6 Conclusions and Future Directions

K. pneumoniae and *E. coli* are clinically important organisms, which have highly diverse populations which include multiple co-circulating lineages across ecological niches and possess very large gene pools [9,115–117,120]. As such, there is ongoing emergence of novel virulent and multi-drug resistant lineages, led predominantly by acquisition of clinically relevant genes via HGT [74,421,422]. In this thesis, the gene pools of these two organisms were examined in order to answer questions regarding their diversity, the distribution of the genes across lineages and the level of gene sharing between lineages. Firstly, these questions were focused on the examination of the distribution of TA systems across a global collection of *K. pneumoniae* isolates. Secondly, the observations from the analysis on TA systems were expanded to look at the entire gene pool of a collection of 10,000 *E. coli* isolates. The availability of a large number of publicly available genomes, generated worldwide and in different settings was utilised to address these questions. The main conclusions and future directions are detailed in a point by point basis below.

6.1 Other use cases of SLING

In Chapter 2, SLING was presented as a tool that can be used to search for genes which are physically linked in large bacterial genomic datasets. Two use cases were presented. The first, to search for TA systems which represent a simple two-component operon, and the second, to search for RND efflux pumps which represent more complex operons where the order of the genes and operon structure vary across isolates. The usefulness of SLING in describing the diversity of these systems was illustrated. An analysis of 90 *E. coli* genomes revealed different distribution patterns of these operons and indicated that some genes undergo more gain and loss than others.

SLING was designed such that the search is not limited to the two use cases presented in Chapter 2. Searches for other important operons can be constructed as detailed on the SLING Wikipedia page (https://github.com/ghoresh11/sling/wiki/create_db). Possible searches could include examining the distributions of restriction-modification systems, secretion systems and CRISPR systems, all important in their contribution to HGT (See Section 1.2.2.1). Importantly, we showed that by applying SLING in Chapter 3 to identify TA systems in *K. pneumoniae*, we discovered novel antitoxins. Hence, applying SLING to search for other gene systems could lead to the discovery of other novel genes. Finally, SLING can further be used as a discovery tool by applying a reverse search (Figure 2.7). For instance, following the discovery of novel

antitoxins, novel antitoxin HMM profiles can be constructed as the query gene in SLING. The genes found in proximity in the new reverse search are candidate novel toxins.

6.2 Further exploration of the biological implications of toxin-antitoxin pairings, the genetic background of the host and their genetic context

Prior to the investigation presented on TA systems in *K. pneumoniae* in Chapter 3, these systems have mostly been studied on small scales in model organisms [253,335,347–350]. The analysis on a global clinical collection showed that while TA systems are all given the same term, the toxins of these systems can be classified based on their distribution patterns in the dataset, as well as based on the level of diversity of their antitoxin repertoire [345]. In addition to this, a high load of orphan antitoxins was observed in the genomes. Finally, some toxins were commonly associated with the presence of plasmid replicons, AMR genes or virulence genes.

The above conclusions require further experiments to explore their implications. These would include testing toxin and antitoxin combinations in order to understand the effect of these combinations on the functionality of the operon. This should be tested across different host genetic backgrounds, as our analysis found that some toxin-antitoxin pairings are specific to a species. The orphan antitoxins should be included in these experiments. These would include RNAseq experiments which would shed light on whether these systems, including the orphan antitoxins, are expressed in their hosts and under which conditions. This will elucidate whether the orphan antitoxins are functional antitoxins which can serve as a protective system against infection with other TA systems or otherwise, if their presence changes the functionality of other TA operons. The coding sequences in proximity to the orphan antitoxins should be further explored as candidates for discovery of novel toxins. Finally, it would be interesting to apply long-read sequencing on selected strains to identify the genetic context of the TA systems. This would shed light on whether ubiquitous or species associated toxins are chromosomally encoded, or whether they are present on plasmids which have persisted across the species, as well as whether toxins which were associated with clinically relevant genes are present on a plasmid with these genes and helping to maintain them.

6.3 Examination of TA systems on even larger scales

The conclusions presented in Chapter 3 regarding the distribution of TA systems were limited to an analysis on the global population of *K. pneumoniae*. The distribution of these systems across other species and genera is still unexplored. Early studies examining the distribution of TA systems across the entire European Nucleotide Archive (ENA) using BItsliced Genomic Signature Index (BIGSI), a tool which enables to query the ENA easily, in combination with SLING [311,423] have been set up. In this study, SLING was applied on over 3,000 genomes, strengthening its usability in searching for TA systems across large genomic datasets. This type of search is agnostic to species or genus boundaries and examines the distribution of these systems across thousands of genomes. This search could be further expanded to search for these systems in metagenomic datasets as well.

