VARIATION OF SURFACE POLYSACCHARIDES IN THE ST25 CLONAL LINEAGE OF *ACINETOBACTER BAUMANNII*

Sarah Maree Cahill

Bachelor of Biomedical Sciences

Submitted in fulfilment of the requirements for the degree of
Master of Philosophy.

Institute of Health and Biomedical Innovation
Faculty of Health
Queensland University of Technology
2021

Keywords

Acinetobacter baumannii, sequence type 25, capsular polysaccharide, lipooligosaccharide, K locus, OC locus

Abstract

Multiply antibiotic resistant infections caused by the globally prominent Gram-negative bacteria *Acinetobacter baumannii* is currently a major public health concern due to the expansion of extensively antibiotic resistant clonal lineages. An increasingly prominent clone, sequence type (ST) 25, possesses various resistance and virulence mechanisms that are of concern. The global expansion of ST25 warrants effective epidemiology tracking, and so investigation into viable epidemiological markers is crucial. It is known that some of the most variable regions of *A. baumannii* genomes belonging to the same clone is the loci responsible for biosynthesis of the capsule polysaccharide and the outer-core component of the lipooligosaccharide, known as the K and OC locus respectively. The variability of these two genomic regions makes them attractive epidemiological markers, and the viability of these regions as markers has been validated previously for major global clones, such as GC1.

Currently, little is known about the variation at the K and OC locus in ST25 and the variability of these genomic loci were explored using *in-silico* tools including *Kaptive*. A newly developed bioinformatics tool capable of identifying surface polysaccharide biosynthesis gene clusters. Genome sequences of ST25 isolates were obtained from public databases, resulting in a collection of 82 geographically distributed ST25 isolates. These genomes were subjected to *Kaptive* analysis using private databases of 140 known KL types and 15 OCL types. Newly identified types were characterised and annotated by further manual examination utilising BLASTp and ACT comparison tools. Quality control procedures lead to the exclusion of 21 genomes from the original 82 isolate pool, leaving 61 assemblies with typeable K and OC loci. KL typing revealed 19 distinct KL types in the 61 ST25 isolates, which included 3 novel locus types (KL132, KL133, KL134) and one variant type (KL32a). OCL typing revealed 5 distinct OCL types within the 61 ST25 isolates, which includes one novel type (OCL15) and one variant

(OCL10a). Comparison of these regions in ST25 to the KL and OCL in other prominent clones (GC1, GC2, ST10) contextualised the extent of genomic variability in the ST25 lineage. Furthermore, the CPS structure produced by 13 of the 19 KL types were gathered from published literature and assessed to ensure genes localised at the KL region were solely responsible for the produced CPS structure. This included the assignment of encoded proteins in the KL32 gene cluster to their biosynthesis roles in the production of the CPS from an *A. baumannii* isolate, LUH 5549. Comparison between the 13 known CPS structure in ST25 revealed a range of different gene modules coding for complex sugars and varied glycosyltransferase produced linkages, indicating a diverse range of cell surface structures in the ST25 lineage. It was concluded that the KL and OCL variation in the ST25 clonal lineage is extensive, with geographically distinct combinations of KL/OCL types evident, highlighting the potential of enhancing current epidemiological tracking methods in *A. baumannii* by using these genomic features.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AA	Amino acid
CRAB	Carbapenem-resistant *Acinetobacter baumannii*
CPS	Capsular polysaccharide
GC	Global clone
GI	Genomic island
ENA	European nucleotide archive
IS	Insertion sequence
KL	Capsule biosynthesis locus (K locus)
LOS	Lipooligosaccharide
LPS	Lipopolysaccharide
MDR	Multidrug resistant
MLST	Multi-locus sequence typing
NCBI	National centre for biotechnology information
NR	Non-redundant
OCL	Outer-core biosynthesis locus (OC locus)
OECD	Organization for Economic Co-operation and Development
ORF	Open reading frame
PCR	Polymerase chain reaction
PFGE	Pulse field gel electrophoresis
SRA	Sequence read archive
ST	Sequence type
WGS	Whole genome sequence

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: QUT Verified Signature

Date: 22/01/2021

Acknowledgements

I would like to thank my supervisory team for their unwavering support over the course of my MPhil project. I would like to especially thank my principle supervisor, Dr Johanna Kenyon for believing in me, allowing me to grow and develop as a scientist and for giving me incredible opportunities over the past year and a half. I would like to thank A/Prof Makrina Totsika for devoting her time and energy into helping see this project through to completion and providing me with much needed support when it seemed this thesis would never end. Many thanks to Prof. Ruth Hall for her mentorship and for sharing her extensive knowledge with me. I greatly look forward to developing as a scientist and continuing my post-graduate journey under the guidance of all three. A special thanks to QUT for awarding me an Australian Government Research Training Program Scholarship during my degree.

I would like to thank all my friends and colleagues at QIMR for their constant encouragement, friendship, and willingness to grab a drink with me most Friday afternoons. I would especially like to thank Dr Emily Bryan and Dr Jacob Tickner for their wisdom and readiness to read through drafts and help me at their own expense.

My deepest gratitude goes to my partner, Jordan Raetz for his constant support and for pushing me to pursue my passions. Thank you for listening to my outburst of frustration and nightly science ramblings

Chapter 1: Introduction

1.1 LITERATURE REVIEW

1.1.1 *Acinetobacter baumannii:* An ESKAPE Pathogen

The ESKAPE pathogens encompass a small, critical group of increasingly prevalent pathogenic bacteria which have developed extensive antimicrobial resistance (AMR) (Rice, 2008). These pathogens, *Enterococcus faecium, Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species, have a huge impact on the healthcare system in both developed and developing countries and account for the majority of nosocomial infections (De Oliveira et al., 2020; Rice, 2010). In 2017, the World Health Organisation (WHO) established a new list of pathogens for which antimicrobial development is urgently required. This list featured the ESKAPE pathogens as the foremost targets for the development of antimicrobials (WHO, 2017). Since then, health authorities such as the Centres for Disease Control and Prevention (CDC) have released antibiotic resistance threat reports heavily featuring these ESKAPE pathogens, with the intention of spreading information regarding current antibiotic resistance trends, hence guiding antimicrobial research efforts (CDC) 2019). The most recent report iteration in 2019 saw carbapenem-resistant *Acinetobacter* as the foremost urgent threat in the U.S, responsible for 8,500 hospitalised cases and 700 deaths in 2017 (CDC, 2019).

Being an ESKAPE pathogen*, Acinetobacter baumannii* (*A. baumannii*) is considered one of the most important Gram-negative bacterial pathogens, accounting for an estimated 2-10% of all Gram-negative hospital-acquired infections worldwide in 2005 (Joly-Guillou, 2005). Concerningly, *A. baumannii* infections are associated with high morbidity and mortality

rates globally, with recent mortality rates as high as ~56% in multi-drug resistant (MDR) *A. baumannii* infections in multiple countries including China and Mexico (Cornejo-Juárez et al., 2020; Zhou et al., 2019). The bacterium is a strictly aerobic, non-motile coccobacillus, commonly found in aquatic environments (Howard et al., 2012). The species was formally designated in 1986 as part of the *Acinetobacter calcoaceticus-baumannii (ACB)* complex (Bouvet & Grimont, 1986). The complex consists of a group of aerobic, non-fermentative coccobacillus Acinetobacteria that are implicated in human infections and are high priority pathogens within intensive care units: *A. baumannii, A. calcoaceticus, Acinetobacter pittii* and *Acinetobacter nosocomalialis* (Blossom & Srinivasan, 2008).

The rise of *A. baumannii* as one of the most significant opportunistic pathogens can be traced back to the 1970s, where advancements in modern medicine, including endotracheal intubation, indwelling catheters and invasive surgical interventions offered a new route for host colonisation (Bergogne-Bérézin & Towner, 1996). *A. baumannii* is now implicated in a variety of site-specific infections within the clinical environment, such as ventilator-associated pneumonia, central-line bacteraemia, and soft tissue infections, particularly in patients in intensive care units and burns patients (McConnell et al., 2013; Roca et al., 2012; Spellberg & Bonomo, 2014). *A. baumannii* infections were also present in traumatic soft tissue injuries common during war-time conflict, primarily in Iraq and Afghanistan (Dallo & Weitao, 2010; O'Shea, 2012). In these cases, it was originally hypothesised that contamination with *A. baumannii* occurred at the time of injury, as this species is commonly found in the environment (Dallo & Weitao, 2010). However, subsequent studies indicate colonisation and infection of patients instead occurs most commonly at the point of treatment within many military medical facilities (Scott et al., 2007).

