Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types

Cummins, Elizabeth A.; Hall, Rebecca J.; Connor, Chris; Mcinerney, James O.; Mcnally, Alan

DOI:
10.1099/mgen.0.000903

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Cummins, EA, Hall, RJ, Connor, C, Mcinerney, JO & Mcnally, A 2022, 'Distinct evolutionary trajectories in the *Escherichia coli* pangenome occur within sequence types', *Microbial Genomics*, vol. 8, no. 11, 000903. https://doi.org/10.1099/mgen.0.000903

MICROBIAL GENOMICS

RESEARCH ARTICLE

Cummins *et al.*, *Microbial Genomics* 2022;8:000903

DOI 10.1099/mgen.0.000903

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Distinct evolutionary trajectories in the *Escherichia coli* pangenome occur within sequence types

Elizabeth A. Cummins[1], Rebecca J. Hall[1], Chris Connor[1,2], James O. McInerney[3] and Alan McNally[1,*]

Abstract

The *Escherichia coli* species contains a diverse set of sequence types and there remain important questions regarding differences in genetic content within this population that need to be addressed. Pangenomes are useful vehicles for studying gene content within sequence types. Here, we analyse 21 *E. coli* sequence type pangenomes using comparative pangenomics to identify variance in both pangenome structure and content. We present functional breakdowns of sequence type core genomes and identify sequence types that are enriched in metabolism, transcription and cell membrane biogenesis genes. We also uncover metabolism genes that have variable core classification, depending on which allele is present. Our comparative pangenomics approach allows for detailed exploration of sequence type pangenomes within the context of the species. We show that ongoing gene gain and loss in the *E. coli* pangenome is sequence type-specific, which may be a consequence of distinct sequence type-specific evolutionary drivers.

DATA SUMMARY

Supporting data and code have been provided within the article or through Supplementary Data files available at 10.6084/m9.figshare.21360108 [1]. Custom Python scripts used to perform analyses are available at github.com/lillycummins/InterPangenome unless otherwise stated in the text. The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary data files.

INTRODUCTION

Escherichia coli is a genotypically and phenotypically diverse species that inhabits a multitude of varying environments and is one of the best-studied bacteria. The species is divided into eight main phylogroups: A, B1, B2, C, D, E, F and G. Phylogroup assignment is a useful process, as the classification can be used to gain ecological and epidemiological insights, such as host specificity and lifestyle [2]. For example, *E. coli* in the microbiota of humans is dominated by phylogroups A and B2, whilst B1 is the most prevalent *E. coli* phylogroup in domestic and wild animal microbiotas [3]. Pathotypes also partially coincide with phylogroup. The majority of extraintestinal pathogenic *E. coli* (ExPEC) lie within phylogroups B2, D and F, whilst those associated with enteric diseases are more generally found within phylogroups A, B1 and D [4, 5]. Phylogroup E has famously been associated with enterohaemorrhagic *E. coli* (EHEC), due to the ownership of the O157:H7 pandemic lineage, but has recently been shown to be a highly diverse phylogroup spanning commensal, environmental and pathogenic lifestyles [6]. The most recently defined phylogroup, G, broadly comprises poultry-associated isolates [7].

Within phylogroups, further subdivision into clonal complexes and sequence types (STs) can be achieved by multilocus sequence typing (MLST) [8, 9]. The Warwick/Achtmann MLST scheme of *E. coli* is based on variations of seven housekeeping genes and has resulted in the generation of a vast multitude of STs. Prominent STs include ST131, a multidrug-resistant pandemic

Received 20 May 2022; Accepted 02 October 2022; Published 23 November 2022

Author affiliations: [1]Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, B15 2TT, UK; [2]Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne 3000, Australia; [3]School of Life Sciences, University of Nottingham, Nottingham, NG7 2UH, UK.

***Correspondence:** Alan McNally, a.mcnally.1@bham.ac.uk

Keywords: comparative pangenomics; *Escherichia coli*; pangenomes; sequence types.

Abbreviations: COG, cluster of orthologous genes; EHEC, enterohemorrhagic E. coli; ExPEC, extraintestinal pathogenic E. coli; GOIs, genes of interest; MLST, multilocus sequence typing; PTS, phosphotransferase system; ST, Sequence type.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary figures are available with the online version of this article.

Impact Statement

The study of pangenomes has gone from strength to strength, coinciding with the exponential increase in available genomic datasets. Traditionally, pangenomes are usually considered in isolation, with the classification of genes as 'core' or 'accessory' being determined on a species level. This approach masks many of the interesting evolutionary processes of gene gain and loss that are occurring within a species pangenome. We introduce comparative pangenomics as a new method for understanding pangenome dynamics within a species by comparing gene classifications between closely related lineages. We use *Escherichia coli* sequence type pangenomes to uncover underlying evolutionary trajectories within the species that would otherwise be masked by traditional solitary pangenome analyses.

ExPEC [10–12]; ST10, a generalist lineage containing commensals and pathogens from a variety of hosts [13, 14]; and ST11, the aforementioned pandemic EHEC O157:H7 [15].

Coinciding with rich phenotypic heterogeneity, there is no gene pool barrier within *E. coli* [16], meaning that genetic material can be freely exchanged between pathogens and commensals. Therefore, some *E. coli* STs could conceivably act as vital genetic repositories in the development of important characteristics such as pathogenicity or antimicrobial resistance within other STs. The extent to which any given ST, or group of STs, acts as a reservoir (genetic source) or recipient (genetic sink) in the exchange of genetic information is currently not well known. The relatively recent development of pangenomics [17] provides a useful perspective we can use to interrogate the genetic contents of available genetic repositories provided by different *E. coli* STs and gain further insight into how these gene collections are structured and evolve.

The pangenome represents the set of all genes present in a given population [17, 18]. Pangenomic studies have been performed to understand *E. coli* at the species level [16, 19–22] but comparative pangenomics analyses between STs within this species can potentially add to our understanding of the evolution of the species. The influence of population structure [21] and the presence of complex epistatic relationships [23] are increasingly being acknowledged to have a major effect on the evolution of prokaryote pangenomes. Whelan and colleagues, for instance, noted that asymmetrical gene dependencies (e.g. the presence of *geneX* first requiring the presence of *geneY*, but not vice versa) cannot be uncovered by the consideration of coincident gene patterns alone. Conditional gene relationships can exist between genes, between sequence variants, or between genes and variants [24]. Inter-pangenome analysis – comparative analysis of closely related pangenomes – provides an excellent mechanism for generating prioritized lists of putative dependences between genes. Inter-pangenome analysis can show whether a gene that is classified as core in one ST is also core in a different ST. Inter-pangenome analysis can also assess differences in functional composition between closely related pangenomes. The functional contents of a pangenome (whether species-level or ST-level) reflects the biological processes occurring within the given population, such as niche adaption [25, 26], or the evolution of important phenotypes, such as drug resistance [27].

An in-depth study of an ST131 pangenome revealed clade-specific diversity in colonization and metabolism genes in the accessory genome of the globally dominant multidrug-resistant sub-lineage of ST131, clade C (H30Rx) [27]. The reported diversity was not due to the presence of unique genes, but rather the presence of unique alleles. Allelic diversity as a signature of selection has now also been observed in ST167 [28] and ST410 [29]. Allelic variation in metabolic genes has been described as an early warning sign of multidrug resistance, with metabolic flexibility potentially being a key trait in multidrug-resistant clones [30]. ST131 is one of the few *E. coli* STs to have undergone detailed pangenome analysis [27, 31, 32]. Understanding of *E. coli* STs on a comparative pangenome level is currently limited in terms of comparative analyses, with little known about how ST-level pangenome evolution is occurring. We wish to test the hypothesis that the different *E. coli* ST-level pangenomes do not evolve in the same way, by the gain and loss of the same kind of genes, but that their evolutionary histories and trajectories differ in significant and meaningful ways.

