Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility

Hall, Rebecca J; Whelan, Fiona J; Cummins, Elizabeth A; Connor, Christopher; McNally, Alan; McInerney, James O

DOI:
10.1099/mgen.0.000650

MICROBIAL GENOMICS

RESEARCH ARTICLE

Hall *et al.*, *Microbial Genomics* 2021;7:000650

DOI 10.1099/mgen.0.000650

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Gene-gene relationships in an *Escherichia coli* accessory genome are linked to function and mobility

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

Abstract

The pangenome contains all genes encoded by a species, with the core genome present in all strains and the accessory genome in only a subset. Coincident gene relationships are expected within the accessory genome, where the presence or absence of one gene is influenced by the presence or absence of another. Here, we analysed the accessory genome of an *Escherichia coli* pangenome consisting of 400 genomes from 20 sequence types to identify genes that display significant co-occurrence or avoidance patterns with one another. We present a complex network of genes that are either found together or that avoid one another more often than would be expected by chance, and show that these relationships vary by lineage. We demonstrate that genes co-occur by function, and that several highly connected gene relationships are linked to mobile genetic elements. We find that genes are more likely to co-occur with, rather than avoid, another gene in the accessory genome. This work furthers our understanding of the dynamic nature of prokaryote pangenomes and implicates both function and mobility as drivers of gene relationships.

DATA SUMMARY

All Supplementary Data files and the Python scripts used in the analyses are available at doi.org/10.17639/nott.7103.

INTRODUCTION

Escherichia coli is one of the most widely used and studied bacterial species in microbiology. Recent efforts have resulted in a cataloguing of the essential genes in this species using transposon-directed insertion site sequencing (TraDIS) [1], thereby defining which genes are indispensable and which are not. It is arguable that the accessory genes found in the *E. coli* pangenome are not essential for the survival of the species, making it somewhat of a curiosity that such a large set of accessory genes is maintained. A separate study of a dataset of 53 *E. coli* genomes identified more than 3000 metabolic innovations that all arose as a consequence of the acquisition via horizontal gene transfer (HGT) of a single piece of DNA less than 30 kb in length, and that 10.6% of innovations were facilitated by earlier acquisitions [2]. This suggests therefore that HGT has the ability to bring together sets of genes that, when combined, provide benefits over and above the benefits that the genes could confer on their own. Indeed, it is likely that there are situations where the genes on their own might be deleterious, but in combination they confer a fitness advantage [3], thereby contributing to the maintenance of a large accessory genome.

Analyses of genome content shows that *E. coli* pangenome evolution is driven by differential gain and loss of accessory genes, including plasmids, phage, and pathogenicity islands [4–6]. What is not yet clear is the underlying structure of the pangenome and which genes are key influencers of the presence or absence of other genes in a given genome. Plasmid mobility, for instance, is anticipated to be an important agent of pangenome structuring [7]. Plasmids engage a diverse range of proteins for the purpose of plasmid partitioning and maintenance, but the presence and absence of these protein encoding genes in accessory genomes has not been established concretely.

Received 26 March 2021; Accepted 10 July 2021; Published 09 September 2021

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

***Correspondence: James O. McInerney,** james.mcinerney@nottingham.ac.uk

Keywords: *Escherichia coli*; evolution; gene co-occurrence; pangenome.

Abbreviations: HGT, horizontal gene transfer; KEGG, kyoto encyclopedia of genes and genomes; MGE, mobile genetic elements; PTS, phosphotransferase system; ST, sequence type; T2SS, type II secretion system; T4SS, type IV secretion system.

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

Variation in overall genome sequence has the potential to influence which other genes can or cannot be present in a genome. A gene might be beneficial to one strain of a species, but deleterious in a different strain. Sets of genes encoding proteins that together form an essential biosynthetic pathway, for example, are expected to co-occur in the same genome, likewise those that form multi-protein complexes. The presence of a gene or set of genes can also exclude, or 'avoid', others from the same genome. Competitive exclusion is exemplified in *Salinispora* species, where only one of two iron-chelating siderophore gene clusters is ever found in a given strain, despite evidence of frequent HGT [8]. An understanding of gene-gene dependencies and influences can therefore provide a real insight into how pangenomes originate and are maintained. We can address at least two questions when we look at gene-gene co-occurrence and avoidance patterns. We can ask whether the pangenome has been structured by random genetic drift or by natural selection, and we can ask whether gene co-occurrence has been more important than avoidance in a pangenome.

In practical terms, the large *E. coli* accessory genome can serve as a testbed for calibrating how much gene co-occurrence and avoidance can teach us about pangenome origins and evolution. A considerable amount of diversity is found across all known sequence types (STs) [4, 9–12], while a large body of experimental validation of gene function has had the effect of reducing the number of unknown open reading frames (ORFs) in the pangenome. It is anticipated that there may be collections of lineage-specific coincident gene-gene relationships, but the process driving these relationships is not yet clear.

Here, we interrogate *E. coli* pangenome dynamics in order to identify coincident gene-gene relationships [13]. In a large sample of the total *E. coli* pangenome, comprising 400 judiciously selected genomes spanning 20 different STs, we have identified significant gene-gene co-occurrences and avoidances. We found more co-occurrence relationships than avoidance relationships in the accessory genome, and found connected components that are enriched in genes that share a common function, suggesting that the role that genes play in a system can partially explain their significant co-occurrence. We also find that MGEs extensively influence gene co-occurrence and avoidance, including for genes linked to detoxification and antimicrobial resistance (AMR). Our results provide an extensive set of possible empirical experiments that, together with our *in silico* predictions, further our understanding of the complex ecological interactions and dynamics in the *E. coli* pangenome.

Impact Statement

The pangenome of a species encompasses the core genes encoded by all genomes, as well as the accessory genes found in only a subset. Much remains to be understood about the relationships and interactions between accessory genes; in particular, what drives pairs of genes to appear together in the same genome, or what prevents them from being in the same genome together, more often than expected by chance. How these co-occurrence and avoidance relationships develop, and what effect they have on the dynamics and evolution of the pangenome as a whole, is largely unknown. Here, we present a springboard for understanding prokaryote pangenome evolution by uncovering significant gene relationships in a model *Escherichia coli* pangenome. We identify mobile genetic elements and the sharing of common function as possible driving forces behind the co-occurrence of accessory genes. Furthermore, this work offers an extensive dataset from which gene relationships could be identified for any gene of interest in this *E. coli* accessory genome, providing a rich resource for the community.

METHODS

E. coli genomes and pangenome

A set of 400 *E. coli* genomes were downloaded from EnteroBase using a custom Python script (github.com/C-Connor/EnterobaseGenomeAssemblyDownload). The set contained 20 genomes from 20 different STs; ST3, ST10, ST11, ST12, ST14, ST17, ST21, ST28, ST38, ST69, ST73, ST95, ST117, ST127, ST131, ST141, ST144, ST167, ST372, ST648. The STs were chosen to include multiple extraintestinal pathogenic *E. coli* (ExPEC), enterohemorrhagic *E. coli* (EHEC), and commensal lineages. The genes within each genome were annotated using Prokka (v1.12) [14] and a gene presence-absence matrix was generated with Panaroo (v1.1.2) [15] using the default settings and the -a core flag to generate a core gene alignment. Panaroo's default definitions of core ($99 \leq \times \leq 100\%$), soft core ($95 \leq \times \leq 99\%$), shell ($15 \leq \times \leq 95\%$), and cloud ($0 < 15\%$) were used.

