Non-negative matrix factorization

The scikit-learn implementation of NMF (55) was used to perform the decomposition. NMF was run 50 times with a rank of 31 (derived from Mash clustering), an initialization of "nndsvd" (which generates sparser output matrices), and a maximum iteration limit of 5,000 (the solution always converged before this limit was reached for all runs). The best run (as defined by the Frobenius norm, sum of squared residuals, and root-mean-square-error metrics) was selected for normalization. For each column in **L**, the 99th percentile was calculated, and every value in that column was divided by this value, which ensured all but a few values were between 0 and 1. To ensure reconstruction consistency, the corresponding rows in the **A** matrix were multiplied by the same normalization values (see SI for more information). The **L** and **A** matrices were then binarized using k-means clustering ($k = 3$), also implemented using scikit-learn. Each column of **L** was segregated into three clusters, and the genes in the cluster with the highest average mean were binarized to 1, with the genes in the other two clusters being set to 0. The same procedure was followed with binarizing the **A** matrix. This protocol ensured the threshold for binarization was always a conservative estimate. In some cases, this estimate in the **A** matrix was clearly too conservative (as evidenced by visualizing the histogram of strain affinities for each phylon), and in those cases, the threshold for binarization was manually lowered.

Phylon characterization

Given the large number of genes in the **L** matrix, phylons were initially characterized using the (binarized) **A** matrix. Phylons were first named based on the phylogroup of the strains with the highest affinity for each phylon, followed by the MLST value of these strains. If at least 90% of strains with high affinity were part of a particular

Clermont-defined phylogroup, that phylon was mapped to that Clermont-defined phylogroup (e.g., if 90% of all strains in phylon2 map to phylogroup B2, that phylon is mapped to B2). In certain cases, the names of the phylons were changed to reflect well-known strains within the phylon (e.g., A-K12, A-BL21, etc.). Some phylons did not follow these patterns and were thus dubbed “uncharacterized” phylons. Strains with “high affinity” are defined as those strains that had an entry of 1 in the binarized A matrix for a particular phylon. For most strains, this only occurred once. For some strains, this occurred multiple times; in all cases, the strain had a high affinity for only one named phylon as the other high affinities were all for the uncharacterized phylons that consisted of mobile genetic elements.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Akanksha Rajput for useful discussions.

This work was funded by the Novo Nordisk Foundation Grant Number NNF20CC0035580.

Conceptualization: B.O.P. and S.M.C., Data curation: S.M.C. and P.V.P., Investigation: S.M.C. and O.A., Methodology: S.M.C., J.M.M., and J.C.H., Mentorship: J.M.M. and B.O.P., Writing and Editing: All authors.

AUTHOR AFFILIATIONS

¹Department of Bioengineering, University of California, San Diego, La Jolla, California, USA

²Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Kongens Lyngby, Denmark

³Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, California, USA

⁴Department of Pediatrics, University of California, San Diego, La Jolla, California, USA

AUTHOR ORCIDs

Siddharth M. Chauhan  <http://orcid.org/0000-0001-6674-895X>

Patrick V. Phaneuf  <http://orcid.org/0000-0002-4122-6589>

Bernhard O. Palsson  <http://orcid.org/0000-0003-2357-6785>

AUTHOR CONTRIBUTIONS

Siddharth M. Chauhan, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing | Omid Ardalani, Formal analysis, Investigation, Visualization, Writing – original draft | Jason C. Hyun, Methodology, Software, Validation | Jonathan M. Monk, Conceptualization, Supervision, Validation, Writing – review and editing | Patrick V. Phaneuf, Data curation, Supervision, Validation | Bernhard O. Palsson, Conceptualization, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

All data (and code) pertaining to this study have been deposited onto Zenodo and can be found with this DOI: [10.5281/zenodo.10575748](https://doi.org/10.5281/zenodo.10575748).

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Supplemental material (mSphere00532-24-s0001.pdf). Supplemental figures and text.
Table S1 (mSphere00532-24-s0002.pdf). NMF model metrics.

