4 Building a collection of 10,000 *E. coli* isolates and defining the gene content in the collection

4.1 Introduction

As of today, there are more than 130,000 *E. coli* and *Shigella* genomes available on public databases. Indeed, recent studies have utilised the availability of these genomes to better understand the population structure and the pan-genome of the species [92,93]. The analysis presented in the Chapter 3 revealed interesting patterns regarding the distribution of a single genetic system in a collection of 259 *K. pneumoniae* genomes. The next two chapters will expand on the analysis presented on TA systems in *K. pneumoniae,* to investigate the distribution of all genes in a collection of 10,000 *E. coli* isolates taken from public databases.

While genomic data is widely available online, the process of building a comprehensive and high-quality collection of genomes is not trivial. The genomic data is stored across different databases which are associated with specific data types. The Sequence Read Archive (SRA) is the main repository which contains all the sequence read data worldwide, and is a collaboration between three read archives worldwide (European Nucleotide Archive; ENA, National Center for Biotechnology Information; NCBI and the DNA Data Bank of Japan; DDBJ) [385]. In some cases, the raw read data is not submitted but only an assembled genome. In these cases, the data will be found elsewhere, for instance, in the NCBI Assembly database. Even more, specific databases have been set up for particular purposes [93,386]. Enterobase, mentioned in Section 1.4, is a database which integrates, assembles and analyses the genomic data of specific enteric pathogens from the SRA, while providing researchers with relevant metadata and software to make these data more accessible [93]. Importantly, when collating the data from these multiple sources, genomes are often duplicated or there are database specific identifiers which need to be matched. Finally, the metadata associated with each genome is often restricted to the publication and is not directly linked to the database from which the genome was downloaded. All of these make the primary process of collating the data challenging.

Following data collation, multiple steps need to be applied to obtain a high-quality collection of genomes and their genes. This includes applying quality control (QC) measures on the downloaded reads to ensure they are of good quality and that there was no contamination. Enterobase, for example, applies its own QC pipeline before importing data from the SRA and after assembly [93]. The reads need to be assembled and annotated for their gene content. Finally, a pan-genome analysis is applied to obtain the gene content across multiple genomes (detailed in Section 1.4.1.4). The most widely used tools for genome assembly, annotation and pan-genome analysis were published anywhere from five to twelve years ago [292,305,356,387]. As the number of genomes has grown exponentially (Figure 1.7), the most commonly used tools can become obsolete as they do not scale well for a very large number of genomes. For instance, a pan-genome genome analysis requires an all-against-all comparison of the CDSs across all isolates being compared. In an analysis of 10,000 isolates, each with 5,000 genes, this would require 1.25 quadrillion pairwise comparisons. For this reason, some (but not all) existing pan-genome analysis tools use an initial step to remove redundant sequences [305,388]. Even so, with a very large dataset of a diverse organism like *E. coli*, the number of unique sequences is large enough that removing redundant sequences does not solve the complexity issues. Therefore, existing studies using very large datasets have compromised on the level of resolution of the analysis applied and were generally limited to high-level descriptive studies with few downstream analyses [92,93].

4.2 Aims

The aim of this Chapter was to build a comprehensive and high-quality collection of *E. coli* and *Shigella* isolates taken from public databases. The work in the Chapter is divided into the following steps which were required to obtain a complete collection of 10,000 *E. coli* isolates and their gene content:

- The data collection process
- The characteristics of the dataset including associated metadata, population structure and AMR and virulence profiles.
- Definition of the gene content across this collection

4.3 Methods

4.3.1 Data collection

The data collection process for this project is summarised in Figure 4.1 and is detailed in the Results section, including specific modifications to the tools used and all the QC measures applied. All scripts for downloading and processing the genomes are available at https://github.com/ghoresh11/ecoli_genome_collection. The final collection of genomes consisted of 10,159 presumptive *E. coli* and *Shigella* genomes.

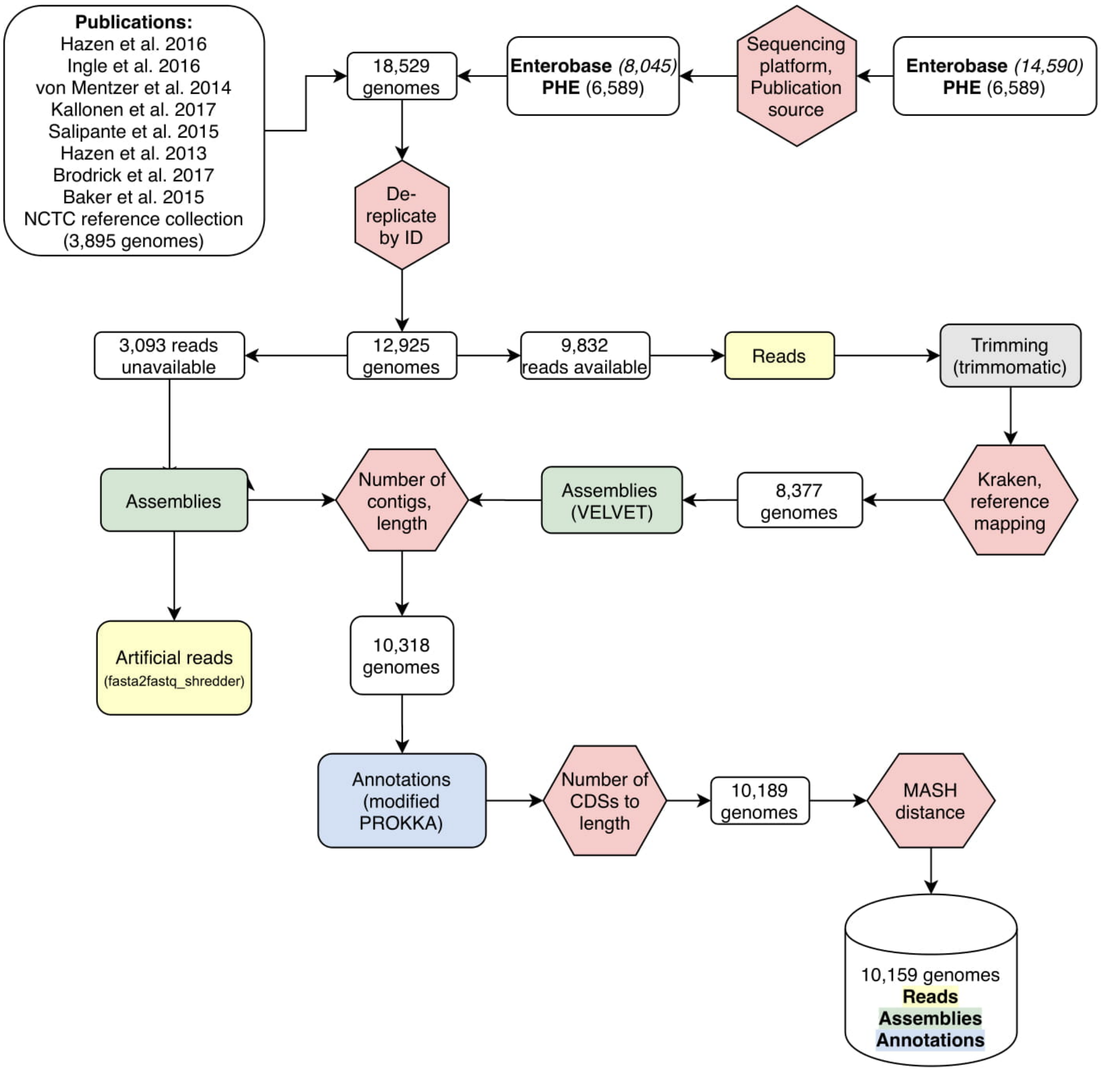

Figure 4.1: Workflow for collating the* E. coli *genome collection*. Steps taken to obtain a final curated, comprehensive and high-quality collection of genomes which include, for all genomes, reads, assemblies and annotation files. QC steps are in red hexagons.*

4.3.1.1 Reads

Reads were downloaded from the SRA using fastq-dump (v2.9.2). Reads which had been Illumina sequenced were trimmed using trimmomatic (v0.33) [389] with the *TruSeq3-PE-2* adaptors, a minimum length of 36 bp, and parameters LEADING=10, TRAILING=10, SLIDING WINDOW=4:15 and quality encoding Phred33. When reads were unavailable, assemblies were shredded into artificial reads (fasta2fastq_shredder.py) with 100bp paired reads from a 350bp insert every 3 bases along a linear genome.

4.3.1.2 Assemblies

Reads were assembled by VELVET (v1.2.09) [356] using the prokaryotic assembly pipeline (v2.0.1) with default setting [357].

4.3.1.3 Gene calling

Predicted CDSs, referred to as "genes" were called using a modified version of Prokka (v1.5). Prodigal (v2.6) was trained using a random selected set of 100 genomes from the entire dataset using the "prodigal.py" script available in Panaroo [292,306]. The training file was then used as the input for Prokka for the predicted genes in the entire dataset. This was compared against running Prokka without using a training file for all genomes. Panaroo was used to compare the gene content of two annotation files by building a synteny graph of the genes [306].

4.3.2 MLST

The ST of all genomes was determined by running "mlst_check" (https://github.com/sanger-pathogens/mlst_check) according to the Achtman MLST scheme downloaded from PubMLST on Jan 22nd, 2019 [390].

4.3.3 Genome Clustering using PopPUNK

Population Partitioning Using Nucleotide K-mers (PopPUNK) (v. 1.1.3) was used to group the assemblies into PopPUNK Clusters [277]. PopPUNK uses Mash to calculate the pairwise distance between every two assemblies. Mash estimates the Jaccard distance between two sequences using a reduced set of k-mers of a defined size *k* [279]. PopPUNK applies Mash with increasing values of *k*. The "core" ($\boldsymbol{\pi}$) and "accessory" (*a*) distances between two assemblies are estimated in PopPUNK by fitting a function which measures the probability of any two sequences matching between the two assemblies across the increasing values of *k* used for Mash (the function: $p_{match} = (1\text{-}a)(1\text{-}\boldsymbol{\pi})^k$). The "core" and "accessory" distances were inferred in this analysis using the *k* values 18, 21, 24, 27 and 31 as these values generated a good fit. Following the distance calculation, the pairwise "core" and "accessory" distances were fitted into clusters using two-dimensional Gaussian mixture models to split the points into K two-dimensional Gaussian distributions and to identify the "core" and "accessory" distance values which represent isolates belonging to the same "strain" or "lineage". The model fitting was applied using six different values of K (5, 8, 11, 14, 17 and 20). The scores generated by PopPUNK for all values of K were compared and these are summarised in Table 4.1. The value of K=11 was chosen for the clustering as it had the overall lowest entropy and comparably high overall score. A network between all assemblies is constructed where each node is an assembly and an edge is drawn between two assemblies only if their “core” and “accessory” distance is within the “within strain” cluster in the result of the two-dimensional Gaussian mixture models. Each connected component in this network is defined as a “PopPUNK Cluster”.

Table 4.1: PopPUNK Clustering statistics. Statistics retrieved from clustering genomes using different values of K when running PopPUNK. Green: The chosen value of K with the lowest entropy.

