5 Redefining the *E. coli* pan-genome reveals new patterns of gene gain/loss and gene sharing between lineages

5.1 Introduction

HGT is common in *E. coli* and is a major contributor to resistance and pathogenicity. *E. coli* has a high plasmid load, with many resistance genes present on these plasmids [4]. The virulence genes which are used as markers to identify the different *E. coli* pathotypes are also horizontally transmitted, either by plasmids, phage or other MGEs [102,221–223]. Additionally, recombination rates have been estimated to be high in *E. coli* [11,77,212,213]. All of the above emphasise the importance of HGT to the lifestyle and pathogenicity of *E. coli* (See Section 1.2.4 of Introduction for more details).

Genome size, plasmid load and recombination rates, along with rates of gene gain and loss, have all been shown to differ across *E. coli* lineages and phylogroups [4,77,92,212]. Indeed, there are differences in the distribution of the pathotypes across the phylogroups and it has been shown that particular genetic backgrounds are required for the acquisition of specific virulence factors [409,410]. Phylogroup F, B2 and D predominantly contain ExPEC isolates whereas phylogroups B1 and E predominantly contain diarrheal *E. coli* pathotypes (See Section 4.4.47, Figure 4.12). Phylogroup A contains isolates from the different *E. coli* pathotypes and has been termed a “generalist” phylogroup [411]. Concordantly, phylogroup A, as well as C, have been estimated to have high rates of HGT with high rates of gene gain/loss and high recombination rates [77,92,212,213]. Conversely, reduced recombination rates were estimated within the global MDR ExPEC lineage, ST131 of phylogroup B2 and the common EHEC lineage ST11/O157:H7 of phylogroup E, suggesting a clonal expansion of these lineages due to their clinical success [213].

These existing studies examining HGT in *E. coli* were mostly focused on high-level descriptions of the relationships between the phylogroups and have not looked at the resolution of specific *E. coli* lineages [92,212]. Even more, these studies have mostly considered only the core genome when estimating recombination rates [77,213], or otherwise, when measuring dynamics of gene gain/loss dynamics or gene sharing, have treated all genes of the gene pool equally [11,92,212]. These approaches are likely to mask particular signals

in the data. When considering only the core genome, the added information of the accessory genome, which represents the main fraction of the gene pool which undergoes HGT, is entirely missing in the analysis. When treating all genes equally in gene gain/loss or gene sharing calculations, rare events would be lost in the background. For instance, if 90% of genes are shared according to phylogenetic relatedness whereas only 10% are not, the signal for the unique 10% would not be observed when events are summed across the entire gene pool. Therefore, a higher resolution approach needs to be applied to understand the dynamics of different genes and how these dynamics differ across lineages.

In the previous chapter, a high-quality collection of *E. coli* genomes was built and the lineages of the collection, termed PopPUNK Clusters, were defined and characterised for their resistance and pathogenic profiles. The described dataset is novel in its resolution as it includes 47 well-characterised lineages (PopPUNK Clusters) with multiple representatives, and the frequency of each gene of the gene pool within each PopPUNK Cluster is known. This dataset provides the ability to identify different types of genes in the *E. coli* gene pool based on their distribution across the 47 lineages, and to unravel the differences between these lineages.

5.2 Aims

The work presented in this Chapter is a novel approach to classifying and analysing the patterns of gene sharing and gene gain and loss in the collection of 7,500 *E. coli* isolates presented in Chapter 4. The specific aims of this chapter were:

- Define a novel approach for describing the *E. coli* pan-genome.
- Unravel the properties of genes from the newly defined gene-classes in terms of their function and dynamics of gain and loss.
- Understand the differences between the PopPUNK Clusters in terms of their gene content and the levels of gene sharing between them.

5.3 Methods

5.3.1 Gene classification into “occurrence classes”

The genes were classified into “occurrence classes” based on their distribution patterns in the dataset. Each gene was assigned to an occurrence class based on its frequency within genomes belonging to the same phylogenetic clusters, termed PopPUNK Clusters. Within each PopPUNK Cluster, a gene was defined as “core” if it was present in more than 95% of

the isolates of that cluster, “intermediate” if present in 15% to 95% of isolates of the cluster, and “rare” if present in up to 15% of the isolates of the cluster. Three main occurrence classes, “Core”, “Intermediate” and “Rare”, contained all the genes that were always observed as being “core”, “intermediate” or “rare” respectively across all PopPUNK Clusters in which they were present. However, within these four occurrence classes, whilst the frequency was maintained within a cluster, genes were seen to be “core”, “intermediate” or “rare” across different numbers of clusters. Hence, to capture the distribution of all genes these occurrence classes were further subdivided into a total of eleven subclasses based on the number of PopPUNK Clusters in which a gene was observed and the frequency of that gene within those clusters (Figure 5.1).

“Dataset core” genes were present and “core” in all PopPUNK Clusters. “Multi-cluster core”, “multi-cluster intermediate” and “multi-cluster rare” genes were present in multiple PopPUNK Clusters in their respective frequencies. “Cluster specific core”, “Cluster specific intermediate” and “Cluster specific rare” genes were present only in one PopPUNK Cluster in their respective frequencies. The final main occurrence class “Varied” included all the genes which were observed as either combination of “core”, “intermediate” or “rare” across multiple PopPUNK Clusters. These combinations were “core, intermediate and rare”, “core and intermediate”, “core and rare” and “intermediate and rare” (Figure 5.1).

5.3.2 Measuring the genetic composition of each PopPUNK Cluster

The number of genes from each of the eleven occurrence classes was counted in each of the 7,693 *E. coli* genomes remaining in the collection described in Section 4.4.6. The mean number of genes and the standard deviation of the number of genes from each occurrence class was calculated per PopPUNK Cluster using built-in R functions. To measure the genetic composition of a typical *E. coli* genome within our dataset, the mean and standard deviations were calculated on the mean counts of all the 47 PopPUNK Clusters.

5.3.3 Phylogenetic analysis

5.3.3.1 Phylogenetic tree construction

A representative sequence from all 47 PopPUNK Clusters was chosen using Treemmer [392]. Treemer greedily prunes leaves off the phylogeny by choosing a random leaf from the two most closely related pairs of leaves in every iteration, until the desired number of leaves in the tree is reached. The core gene alignment of the 47 selected isolates was generated using

Roary [305], and a tree from the SNPs, taken using SNP-sites [332] (v2.3.2), was constructed using RaXML (v8.2.8) using a GTR+gamma model with 100 bootstrap replicates [282].