Here, using one of the biggest collections of *E. coli* genomes to date, we further develop our understanding of *E. coli* pangenome dynamics and evolution by splitting *E. coli* into its constituent STs and comparing and contrasting the fates of these STs in the context of their respective pangenomes. We introduce an ST-focused approach to investigating evolutionary trends of pangenomic structure and contents, including the presence of sequence variants of metabolism genes, within *E. coli*. We addressed the following objectives: (i) to establish a conservative *E. coli* core genome, (ii) to assess whether ST pangenomes vary in structure, (iii) to assess whether some ST pangenomes are enriched for specific biological processes, (iv) to assess the level of metabolic variation across ST pangenomes (given the potential link to multidrug resistance) and (v) to evaluate the potential for STs to act as genetic sources or sinks. We find that the distribution of genes across clusters of orthologous genes (COG) functional categories within an ST core genome is not dictated by being in a given phylogroup and that enrichment occurs in specific functional categories that vary by ST. We also uncover conditional genetic relationships within core genomes and find that sequence variants differ in core classification within and between STs. Inter-pangenome analysis allows us to highlight how pangenome evolution is

Table 1. Summary of pangenome analyses (pangenome size does not include paralogues)

Phylogroup	Sequence type	No. of genomes	DI	No. of core genes	Pangenome Size	Core/Pan. (%)
A	ST10	2370	3.88	3066	27634	11.10
	ST167	115	2.74	3675	9035	40.68
	ST410	1006	3.49	3272	16223	20.17
B1	ST17	1884	2.14	4003	11870	33.72
	ST21	2411	1.98	4058	10671	38.03
	ST3	40	1.46	3749	7933	47.26
B2	ST12	283	2.59	3816	10531	36.24
	ST127	232	2.64	3891	9696	40.13
	ST131	3186	2.34	3460	15665	22.09
	ST14	62	1.60	3881	6825	56.86
	ST141	91	1.91	3879	8217	47.21
	ST144	65	2.15	3908	6938	56.33
	ST28	46	2.50	3630	7346	49.41
	ST372	54	2.51	3752	7514	49.93
	ST73	873	2.22	3789	11865	31.93
	ST95	758	2.89	3815	11933	31.97
D	ST38	617	3.02	3780	14443	26.17
	ST69	696	2.56	3771	14808	25.47
E	ST11	5137	3.84	3944	12904	30.56
F	ST117	269	4.02	3722	11055	33.67
	ST648	382	2.04	3674	11610	31.65

DI, Shannon's diversity index; ST, sequence type.

heterogeneous across a species and is independent of phylogeny, and we further our understanding of how collections of genes vary and evolve between STs.

METHODS

Genome collection and ST pangenome analysis

We downloaded 20577 publicly available *E. coli* assemblies from EnteroBase [33] with a custom Python script (github.com/C-Connor/EnterobaseGenomeAssemblyDownload). EnteroBase employs quality filters when adding draft assemblies to the database: ≤800 contigs, >70% contigs assigned species using Kraken, genome length 3.7–6.4 Mbp and a minimum N50 value of 20 kb [33]. Accession numbers and other identifiers within EnteroBase for these assemblies are provided in File S1 (available in the online version of this article) and as text files within pangenome_data.zip available within the Supplementary Data(https://doi.org/10.6084/m9.figshare.21360108). Genome similarity was estimated using Mash [34] with a sketch size of 1000 and a k-mer size of 21 to ensure that no duplicate entries were included in the dataset.

ST was confirmed with mlst (v2.15) (https://github.com.tseeman/mlst) using the PubMLST database [35] and the phylogroup of the ST was obtained from the published literature. The assemblies covered 6 phylogroups and 21 different STs of *E. coli*: ST3, ST10, ST11, ST12, ST14, ST17, ST21, ST28, ST38, ST69, ST73, ST95, ST117, ST127, ST131, ST141, ST144, ST167, ST372, ST410 and ST648 (Table 1). Phylogroup G is an intermediate group between B2 and F that was characterized in 2019 by Clermont *et al.* [7]. This phylogroup is not included in the current analysis because it was unknown at the time of data collection. Sample source information was collected when available, but the majority of isolates in the dataset had no source of isolate sampling data available on EnteroBase. The source sampling diversity of each ST was determined by Shannon diversity index.

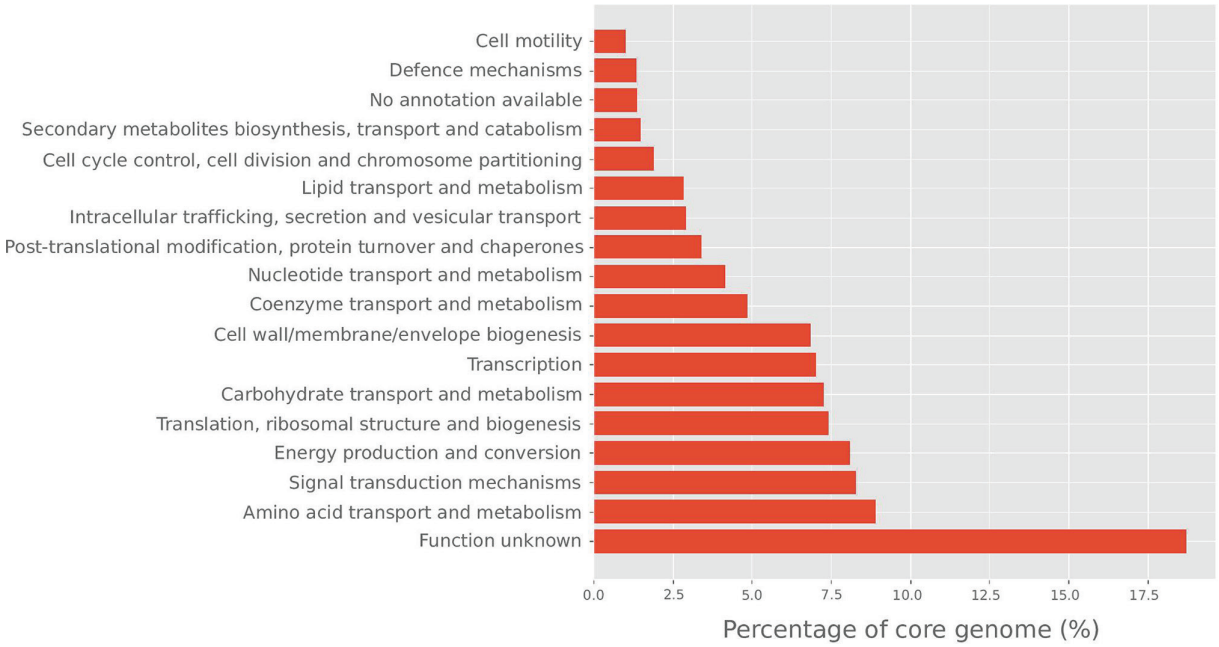

Fig. 1. Functional breakdown of 2172 core *E. coli* genes. Functional classes are based on COG categories.

Genes within each genome were annotated using Prokka (v1.12) [36]. Genomes were grouped by ST (using the EnteroBase classification) for individual ST-level pangenome analyses using Panaroo (v1.1.2) [37] with a 0.95 sequence identity threshold and a 0.99 core genome sample threshold to allow the inclusion of unique core gene alleles in the accessory genome. We use the terms 'gene cluster' or 'gene' to refer to an orthologous gene group constructed by Panaroo. Linear regression was performed using Python scikit-learn (sklearn) LinearRegression module.