K	Components	Density	Transitivity	Score	Entropy
5	920	0.1444	0.9929	0.8496	0.0082
8	1120	0.1405	0.9852	0.8467	0.009
11	1185	0.139	0.982	0.8455	0.0042
14	1918	0.1	0.8973	0.8075	0.0055
17	1856	0.1048	0.9093	0.814	0.0053
20	3361	0.0208	0.6273	0.6143	0.0138

4.3.4 Phylogenetic analysis

The core gene phylogeny was inferred from the core gene alignment generated using Roary for each PopPUNK Cluster [305], and a tree from the SNPs, extracted using SNP-sites [332] (v2.3.2), was constructed using FastTree [391]. Treemer (v0.3) [392] was used to select ten genomes from each PopPUNK cluster as representatives of that cluster and representative of the diversity within that cluster. Treemer greedily prunes leaves off the phylogeny by choosing a random lead from the closest pair of leaves in every iteration, until the number of selected leaves in the tree is reached. Similarly, only a single representative sequence was chosen using Treemer from each of the 50 PopPUNK clusters to generate a minimal tree containing only 50 sequences. In both cases, the core gene phylogeny was inferred from a core gene alignment generated using Roary on the 500 representative genomes [305]. A maximum likelihood tree from the informative SNPs, chosen using SNP-sites (v2.3.2) [332], was constructed using RAxML (v8.2.8) [282] with 100 bootstrap replicates.

4.3.5 Phylogroup assignment

EzClermont (v0.4.5) was used to assign the phylogroup of the 500 representative genomes selected in the previous section [393]. EzClermont applies *in-silico* PCR of marker genes to assign phylogroup according to the phylotyping scheme presented in [271]. PopPUNK clusters were assigned a phylogroup according to the most common phylogroup assignment of the ten representative strains. Phylogroup assignments were corrected based on the phylogeny.

4.3.6 Identification of AMR and virulence genes

A collection AMR genes were obtained from the modified version of ARG-ANNOT available on the SRST2 website (https://github.com/katholt/srst2/tree/master/data, downloaded on 08.03.18) [288,290]. Virulence factors were downloaded from the Virulence Finder Database (https://bitbucket.org/genomicepidemiology/virulencefinder_db/src, downloaded 24/08/18). Read files of genomes (real and artificial) were searched for the presence or absence of genes against the downloaded databases using ARIBA (v2.14) with default settings [283]. A gene was marked as present only if 80% of the database entry was covered, otherwise it was marked as absent.

4.3.7 Pathotype assignments

Each isolate was assigned a pathotype according to the presence and absence of specific virulence genes, as well as the source of isolation (Figure 1.4). If the source of isolation was either “blood” or “urine” it was assigned to “ExPEC”. If any variant of shiga-toxin was present it was assigned to “STEC”. If *eae* was present it was assigned to aEPEC/EPEC. If both shiga-toxin and *eae* were present it was assigned to “EHEC”. If either *aatA*, *aggR* or *aaiC* were present it was assigned EAEC. If *est* or *elt* were present it was assigned to ETEC. If *ipaH9.8* or *ipaD*, characteristic of the invasive virulence plasmid pINV, were present it was assigned to EIEC. A pathotype was assigned to a PopPUNK Cluster if at least half of the isolates of the cluster were assigned to the same pathotype.

4.3.8 Pan-genome analysis

4.3.8.1 Pan-genome analysis on each PopPUNK cluster

A pan-genome analysis using Roary [305] was applied on each PopPUNK Cluster separately using the default identity cut-off of 0.95 with paralog splitting disabled [305]. The gene accumulation curves were generated using the *specaccum* function in the vegan (v2.5.6) library with 100 random permutations [359].

4.3.8.2 Combining the pan-genomes of all PopPUNK Clusters

The outputs of the pan-genome analysis of each PopPUNK Cluster were combined to generate a final collection of gene clusters of the entire dataset according in the following steps:

1. Gene cluster definitions, from the Roary analysis within each PopPUNK cluster, were assumed to be the best approximation of the representation of the genes that are well-defined within a closely related group of genomes. Note that each gene cluster has multiple members, i.e. sequences (Figure 4.2, Step 1). A representative sequence was chosen for each gene cluster as the sequence that had the most common length within that gene cluster (the modal length). If there was no mode, a sequence with the median length was chosen.
2. A pan-genome analysis using Roary was applied on all PopPUNK Clusters in a pairwise manner using an identity threshold of 0.95 and with paralog splitting disabled. Namely, a pan-genome analysis was conducted including all genomes of PopPUNK Clusters 1 and 2, 1 and 3, 1 and 4 etc, leading to a total of 1,081 Roary analyses (47 choose 2). This generated gene clusterings for all pairs of PopPUNK Clusters. Note that each gene cluster in the combined Roary analysis had multiple sequences from both PopPUNK Clusters (Figure 4.2, Step 2).
3. A graph was constructed such that each node was one gene cluster from the original Roary outputs from Step 1, named the “combined Roary graph” (Figure 4.2, Step 3).
4. An edge was drawn between a gene cluster of PopPUNK Cluster “A” to a gene cluster of PopPUNK Cluster “B” if there was a gene clustering in the combined Roary analysis such that 80% of the sequences of the gene cluster of “A” were in the new combined clustering and 80% of the members of the gene cluster of “B” were also in the combined clustering (Figure 4.2, Step 4).
5. The following corrections were applied to remove likely incorrect connections between gene clusters in the combined Roary graph (Figure 4.2, Step 5):
    1. Density based clustering was applied on each connected component of the combined Roary graph using the Jaccard similarity between every two nodes with the `dbscan` method of the python package sci-kit learn[394] with parameters epsilon=0.5 and min_samples=6. Edges between a gene cluster of PopPUNK Cluster A and a gene cluster of PopPUNK Cluster B that do not belong to the same dbscan cluster were removed.
    2. A nucleotide MSA using mafft (v7.310)[364] with default settings was applied to all representative sequences of each gene cluster in a connected component of the combined Roary graph. If the alignment of two representative sequences

had more than 20% mismatches along the length of the longer sequence, the edge between them in the combined Roary graph was removed.

6. To correct for over splitting, the representative sequences of all the gene clusters of the original Roary outputs were aligned to each other using blastp (version 2.9). Representative sequences which were more than 95% identical, over 80% of their length, were merged.
7. Following corrections, the connected components of the combined Roary graph were recalculated and these were the final set of gene clusters in the entire dataset (Figure 4.2, Step 6).

4.3.9 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. The ggplot2 (v3.2.1) package was used for plotting [360].

4.4 Results

4.4.1 Constructing a collection of 10,000 *E. coli* isolates

A collection 18,156 *E. coli* genomes, isolated from human hosts, were downloaded and curated to create a final collection of 10,159 genomes as summarised in (Figure 4.1).

4.4.1.1 Initial collection of 18,156 genomes

For an initial collection of human *E. coli* genomes for which complete metadata is available, whole genome sequences were downloaded and the metadata combined from recent publications describing specific *E. coli* pathotypes. These included 70 EPEC isolates from [115], 398 EPEC isolates from [119], 373 ETEC isolates from [117], 1,509 ExPEC isolates from [397], 302 ExPEC isolates from [121], 113 EHEC and EPEC from [116], 538 ExPEC isolates from [174] and 25 ExPECs from [398]. Additionally, 140 isolates were taken from the Murray collection [399], which includes isolates collected from the pre-antibiotic era. Furthermore, 313 genomes were available from the NCTC reference collection which have been long read sequenced (https://www.sanger.ac.uk/resources/downloads/bacteria/nctc/).

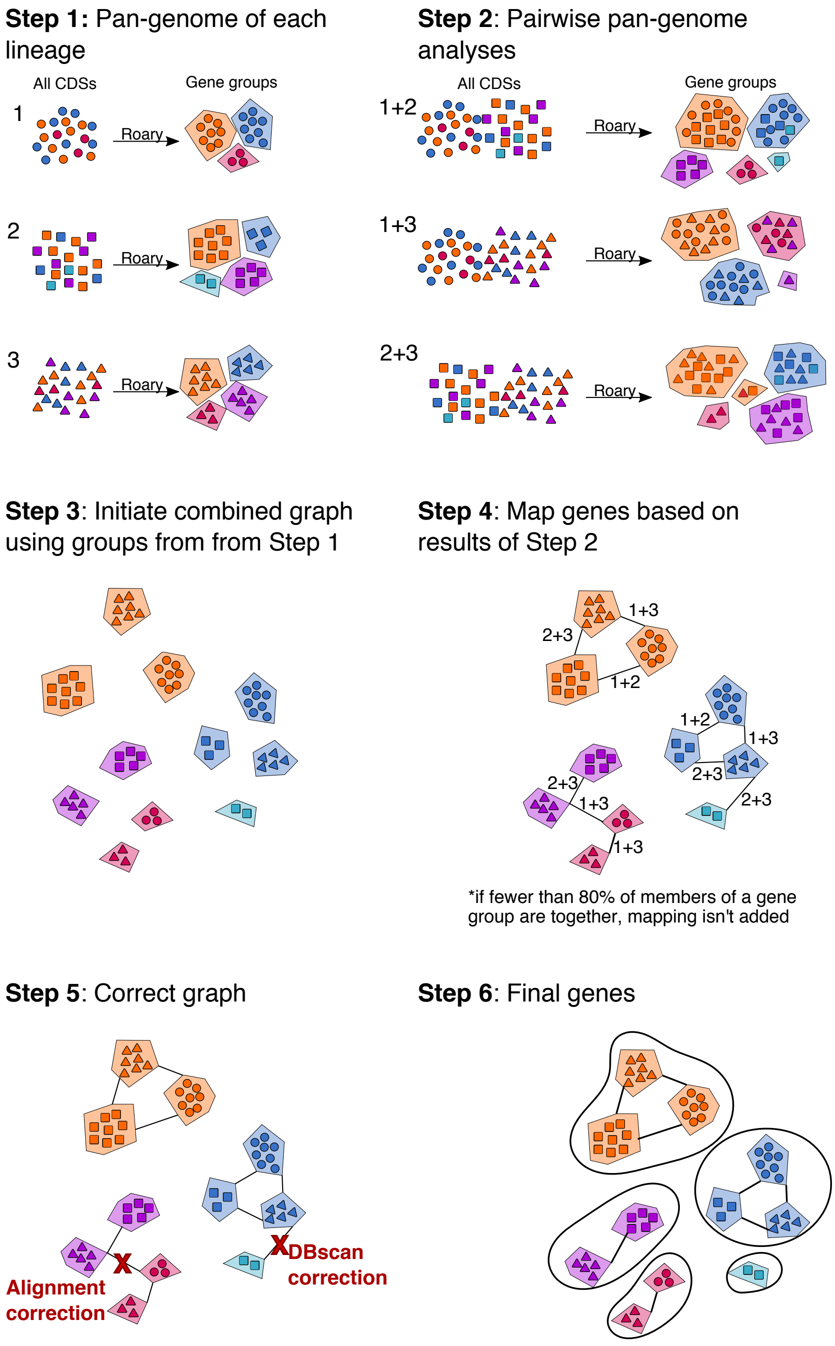
Step 1: Pan-genome of each lineage
All CDSs
Gene groups
1
2
3
Roary
Step 2: Pairwise pan-genome analyses
All CDSs
Gene groups
1+2
1+3
2+3
Roary
Step 3: Initiate combined graph using groups from from Step 1
Step 4: Map genes based on results of Step 2
1+3
2+3
1+2
*if fewer than 80% of members of a gene group are together, mapping isn't added
Step 5: Correct graph
X
Alignment correction
X
DBscan correction
Step 6: Final genes

Figure 4.2: Method for combining the pan-genome analysis of all PopPUNK Clusters. *Step 1: a pan-genome analysis is applied on each PoPPUNK Cluster separately, generating gene clusters from all the CDSs of all genomes in that cluster. Step 2: A pan-genome analysis using Roary was applied on all PopPUNK Clusters reciprocally, generating new gene clusters. Step 3: A graph is constructed where the original gene clusters are the nodes. Step 4: An edge between two gene clusters was added if the members of both gene clusters were grouped together in the pairwise pan-genome analysis. Step 5: Edges were removed from the graph using density-based clustering and sequence alignments. Step 6: Connected components were extracted as the final gene cluster definitions.*

These genomes were supplemented to include other genomes available from public databases for which there was only partial associated metadata available. 14,590 genomes (isolated from human hosts) were downloaded from EnteroBase [400] on August 1st, 2018. EnteroBase searches the NCBI short read archive every day to download (and assemble) newly submitted Illumina reads or complete genomes (See Section 1.4). These genomes were filtered to include only genomes which were sequenced with Illumina, Pacbio or Minion platforms and were open for use, leading to a total of 8,045 genomes. Enterobase's data usage policy states metadata, assemblies and genotyping can only be used for academic purposes following their release. Therefore, the remaining genomes in the dataset were mostly from either publications or otherwise from public surveillance institutions from which we were able to obtain approval to use. These include Public Health England (PHE), the Food and Drug Administration (FDA) and the CDC. An additional 6,589 raw read sequences from Public Health England Routine surveillance bioproject (PRJNA315192) were downloaded on September 17th, 2018.