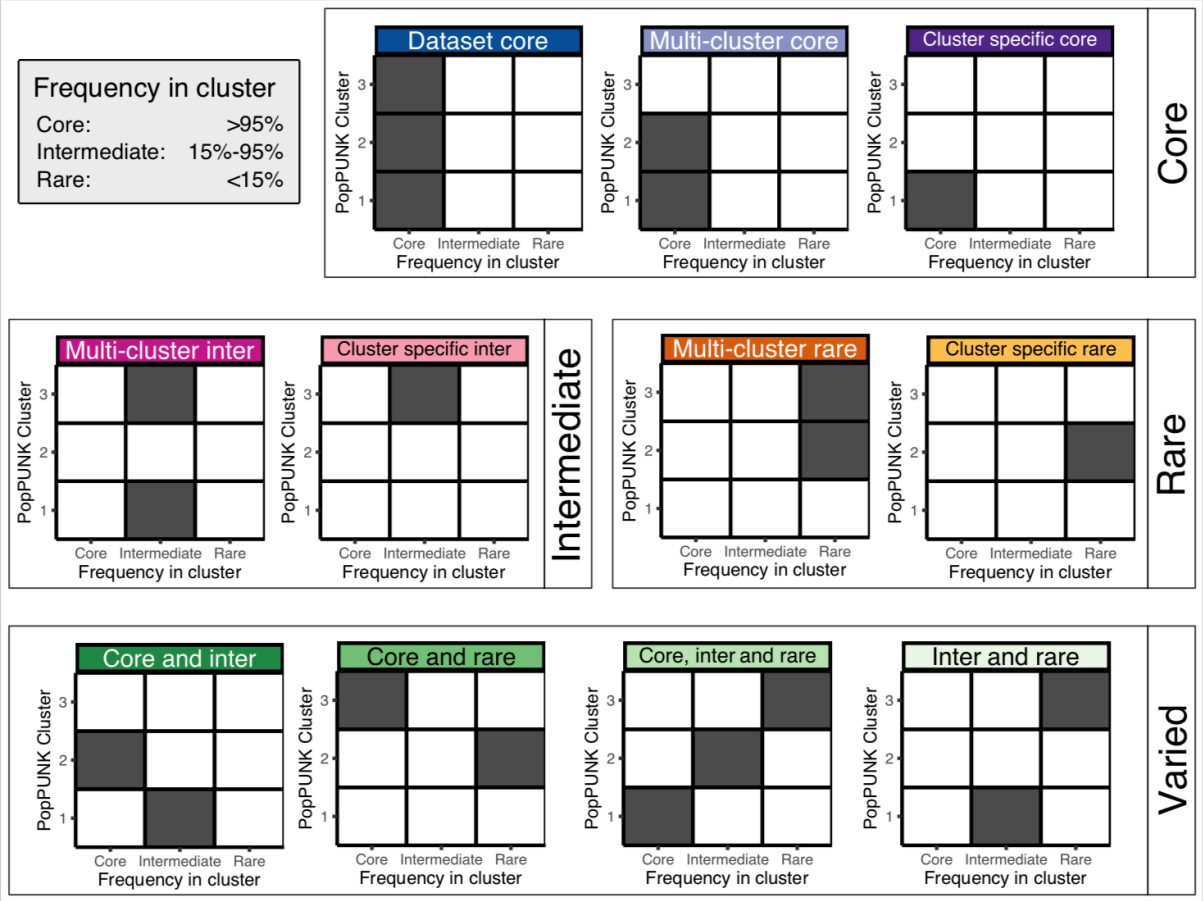

Figure 5.1: Gene classification into occurrence classes*. The figure presents a hypothetical example of comparing a total of 3 PopPUNK Clusters written on the y-axis. The x-axis represents the frequency of the gene in each of the three clusters being compared. A gene is considered "core" in a cluster if it was present in >95% of isolates of the cluster, "intermediate" if it was present in 15%-95% of the the isolates of the cluster, and "rare" if present in <15% of isolates of the cluster. Each panel is an example of a gene from the given occurrence class. A dark square indicates the gene is present in the cluster and the frequency of that gene in the cluster. As there are three clusters, each gene can be observed in any combination of frequencies across the three clusters. "Core" genes were observed in core frequencies in all (dataset core), some (multi-cluster core) or one (cluster specific core) cluster. "Intermediate" genes were observed in intermediate frequencies in some (multi-cluster intermediate) or one (cluster specific intermediate) cluster. "Rare" genes were observed in rare frequencies in some (multi-cluster rare) or one (cluster specific rare) cluster. "Varied" genes were observed in different frequencies across multiple clusters. For instance, the "Core and intermediate" gene*

presented is core in cluster 2 and rare in cluster 1 (and absent in 3). The "Core and rare" gene is core in cluster 3 and rare in cluster 2 (and absent in 1) etc.

5.3.3.2 Phylogenetic distance calculations

The phylogenetic distance between every two PopPUNK Clusters was measured as the patristic distance using the function 'cophenetic' from the R package APE (v5.3) [395]. The patristic distance is the sum of the total distance between two leaves of the tree, which represent the PopPUNK Clusters in this thesis, and hence summarises the total genetic change in the core gene alignment represented in the tree.

5.3.3.3 Ancestral state reconstruction

The leaves or tips of the phylogenetic tree constructed in Section 5.3.3.1 represent the 47 PopPUNK Clusters. Presence of a gene in a PopPUNK Cluster (tree leaf) was defined as the gene being observed at least once in at least one isolate of the PopPUNK Cluster. The presence or absence of a gene in an ancestral node, i.e. an internal node, was determined using accelerated transformation (ACCTRAN) reconstruction implemented in R [412]. ACCTRAN is a maximum parsimony-based approach which minimises the number of transition events on the tree (from absence to presence and vice versa) while preferring changes along tree branches closer to the root of the tree.

5.3.3.4 Counting gain and loss events

Gain and loss events were counted based on the results of the ancestral state reconstruction. If there was a change from absence to presence from an ancestor to a child along a branch in the phylogeny, a gain event was counted. If there was a change from presence to absence a loss event was counted. The total number of gain and loss events was counted for each gene as well as on each branch for all occurrence classes.

5.3.4 Functional assignment of COG categories

The predicted function and COG category of each gene cluster were assigned using eggNOG-mapper (1.0.3) on the representative sequence of each of the gene clusters [413]. Diamond was used for a fast-local protein alignment of the representative sequences against the eggNOG protein database (implemented within eggNOG-mapper). The COG (Clusters of Orthologous Groups) classification scheme comprises 22 COG categories which are broadly divided into functions relating to cellular processes and signaling, information storage and processing, metabolism and genes which are poorly categorised [414]. When no match was found in the eggNOG database, the genes were marked as "?" in their COG category.

Sub-sentences of all lengths were extracted from each of the functional predictions for each gene cluster using the function “combinations” from the python package “itertools”, while ignoring common words. For instance, for the functional prediction “atp-binding component of a transport system”, the words “of”, “a” and “system” were ignored, and the extracted sub-sentences were “atp-binding component”, “atp-binding component transport” and “component transport”. The number of times each sub-sentence appeared in each occurrence class was counted. Overlapping sub-sentences which only had a difference of 3 or smaller in their total counts per occurrence class were merged in the final count to include only the longer sub-sentence. For instance, if “atp-binding component transport” was counted 100 times and “atp-binding component” was counted 103 times, the final count would only include the longer sub-sentence “atp-binding component transport” with a count of 100.

5.3.5 Identifying gene variants

The function makeblastdb from the Blast+ package (v2.9) was used to construct a database from the 50,039 genes of the *E. coli* pan-genome taken from Chapter 4 of this thesis [285,321]. Blastp was used to apply a pairwise all-against-all comparison of all the protein sequences. If two proteins shared more than 95% sequence identity over 95% of the total length of the shorter sequence, they were considered “partner genes”, with one being the “shorter variant” and the other the “longer variant”.

5.3.6 Gene property calculations

The length of each gene cluster was calculated as the mean length of all the members of that gene cluster. The GC content was calculated using Biopython (v1.72) on all the members of a gene cluster and the mean was taken as the value for that gene cluster. The fraction of members in a gene cluster that had ATG as their start codon was measured as the “ATG fraction”. If an alternative start codon was used in more than 50% of the members of a gene cluster then that cluster was considered as starting with an alternative start codon. The contig length was calculated for all the members of a gene cluster and the mean was calculated across all members.

5.3.7 Statistical analysis

Statistical analyses were performed in R (v3.3+). Ape (v5.3) [395] and ggtree (v1.16.6) [396] were used for phylogenetic analysis and visualisation. ggplot2 was used for all plotting [360].

5.4 Results

5.4.1 A novel approach for examining the *E. coli* pan-genome

In a standard pan-genome analysis, genes are classified into four categories: core, soft-core, intermediate and rare. These definitions are based on the frequency of the genes in the dataset. For instance, the default settings in Roary are that genes found in over 99% of the genomes are “core”, between 95% and 99% “soft-core”, between 15% and 95% “intermediate” and fewer than 15% “rare” [305]. In Section 4.4.4.9 of this thesis, these definitions were used to describe the pan-genomes of each of the 47 PopPUNK Clusters individually. Roary was originally designed for a pan-genome analysis of a single *Salmonella enterica* serovar (Typhi), thus the default approach used in Section 4.4.4.9 was valid for a pan-genome analysis on each PopPUNK Cluster which represents a group of closely related isolates. When expanding the pan-genome analysis to examine the pan-genome of an entire species, which in this case includes 47 different PopPUNK Clusters, new definitions needed to be established. Hence, a new set of rules was defined to classify the genes into four broad “occurrence classes”: “core”, “intermediate”, “rare” and “varied” genes. These four occurrence classes could be further subdivided into eleven sub-classes as detailed below. These definitions were based on the number of PopPUNK Clusters in which a gene was present (1 to 47), and the frequency of the gene in the clusters in which it was present (Figure 5.1).

Core genes were always observed in high frequencies (>95%) in one or multiple PopPUNK Clusters (Figure 5.1). These genes represented 9% (4,998/50,039) of the *E. coli* pan-genome (Figure 5.2A). Core genes included 1,426 genes (3% of all genes) which are the “dataset core” genes as they were core in all 47 of 47 PopPUNK Clusters (Figure 5.2B,C, 5.1). On the other side of the spectrum, there were 2,040 genes (4% of all genes) which were “cluster specific core” genes as they were core in a single PopPUNK Cluster. A set of 1,532 genes (3% of all genes) were defined as “multi-cluster core” as they were core to a subset of the PopPUNK Clusters (2-45 PopPUNK Clusters).

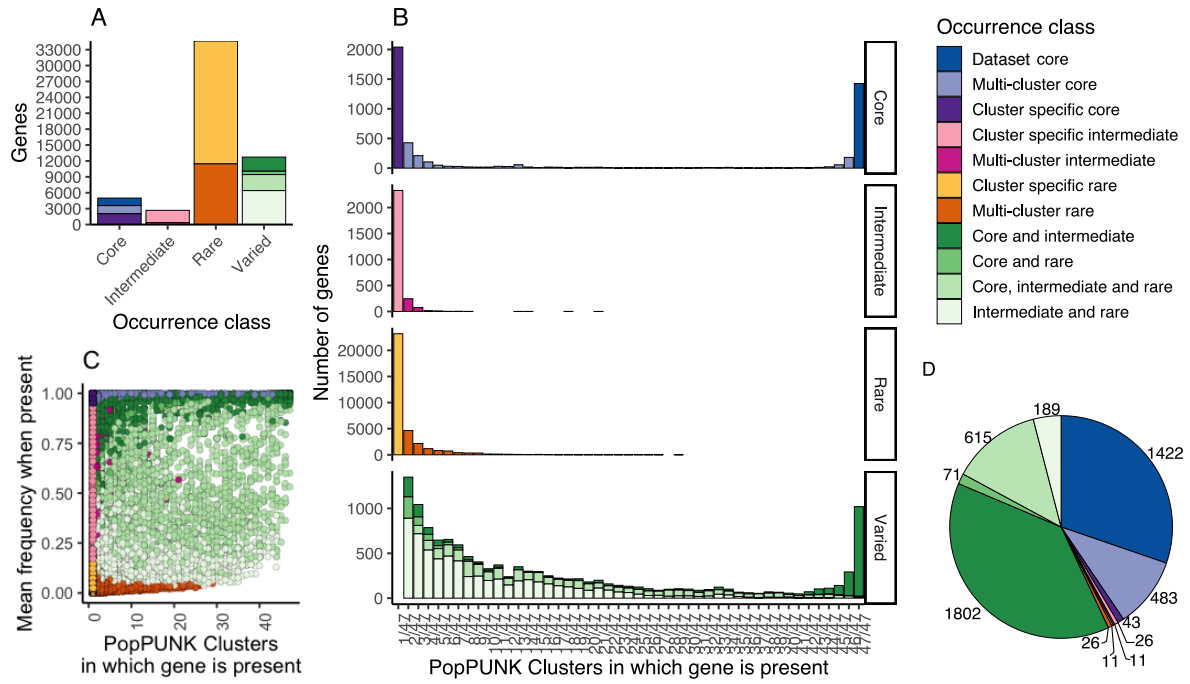

Figure 5.2: Distribution of the* E. coli *gene-pool based on the rules defined. A *Number of genes from each of the occurrence classes.* ***B*** *Distribution of the number of genes in each occurrence class relative to the number of PopPUNK Clusters in which they were found.* ***C*** *Mean frequency of each gene in the PopPUNK Clusters in which it was observed, plotted against the number of PopPUNK Clusters it was observed in, coloured by occurrence class.* ***D*** *The relative abundance and count of genes from each of the occurrence classes in a single representative* E. coli *genome in our dataset.*