As with other problematic nosocomial bacteria, the overuse of broad-spectrum antibiotics remains a key factor in the prevalence of multi-drug resistant strains in healthcare

settings around the world. Although initially considered to have limited virulence potential, *A. baumannii* is now the aetiological agent in many healthcare associated infections that are difficult to treat due to AMR. The recent inclusion of *A. baumannii* as a priority pathogen in 2017 is primarily due to an increase in isolates resistant to carbapenems (drugs of last resort), which have been implicated in hospital outbreaks globally (Carretto et al., 2011; Schultz et al., 2016; Stietz et al., 2013; Villegas et al., 2007).

1.1.2 Growing Antimicrobial Resistance in *Acinetobacter baumannii*

A feature of *A. baumannii* that separates it from other species included in the ESKAPE classification, is the propensity for isolates to rapidly develop AMR. This predisposition is due in part to the frequency of horizontal gene transfer and homologous recombination amongst populations (Holt et al., 2016), aided by the naturally competent nature (the ability to uptake extracellular genetic material) of *A. baumannii* as well as the ability to form biofilms (Madsen et al., 2012; McQueary & Actis, 2011; Wilharm et al., 2013). In many instances, the mobile genetic elements dispensed throughout *A.baumannii* biofilms contain genes essential for the development of AMR, and being naturally competent permits the uptake of these genes.

Consequently, levels of MDR in *A. baumannii* are four times higher than those seen in other ESKAPE pathogens, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* (Xie et al., 2018). The rate of increased resistance for *A. baumannii* continues to climb rapidly, with MDR rates in Organization for Economic Co-operation and Development (OECD) and non-OECD countries reaching 56.9% and 80.4% respectively (Xie et al., 2018), though countries such as Greece and Turkey have recorded MDR prevalence levels as high as 90% and 96%, respectively (Xie et al., 2018). Considering the global increase of MDR isolates, investigations into potential novel drug targets for *A. baumannii* are at the forefront of research. However, studies are complicated by the wide range of both innate and acquired resistance mechanisms present in the nearly all *A. baumannii* strains, including drug inactivation, efflux pumps, outer-

membrane proteins, and decreased membrane permeability (Gordon & Wareham, 2010), which are responsible for resistance to our most utilised antibiotics including aminoglycosides, fluoroquinolones, polymyxins and tetracyclines (Ayoub Moubareck & Hammoudi Halat, 2020).

A major mechanism employed by *A. baumannii* to develop resistance is through the production of enzymes that modify and/or deactivate specific antibiotics particularly β-lactamases that deactivate the β-lactam class of antibiotics (Singh et al., 2013). Chromosomally encoded production of β-lactamases, specifically cephalosporinase is intrinsic to all *A. baumannii* strains (Perilli et al., 1996). Production of this enzyme, known as AmpC, confers resistance to both penicillin and cephalosporin groups of antibiotics, though it differs from AmpC enzymes found in other Gram-negative species (Hamidian & Hall, 2013). Resistance phenotypes can also arise from other resistance mechanisms, working tangentially to contribute to resistance to a single antibiotic class, an example being the mechanisms leading to carbapenem resistance in *A. baumannii* isolates (Lin & Lan, 2014).

Carbapenem class antibiotics (i.e. imipenem and meropenem) were once considered the most important drugs in treating *A. baumannii* infections. The rapid emergence of carbapenem-resistant *Acinetobacter baumannii* (CRAB) globally has imposed a substantial burden on healthcare systems, complicating treatment and increasing mortality rates. Research has revealed multiple mechanisms that contribute to carbapenem resistance, such as the loss and/or alteration of the outer-membrane protein CarO, preventing the uptake of carbapenems into the cell (Y. Lee et al., 2011), as well as modification to efflux pumps, namely the AdeABC resistance nodulation division (RND) efflux pump, which actively removes carbapenems from the cellular periplasm (Coyne et al., 2011). Though these mechanisms contribute to a resistance phenotype, they alone are unable to cause clinically relevant AMR (Hamidian & Nigro, 2019).

The most consequential mechanism for carbapenem resistance is the production of oxacillinase-class enzymes (OXAs), such as OXA-23, OXA-24, OXA-58, OXA-143, OXA-235, mobilised on plasmids and acquired via horizontal gene transfer. Oxacillinase enzymes show diverse substrate preference, and are able to efficiently inactivate not only oxacillins, but also carbapenems and cephalosporins, through hydrolysis of the β-lactam ring (Antunes & Fisher, 2014; Evans & Amyes, 2014). Though the genes encoding for OXA are most often acquired, oxacillinases are also intrinsically encoded by the chromosomal gene *oxaAb,* which has been established as a speciation marker for *A. baumannii* (Nigro & Hall, 2018). Due to the propensity for *A. baumannii* to share and take up genetic content, genomic resistance islands containing various *oxa* genes have been noted in isolates in numerous countries globally (Gao et al., 2017; Hamidian & Hall, 2011; Leungtongkam et al., 2018; Schultz et al., 2016; Villegas et al., 2007).

Though this resistance to one of our last line drugs, carbapenems is of major concern, the polymyxin antibiotic, colistin, remained a viable option for treatment of CRAB isolates, either alone or in combination with other key agents such as tigecycline (Gao et al., 2017; Qureshi et al., 2015). In keeping with the rapidly evolving nature of *A. baumannii*, evidence of colistin resistant strains has begun to emerge in hospitals, particularly within the healthcare systems of Asia (Cai et al., 2012; Gao et al., 2017; Qureshi et al., 2015; Thi Khanh Nhu et al., 2016). The variety of AMR mechanisms in *A. baumannii*, as well as the increasing prevalence of AMR strains in hospitals has established an urgent need for research into relevant therapeutic strategies. Similarly, monitoring global expansion and distribution of strains is crucial for outbreak management and isolate surveillance. Global surveillance strategies are essential for understanding how the species, and in particular, these resistant isolates, circulate both locally and globally.