All downloaded reads were assembled (See Section 4.3.1.2). Artificial reads were generated for assemblies for which reads were unavailable (See Section 4.3.1.1). Annotation files were generated using a modified version of PROKKA, detailed below [293]. By the end of the data collection process, reads, assemblies and annotations were available for all genomes.

4.4.1.2 Modifying the annotation tool PROKKA to remove errors in gene calling between genomes

Prokka combines the use of five other tools to identify features in the assemblies. Importantly, Prokkka uses Prodigal to predict CDSs, or "genes" as they will be referred to in this thesis for simplicity [292,293], By default, Prokka will use the input genome to define properties for gene calling such as the start codon usage, ribosomal binding site motif usage etc. [292]. In this thesis, a collection of 100 randomly sampled genomes from the complete collection of genomes were used to train Prodigal to define these properties (See Section 4.3.1.3). All the genomes were then annotated using the same training properties. This ensured the gene calling was done in a consistent manner for all genomes.

In most cases, the gene content between the modified and default versions Prokka varied by less than 4%, with 96.5% of genes being called the same using both versions (Figure 4.3A). However, there were a number of outlier genomes for which the difference in gene content was much higher. The difference in these cases was mostly driven by genes within each genome which were no longer called when using the same training file across all genomes (Figure 4.3B). In general, the genes which were differentially called were shorter, had a more varied GC content, were often present on shorter contigs and closer the contig edge, and more often began with an alternative start codon (Figure 4.3C-G).

4.4.1.3 Filtering to a high-quality collection of 10,159 genomes

Genomes were removed from the collection in multiple steps along the collection process when they did not pass the QC measures (Figure 4.1).

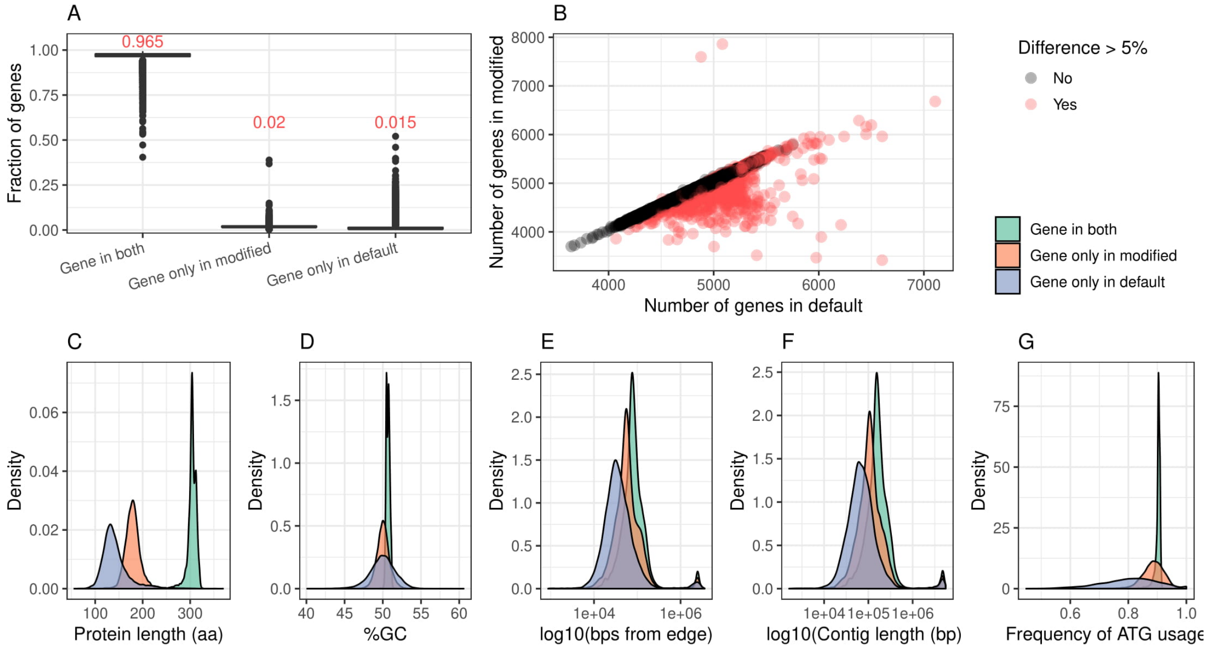

Figure 4.3: Effect of modifying Prokka on the CDS prediction. *The default version generated CDS properties for each genome individually, the modified used the same properties for all genomes.* ***A*** *Fraction of genes in each genome which was found in both runs, only in the modified run and only in the default. Red text: the average fraction of genes in each group across the 10,000 genomes.* ***B*** *Relationship between the number of genes in the default run compared to the modified run for each genome. Red: outliers from A for which there is more than 5% difference in gene content between both runs.* ***C-G*** *Protein length (**C**), GC content (**D**), distance from contig end (**E**), contig length (**F**) and frequency of ATG usage (**G**) of genes that were called in both, modified and default Prokka runs.*

Read filtering: Kraken was used on the reads to determine what organism had been sequenced [401]. Kraken uses a k-mer based search of the reads on a taxonomy tree of RefSeq genomes to find the most likely taxon for each read. If fewer than 30% of reads were assigned to *E. coli* or *Shigella* spp., the genome was removed (Figure 4.1). Following that, reads were mapped to an *E. coli* reference strain cq9 (GCF_003402955.1) and QC stats were calculated. Samples were removed based on the according to the distributions of QC values across all reads (Percentage of reads mapped to the reference >60%, the mean insert size <80bp, percentage of bases mapped that were mismatches was >0.03, percentage of heterozygous SNPs<3%).

Assembly filtering: Assembled genomes were filtered to remove those with more than 600 contigs or those that had a total combined contig length of less than 4 Mbps or larger than 6 Mbps (Figure 4.4A,B, 4.1).

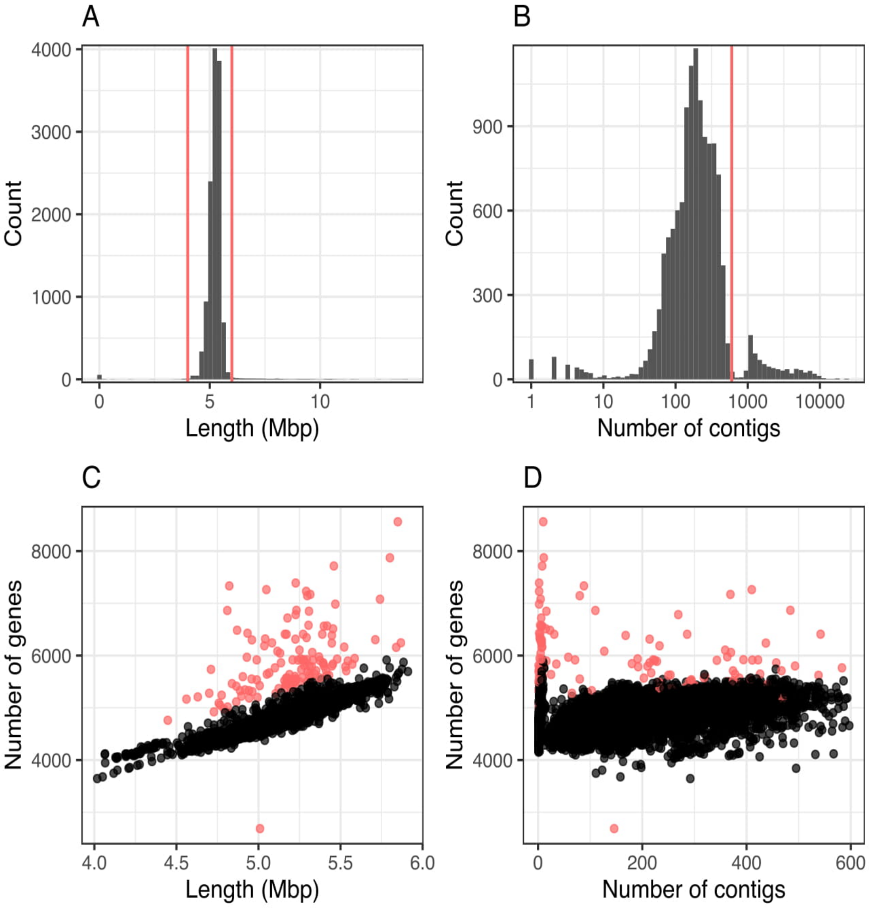

Figure 4.4: Quality control measures used to filter* E. coli *genomes. A *Distribution of genome lengths in the collection. Red lines: genomes shorter than 4 Mbps or longer than 6 Mbps were removed.* ***B*** *Distribution of number of contigs per genome in the collection. Red line: genomes with more than 600 contigs were removed.* ***C*** *Correlation between genome length and number of predicted CDSs using Prokka. Red: Genomes which deviate from the expected number of genes were removed.* ***D*** *Relationship between the number of contigs and number of predicted genes. Red: Genomes which deviate from the expected number of genes presented in C.*

Annotation filtering: The number of genes from each genome was retrieved from the annotations. There was a linear correlation between the size of the genome and the number of genes called (Figure 4.4C). Genomes which deviated from linear correlation by 500 genes were removed (Figure 4.1). These genomes tended to have fewer contigs, i.e. they were long-read sequenced (Figure 4.4D).

Average Nucleotide Identity based filtering: Mash distances were calculated between all the assemblies [279]. Mash uses a minimised database of k-mers to represent each genome (based on the Minhash sketch), and returns the Jaccard distance between the k-mers of every two genomes. A network was constructed so that there was an edge between every two genomes only if their Mash distance was smaller than 0.04 (equivalent to 96% Average ANI) [279]. Isolates from the same species should have an ANI of approximately 95-96%, i.e. Mash distance smaller than 0.04 [402]. Therefore, genomes were removed if they were disconnected from the largest connected component which should represent the *E. coli* species (Figure 4.1).

4.4.2 Characteristics of the filtered dataset

4.4.2.1 Most of the genomes are from developed countries, collected in surveillance in clinical settings

The vast majority of genomes were available from public resources which conduct regular surveillance of *E. coli* in clinical settings. These PHE (5,207 genomes), FDA (883 genomes) and the CDC (561 genomes) (Figure 4.5A). The availability of surveillance data from the United Kingdom and the United States lead to a biased collection from these countries which represented 70% (7,085/10,158) and 15% (1,548/10,158) of the dataset respectively. The rest of the genomes originated mostly from other countries in Europe, with only a small fraction of genomes available from Asia, Africa and Oceania (Figure 4.5A). The continent and country of 336 genomes was unknown.

The source of isolation for 38% of the samples considered here were taken from faeces, blood and urine (Figure 4.5B). However, the remaining samples were simply recorded as having been isolated from unknown “human sources”. Isolates from Africa and Asia include only those collected from faecal samples, whereas isolates from Europe and North America include those causing both intestinal and extra intestinal disease (Figure 4.5B).