Intermediate genes, representing 5% of all genes, were always observed in intermediate frequencies (15%-95%) in one or multiple PopPUNK Clusters (Figure 5.1, 5.2A). 87% of these genes (2,329/2,685) were only observed in a single PopPUNK Cluster and were termed “cluster-specific intermediate” genes (Figure 5.2B,C). The remaining intermediate frequency genes (356) were termed “multi-cluster intermediate”. These were mostly shared between a maximum of five PopPUNK Clusters (97%, 346/356) and their mean frequency within those clusters ranged from 16% to 94% of isolates, representing the full range of possible frequencies for intermediate genes. There were four genes (1%, 4/356) which were observed in intermediate frequencies in more the 10 PopPUNK Clusters. One gene was of particular interest as it was observed in 20 PopPUNK Clusters and its mean frequency across the these clusters was 0.57, appearing be a truly intermediate frequency gene (Figure 5.2C). A closer examination of the precise frequencies in which this gene was observed across the 20 clusters confirmed that it was indeed observed in 30-70% of isolates in all the clusters, with most PopPUNK Clusters having 50-60% of isolates possessing this gene. Further analysis on the

sequence of this gene revealed that this is a a short protein, only 53 aa long, which could not been assigned to any known function using functional annotation tools.

Rares genes were always observed in low frequencies (<15%) in one or multiple PopPUNK Clusters (Figure 5.1). This occurrence class represented the largest fraction of the entire gene pool consisting of a total of 34,624 genes, representing 63% (34,624/55,039) of the entire gene pool (Figure 5.2A). Of these, 67% were “cluster specific rare” genes (23,175/34,624) as they were observed only in a single PopPUNK Cluster (Figure 5.2B,C). The remaining “rare” genes were observed in multiple PopPUNK Clusters, termed “multi-cluster rare”. 76% (8,800/11,449) of these were observed in five PopPUNK Clusters or fewer. There were 651 (5%) genes which were observed in rare frequencies across 10 PopPUNK Clusters or more, i.e. rare genes across multiple PopPUNK clusters were more common than intermediate genes across multiple clusters.

Varied genes were observed in different frequencies across multiple PopPUNK Clusters (Figure 5.1). These genes represented 23% of the gene pool (12,732/55,039) (Figure 5.2A). These were further divided depending on the precise combination of frequencies in which they were found: “Core and intermediate”, “Core, intermediate and rare”, “Core and rare” or “Intermediate and rare” (Figure 5.1). Varied genes which were observed in more PopPUNK Clusters were more commonly observed in higher frequencies within those clusters and thus belonged to the group of “Core and intermediate" genes (Figure 5.2B,C). On the other hand, varied genes which were observed in fewer PopPUNK Clusters were more commonly observed in low frequencies within those clusters and thus belonged to the group of “Intermediate and rare" varied genes (Figure 5.2B,C).

5.4.2 The typical composition of an *E. coli* genome

A typical *E. coli* genome contained 1,422±4 genes (~30%) “dataset core” genes (core across the entire dataset) (Figure 5.2D; see Section 5.3.2). There were 483±66 (~10%) “multi-cluster core” genes which were core to a subset of the population and 43±55 (1-2%) genes which were “cluster specific core” genes, present and core only in a single PopPUNK Cluster (Figure 5.2D). A typical genome also contained 11±7 (~0.3%) “multi-cluster intermediate” and 26±23 (0.5-1%) “cluster specific intermediate” genes (Figure 5.2D). Similarly, there were 26±11 (~0.5%) “multi-cluster rare” genes 5.2D) and 11±9 (~0.3%) “cluster specific rare” genes in each genome (Figure 5.2D). Although the “rare” and “intermediate” genes made up more than 60% of the entire gene pool (34,543/55,039), they each represented fewer than 1% of the genes within a single isolate (Figure 5.2D). The “varied” genes represented approximately

60% of all the genes in a typical *E. coli* genome (Figure 5.2D). Most of these were “core and intermediate” genes (1802±87, ~40%) (Figure 5.2D). Additionally, each genome contained 71±13 (1-2%) “core and rare” genes, 614±116 (10-15%) “core, intermediate and rare” genes, and 189±51 (3-5%) “intermediate and rare” genes.

5.4.3 Rates of gene gain and loss differ across the occurrence classes

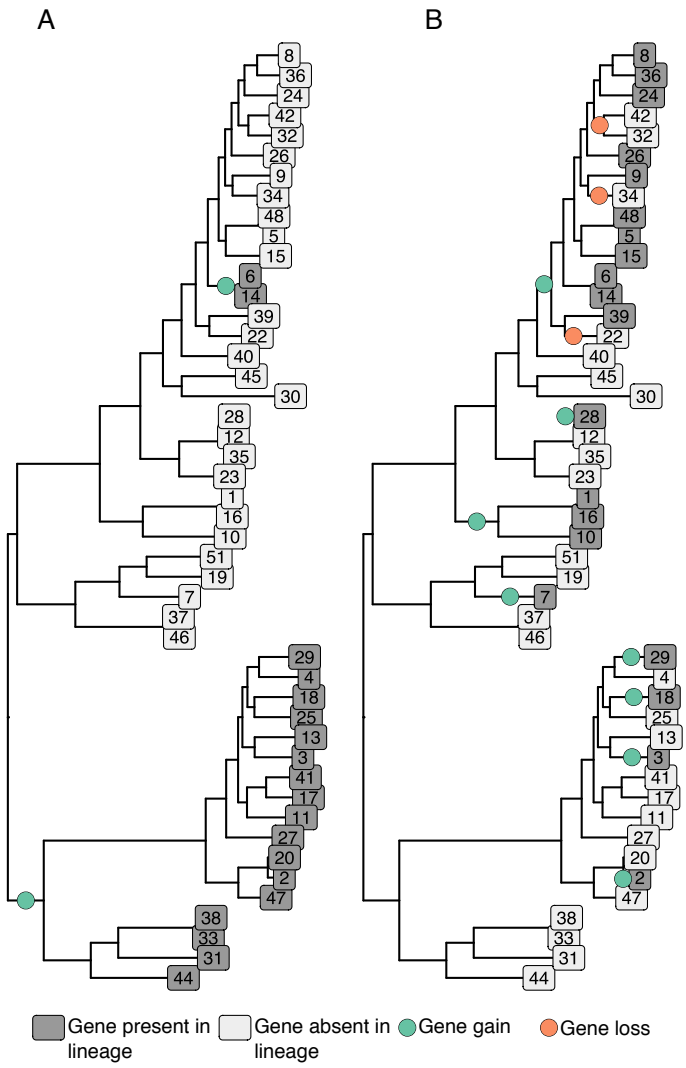

Figure 5.3: Example of the distribution patterns of two genes, along with the number of gain and loss events required to explain their distribution across the tree tips. *A gene is defined as present in a tip (dark grey) if at least one genome of the lineage had the gene. Gain (green circle) and loss (red circle) were estimated using ancestral state reconstruction.* ***A*** *A “multi-cluster core” gene which is associated with two clades and required only 2 gain events to explain its distribution.* ***B*** *An “intermediate and rare” gene which was not clade associated required 8 gain and 3 loss events to explain its distribution along the tree tips.*

The presence and absence patterns of genes which were present across multiple PopPUNK Clusters were used to count the number of gain and loss events estimated to have occurred along the tree branches. This was achieved using a parsimony-based ancestral state reconstruction approach to infer the minimum number of gain and loss events required to explain the distribution of a gene on the tree tips. (See Sections 5.3.3.3-4). For instance, if a gene was present in only two clades (regardless of its frequency when present), its distribution along the tree tips could be explained by two gain events on two branches (Figure 5.3A). If a gene was distributed across

the tree tips with no clear pattern, many more gain or loss events were required to explain its distribution on the tree tips (Figure 5.3B).

The number of gain and loss events which occurred for each gene varied across the occurrence classes (Figure 5.4A). For comparison, the specific combinations of gain and loss events across all genes for each of the occurrence classes are summarised in Figure 5.4B-H. These will be referred to in the following sections. Note that due to the method by which genes from the different PopPUNK Clusters were grouped as described in Section 4.3.8 of this thesis, gene loss could indicate either complete loss, truncation by more than 20% of the gene length or diversification beyond the 95% sequence identity threshold used to group genes together.