Core gene phylogeny

The core gene alignment file produced by Panaroo was trimmed using trimAl [16] with a -gt value of 1. Phylogenetic relationships between the strains were inferred using the core gene set from the pangenome. A core gene phylogeny was constructed from the trimmed alignment using the IQ-Tree software [17] with the GTR+I+G substitution model (as justified in [18]). Phylogenetic tree visualisation was carried out using the Interactive Tree of Life v5 (iTOL) [19].

Coincident gene identification and visualisation

Gene co-occurrences (associations) and avoidance (dissociations) were identified using Coinfinder [13] using default settings with a threshold of 0.01 for low abundance filtering and employing the Bonferroni correction to account for multiple testing. Networks were visualised using the Gephi software program (v0.9.2) with the Fructerman-Reingold algorithm used for graph layout. Heatmaps were generated using seaborn (v0.10.1). Hub genes are identified as genes forming a number of gene co-occurrences or avoidances that

is 1.5 times the interquartile range (IQR), as in [20]. Root excluders were defined as a single gene avoiding two or more genes that co-occur with each other within a connected component. Gene names were taken from the Panaroo gene cluster identifications. Gene functionalities of co-occurring and avoiding genes were separated into biological functional categories through searching a combination of UniProt [21], KEGG [22], and BioCyc [23], with further examples of gene functions (e.g. secretion systems, transposons) determined by comparing Panaroo-defined gene annotations against these databases and the literature (cited where applicable). Hypothetical genes were defined as those without a known annotation in the Panaroo outputs.

Prediction of plasmid-encoded genes

Genes were determined to be chromosomal or plasmid-encoded by collecting each contig containing the gene across all strains in which it was present, and then assessing whether those contigs were predicted to be plasmid- or chromosome-derived sequences using mlplasmids [24].

Phosphotransferase system analysis

A list of all phosphotransferase system (PTS) genes was obtained from the KEGG database. The corresponding genes were subset from the *E. coli* gene presence-absence matrix and Coinfinder outputs to highlight those genes found in the *E. coli* pangenome.

Data availability

All Python scripts used for data analysis and figure construction are freely available at doi.org/10.17639/nott.7103. The gene presence-absence matrix generated by Panaroo, core gene phylogeny, and outputs from Coinfinder are also available as Supplementary Data at the same location. A list of genomes mapped to their ST and phylogroup is provided here as a Supplementary Data File. Descriptions of Coinfinder outputs can be found in the original software publication [13].

RESULTS

The *E. coli* pangenome is highly structured

An *E. coli* pangenome, composed of 400 genomes from 20 different STs, was constructed (Fig. S1, available in the online version of this article). Panaroo was used for this analysis as preliminary investigations found that the clustering produced fewer incidences of false positives in the avoidance network than when the gene presence-absence matrix was constructed using Roary [25]. This pangenome sample consists of 3191 core, 120 soft core, 2935 shell, and 11665 cloud genes (Fig. S2), consistent with previous observations that *E. coli* has an open pangenome [9]. Using Coinfinder [13] we identified a gene co-occurrence network that consisted of 8054 nodes joined by a total of 500654 edges (Fig. 1a), and an avoidance network consisting of 3203 nodes joined by 203503 edges (Fig. 1b). Within these networks, each gene cluster is represented by an individual node, and an edge that joins each node indicates a significant gene relationship (either co-occurrence or avoidance) between the two. The nodes are coloured by connected component, defined as a group of genes that form relationships with one another and not with the rest of the network. The co-occurrence network is therefore larger, both in terms of numbers of nodes and also numbers of interactions, though both networks have similar average numbers of connections per node. Of all gene clusters in the gene presence-absence matrix, 45.0% form at least one co-occurrence pair, and 17.9% at least one avoidance. Of the accessory genes analysed by Coinfinder, 77.8% form at least one co-occurring or avoidant pair.

Co-occurring genes share function

The co-occurrence network contains 224 connected components (Fig. 1a), including one large connected component with extensive interactions that can be found with dark green nodes at the centre of Fig. 1a. If indeed co-occurrence is shaped by functional interactions and dependencies, we could expect co-occurrence analyses to pick out known subsystems. Component 150 consists in part of twelve genes involved in the type II secretion system (T2SS) (*epsC-H, epsLM*, *gspK*, *pppA* and *xcpVW*) that facilitates the translocation of a wide variety of proteins from the periplasm to the exterior of the cell. Similarly, component 50 includes *epsE* (T2SS), and the type IV secretion system (T4SS) genes *vir1,4,8,10,11*. Component 150 has a broader distribution across the STs (n=18) than component 50 (n=6), and is found at a much higher abundance. In fact, for 11 of the STs, all 20 genomes contain component 150 (Fig. 1c). These data illustrate that indeed analyses of co-occurrence patterns, as implemented by Coinfinder, can pick out functional associations.

The constituent genes in components 3, 13, 60, and 149 are outlined in Table 1 and are known to function in DNA replication. Of note in these components are the plasmid copy control gene *repB* and the plasmid partitioning protein-encoding gene *parB*, both known to be plasmid-encoded, as well as the chromosome-encoded *dnaJ* that functions in plasmid replication [26, 27]. The fact that these functions are found across four separate co-occurrence components shows that within this set of functions related to mobility, there are co-occurring communities consisting of distinct groups of genes that preferentially co-occur. These components are found in a variety of STs, though in low abundance (Fig. 1c).

Co-occurrence hub genes are linked to virulence and mobile elements

We found considerable variation in the number of significant co-occurrence or avoidance relationships for any individual gene in the *E. coli* accessory genome. From the degree distribution in the co-occurrence network we identified a large group of 427 highly connected 'hub genes' (Fig. 2a), with a high of 896 co-occurrences recorded for an individual gene cluster (group_2839, identified as dnaC_4 from the nonunique gene name). These hub genes either facilitate or