Assignment of COG functional categories

The linear reference genome provided by Panaroo [37] for each ST pangenome was split into two lists of its respective core and accessory gene clusters. The nucleotide sequence for each gene cluster was translated using a custom Python script (github.com/C-Connor/GeneralTools) to obtain a protein sequence for each cluster. These protein sequences were used to characterize gene function. Gene clusters were assigned COG functional categories [38] using eggNOG-mapper (v2.0.8) bestOG assignment [39] and the eggNOG database [40] with sequence searches performed by DIAMOND (v2.0.7) [41]. Gene clusters that did not return a match within the eggNOG database were categorized under '?'. Heatmaps and clustermaps displaying distribution of COG categories across STs were made with seaborn (v0.11.2).

An ST was labelled as enriched in COG category 'X' if the percentage of ST core genome designated to category X lay above the upper quartile plus 1.5 times the inter-quartile range for all ST core genomes in that category.

Functional domain annotation was performed with InterProScan (v5) [42, 43].

Distribution of ST core genomes

Custom ABRicate databases were made for the core genome of each ST using the representative gene cluster nucleotide sequences from Panaroo and the --setupdb option in ABRicate (v0.8.7) (github.com/tseemann/abricate). The bottom fifth percentile of the average coverage distribution for each set of ST core genes was removed to ensure that any incorrectly called core genes were not included in our analyses. Mass screening across all 20 577 assemblies was carried out for each ST core database with ABRicate (21 searches in total). The results were summarized and partial hits (instances where a gene hit was split over multiple contigs) were accounted for and processed with a custom Python script. The average proportion of gene covered for each core gene cluster per ST was calculated.

Core metabolic reconstructions

Metabolic models were constructed for the core genome of each ST using CarveMe [44]. Representative core gene cluster nucleotide sequences for each ST were used as input and the CarveMe algorithm was executed using the default settings. The number of metabolic reactions and metabolites in each ST core metabolic profile were counted using the Python COBRA package [45].

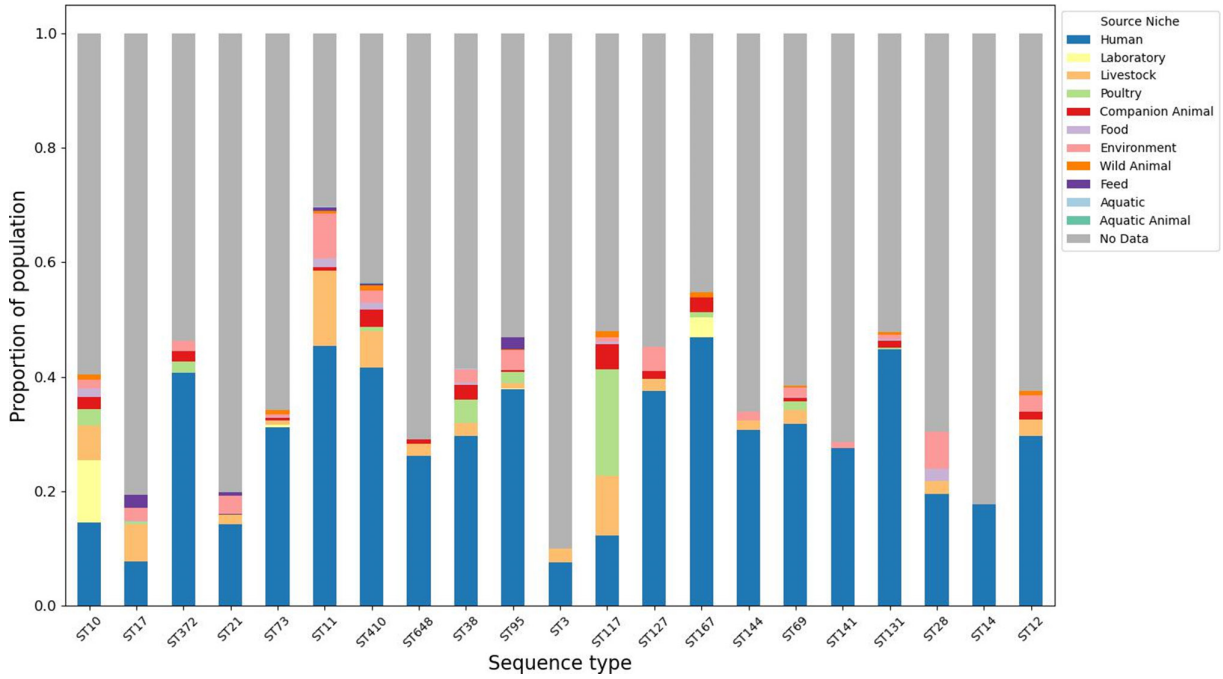

Fig. 2. Breakdown of sampled source information available from EnteroBase for 21 *E. coli* sequence types. Graph coloured by sample source location.

Unique core metabolic reactions and genes

Metabolic reactions uniquely present in a single ST core metabolic reconstruction were extracted using a custom Python script. Unique reaction names were searched manually on the BiGG database website [46] to find the related gene names for each reaction. These related genes of interest (GOIs) were searched for in the descriptors of the ST core gene sequences for the ST the related reaction was uniquely present in. These sequences were combined to construct a custom ABRicate database to perform a mass screening for the GOIs across all 20577 assemblies with ABRicate. ABRicate results were summarized and processed using the same method previously described for the distribution of core genomes. A clustermap displaying the varying presence of the GOIs across 21 STs was made with seaborn (v0.11.2).

RESULTS

A 2172 gene cluster *E. coli* core genome

The size and content of the *E. coli* core genome have been estimated in previous studies [21, 47–49], but not explicitly using a collection of genomes as large as the dataset considered in this work. Here, we provide a representative *E. coli* core gene list. There were n=2, 172 gene clusters identified that had a mean percentage coverage above 98% across all 20577 assemblies. A list of these core genes and their nucleotide sequences is provided in File S2. Grouping these core genes by COG category showed that genes of unknown function (category S) were the largest functional group (18.7%). A breakdown of the functional composition of the core genes can be seen in Fig. 1. This large percentage of species-level core genes with unknown functions highlights that, despite extensive study and characterization, there is still a great deal of information to be uncovered regarding the core genes of *E. coli*.

Pangenome structure varies between *E. coli* sequence types

To assess the level of variation between ST-level pangenomes we first considered variation in the context of structure. We assembled the pangenomes of 21 STs using Panaroo (v1.1.2). The pangenome sizes ranged from 6825 to 27 634 gene clusters (Table 1), with an average size of 11653 gene clusters, and core genome sizes ranged from 3066 to 4058 clusters, with an average of 3738 clusters per ST pangenome. Neither core gene number (r^2=0.005, ordinary least squares) nor total gene number (r^2=0.249, ordinary least squares) were a function of ST sample size. Consequently, ST pangenomes exhibited variation in the core gene number as a percentage of the total pangenome size, which suggests that there is no uniform pangenome structure within *E. coli*. Variation in this percentage was highest for those STs with the fewest genomes, but even when the STs with >100 genomes were considered, the variation in core gene number as a percentage of the total pangenome size extended from 11.1% (ST10) to 56.86% (ST14).