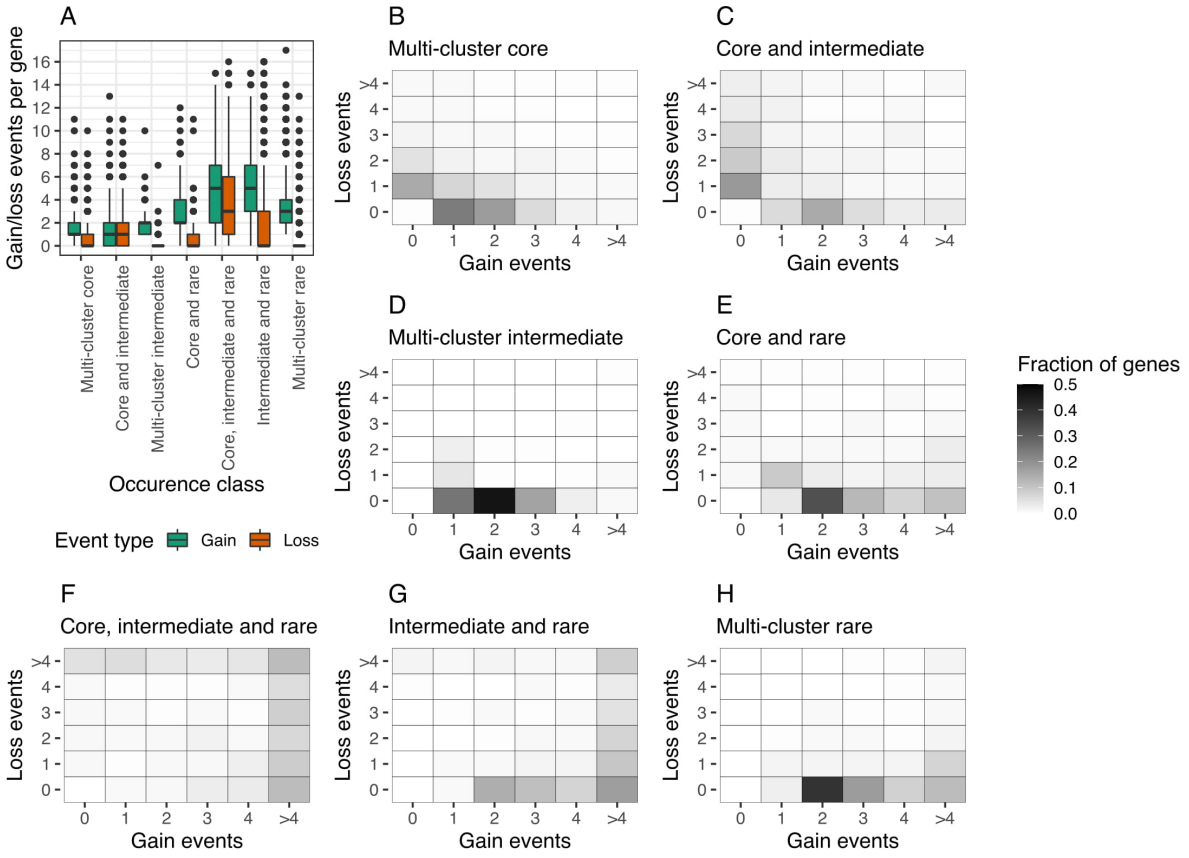

Figure 5.4: Gain and loss events per gene. A *Number of gain and loss events per gene stratified by occurrence class.* ***B-H*** *Fraction of genes which have undergone specific combinations of gain and loss events for each occurrence class. The shade of each square indicates the fraction of genes from the occurrence class which have undergone the specific combination of gain and loss events.*

5.4.4 “Multi-cluster core” genes represent the shifts in core genome of *E. coli* clades

The median number of gain and loss events estimated for “multi-cluster core” genes was a single gain event and a no loss events (Figure 5.4A). The majority (68%) of the presence and absence patterns of these genes could be explained by up to two gain or loss events along the tree branches (Figure 5.4B). Most prominently, a single gain event and no loss events was observed for 24% of these genes, i.e. these genes were gained in a single point in time and were fixed within the lineages downstream from the point of introduction. On the other hand, 15% of these genes were estimated to have been lost in a single event that led to the absence of the gene from a subset of the PopPUNK Clusters. While some genes were estimated to have been gained and lost on more occasions, these were the exception rather than the rule for this occurrence class (Figure 5.4A,B).

The number of gain and loss events predicted to have occurred on each branch were counted (Figure 5.5A,B). Gain and loss events of “multi-cluster core” genes most commonly occurred along the internal branches which define the phylogroups (Figure 5.5A, B, C). A large number of gain events occurred on the branches leading to phylogroups B2 (104 gain events), E (66), F (52) and two clades of phylogroup D (90 and 64) (Figure 5.5A,C). Two PopPUNK Clusters within phylogroup E, Clusters 1 and 16, were also estimated to undergo a large number of gene gain events (97). The branches leading to the clades of phylogroups A, *Shigella*, B1 and C were not estimated to have undergone a large number of gene gain events. Phylogroup B2 was the only phylogroup which had undergone excessive gene loss in addition to gene gain (52 loss events) (Figure 5.5B,C). Otherwise, gene loss occurred most commonly along the tree tips (Figure 5.5C). Most prominently, PopPUNK Clusters 30 and 45 which represent *S. sonnei* and *S. flexneri* respectively, as well as PopPUNK Cluster 18 which has not been assigned to any of the phylogroups, have undergone the largest number of recent loss events (90, 52 and 65) (Figure 5.5B, D).

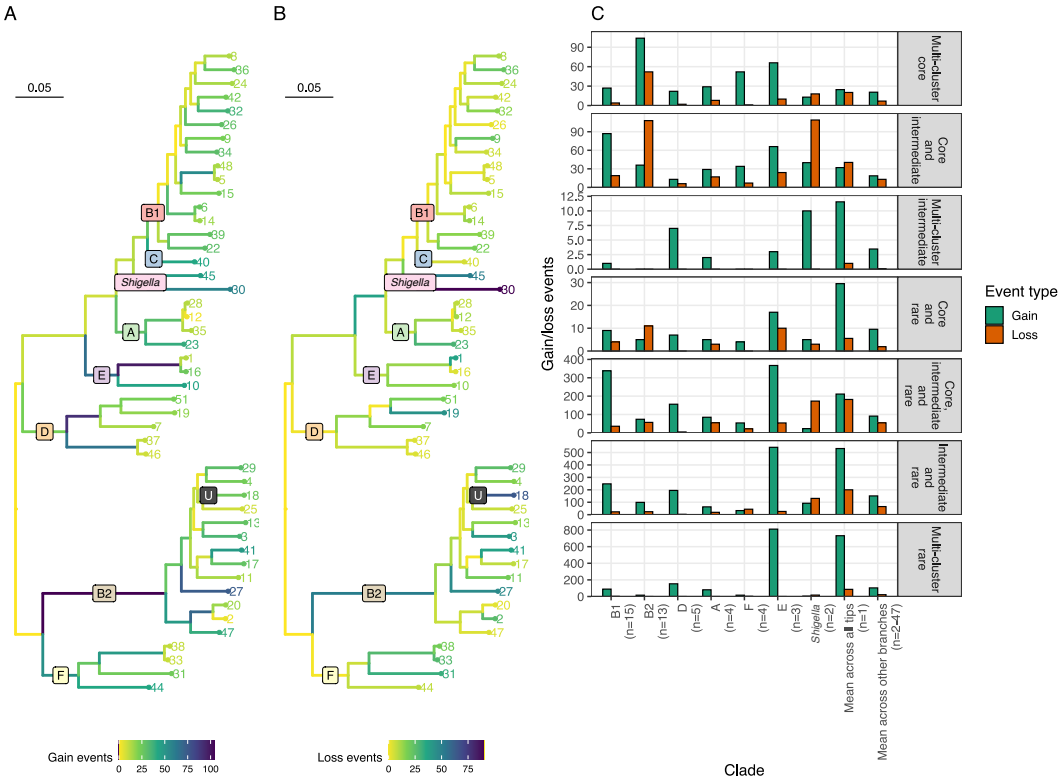

Figure 5.5: Gain and loss events per branch. A,B *Example for "multi-cluster core" genes on the precise counts of "gain" and "loss" events across all genes of this occurrence class predicted to have occurred on each branch. Darker branches indicate a larger number of events occurring on the branch.* ***C*** *Summary of the total number of gain and loss events on key branches for all the occurrence classes. The top panel for the "multi-cluster core" genes summarises panels* ***A*** *and* ***B****.*