1.1.3 Clonal Lineages and Global Distribution of *Acinetobacter baumannii*

Emergence of multiply resistant, epidemic strains of *A. baumannii* has highlighted the need for effective global disease monitoring. This can be enhanced by the identification of clonal lineages, as clonal groups maintain important characteristics that affect antimicrobial sensitivity. Early investigation into outbreak and non-outbreak strains in Europe identified three predominant clonal lineages, originally named 'European clones' I, II and III (Dijkshoorn et al., 1996). These clonal groups have been shown to contribute to most of the epidemic spread throughout other parts of the world, and were subsequently renamed 'Global Clones' I, II and III (GC1, GC2 and GC3). GC1 and GC2 (Fig. 1) account for the majority of extensively antibiotic resistant isolates, and as such, these clonal groups are a primary focus of research. Strains within these prominent clonal lineages in *A. baumannii* are notably phylogenetically robust (Fig. 1), and early research assumed clonal isolates were mostly identical with limited genomic variation (Dijkshoorn et al., 1996). However, it is now known that clonal isolates are highly variable. For example, prior to the 1980s, GC1 and GC2 were shown to confer resistance to early antibiotic classes including tetracycline, sulphonamides and aminoglycosides (Hamidian et al., 2019; Holt et al., 2016). However, following the introduction of new antibiotics such as fluoroquinolones and carbapenems, several recombination and plasmid uptake events have occurred leading to the acquisition of genes conferring resistance to these drugs, indicating variable sublineages within these clonal groups (Holt et al., 2016; Wright et al., 2014).
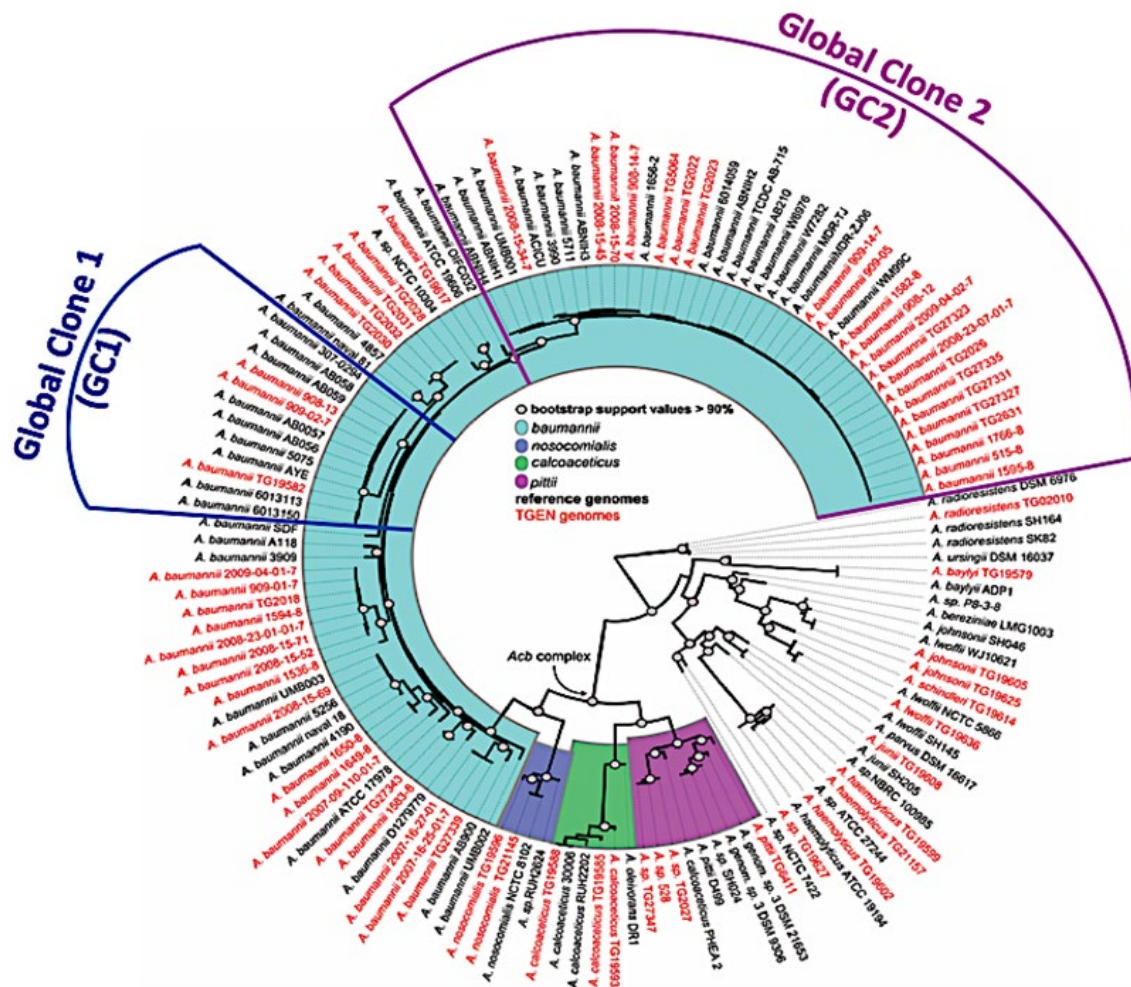
Figure 1. Phylogeny of *Acinetobacter spp.* genomes demonstrating the *A. baumannii* isolates (teal coloured) within the ACB complex. The major global clonal lineages (GC1, GC2) of *A. baumannii* are also indicated. *(Modified from: Sahl, Gillece et al. 2013).*

Though research into these major global clones is prominent, little is known about other rapidly emerging and extensively antibiotic resistant genotypes, namely sequence type (ST) 25 defined by the Insititut Pasteur multi-locus sequence typing (MLST) scheme. Research investigating the virulence traits between epidemic strains of *A. baumannii* revealed that isolates belonging to ST25 demonstrated higher resistance to desiccation, increased adherence on surfaces due to efficient biofilm formation, and greater adherence to alveolar epithelial cells compared to other lineages (Giannouli et al., 2013). These traits have likely led to the success of ST25 isolates in the clinical environment, aiding in the spread within hospital environments and facilitating host colonisation.

In addition to increased pathogenic potential, ST25 isolates have been noted to carry integrative elements, namely the insertion sequence IS*Aba*1 and genomic resistance islands (GRI) containing a complex class 1 integron. These elements are responsible for conferring resistance to multiple antibiotic classes, including third generation cephalosporins and carbapenems (Hamidian & Hall, 2016; M Hamidian et al., 2015). The presence of these antibiotic resistance elements in ST25 isolates has been noted to greatly increase mortality rates (da Silva et al., 2018). To date, ST25 has been implicated in multiple outbreaks across Europe (Carretto et al., 2011; Di Popolo et al., 2011; Gogou et al., 2011; Karah et al., 2011) and South America (Stietz et al., 2013), with sporadic isolates noted in Vietnam (Schultz et al., 2016) and Australia (Kenyon, Hall, et al., 2015; Kenyon et al., 2019). To better understand the true extent of global distribution and spread of ST25, epidemiological tracking is imperative.

1.1.4 Current Epidemiological Tracking Methods in *Acinetobacter baumannii*

For many Gram-negative species, such as *Escherichia coli* and *K. pneumoniae,* serotyping involving structurally variable polysaccharides on the cell surface has traditionally been the preferred method for tracing epidemiological patterns (Brisse et al., 2004; Corbett & Roberts, 2008; Stenutz et al., 2006; Wyres et al., 2016). Similarly, delineation of *A. baumannii* outbreaks was historically performed using serology specific to a highly variable immunogenic polysaccharide on the cell surface (Traub, 1989). A total of 38 different serovars had been identified in the latest update to the scheme (Traub, 1999). In the early literature, this immunogenic target was described as the O-antigen component of the lipopolysaccharide (LPS); a major component of the outer membrane for many Gram-negative species. However, in the last decade, recent work has shown that *A. baumannii* lacks an O-antigen, and alternatively produces only the base structure of the LPS, which is referred to as lipooligosaccharide (LOS) (Kenyon & Hall, 2013; Kenyon, Nigro, et al., 2014). Therefore,

therapeutic technologies and serotyping schemes utilising the O-antigen, as used for *E. coli*, is not a viable option for *A. baumannii*. Instead, it was shown that the major immunogenic polysaccharide present is the capsular polysaccharide (CPS or K-antigen), which structurally resembles O-antigen but is not associated or linked to the LOS.

Advancements in genetic typing methods presented new options for epidemiological typing of *A. baumannii* clinical isolates, which move away from traditional laboratory methods such as serotyping. As such, the introduction of sequencing technologies and innovations in computing capabilities has opened the door to rapid detection of serotypes using *in-silico* methods. An MLST scheme for *A. baumannii,* known as the Institut Pasteur scheme, was developed in 2005 with the intention of providing a highly discriminatory method of characterisation to aid in global surveillance (Bartual et al., 2005). The method was devised based on allelic variations in seven major house-keeping genes (*gltA, gyrB, gdhB, recA, cpn60, gpi and rpoD*), which could be amplified by polymerase chain reaction (PCR) and then sequenced to detect allelic difference (Bartual et al., 2005). GC1, GC2 and GC3 correspond to ST1, ST2 and ST3 in this scheme. While MLST is widely accepted for epidemiological tracking of *A. baumannii*, pulsed-field gel electrophoresis (PFGE) is considered the gold standard for genotyping as it can yield more discriminatory results. However, PFGE has a major drawback due to the time-frame required to generate data (>5 days) and expensive software required for analysis (Ahmed & Alp, 2015). This is indicative of *in-silico* analysis (such as MLST) as a preferred method of sequence typing due to the advantages over other laboratory-based methods.

