University of Melbourne

A genomic framework for enhanced strain identification: to improve outbreak detection and public health surveillance using *Enterobacter cloacae* complex as an exemplar.

by

Susan Noonan
Student No: 1241643

A thesis submitted in fulfillment for the
degree of Master of Science - Bioinformatics
in the
Department of Microbiology and Immunology
The University of Melbourne
The Peter Doherty Institute for Infection and Immunity

Project Supervisors

Dr. Jake Lacey, Dr. Kristy Horan, Prof. Benjamin Howden

June 2024
word count: 13,224

Abstract

Enterobacter cloacae complex (ECC) is a group of opportunistic, nosocomial bacteria that account for 5% of Australian hospital acquired infections. ECC is increasingly showing resistance to carbapenems, a last line treatment. Inconsistency and difficulties in species identification (using MALDI-TOF) may lead to hospital outbreaks going undetected, as confidence in species identification is reduced. Communication of results between laboratories, health staff and public health officials becomes complicated when inaccurate naming occurs. To address this, our study developed a genomic framework to assist confidence in species and strain allocation within ECC.

Utilising Split kmer analysis (SKA) and Mash distance to assess pairwise genomic relatedness, this framework groups isolates into species and subspecies through stepwise distance thresholds of increasing similarity. Species groupings were supported through phylogenetic and pan-genome analyses. Following the determination of species boundaries, a comparison of multi-locus sequence type (MLST) and antimicrobial resistance (AMR) profiles was conducted.

Species boundaries are clearly defined when a Mash threshold ≤ 0.04 is applied and subspecies are differentiated at ≤ 0.02. *Enterobacter cloacae* isolates were often misnamed and found to belong in other species groups. MLST sequence types are unique within species groups, therefore MLST can be employed as a rapid tool to identify a species when whole genome sequencing (WGS) is not available. Antimicrobial resistance (AMR) genes are widespread across the *Enterobacter spp.* isolates analysed.

There is good evidence to suggest many *Enterobacter spp.* in public genome repositories are misidentified. This may contribute to delays and complications in detecting hospital outbreaks and inaccuracies when utilising public databases. MLST can be used diagnostically to determine species in *Enterobacter*. The diversity and prevalence of AMR genes across *Enterobacter* isolates investigated is concerning. The genomic framework developed can be applied to any pathogen where the distinguishing features are not well defined, or the reference database is contaminated.

Declaration

I, Susan Noonan, declare that this thesis titled, **"A genomic framework for enhanced strain identification: to improve outbreak detection and public health surveillance using *Enterobacter cloacae* complex as an exemplar."** and the work presented in it are my own.

I confirm that:

- the research report comprises only my original work towards the Master of Science - Bioinformatics (MC-SCIBIF) except where indicated in the text;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is fewer than the maximum word limit in length, exclusive of tables, maps, references and appendices as approved by the Academic Board.

Signed: _____

Date: 03 June 2024

Preface

The following details the collaborators involved in this research project:

- Dr Jake Lacey and Dr Kristy Horan generated the project research idea and outlined the project aims.
- Assistance downloading the genomes used in this project, advice on research methodology and interpretation was provided by Dr Jake Lacey;
- No work towards this thesis has been submitted for other qualifications or completed prior to enrolment in the degree;
- In preparation of this thesis, editorial assistance was provided by Dr Jake Lacey, Dr Kristy Horan and Professor Benjamin Howden, all of whom are knowledgeable in the discipline of the pathogen genomics;
- All chapters presented in this thesis are unpublished material not currently submitted for publication.

Acknowledgements

First, thanks go to my supervisors, Jake, Kristy, and Ben, for offering me a research project in the Howden Lab in the field of pathogen genomics.

Jake has been the best supervisor and mentor: clever, patient, understanding and encouraging from the outset. With unfailing generosity he answered my questions, solved my code issues, shared his knowledge, and flamed my enthusiasm. I feel extremely lucky to have worked alongside Jake on this project, and enjoyed learning while developing new skills.

My thanks to Kristy for her generous explanations and for validating how challenging speciation can be.

Torsten has always been approachable and willing to share his vast knowledge. His help was greatly appreciated.

CIS team at MDU provided invaluable early assistance solving code problems in the early phase of my research.

Fellow Bioinformatics students have been excellent companions on this journey. I look forward to working with them on future projects.

Hilary encouraged, gave moral support, and was willing to listen. She helped me with editing and expressing my research in plain English. Hilary also designed the beautiful logo for my script.

Boomer has been my canine office companion, who reliably reminded me to get up and leave the desk.

Talking through problem-solving approaches with my two sons and husband was often helpful, especially when completing my analysis and writing code. I have appreciated their support and understanding as I juggled study and family commitments.

Contents

List of Figures

List of Tables

Abbreviations

AGAR	**A**ustralian **G**roup on **A**ntimicrobial **R**esistance
AMR	**A**nti **M**icrobial **R**esistance
ANI	**A**verage **N**ucleotide **I**dentity
BANCSIA	**B**acterial **N**aming for **C**orrect **S**pecies **I**dentification & **A**llocation
CCU	**C**ritical **C**are **U**nit
CDC	**C**entres for **D**isease **C**ontrol
CPE	**C**arbapenemase **P**roducing **E**nterobacterales
CRE	**C**arbepenam **R**esistant **E**nterobacteriaceae
.csv	**C**omma **S**eparated **V**alue
ECC	**E**nterobacter **C**loacae **C**omplex
ESBL	**E**xtended **S**pectrum **B**eta **L**actamase
.gff	**G**eneral **F**eature **F**ormat
GARDP	**G**lobal **A**ntibiotic **R**esearch and **D**evelopment **P**artnership
GML	**G**raph **M**odelling **L**anguage
GTDB	**G**enome **T**axonomy **D**ata **B**ase
HAI	**H**ospital **A**ssociated **I**nfections
HGT	**H**orizontal **G**ene **T**ransfer
HIV	**H**uman **I**mmunodeficiency **V**irus
ICNP	**I**nternational **C**ode of **N**omenclature of **P**rokaryotes
IQR	**I**nter **Q**uartile **R**ange
ISO	**I**nternational **S**tandarisation **O**rganisation
ICU	**I**ntensive **C**are **U**nit
MALDI-TOF	**M**atrix **A**ssisted **L**aser **D**esorption **I**onisation- **T**ime **O**f **F**light
MDR	**M**ulti **D**rug **R**esistant
MDU PHL	**M**icrobiological **D**iagnsotic **U**nit **P**ublic **H**ealth **L**aboratory
MLST	**M**ulti **l**ocus **S**equence **T**yping
MPTF	**M**ulti **P**artner **T**rust **F**und
MRSA	**M**ethicillin **R**esistant **S**taphylococcus **A**ureus
NHSN	**N**ational **H**ealthcare **S**afety **N**etwork
SKA	**S**plit **K**mer **A**nalysis
SNP	**S**ingle **N**ucleotide **P**olymorphism
ST	**S**equence **T**yping
UTI	**U**rinary **T**ract **I**nfection
WGS	**W**hole **G**enome **S**equencing
WHO	**W**orld **H**ealth **O**rganisation

Chapter 1

Introduction

1.1 Background

"*Antimicrobial resistance (AMR) is a global health and development threat. Misuse and overuse of antimicrobials are the main drivers in the development of drug-resistant pathogens*" (World Health Organisation, 2021)

AMR pathogens are microorganisms such as bacteria, viruses, fungi, and parasites that have altered over time enabling them to survive when exposed to antimicrobials. Antimicrobials are medicines used to prevent and treat illness caused by these organisms in humans, animals, and plants. The World Health Organisation (WHO) describe misuse of antimicrobials as a major cause of AMR (World Health Organisation, 2021). A direct relationship between antibiotic use and the emergence and spread of resistant pathogen strains has been demonstrated by epidemiological studies (Ventola, 2015). Figure 1.1 outlines the time association between an antibiotics being introduced and when resistance has been reported in literature (CDC, U.S Department of Health and Human services, 2013). Drug resistance describes when a medicine is no longer effective, and infections are more difficult or impossible to treat.

Bacteria with AMR genes are found ubiquitously. AMR genes can be inherited via vertical transfer from a parent, or via horizontal transfer from related and non-related organisms that contain mobile genetic elements like plasmids or through mutation. The burden of bacterial AMR is an increasing global concern. In 2019, 1.27 million deaths were directly related to AMR, which is greater than malaria and HIV combined (Murray et al., 2022).

Australasia is the region with the lowest AMR burden rate in the world (Murray et al., 2022). However, with rates increasing globally, costs associated with AMR will increase. The speed and ease of international travel means there is a high risk of AMR isolates spreading to Australia. A detailed investigation by Wozniak et al., 2022 demonstrated

the five most common AMR pathogens in Australian hospitals include, *Enterococcus spp.*, *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudamonas aeruginosa* and *Staphylococcus aureus*. (Wozniak et al., 2022). There were over 21,000 infections and 1000 deaths attributed to AMR pathogens in 2020, and an additional 45,876 hospital bed days taken up by patients with an AMR infections (Wozniak et al., 2022). This is a significant cost to both individuals and the health care service.

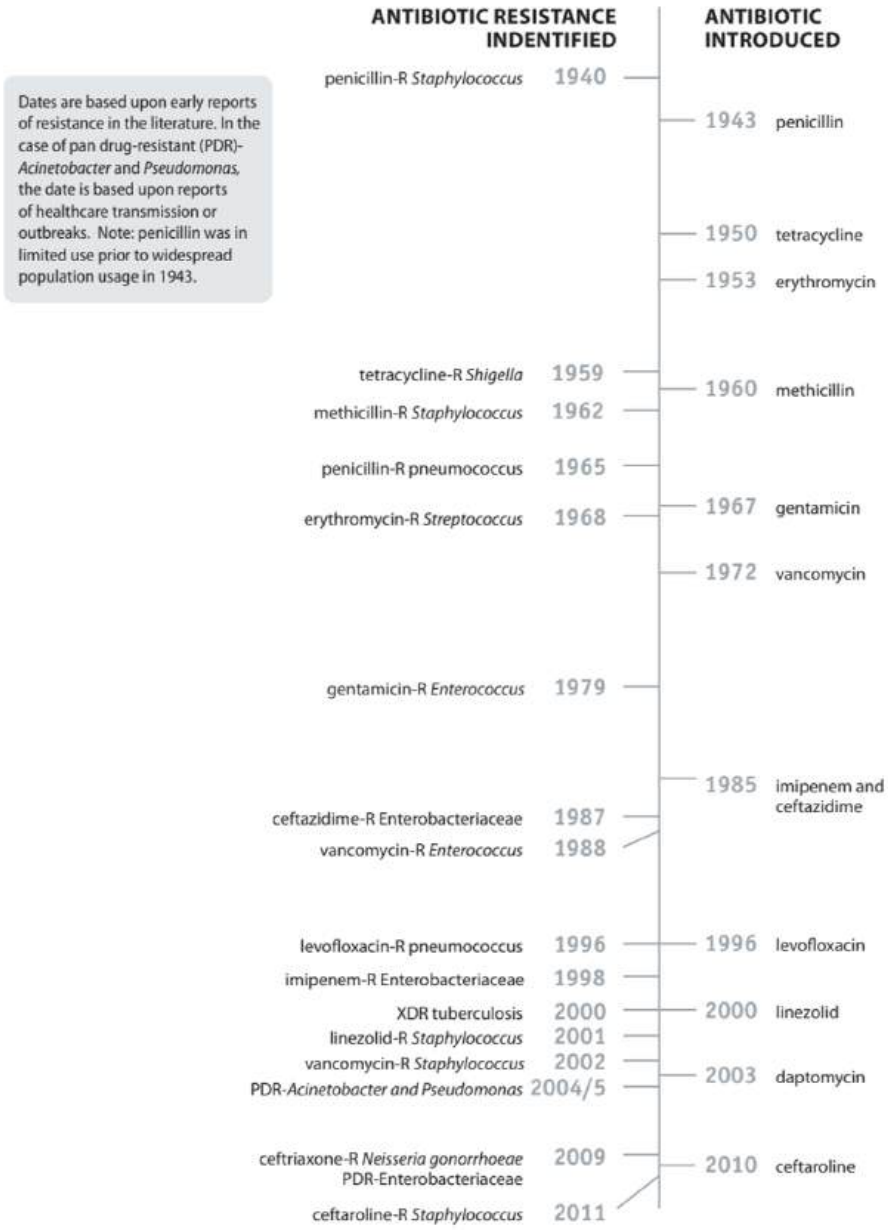
1.2 Literature Review

Healthcare associated infections (HAIs) are the largest contributor to hospital acquired complications (Surveillance and Response for Carbapenemase Producing Enterobacterales (CPE) in NSW Health Facilities, 2019). HAI outbreaks occur when multiple patients are infected with the same pathogen. Early detection is vital to limit the harm caused by these infections, however, recognising an outbreak relies on accurate identification of the pathogen involved.

Australian states have mandated hospitals report key HAIs through the Centres for Disease Control and Prevention (CDC) National Healthcare Safety Network (NHSN), which provides a standardised approach to surveillance methodology. The Australian Group on Antimicrobial Resistance (AGAR) began surveillance of the two most common bacteria species, *E. coli* and *Klebsiella* in 1992. It was only in 2004 that the genus *Enterobacter* was included after showing clinically important resistance (Bell et al., 2020). In 2018, a survey of over 8000 isolates from 36 different Australian institutions was conducted to investigate resistance levels. *Escherichia* accounted for 61%, *Klebsiella* 20%, *Enterobacter* 5%, *P. aeruginosa* 9%, and *Acinetobacter* 1% (Bell et al., 2020). Research is still heavily focused on *E. coli* and *Klebsiella* since they are commonly found in the community. Very little research has focused on *Enterobacter* species. In 2020, the same AGAR survey looked at 8752 isolates from 49 institutions across Australia (Bell et al., 2022). *Enterobacter cloacae* complex (ECC) isolates showed acquired resistance, with most levels remaining stable. However, there was increased resistance to gentamicin of 12.6% (Bell et al., 2022).

Greater innovation and investment are urgently required to target the critical gram-negative bacteria such as carbapenem resistant *Enterobacteriaceae* (CRE). Diagnostic tools, vaccines and the development of new antimicrobial medicines are needed. Research and development in this space has dwindled over the years and WHO reports only six innovative antibiotics are in the development pipeline (World Health

Figure 1.1: Timeline showing the introduction of an antibiotic and when resistance to that dug was reported. Ever since humans discovered the therapeutic value of antimicrobials, the organisms they are designed to kill have been evolving to prevent this. This image is is reproduced courtesy of the CDC, U.S Department of Health and Human services, 2013 and appeared in their Antibiotic resistance threats document.



Organisation, 2021). Various governments including, Sweden, Germany, United States of America, and the United Kingdom are piloting reimbursement models that incentivise pharmaceutical companies to develop new treatments (Gotham et al., 2021). The launch of the Antimicrobial Resistance Multi-Partner Trust Fund (AMR MPTF), the Global Antibiotic Research and Development Partnership (GARDP), and the AMR Action Fund are initiatives that could fill a major funding gap allowing more research to occur ("Antimicrobial Resistance Multi-Partner Trust Fund," 2022; "GARDP," 2022).

Enterobacter belongs to the *Enterobacteriaceae* family and is a genus of gram-negative, rod-shaped, faculatively anaerobic bacteria. Found ubiquitously, in humans it forms part of the natural intestinal flora (Ramirez D, Giron M, 2022; Rogers, Kara, 2022). *Enterobacter spp.* rarely causes disease in healthy hosts but, are opportunistic pathogens commonly associated with nosocomial infections especially in intensive care units (ICUs) and critical care units (CCUs), with increasing AMR being reported (Davin-Regli et al., 2019; Ramirez D, Giron M, 2022).

ECC is a group of opportunistic, nosocomial pathogens of multiple *Enterobacter* species capable of producing a wide variety of infections, including urinary tract infections (UTI), osteomyelitis, soft tissue infections, endocarditis, and respiratory infections (Ramirez D, Giron M, 2022). ECC includes the six *Enterobacter* species that commonly cause infections in humans, *Enterobacter asburiae*, *Enterobacter cloacae*, *Enterobacter hormaechei*, *Enterobacter kobei*, *Enterobacter ludwigii*, and *Enterobacter nimipressuralis* (Paauw et al., 2008). There are currently two subspecies of *E. cloacae* and five subspecies of *E. hormaechei* described in the literature. The predominant species found in humans is *E. hormaechei* which is often misnamed as *E. cloacae* (Sutton et al., 2018). Other *Enterobacter* species do not routinely infect humans (Davin-Regli et al., 2019). Broad-spectrum antibiotic resistance, including the emergence of resistance to last resort carbapenems (Annavajhala et al., 2019), has led to increased interest in this group of organisms.

Taxonomic classification methods have evolved over time as science has progressed along with technological advances. *Enterobacter sp.* and ECC have been known by different names since they were first reported in 1960. This has led to inconsistent name usage within the health and research community. Accurate species identification and classification is vital to track, trace and prevent outbreaks (CDC, U.S Department of Health and Human services, 2013). Consistent name usage by clinicians, infectious disease specialists and laboratory staff will enable consistent concordant communication

across public health networks. Often multiple classification methods are used to identify an organism, which for *Enterobacter spp.* can give conflicting results and low confidence species calls.

Matrix-assisted laser desorption ionization time-of-flight spectrometry (MALDI-TOF) can determine the genus level of *Enterobacter* but not the species (Pavlovic et al., 2012). Multi-locus sequence typing (MLST) is capable of accurately identifying species (Miyoshi-Akiyama et al., 2013; Pérez-Losada et al., 2013) and can be used after Sanger or whole genome sequencing (WGS). When inconsistent names are communicated to or within hospitals, outbreaks may be undetected and sub-optimal surveillance may occur. Accurate classification methods and appropriate database curation will lead to more consistent identification of *Enterobacter* species and improve response times to outbreaks in health care settings.

WGS has many benefits over other sequencing techniques as capturing the entire genome allows all coding and non-coding sections to be examined. Multiple bioinformatic tools can be utilised to investigate the structure and function of the genome for comparative, discovery, and validation studies. Routine WGS by healthcare facilities improves clinical care (Forde et al., 2023). Hospitals can identify outbreaks quickly and implement hygiene measure that have health and cost benefits. Public health interventions are improved with detailed knowledge of pathogens circulating and patterns in resistance profiles (Forde et al., 2023).

1.3 Aims of research project

This study aimed to differentiate the species boundaries for *Enterobacter cloacae* complex, investigate MLST profiles within species groups, complete pan-genome analysis for the genus and report on AMR profiles within the isolates analysed. This project will improve outbreak detection and the development of tools that can assist facilities without access to WGS to identify species quickly and understand the AMR profile of the isolate, informing health responses for individuals and healthcare facilities. The findings will also aid database curation and may lead to improvements in MALDI-TOF specificity within *Enterobacter* species. The methods used and results obtained will be made publicly available to benefit others working in public health and the fight against AMR.

Chapter 2

Methods

2.1 Data Collection and quality control

2.1.1 Datasets

Two datasets were collated and analysed within this research project; the first is based on a high-quality curated set of completed genomes obtained from Refseq (O'Leary et al., 2016), herein referred to the "preliminary dataset" (Appendix A). The second is an extended dataset comprised of varying levels of completeness (complete, scaffold, draft) obtained from Genbank (Clark et al., 2016) herein referred to as the "extended dataset" (Appendix B).

`https://github.com/S-Noonan/MDU.Research.Project/blob/main/ECC_preliminary_genomes.txt`

`https://github.com/S-Noonan/MDU.Research.Project/blob/main/ECC_extended_genomes.txt`

The preliminary dataset was comprised of 389 completed genome sequences from the genus of *Enterobacter* obtained from Refseq (O'Leary et al., 2016). These sequences consist of high quality completed genomes with a single contig representing the circular chromosome and variable present circular contigs representing mobile genetic elements such as plasmids. This dataset represented a non-redundant representative of 14 species and is composed of genomes from multiple different studies from different countries. Selected for the high curation standard, this dataset is free from contamination, and is representative of diverse, non-redundant *Enterobacter* species.