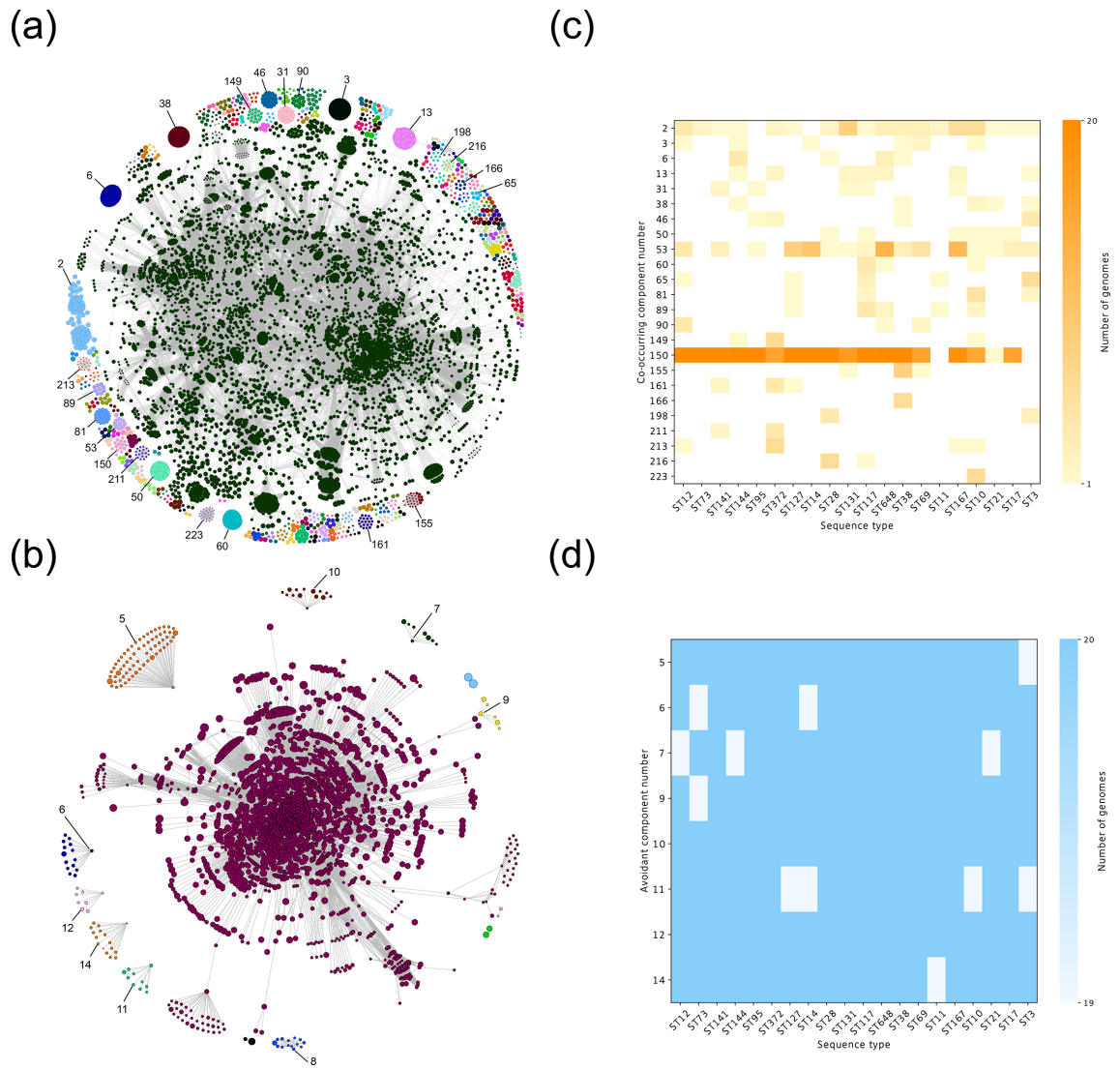

Fig. 1. Gene relationships in the *E. coli* pangenome. (a) An overview of the co-occurrence and (b) avoidance networks, coloured by connected component. Selected connected components are numbered as in the Supplementary Data Files. (c) The number of genomes in which part or all of a co-occurrence and (d) avoidance connected component appears in each ST. An absence of a co-occurrence component in a ST is depicted as colourless.

promote the existence of a large number of other genes in any given genome. The most common known functions of co-occurrence hub genes are attack and defence (including toxin-antitoxin systems, Shiga toxin, CRISPR system subunit, and genes encoding hemolysin and proteins involved in detoxification), metabolism, and DNA and RNA processes (including DNA primase, helicase, and replication proteins, tyrosine recombinases, and proteins involved in DNA repair) (Fig. 2b). Full lists of the number of pairs formed by individual genes are provided as Supplementary information.

A notable subset of the hub genes are linked to DNA exchange and MGEs. Some are known plasmid-encoded genes (*toxB*, *hlyACD*, and *ssb* [27, 28]), and the Shiga toxin subunits *stxAB* [7] are encoded on a single phage [29]. The tyrosine recombinases *xerCD* function in plasmid segregation, as does *parM*, and there are several genes either prophage-encoded (*recE* [30]) or related to phage functions (*intQ*, *tfaE*). The transposase *tnpA* is also a hub gene, forming a co-occurring pair with 841 other genes. These findings indicate a process where hub genes that have a role in lateral mobility of genetic material are specifically co-occurring with genes that confer fitness advantages, on average, when mobilised. This suggests a process of mutual benefit; the mobility enablers co-occurring with the genes most likely to confer positive fitness effects.

The avoidance network is characterised by root excluders

The avoidance network has substantially fewer connected components (n=14) than the co-occurrence network, and many of these components are characterised by the presence

Table 1. Certain co-occurrence components function in fundamental cellular processes. The known gene clusters relating to DNA and RNA replication, regulation and repair found in the co-occurrence connected components 3, 13, 60, and 149. Genes are named as per the Panaroo gene clusters in the gene presence-absence matrix. Where the gene is identified by 'group_', a gene identifier, taken from the Panaroo non-unique gene name, is given in parentheses

Component 3	Component 13	Component 60	Component 149
dnaB_2_dnaB_1_dnaB_3	polA_2	dnaB_3_dnaB_2	group_4304 (*recQ*)
dnaJ_3_dnaJ_1_dnaJ_2	dnaE_1_dnaE_2_dnaE1	ssb_1_ssb_5_ssb_4	srmB_2
ssb_3_ssb_2	dnaG_1	group_3992 (*topB*)	group_296 (*rapA*)
smc__smc_1	dnaQ_1_dnaQ_2		
	group_7368 (*parB*)		
	repB_2		
	lig		
	smc_2_smc		

of a single 'root excluder'; a situation where one gene avoids a gene set that in turn all co-occur with one another (Fig. 1b, Fig. S3). Over half of the components in this network show this pattern, specifically components 5–7, 9–12, and 14 (Supplementary Data Sheet 1). These components are found in at least 19 genomes of all STs (Fig. 1d). None of the root excluders form avoidance hubs (Fig. 2a). As an example, within component 14, the root excluder *dhaR*, a transcriptional regulator of the dihydroxyacetone kinase operon [31] avoids 11 genes involved in the production of T2SS proteins (*gspHK*, *epsEFLM*, *pulDG*, *outC*, *xcpVW*), amongst other known and hypothetical genes. The plasmid-associated nature of some T2SS genes [27] suggests an active process of dissociation or avoidance as a result of their mobility.