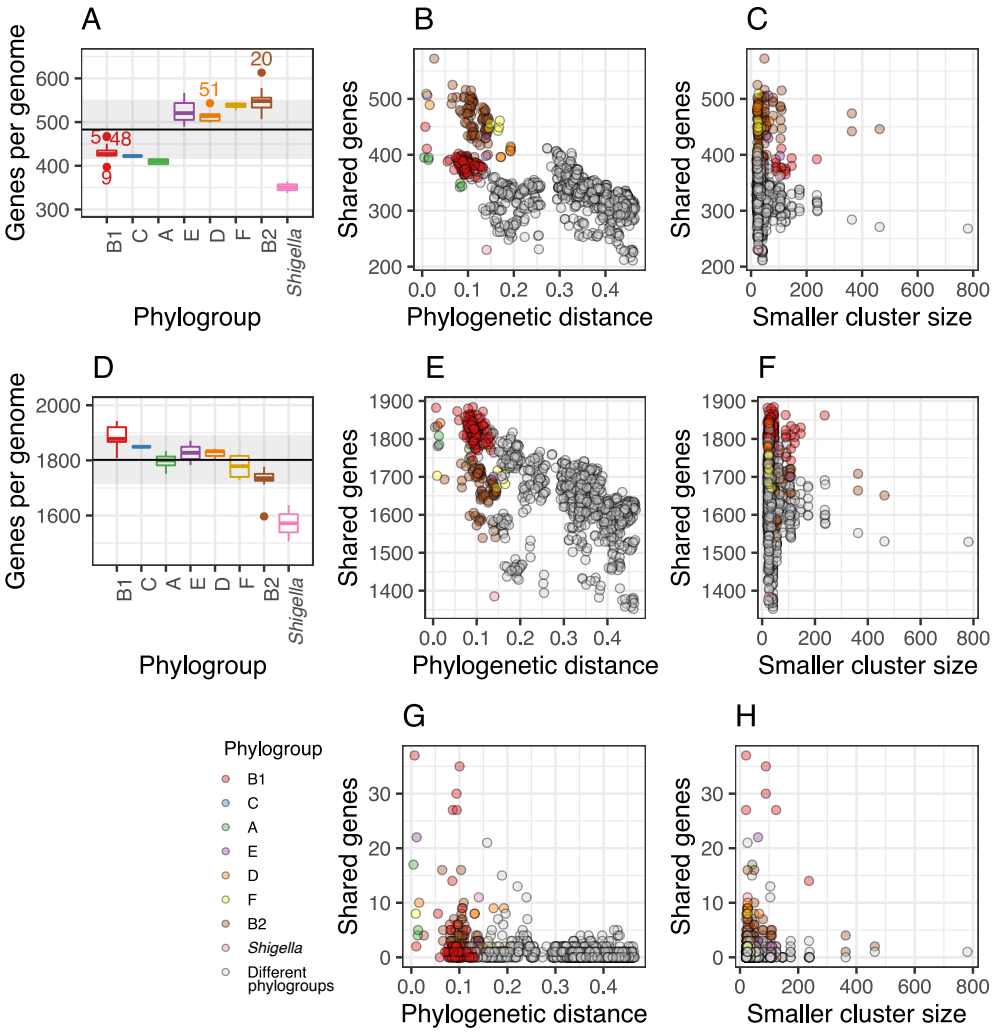

Figure 5.6: Properties of high frequency genes in the* E. coli *dataset. A,D *Number of "multi-cluster core" genes (**A**) and "core and intermediate" genes (**D**) per genome in each of the 47 PopPUNK Clusters, grouped by phylogroup.* ***B, E, G*** *Relationship between the number of genes shared between every two PopPUNK Clusters and phylogenetic distance between them for "multi-cluster core" genes (**B**), "core and intermediate" genes (**E**) and "multi-cluster intermediate" genes (**G**). Coloured dots indicate that the two PopPUNK Clusters being compared are from the same phylogroup, whereas gray dots indicate that the two clusters being compared are from different phylogroups.* ***C, F, H*** *Relationship between the number of genes shared between every two PopPUNK Clusters and the size of the smaller PopPUNK Cluster of the two being compared for "multi-cluster core" genes (**C**), "core and intermediate" genes (**F**) and "multi-cluster intermediate" genes (**H**).*

In agreement with the above, while the mean number of "multi-cluster core" genes was 483 per genome, isolates belonging to phylogroups B1, C, and A tended to have ~400 multi-cluster core genes per genome compared with ~500 for those belonging to phylogroups E, D, F and

B2 (Figure 5.6A). PopPUNK Clusters of *Shigella* spp. had the fewest number of "multi-cluster core" genes per genome with ~350 multi-cluster core genes per genome (Figure 5.6A).

The above analysis suggests that "multi-cluster core" genes represent the changes in the core genome between the clades. Accordingly, the number of "multi-cluster core" genes shared between every two PopPUNK Clusters was correlated negatively with the phylogenetic distance between them (linear regression, R^2=0.42, p<2e-16), i.e. two PopPUNK Clusters which were close phylogenetically shared more "multi-cluster core" genes (Figure 5.6B). There was no connection between the size of the two PopPUNK Clusters being compared and the number of "multi-cluster core" genes they shared (linear regression, R^2=0, p=0.51) (Figure 5.6C).

5.4.5 "Core and intermediate" represent the "soft-core" genome

The properties of the "core and intermediate" genes, which represented 40% of the genes in a single *E. coli* genome and 5% of the entire gene pool (Figure 5.2A,D), prove that these genes present similar distribution patterns, patterns of gain and loss and predicted functions to the defined "multi-cluster core" and "dataset core" genes.

59% of these genes (1,566/2,674) were observed in 40 PopPUNK Clusters or more, and in high frequencies within those clusters (Figure 5.2B,C). In fact, 37% of the "core and intermediate" genes were ubiquitous, i.e. they were present in 47 of 47 PopPUNK Clusters (Figure 5.2B). The median number of gain and loss events occurring per gene for "Core and intermediate" genes was a single gain event and a single loss event (Figure 5.4A). Similar to the "multi-cluster core" genes, 51% of the presence and absence patterns of these genes could be explained by up to two gain and loss events (Figure 5.4C). Gain events of "core and intermediate" genes were largest for the branches leading to phylogroups B1 and E (87 and 66) (Figure 5.5C). Rates of gene loss were generally higher in this occurrence class compared with the "multi-cluster core genes". Similarly, loss events predominantly occurred within *Shigella* and phylogroup B2 (Figure 5.5C). Indeed, these phylogenetic clusters had the lowest number of "core and intermediate" genes per genome relative to the other phylogroups whereas PopPUNK Clusters of Phylogroup B1 had the highest number of these genes per genome (Figure 5.6D).

"Core and intermediate" genes were also more commonly shared between closely related isolates (linear regression, R^2=0.39, p<2e-16) (Figure 5.6E), and there was no connection between the size of the two PopPUNK Clusters being compared and the number of core and

intermediate genes shared between them (linear regression, R^2=0.0007, p=0.18) (Figure 5.6F).

The distribution of predicted functions of this set of genes was similar to the predicted functions of the “dataset core” genes (Figure 5.7). COG categories were assigned to all the genes with eggNOG-mapper on the representative protein sequence of each gene cluster [413,414] (See Section 5.3.6). 34% of the “core and intermediate” genes were assigned to be involved in metabolism, similar to 40% of the “dataset core” genes. 14% and 13% were predicted to be involved in “information storage and processing” and “cellular processes and signalling” relative to 19% and 20% of the “dataset core” genes. Even more, the relative abundance of the specific COG categories was similar between the “dataset core” and the “core and intermediate” genes (Figure 5.7).

5.4.6 “Multi-cluster intermediate” genes are shared between closely related PopPUNK Clusters, but have different functional profiles to the “core” genes

In 89% of cases, “multi-cluster intermediate” genes were gained in 1-3 events and not lost (Figure 5.4D, 5.4C). Additionally, above a certain phylogenetic distance, the number of “multi-cluster intermediate” genes shared between every two PopPUNK Clusters drops to zero, meaning that these genes were only shared between closely related isolates (Figure 5.6G). Shared “multi-cluster intermediate” genes were only observed within PopPUNK Clusters which had fewer than 200 isolates (Figure 5.6H). These findings together suggest that these genes are confined to a phylogenetically close subset of the population, yet were gained multiple times within this subset. Unlike the “core and intermediate” genes, 77% “multi-cluster intermediate” genes were assigned a category of “poorly characterised” in their function prediction, and fewer than 1% were predicted to have a function related to cell metabolism (Figure 5.7). While these genes are shared between closely related PopPUNK Clusters as was observed for the “multi-cluster core” and the “core and intermediate” genes, they evidently differ in their functional profiles.