The extended dataset builds upon the preliminary collection, incorporating the same genomes and adding 3204 additional *Enterobacter spp.* genomes from the Genbank database, resulting in a total of 3,593 genomes. While the extended dataset captures a wider genetic diversity, it also includes genomes of lower quality and higher fragmentation (number of contigs), with many sequences being scaffolded or in draft form. The extended dataset represented 25 species from multiple studies and countries.

2.1.2 Quality assessment of datasets and in silico typing

To determine the quality of each genome used in this research project, we assessed the genome assembly statistics using Seqkit version 2.6.1 (Shen et al., 2016). The seqkit stat option was used to output the summary statistics for each genome assemble including: number of contigs in the assembly (num_seqs) indicating how fragmented the assembly is, sequence length and number of gaps (min_len, av_len, max_len, Q1, Q2, Q3, sum_gaps) highlighting the size of fragments and the intervening gaps, contiguity (N50) which measures the sequence length where half of the total assembly is covered. A visual plot using R software version 2022.12.0+353 (R Foundation for Statistical Computing, n.d.) showed the average number of contigs across all samples and allowed quality threshold values to be derived. These values were used to group all genomes into varying degrees of quality levels before completing subsequent phases of analysis. Multi-locus sequence typing (MLST) were assigned with the 'mlst' tool which uses the pubMLST database, the links are below.

`https://github.com/tseemann/mlst`

`https://pubmlst.org/bigsdb?db=pubmlst\_ecloacae\_isolates`

2.2 Establishing genetic thresholds for determination of species boundaries

Within a genus, isolates belonging to the same species typically exhibit closer genetic relationships to one another than to isolates from different species. There are multiple bioinformatic methods that can measure genomic relatedness between genomes and can be used to determine which isolates are more genetically similar. To assess the genetic distances within the *Enterobacter* genus and species groups we utilised multiple kmer-based pairwise comparison tools, including Mash (Ondov et al., 2016), Mashtree (Katz et al., 2019), and Split kmer analysis (SKA) (Harris, 2018) to establish distance thresholds that can be used to rapidly define species groups.

2.2.1 Mash distance and Mashtree

Mash uses MinHash dimensionality reduction to create sketches that allow rapid computation of an estimated global mutation distance, known as Mash distance (Ondov et al., 2016). Mash reduces large sequences and sequence sets, to small, representative sketches, from which global mutation distances can be rapidly estimated (Ondov et al., 2016) which is correlated closely to average nucleotide identity (ANI). `https://github.com/marbl/Mash` Mash version 2.3 was used to compare each genome

assembly to each other in a pairwise manner. Briefly, mash sketch was used to build a sketch file for each genome (hashed kmers that represent the diversity within each genome) and mash dist was used to estimate the distances (ignoring single copy kmers) (Kim et al., 2014) and significance between each sketch file. Genetic distances between pairs were evaluated. Clustering of genomes was further assessed using Mashtree version 1.4.6 (Katz et al., 2019), the sketch information and corresponding pairwise Mash distances are processed using a neighbour joining algorithm implemented through QuickTree version 2.5 (Howe et al., 2002).

2.2.2 Split kmer analysis

Split kmer analysis (SKA) version 1.0 (Harris, 2018), was utilised to conduct pairwise distance comparisons between genomes in a reference independent manner. Briefly ska fasta was used to create a split kmer file for each genome, and ska distance was used to calculate the pairwise distances and clustering from the split kmer files. The output of ska distance included similarity scores (Jaccard Index & mash-like distance), single nucleotide polymorphisms (SNPs), matches and mismatches, which are used for downstream analysis to determine species group thresholds.

2.2.3 Investigating genetic thresholds for *Enterobacter spp.* differentiation

To determine suitable genetic thresholds for categorising *Enterobacter* species, the results Mash and SKA pairwise comparisons were imported in R version 2022.12.0+353 (R Foundation for Statistical Computing, n.d.). The genomes were grouped by species according to the taxid_id provided in the metadata from Genbank and Refseq. Analyses of genetic distances and phylogenetic relationships were performed both within and between these species groupings. Species separation plots were generated using ggplot2 (Wickham, 2016) to establish thresholds for genetic distinction between species. Different mash-like thresholds were tested to delineate the limits of species clustering and genetic diversity. Network diagrams illustrating genetic relationships were created with ggnetwork (Briatte, 2023), applying criteria such as a maximum of 50,000 SNPs, Jaccard Index of at least 0.4, or a mash-like distance of 0.02 or less.

2.2.4 Species allocation confirmation

To verify the species classification of each genome, a Python 3.9.6 script was developed called BANCSIA (**Ba**cterial **N**aming for **C**orrect **S**pecies **I**dentification and **A**llocation -

https://github.com/S-Noonan/MDU.Research.Project/blob/main/BANCSIA.py). The script analysed a tab separated file containing the genetic distances between isolates and grouped them according to a specified distance threshold. BANCSIA generated a dictionary where each key represented a unique group number, and the corresponding value was a list of isolates that fell within the threshold. Outputted groups were manually inspected to assign the appropriate species name to each group number. Any groupings that could not be clearly determined were categorised under "unknown" species.

2.2.5 Centroid genome determination for rapid species identification

To establish a set of appropriate representative genomes for each species cluster, for future rapid identification of *Enterobacter sp.* we determined the centroid/mediod genome for each species cluster. The most central isolate, or centroid, was deemed as having the smallest average distance to all other isolates in the species group, serving as a representative for the group. After applying the Python script BANCSIA (https://github.com/myscript_details) to the preliminary dataset, species clusters were determined, and R tidyverse (Wickham et al., 2019) was used to identify the centroid for each species, calculated by grouping isolates by species and identifying the one with minimum average SKA mash-like distance within the group.

2.3 Pan-genome analysis to characterise species specific markers operons or loci

To assist in establishing genetic difference between species groups a genome annotation and pan-genome approach was taken to locate potential species-specific markers or operons that could be used for further differentiation and rapid identification.

2.3.1 Gene annotation

Genomes were annotated using Prokka version 1.14.6 using default settings (Seemann, 2014). The general feature format file (.gff) output file was passed as input to Panaroo (Tonkin-Hill et al., 2020).

2.3.2 Pan-genome analysis

A pan-genome analysis was completed on all isolates using Panaroo (Tonkin-Hill et al., 2020). Implementing a graph algorithm, Panaroo represents genes as nodes. The 'strict' mode was enabled, which takes a more aggressive approach to removing nodes due to contamination and erroneous annotation (Tonkin-Hill et al., 2020). Genes that appear beside each other in a contig, are represented as edges between the corresponding nodes in the graph (Tonkin-Hill et al., 2020). Multiple files are created by Panaroo including a gene presence absence matrix used in downstream analysis. The graph modelling language (GML) format file can be viewed in cytoscape (Tonkin-Hill et al., 2020).

An R script called Twilight (Horesh, n.d.) was also implemented to better understand the pan-genome differences between *Enterobacter* species groups in this study. Twilight is a population structure aware approach, which takes a species group file along with the output from Panaroo to assign as gene distribution class (Horesh, n.d.). A gene is identified as 'core', 'intermediate', or 'rare' depending on the proportion of isolates that it contains (>95%, 15-95%, <15% respectively) (Horesh, n.d.). Twilight determines the class for each lineage, the whole collection and multiple lineages for all possible combinations ('core', 'intermediate', 'rare', 'core and intermediate and rare', 'core and intermediate', 'core and rare', 'intermediate and rare') (Horesh, n.d.).

Twilight allows the minimum threshold to be set to accommodate small sample numbers in some lineages. For the preliminary dataset the default of ten was used but this was lowered to five for the extended dataset. Identifying clusters of genes located beside each other in the genome that are unique to one species may be useful to know. During the preliminary analysis, genes of interest were produced by extracting groups of genes collocated in the genome, unique to a species, by analysing the twilight classification output file (classification_output.csv).

2.4 In silico antimicrobial resistance determinants

With AMR such a concern, bioinformatic tools that quickly and accurately identify AMR genes and mutations are essential. AbritAMR is one such tool that uses an ISO endorsed pipeline (Kristy Horan et al., 2023). It takes information produced by AMRFinderPlus (Feldgarden et al., 2021) and uses an additional step to apply local reporting requirements, before compiling its final report in a format meaningful to clinicians. Only exact or close matches (100% or 90-100% identity and coverage respectively) are reported from AMRFinderPlus (Feldgarden et al., 2021).

2.4.1 AbritAMR

Using the command line, each isolate in the extended analysis was provided to AbritAMR version 1.0.14 (Kristy Horan et al., 2023). Information on genes recovered from each isolate with the desired identity threshold using the default of 90% for each drug class were reported. Matches have greater than 90% coverage and between threshold-100% identity. Partial matches occur with 50-90% coverage and above threshold identity. Virulence genes are also reported for the coverage ranges outlined above. Understanding how genotype and the corresponding phenotype relate requires validated research. However, the distribution of AMR genes across species and genus can give clues to the mechanisms involved in acquiring resistance over-time (Kristy Horan et al., 2023). To visually summarise the AbritAMR results plots were created using R (R Foundation for Statistical Computing, n.d.).

Chapter 3

Results

The results for this research project will be split into eight sections; each section focusing on a particular component which addresses part of the project aims. Results section 3.1 focuses on quality control and will present the description and quality assessment of the preliminary and extended datasets. We demonstrate how assessment of quality before completing any analysis was necessary to ensure results were reliable, accurate and meaningful. Section 3.2 reports on the population structure and species boundaries within *Enterobacter* through phylogenetic comparisons and average nucleotide identity (ANI). Sections 3.3 to 3.5 describe a process for testing and determining species boundaries within the *Enterobacter* genus and how this system can be applied for species identification using pairwise comparison methods such as SKA, implementing distance thresholds for species groups allocation (BANCSIA species allocation script) or using distances from a centroid to determine groups.

Following on from species confirmation, we examined for the presence of species specific markers that can also be used for rapid identification of *Enterobacter* species. Section 3.6 investigated the relationship of species clusters to multi-locus sequence typing (MLST) to determining if there was an association between STs and particular species groups as this may aid diagnostics when WGS is not available. Sections 3.7 and 3.8 attempted to find species specific markers through pan-genome and accessory genome comparisons and the profiling of antimicrobial resistance genes (AMR) within *Enterobacter*. Finding trends or key genetic factors that may be restricted to a species or subgroup could be used to improve and support species identification and classification.

3.1 Quality Control and assessment of preliminary and extended datasets.

3.1.1 Description of datasets

The preliminary dataset contained 389 genomes downloaded as completed genomes from Refseq (O'Leary et al., 2016). This dataset was comprised of 13 different species groups

(Table 3.1) and one uncharacterised group. The most common species were *E. hormaechei* (178 genomes, 46%), *E. cloacae* (72, 19%), *E. roggenkampii* (30, 8%), *E. asburiae* (24, 6%), and *E. ludwigii* (18, 5%). There were five species groups with five or fewer samples and seven species groups with less than ten isolates. The uncharacterised group called *E. sp.* contained 27 genomes (7%).

Table 3.1: Species count for the preliminary and extended datasets. Species groups were extracted from metadata associated with the genomes downloaded from Refseq and Genbank. Isolates in the extended dataset that met the quality threshold of ≤50 contigs were included in the downstream analyses. Isolates with "NA" for their contig value were presumed to be above the 50 contig threshold

.

Species	**Preliminary Dataset**	**Extended Dataset**			
	Total	≤21	≤50	>50	Total
E. asburiae	24	51	105	126	231
E. bugandensis	9	21	96	33	129
E. cancerogenus	6	10	14	5	19
E. chengduensis	1	2	2	16	18
E. chuanduensis	2	2	3	1	4
E. cloacae	72	96	183	241	424
E. dykesii	0	0	1	1	2
E. genomosp.	0	0	2	2	4
E. hormaechei	178	361	686	1296	1982
E. huaxiensis	0	1	2	0	2
E. kobei	14	34	65	123	189
E. lignolyticus	2	2	2	0	2
E. ludwigii	18	34	60	26	86
E. mori	5	5	16	6	22
E. oligotrophicus	0	1	3	0	3
E. quasi-hormaechei	0	0	0	2	2
E. quasi-mori	0	0	1	1	2
E.quasi-roggenkampii	0	1	4	1	5
E. roggenkampii	30	45	104	156	260
E. sichuanensis	1	1	3	6	9
E. soli	0	1	1	3	4
E. sp.	27	115	142	47	190
E. timonensis	0	2	2	0	2
E. vonholyi	0	0	1	0	1
E. wuhouensis	0	0	1	0	1
TOTAL	389	785	1499	2094	3593

The extended dataset contained 3593 genomes, which included the complete genomes downloaded from Refseq (O'Leary et al., 2016) and an additional 3204 genomes downloaded from Genbank (Clark et al., 2016) with varying levels of completeness (complete, scaffold, draft). This dataset represented 24 different species groups (Table 3.1) and one uncharacterised group. The most common species were *E. hormaechei*

(1982, 55%), *E. cloacae* (423, 12%), *E. roggenkampii* (259, 7%), *E. asburiae* (231, 6%), and *E. kobei* (189, 5%). There were 11 species groups with five or fewer samples and 12 species groups with less than ten isolates. The uncharacterised group called *E. sp.* contained 190 genomes (5%). Six isolates in the extended dataset were removed from the final analysis as they were identical to another sample based on SKA results (Appendix C).

3.1.2 Quality of Datasets

Each genome in both the preliminary and extended dataset was subject to a quality control assessment. Seqkit stats (Shen et al., 2016) was used to assess the genome assembly statistics and kraken version 2 (Wood et al., 2019) with PlusPF database was used to assess taxonomic assignment. Any genomes found to be of a non-*Enterobacter* species and were highly fragmented were removed from downstream analysis.

The preliminary dataset was made up of genomes with a median of 3.0 contigs (IQR = 1.0 – 5.0, max = 21.0), a median of 5,014,615 bases (IQR = 4,843,965 – 5,194,902, max = 5,652,128), a median sequence length of 1,642,587 bases (IQR = 1,000,649 – 4,546,642, max = 5,369,929), and a median N50 value of 4,793,062 (IQR = 4,705,976 – 4,899,400, max = 5,419,017). See table 3.2

Analysis of all genomes in the extended dataset showed a significantly larger contig number (median = 62.0, IQR = 27.0-85.62, max = 981), a similar median number of bases at 4,946,185, (IQR = 4,785,115 – 5,110,316, max = 5,975,001), a lower median sequence length of 79,125 (IQR = 44,945 – 174,446, max = 5,598,694), and a smaller median N50 size of 277,596 (IQR = 160,697 – 655,082, max = 5,598,694) as outlined in Table 3.2. Using contigs as the main determinant of quality, the results for the extended dataset were visualised using ggplot2 (Wickham, 2016) and three threshold levels were computed. See figures 3.1 and 3.2.

Aligning with the values seen in the preliminary dataset, the first quality level used was contigs between 1-21. In total, 785 isolates from the extended dataset met the threshold of $\leq$ 21 contigs, representing 22% of all samples. When a threshold of $\leq$50 contigs was applied, 1493 isolates met the criteria, equating to 42% of all genomes in the extended dataset. Based on the above results, downstream analysis was completed for the preliminary dataset and the extended dataset genomes that met a $\leq$ 50 contig cut-off. Any samples with $>$ 50 contigs were excluded from further analysis. Seqkit stats were unable to determine the number of contigs for 25 genomes in the extended dataset.

These were presumed to be above the threshold of 50 and therefore were not included in downstream analyses.

Table 3.2: Genome quality statistics produced by Seqkit for the preliminary and extended datasets. The metrics for each genome include; contigs or the number of sequences, the number of bases, the average length for contigs, and the sequence length of the shortest contig that covers 50% of the total genome length. The minimum, median, inter-quartile range and maximum for each metric are recorded.

Metric	**Dataset**	**Min**	**Med**	**IQR**	**Max**
Contigs	Prelim	1.0	3.0	1.0 - 5.0	21.0
	Ext all	1.0	62.0	27.0 - 85.6	981
	Ext≤21	1.0	4.0	2.0 - 8.0	21.0
	Ext ≤50	1.0	20.0	4.0 - 36.0	50.0
No. bases	Prelim	4483402	5014615	4843965 - 5194902	5652128
	Ext all	4129641	4946185	4785115 – 5110316	5975001
	Ext ≤21	4199690	4981029	4788958 - 5167176	5880777
	Ext ≤50	4199690	4863661	4717929 - 5041458	5880777
Avg seq length	Prelim	245862	1642587	1000649 - 4546642	5369929
	Ext all	2747	79125	44945 – 174446	5598694
	Ext ≤21	218098	1195975	648407 - 2394644	5598694
	Ext ≤50	84295	242800	133294 - 1229152	5598694
N50	Prelim	1101888	4793062	4705976 - 4899400	5419017
	Ext all	5354	277596	160697 – 655082	5598694
	Ext ≤21	282819	4736514	4546644 - 4868560	5598694
	Ext ≤50	139359	1093797	409378 - 4748459	5598694

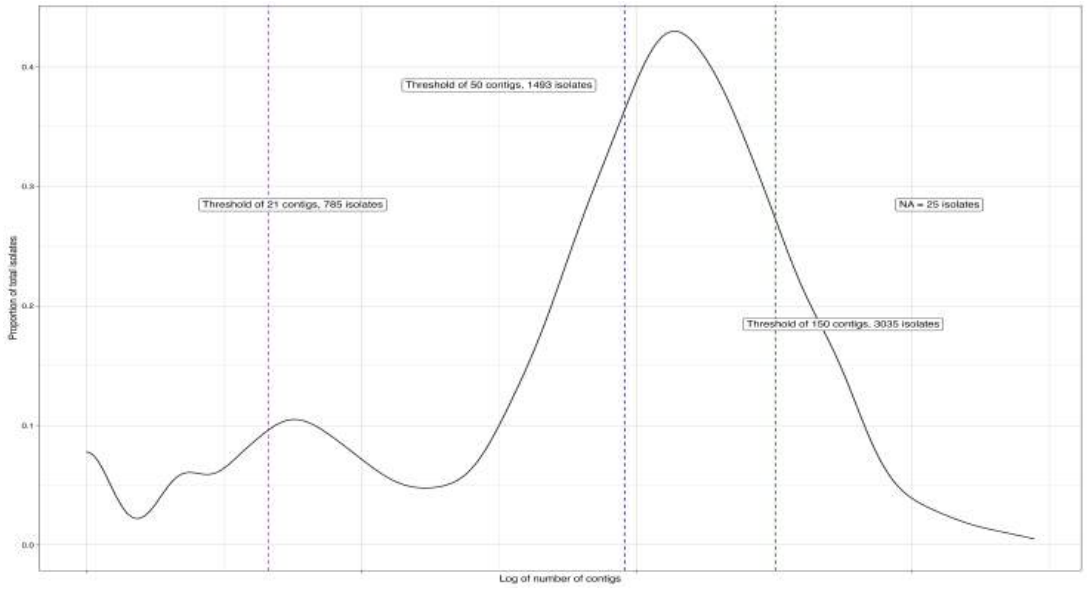


Figure 3.1: Quality assessment for all genomes in the extended dataset based on number of contigs in the assembly. The preliminary dataset contained the most complete genomes. When a threshold of 21 was applied, chosen to align with the maximum observed in the preliminary dataset, 785 isolates were included. When a threshold of 50 contigs was applied 1499 isolates were included.

