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 Corrander, 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

List of Figures

List of Tables

Glossary

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

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