As access to whole genome sequencing (WGS) becomes a more accessible in laboratories worldwide, and increasing advancements in computing power enable reduced analysis time, WGS and computational genomics are beginning to replace traditional laboratory methods including serotyping, PCR and PFGE for a number of bacterial pathogens

(Joensen et al., 2015; Thrane et al., 2016; Wyres et al., 2016; Yachison et al., 2017). Recent genomic studies have shown that the most variable region in genomes belonging to *A. baumannii* GC1 isolates is the locus that includes genes for the synthesis of the CPS (Holt et al., 2016; Kenyon & Hall, 2013), which is known as the K locus or KL (Fig. 2). This suggests that the CPS used as an epidemiological marker in traditional serological studies is also ideal for *in-silico* genotyping. Another highly variable region in *A. baumannii* GC1 genomes is the OC locus (OCL) responsible for the synthesis of the outer-core (OC) component of LOS; which is an important carbohydrate structure on the cell surface. A similar trend has been observed in genomes belonging to GC2 isolates, indicating the potential use of CPS and LOS as epidemiological markers in other *A. baumannii* strains (Adams et al., 2019; Wright et al., 2014).
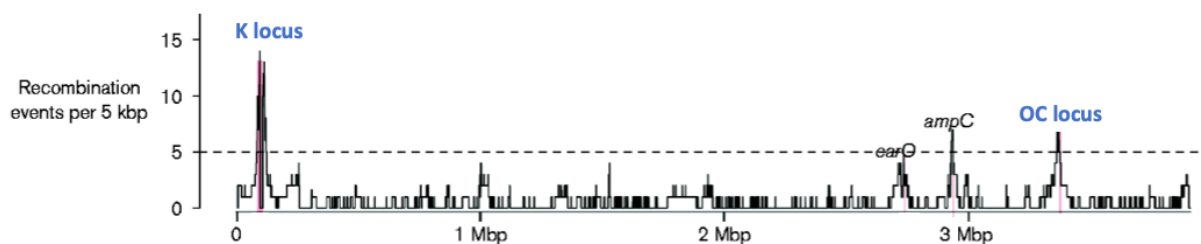
Figure 2. Recombination density plot of *A. baumannii* GC1 genomes. The points of highest recombination are observed to occur at the genomic K and OC loci *(Holt, Kenyon et al. 2016)*.

1.1.5 The Capsule Polysaccharide: A Major Virulence Factor and Therapeutic Target in *Acinetobacter baumannii*

The CPS is composed of long chains of repeating sugar units (oligosaccharides), known as K-units, that include between 2-6 monosaccharides often linked linearly (Fig. 3). The CPS is a primary virulence determinant in a broad range of both Gram-positive and Gram-negative bacterial species, the tightly packed polymers of K-units act as a barrier surrounding the cell, providing protection from environmental stressors, desiccation, phagocytosis and complement-mediated killing (Geisinger & Isberg, 2015; Russo et al., 2010; Tipton et al., 2018).
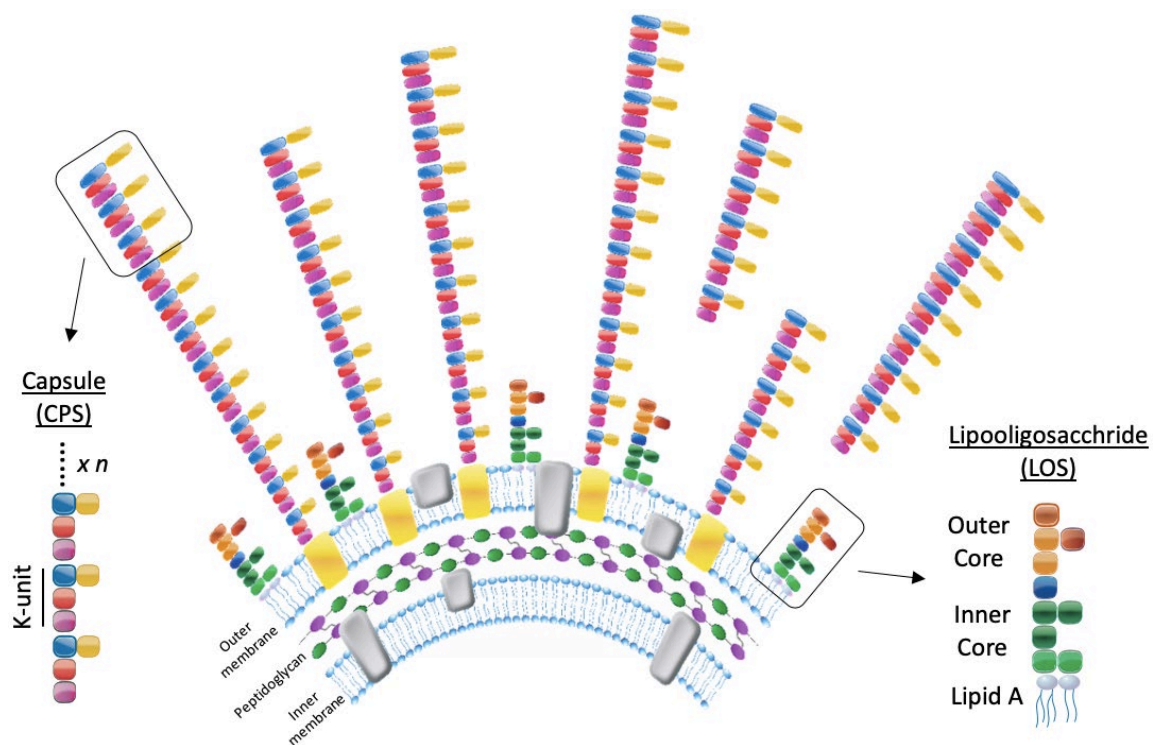
Figure 3. The surface polysaccharides of *Acinetobacter baumannii* extending from the outer membrane. Carbohydrates are illustrated as coloured boxes. The K-unit and outer-core are marked. *(Kenyon., Unpublished).*

The chemical structures of the CPS produced by more than 40 different *A. baumannii* isolates have been elucidated using NMR spectroscopy and sugar composition analyses. The variation observed across these 40 isolates includes differing sugar composition, unique linkages between K-units (N. Arbatsky et al., 2019; N. P. Arbatsky et al., 2019; Kenyon, Shashkov, et al., 2017; Shashkov et al., 2019), varying lengths of structures (Cahill et al., 2020; Kenyon, Kasimova, et al., 2017; Shashkov et al., 2018; Shashkov et al., 2016; Shashkov et al., 2019) and varied location of glycosidic bonds (Kasimova et al., 2017). *A. baumannii* CPS structures often include common UDP-linked sugars such as UDP-glucose and UDP-galactose. However, many also contain complex atypical sugars, such as non-2-ulosonic acids. So far, five distinct non-2-ulosonic acids have been found in the surface polysaccharides of *A. baumannii*, these include: legionaminic acid (N. Arbatsky et al., 2019; N. P. Arbatsky et al., 2019; Shashkov et al., 2015), 8-epilegionaminic acid (Vinogradov et al., 2014) pseudaminic acid (Kasimova et al., 2017; Senchenkova et al., 2015), acinetaminic acid (Kenyon, Kasimova,

et al., 2017; Kenyon, Marzaioli, De Castro, et al., 2015; Kenyon, Marzaioli, Hall, et al., 2015) and 8-epiacinetominic acid (Kenyon, Kasimova, et al., 2017; Kenyon, Notaro, et al., 2017), the latter acinetaminic acid derivatives have only been found in *A. baumannii* (Kenyon, Kasimova, et al., 2017; Kenyon, Marzaioli, De Castro, et al., 2015). In bacterial species, nonulosonic acids are often implicated in enhanced pathogenicity. Several studies have noted that the presence of these constituent glycoconjugates are essential virulence determinants in Gram-negative species, including *A. baumannii.* (Goon et al., 2003; Knirel et al., 2011; Perepelov et al., 2009; Schirm et al., 2003)

