

# Redefining gene distributions in *K. pneumoniae* and *E. coli* using large public datasets



Gal Horesh  
Corpus Christi College, University of Cambridge

The dissertation is submitted for the degree of  
*Doctor of Philosophy*

June 2020



To my parents, who got me here,  
and to Harry, who walked with me every step of the way.



# Declaration

The work presented was carried out at the Wellcome Sanger Institute between October 2016 and June 2020. This work is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. This work is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

This thesis does not exceed the prescribed word limit specified by the Biology Degree Committee.



# Abstract

The work in this thesis is concerned with characterising genes and their distributions in *Escherichia coli* and *Klebsiella pneumoniae*. While both *K. pneumoniae* and *E. coli* are found in the guts of healthy individuals, as well as in animals and in the environment, they are particularly relevant organisms to study, as they represent key players in the dissemination of drug resistance and virulence in bacterial populations. Both organisms were given the highest priority by the World Health Organisation as organisms that pose the greatest threat to human health due to high levels of drug resistance. Additionally, they are both the leading cause of life-threatening extra-intestinal disease worldwide. Finally, some *E. coli* variants are also a major cause of severe diarrheal disease, most commonly in the developing world.

The phenomena that is driving these issues is horizontal gene transfer (HGT); the process by which new genetic material is introduced into a genome from an outside source. Drug resistance is most commonly driven by gene acquisition, and it is through the acquisition of virulence genes that *K. pneumoniae* and *E. coli* can cause disease. Indeed, HGT has been estimated to occur in high rates in *K. pneumoniae* and *E. coli*. Both are highly diverse organisms with very large gene pools and multiple co-circulating lineages. These facts make studying their gene pools on large scales highly relevant, as new genes and lineages are continuously discovered with the sequencing of new genomes.

The aim of this thesis was to utilise the availability of large public genomic datasets to study the gene pools of *K. pneumoniae* and *E. coli* on a scale and resolution not previously possible. Initially, the distribution of toxin-antitoxin (TA) systems was investigated in a collection of 259 *K. pneumoniae* isolates. TA systems are operons where one gene encodes for a toxin which inhibits a cellular process, and the other is an antitoxin which inhibits the toxin's activity. TA systems are relevant to study in the context of HGT as they have been shown to play a role in the maintenance of resistance and virulence genes and to contribute to antibiotic tolerance. The analysis on TA systems in *K. pneumoniae* revealed new insights regarding the distribution TA systems in the species. These insights were then expanded to examine the distribution of all genes of the *E. coli* gene pool in a collection of thousands of genomes. This analysis revealed that genes from different categories undergo different dynamics of gene gain and loss, as well as exposed *E. coli* lineages which may be important in their contribution to gene flow in the population. Due to the novelty and scope of the analyses presented, new computational tools and approaches were developed and are presented.

# Acknowledgements

I would like to thank my supervisors, Nick Thomson and Eva Heinz, who have guided, inspired and encouraged me throughout my PhD. In particular, to Nick for teaching me that every situation, no matter how bad, is an opportunity, and to Eva for convincing me that there are no problems in life, only challenges.

I am also very grateful to Leopold Parts, for his honest feedback, useful advice and consistent support. The other members of my thesis committee: Andres Floto, Simon Harris and Jukka Corander, have given me insightful discussions and constructive criticism. Julian Parkhill and Matt Berriman helped me shape this research by thoroughly questioning my PhD plans during my first year viva.

The work on toxin antitoxin systems would not have been possible without the contributions from my collaborators: Cinzia Fino, Alexander Harms and Kenn Gerdes. Cinzia worked extremely hard on all of the projects that we collaborated on. She was a great student and taught how much I love to teach. Matthew Dorman helped to coordinate the experiments and performed some experimental work himself. Additionally, he has been my considerate office neighbour and my PhD-cohort companion for the last three years.

Other collaborations, beyond the scope of this thesis, have furthered my learning and scientific progression. Gerry Tonkin-Hill and Neil MacAlasdair involved me in the Panaroo project and helped me with the pan-genome analysis. Grace Blackwell and Zam Iqbal entrusted me with the large scale transposon study which has been a good experience.

The work I have achieved would not have been possible on this scale without the help of the Pathogen Informatics team as well as the members of the Sanger Service Desk. I would particularly like to thank Martin Hunt who helped me to build SLING into a package.

For helping me get through my PhD with a smile on my face, with well needed walks and microwave chats, I would like to thank Alex Wailan, Kate Mellor, Alyce Tyler-Brown, Aline Cuenod, Ha My Pham, Sushmita Sridhar, Grace Blackwell and the rest of the members of Team 216 along the years. I am happy to say I have made some good friends along the way.