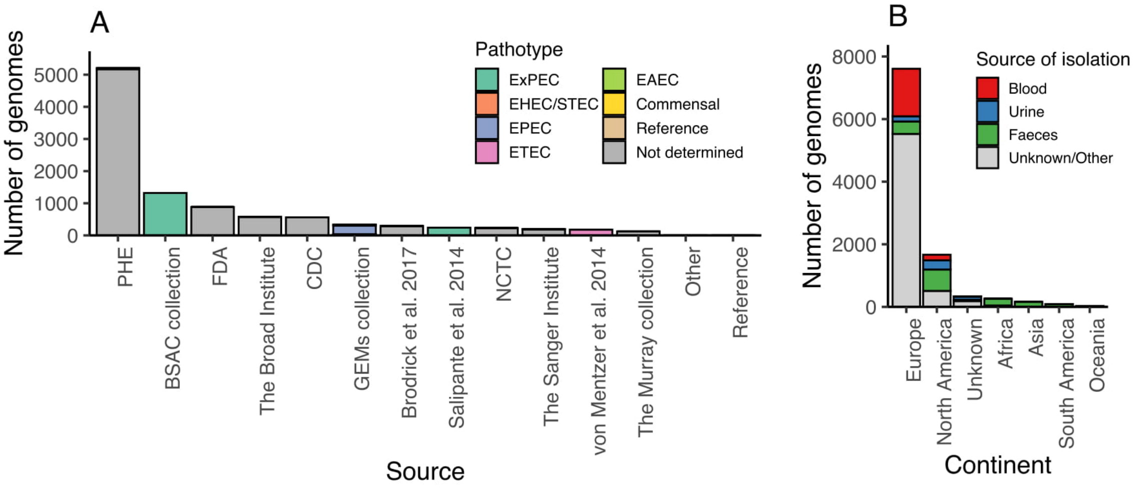

Figure 4.5: Source of **E. coli** ***genomes. A** Source of the* E. coli *genomes in the collection, coloured by the pathotype associated with the specific studies.* ***B*** *Continents from which the* E. coli *genomes were collected, coloured by source of isolation.*

4.4.2.2 Only 5% of all genomes are the cause of diarrheal disease in developing countries

The pathotype for isolates taken from urine and blood samples was assigned as ExPEC (2,299 genomes, 15%). The metadata of 522 (5%) isolates was available and thus the pathotype was known, based on the publication (Figure 4.5A). Within these isolates, the representation of diarrheal disease causing *E. coli* pathotypes, EPECs and ETECs, was very low with only 3% and 2% of the genomes belonging to these pathotypes, taken from the The Global Enteric Multicenter Study (GEMS collection) and from [117] (Figure 4.5A). For the remainder of the genomes, the pathotype could not deterministically be assigned (7,335 genomes). This is due to pathotypes not being defined by clear one to one relationship of presence or absence of specific virulence genes, but by clinical manifestation or phenotype. In Section 4.4.4.7 of this thesis, the virulence profiles of genomes are described as predictive of their pathotype (See Section 1.1.2.3, and Figure 1.4).

4.4.2.3 Six STs represent more than 50% of the genomes in the collection

993 different STs were identified in the collection. 87 STs (9%) alone account for 80% of the isolates. Six STs, 11, 131, 73, 10, 95 and 21, account for 50% of the isolates (Figure 4.6A,B). Many of the latter represent important STs linked to human health. For instance, ST11 (30% of all genomes) is associated with EHEC serotype O157:H7, a major foodborne pathogen that can be contracted by eating contaminated foods, specifically beef products, as since it lives in

the guts of cattle and is the cause of HUS (See Section 1.1.2.2). The collection also includes STs of non-O157 EHECs, including STs 17 (2%) and 21 (2%). STs 131 (8%), 73 (4%), and 95 (3%) are all STs known to be associated with extra-intestinal disease[174,397,403]. ST10 (3%) is a broad host range ST which has been observed in all *E. coli* pathotypes and across hosts [404].

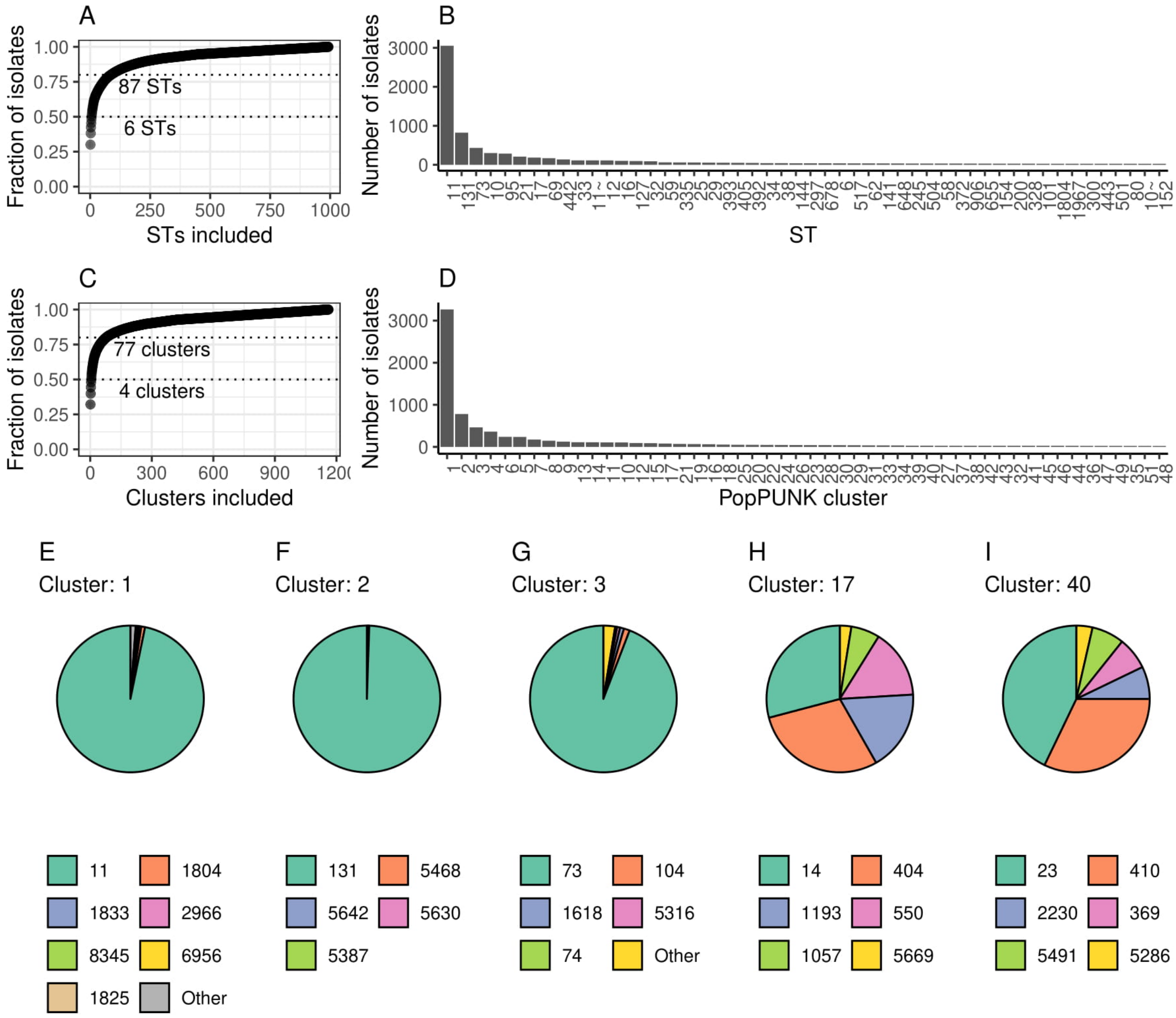

***Figure 4.6: Distribution of STs and PopPUNK Clusters in the collection**. **A,C** Coverage of genome collection by increasing the number of STs (A) or PopPUNK clusters (C) included in the study. Dotted lines: Number of STs (**A**) or PopPUNK clusters (**C**) which account for 0.5 and 0.8 of all isolated in the genome collection. **B,D** Number of genomes in the fifty largest STs (**B**) and PopPUNK clusters (D). **E-I** Examples of ST distributions in five of the PopPUNK Clusters - Cluster 1 (**E**), 2 (**F**), 3 (**G**), 17 (**H**) and 40 (**I**).*

The bias in the collection towards STs which are known to cause severe disease such as HUS or invasive infections emphasises the sampling bias; 80% of isolates originate from developed countries where diarrheal disease caused by EPEC and ETEC is less common. 790 STs (~80% of the STs) are represented by five isolates or fewer and are rarely observed. Thus, this collection is inherently biased towards clinical isolates which are under surveillance in the UK and US, and does not represent the human *E. coli* population.

4.4.3 PopPUNK can be used to group the collection into isolates belonging to the same lineage

In order to examine the gene pool of the *E. coli* genomes considered here, the genomes were grouped into clusters of closely related isolates using PopPUNK [277]). PopPUNK uses a k-mer based comparison of genomes to measure the deviation in gene sequence termed the “core distance”, and the deviation in gene content, termed the “accessory distance” between two genomes (See Section 4.3.3). In *E. coli*, it was shown that the “core distance” estimated by PopPUNK correlates with the pairwise SNP distance between the two genomes being compared, and the “accessory distance” correlates with the Jaccard distance based on the presence and absence of CDSs extracted from a pan-genome analysis [277]. Genomes which were sufficiently similar in both their “core distance” and their “accessory distance” were included in the same PopPUNK Cluster (See Section 4.3.3).

This approach was taken in order to handle the biased sampling of the genomes. For instance, the dataset is over-represented with ST11; had all isolates been treated with the same weight in the analysis, the results would be biased to ST11. By examining the gene content within each subpopulation individually and then merging these results while adding weights for the sampling bias, conclusions can be drawn.

The grouping produced 1,185 PopPUNK Clusters. The partition of the genomes using PopPUNK mostly agreed with partitioning the genomes by ST (rand index of 0.923). Therefore, the distribution of PopPUNK cluster sizes was similar to that of the STs with a few large clusters representing most of the population (Figure 4.6A,B). A single cluster, PopPUNK Cluster 1, contained 34% of all genomes (3,326/10,158) (Appendix E). This cluster was mostly comprised of ST11 (Figure 4.6E), i.e. O157:H7 EHEC. Similarly, PopPUNK Cluster 2 contained 8% of all genomes (781/10,158) consisted mostly of ST131 (Figure 4.6F). The third largest cluster, PopPUNK Cluster 3, contained 5% of all genomes (463/10,158) and was mostly composed of ST73 (Figure 4.6G). See Appendix E for a summary of all other PopPUNK Clusters. There were exceptions for which a higher diversity of STs within a PopPUNK Cluster was observed. For instance, PopPUNK Cluster 17 which had 79 isolates, consisting of four almost equally distributed STs (14, 404, 1193 and 550) (Figure 4.6H). PopPUNK Cluster 40, which had 28 isolates, was composed of two equally common STs (410 and 23) along with another four which were less common (Figure 4.6I).

For this analysis, PopPUNK Clusters of fewer than twenty isolates were removed. There were 50 PopPUNK Clusters in total which met this requirement and together they contained 7,693 genomes (76% of the collection) and 271 different STs (27% of collection) (Appendix E). Whilst the effect of this is a further reduction in the diversity of the dataset, it is not possible to characterise the gene pool of groups for which there were too few representatives. Additionally, this approach would further filter out contaminants and isolates which may not be *E. coli*.