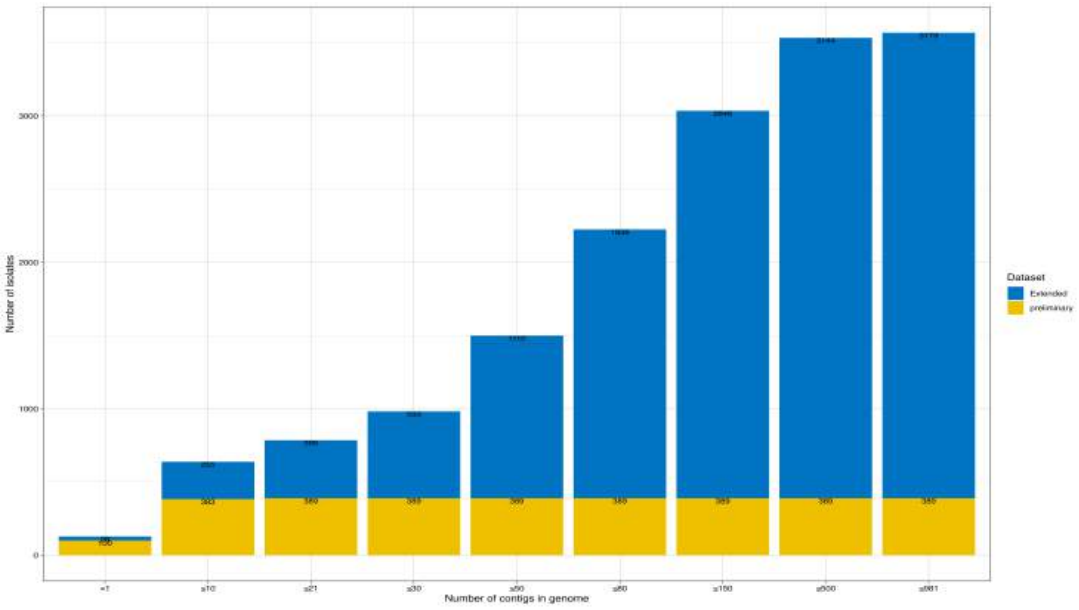


Figure 3.2: Number and relative proportion of isolates from each dataset that contribute to each quality threshold category. In the preliminary dataset 383 samples had ten or fewer contigs. A threshold of ≤ 21 and ≤ 50 contained 389 samples from the preliminary dataset and 396 and 1110 additional genomes respectively.

3.2 Population structure of *Enterobacter* species

We analysed the population structure of the *Enterobacter* genus by assessing phylogenetic relationships and average nucleotide identity. Population structure gives insights into diversity and if populations are subdivided, this can influence how they evolve.

3.2.1 Phylogenetic relationships

The phylogenetic tree (Figure 3.3) generally showed a strong alignment of species groups with the phylogenetic structure, when species were allocated as extracted from Refseq taxid_id metadata (O'Leary et al., 2016). *E. hormaechei* has one monophyletic branch that diverges into multiple distinct subbranches. However, we noted some discrepancies: genomes labelled *E. cloacae* and *E. sp.* are distributed throughout the tree. This suggests a mismatch between species names and their actual position in the phylogenetic tree, likely due to misclassification or incorrect taxonomic identification of those genomes.

As the broader ECC comprised several distinct species beyond *E. cloacae* it seems there is some confusion in classifying and labelling different *Enterobacter* species within public genome databases. Many genomes of *E. cloacae* appeared to cluster with well-defined *E.*

hormaechei and *E. asburiae* isolates. Spread through the tree are *E. sp.* isolates which indicates many of these isolates of unknown species, could be clustered with other well defined species groups.

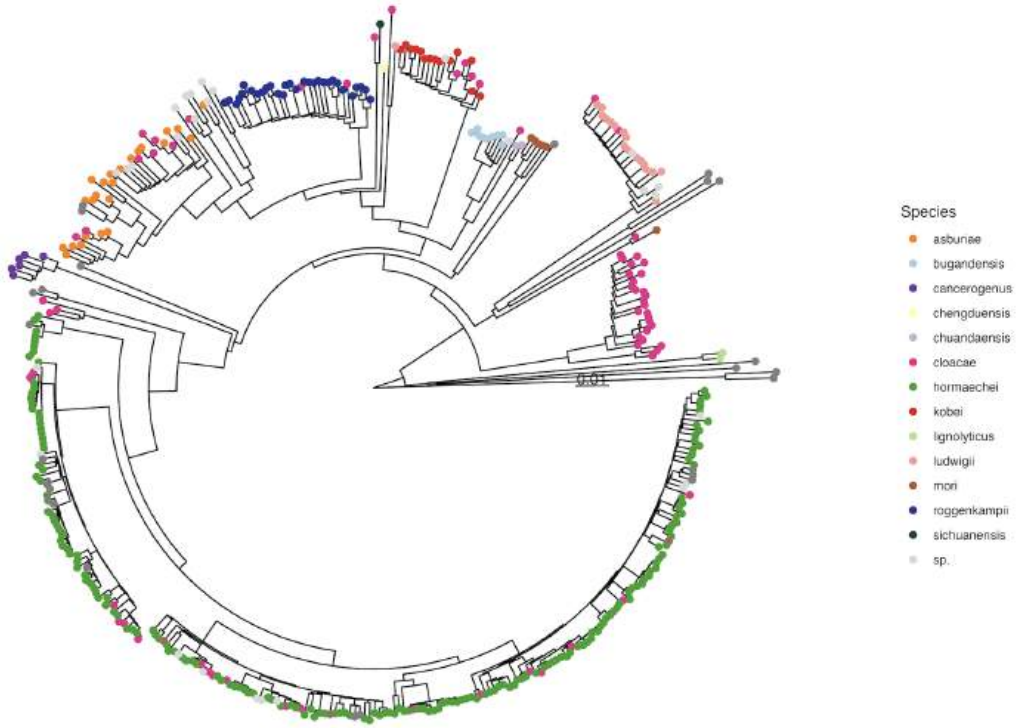


Figure 3.3: Phylogenetic tree created from preliminary dataset which contained 389 closed *Enterobacter* genomes downloaded from Refseq. The isolates are coloured based on the species name assigned in Refseq. Multiple colours in each branch show genomically similar samples that are assigned different species names. Therefore, Refseq species names do not accurately represent phylogenetic relatedness, especially for *E. cloacae* (pink) and *E. sp* (grey).

3.2.2 Defining species boundaries using average nucleotide identity

ANI is a fast comparative measure of genome relatedness but the exact species differentiation thresholds that are appropriate for *Enterobacter* have not been well established. To refine species identification in genome databases, we conducted a comparative analysis using average nucleotide identity to establish this. We experimented with various cut-off points and determined that a mash-like distance of 0.04 effectively delineates species groups.

3.3 Applying 0.04 mash-like threshold

SKA returns a distance value that is similar to Mash distance called mash-like distance. When applying a 0.04 threshold, all isolates in the same group must be less than or equal to 0.04 mash-like distance from other isolates in the group. Using the pairwise comparisons, and applying this threshold, we reassigned all genomes that were initially mismatched into newly defined species groups.

3.3.1 Applying mash-like threshold in preliminary dataset

From the preliminary dataset a total of 76 genomes were reassigned with after the mash-like threshold was applied, 32 being renamed as *E. hormaechei*. All 27 *E. sp.* were successfully reassigned across six species groups and 47 *E. cloacae* were reassigned across 11 different species groups. Furthermore, when a threshold of 0.02 was applied this aligned with four *E. hormaechei* sub-species groups and three each of *E. cloacae* and *E. asburiae*. This supports the phylogenetic relationships seen in Figure 3.3 which showed numerous *E. cloacae* and *E. sp.* interspersed through other species groups.

Applying the mash-like similarity threshold of 0.04, we categorised the preliminary dataset into 16 distinct species groups. When we compared these groups against recognised species names, we found three groups that could not be matched to a known species. These groups which were labelled as "unknown" and included seven isolates as shown in Table 3.3, may represent potential undefined species groups. The within and between species distances were calculated for each group and visualised in R using ggplot2 (Wickham, 2016). See Figure 3.4 which shows clear species separation at 0.04.

Table 3.3: Three groups categorised as unknown after applying the 0.04 species threshold, representing seven isolates from the preliminary dataset.

Isolate	**Dataset**	**Group**
Enterobacter_cloacae_88701 Enterobacter_cloacae_99101 Enterobacter_cloacae_complex_sp_	Preliminary	Unknown 1
Enterobacter_cloacae_CZ-1 Enterobacter_ludwigii_11894-yvys	Preliminary	Unknown 2
Enterobacter_cloacae_WP5-S18-ESBL-01 Enterobacter_mori_HSW1412	Preliminary	Unknown 3

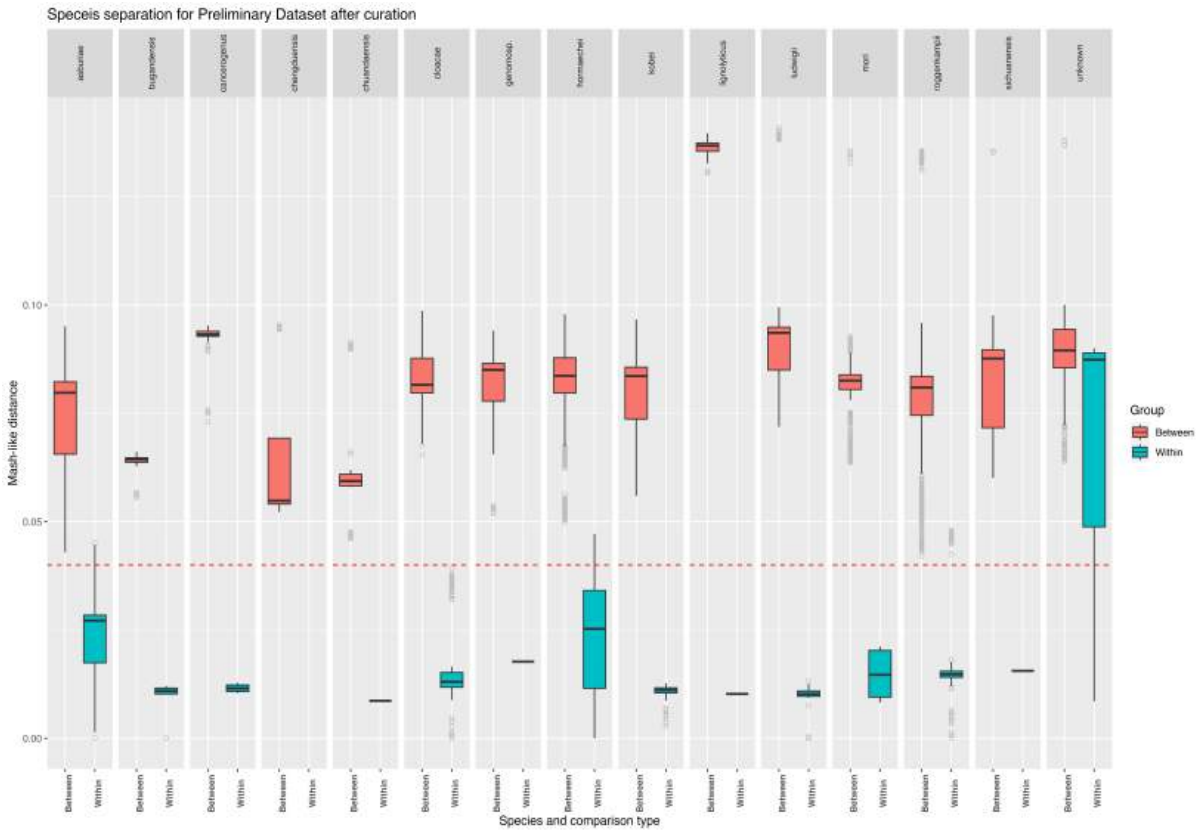


Figure 3.4: Within and between values for each species group after curation (running BANCSIA) of the preliminary dataset. Clear species separation occurs at 0.04 for most groups shown by the red dotted line. Although the mean separation value for *E. asburiae*, *E. hormaechei* and *E. roggenkampii* are well delineated, some maximum, minimum or outlier values in these species groups are close to the threshold of 0.04. The poor separation of the unknown category is expected given this group is a combination of isolates that are not necessarily closely related

When applying a threshold of 0.02 a total of 31 groups were created of which four were *E. hormaechei* sub-species, and three each of *E. cloacae* and *E. asburiae*. Eleven of the groups were unable to be accurately categorised and totalling 17 isolates.

3.3.2 Applying mash-like threshold in extended dataset

Applying the 0.04 threshold to the extended dataset, we separated the 1493 genomes into 43 different groups. Over half of these groups were unable to be confidently categorised when compared to recognised species names and the 35 isolates they represent are outlined in Appendix E. The species separation plot was repeated to show within and between species thresholds and visualisation occurred using R ggplot2 (Wickham, 2016). See Figure 3.5. When applying a threshold of 0.02 a total of 93 groups were created of which 45 groups could not be confidently categorised.

In the extended dataset, 325 isolates below the quality threshold of ≤ 50 contigs were reassigned to different species groups after curation with BANCSIA. Like the preliminary dataset, most isolates reassigned were originally named *E. sp.* or *E. cloacae*. Reassignment occurred across 12 different species groups for *E. sp.* and across 11 groups for *E. cloacae*. Overall, 100 isolates across different species groups were reassigned to *E. hormaechei*. There were 86 isolates that were categorised as both *E. asburiae* (out of 225) and grouped in *E. roggenkampii* (out of 121). To resolve this ambiguity, the distance for each isolate to the centroids were checked and grouped according to the closest match. All 86 samples were determined to be *E. roggenkampii*. See table 3.4 for further details of all re-allocations.

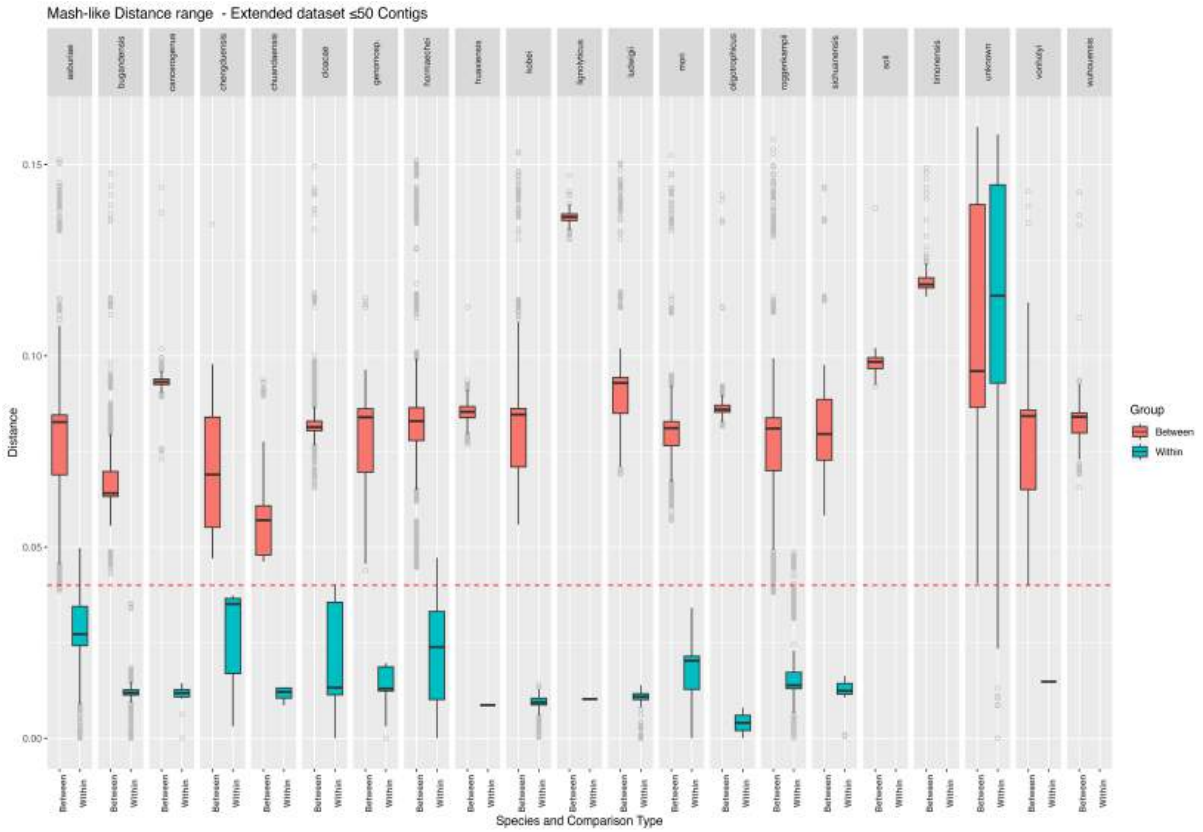


Figure 3.5: Within and between species boxplot. All isolates labelled as unknown are grouped together. Clear separation of species occurs using 0.04 threshold indicated by the dotted red line except for the unknown group. Given the unknown group contains all the isolates that couldn't be confidently characterised similar within and between species values makes sense. Three species groups (*E. asburiae*, *E. hormaechei*, *E. roggenkampii*) have outlying values that cross the threshold line which may indicate these groups are not well delineated or could contain chimeric genes.

Table 3.4: Number of isolates assigned to each species group. Before curation represents the species names extracted from the Refseq taxid_id and Genbank metadata. After curation represents the species grouped after running BANCSIA. Values coloured red indicates an increase and blue indicates a decrease in the number of samples in the group. A large number of *E. cloacae* and *E. sp.* were reassigned after running BANCSIA.

Species	**Preliminary Dataset**			**Extended Dataset**		
	Before curation	After curation	Change	Before curation	After curation	Change
Enterobacter asburiae	24	40	+16	105	139	+34
Enterobacter bugandensis	9	11	+2	95	116	+21
Enterobacter cancerogenus	6	6	0	13	12	-1
Enterobacter chengduensis	1	1	0	2	5	+3
Enterobacter chuanduensis	2	2	0	3	4	1
Enterobacter cloacae	72	26	-46	182	77	-105
Enterobacter dykesii	0	0	0	1	0	-1
Enterobacter genomosp.	0	2	+2	2	9	+7
Enterobacter hormaechei	178	210	+32	686	785	+99
Enterobacter huaxiensis	0	0	0	2	2	0
Enterobacter kobei	14	19	+5	65	78	+13
Enterobacter lignolyticus	2	2	0	2	2	0
Enterobacter ludwigii	18	23	+5	59	73	+14
Enterobacter mori	5	5	0	16	19	+3
Enterobacter oligotrophicus	0	0	0	3	3	0
Enterobacter quasi-mori	0	0	0	1	0	-1
Enterobacter quasi-roggenkampii	0	0	0	4	0	-4
Enterobacter roggenkampii	30	35	+5	103	120	+17
Enterobacter sichuanensis	1	2	+1	3	9	+6
Enterobacter soli	0	0	0	1	1	0
Enterobacter sp.	27	0	-27	142	0	-142
Enterobacter timonensis	0	0	0	1	1	0
Enterobacter vonholyi	0	0	0	1	2	+1
Enterobacter wuhouensis	0	0	0	1	1	0
Enterobacter "unknown"	0	5	+5	0	35	+35
Total	**389**	**389**		**1493**	**1493**	

3.4 Confirming species threshold through an extended dataset of 1493 genomes

Due to low sample numbers in some species groups represented in the preliminary dataset, our investigation of species boundary thresholds was repeated using additional genomes from the extended dataset. We created network plots in ggnetwork (Briatte, 2023) to visualise how isolates clustered together and separated from other groups with various separation thresholds. Applying a mash-like threshold of 0.02 and SNPs $< 50,000$ produced network plots with clusters that validated our previous findings. As seen in Figure 3.6, there are multiple *E. hormaechei*, *E. asburiae* and *E. cloacae* clusters with numerous single or low isolate number groups.

Figure 3.6: Network plot produced when pairwise comparison results were examined using a threshold of 0.02 mash-like distance with fewer than 50,000 SNPs for the preliminary dataset. *E. hormaechei* (green) separate into three large and one small subspecies group. There is one large *E. cloacae* (pink) group and numerous clusters of small groups of or single isolates.

3.5 Preparing a system for species identification from ANI

We could not directly compare every genome in the extended dataset to every other genome using SKA (Harris, 2018) due to computational limitations. Therefore, we devised a system to allow use of this tool for species classification with high sample numbers.

Computationally, the servers used in this research project failed when approximately 750 *Enterobacter* genomes were being compared. Completing all 1493 pairwise comparisons was not possible, hence we needed to be creative in our analysis without losing any valid genomic data. Therefore, we devised the following system to use ANI for species identification.

3.5.1 Centroid determination to select representative genomes for species clusters

Mediods or centroids are isolates that are centralised in species clusters and are representative of the whole group. Extracting these for each species group in the preliminary dataset was completed by determining the isolate with the lowest average mash-like distance in R (R Foundation for Statistical Computing, n.d.) to all other isolates in the cluster. Using centroids to represent each species group reduced the number of pairwise comparisons completed when the extended dataset was introduced and evaluated. The centroids determined from the preliminary dataset are shown in the table (3.5) below.