Last but not least, I would like to thank my partner, Harry Scholes, for unwavering support and all the work-related dinner conversations along the way. My PhD would not have been the same without him.

Finally, I would also like to extend my gratitude to Wellcome for funding my PhD.



# Publications

SLING: a tool to search for linked genes in bacterial datasets, Horesh et al., *Nucleic Acids Research*, 2018, <https://doi.org/10.1093/nar/gky738>

Type II and type IV toxin–antitoxin systems show different evolutionary patterns in the global *Klebsiella pneumoniae* population, Horesh et al., *Nucleic Acids Research*, 2020, <https://doi.org/10.1093/nar/gkaa198>

A comprehensive and high-quality collection of *E. coli* genomes and their genes, Horesh et al., in preparation

A pan-genome analysis of 10,000 *E. coli* genomes reveals new patterns of gene sharing between lineages, Horesh et al., *in preparation*

Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline, Tonkin-Hill et al., *bioRxiv*, 2020, <https://doi.org/10.1101/2020.01.28.922989>

The distribution of toxins containing Gp49 across Gram-negative bacteria, Fino et al., *in preparation*

Horizontal and vertical spread of Tn1-related transposons in Gram-negative bacteria, Blackwell et al., *in preparation*

# Table of Contents

<b>1 Introduction</b>	<b>1</b>
1.1 The organisms: <i>E. coli</i> and <i>K. pneumoniae</i>	1
1.1.1 The species <i>K. pneumoniae</i>	1
1.1.1.1 Taxonomy and classification	1
1.1.1.2 Pathogenicity and resistance	3
1.1.1.3 Genetics	6
1.1.2 The species <i>E. coli</i>	9
1.1.2.1 Taxonomy and classification	9
1.1.2.2 Pathogenicity and resistance	11
1.1.2.3 Genetics	13
1.2 The phenomena: Horizontal gene transfer	17
1.2.1 Mechanisms of HGT	17
1.2.1.1 Inter-cellular mobility	17
1.2.1.2 Intra-cellular mobility	19
1.2.2 Barriers to HGT	20
1.2.2.1 Genetic barriers	20
1.2.2.2 Physical barriers	21
1.2.3 HGT in <i>K. pneumoniae</i> and <i>E. coli</i>	21
1.2.4 Contribution of HGT to virulence and resistance	23
1.3 The genes: Toxin antitoxin systems	25
1.3.1 Classification	25
1.3.2 Mechanisms	26
1.3.3 Role in resistance and pathogenicity	27
1.4 The approach: comparative genomics using public databases	28
1.4.1 Methods for comparative genomics	30
1.4.1.1 Defining the population structure	30
1.4.1.2 Methods for gene detection	31
1.4.1.3 Grouping homologous sequences	32

1.4.1.4 Pan-genome analysis	33
1.5 Thesis outline	33
<b>2 SLING: A tool to Search for LINKed Genes in bacterial datasets</b>	<b>35</b>
2.1 Introduction	35
2.2 Aims	36
2.3 Methods	36
2.3.1 SLING specifications	36
2.3.2 Strains and phylogenetic analysis	39
2.4 Results	39
2.4.1 SLING overview	39
2.4.2 TA systems search	39
2.4.2.1 Construction of profile HMM library and structural requirements	40
2.4.2.2 The process for setting up a TA search are applicable to other operons	43
2.4.3.3 Benchmark on <i>E. coli</i> K-12	43
2.4.2.4 Application on EPEC collection	44
2.4.3 RND efflux pumps search	47
2.4.3.1 Construction of profile HMM library and structural requirements	47
2.4.3.2 Benchmark on <i>E. coli</i> K-12	49
2.4.3.4 Application on EPEC collection	49
2.5 Discussion	50
<b>3 The diversity of type II and type IV toxin-antitoxin systems in the global <i>K. pneumoniae</i> population</b>	<b>53</b>
3.1 Introduction	53
3.2 Aims	54
3.3 Methods	54
3.3.1 Strains and phylogenetic analysis	54
3.3.2 Toxin-antitoxin prediction	55
3.3.3 Statistical analysis	55
3.3.4 Toxin group classification	55
3.3.5 Definition of novel vs known antitoxins	57