Table S2 (mSphere00532-24-s0003.pdf). Specific genes found in each phylon uniquely.**REFERENCES**

- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512. <https://doi.org/10.1126/science.7542800>
- Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–1462. <https://doi.org/10.1126/science.277.5331.1453>
- Perna NT, Plunkett G III, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, et al. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409:529–533. <https://doi.org/10.1038/35054089>
- Kris A, Wetterstrand MS. 2019. The cost of sequencing a human genome. NHGRI. Available from: <https://www.genome.gov/about-genomics-factsheets/Sequencing-Human-Genome-cost>. Retrieved 18 Apr 2023.
- Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, et al. 2023. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* 51:D678–D689. <https://doi.org/10.1093/nar/gkac1003>
- O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733–D745. <https://doi.org/10.1093/nar/gkv1189>
- Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebahia M, Thomson NR, Chaudhuri R, Henderson IR, Sperandio V, Ravel J. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190:6881–6893. <https://doi.org/10.1128/JB.00619-08>
- Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5:e1000344. <https://doi.org/10.1371/journal.pgen.1000344>
- Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, Feist AM, Palsson BØ. 2013. Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc Natl Acad Sci U S A* 110:20338–20343. <https://doi.org/10.1073/pnas.1307797110>
- Fang X, Monk JM, Mih N, Du B, Sastry AV, Kavvas E, Seif Y, Smarr L, Palsson BO. 2018. *Escherichia coli* B2 strains prevalent in inflammatory bowel disease patients have distinct metabolic capabilities that enable colonization of intestinal mucosa. *BMC Syst Biol* 12:66. <https://doi.org/10.1186/s12918-018-0587-5>
- Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 60:1136–1151. <https://doi.org/10.1111/j.1365-2958.2006.05172.x>
- Clermont O, Bonacorsi S, Bingen E. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* 66:4555–4558. <https://doi.org/10.1128/AEM.66.10.4555-4558.2000>
- Clermont O, Christenson JK, Denamur E, Gordon DM. 2013. The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep* 5:58–65. <https://doi.org/10.1111/1758-2229.12019>
- Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, Robeson MS II, Ussery DW. 2021. Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. *Commun Biol* 4:117. <https://doi.org/10.1038/s42003-020-01626-5>
- Achtman M, Zhou Z, Charlesworth J, Baxter L. 2022. EnteroBase: hierarchical clustering of 100 000s of bacterial genomes into species/subspecies and populations. *Philos Trans R Soc Lond B Biol Sci* 377:20210240. <https://doi.org/10.1098/rstb.2021.0240>
- Norsigian CJ, Kavvas E, Seif Y, Palsson BO, Monk JM. 2018. iCN718, an updated and improved genome-scale metabolic network reconstruction of *Acinetobacter baumannii* AYE. *Front Genet* 9:121. <https://doi.org/10.3389/fgene.2018.00121>
- Seif Y, Kavvas E, Lachance J-C, Yurkovich JT, Nuccio S-P, Fang X, Catoiu E, Raffatellu M, Palsson BO, Monk JM. 2018. Genome-scale metabolic reconstructions of multiple *Salmonella* strains reveal serovar-specific metabolic traits. *Nat Commun* 9:3771. <https://doi.org/10.1038/s41467-018-06112-5>
- Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. 2018. ClermontTyping: an easy-to-use and accurate *in silico* method for *Escherichia* genus strain phylotyping. *Microb Genom* 4:e000192. <https://doi.org/10.1099/mgen.0.000192>
- Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 17:132. <https://doi.org/10.1186/s13059-016-0997-x>
- Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102:2567–2572. <https://doi.org/10.1073/pnas.0409727102>
- Lu S, Jin D, Wu S, Yang J, Lan R, Bai X, Liu S, Meng Q, Yuan X, Zhou J, Pu J, Chen Q, Dai H, Hu Y, Xiong Y, Ye C, Xu J. 2016. Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of Marmota himalayana in Qinghai-Tibet plateau of China. *Emerg Microbes Infect* 5:e122. <https://doi.org/10.1038/emi.2016.122>
- Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S, Bridier-Nahmias A, Denamur E, Gordon D. 2019. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ Microbiol* 21:3107–3117. <https://doi.org/10.1111/1462-2920.14713>
- Hyun JC, Monk JM, Palsson BO. 2022. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics* 23:7. <https://doi.org/10.1186/s12864-021-08223-8>
- Rajput A, Seif Y, Choudhary KS, Dalldorf C, Poudel S, Monk JM, Palsson BO. 2021. Pangenome analytics reveal two-component systems as conserved targets in ESKAPEE pathogens. *mSystems* 6:e00981–20. <https://doi.org/10.1128/mSystems.00981-20>
- Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, Takeuchi R, Nomura W, Zhang Z, Mori H, Feist AM, Palsson BO. 2017. iML1515, a knowledge-base that computes *Escherichia coli* traits. *Nat Biotechnol* 35:904–908. <https://doi.org/10.1038/nbt.3956>
- Lee DD, Seung HS. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791. <https://doi.org/10.1038/44565>
- Devarajan K. 2008. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol* 4:e1000029. <https://doi.org/10.1371/journal.pcbi.1000029>
- Cuénod A, Agnetti J, Seth-Smith HMB, Roloff T, Wälchli D, Shcherbakov D, Akbergenov R, Tschudin-Sutter S, Bassetti S, Siegemund M, Nickel CH, Moran-Gilad J, Keys TG, Pflüger V, Thomson NR, Egli A. 2023. Bacterial genome-wide association study substantiates *papGII* of *Escherichia coli* as a major risk factor for urosepsis. *Genome Med* 15:89. <https://doi.org/10.1186/s13073-023-01243-x>
- Biggel M, Xavier BB, Johnson JR, Nielsen KL, Frimodt-Møller N, Matheeußen V, Goossens H, Moons P, Van Puyvelde S. 2020. Horizontally acquired *papGII*-containing pathogenicity islands underlie the emergence of invasive uropathogenic *Escherichia coli* lineages. *Nat Commun* 11:5968. <https://doi.org/10.1038/s41467-020-19714-9>
- Varki A, Gagneux P. 2012. Multifarious roles of sialic acids in immunity. *Ann N Y Acad Sci* 1253:16–36. <https://doi.org/10.1111/j.1749-6632.2012.06517.x>
- Schauer R, Kamerling JP. 2018. Exploration of the sialic acid world. *Adv Carbohydr Chem Biochem* 75:1–213. <https://doi.org/10.1016/bs.accb.2018.09.001>
- The HC, Thanh DP, Holt KE, Thomson NR, Baker S. 2016. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nat Rev Microbiol* 14:235–250. <https://doi.org/10.1038/nrmicro.2016.10>