Avoidance hub genes are lineage-specific

The avoidance network is smaller than the co-occurrence network with fewer hub genes, though high degree nodes can be identified with the highest showing 749 avoidances for a single gene cluster (atoA, atoD, zraR_2_spo0F, and zraS_2) (Fig. 2a). The *ato* genes encode proteins involved in the degradation and transport of short-chain fatty acids [32], and the *zra* genes encode a two-component system linked to antimicrobial tolerance in *E. coli* [33]. The most enriched function of the avoidance hub genes is in metabolism, and when grouped by function, regulation is the sole category where the number of avoidance hubs is greater than the number of co-occurrence hubs (Fig. 2b).

Some genes avoid considerably more genes than they co-occur with. For instance, the putative diguanylate cyclase *ycdT*, which catalyses the production of cyclic di3',5'-guanylate and is reportedly under positive selection in uropathogenic *E. coli* [12], significantly co-occurs only with *pgaA-D*. The proteins encoded by *pgaA-D* are involved in biofilm formation. In contrast, *ycdT* avoids 83 other gene clusters including the CRISPR system Cascade subunits *casACDE*. All 400 genomes have at least one of either *ycdT* or the *casACDE* genes, but they are found in isolation in 298 genomes. The only STs where this *ycdT-casACDE* avoidance pattern is not observed in any of the genomes are ST3, ST17, ST11, ST38, and ST28, and therefore does not demonstrate a clear phylogenetic split (Fig. S1); ST3 and ST17 are phylogroup B1, ST11 belongs to phylogroup E, and ST38 and ST28 are phylogroups D and B2, respectively (Supplementary Data File). We might speculate that there is another element of genome variation that modulates whether *ycdT-casACDE* can or cannot be present in the same genome.

The avoidance hub genes are also enriched in functions related to secretion systems (Fig. 2b). We found that between 627 and 661 gene clusters avoid the T2SS-related genes *gspA-M* and the probable bifunctional chitinase/lysozyme that is secreted by the T2SS [34]. Of these, one is particularly pertinent; *spiA* (also known as *escC*), encoding a type III secretion system (T3SS) outer membrane secretin [35]. Incidences of *spiA* avoiding *gspA-M* gene clusters are found in all 20 STs. These data provide some evidence that secretion systems may have a substantial genome-wide influence.

Repeated co-occurrence of resistance genes and transposon genes

We have demonstrated that genes that share functions or pathways will commonly co-occur, and that certain highly connected co-occurrences and avoidances involve genes associated with plasmids. We hypothesised that other agents of horizontal transfer could therefore form the centre of co-occurrence hubs as a result of their mobility. The transposase *tnpA*, for example, is a co-occurrence hub gene (n=841 gene pairs) (Fig. 2a). This is the only transposon of the 22 within the co-occurrence network that is classified as a hub gene. We found that three connected components (excluding the large component in the centre of Fig. 1a) contain transposons.

The first, component 81, is comprised of a collection of genes enriched in functions related to metal detoxification. Eight genes relate to copper (*copABDR*, *pcoCE*) and silver (*silPE*) resistance, and six encode proteins involved in producing a

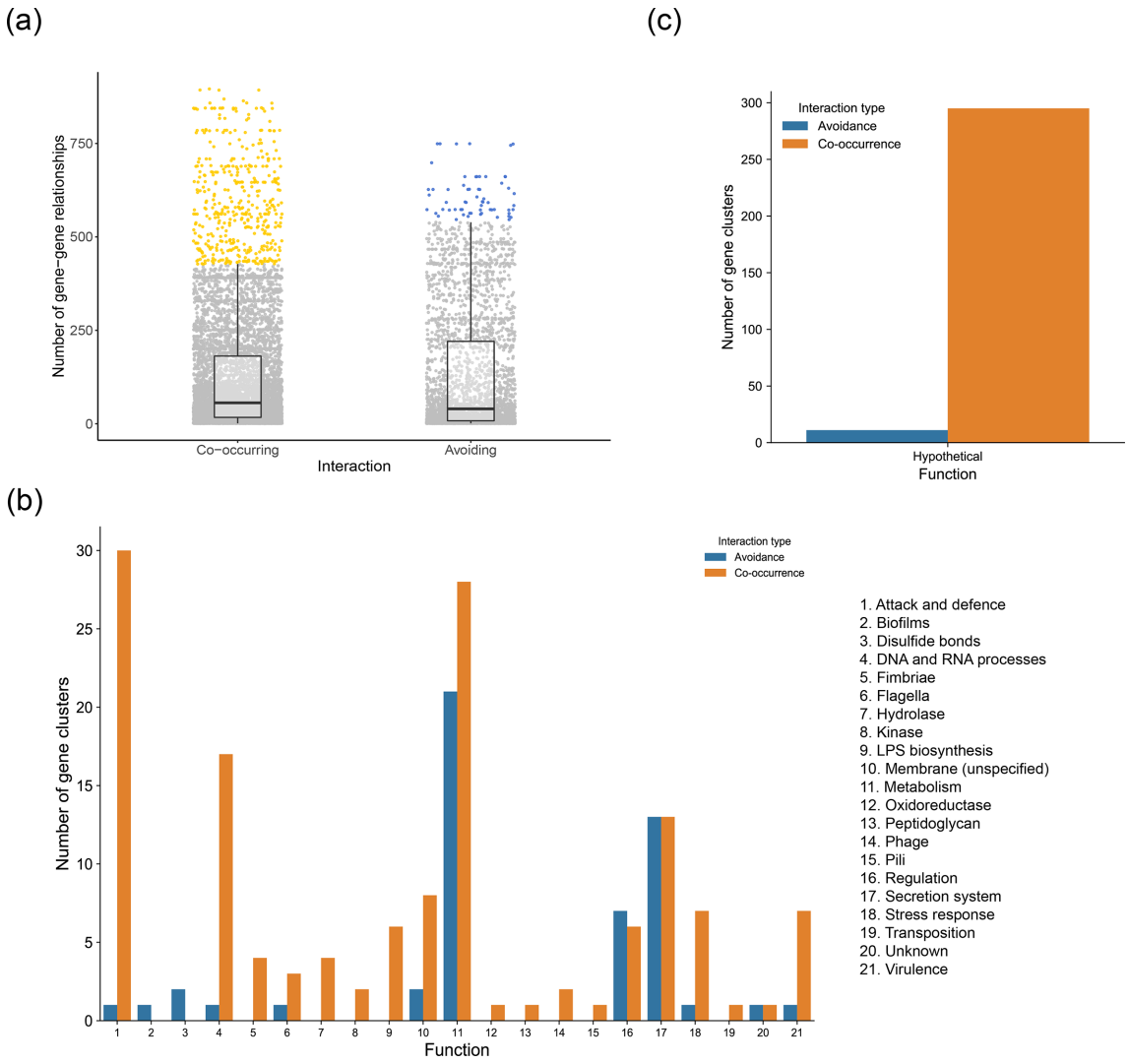

Fig. 2. Highly connected hub genes in the accessory genome. (a) The number of gene-gene relationships formed by individual genes. Hub genes (1.5 times the upper IQR) are coloured orange (co-occurrence) and blue (avoidance). (b) Co-occurrence (orange) and avoidance (blue) hub genes categorised by loose biological function. Attack and defence includes CRISPR system subunits, toxin-antitoxin systems, toxin production, detoxification. DNA and RNA processes includes replication, repair, recombination, and plasmid segregation. (c) The number of hub genes encoding hypothetical proteins.

cation efflux system (*cusABCFSR*) (Fig. 3a). A Tn7 transposition protein, *tnsA*, is also present. These systems may be acquired and retained together in order to provide a broader spectrum of metal detoxification. This connected component is only found in four of the 20 STs; ST3 (number of individual genomes=2), ST10 (*n*=5), ST117 (*n*=2), and ST127 (*n*=1) (Fig. 1c). These STs are not close phylogenetically (Fig. S1), belonging to phylogroups B1 (ST3), A (ST10), F (ST117) and B2 (ST127), indicating that it is not simply a lineage-dependent characteristic.