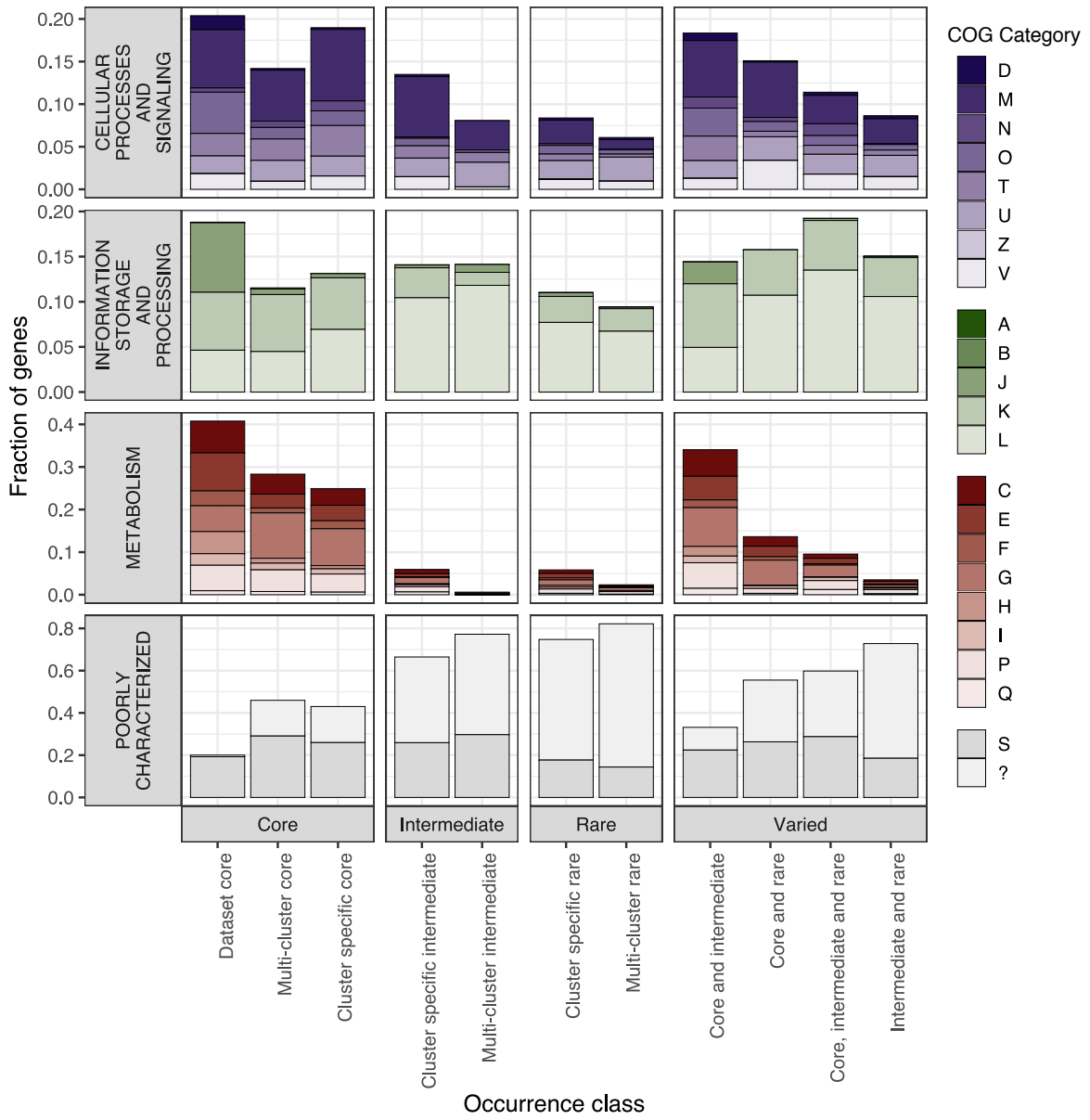

Figure 5.7: Fraction of genes from each occurrence class which were assigned each of the COG categories. *D (Cell cycle control, cell division, chromosome partitioning), M (Cell wall/membrane/envelope biogenesis), N (Cell motility), O (Post-translational modification, protein turnover, and chaperones), T (Signal transduction mechanisms), U (Intracellular trafficking, secretion, and vesicular transport), Z (Cytoskeleton), V (Defense mechanisms), A (RNA processing and modification), B (Chromatin structure and dynamics), J (Translation, ribosomal structure and biogenesis), K (Transcription), L (Replication, recombination and repair), C (Energy production and conversion), E (Amino acid transport and metabolism), F (Nucleotide transport and metabolism), G (Carbohydrate transport and metabolism), H (Coenzyme transport and metabolism), I (Lipid transport and metabolism), P (Inorganic ion transport and metabolism), Q (Secondary metabolites biosynthesis, transport, and catabolism), S (Function unknown) and "?" (unassigned).*

5.4.7 Low frequency genes are gained and lost at high rates, and their sharing is independent of the phylogeny

Shared low frequency genes include “multi-cluster rare”, “intermediate and rare” and “core, intermediate and rare” and “core and rare” genes as these were most commonly found in a small number of PopPUNK Clusters and in a low frequency within those clusters (Figure 5.2B,C). Unlike their high frequency counter-parts (“multi-cluster core”, “multi-cluster intermediate” and “core and intermediate” genes), the estimated number of gain and loss events predicted to have occurred for these occurrence classes was often estimated to be as high as four events and more (Figure 5.4A,E-H). “Multi-cluster rare” genes were not commonly lost as they were generally observed across a smaller number of PopPUNK Clusters and hence were mostly commonly gained 2-3 times along the tree tips (Figure 5.2B, 5.3H, 5.4C). Gain events of low frequency genes mostly occurred recently along the tree tips (Figure 5.5C). Phylogroup E was an exception which presented a large number of acquisition events of low frequency genes. A large number of gain events of “Core, intermediate and rare” genes were predicted to have occurred on the branch leading to Phylogroup B1.

The number of genes shared between every two PopPUNK Clusters for the “multi-cluster rare”, “intermediate and rare” and “core, intermediate and rare” genes did not correlate with the phylogenetic distance between the clusters (linear regression, R^2<0.03, Figure 5.8A-C). On the other hand, the number of shared genes was positively correlated with the size of the two PopPUNK Clusters being compared, with larger clusters sharing more genes (linear regression, R^2=[0.566,0.349,0.22], p<2.2e-16) (Figure 5.8D-F). This is because more genomes need to be sampled in order for the same low frequency gene to be observed in two PopPUNK Clusters. However, the number of genes shared plateaued after a particular PopPUNK Cluster size (Figure 5.8D-F). This number was smaller when the PopPUNK Clusters were from two different phylogroups, compared to when they were from the same phylogroup (Figure 5.8D-F).

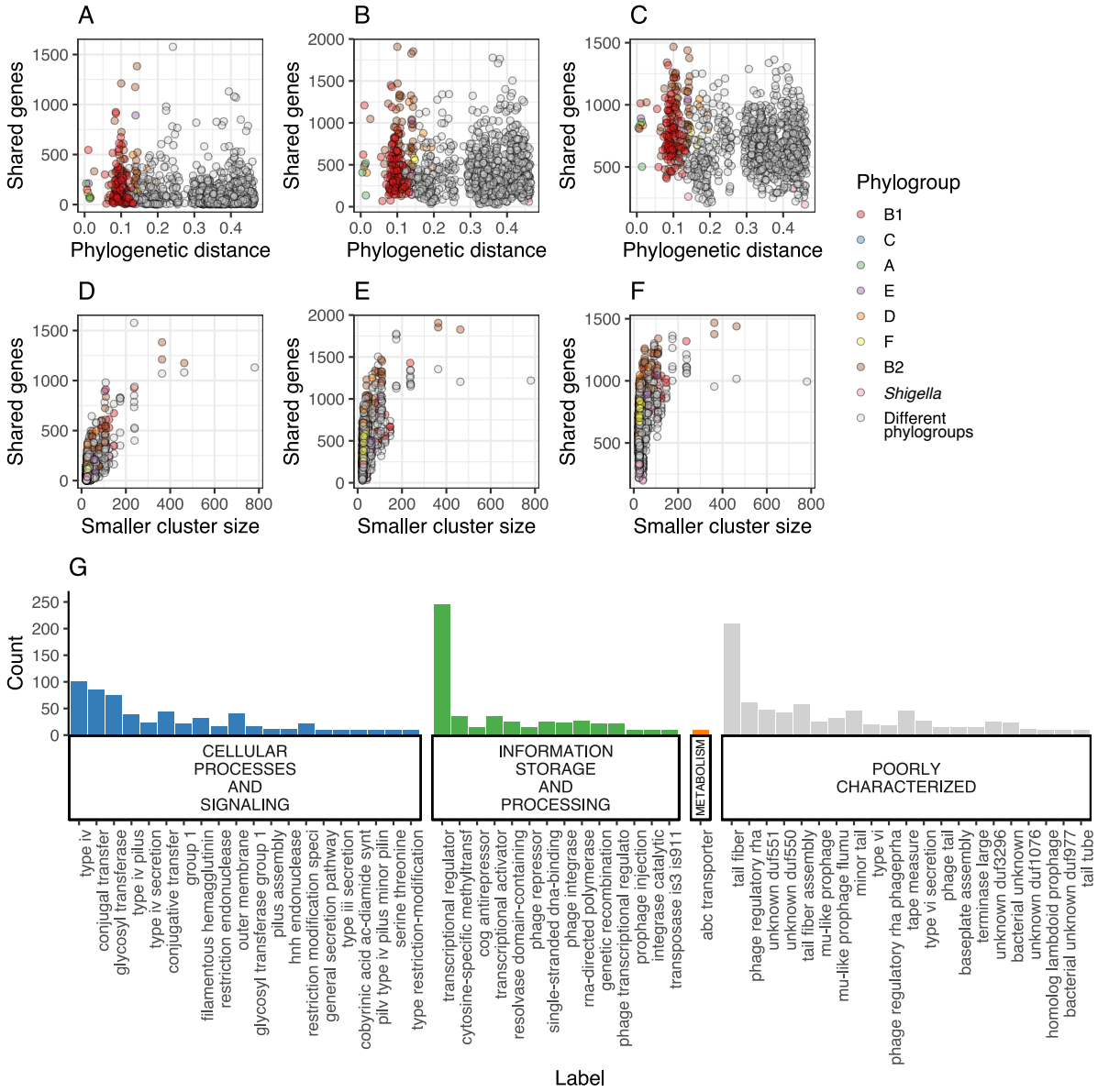

Figure 5.8: Properties of low frequency genes in the* E. coli *dataset*.** ***A-C *Relationship between the number of genes shared between every two PopPUNK Clusters and phylogenetic distance between for "multi-cluster rare" genes (**A**), "intermediate and rare" genes (**B**) and "core, intermediate and rare" genes (**C**). Coloured dots indicate that the two PopPUNK Clusters being compared are from the same phylogroup whereas gray dots indicate the two clusters being compared are from different phylogroups.* ***D-F*** *Relationship between the number of genes shared between every two PopPUNK Clusters and the size of the smaller PopPUNK Cluster of the two being compared between for "multi-cluster rare" genes (**D**), "intermediate and rare" genes (**E**) and "core, intermediate and rare" genes (**F**).* ***G*** *Most common phrases taken from the predicted functional annotations of the "multi-cluster rare", "intermediate and rare" and "core, intermediate and rare", divided into the four main COG categories.*

A large fraction of genes from these three gene categories were assigned a COG category of “Poorly Characterised” (Figure 5.7). The most common predicted terms for these genes were prophage related (Figure 5.8G). These included terms such as “tail fiber”, “baseplate assembly”, “terminase” and “Mu-like prophage”. Other common annotations in the other COG categories included “conjugal transfer”, “type IV pilus”, “restriction endo-nuclease”, “integrase catalytic”and “transposase”.