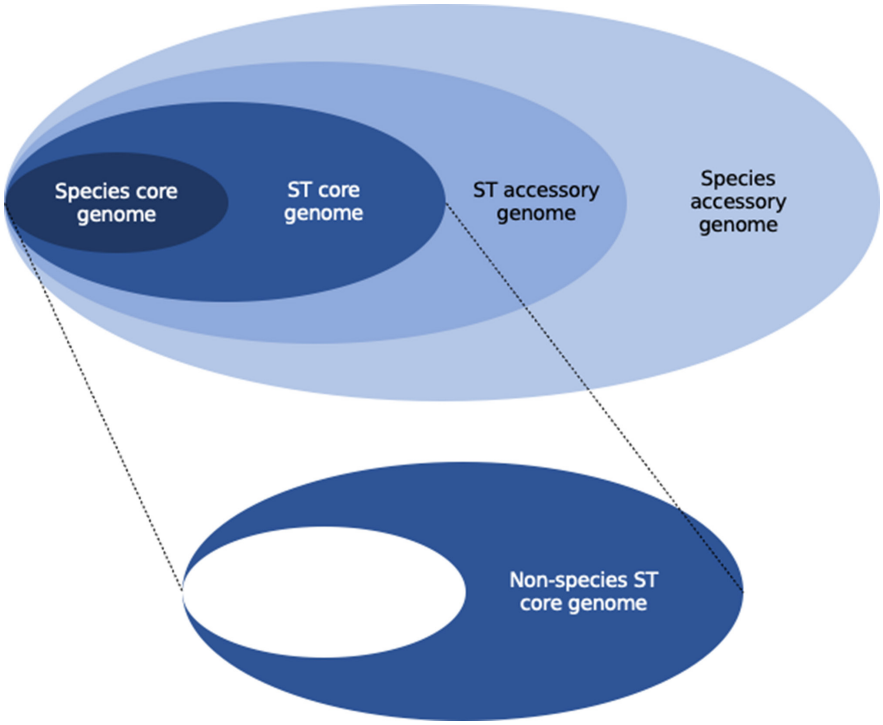

Fig. 3. Depiction of pangenome segments used in this analysis.

The range of sampled sources per ST varied from one known sample source (ST14) to nine known sample sources (ST10, ST11, ST410). Isolates with no source data made up the majority of samples in all STs, with the exceptions of ST11, ST167 and ST410 (Fig. 2). The Shannon diversity index was calculated as a measure of sample source diversity for each ST (Table 1). Neither pangenome size (r^2=0.351, ordinary least squares) nor ST core genome size (r^2=0.214, ordinary least squares) were a function of sample source diversity, indicating that sampling bias should not strongly affect our pangenome analyses.

ST-specific core functions vary between sequence types

For the purposes of this study, we define three pangenome segments that are analysed and discussed throughout this work. Firstly, the 'species core genome' is the set of genes common to all genomes in this study. Next, the 'ST*X*-specific core genome' is the set of all genes considered core to ST*X*, with the species core genome removed. Finally, the 'unique ST*X*-specific core genome' is the set of genes that are found to be core only in ST*X* and no other ST. These pangenome segments are visualized conceptually in Fig. 3. We calculated the percentage of each ST-specific core genome that was assigned to each of the COG functional categories. As we were interested primarily in functional differences between STs, category 'S' (function unknown) and '?' (no functional annotation available) were masked from visualization in Fig. 4, as they were always the largest two categories.

Hierarchical clustering of the percentage of each ST core genome assigned to 20 COG functional categories highlighted the ST131- and ST10-specific core genomes as having the most distinct functional profiles (Fig. 4). The accessory genomes were also functionally categorized, however in all ST pangenomes the accessory genome was dominated by genes of unknown function (data not shown). The data show that ST pangenomes do not possess uniform core functional profiles and additionally, this observed variation is not heavily influenced by the identity of the phylogroup.

We also examined variation in ST-specific core genomes, and in particular their propensity to be differentially enriched in specific biological processes of a particular function. ST95, ST410, ST648, ST131 and ST10 exhibited functional enrichment in four COG categories in their ST-specific core genomes (see Methods). ST95 was enriched in genes linked to cell membrane biogenesis (category M); ST10 was enriched in genes pertaining to transcription (category K) and carbohydrate metabolism and transport (category G); ST410 and ST648 were enriched in energy production and conversion genes (category C); and ST131 was enriched in genes pertaining to cell membrane biogenesis (category M) and carbohydrate metabolism and transport (category G). These enriched categories are highlighted in Fig. 4. This suggests that genes encoding these functions may be particularly influential in these STs.

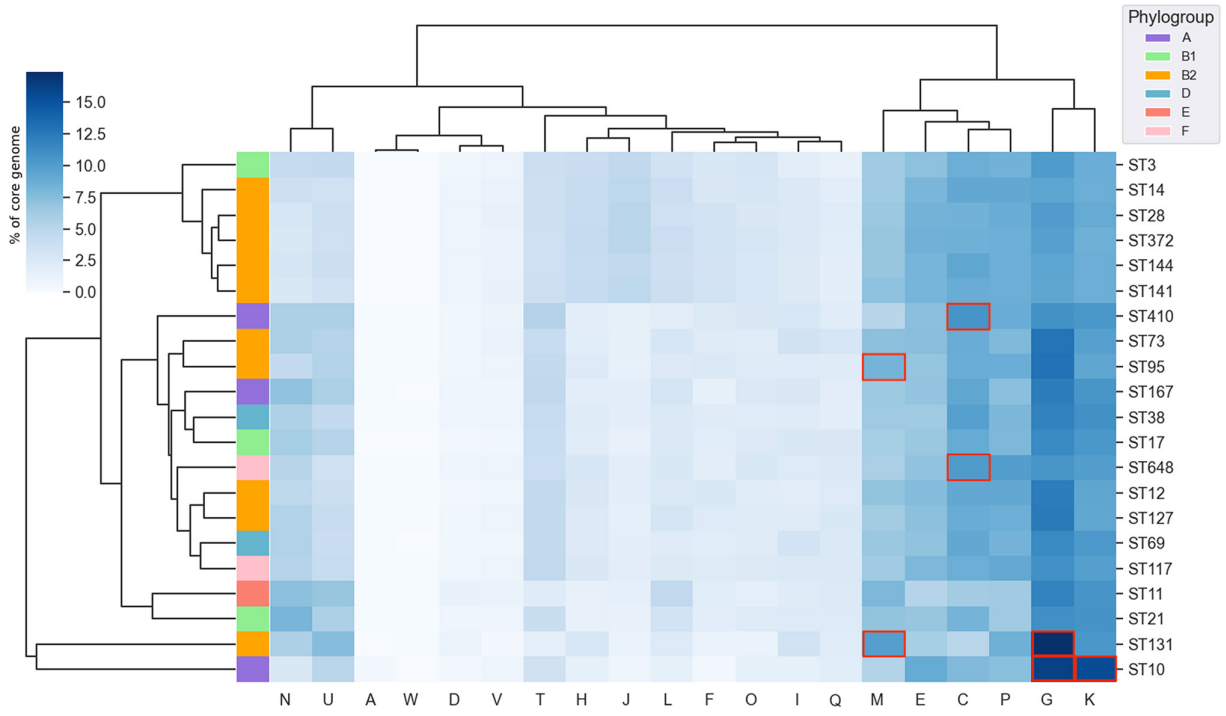

Fig. 4. Hierarchically clustered (by percentage presence in core genome row-wise and column-wise) heatmap showing percentage of non-species ST core genomes classified into 20 functional COG categories. COG categories that are enriched in a sequence type are outlined in red. Functional COG categories: 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, ribosomal structure and biogenesis; K, transcription; L, replication and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, post-translational modification, protein turnover and chaperone functions; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; T, signal transduction; U, intracellular trafficking and secretion; V, defence mechanisms; W, extracellular structures.

ST131 and ST10 pangenomes possess multiple alleles related to carbohydrate metabolism and transport

We have identified two ST-specific core genomes, ST10 and ST131, that are enriched in carbohydrate metabolism and transport (category G) genes. To investigate whether this enrichment was related to metabolic diversity, we explored the presence of alleles within the category G genes in the ST131 and ST10 pangenomes. The ST131-specific core genome includes n=100 gene clusters linked to carbohydrate metabolism and transport (category G), of which 64% are indicated to have multiple alleles present, as different gene clusters, in the ST131 pangenome (File F3). These include, but are not limited to, *manRXZ*, *sorABFM*, *malPX* and *gatABCYZ*, involved in the mannose, sorbose, maltose and galactose phosphotranserfase systems [50–53]. The ST10-specific core genes in category G (File S3) that have multiple alleles present across the ST10 pangenome include *mngAB*, involved in mannose transport and metabolism [54] and sugar efflux transporters *setAC* [55]. A full summary of the number of genes present as multiple alleles per enriched COG category is provided in File S4. Beyond multiple alleles of carbohydrate metabolism genes being present across the pangenome, certain genes were present as multiple alleles within the ST131 core genome. The non-species ST131 core genome possessed two alleles of each of the following genes: *fruA*, *gatC*, *kdgK*, *nagB*, *tabA* and *uxaA*. Multiple alleles of carbohydrate metabolism genes were not present within the ST10-specific core genome, but there were multiple alleles present of three transcription genes; *glpB*, *hcaR* and *mngR*.