Component 53 consists predominantly of the Tn10 transposon proteins *tetCD* that confer tetracycline resistance, as well as *gltS* (a sodium/glutamate symporter) and five hypothetical proteins (Fig. 3b). This component is found in more than half of the genomes in the multidrug resistance-associated ST648 (*n*=13, phylogroup F) and ST167 (*n*=12, phylogroup A), and in at least one genome in 16 of the 20 STs (Fig. 1c). The co-occurrence of *gltS* and the hypothetical proteins with the tetracycline resistance genes could suggest they may either be linked in a novel way to AMR, or, alternatively, that they have been transferred with the resistance genes and their co-occurrence is purely as a result of this hitchhiking effect.

The remaining connected component that contains a transposon is component 2. This component is comprised of two quasi-cliques and consists, in part, of the T4SS genes *virB1,2,4,6,8–11*, a putative transposon Tn552 DNA invertase (*bin3*), the conjugal transfer protein *traG*, and the tyrosine recombinase *xerD* (Fig. 3c). The presence of these genes throughout both of the quasi-cliques strongly suggests that MGEs are influencing the co-occurrence

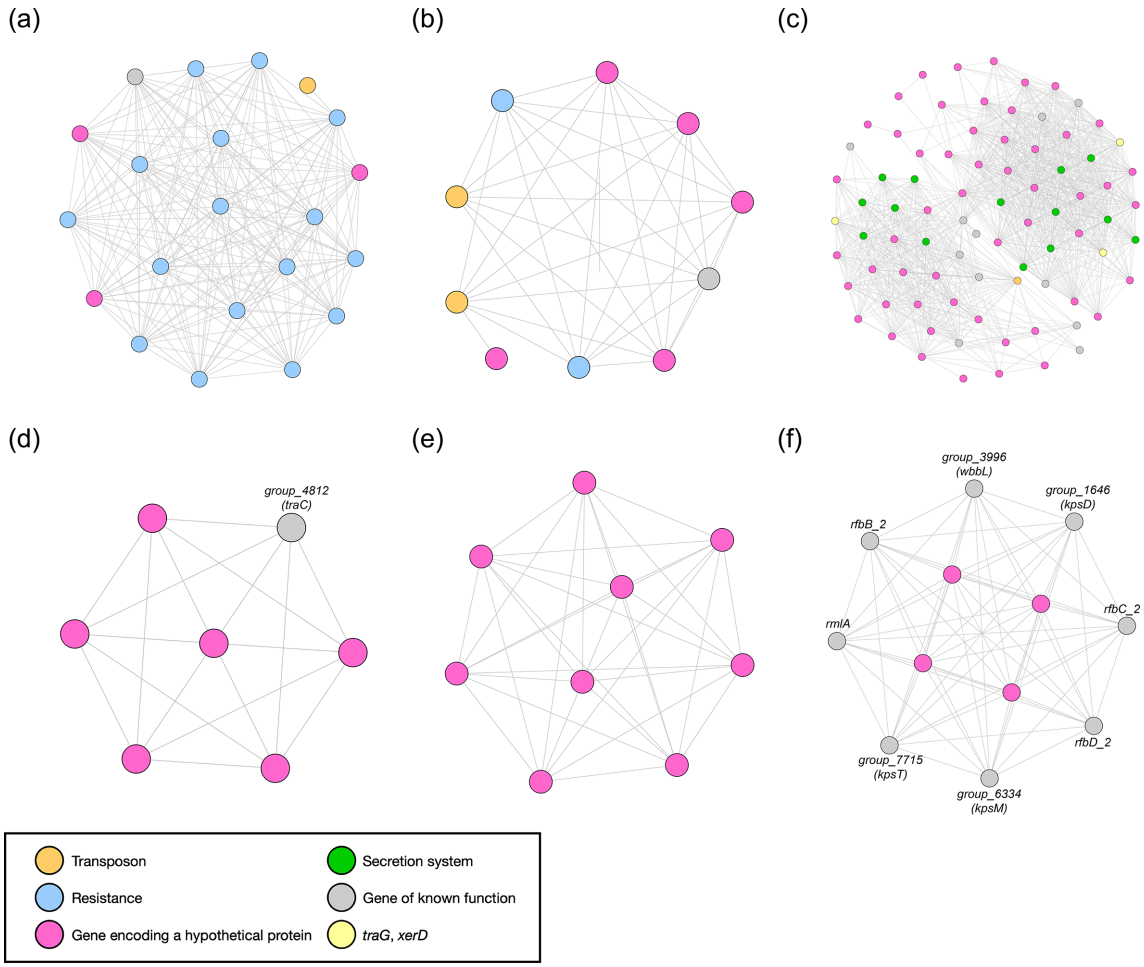

Fig. 3. Some co-occurrence connected components are linked to transposons (orange nodes); (a) co-occurrence component 81, enriched in genes related to metal detoxification (blue). (b) Co-occurrence component 53, containing genes linked to tetracycline resistance (blue). (c) Co-occurrence component 2, comprised in part of genes encoding secretion systems (green). Other components are enriched in hypothetical genes (pink), including co-occurrence components 198 (d) and 89 (e). (f) Co-occurrence component 90 contains genes relating to capsule formation, as well as four hypothetical genes. Select gene clusters are labelled. Genes with a function that is known but unrelated to the highlighted functions are depicted with a grey node.

relationships in this component. This component is found in at least one genome in 18 different STs (Fig. 1c), providing further evidence for gene mobility. The transposons could influence the transfer of the genes in these components, or they could be hitchhiking between genomes alongside the rest of the component.