3.3.6 Orphan antitoxins	57
3.3.7 Identification of AMR genes, virulence genes and plasmid replicons	58
3.3.8 Phenotypic testing	58
3.4 Results	61
3.4.1. Type II and type IV TA systems are highly abundant in the <i>K. pneumoniae</i> species complex	61
3.4.2 Redefining toxins based on their distribution patterns	65
3.4.3 Prediction of novel antitoxins	66
3.4.4 Fluid association and distribution of toxin-antitoxin pairings	68
3.4.5 Phenotypic testing <i>in silico</i> predictions of toxins and confirmation of novel antitoxins	70
3.4.6 Orphan antitoxins are abundant in the dataset	74
3.4.7 The association between toxins and antimicrobial resistance genes, virulence genes or plasmid replicons	76
3.5 Discussion	78
<b>4 Building a collection of 10,000 <i>E. coli</i> isolates and defining the gene content in the collection</b>	<b>81</b>
4.1 Introduction	81
4.2 Aims	82
4.3 Methods	82
4.3.1 Data collection	82
4.3.1.1 Reads	83
4.3.1.2 Assemblies	84
4.3.1.3 Gene calling	84
4.3.2 MLST	84
4.3.3 Genome Clustering using PopPUNK	84
4.3.4 Phylogenetic analysis	85
4.3.5 Phylogroup assignment	86
4.3.6 Identification of AMR and virulence genes	86
4.3.7 Pathotype assignments	86
4.3.8 Pan-genome analysis	86

4.3.8.1 Pan-genome analysis on each PopPUNK cluster	86
4.3.8.2 Combining the pan-genomes of all PopPUNK Clusters	87
4.3.9 Statistical analysis	88
4.4 Results	88
4.4.1 Constructing a collection of 10,000 <i>E. coli</i> isolates	88
4.4.1.1 Initial collection of 18,156 genomes	88
4.4.1.2 Modifying the annotation tool PROKKA to remove errors in gene calling between genomes	90
4.4.1.3 Filtering to a high-quality collection of 10,159 genomes	91
4.4.2 Characteristics of the filtered dataset	93
4.4.2.1 Most of the genomes are from developed countries, collected in surveillance in clinical settings	93
4.4.2.2 Only 5% of all genomes are the cause of diarrheal disease in developing countries	94
4.4.2.3 Six STs represent more than 50% of the genomes in the collection	94
4.4.3 PopPUNK can be used to group the collection into isolates belonging to the same lineage	96
4.4.4 Characteristics of the selected 50 largest PopPUNK Clusters	97
4.4.4.1 Genetic diversity	97
4.4.4.2 Population structure	97
4.4.4.3 Pathogenic and geographic association	99
4.4.4.4 Sampling time	100
4.4.4.5 Genome size and number of predicted genes	101
4.4.4.6 Antimicrobial resistance profiles	102
4.4.4.7 Markers of virulence	104
4.4.4.8 Relationship between resistance and virulence	106
4.4.4.9 Pan-genomes	107
4.4.5 Combining pan-genomes of the PopPUNK Clusters	107
4.4.6 Final collection of 55,039 genes	107
4.5 Discussion	109

<b>5 Redefining the <i>E. coli</i> pan-genome reveals new patterns of gene gain/loss and gene sharing between lineages</b>	<b>111</b>
5.1 Introduction	111
5.2 Aims	112
5.3 Methods	112
5.3.1 Gene classification into “occurrence classes”	112
5.3.2 Measuring the genetic composition of each PopPUNK Cluster	113
5.3.3 Phylogenetic analysis	113
5.3.3.1 Phylogenetic tree construction	113
5.3.3.2 Phylogenetic distance calculations	115
5.3.3.3 Ancestral state reconstruction	115
5.3.3.4 Counting gain and loss events	115
5.3.4 Functional assignment of COG categories	115
5.3.5 Identifying gene variants	116
5.3.6 Gene property calculations	116
5.3.7 Statistical analysis	116
5.4 Results	117
5.4.1 A novel approach for examining the <i>E. coli</i> pan-genome	117
5.4.2 The typical composition of an <i>E. coli</i> genome	119
5.4.3 Rates of gene gain and loss differ across the occurrence classes	120
5.4.4 “Multi-cluster core” genes represent the shifts in core genome of <i>E. coli</i> clades	122
5.4.5 “Core and intermediate” represent the “soft-core” genome	125
5.4.6 “Multi-cluster intermediate” genes are shared between closely related PopPUNK Clusters, but have different functional profiles to the “core” genes	126
5.4.7 Low frequency genes are gained and lost at high rates, and their sharing is independent of the phylogeny	128
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	130