The biosynthesis and export of the CPS in *A. baumannii* is directed by clusters of genes at the K locus (KL) located between the *fkpA* and *lldP* genes in the chromosome (Fig 4). The arrangement of gene clusters found at this location have been shown to be highly variable (Kenyon & Hall, 2013) with more than 128 different KL gene clusters identified to date (Wyres et al., 2020). Though variable in genetic content, all gene clusters include two largely conserved regions (region 1 and region 3 in Fig. 4). Region 1 includes an operon of genes (*wzc-wzb-wza*) encoding enzymes for CPS export (on the left as drawn in Fig. 4). These three proteins form a multi-protein complex, establishing a channel between inner and outer membranes to transport the capsule structure from the periplasm to the cell surface (Kenyon & Hall, 2013). Region 3 is a module of genes responsible for simple sugar synthesis, namely common UDP-linked sugars, which is transcribed at the end region of the K locus (on the right as drawn in Fig. 4). The central region (region 2), located between the conserved flanking modules, includes genes coding for complex synthesis of specific CPS structures. This central region features genes encoding glycosyltransferases (Gtrs), acetyltransferases (Atrs), an initiating transferase (Itr), processing proteins (Wzy, Wzx), and often genes for complex sugars, such as a non-2-ulosonic acid (Kenyon & Hall, 2013).
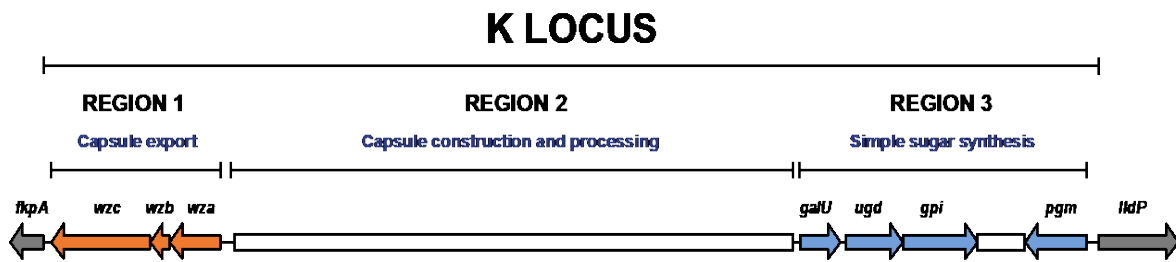
Figure 4. A simplified illustration of the general gene cluster arrangement of the K locus. Flanking genes are in grey. Genes common to all K types are shown (i.e. genes for capsule export), and the variable regions are depicted by white bars (Wyres et al., 2020).

As an immunogenic dominant surface structure, the polysaccharide capsule has been a major antigenic target in vaccine development for various encapsulated bacteria. Currently, polysaccharide vaccines exist for *Neisseria meningitidis* serogroups A, C, W and Y, providing high efficacy (>85%) and short-term immunity (Harrison et al., 2018; Rouphael et al., 2009). Polysaccharide vaccines against multiple *Streptococcus pneumoniae* serotypes also exists as both unconjugated and conjugated to toxins in order to elicit a protective immune response in at risk populations (Daniels et al., 2016). Studies in mice have explored the potential of passive immunisation against *A. baumannii* using a CPS-specific antibody, which succeeded in eliciting protection against various clinical isolates (Russo et al., 2013), though efficacy was enhanced by a conjugate vaccine involving CPS glycans and a protein carrier (Yang et al., 2017). However, vaccine development utilising CPS is complicated by the extensive structural diversity seen across the species (Wang-Lin et al., 2017).

More recently, the CPS of *A. baumannii* has been an important target in other novel therapeutic strategies. For example, Bacteriophage, which recognise specific CPS epitopes, have been explored as a treatment for infections caused by many MDR bacterial species. Bacteriophage (or simply phage) have many advantages over antibiotics, namely due to their narrow spectrum of activity, which in turn is thought to make phage therapy a better tolerated

and safer alternative to antibiotics (Principi et al., 2019). The rise of CRAB isolates globally has shifted focus towards the viability of phage therapy in treating *A. baumannii* infections. Current phage therapy strategies are reliant on phage with structural depolymerases, which can recognise and digest specific CPS structures (Gordillo Altamirano & Barr, 2019; Knirel et al., 2020). Without the CPS, *A. baumannii* becomes more susceptible to antibiotics, the host immune system and environmental factors (Russo et al., 2010; Wang-Lin et al., 2017). However, the specificity of phage to their target means that the exact structures of CPSs produced by the species must be known to ensure depolymerases encoded by phage are targeted and viable for the treatment of infections (Ghajavand et al., 2017).

1.1.6 The Outer-core Oligiosaccharide: A Variable Antigenic Component of the *A. baumannii* Lipid Membrane

For many Gram-negative bacteria, a second surface structure separate from the CPS exists, known as the lipopolysaccharides (LPS), which is anchored in the outer membrane of the cell wall. The LPS is a high-molecular weight structure consisting of a lipid-A portion and an extending carbohydrate structure consisting of a core oligosaccharide with distinct inner core and outer core (OC) segments, and an immunogenic O-antigen polysaccharide (Raetz & Whitfield, 2002). However, certain bacterial species including *Neisseria* spp. and *A. baumannii* express only the lipid A and core oligosaccharide portions forming a structure known as the lipooligosaccharide (LOS), which lacks the O-antigen component (Kenyon & Hall, 2013; Kenyon, Nigro, et al., 2014; Preston et al., 1996). For *A. baumannii*, the LOS is also an important virulence factor involved in resistance to antimicrobial peptides (Luke et al., 2010), cell motility (McQueary et al., 2012) and surface adhesion (Soon et al., 2011).

In *A. baumannii*, the outermost OC component of the LOS exhibits structural variation in the carbohydrate residues and linkages between sugars (Kenyon, Nigro, et al., 2014). This correlates to genetic variation seen at the OC locus (OCL) between the *aspS* and *ilvE* genes in the *A. baumannii* chromosome, which includes genes that direct the synthesis of the OC structure. A total of 14 OCL gene clusters have been reported in the species to date, with the lengths ranging between 8 and 12 kb depending on gene content (Kenyon, Holt, et al., 2014; Wyres et al., 2020). OCL gene clusters fall into two families (Group A and Group B) defined by a conserved region (Region 1 on the left in Fig. 5), which includes glycosyltransferases (named *gtrOC* to distinguish from *gtr* genes at the K locus) and a gene for a polysaccharide deacetylase (*pda*). The variable region (Region 2 in Fig. 5) in both Group A and Group B is located downstream of the Region 1 and may include genes for nucleotide-sugar biosynthesis, acetyltransferases, various glycosyltransferases, or genes of unknown function (Kenyon, Holt, et al., 2014).
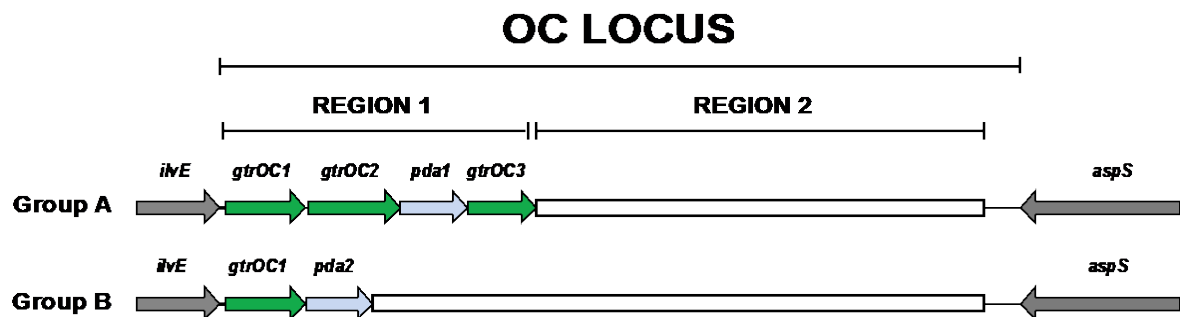
Figure 5. A simplified illustration of the general gene cluster arrangement for the OC locus. Flanking genes are shown in grey, and the two OC gene cluster families are indicated. Genes common to OC types are included, and the variable region is depicted by white bars (Wyres et al., 2020).

1.1.7 **The Viability of the K and OC locus as Epidemiological Markers in *A. baumannii***

Due to the extreme genetic variability seen at the K and OC loci, these genomic regions have proven valuable in the tracing and surveillance of *A. baumannii* outbreaks. Specifically, the combination of MLST with KL and OCL typing has been used successfully to analyse the

evolution of the major global clonal lineages, GC1 and GC2 (Hamidian et al., 2019; Hamidian & Nigro, 2019; Holt et al., 2016). Unfortunately, many of these studies have relied on manual KL and OCL typing as *in silico* typing tools have been lacking. However, recently, curated databases of reference sequences for both the KL and OCL types found in *A. baumannii* have been developed (Wyres et al., 2020) to be used in conjunction with a bioinformatics tool named '*Kaptive*' (Wick et al., 2018). These databases have the potential to rapidly screen thousands of *A. baumannii* genome sequences using *Kaptive*, providing a broader view on the extent of KL and OCL diversity in the species. By combining this screening with MLST data, the genetic diversity of KL and OCL found within the ST25 lineage can be determined, which can aid in assessing if the KL and/or OCL can be used as epidemiological markers for future ST25 outbreaks.