4.4.4 Characteristics of the selected 50 largest PopPUNK Clusters

4.4.4.1 Genetic diversity

The median "core distance" and median "accessory distance" estimated within each of the remaining PopPUNK Clusters were correlated, with higher deviations in the core indicating higher deviations in gene content, i.e. in the accessory genome (linear regression, *p*=1.342e-11, R^2=0.61) (Figure 4.7). However, differences between the PopPUNK Clusters were evident, with some PopPUNK Clusters presenting higher diversity in their accessory genome relative to their core genome, and vice versa. For instance, PopPUNK Cluster 40, which contains isolates of STs 410 and 23, had high diversity in its accessory genome relative to the core genome. There was no connection between the size of the PopPUNK Cluster and the median "core" or "accessory" distances (not shown).

4.4.4.2 Population structure

The phylogeny of the 50 selected PopPUNK Clusters was examined by selecting ten genomes from each PopPUNK cluster that captured most of the diversity of that cluster (See 4.3.4), leading to a total of 500 genomes representing the complete dataset. The core genome of these 500 genomes was extracted and the phylogenetic tree of the core gene alignment was built (Figure 4.8). PopPUNK separated the genomes into clearly distinct lineages based on their core genome. The effect of the "accessory distance" between every two isolates was minimal as there was a correlation between "core" and "accessory" distance across the isolates (Figure 4.7). The exception to this was PopPUNK Cluster 12 which was split into two closely related clades. One clade was more closely related to PopPUNK Cluster 28 whereas the other to PopPUNK Cluster 35. The "core" and "accessory" distances estimated by PopPUNK showed that indeed the "core" distance between PopPUNK Clusters 12, 28 and 35 were low and these could be viewed as a single clade according to their core distances. However, PopPUNK Clusters 12, 28 and 35 deviate in their accessory gene content from PopPUNK Cluster 12 whereas the two clades of PopPUNK Cluster 12 are sufficiently low in their accessory distance. That said, PopPUNK Cluster 12 presented the highest median “core distance” and median “accessory distance” between every two isolates (Figure 4.7).

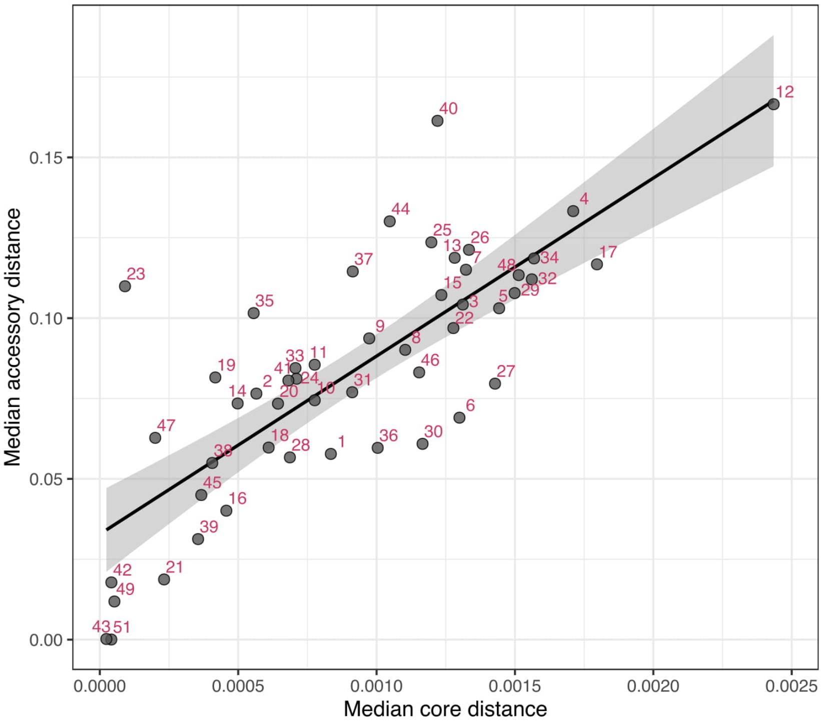

Figure 4.7: ***PopPUNK Clusters’ genetic diversity.*** *Median “core distance” and “accessory distance” between all isolates of the same PopPUNK Cluster. Line fitted using linear regression, showing 0.95 confidence interval.*

Although the dataset was substantially reduced to include only PopPUNK Clusters with 20 genomes or more, the remaining genomes spanned the complete *E. coli* population, defined by having PopPUNK Clusters representing the well described *E. coli* phylogroups (18 from B1, 12 from B2, 4 from A, 5 from D, 4 from F, 3 from E, 1 from C, 2 of *Shigella* representing *S. sonnei* (45) and *S. flexneri* (30) and one phylogroup which was undefined according to the Clermont 2013 phylotyping scheme (18) [271,393]).

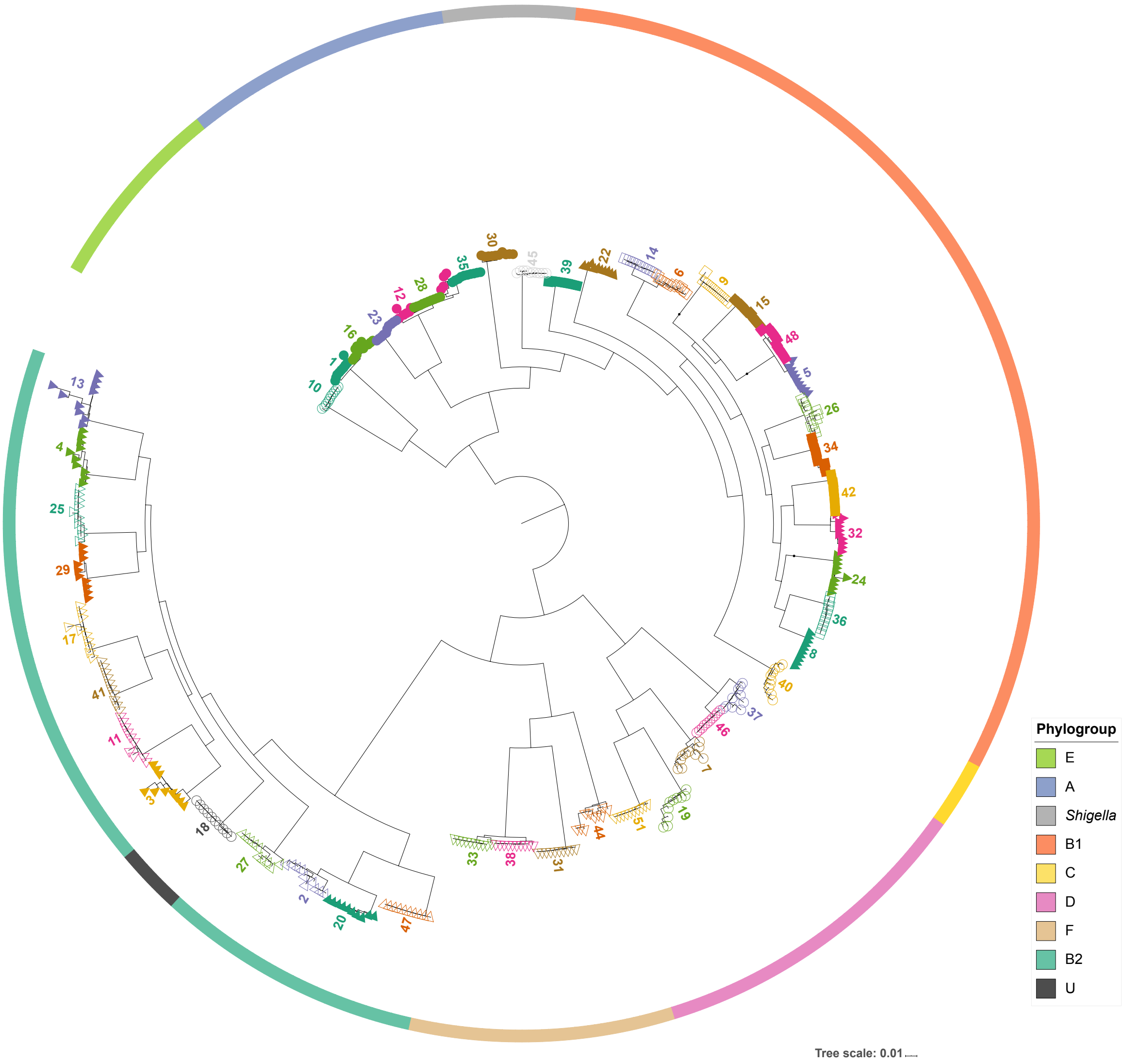

Figure 4.8: Population structure of the PopPUNK Clusters. *Core gene phylogeny of 10 representatives from each of the 50 PopPUNK clusters chosen using Treemer [392]. Coloured bar indicates the phylogroup assignment of the representatives of that PopPUNK Cluster.*

4.4.4.3 Pathogenic and geographic association

The PopPUNK Clusters broadly divided into those enriched for isolates collected from faecal samples (2, 5, 6, 14, 21, 26, 34, 42, 43, 48, 49 and 51) and those collected from blood and urine samples (2, 3, 4, 7, 11, 13, 17, 19, 20, 25, 29, 31, 33, 37, 40, 41, 46, and 47), i.e. those causing intestinal or extra-intestinal disease (Figure 4.9A). Only PopPUNK Clusters 26, 34 and 48 of the intestinal causing disease clusters were enriched for samples collected from Africa and Asia (Figure 4.9B). These clusters mostly represented EPEC and ETEC isolates which had been collected from faecal samples in developing countries as part of the GEMS collection, in contrast to the other PopPUNK Clusters containing faecal samples which include

STECs or EHECs collected in the developed world. PopPUNK Cluster 12, which consisted of 78% isolates from ST10, was the only PopPUNK Cluster that spanned all continents and consisted of all types of isolation source samples (faecal, blood, urine or unknown).