Further investigation revealed that the fructose phosphotransferase system (PTS) gene, *fruA*, beyond being present as two alleles in the ST131-specific core genome (as clusters 'fruA_2' and 'fruA_3_fruA_1'), was in fact present as four gene clusters across the ST131 core genome; 'fruA_1' and 'manP_fruA_4_fruA_1' clusters were found to be present in the species core genome. To investigate the functionality of these four gene clusters, functional domain analysis was carried out using InterProScan [42]. Three distinct functional domains corresponding to the PTS system EIIA, EIIB and EIIC components [56] were identified within the four *fruA* clusters. 'fruA_1' possessed both EIIB and C, whilst 'manP_fruA_4_fruA_1' and 'fruA_3_fruA_1' possessed only EIIC (Fig. 5) . 'fruA_2' encodes the PTS EIIA component that is associated with *fruB* rather than *fruA* [57]. This discrepancy is likely attributable to the Panaroo cluster naming algorithm rather than being of biological significance. The presence of these four gene clusters is shown in Fig. 5. The three correctly named *fruA* clusters are distinct (>5% sequence divergence), highly conserved

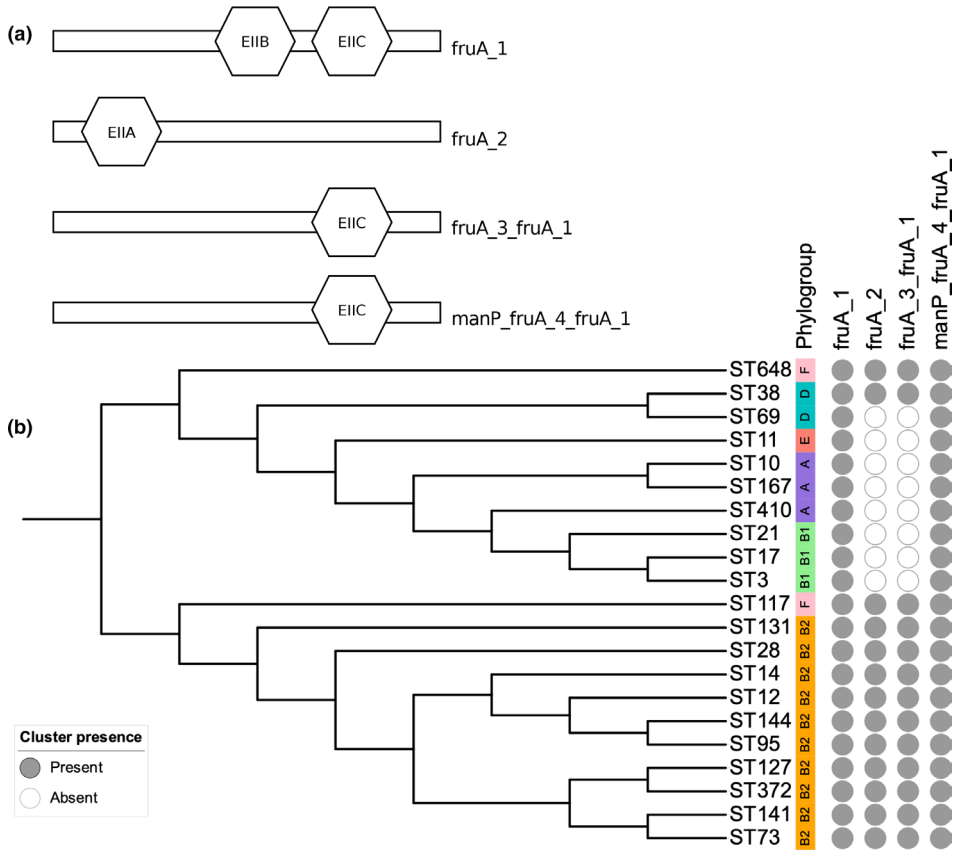

Fig. 5. (a) Phosphotransferase system components encoded by four *fruA* annotated clusters and (b) cluster presence across an *E. coli* phylogeny. Created using iTOL [75].

and seemingly functional, and encode non-truncated peptide sequences, which suggests that annotation error, degradation, or pseudogenization are unlikely to be responsible for this multiplicity.

The ST131 pangenome has distinct gene presence patterns at phylogroup and species level

Further investigation of the presence patterns of genes from ST pangenomes enriched in specific COG functional categories (outlined in Fig. 4) across other ST pangenomes revealed that the ST131 pangenome displays gene presence and absence patterns that are distinct from those of other phylogroup B2 ST pangenomes. The 'mhpA' gene cluster from the ST410 pangenome, or 'mhpA_1' gene cluster from the ST648 pangenome, is only present in the ST131 pangenome out of the 10 B2 phylogroup ST pangenomes (Figs S1 and S2). *mhpA* encodes a 3-(3-hydroxyphenyl)propionate hydroxylase involved in phenylalanine metabolism [58]. Similarly, the 'mngA_1', 'mngB' (Fig. S3) and 'mngR_1' (Fig. S4) clusters from the ST10 pangenome are only present in the ST131 pangenome out of all B2 phylogroup ST pangenomes. There are also gene clusters ('rspA_1', 'hbp', 'tsx_1', 'fimC_1') from the ST95 pangenome that are only absent in the ST131 pangenome out of the B2 phylogroup ST pangenomes (Fig. S5).

The ST131 pangenome possesses core genes that are not seen in any other *E. coli* ST pangenome considered in this study. Notable presence patterns within ST131's enriched carbohydrate transport and metabolism core genes (Fig. S6) include the uniquely present 'group_3501' and 'yihP_yicJ_3_yicJ_1' gene clusters. These clusters were not detected in any other ST pangenome. Functional annotation of 'group_3501' suggests that this gene encodes a glycosyl hydrolase and eggNOG provided *xylS* as an annotation for this gene cluster. In the KEGG orthology database [59, 60], *xylS* is synonymous with *yicI* [61]. With this connection, we postulate that these two gene clusters, uniquely present in the ST131 pangenome, are involved in the same xyloside metabolic pathway. The nucleotide sequence was searched against the uniprot [62] database, resulting in a top hit of 85.3% similarity to a putative glycosyl hydrolase from *Citrobacter rodentium* and *Citrobacter freundii*.