Hypothetical proteins form a hidden network with a large influence on the *E. coli* accessory genome

A theme that emerged within the connected components is the central role in structuring the pangenome that is clearly being played by many genes with unknown function. First, 67.5% of the co-occurrence hub genes (n=295) and 17.5% (n=11) of the avoidant hub genes encode hypothetical proteins (Fig. 2c). This was surprising given the extent to which *E. coli* has been studied. While many genes in our dataset have assigned functions, 11491 ORFs in the gene presence-absence matrix used as input to Coinfinder were not ascribed a function, 64.2% of the total. Second, certain connected components are enriched in genes encoding hypothetical proteins, with several connected components consisting solely of unknown ORFs. Examples of this are co-occurrence components 198 and 89 (Fig. 3d, e), both of which are found in several different STs, though no ST contains both components (Fig. 1c). The entirety of component 89 consists of hypothetical proteins, while all but one of component 198 is a hypothetical protein. All gene clusters within these two components are predicted by mlplasmids to be chromosomal. It should also be noted that both of these connected components also form a clique; every node in the connected component is connected to every other node. This means that every gene in the component shows a significant co-occurrence pattern with every other gene. It is therefore highly likely that these groups of genes are tightly, functionally linked to one another, though there is no published information related to the function of any of these genes.

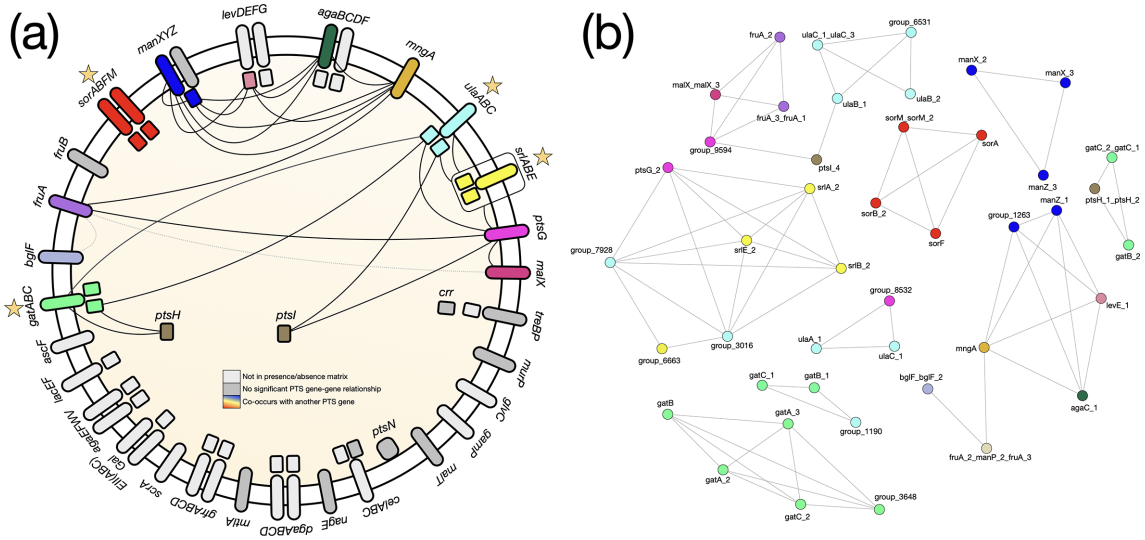

Fig. 4. PTS gene relationships are not universal. (a) A schematic overview of the co-occurrence PTS gene relationships. PTS gene systems found in the KEGG database are coloured by those not present in this *E. coli* pangenome (white), those which do not form a coincident PTS-PTS pair in the accessory genome (grey), and the clusters that form a co-occurring pair with another PTS gene (coloured by system). A significant relationship is indicated by a solid black line. Where a relationship is observed with all genes in a PTS, the line connects to a box around the system. Transporters that consist of genes that all significantly co-occur with one another are marked with a yellow star. The grey dashed line connecting *fruA* with *bglF* and *malX* indicates a significant relationship in both the co-occurrence and avoidance datasets for different clusters of the same gene identification. (b) The specific co-occurrences by gene cluster. Nodes are coloured by system as in (a).

We observe several connected components where the majority of genes in a component share a similar function. This makes it tempting to imply a putative role for the genes encoding the remaining hypothetical proteins in those components. Co-occurrence component 90 consists of five genes with known functions; three genes that are found in the dTDP-rhamnose biosynthesis pathway, a rhamnoslytransferase, and a polysialic acid transporter. This leaves four genes in this connected component that encode hypothetical proteins (Fig. 3f). All gene clusters that form component 90, regardless of function, are predicted by mlplasmids to be chromosomal. We could speculate that these hypothetical proteins may therefore function in lipopolysaccharide or capsule formation [36]. Component 2 consists of 82 gene clusters, including 53 that encode hypothetical proteins and 14 that relate to secretion systems (Fig. 3c). These observations show that there is an important network of genes of unknown function that exert a large influence on the *E. coli* accessory genome, and that our approach can generate meaningful hypotheses to inform investigation of the function of those genes.

The drivers of gene-gene relationships are multifaceted

We have identified co-occurrence connected components that entirely consist of, or are enriched in, genes that share a common, identifiable role. We have also highlighted gene relationships that appear to be influenced by the presence of MGEs. If function and mobility were the only drivers behind significant gene-gene relationships, it might be expected that genes encoded on the same operon, or that function in a discrete system, will share the same patterns of co-occurrence. To test this, we focused on phosphotransferase system (PTS) genes, which are used to import carbohydrates [37]. These systems were chosen because several have been identified in *E. coli*, they are typically multi-component [38], and the presence of one system might logically be linked to the presence or absence of another.

We collated all genes encoding PTS components and identified those that form a co-occurring pair with at least one other PTS gene. We found no correlation between whether the encoded proteins are membrane-bound or cytoplasmic and the likelihood that they will co-occur with another PTS gene. Of the 67 PTS genes detailed in the KEGG database [22], 29 were not observed in our gene presence-absence matrix, and a further ten did not manifest a significant co-occurrence with any other PTS gene (Fig. 4a).

We found a complex pattern of co-occurrence across the systems. Certain PTS genes do show a consistent pattern for the complete system. For example, *srlABE* all co-occur with the same genes, and *sorABFM* only co-occur with each other (Fig. 4a, b). In contrast, the *manXYZ*, *levDEFG*, *agaBCDF*, and *ulaABC* gene relationships are not system-specific, varying instead by individual gene. Selection pressures on these systems are complex and heterogeneous, and gene co-occurrence relationships are likely driven by multiple factors.

DISCUSSION

The open pangenome of *E. coli* [9] provides a rich testbed to help understand genome and pangenome dynamics. Factors such as the immediate microenvironment in which a strain lives are known to influence the accessory gene content of any given genome and, consequently, the pangenome [39–41]. Though there are known examples of how the fitness of one gene in a genome is influenced by the presence or absence of other genes, there has been no systematic, large-scale study of how genetic background, in terms of the presence or absence of genes in a genome, influences the fitness effect of an incoming gene [13]. Recently, however, it has been shown that the genetic background of a genome has a direct effect on whether or not a gene is essential [42]. We have little knowledge of why some lineages encode genes that others do not, and the extent to which the observed encoded genes influence the likelihood of successfully integrating other incoming genes. We also know little about whether pangenomes arise as a result of drift, or whether they are maintained by selection [43–45]. Significant co-occurrence of genes that share common function have been identified recently in a *Pseudomonas* pangenome, providing support for selection as a driver of pangenome evolution [20].