1.1.8 The Current Known Variability of the KL and OCL Regions in ST25

While studies are yet to describe the KL and OCL diversity in ST25, some ST25 isolates have been the subject of CPS structure analysis and KL gene cluster correlation and so the KL gene cluster for these isolates is known. The ST25 isolate, D4 has been the subject of many studies due to various genomic features, including genomic islands and plasmids which harbour a multitude of antibiotic resistance genes, which confer resistance to gentamicin, streptomycin, and to sulphonamides, amongst other antimicrobials (Hamidian & Hall, 2016; Mohammad Hamidian et al., 2015). Most recently the CPS produced by D4 has been elucidated and characterised, and the KL gene cluster which drives this production had been characterised and typed as KL16 in the same 2019 publication (Kenyon et al., 2019).

Another ST25 isolate, D46 was known to be an extensively antibiotic resistant clinical isolate, and has been sequenced to determine the causes of resistance (Nigro et al., 2015). It was also found to contain mobile genetic elements, namely a conjugative plasmid which

conferred resistance to both carbapenems and aminoglycosides (Nigro et al., 2015). Later, D46 was also the subject of CPS structural determination and KL gene cluster correlation in 2015 (Kenyon, Hall, et al., 2015). Like the D4 isolate, D46 and was found to harbour a novel KL gene cluster arrangement, which was designated KL14 (Kenyon, Hall, et al., 2015).

Interestingly, both D4 and D46 are sporadic ST25 isolates originating from Sydney, Australia in 2006 and 2010 respectively (Hamidian & Hall, 2016). Research regarding these isolates has characterised distinct resistance profiles, though both are considered MDR strains. Furthermore, the KL types of these isolates has been shown to differ significantly, with the KL16 gene cluster of D4 including a complex sugar synthesis cluster and *gtr* genes encoding for novel glycosyltransferases (Kenyon et al., 2019), and D46 including a less complex KL type, KL14 (Kenyon, Hall, et al., 2015). These findings set a precedent for the extent of genomic variation in the ST25 clonal lineages and allude to diversity of the KL and OCL regions of the genome which could aid in epidemiological tracking efforts.

1.2   PROJECT OUTLINE

1.2.1 Project Significance

This project aims to address the lack of current published research on the diversity of surface polysaccharides in the ST25 clonal lineage of *A. baumannii*. Currently, no studies have examined the full extent of KL and OCL variation in a complete collection of available ST25 genomes, or the consequent effect that these genetic differences have on CPS structures produced by ST25 isolates. More so, studies involving the ST25 lineage often work to build comparisons between isolates from different serotypes, rather than addressing the diversity of isolates belonging to a single sequence type (Giannouli et al., 2013; Zarrilli et al., 2011). Only

one study exists that characterises the genomic diversity in the ST25 lineage (Sahl et al., 2015). However, this study from 2015 did not address the genomic variation of the KL and OCL regions, though they are amongst the most variable regions in the *A. baumannii* genome (Holt et al., 2016) and the polysaccharide structures they encode for are crucial virulence determinants (Luke et al., 2010; McQueary et al., 2012; Russo et al., 2010).

By utilising current publicly available genomes belonging to ST25 isolates, a diverse pool of genome sequences obtained from various geographical locations can be examined, which will result in a more complete understanding of the global evolution of this lineage, as well as potential trends in local outbreaks. As studies have already confirmed the viability of the KL and OCL as epidemiological markers in other major clones (Holt et al., 2016), it is crucial that this is explored in other rapidly emerging lineages such as ST25 to assess global spread. This will aid surveillance and subsequent containment for infection control and will also guide future phylogenomic research on this important lineage. KL and OCL typing of a diverse set of genomes will likely also unearth novel locus types, novel genes, and potentially novel complex sugar products that may assist with therapeutics development. This can be further guided by correlating the genetic content at the K locus with the CPS structure produced to identify protein functions in the CPS biosynthesis pathway.

1.2.2 Hypothesis and Aims

Hypothesis

The surface polysaccharide variation in the ST25 clonal lineage of *A. baumannii* will be valuable as an epidemiological marker and in understanding the global spread and evolution of this clone.

Aim 1- Identify, characterise, and annotate the KL and OCL regions in all available genomes of *A. baumannii* ST25 isolates to assess the extent of sequence diversity in comparison to other globally important clonal lineages.

Aim 2- Correlate elucidated CPS structural information with KL sequences from *A. baumannii* ST25 isolates to aid in identification of protein function.

Chapter 2: Genotypic Variation of the KL and OCL Regions of ST25

This chapter outlines the standard and novel bioinformatics approaches used to identify and annotate the genetic content of KL and OCL regions in publicly available *A. baumannii* ST25 genome assemblies. Here, a comprehensive picture of the variation of these loci within the clone is presented, which is further compared to the level of KL and OCL variation within other globally distributed *A. baumannii* clonal lineages, including GC1 and GC2. A workflow diagram outlining the methodology and processed followed for Aim 1 is available in Appendix 1 of this document. Furthermore, the results presented in Section 2.2.9 have been recently published and are included in Wyres et al. 2020 (Wyres et al., 2020).

2.1   METHODOLOGY

2.1.1   Identification of *Acinetobacter baumannii* ST25 Genomes from Public Databases

The NCBI whole genome sequence and non-redundant databases (https://www.ncbi.nlm.nih.gov/genbank/wgs/) were used to bulk download all publicly available genome assemblies (3412 as of 21/02/2019) listed under the '*Acinetobacter baumannii*' organism classification. Independent verification of the *A. baumannii* species assignment of each genome assembly was achieved by searching for the *oxaAb* gene (GenBank accession number CP010781.1, base positions 1753305–1754129) specific to the *A. baumannii* species in the *Acinetobacter* genus using command-line BLASTn (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (Zhang et al., 2000). Genomes were confirmed as *A. baumannii* when BLASTn hits for *oxaAB* exceeded 95% for nucleotide sequence similarity,

and coverage was greater than 90% (Wyres et al., 2020). Confirmed *A. baumannii* genomes were retained for further analyses described in the following sections.

To identify genome assemblies belonging to ST25 isolates, multi-locus sequence typing (MLST) was performed using a command-line MLST script (https://github.com/tseemann/mlst) with the PubMLST Institut Pasteur *A. baumannii* typing scheme (abaumannii_2_scheme) (Jolley et al., 2018). All confirmed *A. baumannii* genome assemblies with the distinct ST25 allelic profile (*cpn60*-3, *fusA*-3, *gltA*-2, *pyrG*-4, *recA*-7, *rplB*-2, *rpoB*-4) were isolated from the bulk genome pool, and metadata (country, isolation date, sample source) for these isolates available in the original NCBI entries were recorded. Short read sequence data for a further ten ST25 *A. baumannii* isolates that did not have a genome assembly available were obtained from the Sequence Read Archive (SRA) and were *de novo* assembled into contigs using the short-read genome assembly program, *SPAdes v. 3.14* (Bankevich et al., 2012).