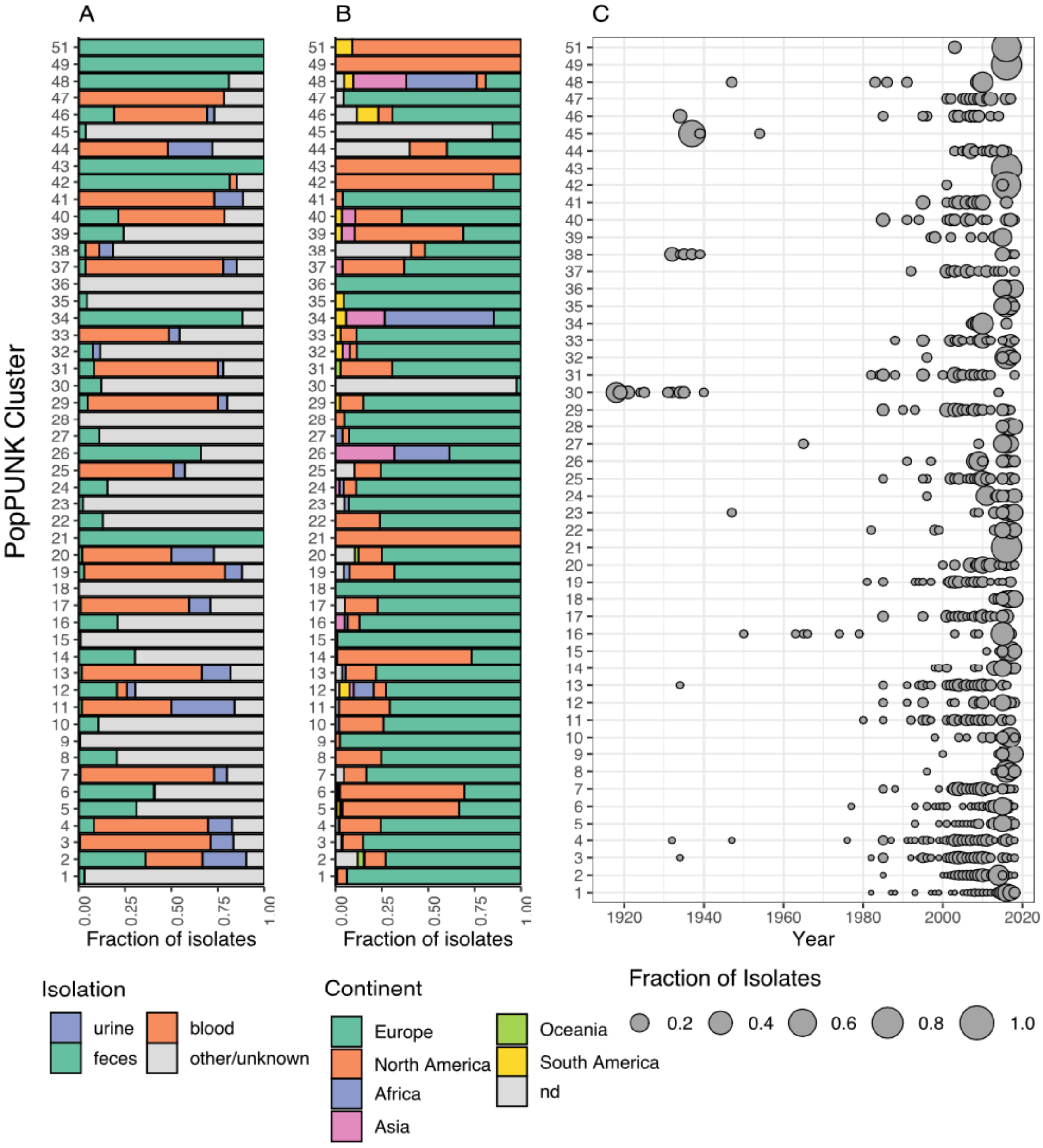

Figure 4.9: Metadata associated with the PopPUNK Clusters. A,B *Source of isolation (**A**) and continent (**B**) of isolates from the fifty PopPUNK Clusters.* ***C*** *Fraction of genomes from each of the PopPUNK Clusters collected from each year (where metadata was available).*

4.4.4.4 Sampling time

A number of PopPUNK Clusters consisted of older isolates taken from the Murray collection. Notably, PopPUNK Cluster 30, with contains *S. flexneri* isolates, had a higher proportion of isolates sampled before 1980 relative to the rest of the collection (Wilcox summed rank test, p<0.05, Bonferroni corrected, Figure 4.9C). 39% of the rest of the genomes for which sampling date was available, were collected in the last 10 years.

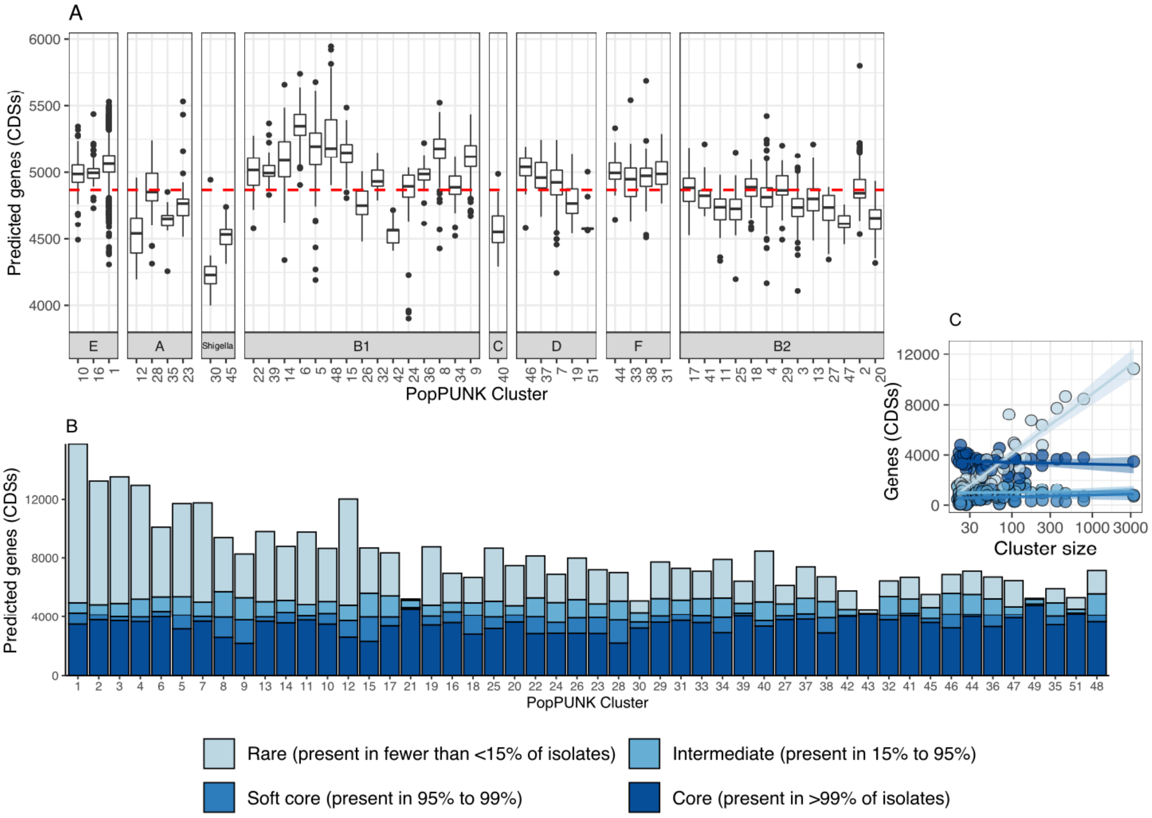

***Figure 4.10: Gene content in the 50 PopPUNK clusters. A** Number of genes (predicted CDSs) per isolate across the PopPUNK clusters, divided by their phylogroup. Dashed line: mean number of predicted genes across the entire population. **B** Number of core (>99% of isolates), soft-core (95%-99% of isolates), intermediate (15%-95% of isolates) and rare genes (<15% of isolates) in each PopPUNK Cluster. Clusters on the x-axis are ordered by their size. **C** Size of PopPUNK Cluster against number of core, soft-core, intermediate and rare genes. Line is fitted using a generalised log-linear model with 0.95 confidence interval.*

4.4.4.5 Genome size and number of predicted genes

The number of genes in a single isolate and the size of the genome varied significantly between the PopPUNK clusters (Figure 4.10A). The mean number of genes corrected across all PopPUNK Clusters was 4,869 genes and a genome length of 5.2 Mbp. Smaller genomes had fewer genes as we used the correlation between genome length and the number of genes as a measure of QC, thus these measures are interchangeable (See Section 4.4.1.3). Isolates from the *Shigella* PopPUNK Clusters 30 and 45 had the smallest genomes with a median of

4,231 genes per isolate and a genome size of only 4.3 and 4.7 Mbp. PopPUNK Clusters 12, 40 and 48, had the second smallest genome lengths with a mean of ~4,500 genes and genome length of ~4.85Mbp. On the other hand, PopPUNK clusters 5, 6, 8, 15, and 48, all from phylogroup B1, had a mean of over 5,100 genes per isolate (200 genes more than the population mean). The number of predicted genes/length of the genome was affected by the phylogroup (Figure 4.10A). Isolates from phylogroups E, F and B1 tended to have larger genomes with a few exceptions. Isolates from phylogroup C, B2 and A tended to have smaller genomes, whereas within phylogroup D a wider range of genome sizes was observed.

4.4.4.6 Antimicrobial resistance profiles

A total of 153 known resistance genes were identified in the collection (See Section 4.3.6), conferring resistance to beta-lactamases, aminoglycosides, macrolides, sulfonamides, fluoroquinolones and other antimicrobial classes (Appendix E) [286]. The number of known resistance genes found within each isolate ranged from no resistance genes detected to a maximum of 18 resistance genes present in a single isolate, conferring resistance to up to nine different antimicrobial classes in a single isolate (Figure 4.11A). The median number of resistance genes per isolate in the complete dataset was one gene. This was because 99% of isolates possess the multidrug resistance efflux pump gene *mdfA[405]* (Figure 4.11B).

Multidrug resistance in an isolate has been defined as resistance to three classes of antibiotics or more [406]. All but six PopPUNK Clusters (21, 28, 36, 43, 47 and 49) had at least one isolate which was MDR. An MDR PopPUNK Cluster was defined as one where half of the isolates or more were MDR, i.e. resistant to three classes of antibiotics or more (Figure 4.11A, Appendix E). 16 of the 50 PopPUNK Clusters investigated in this thesis were MDR. Half of these were PopPUNK Clusters which were isolated predominantly from blood and urine sample, i.e. ExPECs (Clusters 2, 20, 44, 40, 17, 7, 37 and 9). These include PopPUNK Clusters 2 and 20 which both contain isolates of the global ExPEC lineage ST131. Three of the ExPEC MDR PopPUNK Clusters belong to phylogroup D (Clusters 19, 7 and 37). These three PopPUNK Clusters possessed the same set of genes which confer resistance to ESBLs, sulfonamides, tetracycline and aminoglycosides (Figure 4.11B). Three other MDR PopPUNK Clusters predominantly contained EPEC isolates from the GEMS collection (Clusters 26, 34 and 48). The remaining five PopPUNK Clusters (Clusters 32, 35, 18, 16 and 24) were isolated from unknown sources. Resistance to carbapenems was most common within PopPUNK Cluster 44 of phylogroup F with 44% of the isolates of this Cluster possessing the carbapanemase *bla*KPC-2. Resistance in PopPUNK Clusters 44 as well as PopPUNK Cluster 37 were generally high, with most of the isolates in these PopPUNK Clusters resistant to seven classes of antibiotics or more, comparable and even higher to the resistance observed for

ST131 in PopPUNK Clusters 2 and 20. Resistance to colistin was not observed within any of the isolates in this dataset.