We build upon this work by analysing a network of coincident gene relationships in a model *E. coli* pangenome, with the observation that many co-occurring connected components are enriched in genes that share function, broadly defined. We found that nearly half of all gene clusters in the pangenome form a significant pair, and that this relationship is more likely to be one of co-occurrence than avoidance. Genes co-occur at least in part because they share function, whereas direct, antagonistic avoidances are less common. This shows that the *E. coli* pangenome is highly structured. The presence of a number of gene clusters related to DNA and RNA processes, including replication and repair, forming co-occurrence pairs in the accessory genome was perhaps surprising given their core roles within a cell, but similar genes have been identified previously in *Neisseria* species accessory genomes [46] and unique alleles of such genes have also been found in an *E. coli* ST131 pangenome [41]. We suggest sharing common function is one contributor to the overall genegene co-occurrence network, but acknowledge that it is not the only factor. For co-occurrences that are, for example, only observed in a single ST, it is possible that this is instead a product of co-inheritance. This could be mitigated by manually selecting an appropriate D value cut-off, or by using a ClonalFrameML tree [47] as input to Coinfinder to reduce recombination signal. Co-localisation may also be a factor, although recent work into gene-gene relationships in *Pseudomonas* found that many co-occurrence relationships are still statistically significant even when synteny is considered [20]. Linkage can occur but is not a strong feature of co-occurrence [20], and so was therefore not considered here to be the primary driving force behind the significant relationships. The variation in co-occurrence patterns in PTS genes shown here also indicates that function and co-localisation are not the only drivers of these relationships. An interesting progression may be to extend the gene-by-gene work presented in PTS genes to all genes forming coincident relationships, thereby possibly uncovering relationships that may violate functional linkage assumptions and, subsequently, as yet unknown drivers of gene-gene relationships.

We also propose, alongside the role that sharing function plays, that gene mobility could be an additional mechanism behind the formation of these gene-gene relationships. MGEs are a known link to gene essentiality and virulence in *E. coli* [4, 42, 48], and they have been implicated in driving accessory genome differences in a *Listeria monocytogenes* pangenome [49]. Here, we have demonstrated that they also influence gene co-occurrences by uncovering hub genes and connected components linked to or encoded on MGEs [7, 27, 28, 50–54]. This includes known phage-encoded virulence factors [29, 55], transposons, and genes involved in conjugation. It is possible that a transposon that co-localises and co-occurs with a number of genes, for example those related to metal detoxification, could mobilise these genes within a given niche in which they are beneficial; this is particularly relevant given the high clinical importance of *E. coli*. Together, this progresses current understanding of prokaryote pangenomes by suggesting that they are structured and dynamic, but also further underscores the importance of HGT in driving pangenome evolution [56].

Furthermore, we found that these gene relationships are often non-randomly distributed across the pangenome, with some being more frequently observed in specific STs. For example, co-occurrences related to resistance phenotypes are found through the different STs, but are particularly evident in the MDR-associated ST167 and ST648 [57, 58]. ST-specific differences that confer pathogenicity and resistance in *E. coli* are well-studied [48, 58, 59], but this work provides a new layer of understanding into how the accessory genome may interact to this end. We suggest that certain gene collections are required by specific lineages and that this may be driven in part by MGEs.

Alongside these gene co-occurrences of known function, we have also uncovered a network of genes with unknown function that influence the structure of the pangenome through the high number of genes that they co-occur with and avoid. *E. coli* has fewer genes of unknown function than most prokaryotes, although there is evidence that many of the accessory genes in lineages such as ST131 are of hypothetical function [43]. The number of coincident gene relationships formed by such genes here highlights the challenges in understanding the global prokaryote pangenome. This potential issue should be considered when approaching this pipeline. It may be possible, however, to use gene-gene co-occurrences to assign putative functions to hypothetical genes; by discerning the identity of sets of genes that share identical patterns of co-occurrence, for example, a tentative function could be suggested for those about which little is known. Given the prevalence of MGEs in the co-occurrence network, it is tempting to conclude that at least some of the

high proportion of co-occurrence hub genes identified as encoding hypothetical proteins may be related to mobility.

The data presented here support the concept that pangenomes create pangenomes. The diversity of gene content in a cosmopolitan species such as *E. coli* means that the fitness effect of gaining and losing individual genes is not the same for all constituent genomes. We observed the presence or absence of some genes only when other genes are also present or absent. We consistently observed some pairs of genes co-occurring or avoiding repeatedly across the diversity of the group of genomes, and root excluders consistently avoiding certain cliques. We observed unknown ORFs that significantly co-occur in the same genome, even though their distribution in the pangenome is patchy. There is therefore an emerging logic to the *E. coli* pangenome that clearly identifies natural selection to have frequently dominated over genetic drift.

Funding information
R.J.H., was supported by the BBSRC (BB/N018044/2), awarded to J.O.M. F.J.W., was funded by a Marie Skłodowska-Curie Individual Fellowship (GA no. 793818). E.A.C., was funded by the Wellcome AAMR DTP and CC by the Wellcome Midas DTP.

Acknowledgements
The authors thank M.R. Domingo-Sananes and S. Thorpe for their insightful feedback on the work presented here, and the two anonymous reviewers for their thoughtful suggestions.