2.1.1 Identification and Characterisation of KL and OCL Regions in *A. baumannii* ST25 Genomes

The K locus (~20-40 kb in size) and OC locus (~8-15 kb in size) regions in all ST25 genome assemblies were identified and typed using the bioinformatics tool, *Kaptive v. 0.7* (https://github.com/katholt/*Kaptive*), with default parameters. Kaptive was developed by the Holt Lab at Monash University foremost as an epidemiological typing tool for both *Klebsiella pneumoniae* and *Acinetobacter baumannii*. Briefly, *Kaptive* runs each genome assembly as a query against two separate curated reference databases containing either 128 KL types or 14 OCL types known for *A. baumannii* to respectively identify the closest KL or OCL match using the BLAST algorithm (Wyres et al., 2020). The results output includes the closest match, confidence score (graded as 'Perfect', 'Very high' 'High', 'Good', 'Low', 'None'), length

discrepancies, coverage and identity percentage, genes identified within and outside the locus, and problems detected with the sequence match. The match confidence scores graded by *Kaptive* are assigned on the basis of multiple factors (Wick et al., 2018). 'Perfect' match score indicates that the locus was found on a single contig with 100% coverage and 100% identity. 'Very high' indicates that the locus was found on a single contig, with ≥99% coverage and ≥95% identity. 'High' is assigned when the locus is located on a single contig with ≥99% coverage but with ≤3 missing genes and no unexpected genes identified. 'Good' confidence matches indicate that the locus was found on a single contig or split across multiple assembly contigs with ≥ 95% coverage to the best-match locus, ≤3 missing genes and ≤1 unexpected gene within the locus. A 'Low' confidence match indicates that the locus was found on a single contig or split across multiple, with ≥90% coverage and ≤3 missing genes and ≤2 unexpected genes within the locus. Finally, a match confidence assignment of 'None' indicates that the match does not meet the criteria for any other confidence level. Further, Kaptive employs symbolism to denote specific problems identified in each query genome sequence when compared to the closest in database match, which reflects the match confidence score. These symbols as noted in the raw output in *Appendix 2*, include '**?**' which indicates that the match was not in a single piece, possibly due to either a poor match or discontiguous assembly. A '**-**' symbol denotes that genes found in the closest match locus were not found in the query locus. A '**+**' symbol indicates that genes not present in the closest match locus were found in the query locus, if these additional genes are within the Kaptive database, these genes are noted in the output with an identity score to the closest known K-locus. Finally, the inclusion of '*****' in the output denotes that one or more expected genes in the query locus have a low identity match to the best match locus, an identity percent below 100% will be flagged by Kaptive. The tool also extracts the sequence of the query locus in a FASTA file for further analysis.

Following *Kaptive* analysis, output results were manually checked and imperfect locus matches ('very high' confidence level or lower) were subjected to further investigation as described below.

2.1.2 Assessment of Genome Sequence Assembly Quality

As a general rule, *Kaptive* matches with confidence levels of 'good', 'low' or 'none' (<97% identity) can suggest genome assemblies of poor quality (Wyres et al., 2020). Often these locus sequences are found over multiple contigs making accurate locus typing difficult. Thus, the sequence quality of ST25 genome assemblies with *Kaptive* hits of 'good' or lower confidence levels were assessed to remove genomes of poor quality from further analyses. It is important to note that as the methods of this study concern specific loci on the genome, it is possible that some overall poor quality genome assemblies may have typable K and/or OC loci, or that overall good quality genomes may have assembly problems isolated to the K and/or OC loci. Therefore, genome quality testing was conducted after K and OC typing as a means of determining why *Kaptive* failed to type loci in specific isolate genomes.

Here, sequence quality was assessed using the genome assembly metric tool, QUAST (https://github.com/ablab/quast) (Gurevich et al., 2013). Exclusion of genome assemblies for analysis in this study was based on one or more of the following criteria: a) a contig count >300 for each individual assembly, and/or b) total genome size less than 3.7 Mbps. Genome sequences with KL or OCL matches present across more than 2 contigs (indicating multiple sequence breaks in the original genome assembly) that passed the quality criteria underwent manual examination.

2.1.3 Manual Examination of KL and OCL Regions in ST25 Genomes.

For ST25 genome assemblies that passed the sequence quality criteria for inclusion in the study, the KL and OCL matches called by *Kaptive* with 'Very good' or lower confidence levels, which were also present on a single unbroken contig, were tentatively considered new locus types. These regions were first compared against the best match locus type using the Artemis Comparison tool (ACT) included in the Artemis package software (Carver et al., 2012). Loci that were confirmed to include large sequence insertions or deletions, including open reading frame (ORF) replacements, were further examined using ORFfinder (https://www.ncbi.nlm.nih.gov/orffinder/). Through this tool, the translated sequence of each ORF within a locus sequence was subjected to BLASTp analysis against the NCBI non-redundant protein sequence database. Predicted proteins were assigned the same name as the best identified homologue when the pairwise sequence alignment was considered co-linear (>90% coverage) and the amino acid sequence identity exceeded 85% (Kenyon & Hall, 2013). Protein sequences with no identified match were further compared against an in-house database that includes published and unpublished KL and OCL protein sequences included in the current nomenclature system (*Kenyon, unpublished data*) using stand-alone BLAST. If a co-linear, high-identity match (>90% coverage, >85% amino acid sequence identity) was not detected for a sequence, the protein was considered novel and assigned a new name in accordance with the standardised nomenclature system (Kenyon & Hall, 2013; Kenyon, Nigro, et al., 2014; Wyres et al., 2020). Loci found to contain new ORFs between the defined flanking genes were considered novel types and further assigned a new KL or OCL number.

In instances where *Kaptive* could not confidently assign the KL and/or OCL types due to the matched sequence being found on more than one contig, the contigs containing the locus region were manually recovered from the original genome assembly. The transcription direction of

each contig sequence was first checked and corrected, if necessary, using the online reverse-complement tool (https://www.bioinformatics.org/sms/rev_comp.html). At each break site within the locus region, 100 bp of sequence from the end of each contig was submitted to ISfinder (Siguier et al., 2006) to determine if the insertion of an insertion sequence (IS) was the cause of the break. In the event that an IS segment was identified on either side of the break and was the only identifiable difference to the best match locus, the region was considered a variant of the original and was assigned a letter after the locus number i.e. KL1a.

For all other KL or OCL types, the contig ends were assessed to determine if the sequences could be directly abutted when compared to the best match locus. These contigs were joined with a series of 100 'N' bases to manually assemble the locus region in a single FASTA file, which was then manually assessed as detailed above to determine if the locus was novel.

2.1.4 Comparison of K and OC Locus Regions Found in ST1, ST2, ST10 and ST25 Isolates

To generate comparative data, genome assemblies of clonal groups of interest were recovered from the complete *A. baumannii* genome assembly pool (*n*= 3,378) obtained from methods outlined in Section 2.1.1. Clonal groups of interest were collected after the original *A. baumannii* genomes (n= 3,412) were bulk download and subjected to in-house speciation assignment via the BLASTn search for *oxaAb* gene. Only major globally distributed clonal lineages (GC1/ST1, GC2/ST2) and an emerging global clone (ST10) that includes a large proportion of multiply and extensively antibiotic resistant isolates were included in the comparative analysis. The ST1, ST2 and ST10 isolates underwent quality control analysis using an abridged QC protocol as seen in Section 2.1.2, where an efficient in-house command-line script similar to QUAST was used to inspect genomes against an exclusion criteria of a

contig count of >300 and/or an assembly length of < 3.6 Mbp (Wyres et al., 2020). An in-house script for genome assembly quality was chosen for this larger dataset due to efficiency of selecting for only these two exclusion metrics, as QUAST utilises many genome quality metrics, complicating the output for large datasets. Genome assemblies from isolates belonging to these three clonal lineages were subjected to *Kaptive* analysis to type KL and OCL regions as described above. Assemblies that had a *Kaptive* KL and OCL match with a confidence level of 'good' or above were compared with the previously typed ST25 assemblies by calculating KL and OCL frequencies visualised on a frequency heatmap generated in R using ggplot2 (https://ggplot2.tidyverse.org) (Wickham, 2016).

2.2   RESULTS

2.2.1 Acquisition of 3,422 *'Acinetobacter baumannii'* Genomes from Public Databases Includes 82 ST25 Isolates

A total of 3,412 genome assemblies listed under the *Acinetobacter baumannii* species classification were downloaded from the NCBI non-redundant and Whole Genome Sequence (WGS) databases. The *oxaAb* gene was absent from 34 of the 3,412 genome assemblies (0.99%), indicating a total of 3,378 confirmed *A. baumannii* genome assemblies. MLST analysis revealed that 71 genome assemblies belonged to the Institut Pasteur ST25 clone (*cpn60*-3, *fusA*-3, *gltA*-2, *pyrG*-4, *recA*-7, *rplB*-2, *rpoB*-4). An additional single locus variant (SLV), ST402, was also identified with an allelic profile differing from ST25 at *pyrG* (*cpn60-3, fusA-3, gltA-2, pyrG-1, recA-7, rplB-2, rpoB-4*). This SLV was found to be *A. baumannii* isolate LUH_7841 (Assembly accession number ASM100768v1) and was incorporated into further analyses due to its close genetic relatedness to ST25, and is hereafter referred to as an

ST25 isolate. The short reads for an additional 10 isolates previously reported to belong to ST25 (Schultz et al., 2016) were obtained from the European Nucleotide Archive (ENA) database (https://www.ebi.ac.uk/ena), assembled and included in the analysis. Overall, a total of 82 ST25 genome assemblies were obtained from these public databases.