Table 3.5: Centroids identified from the preliminary dataset are isolates that are most representative of the whole species group. Each isolate had the lowest average SKA mash-like distance of all samples in that species group.

Species	**Preliminary Centroid**	**Dataset**
E. asburiae	Enterobacter_asburiae_161373-yvys	Preliminary
E. bugandensis	Enterobacter_sp._Colony324	Preliminary
E. cancerogenus	Enterobacter_cancerogenus_HAEC1	Preliminary
E. chengduensis	Enterobacter_chengduensis_WCHECl-C4_WCHECh050004	Preliminary
E. chuandsensis	Enterobacter_chuandaensis_Colony355	Preliminary
E. cloacae	Enterobacter_cloacae_CZ862	Preliminary
E. hormaechei	Enterobacter_hormaechei_3804	Preliminary
E. kobei	Enterobacter_kobei_12379-yvys	Preliminary
E. lignolyticus	Enterobacter_lignolyticus_G5	Preliminary
E. ludwigii	Enterobacter_sp._RHBSTW-00593	Preliminary
E. mori	Enterobacter_mori_BC01	Preliminary
E. roggenkampii	Enterobacter_roggenkampii_12795-yvys	Preliminary
E. sichuanensis	Enterobacter_sichuanensis_SGAir0282	Preliminary

3.5.2 Comparison against Centroids

Centroids extracted from the preliminary dataset along with isolates categorised and labelled as "unknown" (Table 3.3), were compared to the additional 396 genomes from the extended dataset that met the first quality threshold range of ≤ 21 contigs. In total, 416 pairwise comparisons occurred using SKA (Harris, 2018) in the second phase of this analysis.

After applying the 0.04 mash-like threshold for all isolates ≤ 21 contigs, species groups were allocated, and new centroids extracted (Table 3.6). These were compared to the centroids determined from the preliminary dataset to investigate the influence exerted by new data points on centroid position. Four species groups had no change to their centroid

and three new isolates were identified to represent the new species groups included in this phase of the analysis. The new centroids determined for *E. asburiae* and *E. roggenkampii* were both > 0.034 distance from the centroid in that species from the preliminary data.

Table 3.6: Centroids extracted after comparing the additional 386 isolates from the extended dataset that met the ≤ 21 contigs threshold to the preliminary centroids and unknown isolates. Distance is mash-like distance between this centroid and the preliminary centroid for that species. Four new species groups are represented and 0.0* indicates the same isolate. New isolates exerted a large influence on centroid position within *E. asburiae*, *E. hormaechei* and *E. roggenkampii*.

Species	**Under 21 contigs Centroid**	**Distance**
E. asburiae	Enterobacter_sp._SM1	0.0387195
E. bugandensis	Enterobacter_bugandensis_UMB0819	0.0113622
E. cancerogenus	Enterobacter_cancerogenus_HAEC1	0.0*
E. chengduensis	Enterobacter_cloacae_BWH_43	0.00310023
E. chuandaensis	Enterobacter_chuandaensis_Colony355	0.0*
E. cloacae	Enterobacter_sp._RC4	0.0161005
E. hormaechei	Enterobacter_hormaechei_015	0.024159
E. huaxiensis	Enterobacter_huaxiensis_090008_WCHEHu090008	New group
E. kobei	Enterobacter_sp._SECR18-0236	0.0111908
E. lignolyticus	Enterobacter_lignolyticus_G5	0.0*
E. ludwigii	Enterobacter_sp._RHBSTW-00131	0.0125199
E. mori	Enterobacter_mori_BC01	0.0*
E. oligotrophicus	Enterobacter_oligotrophicus_CCA6	New group
E. roggenkampii	Enterobacter_sp._TCD1-1	0.0349852
E. sichuanensis	Enterobacter_sp._BIDMC92	0.011916
E. soli	Enterobacter_soli_LF7a	New group

Centroids representing the preliminary dataset and the isolates classified as unknown were then compared to the additional 708 isolates from the extended dataset that met the second quality threshold range (between 21 – 50 contigs). All ska distance results were collated to assist final species group allocation.

3.5.3 Species group assignment based on standardised distances

Naming conventions for *Enterobacter* have changed over time leading to inconsistencies, yet names are the fundamental way clinicians, laboratories and researchers identify isolates, especially in outbreak detection. To accurately assign species groups we generated a python script (Van Rossum & Drake, 1995) to allocate isolates to groups based on their genetic relatedness called BANCSIA (**Ba**cterial **N**aming for **C**orrect **S**pecies **I**dentification & **A**llocation – Appendix D). BANCSIA takes output from ska distance (Harris, 2018) and a threshold value and outputs a .csv file that contains a list of isolate names, and the group number they were assigned. Manual examination of the

output was required to assign recognised species names to each group. The number of groups assigned varies depending on the threshold provided to the script. Fast, scalable, and reliable, it required no prior knowledge of species names. See also `https://github.com/S-Noonan/MDU.Research.Project/blob/main/BANCSIA.py`

Deciding which species name was most appropriate for each group was relatively simple for the preliminary dataset as within each group a clear majority ruled. However, after introducing data from the additional genomes in the extended dataset, the increased sample size created added complexity. Some species groups were obvious but other groups could be interpreted differently depending on perspective and rationale. We tested four variations of species group assignment to determine if categorising via one approach altered downstream analysis. We found no difference between the four variations. Hence, settled on the interpretation of one *E. hormaechei*, one *E. cloacae* and one *E. genomosp.* group, with all hard to determine groups labelled "unknown" combined together.

BANCSIA works similar to a density clustering algorithm. Working sequentially through all isolates it will add new isolates to an established group provided at least one sample is within the threshold distance. Thus, different groups may overlap and manual curation is required to assign species names. No overlap occurred in the lower thresholds of ≤ 21 contigs but did happen after combining all SKA data for genomes with ≤ 50 contigs. Running the combined SKA data through BANCSIA for all genomes ≤ 50 contigs, created 42 groups, of which 22 were unable to be determined. Interestingly, 86 isolates appeared in more than one group. Closer inspection revealed these were shared across *E. asburiae* (with a total of 225 isolates) and *E. roggenkampii* (with a total of 121). Resolving the overlap was completed by determining the closest preliminary centroid, which was *E. roggenkampii* in all cases. The overlap was due to one isolate that fell within the threshold distance for both species groups; Enterobacter_asburiae_LH74. All the isolates categorised as unknown are listed in Appendix E.

The species separation plot was re-computed showing *E. asburiae*, *E. hormaechei* and *E. roggenkampii* with outliers that cross the threshold value highlighting the variation within these species. All other species separated well at a distance of 0.04 as shown in Figure 3.5.

3.6 Assessment of the association of multi-locus sequence typing and *Enterobacter* species assignment

Typing in *Enterobacter* uses seven housekeeping genes; dnaA, fusA, gyrB, leuS, pyrG, rplB, and rpoB. Profiles for all samples in the extended dataset were generated by running in-silico mlst (Jolley et al., 2018, Seemann, n.d.).

3.6.1 MLST analysis of preliminary dataset

Four of the 389 preliminary dataset isolates did not have a complete MLST profile. Another 45 isolates did have a complete profile but were unable to be assigned a sequence type (ST) after comparison to the PubMLST database (PubMLST, n.d.) on 6th March 2024. These 45 isolates were spread across 11 different species groups. Uniqueness of STs across species groups was investigated and our results showed no ST appears in more than one species. Sequence typing for each species is summarised in Table 3.7.

Table 3.7: The sequence types determined for each species group in the preliminary dataset. ST was determined by running mlst and searching PubMLST (current on 6th March 2024) after curating species groups. 49 isolates out of 389 had no ST, either due to an incomplete MLST profile or a profile with no corresponding ST in the current PubMLST database

Species	**ST included - Preliminary Dataset**
E. asburiae	23, 24, 25, 27, 28, 53, 162, 252, 261, 484, 523, 650, 657, 709, 807, 930, 1108, 1407, 1578, 1585, 1586, 1587
E. bugandensis	495, 795, 1052, 1140
E. cancerogenus	Nil
E. chengduensis	414
E. chuandaensis	1900
E. cloacae	1, 84, 432, 513, 524, 736, 765, 837, 922, 932, 1385, 1421, 1523, 2094
E. hormaechei	4, 46, 50, 61, 62, 65, 78, 88, 89, 90, 92, 93, 97,102, 104, 106, 108, 109, 110, 113, 114, 124, 133, 134, 138, 141, 145, 150, 168, 171, 175, 177, 190, 200, 231, 254, 269, 278, 279, 286, 303, 310, 344, 354, 382, 418, 425, 451, 461,535, 544, 594, 662, 664, 696, 756, 764, 765, 813, 816, 906, 911, 1015, 1019, 1103, 1344, 1439, 1643, 1723, 1734, 1735, 1736, 1795, 1862, 2412, 2636
E. kobei	3, 32, 54, 56, 57, 99, 191, 280, 365, 520, 770, 806, 910, 1605
E. lignolyticus	811
E. ludwigii	2, 13, 282, 374, 446, 714, 735, 1145, 1281, 1507, 1708, 1724, 2751
E. mori	1006
E. roggenkampii	166, 232, 272, 466, 486, 595, 702, 905, 929, 997, 1010, 1059, 1142, 1168, 1237, 1778, 2399, 2661, 2662
E. sichuanensis	738, 1330
E. unknown	873, 911

Table 3.8: The sequence types determined after using mlst and searching PubMLST for each species group following curation of the extended dataset. 258 isolates out of 1493 had no ST, either due to an incomplete MLST profile or a profile with no corresponding ST in the current PubMLST database (as of 22nd April 2024).

Species	**ST included - Extended Dataset**
E. asburiae	23, 24, 25, 27, 28, 29, 53, 59, 162, 252, 261, 277, 290, 319, 484, 531, 562, 643, 644, 650, 657, 684, 685, 709, 713, 720, 727, 733, 745, 807, 879, 919, 930, 1057, 1108, 1249, 1253, 1407, 1578, 1585, 1586, 1587, 1602, 1622, 1639, 1673, 2431, 2448, 2478
E. bugandensis	35, 386, 431, 495, 499, 695, 701, 784, 791, 795, 921, 1052, 1080, 1084, 1085, 1087, 1090, 1096, 1098, 1099, 1140, 1394, 1675, 1694, 1943, 2301, 2503, 2504, 2506, 2509
E. cancerogenus	2745
E. chengduensis	414, 1966
E. chuandaensis	944, 1900
E. cloacae	1, 84, 167, 412, 432, 477, 513, 524, 609, 627, 700, 712, 721, 736, 765, 789, 820, 837, 922, 932, 952, 976, 1382, 1385, 1421, 1511, 1513, 1515, 1516, 1517, 1519, 1521, 1523, 1524, 1551, 1923, 2092, 2094, 2096
E. genomosp.	873, 2390
E. hormaechei	4, 45, 46, 48, 49, 50, 51, 61, 62, 63, 65, 66, 68, 78, 79, 88, 90, 92, 93, 94, 97, 102, 104, 106, 108, 109, 110, 112, 113, 114, 116, 120, 121, 124, 127, 133, 134, 135, 136, 138, 141, 145, 146, 150, 151, 152, 158, 168, 170, 171, 175, 177, 182, 190, 200, 204, 231, 233, 254, 264, 268, 269, 278, 279, 286, 295, 303, 304, 310, 316, 331, 335, 344, 346, 354, 382, 395, 407, 418, 421, 425, 451, 459, 461, 511, 517, 521, 528, 535, 542, 544, 550, 554, 557, 568, 592, 594, 597, 604, 636, 654, 662, 664, 677, 678, 682, 683, 687, 688, 693, 696, 697, 699, 705, 706, 722, 724, 728, 729, 740, 742, 744, 756, 758, 764, 766, 772, 776, 782, 785, 787, 792, 793, 797, 798, 800, 801, 805, 809, 813, 816, 828, 906, 927, 947, 948, 968, 973, 974, 982, 1015, 1017, 1019, 1078, 1081, 1103, 1115, 1160, 1165, 1196, 1197, 1240, 1254, 1257, 1260, 1283, 1344, 1350, 1401, 1439, 1476, 1480, 1643, 1723, 1734, 1735, 1736, 1795, 1854, 1862, 1902, 2078, 2246, 2412, 2582, 2584, 2596, 2597, 2604, 2614, 2617, 2621, 2622, 2635, 2636, 2646, 2648, 2649, 2749
E. huaxiensis	Nil
E. kobei	3, 32, 54, 56, 57, 99, 125, 191, 280, 365, 480, 520, 555, 639, 691, 694, 708, 726, 731, 741, 759, 770, 773, 777, 790, 806, 910, 914, 1001, 1034, 1204, 1352, 1409, 1605, 2451, 2467, 2485
E. lignolyticus	811
E. ludwigii	2, 12, 13, 14, 16, 20, 253, 258, 282, 374, 409, 446, 675, 692, 698, 714, 735, 775, 781, 895, 1102, 1145, 1200, 1252, 1281, 1299, 1507, 1708, 1724, 1803, 1838, 2402, 2443, 2737, 2738, 2740, 2751
E. mori	624, 1006
E. oligotrophicus	Nil
E. roggenkampii	40, 96, 165, 166, 232, 272, 466, 476, 486, 523, 526, 590, 595, 634, 681, 702, 715, 725, 730, 732, 743, 746, 747, 767, 905, 929, 963, 997, 1010, 1055, 1059, 1066, 1142, 1168, 1237, 1255, 1256, 1258, 1292, 1403, 1594, 1652, 1778, 1911, 2053, 2085, 2392, 2396, 2399, 2434, 2447, 2464, 2479, 2661, 2662
E. sichuanensis	676, 738, 1330, 1392
E. soli	723
E. timonensis	1464
E. vonholyi	Nil
E. wuhouensis	Nil
E. "unknown"	680, 707, 719, 774, 911, 928, 1920, 2005, 2442

3.6.2 MLST analysis of extended dataset

After curating species groups for the extended dataset, the MLST results were evaluated again. Of the 1493 isolates meeting the minimum quality threshold, 56 did not have a complete MLST profile when compared to data uploaded on PubMLST (Jolley et al., 2018) as of 22nd April 2024. Another 202 were not assigned a ST, even though they did have a complete profile of alleles indicating novel combinations of alleles. ST undefined isolates were spread across 17 different species groups. Each species group contained unique STs which is summarised in Table 3.8. We did not observe a sharing of MLST profiles across species groups. Our evidence suggests MLST can be used to identify species due to this uniqueness.

3.7 Pan-genome analysis to define species specific markers

Pan-genome analysis can assist in identifying species-specific markers by analysing the genetic diversity within and between species isolates, focusing on conserved and unique genomic regions to the total population or a subgroup of genomes. Following genome annotation with Prokka (Seemann, 2014) a pan-genome assessment using Panaroo (Tonkin-Hill et al., 2020) and the R script Twilight (Horesh, n.d.) was implemented, initially on the preliminary dataset and then on the quality controlled isolates in the extended dataset. We then assessed if there were any group specific genes associated with the dominant species groups (13 species groups).

3.7.1 Core and lineage specific genes

The pan-genome of the *Enterobacter* was determined to be 82,176 gene families with 1094 gene families assigned as core (contained in >95% genomes) across the dominant 13 species groups. The more highly sampled species groups contained larger pan-genomes with *E. hormaechei* reported as having 33,107 gene families, *E. asburiae* and *E. roggenkampii* as having 22,869 and 20,716 respectively.

To ensure robust lineage specific genes were identified from Twilight (Horesh, n.d), we only investigated the 13 species groups with sufficient sample numbers. There are 38,007 lineage specific rare genes and 89 lineage specific core genes across eight species groups. This highlights the small core genome in *Enterobacter* shared between species, but each individual species maintains a much larger proportion of genes (1025-3790) as seen in table 3.9.

3.7.2 Genes of interest

After extracting the lineage specific core genes from Twilight (Horesh, n.d.), their position in the genome was investigated using output from Panaroo (Tonkin-Hill et al., 2020). Clusters of genes adjacent in the genome known as operons, often co-transcribe multiple proteins, and are usually well conserved. Identifying if gene clusters exist in a species group, can allow the operon to be targeted for diagnostics. Single genes are more likely to be gained or lost over time whereas operons, particularly those with similar metabolic functions, can be targeted by sequencing panels and may provide another form of confirmation of species group when WGS has been performed.

Table 3.9: Gene count produced following Twilight pan-genome analysis. For each species group the total gene diversity count is included (Total genes) along with the count of genes found in all samples of that species (Core genes), and the number of lineage specific core genes.

Species	**Total genes**	**Core genes**	**Lineage specific core**	**number of genomes**
Enterobacter asburiae	22869	2886	0	139
Enterobacter bugandensis	12695	3244	4	116
Enterobacter cancerogenus	7532	3725	26	12
Enterobacter chengduensis	6815	3548	0	5
Enterobacter chuanduensis	Not reported	Not reported	0	4
Enterobacter cloacae	16930	3052	8	77
Enterobacter genomosp.	6332	3516	1	9
Enterobacter hormaechei	33107	1706	0	785
Enterobacter huaxiensis	Not reported	Not reported	0	2
Enterobacter kobei	16691	2819	13	78
Enterobacter lignolyticus	Not reported	Not reported	0	2
Enterobacter ludwigii	14649	3553	29	73
Enterobacter mori	10158	3447	4	19
Enterobacter oligotrophicus	Not reported	Not reported	0	3
Enterobacter roggenkampii	20716	2662	0	120
Enterobacter sichuanensis	6024	3790	4	9
Enterobacter soli	Not reported	Not reported	0	1
Enterobacter timonensis	Not reported	Not reported	0	1
Enterobacter vonholyi	Not reported	Not reported	0	2
Enterobacter wuhouensis	Not reported	Not reported	0	1
Enterobacter unknown	25216	1025	0	35

The lineage specific genes identified by Twilight (Horesh, n.d.) are detailed in Appendix J. When a minimum operon length of three genes was considered, only fours species group

contained operons. *E. cancerogenus* contained three operons, *E. cloacae* contained one operon, *E. kobei* contained three operons, and *E. ludwigii* contained two operons. Most of the individual genes reported are named "hypothetical" with no currently known function. A UniProt (The UniProt Consortium et al., 2023) search confirmed the function reported by Panaroo and highlighted many of these genes are recorded across a wide number of taxa. Further annotation is required to understand the function of these operons and if they can provide useful diagnostic markers.

3.8 Anti-microbial resistance gene presence and species profiling

Finally, we report on anti-microbial resistance gene carriage across *Enterobacter* as this bacterial group is a WHO pathogen of concern. *Enterobacter spp.* are intrinsically resistant to ampicillin, amoxicillin, first-generation cephalosporins, and cefoxitin, owing to the presence of a constitutive AmpC β-lactamase (Intra et al., 2023). Resistance to carbapenems, a last line drug, is becoming more common, therefore we ran AbritAMR (Kristy Horan et al., 2023) to understand the complete AMR profile within our dataset.

3.8.1 Results from AbritAMR

AbritAMR (Kristy Horan et al., 2023) results from the extended dataset have been filtered and the summary information for those that met the quality threshold ≤ 50 contigs are outlined. Zero antimicrobial genes were found in 15 isolates and the median value was four genes (IQR = 4-9, maximum = 20) See Figure J.1. *E. asburiae*, *E. bugandensis*, *E. chengduensis*, *E. cloacae*, *E. genomosp.*, *E. hormaechei*, *E. kobei*, *E. ludwigii*, *E. mori*, *E. roggenkampii* and the unknown group contain isolates with greater than nine AMR genes which was the third-quartile value.

The largest species groups were analysed in detail and Table 3.10 outlines the proportion of isolates (as a percentage of the total isolates in that species) that contain at least one resistance gene for each drug resistance type. Appendix N lists the exact resistance genes reported for each drug class after running AbritAMR (Kristy Horan et al., 2023).