33. Di Martino ML, Fioravanti R, Barbabella G, Prosseda G, Colonna B, Casalino M. 2013. Molecular evolution of the nicotinic acid requirement within the *Shigella*/EIEC pathotype. *Int J Med Microbiol* 303:651–661. <https://doi.org/10.1016/j.ijmm.2013.09.007>
34. Dogan B, Suzuki H, Herlekar D, Sartor RB, Campbell BJ, Roberts CL, Stewart K, Scherl EJ, Araz Y, Bitar PP, Lefébure T, Chandler B, Schukken YH, Stanhope MJ, Simpson KW. 2014. Inflammation-associated adherent-invasive *Escherichia coli* are enriched in pathways for use of propanediol and iron and M-cell translocation. *Inflamm Bowel Dis* 20:1919–1932. <https://doi.org/10.1097/MIB.0000000000000183>
35. Viladomiu M, Metz ML, Lima SF, Jin W-B, Chou L, Bank JLC, Guo C-J, Diehl GE, Simpson KW, Scherl EJ, Longman RS. 2021. Adherent-invasive *E. coli* metabolism of propanediol in Crohn's disease regulates phagocytes to drive intestinal inflammation. *Cell Host Microbe* 29:607–619. <https://doi.org/10.1016/j.chom.2021.01.002>
36. Cooper KK, Mandrell RE, Louie JW, Korch J, Clark TA, Parker CT, Huynh S, Chain PS, Ahmed S, Carter MQ. 2014. Comparative genomics of enterohemorrhagic *Escherichia coli* O145:H28 demonstrates a common evolutionary lineage with *Escherichia coli* O157:H7. *BMC Genomics* 15:17. <https://doi.org/10.1186/1471-2164-15-17>
37. Korea C-G, Badouraly R, Prevost M-C, Ghigo J-M, Beloin C. 2010. *Escherichia coli* K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. *Environ Microbiol* 12:1957–1977. <https://doi.org/10.1111/j.1462-2920.2010.02202.x>
38. Larssonneur F, Martin FA, Mallet A, Martinez-Gil M, Semetey V, Ghigo J-M, Beloin C. 2016. Functional analysis of *Escherichia coli* Yad fimbriae reveals their potential role in environmental persistence. *Environ Microbiol* 18:5228–5248. <https://doi.org/10.1111/1462-2920.13559>
39. Cuzon G, Naas T, Nordmann P. 2011. Functional characterization of Tn4401, a Tn3-based transposon involved in *bla*_{KPC} gene mobilization. *Antimicrob Agents Chemother* 55:5370–5373. <https://doi.org/10.1128/AAC.05202-11>
40. Nordmann P, Dortet L, Poirel L. 2012. Carbapenem resistance in *Enterobacteriaceae*: here is the storm! *Trends Mol Med* 18:263–272. <https://doi.org/10.1016/j.molmed.2012.03.003>
41. Borowiak M, Fischer J, Hammerl JA, Hendriksen RS, Szabo I, Malorny B. 2017. Identification of a novel transposon-associated phosphoethanolamine transferase gene, *mcr-5*, conferring colistin resistance in *d*-tartrate fermenting *Salmonella enterica* subsp. *enterica* serovar Paratyphi B. *J Antimicrob Chemother* 72:3317–3324. <https://doi.org/10.1093/jac/dkx327>
42. Zhang H, Zong Z, Lei S, Srinivas S, Sun J, Feng Y, Huang M, Feng Y. 2019. A genomic, evolutionary, and mechanistic study of MCR-5 action suggests functional unification across the MCR family of colistin resistance. *Adv Sci (Weinh)* 6:1900034. <https://doi.org/10.1002/adv.201900034>
43. Sheng Y, Wang H, Ou Y, Wu Y, Ding W, Tao M, Lin S, Deng Z, Bai L, Kang Q. 2023. Insertion sequence transposition inactivates CRISPR-Cas immunity. *Nat Commun* 14:4366. <https://doi.org/10.1038/s41467-023-39964-7>
44. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 3:124. <https://doi.org/10.12688/wellcomeopenres.14826.1>
45. Catoiu EA, Phaneuf P, Monk J, Palsson BO. 2023. Whole-genome sequences from wild-type and laboratory-evolved strains define the alleloome and establish its hallmarks. *Proc Natl Acad Sci U S A* 120:e2218835120. <https://doi.org/10.1073/pnas.2218835120>
46. Clermont O, Condamine B, Dion S, Gordon DM, Denamur E. 2021. The *E. coli* phylogroup of *Escherichia coli* is highly diverse and mimics the whole *E. coli* species population structure. *Environ Microbiol* 23:7139–7151. <https://doi.org/10.1111/1462-2920.15742>
47. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
48. Seemann T. mlst: :id: scan contig files against PubMLST typing schemes. Github. Available from: <https://github.com/tseemann/mlst>. Retrieved 14 Nov 2023.
49. Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17:282–283. <https://doi.org/10.1093/bioinformatics/17.3.282>
50. Li W, Fu L, Niu B, Wu S, Wooley J. 2012. Ultrafast clustering algorithms for metagenomic sequence analysis. *Brief Bioinform* 13:656–668. <https://doi.org/10.1093/bib/bbs035>
51. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47:D309–D314. <https://doi.org/10.1093/nar/gky1085>
52. Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18:366–368. <https://doi.org/10.1038/s41592-021-01101-x>
53. 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. *Mol Biol Evol* 38:5825–5829. <https://doi.org/10.1093/molbev/msab293>
54. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, Edalatmand A, Petkau A, Syed SA, Tsang KK, et al. 2023. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* 51:D690–D699. <https://doi.org/10.1093/nar/gkac920>
55. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830. <https://doi.org/10.48550/arXiv.1201.0490>