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

References

1. Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, *et al*. The essential genome of *Escherichia coli* K-12. *mBio* 2018;9:02096–17.
2. Pang TY, Lercher MJ. Each of 3,323 metabolic innovations in the evolution of *E. coli* arose through the horizontal transfer of a single DNA segment. *Proc Natl Acad Sci U S A* 2019;116:187–192.
3. Domingo-Sananes MR, McInerney JO. Mechanisms that shape microbial pangenomes. *Trends Microbiol* 2021;29:493–503.
4. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, *et al*. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci U S A* 2009;106:17939–17944.
5. Hentschel U, Hacker J. Pathogenicity islands: The tip of the iceberg. *Microbes Infect* 2001;3:545–548.
6. Johnson TJ. Separate F-type plasmids have shaped the evolution of the H30 subclone of *Escherichia coli* sequence type 131. *mSphere* 2016;1:00121–16.
7. Nakamura K, Murase K, Sato MP, Toyoda A, Itoh T, *et al*. Differential dynamics and impacts of prophages and plasmids on the pangenome and virulence factor repertoires of Shiga toxin-producing *Escherichia coli* O145:H28. *Microb Genom* 2020;6.
8. Bruns H, Crüsemann M, Letzel A-C, Alanjary M, McInerney JO, *et al*. Function-related replacement of bacterial siderophore pathways. *ISME J* 2018;12:320–329.
9. 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.
10. Dobrindt U. (Patho-)Genomics of Escherichia coli. *Int J Med Microbiol* 2005;295:357–371.
11. Decano AG, Downing T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci Rep* 2019;9:17394.
12. Chen SL, Hung C-S, Xu J, Reigstad CS, Magrini V, *et al*. Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc Natl Acad Sci U S A* 2006;103:5977–5982.
13. Whelan FJ, Rusilowicz M, McInerney JO. Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics* 2020;6:e000338.
14. Seemann T. PROKKA: Rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
15. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, *et al*. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* 2020;21:180.
16. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
17. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
18. Abadi S, Azouri D, Pupko T, Mayrose I. Model selection may not be a mandatory step for phylogeny reconstruction. *Nat Commun* 2019;10:934.
19. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
20. Whelan FJ, Hall RJ, McInerney JO. Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Mol Biol Evol* 2021.
21. The UniProt Consortium. UNIPROT: The universal protein knowledge base. *Nucleic Acids Research* 2017;45.
22. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Tech Rep* 2000;1.
23. Caspi R, Altman T, Billington R, Dreher K, Foerster H, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2014;42:D459-71.
24. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, *et al*. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb Genom* 2018;4:e000224.
25. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al*. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.
26. Kawasaki Y, Wada C, Yura T. Roles of *Escherichia coli* heat shock proteins DnaK, DnaJ and GrpE in mini-F plasmid replication. *Molec Gen Genet* 1990;220:277–282.
27. Makino K, Ishii K, Yasunaga T, Hattori M, Yokoyama K, *et al*. Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an enterohemorrhagic *Escherichia coli* O157:H7 derived from Sakai outbreak. *DNA Res* 1998;5:1–9.
28. Schmidt H, Beutin L, Karch H. Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL 933. *Infect Immun* 1995;63:1055–1061.
29. Waldor MK. Bacteriophage biology and bacterial virulence. *Trends Microbiol* 1998;6:295–297.
30. Handa N, Kobayashi I. Type III Restriction Is Alleviated by Bacteriophage (RecE) Homologous Recombination Function but Enhanced by Bacterial (RecBCD) Function. *J Bacteriol* 2005;187:7362–7373.
31. Bachler C, Schneider P, Bahler P, Lustig A, Erni B. *Escherichia coli* dihydroxyacetone kinase controls gene expression by binding to transcription factor DhaR. *EMBO J* 2005;24:283–293.
32. Jenkins LS, Nunn WD. Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in *Escherichia coli*: The ATO system. *J Bacteriol* 1987;169:42–52.
33. Rome K, Borde C, Taher R, Cayron J, Lesterlin C, *et al*. The two-component system ZRAPSR is a novel ESR that contributes to intrinsic antibiotic tolerance in *Escherichia coli*. *J Mol Biol* 2018;430:4971–4985.

34. Francetic O, Belin D, Badaut C, Pugsley AP. Expression of the endogenous type II secretion pathway in *Escherichia coli* leads to chitinase secretion. *EMBO Journal* 2000;19:6697–6703.
35. Miki T, Okada N, Kim Y, Abe A, Danbara H. DsbA directs efficient expression of outer membrane secretin EscC of the enteropathogenic *Escherichia coli* type III secretion apparatus. *Microb Pathog* 2008;44:151–158.
36. Silver RP, Aaronson W, Vann WF. The K1 capsular polysaccharide of *Escherichia coli*. *Rev Infect Dis* 1988;10 Suppl 2:S282-6.
37. Postma PW, Lengeler JW. Phosphoenolpyruvate: Carbohydrate phosphotransferase system of bacteria. 1985.
38. Tchieu JH, Norris V, Edwards JS, Saier MH. The complete phosphotransferase system in *Escherichia coli*. *J Mol Microbiol Biotechnol* 2001;3:329–346.
39. Manges AR, Geum HM, Guo A, Edens TJ, Fibke CD, *et al*. Global extraintestinal pathogenic *Escherichia coli* (ExPEC) lineages. *Clin Microbiol Rev* 2019;32.
40. Sokurenko EV, Chesnokova V, Dykhuizen DE, Ofek I, Wu XR, *et al*. Pathogenic adaptation of *Escherichia coli* by natural variation of the FIMH adhesin. *Proc Natl Acad Sci U S A* 1998;95:8922–8926.
41. McNally A. Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency- dependent selection. *mBio* 2019;10:00644–19.
42. Rousset F. The impact of genetic diversity on gene essentiality within the *E. coli* species. *bioRxiv* 2020.
43. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. *Nature Microbiology* 2017;2:1–5.
44. Shapiro BJ. The population genetics of pangenomes. 2017.
45. Andreani NA, Hesse E, Vos M. Prokaryote genome fluidity is dependent on effective population size. *ISME Journal* 2017;11:1719–1721.
46. Lu Q-F, Cao D-M, Su L-L, Li S-B, Ye G-B, *et al*. Genus-wide comparative genomics analysis of Neisseria to identify new genes associated with pathogenicity and niche adaptation of Neisseria pathogens. *Int J Genomics* 2019;2019:6015730.
47. Didelot X, Wilson DJ. Clonalframeml: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 2015;11:e1004041.
48. Cusumano CK, Hung CS, Chen SL, Hultgren SJ. Virulence plasmid harbored by uropathogenic *Escherichia coli* functions in acute stages of pathogenesis. *Infect Immun* 2010;78:1457–1467.
49. Kuenne C. Reassessment of the Listeria monocytogenes pan-genome reveals dynamic integration hotspots and mobile genetic elements as major components of the accessory genome. *BMC Genomics* 2013;14:47.
50. Firth N, Skurray R. Characterization of the F plasmid bifunctional conjugation gene, traG. *MGG Molecular & General Genetics* 1992;232:145–153.
51. Wallden K, Rivera-Calzada A, Waksman G. Type IV secretion systems: Versatility and diversity in function. 2010.
52. Schmidt H, Henkel B, Karch H. A gene cluster closely related to type II secretion pathway operons of Gram-negative bacteria is located on the large plasmid of enterohemorrhagic *Escherichia coli* O157 strains. *FEMS Microbiology Letters* 2006;148:265–272.
53. Arciszewska L, Sherratt D. Xer site-specific recombination *in vitro*. *EMBO J* 1995;14:2112–2120.
54. Cornet F, Hallet B, Sherratt DJ. Xer recombination in *Escherichia coli*: Site-specific DNA topoisomerase activity of the XerC and XerD recombinases. *Journal of Biological Chemistry* 1997;272:21927–21931.
55. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2020;19:37–54.
56. Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, *et al*. The ecology and evolution of pangenomes. *Current Biology* 2019;29:R1094–R1103.
57. Zhang X. First identification of coexistence of blaNDM-1 and bla CMY42 among *Escherichia coli* ST167 clinical isolates. *BMC Microbiol* 2013;13:282.
58. Schaufler K. Genomic and functional analysis of emerging virulent and multidrug-resistant Escherichia coli lineage sequence type 648. *Antimicrob Agents Chemother* 2019;63.
59. Hibbing ME, Dodson KW, Kalas V, Chen SL, Hultgren SJ. Adaptation of arginine synthesis among uropathogenic branches of the *Escherichia coli* phylogeny reveals adjustment to the urinary tract habitat. *mBio* 2020;11.