Over 90% of isolates in most highly sampled species classes show resistance types ESBL AmpC, fosfomycin, Phenicol/Quinolone. *E. hormaechei* had significantly higher proportions across many of the resistance types compared to the other species and also

contained the largest count of genes in most categories. *E. hormaechei* had the highest proportion of metallo-β-lactamase carbapenemase. These results are not surprising given *E. hormaechei* is the most prevalent clinical isolate. *E. bugandensis* had the lowest proportion of resistance to carbapenems. A worrying finding is the high proportion of isolates that contain genes that may confer resistance to colistin (30.8% in *E. kobei*, 19.9% in *E. hormaechei*, and 17.3% in *E. asburiae*). Colistin is a last resort therapeutic option in *Enterobacter* (Doijad et al., 2023). The WHO Global Antimicrobial Resistance Surveillance System report (World Health Organization, 2022), indicates *Enterobacter* is naturally susceptible, but a recent paper reported that colistin resistance patterns were strongly associated with the presence of the arn gene cassette (Doijad et al., 2023). This gene group was not reported in our results.

Figure K.1 in Appendix K has a summary plot produced in ggplot2 (Briatte, 2023) that shows the isolates that contain resistance genes that may confer resistance to common drug classes, grouped by species type. For each coloured species group, the bars represent the per isolate resistance gene count. Disturbingly, all species groups contain isolates with resistance to each resistance type. Validation to relate resistance gene presence and non-susceptibility to antibiotics has not yet occurred for *Enterobacter*, however our findings indicate the diversity of resistance mechanisms present in this genus. AMR is not restricted to hospital sources samples but occurs boradly in the genus.

Table 3.10: Number of isolates that contain genes that may confer resistance to a drug type. The value given is the percentage of isolates in that species that contain a gene for that drug class, as reported by AbritAMR. Some isolates contained more than one resistance gene.

Resistance Type	***E. asburiae***	***E. bugandensis***	***E. cloacae***	***E. hormaechei***	***E. kobei***	***E. ludwigii***	***E. roggenkampii***
Amikacin/Kanamycin	0	0	0	1.9	1.3	0	0
Amikacin/Kanamycin/Tobramycin	2.2	0	2.6	8.2	2.6	0	1.7
Amikacin/Kanamycin/Tobramycin/Quinolone	5.0	3.4	2.6	11.8	2.6	2.7	2.5
Amikacin/Gentamicin/Kanamycin/Tobramycin	0	0	0	0.1	0	0	0
Aminoglycosides (Ribosomal methyltransferase)	0	0	3.9	3.1	2.6	0	0
Other aminoglycoside resistance (non-RMT)	0	0	0	0.5	1.3	0	0.8
Beta-lactamase (not ESBL or carbapenemase)	10.8	5.2	15.6	36.2	12.8	2.7	5.0
Beta-lactamase (unknown spectrum)	0	0	0	1.4	1.3	0	0
Carbapenemase	18.0	3.4	13.0	12.4	10.3	11.0	15.8
Carbapenemase (MBL)	7.2	0.9	6.5	20.4	7.7	0	0.8
Chloramphenicol	7.2	3.4	7.8	20.1	9.0	1.4	2.5
Chloramphenicol/Florfenicol	0	0	3.9	4.3	1.3	0	1.7
Colistin	17.3	0	5.2	19.9	30.8	4.1	12.5
Erythromycin	0	0	0	0	0	0	0
ESBL	6.5	6.0	7.8	23.8	16.7	1.4	2.5
ESBL (AmpC type)	100	100	100	98.9	97.4	100	98.3
Fosfomycin	93.5	100	96.1	86.0	94.9	100	90.8
Gentamicin	9.4	4.3	7.8	28.9	9.0	1.4	4.2
Gentamicin/Kanamycin/Tobramycin	6.5	0	1.3	12.9	10.3	0	2.5
Gentamicin/Tobramycin/Apramycin	0	0	0	0.3	0	0	0
Kanamycin	1.4	0	2.6	10.3	0	1.4	0.8
Lincosamides	0	0	1.3	0.4	0	0	0
Macrolide	7.2	0	2.6	14.1	3.8	4.3	0.8
Other antimicrobial	0.7	0.9	6.5	11.5	2.6	0	0
Phenicol/Quinolone	92.8	100	96.1	92.5	98.7	93.1	95.0
Quinolone	36.0	5.2	15.6	30.0	17.9	2.7	8.3
Rifamycin	5.8	0	0	14.1	3.8	2.7	3.3
Streptomycin	10.1	4.3	15.6	44.1	23.0	5.5	14.2
Streptomycin/Spectinomycin	0	0	0	0	0	0	0
Sulfonamide	12.2	4.3	22.1	44.8	21.8	5.5	18.3
Tetracycline	2.2	4.3	14.3	25.6	5.1	4.1	11.7
Trimethoprim	10.1	4.3	13.0	40.4	11.5	4.1	12.5
Number of isolates	**139**	**116**	**77**	**785**	**78**	**73**	**120**

Chapter 4

Discussion

4.1 How genome sequencing and new methods can improve identification of bacterial pathogens including *Enterobacter* species

Whole genome sequencing is increasingly being utilized in research and public health laboratories as a cost-effective and time-efficient option for tracking pathogens. Sequencing the entire genome provides comprehensive information about the genetic makeup of an organism, including genes involved in virulence, antibiotic resistance, and metabolic pathways. Traditional methods of identification using morphological traits or biochemical tests cannot distinguish bacteria that exhibit phenotypic plasticity. Some bacteria modify their phenotypic characteristics in response to environmental conditions, leading to misidentification. Additionally, genotypically different bacteria can demonstrate the same phenotype. Genome sequencing can help improve strain identification, complementing and even replacing some existing laboratory methods.

The *Enterobacter cloacae* complex (ECC) encompasses multiple species that are clinically significant and often associated with multidrug resistance (MDR) including resistance to last-resort carbapenems (Annavajhala et al., 2019). The classification of ECC has evolved significantly due to advancements in genomic techniques. ECC comprises multiple species, including *Enterobacter cloacae*, *Enterobacter hormaechei*, *Enterobacter roggenkampii*, and *Enterobacter kobei*, among others. These species exhibit significant genetic diversity, making phenotypic identification unreliable (Paauw et al., 2008). The use of 16S rRNA amplicon sequencing is problematic as many genera in *Enterobacteriaceae* cannot be separated unless variable regions V3 and V4 are sequenced, and species level differentiation is not possible with this method (Gupta et al., 2019). MALDI-TOF has also historically been very inaccurate at differentiating ECC however, progress is developing in this area (Candela et al., 2023).

Whole genome sequencing is now regarded as essential for accurate species identification within the ECC but other methods are being developed for rapid identification to assist clinical decision-making (Ji et al., 2021). Phylogenetic analysis have led to several species being redefined including *E. cloacae subspecies dissolvens* (Hoffmann, Stindl, Ludwig, Stumpf, Mehlen, Heesemann, et al., 2005) and multiple *E. hormaechei* subspecies (Sutton et al., 2018). Despite the advances of genome sequencing to differentiate these species, there is still a lack of a comprehensive understanding on where species boundaries occurs within the ECC and even WGS based identification tools suffer from the flaws in previous classification systems.

Existing reference databases for bacterial taxonomy are often incomplete and biased toward well-studied species. As a result, novel or less studied species may be misclassified or incorrectly identified. These databases rely on current knowledge of the genetic landscape for pathogens and are subject to curation. In the case of the ECC, many WGS based identification systems are contaminated or involve misclassified genomes, which can lead to erroneous results and identification. The analyses of 1493 genomes of the 3593 *Enterobacter* species presented in this thesis, confirmed this observation with our results showing 105 of 182 *E. cloacae* genomes within genome repositories are stored under the wrong species name.

We conducted pairwise comparisons between a curated selection of genomes to examine whether clear species boundaries exist between ECC species and if a standardised threshold of genetic distance can be applied for future identification and classification. We determined a distance measure of 0.04 was a reliable threshold for species group separation and within this dataset, we found 20 species groups and 23 novel groups (to be discussed later). The species threshold of 0.04 is similar to what has been reported in other *Enterobacteriaceae* such as *Klebsiella pneumonaie* (Wyres et al., 2020), another pathogen that has experienced naming and taxonomic issues which has undergone several stages of taxonomic re-classification heavily driven by genomic studies (Lam et al., 2021; Wyres et al., 2020).

Based upon our established 0.04 threshold we utilised the BANCSIA script to reassign genomes to their appropriate species group. This method reallocated 325 isolates from the extended dataset of which 105 were originally assigned as *E. cloacae* and another 142 were originally *E. sp.*. We suggest these have occurred because it is likely researchers and public health analysts have either uploaded their genomes as *E. sp.* because of a lack of identification, or they have misinterpreted what ECC means and

think its means *E. cloacae* the species by default.

Our results also indicate there is a population sub-structure within the most common species *E. hormaechei* which creates four or five subspecies, depending on the threshold applied. This supports previous work completed in this genus (Hoffmann, Stindl, Ludwig, Stumpf, Mehlen, Monget, et al., 2005; Sutton et al., 2018; Wu et al., 2020). Our curated phylogenetic tree (Appendix G) shows four clades in one portion of the tree and another clade between *E. cloacae* and *E. ludwigii*. The network plots generated in our study, from *E. hormaechei* isolates in the preliminary dataset, showed five subspecies groups exist at 0.015 and four groups at 0.02 mash-like distance, which appears to be a good marker for differentiating these subspecies. However, further work within the *E. hormaechei* species needs to be conducted to refine this.

Generating sufficient data to extract subspecies level centroids in *E. hormaechei* would be extremely informative. The preliminary dataset was represented by one *E. hormaechei* centroid which may have been too broad, given the large sample numbers and evidence to suggest there are clear subspecies. Additional isolates from the extended dataset were compared to this single centroid. Completing more SKA comparisons between *E. hormaechei* isolates from the preliminary and extended datasets, then extracting subspecies centroids, may assist in delineating this group. Given the high portion of clinical samples that are *E. hormaechei*, clearer differentiation within this group would be beneficial.

We found 23 clusters (35 isolates) that sat outside the 0.04 threshold to all other known species groups, which aligned closely to results recorded by Wu et al., (2020). These clusters may represent novel species that are yet unnamed. Individual isolates were contained in 14 of these clusters and the remaining clusters contained between 2-5 genomes, with no obvious pattern relating to quality of the assembly. Genotypic differences can influence traits like pathogenicity which further biases which organisms are commonly sampled in healthcare environments. The discovery of novel bacterial species, not be identifiable through traditional methods is facilitated through genome sequencing when sufficient numbers of diverse genome assemblies exist. Of the species groups that aligned with known species names, we reported nine with five or fewer genomes. Low sample numbers for these groups may have influenced our results. Sequencing more high-quality genome assemblies for these less common species groups will improve our ability to identify novel species, enable the exploration of bacterial diversity in various environments, and lead to a more comprehensive understanding of

the microbial world.

When allocating species groups using BANCSIA, 85 isolates appeared in both *E. asburiae* and *E. roggenkampii*. ANI is global measure of similarity. It does not indicate which sections of sequence are similar or different. Closer investigation of these isolates showed they were all closer to the *E. roggenkampii* centroid and linked by one isolate that fell within the threshold distance for both species groups (Enterobacter_asburiae_LH74). Completing detailed sequence alignment for this sample in relation to other isolates in the *E. asburiae* and *E. roggenkampii* groups may reveal if the genomes have converged or if it is a chimeric genome.

We further confirmed species boundaries in *Enterobacter* by analysing the multi-locus sequence typing (MLST) profile and pan-genome to establish whether any additional genome markers can be used to assign species. MLST can be conducted independently of WGS, so it enables laboratories or research groups to differentiate species groups utilising the typing scheme, thanks to advances in WGS data. The current MLST scheme used in *Enterobacter* utilises seven genes; dnaA, fusA, gryB, leuS, pyrG, rplB, rpoB (Miyoshi-Akiyama et al., 2013) and expanded on previous work (Paauw et al., 2008). MSLT is a valuable reference, provided researchers and laboratories upload their sequence to pubMLST (Jolley et al., 2018) regularly so allelic profiles are comprehensively represented for each sequence type. Our results showed MLST sequence types are unique within species, yet further sampling to dramatically increase the database size will confirm if isolates with the same ST are from the same species group. Encouraging regular additions to the pubMLST database (Jolley et al., 2018) will ensure ST can be used to aid species differentiation.

Identifying species specific genes to enable accurate, cost-effective diagnosis using PCR tests in clinical settings, improves patient care. Our results showed species specific markers could be identified in some groups but these results were not consistent with the findings reported by (Ji et al., 2021) who designed a multiplex PCR amplification method that could accurately identify clinically significant *E. cloacae*, *E. hormaechei*, *E. roggenkampii*, and *E. kobei* isolates.

Expanding on the lineage specific gene markers identified in our research, we searched for species specific operons that appeared to be stable within species groups. If found, these operons could enable rapid species confirmation, when WGS was used, or alternatively allow real-time PCR or standard amplicon PCR tests to be developed. We

found 21 genes with a unique signal that made up six operons, as they occurred in runs of at least three genes. *E. kobei* contained three operons, *E. ludwigii* contained two operons, and *E. cloacae* contained one operon. There were no lineage specific markers identified in *E. hormaechei* but this could be due to the diverse clades contained within this species. A more detailed analysis at the subspecies level for *E. hormaechei* may reveal unique gene markers to assist rapid diagnostics.

The detailed information obtained from our results is a clear example of how we can refine taxonomic classifications and better understand bacterial diversity. We have demonstrated how applying ANI thresholds in *Enterobacter spp.*, in the same way as has been applied in *Klebsiella* improves taxonomic classification. Groups such as Genome Taxonomy Database (GTDB) apply global thresholds to bacterial species (Parks et al., 2022), but local, species specific approaches to finesse boundaries need to be used for different species and genus groups.

4.2 Improvement of Reference Databases and Communication

Within this thesis we address the current issues in species identification and limitations in reference databases for whole genome sequencing of *Enterobacter*. A common issue within public health and research programs is where the communication and understanding of pathogens and their taxonomy is based on outdated schemes for which even curated repositories are lagging behind (Larsen et al., 2014). The traditional species concepts and taxonomic boundaries are often inconsistent with genomic data. Tools like Average Nucleotide Identity (ANI) and new methods like FastANI help address these inconsistencies but require widespread adoption and integrating them into taxonomic practices will be an ongoing state of development for the next few years (Prinzi & Moore, 2023; Varghese et al., 2015).

In the case of *Enterobacter* species boundaries, there are groups tackling this issue (Sutton et al., 2018; Wu et al., 2020) for which similar results were reported to what we have found. Despite several research groups tackling this, reference databases still often contain mislabelled or erroneous sequences, leading to incorrect species identification by other groups or individuals who may not be aware of these errors (Cabezas et al., 2023; Keck et al., 2023). Additionally, databases utilised may be incomplete and contain significant gaps, have key taxa missing or intraspecific variants (Cabezas et al., 2023; Keck et al., 2023). Unfortunately, the solution to this problem is not simple but rather

requires strategies with extensive collaboration between public health, research, and repository curators to provide updated taxonomic practises and maintain improve database management.

So where do we go for *Enterobacter* identification?

Although the International Code of Nomenclature of Prokaryotes (ICNP) was revised in 2022 (Oren et al., 2023), there is no accepted standard for naming new species. To validate species names before publishing, phenotypic characterisation and deposition of a type strain is required (Sutton et al., 2018) but many ECC clades do not have a named type strain (Sutton et al., 2018). Sutton et al., (2018) suggests using placeholder names that indicate the species is provisional whilst waiting for someone to complete the hard work required to type strain the isolate. The species delineation suggested by Wu et al., (2020) separates the large group we referred to as *E. hormaechei* into distinct species (*E. hoffmannii*, *E. xianfangensis* and *E. hormaechei*) but their suggested name for the novel species *E. quasiroggenkampii* is confusing, as it incorporates another species names within it. Despite being a separate species according to their findings (Wu et al., 2020) our results didn't show any significant divergence of "quasi" species from established species groups. Other examples of taxonomic flux include the removal *E. nimipressuralis* (reclassified to *Lelliottia nimpressuralis* (Brady et al., 2013) after being included in ECC by Paauw et al., 2008) and *E. timonensis* to a novel genus *Pseudenterobacter* (Wu et al., 2020). Given naming of subspecies within *Enterobacter* is not clearly defined either, for public health purposes, it may be more useful referring to species name and sequence type.

It has become clear there is a disparity between the use of *Enterobacter cloacae* complex (ECC) and *Enterobacter cloacae* when literature and stakeholders refer to *Enterobacter*. ECC refers to a group of species that share the trait of causing human infections and *E. cloacae* is a specific species within the *Enterobacter genus*. The term ECC is misleading as it implies a level of genetic relatedness which was not supported by our study or other studies (Sutton et al., 2018; Wu et al., 2020). The between species distance findings were no different in ECC species compared to the rest of the species groups within the genus, therefore its utility as a naming convention does not provide added value to research or diagnostics.

Given the predominance of *E. hormaechei* in clinical settings and our suggestion that ECC is an outdated and confusing term, communicating if an isolate is *E. hormaechei*

or another *Enterobacter* species is likely more relevant and useful in reporting and diagnostics. Wu et al., (2020) proposed reclassifying an *E. cloacae* subspecies to a separate species called *E. dissolvens*, returning to the previously published species group *E. xiangfangensis* and creating *E. hoffmanii* instead of these being a subspecies of *E. hormaechei* (Wu et al., 2020). The biggest challenge is how to communicate taxonomic changes to the scientific community and increase usage of updated taxonomic classification systems in the future.

Inaccurate communication also occurs when methods like MALDI-TOF (Lasch, Peter et al., 2023) or MLST (Jolley & Maiden, 2014) are used to confirm species. Now with enhanced understanding of genomics and species groups, our work can provide a better strain list for *Enterobacter* to build the MALDI profiles for this group and improve the organism database. As key components of the commercial MALDI systems (Singhal et al., 2015), organism databases must reflect updated taxonomic names and the discovery of new microbial species and annotations. Expanding the pubMLST database to include new allelic combinations that currently have no ST associated, would improve our understanding of associations between species and MLST profile, improving this tools usefulness in identification.

4.3 Limitations and restrictions within this study

This study was a preliminary glance at the speciation boundaries in *Enterobacter*. The datasets used for our analysis were a snapshot of genome assemblies online. There are tens of thousands of read sets available for further interrogation. The isolates we downloaded and examined were not evenly distributed across geography or time, which may have important impacts on the conclusions.

Another limitation is the uneven sampling across species groups. Some species groups contained hundreds of isolates and others only held 2-5 samples. Ideally, the representation of within and between groups numbers would be more even. Looking more closely at the diversity and differences within each species group could help establish improved within-species metrics.

Despite successfully grouping isolates, our genomic framework required manual curation to confidently allocate species names. Many samples that were labelled “unknown” may represent novel species, and the allocation of species names was influenced by previous naming conventions. Cross referencing the isolates evaluated by Wu et al., (2020) and the species names they proposed with our data, may resolve some of these difficult to

classify groups. Using incorrect naming conventions that contain errors will be passed forward in the manual curation step our genomic framework.

We lack full knowledge about the function of many genes and do not understand the complex processes that influence gene regulation and transcription. This was highlighted by our Twilight (Horesh, n.d.) and Panaroo (Tonkin-Hill et al., 2020) results that reported numerous "hypothetical" proteins with unknown function. Low sample numbers in some species groups made pan-genome and lineage specific inferences challenging. Species where subspecies groups exist, may need to be investigated at this deeper level to make meaningful conclusions about unique gene markers or core genes.

4.4 Conclusions and future directions

In summary, whole genome sequencing offers powerful tools to address many of the challenges in bacterial taxonomy, providing a more precise, comprehensive, and evolutionary-based framework for classifying and understanding bacterial diversity. We confirmed that 785 (52%) of the isolates analysed were *E. hormaechei*, and 105 (58%) isolates called *E. cloacae* are misnamed in the reference databases we accessed. A threshold of 0.04 does successfully delineate species group in *Enterobacter* and there is good evidence to suggest MLST sequence types are unique within each species groups. Diagnostically this is important as the vast majority of isolates in clinical setting are *E. hormaechei* and MLST can be used to exclude samples from this group or differentiate isolates within it. Although species specific markers exist for some groups, our research identified none for the most common *E. hormaechei* group, however, reassigning species groups to match the naming system used by Wu et al., (2020) may reveal different markers. Antimicrobial genes are present in all except 15 isolates and genes that could confer resistance to each drug class are present in each species group.