5.4.8 PopPUNK Clusters of broad host range lineage ST10 and MDR lineage ST410 share more low frequency genes with distantly related PopPUNK Clusters than expected

To explore the distribution of low frequency genes further I identified PopPUNK Clusters which share a large number of low frequency genes with other PopPUNK Clusters that are distantly related to them. The median number of low frequency genes each cluster shares with all other clusters that are distant from it (patristic distance higher than 0.4) was compared against the size of the cluster (Figure 5.9A). As expected, there was a linear relationship between the size of the PopPUNK Cluster and the median number of low frequency genes that a cluster shared with distantly related PopPUNK Clusters (log linear regression, R^2=0.547, p=2.965e-08). However, there were also a number of PopPUNK Clusters that shared more low frequency genes with distant PopPUNK Clusters than expected for their size. These include PopPUNK Clusters 12 and 40 (Figure 5.9A). Accordingly, the branches leading to these PopPUNK Clusters had been predicted to have undergone a large number of gain-events of “core, intermediate and rare” and “intermediate and rare” genes relative to the rest of the tips (Cluster 12: 682 and 1574 events, Cluster 40: 333 and 836 events, Tip-mean: 182 and 530 events, not shown). 78% of the isolates from PopPUNK Cluster 12 are of ST10, members of which are known to have a broad host-range. 30% of the isolates in PopPUNK Cluster 40 are from ST410 known as an MDR lineage and another 43% are from ST23. Multidrug resistance was common amongst the other PopPUNK Clusters which deviated from the expected number of shared genes, including PopPUNK Clusters 19, 26 and 34 with resistance observed to aminoglycoside, sulfonamides, beta-lactams and more (See Appendix E). Clusters 26 and 34 predominantly contained EPEC isolates from the GEMs collection (see Section 4.4.4.7).

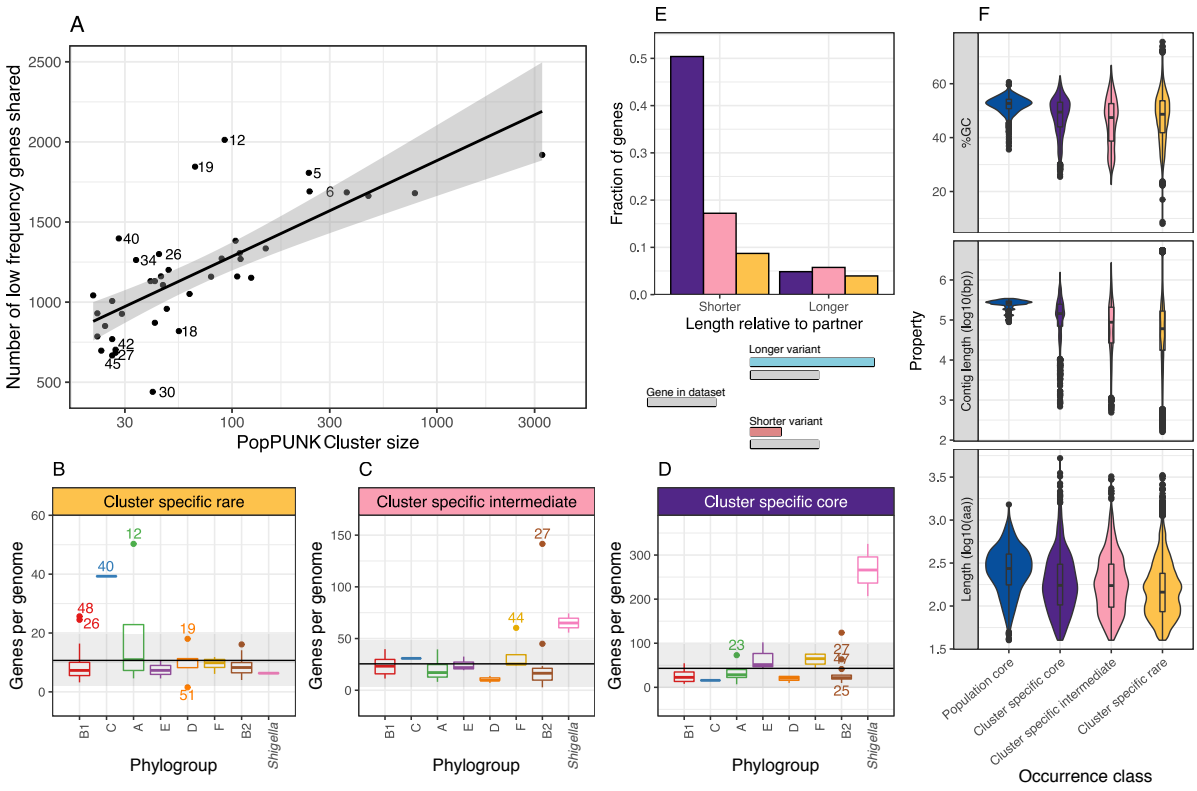

Figure 5.9: Cluster specific genes in the* E. coli *dataset. A *Median number of low frequency genes shared by each PopPUNK Cluster, with other clusters which are phylogenetically distant from it, relative to the size of the cluster. Line fitted using linear regression, shaded area is the 95% confidence interval.* ***B-D*** *Number of "cluster specific rare" genes (**B**), "cluster specific intermediate" genes (**C**), and "cluster specific core" genes (**D**) per genome in each of the 47 PopPUNK Clusters, grouped by phylogroup.* ***E*** *Fraction of cluster specific genes that were found to either be a short variant or a long variant of another gene in the dataset.* ***F*** *Distribution of GC content, contig length and protein length of the genes of cluster specific occurrence classes, compared to the "dataset core" genes.*

5.4.9 Hyper-sharing PopPUNK Clusters possess more "cluster specific rare" genes in a single genome relative to the rest of the clusters

PopPUNK Clusters 48, 26, 40, 12 and 19 had more "cluster specific rare" genes per genome relative to the rest of the PopPUNK Clusters (Figure 5.9B). There was an overlap between clusters which had a high number of "cluster specific rare" genes in each genome and the clusters which shared more low frequency genes with distant PopPUNK Clusters in the dataset. Similar to the "multi-cluster rare" genes, the "cluster specific rare" genes were most commonly predicted to be phage derived or otherwise had other annotations related to HGT such as "conjugational transfer", "restriction modification", "resolvase" and more (not shown).

5.4.10 PopPUNK Clusters which shared fewer low frequency genes than expected also had the largest number of “cluster specific core” genes

PopPUNK Clusters which were not assigned a phylogroup based on the Clermont phylotyping scheme (18) and the *Shigella* PopPUNK Clusters (30,40) shared fewer low frequency genes with distantly related PopPUNK Clusters than expected for their size. The branches leading to these PopPUNK Clusters were estimated to have undergone a large number of gene loss events of “multi-cluster core” genes (Figure 5.5B,C). Additionally, these clusters possessed more “cluster specific core” and “cluster specific intermediate” genes relative to the rest of the PopPUNK Clusters (Figure 5.9A,C,D). While this was expected for *Shigella* spp., PopPUNK cluster 18 is nested within phylogroup B2. This cluster had a mean of 123 “cluster-specific core” genes, relative to a mean of 25 cluster-specific core genes in the rest of the clusters in phylogroup B2. 60% of the isolates of this cluster are from ST504 which has been described in the past as atypical STEC as they have been misclassified as *Shigella* spp. due to the biochemical phenotype these present [415]. Indeed, 100% of the isolates in PopPUNK Cluster 18 were positive for the shiga-toxin gene *stx1B*.

5.4.11 Cluster specific core genes are often truncated variants of other genes in the collection

The sequences of the cluster specific genes, including “cluster specific core”, “cluster specific intermediate” and “cluster specific rare” genes, were aligned against all the other genes in the collection (See Section 5.3.5). Strikingly, 50% of the “cluster specific core” genes were identical along their full length to a region of another gene in the collection (Figure 5.9E). 17% of the “cluster specific intermediate” genes were also identified as shorter variants of other genes in the dataset (Figure 5.9E). Shorter variants of other genes more commonly had an alternative start codon relative to other genes (22% of short variants versus 10% of the rest). Even though only a subset of genes from these occurrence classes were identified as shorter variants of other genes, the length of all cluster specific genes was an order of magnitude shorter than the observed lengths of the “dataset core” genes (Figure 5.9F) The “cluster specific core” genes shared a similar predicted functional profile to those given to the “dataset core and the “multi-cluster core” genes, suggesting these are variants of this same subset of genes (Figure 5.7). Conversely, cluster specific genes had more extreme values in their GC content, particularly the “cluster specific rare genes”, and were more commonly found on shorter contigs (Figure 5.9F).

5.4.12 STEC PopPUNK Cluster 27 and ExPEC PopPUNK Cluster 44 possess a large number of “cluster specific intermediate” genes.