6.4 Therapeutic potential of TA systems

The analysis on TA systems presented in Chapter 3 revealed that TA systems are highly abundant in the *K. pneumoniae* species complex. The discovery of the range of TA systems present in clinical isolates can be used as a potential therapeutic against *K. pneumoniae* infections. This avenue of research is particulatly relevant to further explore in *K. pneumoniae* due to the increasing levels of multi-drug resistance and the inability to treat infections. Assuming these systems are expressed in the host, and that the expression of the toxin inhibits growth or leads to cell death in physiological conditions, new drugs, using peptides or small molecules, can be designed to inhibit the interaction between the toxin and the antitoxin [352]. The classification of the TA systems based on their distribution patterns, presented in Chapter 3, would lead to different outcomes based on the TA system being targeted. Targetting of ubiquitous or species associated TA systems would lead to growth arrest or death of all members of the species complex or one of the species. Alternatively, targettting the sporadic TA systems, which were found to be associated with the presence of AMR and virulence genes and plasmids, can be targeted to prevent the maintenance of these genes and thus lead to their loss. These differences between the TA systems highlight the need to better characterise the distribution of these systems across both clinical and non-clinical isolates in order to better understand their therapeutic potential.

6.5 More reliable databases and scalable tools are required

Through the research presented in this thesis, issues were raised regarding the accessibility of available genomic data and metadata, as well as the scalability of existing tools. While tools

have been developed to address the research questions presented in this thesis, they were either not applicable or not scalable to the size of the datasets used. Existing tools to search for TA systems could only be applied on a small number of genomes, and hence SLING was developed (Chapter 2) [311]. Prokka, the genome annotation tool, was not originally designed for comparative genomics, but rather for the annotation of a single genome (Chapter 4) [293]. The pan-genome analysis tool Roary, was designed to be applied on relatively clonal populations and had to be modified for the purpose of this thesis (Chapter 4) [305]. Furthermore, the data collection process was not straightforward and required programming skills and computational resources (Chapter 4).

The dataset of *E. coli* genomes presented in Chapters 4 and 5 should be made available for others to easily access without barriers of computational ability or resources. Ideally, this would be an online resource which enables users to query an *E. coli* genome in order to investigate its context in the *E. coli* pan-genome, for instance, by providing an R Shiny app [424]. Otherwise, a gene could be used as a query to investigate its distribution across the collection. This can provide the context required when working on a single gene system relative to *E. coli* clinical isolates.

6.6 More systematic sampling of under-represented *E. coli* lineages

The *E. coli* genome collection presented was limited to publicly available data, and hence was heavily biased towards clinical isolates causing disease in the developed world. The vast majority of isolates were collected from Europe and North America and include almost exclusively EHECs and ExPECs. This dataset does not represent natural *E. coli* populations nor does it represent the global clinical burden of *E. coli*, but rather it represents isolates that have been heavily sequenced due to their clinical significance where sequencing is available.

Systematic sampling of *E. coli* isolates is required to cover the full breadth of the *E. coli* diversity. This includes more sampling from under-represented areas of the world, as well as increased sampling of non-pathogenic isolates to better understand commensal *E. coli* populations. One possibility is to expand the research presented to include metagenomic assembled *E. coli* genomes, which could represent commensal populations [425]. Additionally, isolates from other hosts and environments beyond human isolates should be included. The analysis of these genomes can be compared to the dataset presented in this thesis to test whether the same gene distribution or lineages are observed across different

niches. When investigating gene movement, it is essential to include commensal *E. coli* and isolates from different environments as these could be facilitating the movement of genes between lineages and environments.

6.7 Further genomic analysis, as well as functional studies, to understand the differences and commonalities between *E. coli* lineages

In Chapter 5 of this thesis, it was revealed that a large part of the accessory genome is in fact comprised of genes which are core to one lineage or multiple lineages in the dataset. This emphasises the need to expand on traditional pan-genome analyses, as genes which on the surface are part of the accessory genome are core to part of the dataset. These genes are important both to better understand the evolution of the species, and in their potential in diagnostics.

The genes which were identified as core to a subset of the dataset should be further explored both *in-silico* and experimentally for their biological implications. A genomic analysis can be used to investigate genes which differ across the lineages and explore their predicted functions and what implications their presence and absence might have, followed up by functional experiments. The “multi-cluster core” genes should be explored as they represent the shifts in the core genome between *E. coli* lineages. “Multi-cluster core genes” which were acquired independently in multiple evolutionary events are interesting to explore as cases of parallel evolution. The “cluster specific core” should be further explored as they were only observed in a single lineage and were core to that lineage.