This research is a starting point for many other areas of work, which could include; sequencing a wider variety of *Enterobacter* genomes, improving MALDI-TOF, improving pubMLST, and investigating the association of genotype and phenotype in AMR.

Increasing the number and diversity of quality genomes held in databases from animal and environmental sources, will help identify novel species. Sufficient numbers of complete genomes in less common species will improve the reference capabilities of these databases. Tools like Twilight (Horesh, n.d.) will have increased robustness when pan-genome analysis is completed, if the minimum of ten samples exists in each species

group.

MALDI-TOF is a quick, cost-effective way for hospitals to identify a pathogen but in *Enterobacter* it's only accurate to genus level (Pavlovic et al., 2012). MALDI-TOF MS has been shown to improve identification of microorganisms and improve antimicrobial stewardship and disease prevention to benefit public health (Rodríguez-Sánchez et al., 2019). If the sensitivity of this method could be improved to accurately identify species in *Enterobacter* and determine antimicrobial susceptibility patterns to provide same day results, this would be a huge advantage in clinical practice. Clinicians could confidently prescribe the most effective antimicrobial and suspected outbreaks would be identified more quickly. All providing cost-benefits to the health service (Forde et al., 2023).

Expansion of the pubMLST database could also be implemented following our work. Many of the isolates did not have a sequence type to match their allelic variation. When more high-quality samples are uploaded to pubMLST the typing scheme will improve. There is also an argument to support renaming of the collection from *Enterobacter cloacae* to *Enterobacter spp.* as this more accurately reflects the diversity of species in the collection.

AMR profiling was only a small component of our study but the alarming prevalence of AMR genes across all species groups warrants further investigation. Completing susceptibility studies to understand the association of antimicrobial phenotype with genotype would be extremely helpful for public health officials and clinicians, especially if this could be linked to a reliable accurate species identification method. Given the predominate species group is *E. hormaechei*, and it shows the largest proportion of isolates with resistance genes across most drug types, investigating AMR patterns within this species would provide public health officials greater insights into the relationship of healthcare environments and resistance evolution.

Following this research, a direct improvement of public health protocols and decisions in relation to *Enterobacter* in Victoria, Australia will occur. MDU PHL standard operating procedures will be updated to include the curated species names for isolates used in this study.

Appendix A

Preliminary Dataset Genomes

https://github.com/S-Noonan/MDU.Research.Project/blob/main/ECC_preliminary_genomes.txt

Appendix B

Extended Dataset Genomes

https://github.com/S-Noonan/MDU.Research.Project/blob/main/ECC_extended_genomes.txt

Appendix C

Duplicated genomes

Duplicate of	Removed
Enterobacter_bugandensis_ECH10	Enterobacter_bugandensis_ECH9
Enterobacter_cancerogenus_ATCC_35316	Enterobacter_cancerogenus_GCF_902373965.1
Enterobacter_ludwigii_DLL7524	Enterobacter_ludwigii_GCF_902387865.1
Enterobacter_roggenkampii_DS05262	Enterobacter_roggenkampii_GCF_902387855.1
Enterobacter_sichuanensis_GCF_902387735.1	Enterobacter_cloacae_DS15987
Enterobacter_timonensis_GCF_902375915.1	Enterobacter_timonensis_mt20

Appendix D

BANCSIA Species Allocation Script

https://github.com/S-Noonan/MDU.Research.Project/blob/main/BANCSIA.py

```
# BANCSIA - Bacterial Naming for Correct Species Identification and
    Allocation
# created by: Susan Noonan
# Python script to determine species groups for ECC Masters research
    project.  This script can be used to determine species groups based
    on genetic relatedness and a set threshold distance.

# Input1 = the output from running ska dist (.tsv) file. \url{https://
    github.com/simonrharris/SKA}
# Input2 = threshold distance relevant to your analysis (eg. 0.02 for
    subspecies, 0.04 for species)
# Output = .csv file with isolate name and group number


!pip install pandas
import pandas as pd

# Functions
# Extract samples within a set distance
def extract\_groups(df, sample\_of\_interest, max\_distance):
    filtered\_df = df[(df['S1'] == sample\_of\_interest) \& (df['dist']
    <= max\_dist) | (df['S2'] == sample\_of\_interest) \& (df['dist']
   <= max_dist)]\\
     samples\_within\_dist1 = list(set(filtered_df['S1']))\\
     samples\_within\_dist2 = list(set(filtered_df['S2']))\\
     samples\_with\_dist = list(set(samples\_within\_dist1 + samples\
   _within\_dist2))\\
     if len(samples\_with\_dist) == 0:\\
         samples\_with\_dist = [sample\_of\_interest]\\
     return samples\_with\_dist}\\
```

```
# Check if two lists overlap
def check_overlap(list1, list2):
    # Convert the lists to sets and check if their intersection is non-empty
    return bool(set(list1) & set(list2))

# Find new items to add from two overlapping lists
def list_diff(list1, list2):
    # Convert lists to sets and get all the unique items
    diff = sorted(list(set(list1 + list2)))
    return diff

# Check dictionary to see if list already exists, if not add new key
def check_dict(dict, list):
    for key in dict.keys():
        if check_overlap(list, dict[key]) == True:
            return key
        else:
            continue
    return False
    # if there are no matches, add a new item to the dict

# Allocate each item in input data to a species group and output results
def species_groups(df):
    # Dictionary to store results
    result_samples = {}
    unique_sample_names = pd.unique(df[['S1', 'S2']].values.ravel('K'))
    i = 1
    for sample in unique_sample_names:
        group = sorted(extract_groups(df, sample, max_dist))
        # check dictionary
        if len(result_samples) == 0:
            # there are no items in dict, so add list of samples
            result_samples[i] = group
            i+=1
        else:
            if check_dict(result_samples, group) == False:
                # there were no values that overlap, so add new item to dict
                result_samples[i] = group
                i+=1
            else:
                index = check_dict(result_samples, group)
```

```
                    new = list_diff(group, result_samples[index])
                    result_samples[index] = new
                    i+=1
        return result_samples

# Create a dataframe from the dictionary created with each isolates
    allocated a species group number
def df_with_spec(dict):
        # create empty dataframe
        empty_df = pd.DataFrame(columns = ["isolate", "spec_no"])
        for i, entry in enumerate(dict):
            for j in dict[entry]:
                empty_df.loc[len(empty_df)] = [j, i+1]
        return empty_df
```

Listing D.1: BANCSIA species allocation script

Code to run script

To run python species grouping analysis on output from the ska results data-frame use the following:

```
# read file
file_path = "/path_to_file/ska_output.csv"
# Turn CSV file into a DataFrame
df = pd.read_csv(file_path)
# set max_dist
max_dist = 0.04
# run function to get species groups
grouping = species_groups(df)
# turn output into dataframe where each isolate has a grouping number
df_ouput = df_with_spec(grouping)
# save this file as csv and use for further analysis in R
df_ouput.to_csv('/path_to_file/allocation_result.csv', index=True)

# Manually check your results to determine which species names are
    appropriate or complete further analysis using group number
```

Listing D.2: running BANCSIA

Appendix E

Genomes categorised as "unknown"

Table E.1: Genomes with ≤ 50 contigs classified as "unknown" after curation using BANCSIA. These isolates were unable to be categorised to a known species group after applying the 0.04 threshold.

Isolate	Dataset	Group
Enterobacter_cloacae_CZ-1 Enterobacter_ludwigii_11894-yvys Enterobacter_sp._A11 Enterobacter_sp._E1	Preliminary Extended Extended	Unknown 11
Enterobacter_cloacae_WP5-S18-ESBL-01 Enterobacter_mori_HSW1412 Enterobacter_sp._RHBSTW-00901	Preliminary Preliminary Extended	Unknown 14
Enterobacter_asburiae_E.a101	Extended	Unknown 17
Enterobacter_cloacae_D41-sc-1712200 Enterobacter_cloacae_DSM_26481 Enterobacter_sp._I4	Extended	Unknown 18
Enterobacter_sp._AD2-3	Extended	Unknown 22
Enterobacter_sp._BIGb0359 Enterobacter_sp._BIGb0383	Extended	Unknown 23
Enterobacter_sp._CC120223-11	Extended	Unknown 43
Enterobacter_sp._Colony194 Enterobacter_sp._E76	Extended	Unknown 24
Enterobacter_sp._JGM127	Extended	Unknown 25
Enterobacter_sp._KPR-6	Extended	Unknown 26
Enterobacter_sp._RHBSTW-00175	Extended	Unknown 27
Enterobacter_sp._RHBSTW-00994	Extended	Unknown 28
Enterobacter_sp._RIT712	Extended	Unknown 29
Enterobacter_sp._SA187	Extended	Unknown 30
Enterobacter_cloacae_CH1	Extended	Unknown 33
Enterobacter_cloacae_JD6301 Enterobacter_sp._9-2	Extended	Unknown 34
Enterobacter_cloacae_P40C Enterobacter_cloacae_P40C2 Enterobacter_cloacae_P40RS	Extended	Unknown 35
Enterobacter_cloacae_S18121600014	Extended	Unknown 36
Enterobacter_cloacae_e483 Enterobacter_genomosp._GN03164	Extended	Unknown 37
Enterobacter_kobei_GCF_900185885.1	Extended	Unknown 39
Enterobacter_sp._CC120223-11	Extended	Unknown 41
Enterobacter_sp._RIT_418	Extended	Unknown 42
Enterobacter_sp._Tr-810	Extended	Unknown 43

Appendix F

Species Separation - Extended Refseq

When the 0.04 threshold is applied each species should have within values below the threshold and between values above the threshold. When outliers cross the threshold values this may indicate inaccurate species allocation.

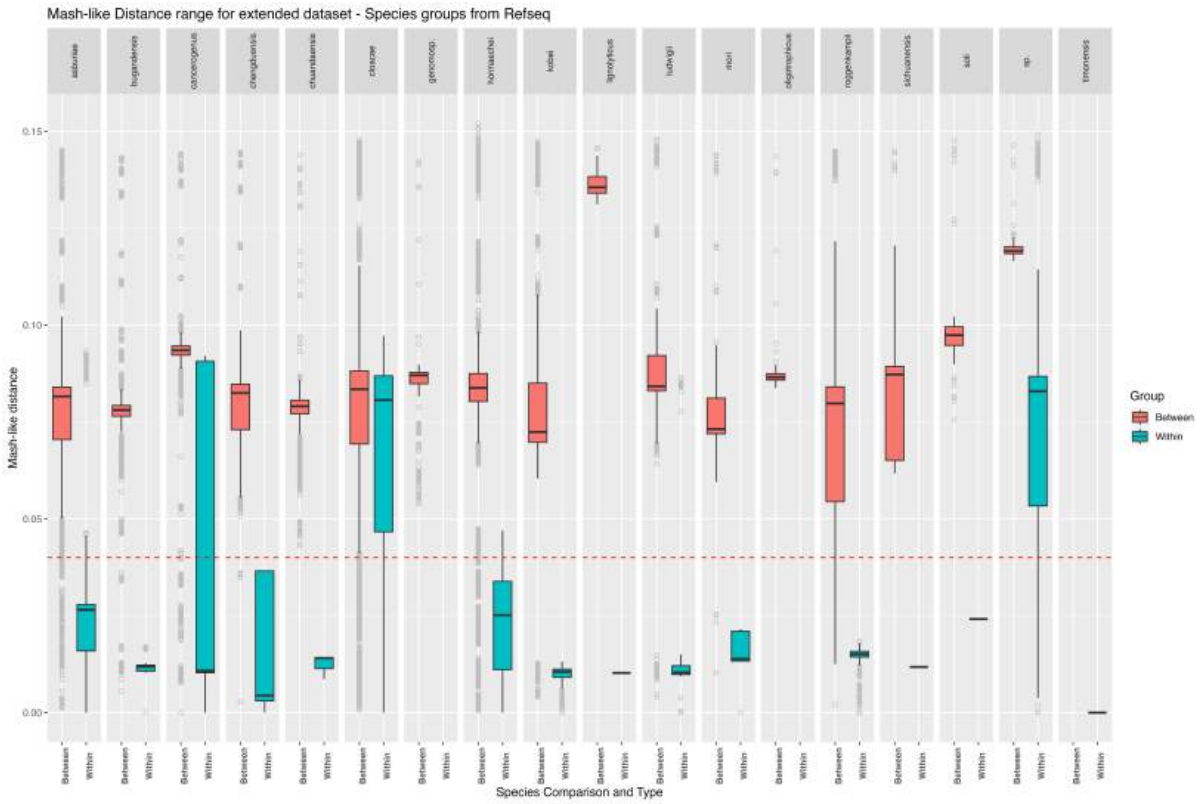


Figure F.1: Within and between species boxplot for 759 high quality isolates from the extended dataset. Species groups were extracted from the Refseq taxid_id metadata. Clear separation of species should occur at 0.04 threshold indicated by the dotted red line. This is not the case for a majority of the species groups, indicating species names may be inaccurate or mis-assigned

Appendix G

Phylogenetic Tree Extended Curated

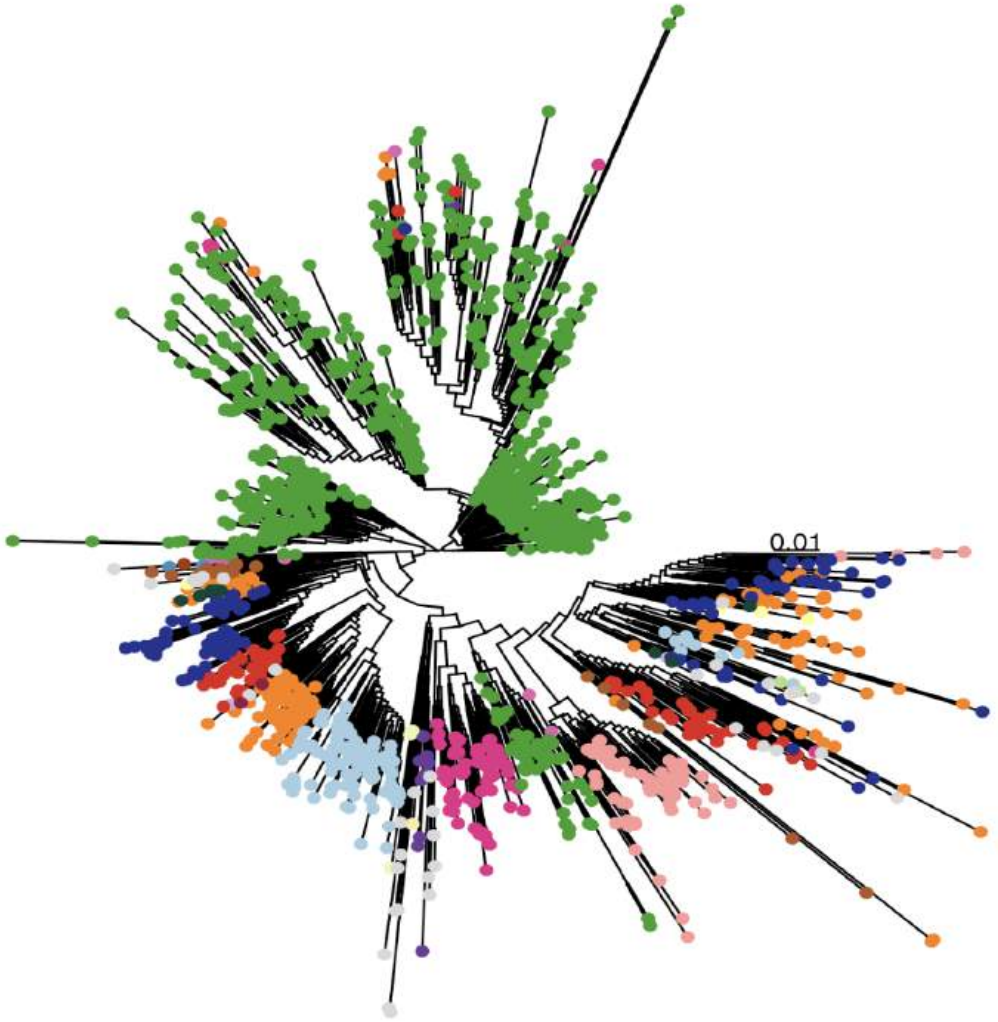


Figure G.1: Phylogenetic tree for isolates from the extended dataset that met the quality threshold of ≤ 50. The tree is coloured by curated species name and the tree in Appendix H is coloured by species names extracted from Refseq tax_id. Many of the *E. sp.* (grey) and *E. cloacae* (pink) species from the preliminary data phylogenetic tree are resolved in this curated tree. There are a few branches where species allocation and phylogenetic relatedness do not align as many coloured isolates are still present in one branch.

Appendix H

Phylogenetic Tree Extended Refseq

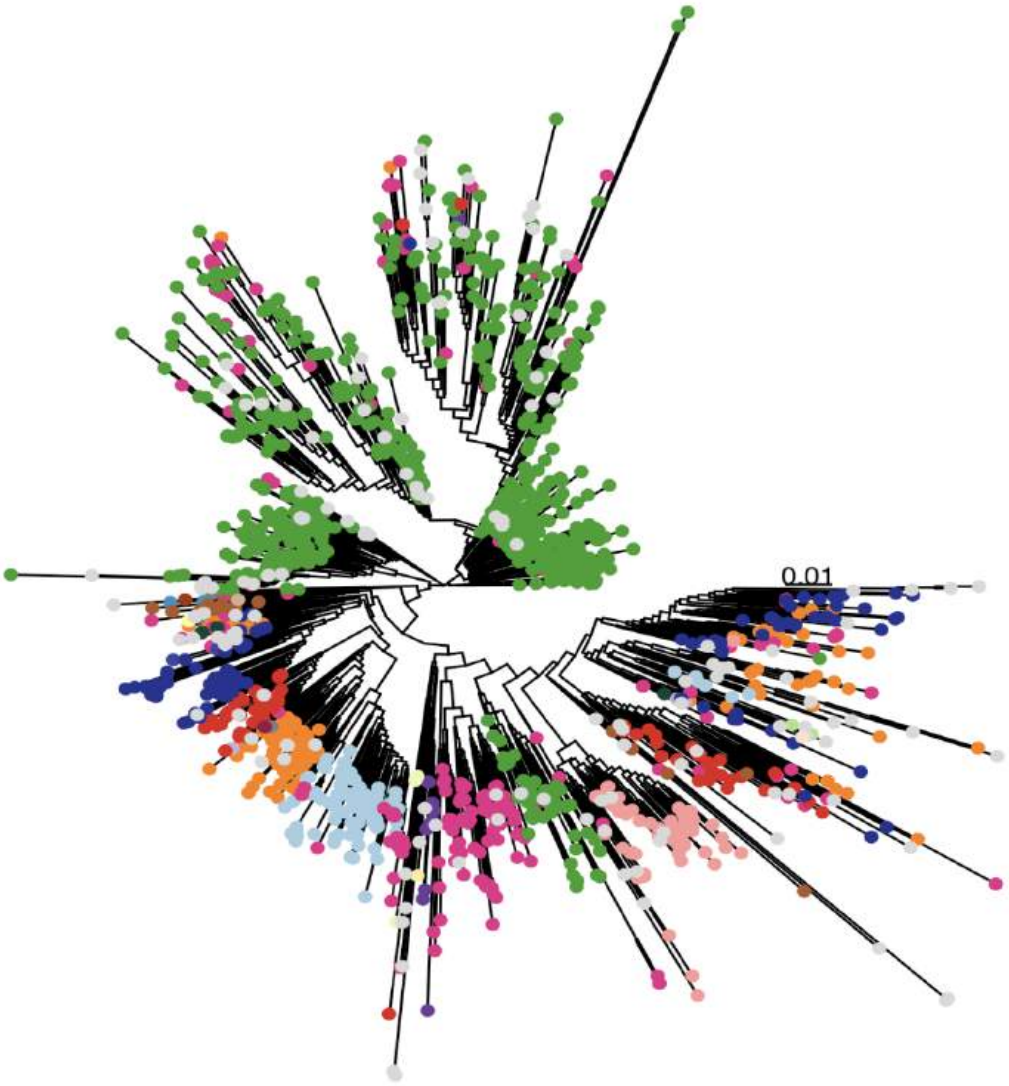


Figure H.1: Phylogenetic tree for isolates from the extended dataset that met the quality threshold of ≤ 50. The tree is coloured by Refseq_tax_id names. Many of the *E. sp.* (grey) and *E. cloacae* (pink) species in this tree are resolved in the previous curated tree (Appendix G).