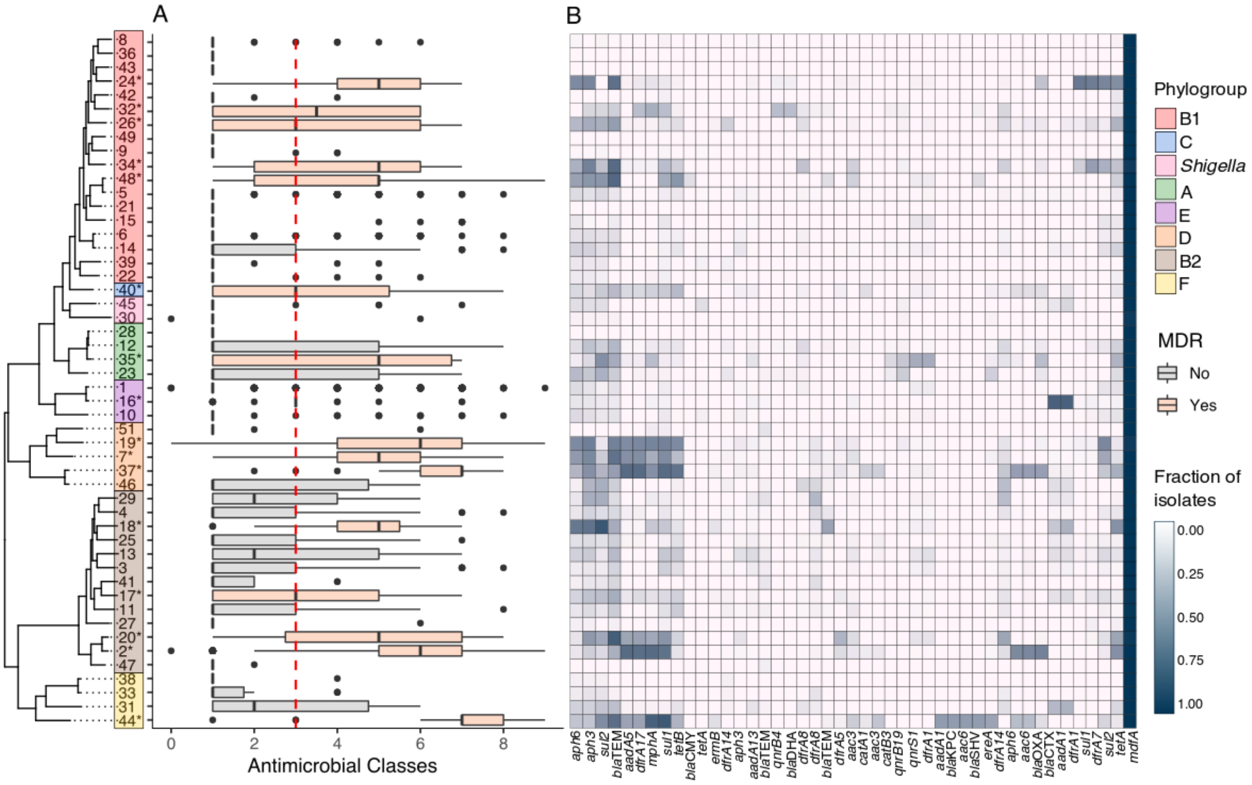

***Figure 4.11: Antimicrobial resistance profiles of the PopPUNK Clusters. A** Number of antimicrobial classes each isolate is resistant to, stratified by PopPUNK Cluster. Dashed red line indicates threshold for multidrug resistance. **B** Heatmap presenting the frequency of each resistance gene within each of the 50 PopPUNK Clusters. (Presenting only genes which were found in at least 10% of isolates of one PopPUNK Cluster.) Darker squares indicate higher prevalence of a gene in the PopPUNK Cluster. Phylogenetic tree constructed by selecting one isolate from each PopPUNK Cluster using Treemmer [392] (See Methods 4.3.4). Asterisk by PopPUNK Cluster name indicates MDR cluster.*

The presence and absence patterns of antibiotic resistance genes are presented in Figure 4.11B. Particular resistance genes are widespread in the dataset, these include *sul2* and *blaTEM*. Certain resistance gene combinations tended to co-occur multiple times in distantly related PopPUNK Clusters. For instance, resistance genes *aac6*, *blaOXA* and *blaCTX* co-occur in the MDR PopPUNK Clusters 20, 37 and 44. The genes *aadA1* and *dfrA1* are present together in PopPUNK Clusters 31, 17, 18 and 16. Finally, most of the resistance genes observed were in fact observed rarely and only present in very low frequencies in this dataset.

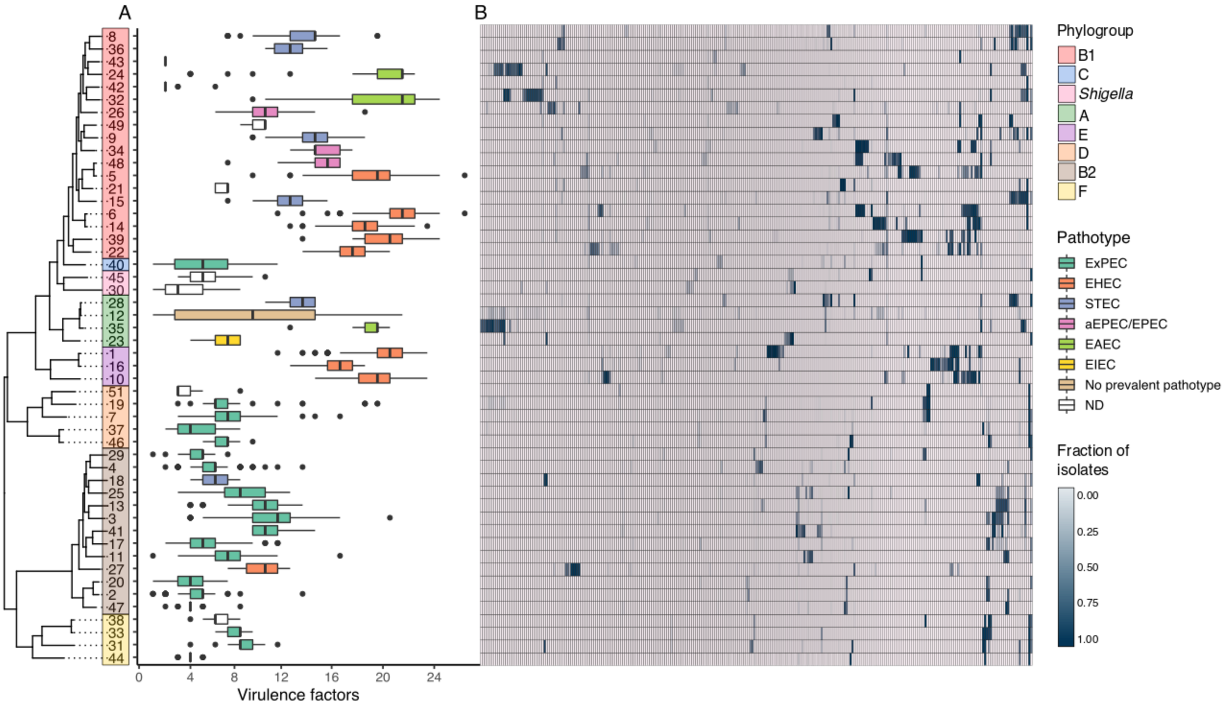

Figure 4.12: Markers of virulence in the PopPUNK Clusters. A *Number of virulence genes per isolate, stratified by PopPUNK cluster and coloured according to the most prevalent predicted pathotype in the cluster. ND = "Not Determined"* ***B*** *Heatmap presenting distribution of the virulence genes across the 50 PopPUNK clusters. Darker squares indicate higher prevalence of a gene in a lineage. (Presenting only genes which were found in at least 10% of isolates of one PopPUNK Cluster.) Phylogenetic tree constructed by selecting one isolate from each PopPUNK cluster using Treemmer [392] (See Methods XX).*

4.4.4.7 Markers of virulence

Consistent with the collection of *E. coli* isolates being from human hosts and mostly from clinical samples, 439 known virulence factors were observed in our dataset. The isolates had a median of nine known virulence factors in a single genome, with a maximum value of 26 virulence factors present in a single isolate (Figure 4.12A).

A combination of the source of isolation as well as the presence of key virulence factors were used to find the most prevalent predicted pathotype of each PopPUNK Cluster (See Section 4.3.7). 41 of 50 PopPUNK Clusters were identified as predominantly containing one of the defined *E. coli* pathotypes (See Section 1.1.2.2) (Figure 4.12A). Two of the PopPUNK Clusters without a prevalent pathotype were PopPUNK Clusters 30 and 45 which represent the *Shigella* species. PopPUNK Cluster 12, which mostly consists of *E. coli* isolates typing as ST10, was the only PopPUNK Cluster which contained isolates assigned to different pathotypes with no single pathotype dominating (11% ExPEC, 29% EAEC, 24% EPEC, 9% STEC, 2% EHEC, 1% ETEC, and 24% Not Determined (ND)). Indeed, PopPUNK Cluster 12 had the highest variability in number of virulence genes per isolate, relative to the rest of the clusters (Figure 4.12A). The remaining six PopPUNK Clusters which were not assigned a pathotype (21, 38, 42, 43, 45, 49 and 51) had relatively few virulence factors per isolate as well as low levels of predicted resistance, perhaps representing non-virulent lineages (Figure 4.12A).

Phylogroups B2, F, and D predominantly contained ExPEC isolates. PopPUNK Cluster 27 was the only cluster in phylogroup B2 which contained 67% EHEC isolates and 33% aEPEC/EPECs. PopPUNK Cluster 18, also nested within phylogroup B2 but not assigned a phylogroup according to the Clermont typing scheme, contained 100% STEC isolates. All PopPUNK clusters of phylogroup E contained predominantly EHEC isolates (Figure 4.12A, Appendix E). Phylogroups A and B1 had more diversity of pathotypes, containing PopPUNK Clusters which were assigned to the range of diarrheagenic pathotypes (EPEC, EHEC, EAEC and EIEC). PopPUNK Cluster 24 of phylogroup B1 also contained 38% isolates which were *stx* and *eae* positive. These are isolates of *E. coli* serotype O104:H4 taken from the 2011 German outbreak, which were classified as both shiga-toxin producing EAEC [407] (See Section 1.1.2.2). PopPUNK Cluster 40, the only cluster assigned to phylogroup C, was the only ExPEC cluster within the B1-C-A clade (Figure 4.12A).

The number of virulence factors per isolate differed between the phylogroups depending on their predominant pathotype (Figure 4.12A). Phylogroups containing ExPEC isolates (B2, D, F and C) had fewer virulence factors per isolate, relative to phylogroups containing the PopPUNK Clusters of the diarrheagenic *E. coli* (E, B1 and A). This could be a result of biases in the virulence factor database and our lack of complete understanding of ExPEC virulence factors.

The virulence factors identified in this dataset were more commonly specific to a PopPUNK Cluster and were generally not widespread across the whole dataset. PopPUNK Clusters which had a large number of virulence genes per isolate tended to possess a set of virulence factors which were otherwise not shared with other PopPUNK Clusters. This is exemplified in Figure 4.12B for PopPUNK Cluster 27, 10, 35 and more. Exceptions to this exist for virulence factors which were shared across PopPUNK Clusters which were assigned to the same pathotype, such as the ExPEC PopPUNK Clusters in Phylogroup B2 or the EHEC PopPUNK Clusters in phylogroups E and B1.

4.4.4.8 Relationship between resistance and virulence

The PopPUNK Clusters divided into clear groups based on their pathotype when comparing the median number of antimicrobial classes each isolate was resistant to against the median number of virulence factors identified per isolate for each PopPUNK Cluster (Figure 4.13). PopPUNK Clusters which were not assigned to a pathotype were resistant only to a single class of antimicrobials, i.e. these were predicted to be non-virulent and non-resistant. PopPUNK Clusters containing mostly ExPEC isolates ranged in the number of antimicrobial classes they were resistant to, with the most resistant PopPUNK Clusters, 2, 44 and 37, containing predominantly ExPEC isolates. However, more than half of the ExPEC PopPUNK Clusters (11/19) showed only low levels of resistance. Shiga-toxin producing isolates, EHECs and STECs, showed low levels of resistance relative to a high load of virulence factors. Exceptions to this were PopPUNK Clusters 16 and 18 which were the only MDR STEC and EHEC Clusters. PopPUNK Cluster 18 was particularly peculiar for an STEC as it is nested within phylogroup B2 and had low number of virulence factors per isolate relative to other STECs. PopPUNK Cluster which contained predominanly EAEC and EPEC isolates were all MDR and highly virulent.