For instance, the analysis presented revealed that approximately 100 genes were gained on the branch leading to phylogroup B2, and approximately 70 genes were, on the other hand, lost on that branch. This could be a result of compensatory relationships between these genes, or otherwise, may reveal adaptation of this phylogroup to a particular niche which manifests in major changes in the core genome. Furthermore, the genetic context of these genes can be explored to better understand whether the shifts in the core genome occurred in a single evolutionary event which was beneficial and led to the expansion of this phylogroup, or alternatively, whether this was the the result of the accumulation of many changes, spread across the genome, which occurred over time.

The importance of further investigating these genes is particularly relevant for their potential use in diagnostics and epidemiology. Whole genome sequencing is often not available in clinical laboratories which need to identify the specific causative agent of an infection, nor are they always available during epidemiological investigations. The “cluster specific core” genes and “multi-cluster core” genes were only observed in specific lineages and were core to them, and therefore should be further investigated for their potential as marker genes to identify any of the lineages presented in this thesis using simple assays such as PCR. Importantly, the biased nature of the dataset presented is a major caveat to this potential. The lack of representation of most *E. coli* lineages means that we cannot rule out that the lineage specific genes identified here are not present in any other member of the species. We also cannot rule out that other members of a lineage, not sampled in this study, would possess the lineage specific genomes observed in this collection as sampling needs to be expanded to include other hosts and geographical locations. This emphasises the need to continue to apply similar analyses on much broader collections.

6.8 Examining the routes of movement of the shared low and intermediate frequency genes

Low frequency genes in the *E. coli* dataset were frequently gained and lost and their sharing was independent of the phylogenetic distance between lineages. A number of lineages were identified which shared more of these genes than expected, which were termed “hyper-sharers”.

The routes of movement of these genes should be further investigated. Understanding the precise routes of gene movement in the population would reveal and confirm the hypothesis presented that some lineages are facilitating the movement of genes in the population more than others. By understanding this, we could begin to tackle the problem of the introduction and propagation of novel resistance and virulence genes in the population. Understanding the complete routes by which genes are moving in the population is a harder problem to address, especially given that the dataset is biased and not densely sampled. While the bias in the dataset and under sampling cannot be resolved without more systematic sampling, this question can begin to be tackled in various ways. For instance, comparing gene-trees to species trees for the mobile genes could unravel whether the “hyper-sharers” are the source of many genes, or whether they are a hub and that genes in the dataset are passing through them. Additionally, the genetic context of these genes should be investigated, as a shared context would imply that genes are moving on the same element. Examining the co-

occurrence of these genes across the dataset could shed light on whether some of these genes are moving together as a unit. Furthermore, additional genomic analysis should be applied on the “hyper-sharers” to look at levels of recombination within these isolates, as well as presence of particular MGEs or genes which facilitate HGT that may be contributing to the observed property.

6.9 Further exploration of the rare and intermediate genes

A large proportion of the *E. coli* gene pool is represented by rare genes that were only observed in a single lineage and observed in a low frequency within that lineage. The function of most of these rare genes is unknown. The lineages termed “hyper-sharers” tended to have more of these rare genes within a single isolate genome relative to the other lineages in the dataset. Additionally, we found genes which were found in approximately 50% of isolates, across 50% of the lineages, representing genes which were truly intermediate across the collection. When we examined one such gene, it was a short protein with unknown function.

The origin and function of these rare and intermediate genes should be further explored. BIGSI can be used to search the ENA for these genes to see if they are present in other genera, in addition to searching for them in metagenomic samples. Computational approaches could be used to reveal more of their function, for instance via a “guilt-by-association” approach, followed by functional experiments. Furthering our understanding of these genes is highly relevant. First and foremost, they represent the majority of the *E. coli* gene pool and they have mostly been unexplored. Resistance and pathogenicity, as well as colonisation of different niches, are almost exclusively driven by the accessory genome in *E. coli*, therefore a better characterisation of the genes that make up the majority of the gene pool is highly relevant. As shown in Chapter 4, many AMR and virulence genes were observed in low or intermediate frequencies across the lineages (Figures 4.11, 4.12). This suggests that it is beneficial that only a fraction of the population would possess these genes as in this way the potential for their propagation exists under selective pressure, yet the metabolic burden of possessing them does not inhibit the growth of the whole population. Thus, understanding their function is essential in order to understand the full potential of the gene pool. Secondly, more of these genes were observed in isolates which tended to share more genes with other lineages- it is possible that these genes themselves are contributing to gene movement in the population.

To conclude, this thesis sets the basis for a range of future studies. These include examining TA systems experimentally to investigate the implications of the work presented here, or

otherwise to expand the search even more and examine these systems across more organisms. Within *E. coli*, the high-quality dataset presented sets an opportunity to address more questions regarding the movement of genes between lineages, the differences between the lineages and the function of these genes. This thesis sets a baseline to begin our understanding. With the availability of this resource to the broader scientific community, I hope that the future directions mentioned above and more will be addressed by us and others.