Appendix I

Lineage specific gene markers

Table I.1: Lineage specific genes identified from Twilight and Panaroo. Operons are gene clusters that are a minimum of tree genes co-located in the genome and are indicated by *first operon, **second operon, ***third operon.

Species	**Gene**	**Function**
E. bugandensis	group_49098	hypothetical
	* group_49603	hypothetical
	* group_9813	hypothetical
	lpfD_1∼∼∼lpfD_2	minor fimbrial subunit
E. cancerogenus	group_5235	hypothetical
	group_6811	hypothetical
	group_3293	hypothetical
	group_11166	hypothetical
	gltF_2∼∼∼gltF_3 ∼∼∼gltF_4	glutamate synthesis
	group_20896	hypothetical
	ydeP_2	oxidoreductase, acid resistance
	group_22107	hypothetical
	* group_3891	hypothetical
	* group_39211	hypothetical
	* group_4712	hypothetical
	* group_4724	hypothetical
	* group_51449	hypothetical
	group_6783	hypothetical
	group_49641	hypothetical
	group_19031	hypothetical
	group_45634	hypothetical
	group_35763	hypothetical
	group_41525	hypothetical
	group_16863	hypothetical
	** ptrB_3∼∼∼ptrB_2 ∼∼∼ptrB_1	protease
	** group_9413	hypothetical
	** group_12685	hypothetical
	*** group_26188	hypothetical
	*** group_48886	hypothetical
	*** group_33469	hypothetical

Species	Gene	Function
E. cloacae	group_38017	hypothetical
	group_1509	hypothetical
	group_24872	hypothetical
	group_46526	hypothetical
	group_47529	hypothetical
	* group_49488	soluble epoxide hydrolase
	* group_45076	HTH type transcriptional repressor ComR
	* group_49461	hypothetical
E. genomosp.	group_38507	hypothetical
E. kobei	aat	tRNA acyltransferase
	* fimG_3∼∼∼fimG_2	regulates fimbriae length
	* group_45644	hypothetical
	* fimC∼∼∼fimC_2∼∼∼fimC_1 ∼∼∼fimC_3	Biogenesis of type I fimbriae
	* group_32930	hypothetical
	** arsA_1∼∼∼arsA	aresenical pump-driving ATPase
	** arsD_1∼∼∼arsD∼∼∼arsD_2	Arsenical resistance operon
	** group_12501	hypothetical
	*** group_42945	type II secretion system protein
	*** group_53193	type II secretion system protein
	*** group_6445	type II secretion system protein
	*** gspL_2∼∼∼gspL_1∼∼∼gspL	type II secretion system protein
	*** gspM	type II secretion system protein
E. ludwigii	group_48758	hypothetical
	group_25897	hypothetical
	gltF_3∼∼∼gltF_2∼∼∼gltF_4	glutamate biosynthesis
	* group_21282	hypothetical
	* group_13501	hypothetical
	* group_26420	Lectin A
	group_12791	hypothetical
	rarA	replication assoc. recombination
	group_48851	hypothetical
	group_42308	hypothetical
	group_50244	hypothetical
	group_39311	hypothetical
	** group_50610	hypothetical
	** dhaA	soluble expoxide hydrolase
	** group_50443	HTH type transcriptional repressor ComR
	dmlR_12	HTH type transcriptional regulator DmlR
	group_31329	hypothetical
	group_54009	hypothetical
	pcpR_1∼∼∼pcpR_2	PCP degradation transcriptional activation
	group_42734	hypothetical
	attM	N-acyl homoserine lactonase
	group_53432	hypothetical
	group_39542	hypothetical
	group_13283	hypothetical
	group_52164	hypothetical
	group_50532	mannose-6 phosphate isomerase
	group_12569	L-rhamnose mutarotase
E. mori	group_47532	hypothetical
	group_2177	hypothetical
	group_2684	hypothetical
	group_50565	hypothetical
E. sichuanensis	group_54463	HTH type transcriptional regulator VirS
	group_53307	Hypothetical
	group_46810	oxygen regulatory protein NreC

Appendix J

AbritAMR gene count per species

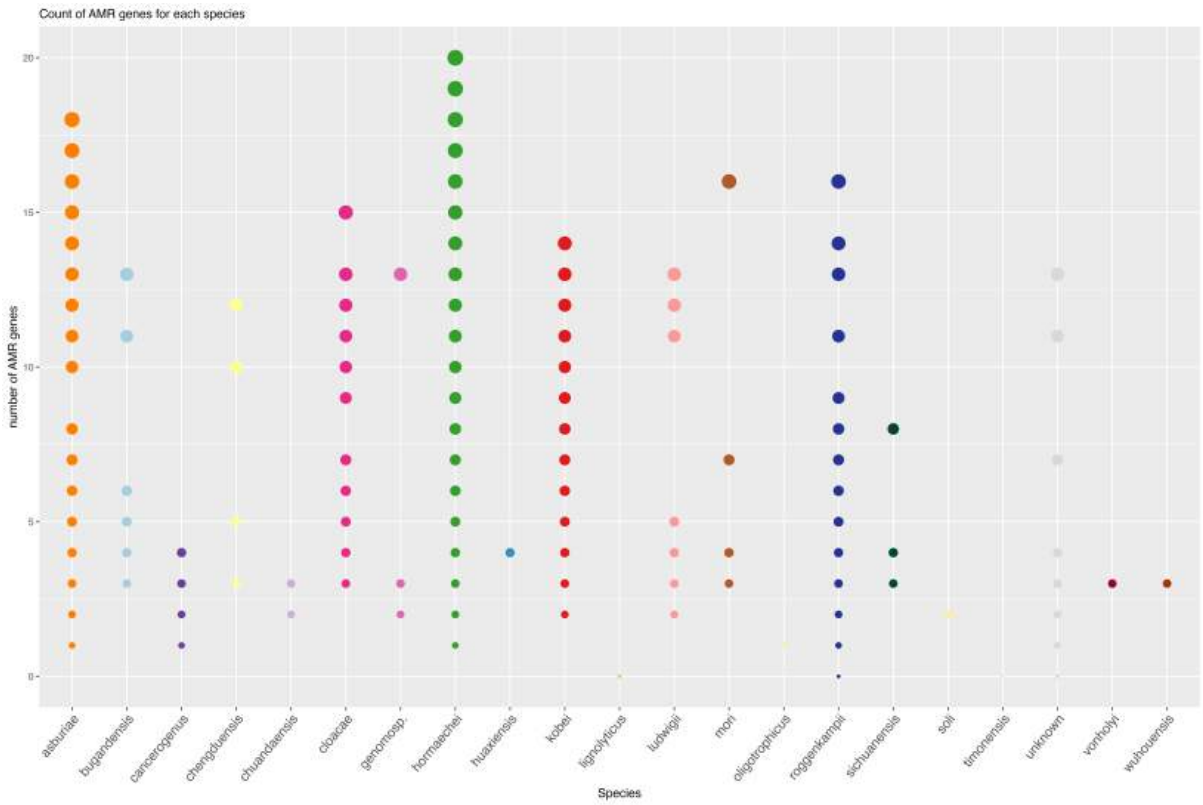


Figure J.1: Count of AMR genes per isolate for species group. The size of the dot represents the number of isolates with that count. *E. asburiae*, *E. bugandensis*, *E. chengduensis*, *E. cloacae*, *E. genomosp.*, *E. hormaechei*, *E. kobei*, *E. ludwigii*, *E. mori*, *E. roggenkampii* and the unknown group contain isolates with more then the Q3 value of nine AMR genes.

Appendix K

Drug class resistance per species

Figure K.1: Count of resistance genes per isolate that could confer resistance to each drug class reported by AbritAMR. The isolates are grouped into species class represented by colour. *E. asburiae* is orange, *E. bugandensis* is light blue, *E. cloacae* is pink, *E. hormaechei* is bright green, *E. kobei* is red, *E. ludwigii* is peach, *E. mori* is brown, *E. roggenkampii* is dark blue, *E. sichuanensis* is dark green. Each species contains isolates with one or multiple resistance genes across all drug class types.

Appendix L

Resistance genes by Species

The resistance genes reported by AbritAMR are outlined tables for the following species

- *Enterobacter asburiae*
- *Enterobacter bugandensis*
- *Enterobacter cloacae*
- *Enterobacter hormaechei*
- *Enterobacter kobei*
- *Enterobacter ludwigii*
- *Enterobacter roggenkampii*

Table L.1: Antimicrobial resistance genes identified by AbirtAMR for each drug class, resistance type and species. Results are limited to samples with ≤ 50 contigs for the largest species groups in our study. Each drug resistance type reported by AbritAMR is grouped by drug class. Some isolates contained multiple genes and combinations of the genes listed. For simplicity the gene names only are reported.

Class	**Resistance Type**	**List of genes**			
		E. asburiae	***E. bugandensis***	***E. cloacae***	***E. hormaechei***
Penicillins	Beta-lactamase (not ESBL or carbapenemase)	blaCARB-2 blaOXA-1 blaTEM-1 blaOXA-9	blaOXA-1 blaTEM-1 blaLAP2	blaOXA-1 blaTEM-1 blaOXA-9 blaLAP-2 blaOXA-10	blaOXA-1 blaTEM-1 blaOXA-9 blaLAP-2 blaOXA-10 blaCARB-2 blaOXA-2 blaSCO-1 blaSED blaTEM-2
	Beta-lactamase (unknown spectrum)				blaOXA-129 blaOXA-216 blaTEM-238
Aminoglycosides	Amikacin/Gentamicin/ Kanamycin/Tobramycin				aac(6')-Ie aph(2")-Ia
	Amikacin/Kanamycin				aph(3')-VI aph(3')-VIa aph(3')-VIb aph(3')-XV
	Amikacin/Kanamycin/ Tobramycin	aac(6')-Ib aac(6')-Ib3		aac(6')-Ib3	aac(6')-Ib aac(6')-Ib3 aac(6')-Il aac(6')-Ian
	Amikacin/Kanamycin/ Tobramycin/Quinolone	aac(6')-Ib-cr5 aac(6')-Ib-cr7	aac(6')-Ib-cr5	aac(6')-Ib-cr5	aac(6')-Ib-cr5 aac(6')-Ib-cr7
	Aminoglycosides (Ribosomal methyltransferase)			rmtB1 rmtC	rmtB1 rmtC armA rmtB4
	Other aminoglycoside resistance (non-RMT)	aac(6')-Ia			aac(6')-30 aac(6')-lae aph(4)-la
	Gentamicin	aac(6')-Ib aac(3)-IIb aac(3)-IId aac(3)-IIe aac(3)-IIg aac(6')-Ib4	aac(3)-IId aac(3)-IIe	aa(3)-IId aac(3)-IIe aac(6')-Ib	aac(6')-Ib aac(3)-IId aac(3)-IIe aac(3)-IIg aac(6')-Ib4 aac(3)-Ia aac(3)-Ib
	Gentamicin/Kanamycin/ Tobramycin	aac(6')-IIc ant(2")-Ia		ant(2")-Ia	aac(6')-IIc ant(2")-Ia
	Gentamicin/Tobramycin/ Apramycin				aac(3)-IVa
	Streptomycin	aadA1 aadA2 aadA15 aph(3")-Ib aph(6)-Id	aadA1 aph(3")-Ib aph(6)-Id	aadA2 aph(3")-Ib aph(6)-Id aadA1 aadA21 aadA15 aadA22 aadA5	aadA1 aadA2 aph(3")-Ib aph(6)-Id aadA16 aadA13 aadA22 aadA5 aph(6)-Id aadA6

Table L.2: Antimicrobial genes continued.

Class	**Resistance Type**	**List of genes**			
		E. asburiae	***E. bugandensis***	***E. cloacae***	***E. hormaechei***
Carbapenems	Carbapenemase	blaFRI-11 blaFRI-4 blaFRI-6 blaFRI-8 blaFRI-9 blaGES-5 blaIMI-1 blaIMI-2 blaIMI-12 blaKPC-2 blaKPC-3 blaNMC-A blaOXA-48	blaIMI-1 blaIMI-20	blaGES-5 blaIMI-1 blaKPC-2 blaOXA-48	blaKPC-2 blaKPC-3 blaOXA-48 blaBKC-2 blaFLC-1 blaIMI-2 blaKPC-4 blaKPC-6
	Carbapenemase (MBL)	blaIMP-1 blaIMP-11 blaIMP-26 blaIMP-4 blaIMP-60 blaIMP-8 blaNDM-1 blaVIM-1	blaNDM-5	blaNDM-1 balNDM-5	blaIMP-1 blaIMP-26 blaIMP-4 blaIMP-8 blaNDM-1 blaVIM-1 blaNDM-5 blaVIM-4 blaGIM-1 blaIMP-13 blaIMP-96 blaKHM-1 blaNDM-7
Amphenicol	Chloramphenicol	catA1 catB3 catA2 cmlA6	catA1	catA2 catA1 catB8	catA1 catB3 catA2 catB11 cmlA6 cmlA1 catB2 cmlA10 cmlA5 cmlB1
	Chloramphenicol/ Florfenicol			floR	floR floR2
Polymyxins	Colistin	mcr-10.1 mcr10.4 mcr-9		mcr-10.1 mcr-9 mcr-10.1	mcr-10.1 mcr-9
Macrolides	Erythromycin				
	Rifamycin	arr arr-3			arr arr-3 arr-2
	Macrolide			mph(A)	ere(A) mph(A) mph(E) msr(E)

Table L.3: Antimicrobial genes continued.

Class	**Resistance Type**	**List of genes**			
		E. asburiae	***E. bugandensis***	***E. cloacae***	***E. hormaechei***
Cephalosporins	ESBL	blaCTX-M-14 blaCTX-M-3 blaCTX-M-15 blaCTX-M-2 blaCTX-M-9 blaSFO-1 balSHV-12	blaCTX-M-15 blaCTX-M-3 blaCTX-M-55 blaSHV-2	blaCTX-M-15 blaSHV-12 blaCTX-M-65 blaGES-1	blaCTX-M-15 blaCTX-M-2 blaCTX-M-9 blaSHV-12 blaCTX-M-3 blaCTX-M-55 blaCTX-M-65 blaCTX-M-14 blaCTX-M-236 blaCTX-M-3 blaCTX-M-36 blaSFO-1 blaSHV-30 blaSVH-7 blaVEB-3 blaSVH-5 blaSVH-7 blaTEM-12
	ESBL (AmpC type)	blaACT-1 blaACT-105 blaACT-13 blaACT-2 blaACT-29 blaACT-3 blaDHA-1 blaACT-34 blaACT-38 blaACT-4 blaACT-57 blaACT-58 blaACT-6 blaACT-62 blaACT-68 blaACT-8 blaFOX-3 blaMIR-14	blaACT-49 blaACT-76 blaACT-77 blaACT-78 blaACT-80 blaACT-82	blaCMH-2 blaCMH-3 blaCMY-3 blaCMY-6 blaCMH-4 blaCMH-5 blaCMH-7	blaACC-1 blaACT-17 blaACT-24 blaACT-106 blaDHA-1 blaACT-14 blaACT-15 blaACT-16 blaFOX-5 blaACT-17 blaACT-18 blaACT19 blaACT-23 blaACT-24 blaACT-25 blaACT-27 blaACT-32 blaACT-35 blaACT-36 blaACT-37 blaACT-40 blaACT-41 blaACT-42 blaACT-43 blaACT-44 blaACT-45 blaACT-46 blaACT-47 blaACT5 blaACT55 blaACT-56 blaACT-65 blaACT-66 blaACT-67 blaACT-69 blaACT-70 blaACT-72 blaACT-74 blaACT-75 blaACT-84 blaACT-85 blaACT-89 blaACT-90

Table L.4: Antimicrobial genes continued.

Class	**Resistance Type**	**List of genes**			
		E. asburiae	***E. bugandensis***	***E. cloacae***	***E. hormaechei***
Phosphonic	Fosfomycin	fosA fosA2	fosA fosA2 fosA7.2	fosA fosA2 fosA7.3 fosA7.4	fosA fosA2 fosA7.4 fosA5 fosA3 fosG
Lincosamides	Lincosamides			Inu(G)	InuG InuF
Other	Other antimicrobial	ble	ble	ble	ble bleO sat2
Quinolones	Phenicol/Quinolone	oqxA oqxB9	oqxA oqxB9 oqxB17	oqxA oqxB9 oqxB oqxB3	oqxA oqxB9 oqxB oqxB27
	Quinolone	qnrA1 qnrE4 qnrB10 qnrB2 qnrB4 qnrE1 qnrS1	qnrB1 qnrE3 qnrS1	qnrB2 qnrS1 qnrB1 qnrE3	qnrA1 qnrB2 qnrB4 qnrS1 qnrB1 qnrA9 qnrB39 qnrB19 qnrB5 qnrB6 qnrB7
Sulfonamides	Sulfonamide	sul1 sul2	sul2	sul1 sul2	sul1 sul2 sul3
Tetracyclines	Tetracycline	tet(A) tet(D)	tet(A) tet(D)	tet(A) tet(D) tet(C) tet(X4) tet(B)	tet(A) tet(D) tet(C) tet(X4) tet(B) tmexD2 toprJ2 tet(M) tet(G) tet(X5)
Diaminopyrimidines	Trimethoprim	dfrA14 dfrA16 dfrB1 dfrA17 dfrA19 dfrA8 dfrB1	dfrA14	dfrA14 dfrA12 dfrA15 dfrA17	dfrA14 dfrA16 dfrA19 dfrA12 dfrA15 dfrA17 dfrA1 dfrB1 dfrA27 dfrA21 dfrA8 dfrA22 dfrA25 dfrA5 dfrB3

Table L.5: Antimicrobial genes continued.

Class	**Resistance Type**	**List of genes**		
		E. kobei	***E. ludwigii***	***E. roggenkampii***
Penicillins	Beta-lactamase (not ESBL or carbapenemase)	blaOXA-1 blaTEM-1 blaLAP-2 blaOXA-2	blaOXA-1 blaTEM-1	blaTEM-1 blaLAP-2 blaLAP-2 blaTEM-1
	Beta-lactamase (unknown spectrum)	bla		
Aminoglycosides	Amikacin/Gentamicin/ Kanamycin/Tobramycin			
	Amikacin/Kanamycin	aph(3')-XV		
	Amikacin/Kanamycin/ Tobramycin	aac(6')-Ib3		aac(6')-Ib3 aac(6')-I
	Amikacin/Kanamycin/ Tobramycin/Quinolone	aac(6')-Ib-cr5	aac(6')-Ib-cr5	aac(6')-Ib-cr5
	Aminoglycosides (Ribosomal methyltransferase)	armA		
	Other aminoglycoside resistance (non-RMT)	aac(2')-IIa		aac(6')-31
	Gentamicin	aac(6')-Ib aac(3)-IIg aac(6')-Ib4 aac(3)-IId	aac(3)-IId	aac(6')-Ib aac(3)-IId aac(3)-IIg
	Gentamicin/Kanamycin/ Tobramycin	aac(6')-IIc ant(2")-Ia		aac(6')-IIc ant(2")-Ia
	Gentamicin/Tobramycin/ Apramycin			
	Kanamycin		aph(3')-Ia	aph(3')-Ia
	Streptomycin	aadA1 aadA2 aph(3")-Ib aph(6)-Id aadA11 aadA5	aph(3")-Ib aph(6)-Id aadA16 aadA5 aadA2	aadA1 aadA2 aph(3")-IIb aph(6)-Id aadA16 aadA5
Carbapenems	Carbapenemase	blaKPC-2 blaOXA-48 blaGES-24	blaNMC-A	blaFRI-8 blaGES-5 blaIMI-1 blaKPC-2
	Carbapenemase (MBL)	blaIMP-1 blaIMP-4 blaNDM-1 blaVIM-1		blaIMP-1
Amphenicol	Chloramphenicol	catA1 catB3 catA2 catB2	catB3	catA2 catA1
	Chloramphenicol/ Florfenicol	floR		floR
Polymyxins	Colistin	mcr-10.1 mcr-9 mcr-10.2 mcr-4.3	mcr-10.1 mcr-9	mcr-10.1 mcr-9
Macrolides	Erythromycin			
	Rifamycin	arr arr-3	arr-3	arr", "arr-3
	Macrolide	mph(A) mph(E) msr(E)	mph(A)	mph(A)

Table L.6: Antimicrobial genes continued.