Alleles of metabolism genes vary in core status across STs

To examine metabolic diversity within ST core genomes in more detail, we performed metabolic reconstructions for the core genome of each ST pangenome using CarveMe [44] to create a 'ST core metabolic profile' so that gene information could be extrapolated to

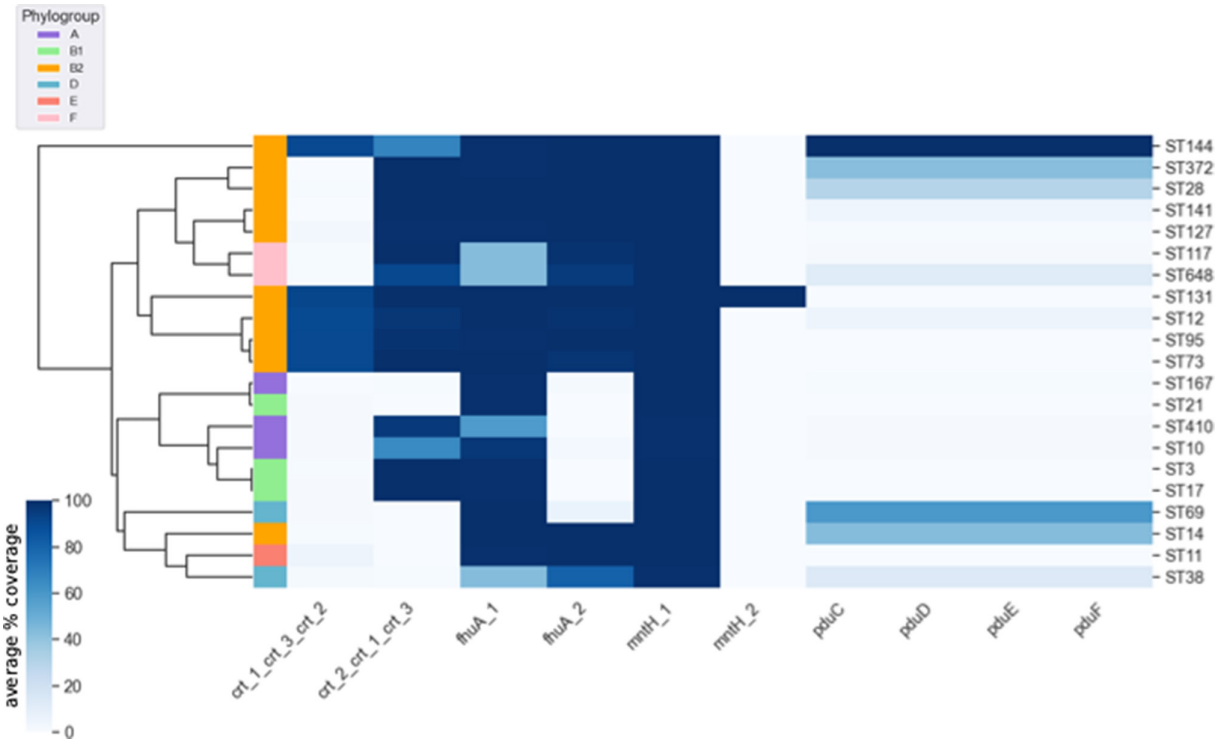

Fig. 6. Hierarchical row-wise clustering of the average presence of *crt*, *fhuA*, *mntH* and *pduCDEF* gene clusters across 21 sequence types of *E. coli*.

utilization and specific metabolic pathways. Comparison of the 21 ST core metabolic profiles uncovered 825 metabolic reactions that were found in at least one ST core metabolic profile, but not common to all STs (i.e. not a species-core reaction). We focused further analysis on metabolic reactions that were uniquely present in a single ST core metabolic profile. Tracing uniquely present reactions within the BiGG database [46] back to their related gene names, and then searching for these names within our dataset led us to a subset of gene clusters with non-ubiquitous presence patterns (Fig. 5) . These selected clusters were *fhuA* (iron acquisition) [63], *pduCDEF* (propanediol utilization) [64], *mntH* (manganese transport) [65] and the hydratase *crt* [66].

The manganese transporter *mntH* [65] has two alleles present in the core genome of ST131 (Fig. 6), which raises the question of why there is a fixed second allele and also why this has not happened in another ST pangenome. Functional domain analysis of 'mntH_1' and 'mntH_2' returned the same InterPro annotation accession number for both alleles. Similarly, *fhuA* and *crt* have two alleles simultaneously present in the core genomes of multiple STs Fig. 6. The *crt* alleles are also involved in a conditional relationship. From the clustermap in Fig. 6, we see that the 'crt_1_crt_3_crt_2' cluster is only present in an ST pangenome when 'crt_2_crt_1_crt_3' is also present, with the possible exception of ST144. Additionally, the 'fhuA_2' cluster is only present, excluding phylogroup F STs, when 'fhuA_1' is also present in the ST, with the possible exception of ST38.

Fixation of the propanediol utilization operon *pduCDEF* has occurred uniquely in ST144 (Fig. 6) . These genes are reported in other STs (ST372, ST28, ST141, ST648, ST12, ST69, ST14, ST38) at lower average frequencies, showing that *pduCDEF* are accessory genes intermittently present within these STs. The *pdu* operon is involved in anaerobic respiration, which is used by enteropathogenic *Enterobacteriaceae* to out-compete existing intestinal microbiota during infection and is frequently reported in *Yersinia enterocolitica* and *Salmonella* Typhimurium [67]. However, this was considered to be a rare phenotype in *E. coli* [19]. Each of the four genes presented here provide additional evidence for sequence-level variation in metabolism between *E. coli* STs.

ST10 has the potential to be a genetic source for other *E. coli* sequence types

Evidence for a phylogroup or ST acting as a genetic source for other *E. coli* may arise in the form of an ST pangenome possessing a low (or no) amount of ST-specific core genes (genes that are classified as core in only one specific ST). To this end, gene clusters that were uniquely core to a specific phylogroup were examined first. The three ST pangenomes in phylogroup A had no unique core genes (Table 2). Alleles including those of flagellar genes *fliDS* are present amongst the 52 B1 unique core genes. B2 ST pangenomes possessed 99 uniquely core gene clusters, including alleles of genes involved in central metabolism, *sucABCD*, fructose metabolism, *fruA* and the decarboxylase *tabA*. The 14 unique core genes of the 2 phylogroup D ST pangenomes included alleles of a putative fimbrial protein, *yadNMV*. Extending this analysis from phylogroup to ST, we also considered unique ST-specific

Table 2. Numbers of genes uniquely core to a single phylogroup and sequence type

Phylogroup	Unique phylogroup core genes	ST	Unique ST core genes
A	0	ST10	0
		ST167	28
		ST410	6
B1	52	ST3	68
		ST17	14
		ST21	74
B2	99	ST131	21
		ST28	44
		ST14	83
		ST12	10
		ST95	8
		ST144	56
		ST141	38
		ST73	26
		ST127	14
		ST372	23
D	14	ST38	23
		ST69	43
E	99	ST11	99
F	1	ST117	24
		ST648	40

ST, sequence type.

core genes. The number of alleles uniquely core to a single ST varied within and between phylogroups, with ST14 (phylogroup B2, *n*=83) and ST21 (phylogroup B1, *n*=74) encoding the largest number of unique core genes (Table 2). Whilst phylogroup A has no unique core genes, within this phylogroup only the ST10 pangenome had no reported unique ST-specific core genes; the ST167 and ST410 pangenomes were found to have 28 and 6 unique core genes, respectively. The ubiquity of the ST10 core genome across all other STs may be an indicator that this ST is likely to be capable of acting as a genetic source within *E. coli*.

DISCUSSION

Extensive phenotypic variation and the existence of diverse STs within *E. coli* are well documented [19, 68]. However, little is known about how the genetic repertoire of each ST varies in terms of pangenome structure and content, and consequently which genes are given core status within different ST pangenomes. We build upon previous work analysing a single *E. coli* species pangenome [19, 20] or *E. coli* ST pangenome [27, 31, 32] by performing large-scale comparative analysis on 21 ST pangenomes constructed from over 20 000 genomes. We introduce the concept of comparative pangenomics with a method that interrogates ST pangenome content and structure variation across the species. We also classified the non-species ST core genome of each ST pangenome into COG functional groups. Our study revealed variation in pangenome structure and core genome functionality both across and within *E. coli* phylogroups.

Previous estimates of the size of the *E. coli* core genome fell in the range of 1000 to 3000 gene clusters and were extrapolated from small genome collections ranging from 14 to 186 isolates [47–49]. We build upon this earlier work by estimating an *E. coli* core genome with a larger dataset. The bias within EnteroBase towards human pathogens, due to clinical relevance, may impact on our defined core genome as there could conceivably be a false over-representation of genes relating to, say, virulence and antibiotic resistance as a result of our sampling. Many of the STs sampled in this work are ExPEC lineages, but there are also

representative lineages from other pathotypes and commensals, and the lineages represent all possible phylogroups and a range of antimicrobial resistance prevalence, and therefore we believe are a broad representative sampling of *E. coli* from many common genetic backgrounds. Nonetheless, the non-uniformity in structure between *E. coli* ST pangenomes demonstrates the extent of the flexibility within this species and is a valuable lesson gained from comparative pangenomics.