5.4.9 Hyper-sharing PopPUNK Clusters possess more “cluster specific rare” genes in a single genome relative to the rest of the clusters	131
5.4.10 PopPUNK Clusters which shared fewer low frequency genes than expected also had the largest number of “cluster specific core” genes	132
5.4.11 Cluster specific core genes are often truncated variants of other genes in the collection	132
5.4.12 STEC PopPUNK Cluster 27 and ExPEC PopPUNK Cluster 44 possess a large number of “cluster specific intermediate” genes.	133
5.5 Discussion	133
<b>6 Conclusions and Future Directions</b>	<b>138</b>
6.1 Other use cases of SLING	138
6.2 Further exploration of the biological implications of toxin-antitoxin pairings, the genetic background of the host and their genetic context	139
6.3 Examination of TA systems on even larger scales	140
6.4 Therapeutic potential of TA systems	140
6.5 More reliable databases and scalable tools are required	140
6.6 More systematic sampling of under-represented <i>E. coli</i> lineages	141
6.7 Further genomic analysis, as well as functional studies to understand the differences and commonalities between <i>E. coli</i> lineages	142
6.8 Examining the routes of movement of the shared low frequency genes	143
6.9 Further exploration of the rare genes	144
<b>References</b>	<b>146</b>
<b>Appendix</b>	<b>174</b>
A Strains, plasmids and oligonucleotides used in this study	174
B Identified toxin groups	177
C Identified antitoxin groups	185
D Identified orphan antitoxins	198
E Summary of <i>E. coli</i> PopPUNK Clusters	203

# List of Figures

## Chapter 1

1.1 Incidence of bloodstream infections caused by eight major pathogens in England.	4
1.2 Pan-genome definition.	7
1.3 Population structure of <i>E. coli</i> .	10
1.4 Decision network of the virulence factors defining the <i>E. coli</i> pathotypes.	16
1.5 Main mechanisms of HGT.	18
1.6 Types of TA systems.	26
1.7 Number of bacterial and archeal genomes released each year on NCBI.	29

## Chapter 2

2.1 Overview of the SLING pipeline.	38
2.2 Defining the HMM collection and structural requirements for toxins.	41
2.3 General construction of HMM profiles and structural requirements for SLING input.	43
2.4 Identification of TA systems using SLING.	45
2.5 Identification of RND efflux pumps using SLING.	48
2.6 Defining the HMM collection and structural requirements for RND efflux pumps.	49
2.7 Utility of SLING.	51

## Chapter 3

3.1 Effect of modifying the blastp identity threshold in SLING on the toxin group clustering.	56
3.2 Diversity of toxins in <i>K. pneumoniae</i> species complex.	62
3.3 Number of unique toxin groups for each of the toxin Pfam profiles used in the search.	63
3.4 Example of diversity of toxins containing a HicA toxin Pfam profile domain.	64
3.5 Nucleotide identity of toxins within and between species.	65
3.6 Copy number of species-associated toxins.	67
3.7 Identification of novel antitoxins in the <i>K. pneumoniae</i> genomes.	68
3.8 Diversity in the observed operon structures for the different toxin categories.	70
3.9 Phenotypic testing of selected toxins.	72



3.10 Phenotypic testing of predicted toxin-antitoxin combinations.	73
3.11 Orphan antitoxins in <i>K. pneumoniae</i> genomes.	75
3.12 Toxin groups associated with AMR genes, virulence genes and plasmid replicons	77
<b>Chapter 4</b>	
4.1 Workflow for collating the <i>E. coli</i> genome collection.	83
4.2 Method for combining the pan-genome analysis of all PopPUNK Clusters.	89
4.3 Effect of modifying Prokka on the CDS prediction.	91
4.4 Quality control measures used to filter <i>E. coli</i> genomes.	92
4.5 Source of <i>E. coli</i> genomes.	94
4.6 Distribution of STs and PopPUNK Clusters in the collection.	95
4.7 PopPUNK Clusters' genetic diversity.	98
4.8 Population structure of the PopPUNK Clusters.	99
4.9 Metadata associated with the PopPUNK Clusters.	100
4.10 Gene content in the 50 PopPUNK Clusters.	101
4.11 Antimicrobial resistance profiles of the PopPUNK Clusters.	103
4.12 Markers of virulence in the PopPUNK Clusters.	104
4.13 Relationship between resistance and virulence.	106
4.14 Gene frequencies across the PopPUNK Clusters.	108
<b>Chapter 5</b>	
5.1 Gene classification into occurrence classes.	114
5.2 Distribution of the <i>E. coli</i> gene-pool based on the rules defined.	118
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	120
5.4 Gain and loss events per gene.	121
5.5 Gain and loss events per branch.	123
5.6 Properties of high frequency genes in the <i>E. coli</i> dataset.	124
5.7 Fraction of genes from each occurrence class which were assigned each of the COG categories.	127
5.8 Properties of low frequency genes in the <i>E. coli</i> dataset.	129
5.9 Cluster specific genes in the <i>E. coli</i> dataset.	131