Class	**Resistance Type**	**List of genes**		
		E. kobei	***E. ludwigii***	***E. roggenkampii***
Cephalosporins	ESBL	blaCTX-M-9 blaSHV-12 blaCTX-M-3 blaSFO-1 blaOXA-17	blaSFO-1	blaCTX-M-9 blaSHV-12
	ESBL (AmpC type)	blaACT-102 blaACT-103 blaACT-104 blaACT-28 blaACT-51 blaACT-52 blaACT-64 blaACT-87 blaACT-9 blaACT-95 blaACT-98 blaACT-99	blaACT-109 blaACT-12 blaACT-54	blaACT-29 blaACT-62 blaMIR-10 blaMIR-11 blaMIR-12 blaMIR-13 blaMIR-15 blaMIR-16 blaMIR-17 blaMIR-18 blaMIR-20 blaMIR-21 blaMIR-22 blaMIR-23 blaMIR-3 blaMIR-5 blaMIR-7 blaMIR-9
Phosphonic	Fosfomycin	fosA fosA2	fosA2 fosA3 fosA7.3	fosA fosA2 fosI
Lincosamides	Lincosamides			
Other	Other antimicrobial	sat2 ble		
Quinolones	Phenicol/Quinolone	oqxB9 oqxA oqxB	oqxA oqxB9	oqxA oqxB9
	Quinolone	qnrA1 qnrB2 qnrS1 qnrB19 qnrS2	qnrA1 qnrS1 qnrB6	qnrA1 qnrE1 qnrE4 qnrS1 qnrB19 qnrB6 qnrS2 qnrA1 qnrB6
Sulfonamides	Sulfonamide	sul1 sul2	sul1 sul2	sul1 sul2
Tetracyclines	Tetracycline	tet(A) tet(D) tet(C)	tet(C) tet(B) tet(D)	tet(A) tet(D)
Diaminopyrimidines	Trimethoprim	dfrA14 dfrA16 dfrA12 dfrA1 dfrA25	dfrA17 dfrA12 dfrA14 dfrA27	dfrA14 dfrA16 dfrA19 dfrA1 dfrA27 dfrA15

References

Annavajhala, M. K., Gomez-Simmonds, A., & Uhlemann, A.-C. (2019). Multidrug-Resistant *Enterobacter cloacae* Complex Emerging as a Global, Diversifying Threat. Frontiers in Microbiology, 10, 44. https://doi.org/10.3389/fmicb.2019.00044

Antimicrobial Resistance Multi-Partner Trust Fund. (2022). Countering Antimicrobial Resistance with a 'One Health' Approach. https://mptf.undp.org/fund/amr00

Arnold, B. J., Huang, I.-T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. Nature Reviews Microbiology, 20(4), 206–218. https://doi.org/10.1038/s41579-021-00650-4

Bell, J. M., Gottlieb, T., Daley, D. A., & Coombs, G. W. (2020). Australian Group on Antimicrobial Resistance (AGAR) Australian Gram-negative Sepsis Outcome Programme (GNSOP) Annual Report 2018. Communicable Diseases Intelligence, 44. https://doi.org/10.33321/cdi.2020.44.79

Bell, J. M., Lubian, A. F., Partridge, S. R., Gottlieb, T., Iredell, J., Daley, D. A., & Coombs, G. W. (2022). Australian Group on Antimicrobial Resistance (AGAR) Australian Gram-negative Sepsis Outcome Programme (GnSOP) Annual Report 2020. Communicable Diseases Intelligence, 45. https://doi.org/10.33321/cdi.2022.46.11

Brady, C., Cleenwerck, I., Venter, S., Coutinho, T., & De Vos, P. (2013). Taxonomic evaluation of the genus *Enterobacter* based on multilocus sequence analysis (MLSA): Proposal to reclassify *E. nimipressuralis* and *E. amnigenus* into *Lelliottia gen. nov.* as *Lelliottia nimipressuralis comb. nov.* and *Lelliottia amnigena comb. nov.*, respectively, *E. gergoviae* and *E. pyrinus* into *Pluralibacter gen. nov.* as *Pluralibacter gergoviae comb. nov.* and *Pluralibacter pyrinus comb. nov.*, respectively, *E. cowanii*, *E. radicincitans*, *E. oryzae* and *E. arachidis* into *Kosakonia gen. nov.* as *Kosakonia cowanii comb. nov.*, *Kosakonia radicincitans comb. nov.*, *Kosakonia oryzae comb. nov.* and *Kosakonia arachidis comb. nov.*, respectively, and *E. turicensis*, *E. helveticus* and *E. pulveris* into *Cronobacter* as *Cronobacter zurichensis nom. nov.*, *Cronobacter helveticus comb. nov.* and *Cronobacter pulveris comb. nov.*, respectively, and emended description of the genera *Enterobacter* and *Cronobacter*. Systematic and Applied Microbiology, 36(5), 309–319. https://doi.org/10.1016/j.syapm.2013.03.005

Briatte, F. (2023). ggnetwork: Geometries to Plot Networks with "ggplot2" (R package version 0.5.12) [Computer software]. https://github.com/briatte/ggnetwork

Cabezas, M. P., Fonseca, N. A., & Muñoz-Mérida, A. (2023). MIMt – A curated 16S rRNA reference database with less redundancy and higher accuracy at species-level identification. https://doi.org/10.1101/2023.12.15.571851

Candela, A., Guerrero-López, A., Mateos, M., Gómez-Asenjo, A., Arroyo, M. J., Hernandez-García, M., Del Campo, R., Cercenado, E., Cuénod, A., Méndez, G., Mancera, L., Caballero, J. D. D., Martínez-García, L., Gijón, D., Morosini, M. I., Ruiz-Garbajosa, P., Egli, A., Cantón, R., Muñoz, P., ... Rodríguez-Sánchez, B. (2023). Automatic Discrimination of Species within the *Enterobacter cloacae* Complex Using Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry and Supervised Algorithms. Journal of Clinical Microbiology, 61(4), e01049-22. https://doi.org/10.1128/jcm.01049-22

CDC, U.S Department of Health and Human services. (2013). ANTIBIOTIC RESISTANCE THREATS in the United States, 2013. https://www.cdc.gov/drugresistance/pdf/ar-threats-2013-508.pdf

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. Nucleic Acids Research, 44(D1), D67–D72. https://doi.org/10.1093/nar/gkv1276

Davin-Regli, A., Lavigne, J.-P., & Pagès, J.-M. (2019). Enterobacter spp.: Update on Taxonomy, Clinical Aspects, and Emerging Antimicrobial Resistance. Clinical Microbiology Reviews, 32(4), e00002-19. https://doi.org/10.1128/CMR.00002-19

Doijad, S. P., Gisch, N., Frantz, R., Kumbhar, B. V., Falgenhauer, J., Imirzalioglu, C., Falgenhauer, L., Mischnik, A., Rupp, J., Behnke, M., Buhl, M., Eisenbeis, S., Gastmeier, P., Gölz, H., Häcker, G. A., Käding, N., Kern, W. V., Kola, A., Kramme, E., ... Chakraborty, T. (2023). Resolving colistin resistance and heteroresistance in *Enterobacter* species. Nature Communications, 14(1), 140. https://doi.org/10.1038/s41467-022-35717-0

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Scientific Reports, 11(1), 12728. https://doi.org/10.1038/s41598-021-91456-0

Forde, B. M., Bergh, H., Cuddihy, T., Hajkowicz, K., Hurst, T., Playford, E. G., Henderson, B. C., Runnegar, N., Clark, J., Jennison, A. V., Moss, S., Hume, A., Leroux, H., Beatson, S. A., Paterson, D. L., & Harris, P. N. A. (2023). Clinical Implementation of Routine Whole-genome Sequencing for Hospital Infection Control of Multi-drug Resistant Pathogens. Clinical Infectious Diseases, 76(3), e1277–e1284. https://doi.org/10.1093/cid/ciac726

GARDP. (2022, October 17). Germany and Other Funders Pledge Support to GARDP to Ramp up Efforts in Countering Antibiotic Resistance. https://gardp.org/news-resources/germany-and-other-funders-pledge-support-to-gardp-to-ramp-up-efforts-in-countering-antibiotic-resistance/

Gotham, D., Moja, L., van der Heijden, M., Paulin, S., Smith, I., & Beyer, P. (2021). Reimbursement models to tackle market failures for antimicrobials: Approaches taken in France, Germany, Sweden, the United Kingdom, and the United States. Health Policy, 125(3), 296–306. https://doi.org/10.1016/j.healthpol.2020.11.015

Gupta, S., Mortensen, M. S., Schjørring, S., Trivedi, U., Vestergaard, G., Stokholm, J., Bisgaard, H., Krogfelt, K. A., & Sørensen, S. J. (2019). Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. Communications Biology, 2(1), 291. https://doi.org/10.1038/s42003-019-0540-1

Harris, S. R. (2018). SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology [Preprint]. Genomics. https://doi.org/10.1101/453142

Hoffmann, H., Stindl, S., Ludwig, W., Stumpf, A., Mehlen, A., Heesemann, J., Monget, D., Schleifer, K. H., & Roggenkamp, A. (2005). Reassignment of *Enterobacter dissolvens* to *Enterobacter cloacae* as *E. cloacae* subspecies *dissolvens comb. Nov.* And emended description of *Enterobacter asburiae* and *Enterobacter kobei.* Systematic and Applied Microbiology, 28(3), 196–205. https://doi.org/10.1016/j.syapm.2004.12.010

Hoffmann, H., Stindl, S., Ludwig, W., Stumpf, A., Mehlen, A., Monget, D., Pierard, D., Ziesing, S., Heesemann, J., Roggenkamp, A., & Schleifer, K. H. (2005). *Enterobacter hormaechei subsp. Oharae subsp. Nov.*, *E. hormaechei subsp. Hormaechei comb. Nov.*, and *E. hormaechei subsp. Steigerwaltii subsp. Nov.*, Three New Subspecies of Clinical Importance. Journal of Clinical Microbiology, 43(7), 3297–3303. https://doi.org/10.1128/JCM.43.7.3297-3303.2005

Horesh, G. (n.d.). Twilight [R]. https://github.com/ghoresh11/twilight?tab=readme-ov-file

Howe, K., Bateman, A., & Durbin, R. (2002). QuickTree: Building huge Neighbour-Joining trees of protein sequences. Bioinformatics, 18(11), 1546–1547. https://doi.org/10.1093/bioinformatics/18.11.1546

Ji, Y., Wang, P., Xu, T., Zhou, Y., Chen, R., Zhu, H., & Zhou, K. (2021). Development of a One-Step Multiplex PCR Assay for Differential Detection of Four species (*Enterobacter cloacae*, *Enterobacter hormaechei*, *Enterobacter roggenkampii*, and *Enterobacter kobei*) Belonging to *Enterobacter cloacae* Complex With Clinical Significance. Frontiers in Cellular and Infection Microbiology, 11, 677089. https://doi.org/10.3389/fcimb.2021.677089

Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Research, 3, 124. https://doi.org/10.12688/wellcomeopenres.14826.1

Jolley, K. A., & Maiden, M. C. (2014). Using MLST to study bacterial variation: Prospects in the genomic era. Future Microbiology, 9(5), 623–630. https://doi.org/10.2217/fmb.14.24

Katz, L. S., Griswold, T., Morrison, S. S., Caravas, J. A., Zhang, S., den Bakker, H. C., Deng, X., & Carleton, H. A. (2019). Mashtree: A rapid comparison of whole genome sequence files. Journal of Open Source Software, 4(44). https://doi.org/10.21105/joss.01762

Keck, F., Couton, M., & Altermatt, F. (2023). Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. Molecular Ecology Resources, 23(4), 742–755. https://doi.org/10.1111/1755-0998.13746

Kim, M., Oh, H.-S., Park, S.-C., & Chun, J. (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal of Systematic and Evolutionary Microbiology, 64(Pt_2), 346–351. https://doi.org/10.1099/ijs.0.059774-0

Kristy Horan, Anders Goncalves da Silva, & Andrew Perry. (2023). MDU-PHL/abritamr: DB updater (v1.0.15) [Computer software]. [object Object]. https://doi.org/10.5281/ZENODO.10369242

Lam, M. M. C., Wick, R. R., Watts, S. C., Cerdeira, L. T., Wyres, K. L., & Holt, K. E. (2021). A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. Nature Communications, 12(1), 4188. https://doi.org/10.1038/s41467-021-24448-3

Larsen, M. V., Cosentino, S., Lukjancenko, O., Saputra, D., Rasmussen, S., Hasman, H., Sicheritz-Pontén, T., Aarestrup, F. M., Ussery, D. W., & Lund, O. (2014). Benchmarking of Methods for Genomic Taxonomy. Journal of Clinical Microbiology, 52(5), 1529–1539. https://doi.org/10.1128/JCM.02981-13

Lasch, Peter, Stämmler, Maren, & Schneider, Andy. (2023). Version 4 (20230306) of the MALDI-ToF Mass Spectrometry Database for Identification and Classification of Highly Pathogenic Microorganisms from the Robert Koch-Institute (RKI) (Version 4 (20230306)) [dataset]. Zenodo. https://doi.org/10.5281/ZENODO.7702375

Miyoshi-Akiyama, T., Hayakawa, K., Ohmagari, N., Shimojima, M., & Kirikae, T. (2013). Multilocus Sequence Typing (MLST) for Characterization of *Enterobacter cloacae*. PLoS ONE, 8(6), e66358. https://doi.org/10.1371/journal.pone.0066358

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., ... Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. The Lancet, 399(10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biology, 17(1), 132. https://doi.org/10.1186/s13059-016-0997-x

Oren, A., Arahal, D. R., Göker, M., Moore, E. R. B., Rossello-Mora, R., & Sutcliffe, I. C. (2023). International Code of Nomenclature of Prokaryotes. Prokaryotic Code

(2022 Revision). International Journal of Systematic and Evolutionary Microbiology, 73(5a). https://doi.org/10.1099/ijsem.0.005585

Paauw, A., Caspers, M. P. M., Schuren, F. H. J., Leverstein-van Hall, M. A., Delétoile, A., Montijn, R. C., Verhoef, J., & Fluit, A. C. (2008). Genomic Diversity within the *Enterobacter cloacae* Complex. PLoS ONE, 3(8), e3018. https://doi.org/10.1371/journal.pone.0003018

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research, 50(D1), D785–D794. https://doi.org/10.1093/nar/gkab776

Pavlovic, M., Konrad, R., Iwobi, A. N., Sing, A., Busch, U., & Huber, I. (2012). A dual approach employing MALDI-TOF MS and real-time PCR for fast species identification within the *Enterobacter cloacae* complex. FEMS Microbiology Letters, 328(1), 46–53. https://doi.org/10.1111/j.1574-6968.2011.02479.x

Pérez-Losada, M., Cabezas, P., Castro-Nallar, E., & Crandall, K. A. (2013). Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infection, Genetics and Evolution, 16, 38–53. https://doi.org/10.1016/j.meegid.2013.01.009

Prinzi, A. M., & Moore, N. M. (2023). Change of Plans: Overview of Bacterial Taxonomy, Recent Changes of Medical Importance, and Potential Areas of Impact. Open Forum Infectious Diseases, 10(7), ofad269. https://doi.org/10.1093/ofid/ofad269

PubMLST (1.42.0). (n.d.). [dataset]. Organisms/*Enterobacter cloacae*/ *Enterobacter cloacae* typing. https://pubmlst.org/bigsdb?db=pubmlst_ecloacae_seqdef&page=profiles&scheme_id=1

R Foundation for Statistical Computing. (n.d.). R: A language and environment for statistical computing. [Computer software]. https://www.R-projcet.org/

Ramirez D, Giron M. (2022). *Enterobacter* Infections. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK559296/

Rodríguez-Sánchez, B., Cercenado, E., Coste, A. T., & Greub, G. (2019). Review of the impact of MALDI-TOF MS in public health and hospital hygiene, 2018. Eurosurveillance, 24(4). https://doi.org/10.2807/1560-7917.ES.2019.24.4.1800193

Rogers, Kara. (2022). *Enterobacter*. In Encyclopedia Britannica. https://www.britannica.com/science/Enterobacter

Seemann, T. (n.d.). mlst Tool [Computer software]. https://github.com/tseemann/mlst

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE, 11(10), e0163962. https://doi.org/10.1371/journal.pone.0163962

Singhal, N., Kumar, M., Kanaujia, P. K., & Virdi, J. S. (2015). MALDI-TOF mass spectrometry: An emerging technology for microbial identification and diagnosis. Frontiers in Microbiology, 6. https://doi.org/10.3389/fmicb.2015.00791

Surveillance & Response for Carbapenemase-Producing Enterobacterales (CPE) in NSW Health Facilities. (2019). NSW Department of Health. https://www1.health.nsw.gov.au/pds/ActivePDSDocuments/GL2019_012.pdf

Sutton, G. G., Brinkac, L. M., Clarke, T. H., & Fouts, D. E. (2018). *Enterobacter hormaechei subsp. Hoffmannii subsp. Nov.*, *Enterobacter hormaechei subsp. Xiangfangensis comb. Nov.*, *Enterobacter roggenkampii sp. Nov.*, and *Enterobacter muelleri* is a later heterotypic synonym of *Enterobacter asburiae* based on computational analysis of sequenced *Enterobacter* genomes. F1000Research, 7, 521. https://doi.org/10.12688/f1000research.14566.2

Tao, S., Chen, H., Li, N., Wang, T., & Liang, W. (2022). The Spread of Antibiotic Resistance Genes In Vivo Model. Canadian Journal of Infectious Diseases and Medical Microbiology, 2022, 1–11. https://doi.org/10.1155/2022/3348695

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., ... Zhang, J. (2023). UniProt: The Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pan-genomes

with Panaroo pipeline. Genome Biology, 21. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02090-4#citeas

Van Rossum, G., & Drake, J. F. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam. https://docs.python.org/3/reference/index.html

Varghese, N. J., Mukherjee, S., Ivanova, N., Konstantinidis, K. T., Mavrommatis, K., Kyrpides, N. C., & Pati, A. (2015). Microbial species delineation using whole genome sequences. Nucleic Acids Research, 43(14), 6761–6771. https://doi.org/10.1093/nar/gkv657

Ventola, C. L. (2015). The antibiotic resistance crisis: Part 1: causes and threats. P & T: A Peer-Reviewed Journal for Formulary Management, 40(4), 277–283.

Wickham, H. (2016). Ggplot2. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biology, 20(1), 257. https://doi.org/10.1186/s13059-019-1891-0

World Health Organisation. (2021, November 17). Antimicrobial resistance. Why Is Antimicrobial Resistance a Concern? https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistanceworls

World Health Organization. (2022). Global Antimicrobial Resistance and Use of Surveillance System (GLASS) Report 2022. WHO. https://www.who.int/publications/i/item/9789240062702

Wozniak, T. M., Dyda, A., Merlo, G., & Hall, L. (2022). Disease burden, associated mortality and economic impact of antimicrobial resistant infections in Australia. The Lancet Regional Health - Western Pacific, 27, 100521. https://doi.org/10.1016/j.lanwpc.2022.100521

Wu, W., Feng, Y., & Zong, Z. (2020). Precise Species Identification for *Enterobacter*: A Genome Sequence-Based Study with Reporting of Two Novel Species, *Enterobacter quasiroggenkampii sp. nov.* and *Enterobacter quasimori sp. nov.* mSystems, 5(4), e00527-20. https://doi.org/10.1128/mSystems.00527-20

Wyres, K. L., Lam, M. M. C., & Holt, K. E. (2020). Population genomics of *Klebsiella pneumoniae*. Nature Reviews Microbiology, 18(6), 344–359. https://doi.org/10.1038/s41579-019-0315-1