PopPUNK Cluster 44 of phylogroup F and PopPUNK Cluster 27 of phylogroup B2 possessed a high number of “cluster specific intermediate” genes relative to the rest of the clusters (Figure 5.9EC). These clusters had a mean of 60 and 142 “cluster specific intermediate” genes per genome, relative to the mean in the dataset of only 10 “cluster specific intermediate” genes per genome. 100% of the isolates of cluster 44 were from ST648 and were mostly (72%) ExPECs collected from either blood or urine samples. These isolates are multi-drug resistant, with observed resistance to fluoroquinolones, macrolides, aminoglycosides and beta-lactams including ESBLs and carbapenems (See Section 4.4.4.6). ST648 has been described as an emerging multi-drug resistant lineage of phylogroup F, present both in humans and animals [416,417]. Cluster 27, on the other hand, contains 88% isolates from ST583 and 66% of the isolates were collected from fecal samples. Additionally, 66% of isolates from PopPUNK Cluster 27 were positive for shiga toxin gene *stx2B* and 100% positive for *eae* (See Sections 1.1.2.2-3 of Introduction for pathotype definitions). No resistance was observed in this cluster. Thus, these two PopPUNK Clusters with high loads of “cluster specific intermediate” frequency genes are different in their pathogenic and resistance profiles. Their shared property is that they are both out-groups of other clades; PopPUNK Cluster 27 is an out-group of a clade in phylogroup B2 and PopPUNK cluster 44 an out-group in phylogroup F (Chapter 4, Figure 4.8). This resembles the phylogenetic locations of the *Shigella* PopPUNK Clusters 30 and 45 relative to phylogroup B1. PopPUNK Cluster 27 is also similar to these clusters as it shares fewer low frequency genes with distantly related PopPUNK Clusters than expected for its size and the branch leading to PopPUNK Cluster 27 has been estimated to undergone a large number of gain and loss events of “multi-cluster core” genes (Figure 5.9A, 5.4A,B).

5.5 Discussion

An accurate description of the pan-genome of thousands of *E. coli* genomes, when considering all the biases in public genome datasets, required redefining the approach used to understand the distribution of the genes in that dataset. The new approach presented is an extension of previous approaches used for the exploration of the pan-genome in a single species or lineage. In addition to classifying the genes based on their frequency in a lineage, the rules extend to examine the number of lineages, or PopPUNK Clusters, each gene was observed in. The classification presented in this thesis is appropriate given the diversity of the dataset used; Roary, for instance, was designed to handle a dataset with low gene content and sequence diversity and thus would not be applicable to this dataset [305]. Additionally,

this approach corrects for the over-representation of particular lineages in the dataset. For instance, genes which were core and specific to a single PopPUNK Cluster that has a low representation in the dataset would have been mistaken for “rare” genes had we treated all gene-counts equally. However, it is still important to note that the analysis presented here is still an approximation to our understanding of the true distribution of genes in the *E. coli* population. The true representation of each lineage in the natural *E. coli* population is unknown because most of the sequenced isolates in this study, and indeed the public databases have clinical relevance and as such were highly biased in their sampling. Notwithstanding this, as this approach uses two metrics, it provides a higher-resolution to classify the genes in the dataset into occurrence classes which were fully characterised in this thesis, revealing their different functions and dynamics of gain and loss.

There were only 1,426 “dataset core” genes which are the set of genes which are present in every single *E. coli* PopPUNK Cluster and in more than 95% of the isolates of that cluster. These only represent ~30% of the genes in a typical *E. coli* genome. However, there were twice as many genes which were observed in both “core and intermediate” frequencies in multiple PopPUNK Clusters (2,674) and these represent ~40% of the genes in a single *E. coli* genome. The number of PopPUNK Clusters in which these genes were most commonly observed, their mean frequency within those clusters, their predicted functions and their level of association with the population structure revealed that these genes resemble the “dataset core” and the “multi-cluster core” genes, more than they do to the other occurrence classes. Thus, the "core and intermediate" genes represent a level of error that is tolerated using our approach, and they likely represent the "soft-core" genome of the dataset. The fact that these genes were at times observed in intermediate frequencies in particular clusters could be the result of mistakes in sequencing, assembly, annotation or pan-genome pipelines. Alternatively, these genes may be in the process of being lost in some clades. We observed the loss of these genes in PopPUNK Clusters which are undergoing gene degradation like the *Shigella* spp. clusters strengthening the hypothesis that they may be undergoing loss (Figures 5.4C). Importantly, setting a single cut-off between “intermediate” and “core” genes across the entire dataset removes the additional level of understanding of the intricate differences between the genes. Including the “core and intermediate” genes which were observed in 40 PopPUNK Clusters or more as part of the core genome would double the size of the *E. coli* core-genome in this analysis and its relative proportion in a single genome.

Genes which were either core and specific multiple PopPUNK Clusters, i.e. “multi-cluster core” genes, were most commonly found to be gained or lost in a single event on an internal branch in the phylogeny (Figure 5.4,5.5). Genes from these occurrence classes should be further

investigated as they represent the changes in gene content between the clades in the *E. coli* dataset, including the differences between the phylogroups. The fact that these genes had mostly undergone a single gain or loss event suggests that independant shifts in the "core" genome of two or more unrelated lineages are less common. Even so, in 32% of cases changes in the core occur on 3 or more events and in 25% of cases "multi-cluster core" genes are shared between distantly related PopPUNK Clusters. It would be interesting to explore these cases as these could shed light on the commonality of distantly related PopPUNK Clusters and whether they are likely to share similar ecological environments or pressures that lead to the selection of the same genes under different genetic backgrounds.

In most cases, gene sharing of low frequency genes was found to be independent of the phylogenetic distance between the two PopPUNK Clusters being compared. This is an indication of a lack of barrier for movement of these genes between distantly related isolates, for instance, compatibility of phage receptors across the species. Additionally, low frequency genes were estimated to have undergone a large number of gain and loss events along the tree branches, mostly commonly on the tree tips. This means that low frequency genes transfer between distantly related isolates and this happens on short evolutionary timescales. The dependency between the size of the two PopPUNK Clusters being compared and the number of low frequency genes shared between them means that we do not observe sharing of genes due to under-representation of particular lineages rather than lack of sharing between them. This is a likely scenario in the case of low frequency genes as more isolates need to be sampled for these genes to be observed. We have not sampled enough from most of the PopPUNK Clusters in this study in order to truly understand the level of gene sharing of low frequency genes between them. For the largest clusters, we observed a plateau in the number of shared low-frequency genes, meaning that from a specific sample size we were able to capture most of the low frequency genes that are shared between these clusters.

Particular PopPUNK Clusters shared more low-frequency genes with distantly related PopPUNK Clusters than expected for their size and appeared to have an increased ability to acquire genes. Most prominently, these include PopPUNK Cluster 12 which contains isolates from ST10 and PopPUNK Cluster 40 which contains isolates from ST23 and ST410, as well as other PopPUNK Clusters which contain MDR isolates. Interestingly, these same PopPUNK Clusters also contained a high number of "cluster specific rare" genes per genome relative to the rest of the dataset. The correlation between the number of rare genes per genome and enhanced sharing of low frequency genes suggests that a high frequency of rare variants in a single genome can be seen as an enhanced ability to contain low frequency genes in the genome and perhaps to donate them. This assumption appears to be particularly relevant as

many of the cluster specific rare genes were predicted to be mobile elements, vectors of HGT and defense mechanisms that may all contribute to the levels of HGT within these clades and with other clades. ST10 and ST23 are known for their ubiquity as they have been described as both commensal and pathogenic, MDR, as well as isolated from human and animal sources [404,418]. These properties have labelled these lineages as potential facilitators of gene movement in the population [419]. The results in this thesis strengthen these hypotheses. Even more, other PopPUNK Clusters which share similar properties to PopPUNK Clusters 12 and 40 can be viewed as having a high potential to either acquire multidrug resistance or to facilitate movement of genes in the population. Interestingly, PopPUNK Clusters 12 and 40 tended to have smaller genomes relative to the rest of the PopPUNK Clusters in the dataset, suggesting a small genome is not necessarily an indication of a small gene pool or lower levels of HGT (See Section 4.4.4.5)

Particular PopPUNK Clusters shared fewer low-frequency genes with distantly related PopPUNK Clusters than expected for their size. This was particularly apparent in PopPUNK Clusters 30, 45 and 18 which either belong to *Shigella* spp. (30, 45) or were not assigned a phylogroup using the Clermont typing scheme (18). These same clusters had a much larger proportion of “cluster specific core” genes in a single genome and had lost a large number of “multi-cluster core genes”. These results indicate that these lineages are evolving in a separate trajectory to the rest of the PopPUNK Clusters, with little gene sharing and large shifts in their core genome that is specific to them. While on the surface the “cluster specific core” genes appear to represent the acquisition of new genetic material, we found that these genes are commonly short variants of other genes in the dataset, share a similar functional profile to the “dataset core” genes and were an order of magnitude shorter than the “dataset core” genes. Hence, these genes likely represent the process of loss of function and gene degradation rather than gain of function in these clusters. Indeed, major gene degradation has been described in *Shigella* spp. and thus this is an expected result for PopPUNK Clusters 30 and 45 [420]. PopPUNK Cluster 18, on the other hand, contains STEC isolates of ST504 which has been mistaken for *Shigella* spp. in phenotypic testing [420]. Additionally, clusters 30 and 45 have smaller genome sizes relative to the rest of the PopPUNK Clusters, fitting with a model of gene-degradation (Chapter 4, Figure 4.10A). PopPUNK Cluster 18, on the other hand, has a similar genome size to the rest of the clusters in the dataset. This suggests it is undergoing an evolutionary process leading to a phenotype that differs from the rest of the dataset and resembles *Shigella* spp. while maintaining production of shiga-toxin and a large genome.

The new approach for investigating the pan-genome in this study is simple and based on the expansion of the existing approaches however, this analysis provides valuable novel insights regarding gene-sharing and evolutionary dynamics of the lineages in this dataset. Future studies for pan-genomes analysis can use the insights from this study to use more relevant properties beyond the frequency, such as gain and loss rates, clade-association and function, to better define the gene-pools in large collections.