We also found that the function of genes given core status within an ST pangenome (the ST-specific core genome) varied between STs, with certain ST pangenomes having higher percentages of genes in four functional COG categories: energy production, carbohydrate metabolism, transcription and cell membrane biogenesis. This enrichment may signpost ST-specific adaptive evolutionary processes as a signature of selection via accumulation of allelic diversity. It is already known that there are large ecological variations in *E. coli*; isolates have been found as gut commensals in most animals, as well as in environmental samples, and can exist on a spectrum of pathogenicity, ranging from complete commensal to strict pathogen.

Going beyond consideration of ST core genomes as functional units, we attained more nuanced findings regarding ST core gene variants. We found possession of multiple variants of carbohydrate metabolism genes in ST pangenomes. More diverse genes relating to metabolism, including clone-specific SNPs in anaerobic metabolism loci within ST410 [29] and ST131 [27], have been reported previously. In these cases, the sequence diversity was attributed to differential evolution whereby selection for a process (enhanced anaerobic metabolism capabilities) rather than selection for a gene was occurring. Metabolic flexibility has also been proposed as a precursory stage to multidrug resistance [30]. We could conceivably extrapolate our findings, such as our observed diversity in a fructose metabolism gene within an ST core genome, as a potential signature of an evolutionary selection pressure.

The fixation of the *pdu* operon in the ST144 pangenome suggests a unique evolutionary history. ST144 is a uropathogenic *E. coli* that shares the closest common ancestor with ST95 [69]. 1,2-propanediol is enriched in the mucosal lining of the intestine, so the ability to utilize this alternative carbon source is advantageous in an inflamed gut [70]. Similarly, the ST131 pangenome has a second *mntH* allele and a glyosyl hydrolase linked to xyloside metabolism, among other distinct gene presence patterns which suggests a separate evolutionary trajectory for this ST. Recent mash-based analysis by Abram and colleagues has demonstrated notable differences between ST131 and other B2 strains that were significant enough to classify ST131 within the subgroup B2-1, said to have recently emerged from B2-2 [71]. The ability to discriminate between ST131 and the rest of the B2 phylogroup was attributed to the differential, rapid uptake of unique virulence factors and mobile genetic elements by ST131 [10]. The unique gene presence patterns we reported within the ST131 pangenome are consistent with this previous study [71].

Pangenomes can reflect the ecology of an organism [72, 73], so insight may be gained by translating gene presence/absence to, for example, niche occupation. Genes core to an ST, or a group of STs, provide indicators of evolutionary advantages in certain ecological settings and genetic backgrounds [74]. From our dataset, ST10 was the only ST pangenome to have no unique core genes. This weak unique core signature may reflect the heterogeneous nature of ST10 [71]. This aligns with previous work that has identified ST10 as a generalist lineage and a potential genetic reservoir for other *E. coli* lineages [19, 21]. However, it may be possible that other STs in our dataset are less well sampled than ST10 and are therefore less representative of their ecological realities. An underlying caveat of all pangenome analysis is sampling bias. There is almost certainly incomplete coverage of all possible source diversity in our data set, which is impossible to fully capture. This is primarily caused by oversampling of particular niches, such as human clinical samples within *E. coli* genomic datasets. Inadequacy in sampling will also affect the classification of a gene as core. For instance, it may be poor sampling that makes a gene appear as core, when denser sampling would have moved it to the accessory category. Ecological under-representation within sampling must always be considered when interpreting any results from pangenomic analysis.

Our goal was to test whether *E. coli* ST pangenomes are evolving in a uniform way. Our data show that variation in core functions between ST pangenomes is a clear signal of ST-specificity, and we show that ST pangenomes are distinct in different ways, from structure to alleles of genes varying in core status across ST pangenomes. We have also provided a putative list of core gene clusters from a dataset of over 20 000 *E. coli* genomes. We believe that this comparative pangenomics approach represents a valuable tool in the future analysis of microbial genomics and population genomics.

Funding information
E.A.C. was funded by the Wellcome Antimicrobial and Antimicrobial Resistance (AAMR) DTP (108 876B15Z).

Acknowledgements
We would like to thank Dr Steven Dunn for fruitful discussions on the visualization process used in the work presented here.

Author contribution
E.A.C.: methodology, formal analysis, visualization, writing – original draft preparation and review and editing. R.J.H.: writing – original draft preparation and review and editing. C.C.: data curation, writing – review and editing. J.O.M.: conceptualization, writing – review and editing. A.M.: conceptualization, supervision, writing – review and editing.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. Distinct evolutionary trajectories in the *escherichia coli* pangenome occur within sequence types. *FigShare*. 2022. DOI: 10.6084/m9.figshare.21360108.
2. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genom* 2018;4:1–8.
3. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* 2010;8:207–217.
4. Dale AP, Woodford N. Extra-intestinal Pathogenic Escherichia coli (ExPEC): disease, carriage and clones. *J Infect* 2015;71:615–626.
5. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2021;19:37–54.
6. Clermont O, Condamine B, Dion S, Gordon DM, Denamur E. The E phylogroup of *Escherichia coli* is highly diverse and mimics the whole *E. coli* species population structure. *Environ Microbiol* 2021;23:7139–7151.
7. Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S, *et al*. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ Microbiol* 2019;21:3107–3117.
8. Wirth T, Falush D, Lan R, Colles F, Mensa P, *et al*. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60:1136–1151.
9. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, *et al*. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* 2008;9:1–14.
10. Petty NK, Ben Zakour NL, Stanton-Cook M, Skippington E, Totsika M, *et al*. Global dissemination of a multidrug resistant Escherichia coli clone. *Proc Natl Acad Sci* 2014;111:5694–5699.
11. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, *et al*. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *mBio* 2016;7:e02162.
12. Pitout JDD, DeVinney R. *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Res* 2017;6:1–7.
13. Matamoros S, van Hattem JM, Arcilla MS, Willemse N, Melles DC, *et al*. Global phylogenetic analysis of *Escherichia coli* and plasmids carrying the mcr-1 gene indicates bacterial diversity but plasmid restriction. *Sci Rep* 2017;7:1–9.
14. Oteo J, Diestra K, Juan C, Bautista V, Novais A, *et al*. Extended-spectrum beta-lactamase-producing *Escherichia coli* in Spain belong to a large variety of multilocus sequence typing types, including ST10 complex/A, ST23 complex/A and ST131/B2. *Int J Antimicrob Agents* 2009;34:173–176.
15. Dallman TJ, Greig DR, Gharbia SE, Jenkins C. Phylogenetic structure of Shiga toxin-producing *Escherichia coli* O157:H7 from sublineage to SNPs. *Microb Genom* 2021;7.
16. Gordienko EN, Kazanov MD, Gelfand MS. Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*. *J Bacteriol* 2013;195:2786–2792.
17. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, *et al*. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc Natl Acad Sci* 2005;102:13950–13955.
18. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, *et al*. The ecology and evolution of pangenomes. *Curr Biol* 2019;29:R1094–R1103.
19. Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, *et al*. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 2008;190:6881–6893.
20. Hall RJ, Whelan FJ, Cummins EA, Connor C, McNally A, *et al*. Gene-gene relationships in an *Escherichia coli* accessory genome are linked to function and mobility. *Microb Genom* 2021;7.
21. Horesh G, Taylor-Brown A, McGimpsey S, Lassalle F, Corander J, *et al*. Different evolutionary trends form the twilight zone of the bacterial pan-genome. *Microb Genom* 2021;7.
22. Horesh G, Blackwell GA, Tonkin-Hill G, Corander J, Heinz E, *et al*. A comprehensive and high-quality collection of *Escherichia coli* genomes and their genes. *Microb Genom* 2021;7:1–15.
23. Whelan FJ, Hall RJ, McInerney JO. Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Mol Biol Evol* 2021;38:3697–3708.
24. Domingo-Sananes MR, McInerney JO. Mechanisms that shape microbial pangenomes. *Trends Microbiol* 2021;29:493–503.
25. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. *Nat Rev Genet* 2018;19:549–565.
26. Liao J, Guo X, Weller DL, Pollak S, Buckley DH, *et al*. Nationwide genomic atlas of soil-dwelling Listeria reveals effects of selection and population ecology on pangenome evolution. *Nat Microbiol* 2021;6:1021–1030.
27. McNally A, Kallonen T, Connor C, Abudahab K, Aanensen DM, *et al*. Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *mBio* 2019;10:1–19.
28. Zong Z, Fenn S, Connor C, Feng Y, McNally A. Complete genomic characterization of two *Escherichia coli* lineages responsible for a cluster of carbapenem-resistant infections in a Chinese hospital. *J Antimicrob Chemother* 2018;73:2340–2346.
29. Feng Y, Liu L, Lin J, Ma K, Long H, *et al*. Key evolutionary events in the emergence of a globally disseminated, carbapenem resistant clone in the *Escherichia coli* ST410 lineage. *Commun Biol* 2019;2:322.
30. Cummins EA, Snaith AE, McNally A, Hall RJ. The role of potentiating mutations in the evolution of pandemic *Escherichia coli* clones. *Eur J Clin Microbiol Infect Dis* 2021.
31. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, *et al*. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet* 2016;12:1–16.
32. Decano AG, Downing T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci Rep* 2019;9:1–13.
33. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. The enterobase user's guide, with case studies on salmonella transmissions, yersinia pestis phylogeny, and escherichia core genomic diversity. *Genome Res* 2020;30:138–152.
34. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, *et al*. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;17:1–14.
35. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 2010;11:595.
36. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
37. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:1–21.
38. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;28:33–36.
39. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, *et al*. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* 2017;34:2115–2122.
40. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–D314.
41. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.