# List of Tables

## **Chapter 1**

1.1 <i>K. pneumoniae</i> species and subspecies.	3
--	---

## **Chapter 2**

2.1 Search parameters used in SLING.	42
--------------------------------------	----

## **Chapter 3**

3.1 Phenotypic testing of identified toxins.	59
--	----

3.2 Combinations of toxin-antitoxins tested for antitoxin inhibition.	60
---	----

## **Chapter 4**

4.1 PopPUNK Clustering statistics.	85
------------------------------------	----

# Glossary

<b>aa</b>	amino acids
<b>ACCTRAN</b>	Accelerated Transformation
<b>aEPEC</b>	atypical Enteropathogenic <i>E. coli</i>
<b>AIEC</b>	Adherent Invasive <i>E. coli</i>
<b>AMR</b>	Antimicrobial Resistance
<b>ANI</b>	Average Nucleotide Identity
<b>BLAST</b>	Basic Local Alignment Search Tool
<b>bp</b>	basepairs
<b>BSI</b>	Bloodstream Infection
<b>CDC</b>	Centers for Disease Control and Prevention
<b>CDS</b>	Coding Sequence
<b>COG</b>	Clusters of Orthologous Groups
<b>contig</b>	contiguous assembled sequence
<b>DAEC</b>	Diffusely Adherent <i>E. coli</i>
<b>EAEC</b>	Enterogaggaragive <i>E. coli</i>
<b>ECOR</b>	<i>E. coli</i> Reference Collection
<b>EHEC</b>	Enterohaemorrhagic <i>E. coli</i>
<b>EIEC</b>	Enteroinvasive <i>E. coli</i>
<b>ENA</b>	European Nucleotide Archive
<b>EPEC</b>	Enteropathogenic <i>E. coli</i>
<b>ESBL</b>	Extended Spectrum Beta Lactams
<b>ETEC</b>	Enterotoxigenic <i>E. coli</i>
<b>FDA</b>	Food and Drug Administration
<b>FDR</b>	False Discovery Rate
<b>GEMS</b>	The Global Enteric Multicenter Study
<b>HGT</b>	Horizontal Gene Transfer
<b>HMM</b>	Hidden Markov Model
<b>HUS</b>	Hemolytic uremic syndrome
<b>hvKp</b>	hyper-virulent <i>K. pneumoniae</i>
<b>ICE</b>	Integrative and Conjugative Elements
<b>IncA/C</b>	Plasmid incompatibility type A/C
<b>IPTG</b>	isopropyl $\beta$ -D-thiogalactopyranoside
<b>LB</b>	Lysogeny Broth
<b>LEE</b>	Locus of Enterocyte Effacement

<b>Mbp</b>	Million basepairs
<b>MDR</b>	Multidrug resistant
<b>MFP</b>	Membrane Fusion Protein
<b>MGE</b>	Mobile Genetic Element
<b>MLST</b>	Multi-locus Sequence Type
<b>MSA</b>	Multiple Sequence Alignment
<b>NCBI</b>	National Center for Biotechnology Information
<b>ND</b>	Not Determined
<b>OMF</b>	Outer Membrane Protein
<b>PBS</b>	phosphate-buffered saline
<b>PCA</b>	Principal Component Analysis
<b>PCR</b>	Polymerase Chain Reaction
<b>PHE</b>	Public Health England
<b>PopPUNK</b>	Population Partitioning Using Nucleotide K-mers
<b>PSK</b>	Post Segregational Killing
<b>QC</b>	Quality Control
<b>RND</b>	Resistance-Nodulation-Division
<b>SNP</b>	Single Nucleotide Polymorphism
<b>SSN</b>	Sequence Similarity Network
<b>ST</b>	Sequence Type
<b>ST10</b>	Assigned to Sequence Type 10
<b>STEC</b>	Shiga toxin-producing <i>E. coli</i>
<b>TA</b>	Toxin Antitoxin
<b>TADB</b>	Toxin Antitoxin Database
<b>UTI</b>	Urinary Tract Infection
<b>VTEC</b>	Verotoxigenic <i>E. coli</i>
<b>WGS</b>	Whole Genome Sequencing
<b>WHO</b>	World Health Organisation