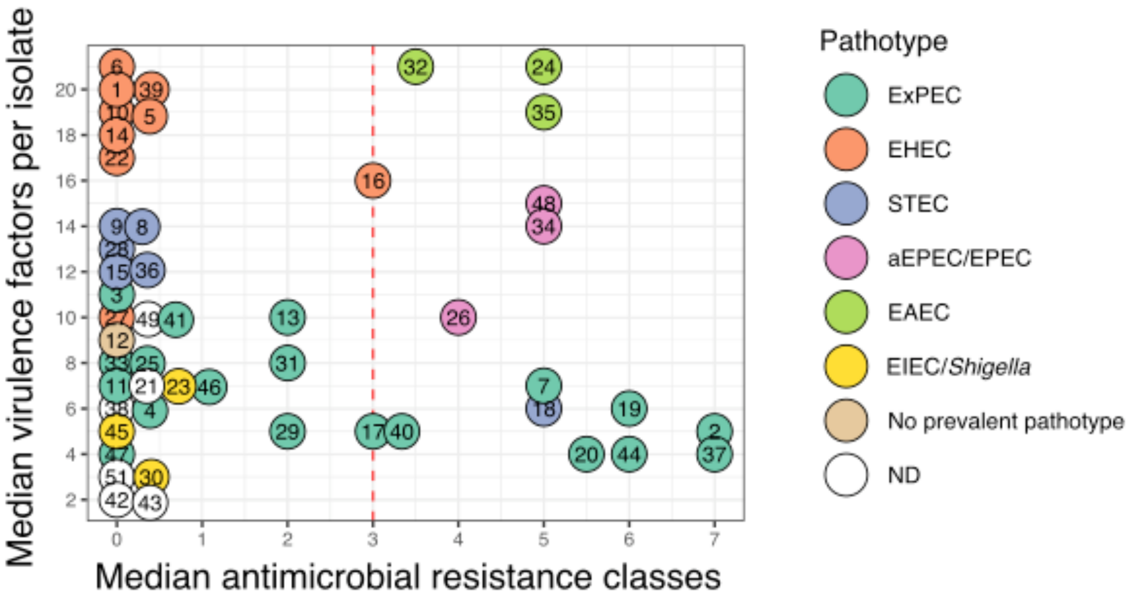

Figure 4.13: Relationship between resistance and virulence. *Each numbered dot represents a PopPUNK Cluster. Clusters are coloured by the most prevalent predicted pathotype in the cluster.*

4.4.4.9 Pan-genomes

A pan-genome analysis was applied on the isolates of each of the PopPUNK Clusters separately (See Section 4.3.8.1). Genes found within each PopPUNK cluster were divided into 4 categories based on their frequency within the cluster: genes present in more than 99% of isolates of a PopPUNK Cluster were labelled “core”, 95% to 99% of isolates were labelled “soft-core”, 15% to 95% of isolates labelled “intermediate” and “rare” were those present in fewer than 15% of isolates of a PopPUNK Cluster. The number of “core”, “soft-core” and “intermediate” genes in each PopPUNK cluster was stable across the clusters, regardless of the number of genomes in the cluster (Figure 4.10B,C). The number of “rare” genes per PopPUNK Cluster varied and was dependent on the cluster size, with larger PopPUNK clusters possessing more “rare” genes in their pan-genomes than smaller clusters (Figure 4.10C).

The pan-genome analysis on the PopPUNK clusters showed that there was low genetic diversity within PopPUNK clusters 21, 43 and 49. Therefore, these clusters were removed from the analysis, as they contain multiple isolates which were all collected at the same time and were all collected by the FDA (possibly representing an outbreak investigation).

4.4.5 Combining pan-genomes of the PopPUNK Clusters

Following the analysis of the pan-genome of each PopPUNK cluster individually, the outputs of all the analyses were combined in order to provide a description of the gene pool in the entire *E. coli* dataset analysed in this thesis. The precise steps taken are detailed in Section 4.3.7.2. Briefly, a reciprocal pairwise pan-genome analysis was run on every two PopPUNK clusters (Figure 4.2). The grouping of genes in every pairwise pan-genome analysis was examined to determine whether two genes from two separate PopPUNK clusters should be labelled as the same gene in the complete dataset. Since every pairwise comparison between genes was applied, it was possible to identify spurious matches between genes that were identified in single pan-genome analysis but were not supported across other pairwise gene comparisons. In addition, all representative sequences of a gene group were aligned and incorrect gene-groupings removed based on the SNP distances between the members.

4.4.6 Final collection of 55,039 genes

There were 55,039 genes (predicted CDSs) in the dataset after combining the genes of the pan-genomes of the 47 PopPUNK Clusters. As there were 47 PopPUNK clusters, and a varying number of isolates per cluster, each gene had its own frequency within each of the 47 PopPUNK Clusters. For instance, the *intA* gene, encoding a prophage integrase, was observed in 20 of the PopPUNK Clusters. In two clusters (6 and 9) it was present in over 95% of isolates, in another 8 clusters it was present in intermediate frequency (between 15% and 95%) and in the final 10 clusters it was present in fewer than 15% of isolates (A). In contrast, the gene *wzyE*, a gene involved in antigen biosynthesis, is a core gene which was observed across all PopPUNK Clusters in a frequency of over 95% (Figure 4.14B). Principal component analysis was applied to all gene frequencies across the PopPUNK Clusters (Figure 4.14C). The first and second principal components explained 17.93% and 7.49% of the variance and separated the PopPUNK clusters by the phylogeny.

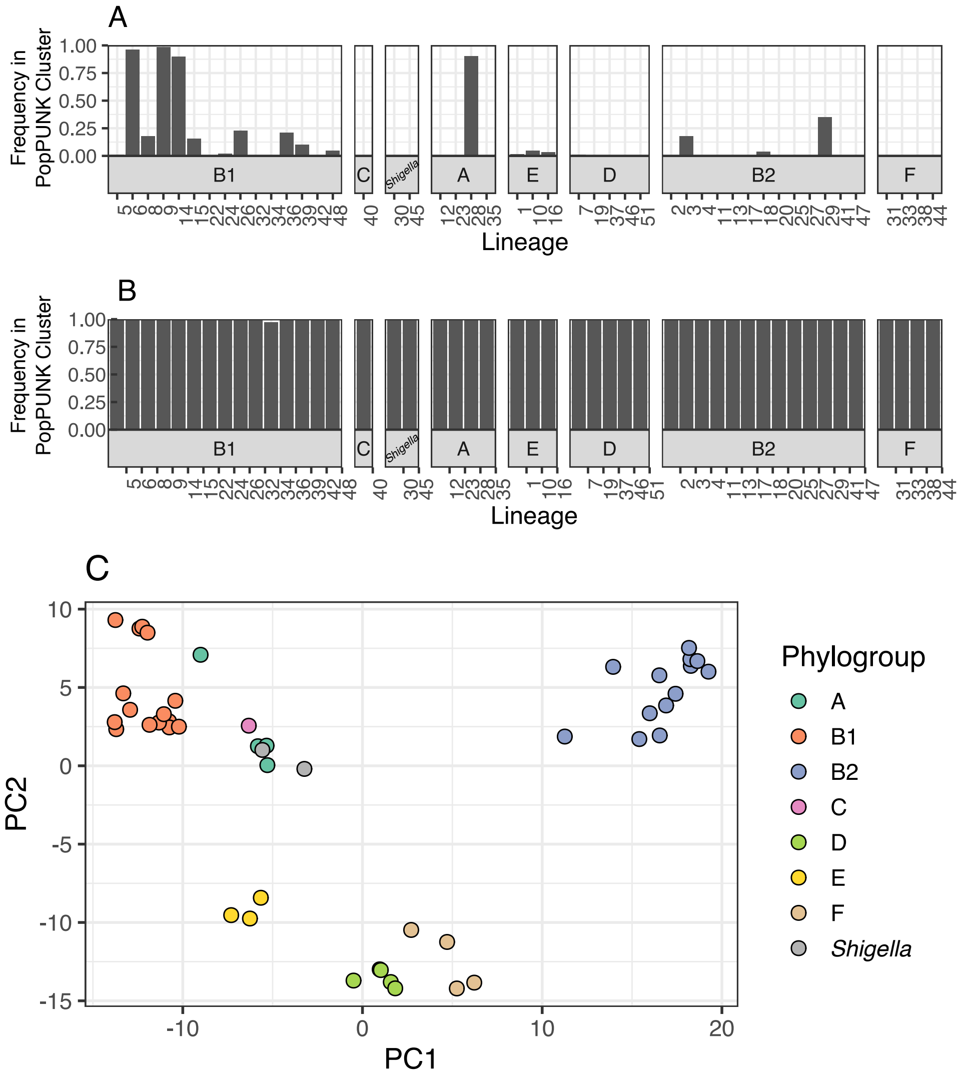

***Figure 4.14: Gene frequencies across the PopPUNK Clusters. A,B** Examples of the frequencies of two genes across the 47 PopPUNK Cluster, stratified by phylogroup.* intA *(**A**) is present in some PopPUNK Clusters and is found in different frequencies within them.* wzyE *(**B**) is core across all clusters. **C** PCA plot of the gene frequencies across all clusters.*

4.5 Discussion

The process of building and processing a high-quality dataset of thousands of *E. coli* genomes was described, along with the properties of the lineages that are present within the dataset and their gene (predicted CDS) content. The construction of this collection presented challenges in data accessibility, the scalability of existing tools and general biases in available sequencing data.

Aggregating data from diverse sources along with their associated metadata was a time-consuming effort. Genome identifiers and data formats across publications and databases do not always match leading to many conversions which are error prone and require knowledge of programming. In addition, computational resources are required in order to apply thousands of assembly and annotation calculations. These are all limiting factors to research. This emphasises the need to build new resources which maintain high quality genome collections where users would more easily be able to both retrieve and apply analysis on large collections. Without such resources, we have a mountain of information that is on the one hand available, but on the other hand practically unusable. Enterobase has proved to be one of these valuable resources, collating genomic data, providing assemblies and complete metadata tables for all genomes [93,400]. However, Enterobase currently only includes seven species.

The collection we obtained is biased towards *E. coli* lineages which have clinical significance. Not only that, the vast majority of genomes were available from Europe and North America, such that the pathotypes comprising the dataset are those which predominantly affect these areas. This was exacerbated by the fact that Enterobase's policy on data usage was ambiguous regarding the correct use of genomes which had been uploaded to public databases and have not yet published (or it is hard to confirm if they had been published). In the analysis presented here all genomes which were not taken directly from publications or from institutions from which approval was acquired were removed. This led to the removal of thousands of genomes. Finally, in the final collection, lineages or PopPUNK Clusters which had fewer than 20 isolates were also removed. Of the 1,185 PopPUNK Clusters, only 50 remained. This emphasises our lack of understanding of the true diversity of *E. coli* as a species. Hence, sampling should be increased in under-represented areas in the world as well as sampling of non-clinical isolates.

Existing tools were often designed to handle smaller collections or were not suitable for the analysis of a biased and diverse collection. Division of the dataset into groups of closely related isolates had been applied before when analysing diverse collections [408]. Indeed, Roary was designed to define the pan-genome of groups of closely related isolates, and thus was suitable when investigating the pan-genome of each PopPUNK Cluster [305]. However, an option to merge results of multiple pan-genome analyses had not been implemented and hence was built in this thesis. Additionally, Prokka, a commonly annotation tool, was not originally designed for genome comparison but rather for the annotation of a single genome [293]. A modified version of Prokka needed to be designed in order to remove artefacts when comparing multiple genomes. With more genomes, new methods need to be designed that are scalable when analysing diverse and large datasets.

Biological differences between the PopPUNK Clusters (lineages) were revealed from the initial investigation presented in this chapter. There were clear differences in the genome size between different phylogroups and PopPUNK Clusters. Higher variability in genome size with a phylogroup or PopPUNK Cluster could be an indication of higher rates of gene gain and loss within that cluster, as observed in phylogroup D. A larger genome size may also help to equip a lineage to survive in a range of niches as observed for PopPUNK Clusters of phylogroups E, F and B1 [4] (Figure 4.10A). Considering the major discrepancies in genome size between PopPUNK Clusters, it is interesting that the size of the core-genome across all the clusters is stable. This suggests that within a closely related group of genomes there is a specific number of genes, approximately 4,000 genes, that are required to define a lineage of closely related isolates (Figure 4.10B). The number of rare genes in a pan-genome was dependent on the cluster size, suggesting that the pan-genome of all lineages is open and is driven by continuous discovery of rare variants. A PCA plot of the gene frequencies as extracted from the complete dataset suggests that the phylogeny is driving the differences in gene content between the PopPUNK clusters. Questions regarding the distributions of different genes and the levels of gene sharing between the PopPUNK Clusters are further examined in Chapter 5 of this thesis.