42. Jones P, Binns D, Chang HY, Fraser M, Li W, *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236–1240.
43. Blum M, Chang H-Y, Chuguransky S, Grego T, Kandasaamy S, *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021;49:D344–D354.
44. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* 2018;46:7542–7553.
45. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for python. *BMC Syst Biol* 2013;7:74.
46. King ZA, Lu J, Dräger A, Miller P, Federowicz S, *et al*. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res* 2016;44:D515–22.
47. Chattopadhyay S, Weissman SJ, Minin VN, Russo TA, Dykhuizen DE, *et al*. High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection. *Proc Natl Acad Sci U S A* 2009;106:12412–12417.
48. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* 2010;60:708–720.
49. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. *BMC Genomics* 2012;13:577.
50. Saris PEJ, Palva ET. Regulation of manX (ptsL) and manY (pel) genes required for mannose transport and penetration of λDNA in *Escherichia coli* K12. *FEMS Microbiology Letters* 1987;44:377–382.
51. Wehmeier UF, Nobelmann B, Lengeler JW. Cloning of the *Escherichia coli* sor genes for L-sorbose transport and metabolism and physical mapping of the genes near metH and iclR. *J Bacteriol* 1992;174:7784–7790.
52. Boos W, Shuman H. Maltose/maltodextrin system of *Escherichia coli*: transport, metabolism, and regulation. *Microbiol Mol Biol Rev* 1998;62:204–229.
53. Nobelmann B, Lengeler JW. Sequence of the gat operon for galactitol utilization from a wild-type strain EC3132 of *Escherichia coli*. *Biochim Biophys Acta* 1995;1262:69–72.
54. Sampaio M-M, Chevance F, Dippel R, Eppler T, Schlegel A, *et al*. Phosphotransferase-mediated transport of the osmolyte 2-O-alpha-mannosyl-D-glycerate in *Escherichia coli* occurs by the product of the mngA (hrsA) gene and is regulated by the mngR (farR) gene product acting as repressor. *J Biol Chem* 2004;279:5537–5548.
55. Liu JY, Miller PF, Willard J, Olson ER. Functional and biochemical characterization of *Escherichia coli* sugar efflux transporters. *J Biol Chem* 1999;274:22977–22984.
56. Saier MH, Reizer J. Proposed uniform nomenclature for the proteins and protein domains of the bacterial phosphoenolpyruvate: sugar phosphotransferase system. *J Bacteriol* 1992;174:1433–1438.
57. Deutscher J, Francke C, Postma PW. How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiol Mol Biol Rev* 2006;70:939–1031.
58. Xu Y, Zhou NY. MhpA Is a hydroxylase catalyzing the initial reaction of 3-(3-hydroxyphenyl)propionate catabolism in *Escherichia coli* K-12. *Appl Environ Microbiol* 2020;86:e02385-19.
59. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
60. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* 2019;28:1947–1951.
61. Okuyama M, Mori H, Chiba S, Kimura A. Overexpression and characterization of two unknown proteins, YicI and YihQ, originated from *Escherichia coli*. *Protein Expr Purif* 2004;37:170–179.
62. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, *et al*. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;49:D480–D489.
63. Ferguson AD, Hofmann E, Coulton JW, Diederichs K, Welte W. Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide. *Science* 1998;282:2215–2220.
64. Bobik TA, Havemann GD, Busch RJ, Williams DS, Aldrich HC. The propanediol utilization (pdu) operon of *Salmonella enterica* serovar *Typhimurium* LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B(12)-dependent 1, 2-propanediol degradation. *J Bacteriol* 1999;181:5967–5975.
65. Makui H, Roig E, Cole ST, Helmann JD, Gros P, *et al*. Identification of the *Escherichia coli* K-12 Nramp orthologue (MntH) as a selective divalent metal ion transporter. *Mol Microbiol* 2000;35:1065–1078.
66. Waterson RM, Conway RS. Enoyl-CoA hydratases from *Clostridium acetobutylicum* and *Escherichia coli*. *Methods Enzymol* 1981;71 Pt C:421–430.
67. McNally A, Thomson NR, Reuter S, Wren BW. "Add, stir and reduce": *Yersinia* spp. as model bacteria for pathogen evolution. *Nat Rev Microbiol* 2016;14:177–190.
68. Chaudhuri RR, Henderson IR. The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol* 2012;12:214–226.
69. Hastak P, Cummins ML, Gottlieb T, Cheong E, Merlino J, *et al*. Genomic profiling of *Escherichia coli* isolates from bacteraemia patients: a 3-year cohort study of isolates collected at a Sydney teaching hospital. *Microb Genom* 2020;6:1–16.
70. Viladomiu M, Metz ML, Lima SF, Jin W-B, Chou L, *et al*. Adherent-invasive *E. coli* metabolism of propanediol in Crohn's disease regulates phagocytes to drive intestinal inflammation. *Cell Host Microbe* 2021;29:607–619.
71. Abram K, Udaondo Z, Bleker C, Wanchai V, Wassenaar TM, *et al*. Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups. *Commun Biol* 2021;4:117.
72. Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, *et al*. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J* 2020;14:1247–1259.
73. Cummins EA, Hall RJ, McInerney JO, McNally A. Prokaryote pangenomes are dynamic entities. *Curr Opin Microbiol* 2022;66:73–78.
74. Gori A, Harrison OB, Mlia E, Nishihara Y, Chan JM, *et al*. Pan-GWAS of *Streptococcus agalactiae* highlights lineage-specific genes associated with virulence and niche adaptation. *mBio* 2020;11:e00728-20.
75. Letunic I, Bork P, Gmbh BS. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;49:W293–W296.