To develop a broader understanding of the isolates included in this study, metadata on each isolate were collected from the bio-sample summary in their NCBI data entries. Metadata obtained included year of isolate collection, collection source and country of origin when available. The metadata for all 82 ST25 isolates, including Assembly or ENA accession numbers, are listed in Table 1.

Table 1. Data summary for all 82 ST25 isolates included in this study.

Strain	Assembly	Year	Country	Isolation Sample
741019	ASM100777v1 (NCBI)	2011	Argentina	Pleural fluid
D46	7468_2_59 (NCBI)	2010	Australia	N/A
D4	7521_8_34 (NCBI)	2006	Australia	N/A
MC104	ASM358425v1 (NCBI)	2016	Bolivia	Wound secretion
MC102	ASM358427v1 (NCBI)	2016	Bolivia	Ulcer
MC100	ASM358431v1 (NCBI)	2016	Bolivia	Tracheal secretion
MC89	ASM358442v1 (NCBI)	2016	Bolivia	Ulcer
MC29	ASM358448v1 (NCBI)	2016	Bolivia	Urine culture
MC93	ASM358451v1 (NCBI)	2016	Bolivia	Wound secretion
MC63	ASM359568v1 (NCBI)	2016	Bolivia	Fluid (Bone tissue)
MC51	ASM359581v1 (NCBI)	2016	Bolivia	Lumbosacral ulcer
MC39	ASM359586v1 (NCBI)	2016	Bolivia	Urine culture
MC32	ASM359593v1 (NCBI)	2016	Bolivia	Ulcer
MC57	ASM359612v1 (NCBI)	2016	Bolivia	Tracheal secretion
MC14	ASM359623v1 (NCBI)	2016	Bolivia	Blood
MC103	ASM358353v1 (NCBI)	2016	Bolivia	Pleural fluid
MC95	ASM358356v1 (NCBI)	2016	Bolivia	Tracheal secretion
MC94	ASM358357v1 (NCBI)	2016	Bolivia	Wound secretion
MC87	ASM358360v1 (NCBI)	2016	Bolivia	Tracheal secretion
MC98	ASM358362v1 (NCBI)	2016	Bolivia	Wound secretion
MC101	ASM358430v1 (NCBI)	2016	Bolivia	Blood culture
MC96	ASM358432v1 (NCBI)	2016	Bolivia	Tracheal secretion
MC91	ASM358436v1 (NCBI)	2016	Bolivia	Urine Culture
MC77	ASM358437v1 (NCBI)	2016	Bolivia	Tracheal secretion

MC78	ASM358438v1 (NCBI)	2016	Bolivia	N/A
MC90	ASM358444v1 (NCBI)	2016	Bolivia	Blood culture
MC71	ASM358445v1 (NCBI)	2016	Bolivia	Exudate (leg)
MC69	ASM359566v1 (NCBI)	2016	Bolivia	Wound secretion
MC64	ASM359567v1 (NCBI)	2016	Bolivia	Urine Culture
MC59	ASM359576v1 (NCBI)	2016	Bolivia	Catheter
MC34	ASM359592v1 (NCBI)	2016	Bolivia	Tracheal secretion
MC31	ASM359595v1 (NCBI)	2016	Bolivia	Catheter
MC48	ASM359613v1 (NCBI)	2016	Bolivia	Secretion
MC27	ASM359617v1 (NCBI)	2016	Bolivia	Pressure ulcer
MC18	ASM359620v1 (NCBI)	2016	Bolivia	Foley Catheter
107m	ABIBUN_1 (NCBI)	N/A	Colombia	N/A
NIPH146	Acin_baum_NIPH_146_V1 (NCBI)	1993	Czech Republic	Wound
161/07	ASM100781v1 (NCBI)	2007	Germany	Respiratory tract
BAuABod-3	ASM257380v1 (NCBI)	2015	Germany	Environment/Turkey
IHIT38008	ASM415359v1 (NCBI)	2018	Germany	Urine/Dog
4390	ASM100776v1 (NCBI)	2003	Greece	Bronchial aspirate
HEU3	ASM292785v1 (NCBI)	2016	Honduras	Cornea secretion
MCR10179	ASM292796v1 (NCBI)	2015	Honduras	Endobronchial tube
CI86	ASM51657v2 (NCBI)	2005	Iraq	Superficial Wound
CI79	ASM51663v2 (NCBI)	2005	Iraq	Respiratory Tract
4190	ASM18971v2 (NCBI)	2009	Italy	N/A
OCU_Ac2	ASM356944v1 (NCBI)	2014	Japan	Blood
7804	ASM343138v1 (NCBI)	2006	Mexico	Bronchoalveolar lavage
AB_2008-15-69	ASM30133v1 (NCBI)	N/A	N/A	N/A
PR388	ASM213810v1 (NCBI)	N/A	N/A	N/A
ARLG1317	ASM214397v1 (NCBI)	N/A	N/A	N/A
AR_0088	ASM300603v1 (NCBI)	N/A	N/A	N/A
R348	ASM395535v1 (NCBI)	N/A	N/A	N/A
984213	ASM58105v1	N/A	N/A	N/A
LUH_7841[1]	ASM100768v1 (NCBI)	2002	Netherlands	Venous Catheter
RUH1486	ASM100770v1(NCBI)	1985	Netherlands	Umbilicus
LUH_6220	ASM100774v1 (NCBI)	2000	Netherlands	Respiratory Tract
HWBA8	ASM208278v1 (NCBI)	2013	South Korea	Sputum
4300STDY7045700	24276_2_166 (NCBI)	2016	Thailand	N/A
4300STDY7045786	24276_2_251 (NCBI)	2016	Thailand	N/A
4300STDY7045802	24276_2_267 (NCBI)	2016	Thailand	N/A
4300STDY7045824	24276_2_289 (NCBI)	2016	Thailand	N/A
4300STDY7045840	24276_2_305 (NCBI)	2016	Thailand	N/A
4300STDY7045893	24276_3_50 (NCBI)	2016	Thailand	N/A
NM3	ASM100772v1 (NCBI)	2008	United Arab Emirates	Sputum
AB5256	ASM24170v2 (NCBI)	2009	USA	Blood
OIFC143	AcbauOIFC143v1.0 (NCBI)	2003	USA	N/A
1429530	ASM58113v1 (NCBI)	2014	USA	Perirectal
ABBL018	ASM143254v1 (NCBI)	2005	USA	Blood

AB2828	ASM161199v1 (NCBI)	2006	USA	Blood
AB3638	ASM161209v1 (NCBI)	2007	USA	Trauma Wound
AB3806	ASM161210v1 (NCBI)	2007	USA	Trauma Wound
UV_1036	SAMEA1569559 (ENA)	2003	Vietnam	Tracheal washes
UV_964	SAMEA1569507 (ENA)	2003	Vietnam	Tracheal washes
UV_973	SAMEA1569410 (ENA)	2003	Vietnam	Tracheal washes
238_an	SAMEA1569453 (ENA)	2005	Vietnam	Anal swab
259_an	SAMEA1569518 (ENA)	2005	Vietnam	Anal swab
266_an	SAMEA1569452 (ENA)	2005	Vietnam	Anal swab
276_ax	SAMEA1569532 (ENA)	2005	Vietnam	Axilla swab
295_c	SAMEA1569392 (ENA)	2005	Vietnam	Cannula entry
338_an	ERR263723 (ENA)	2006	Vietnam	Anal swab
338_ax	ERR263724 (ENA)	2006	Vietnam	Axilla swab

[1] ST402 isolate

Examination of the metadata for each isolate indicated that the genome collection was largely varied, and included samples collected as early as 1985 and as recently as 2018 (Table 1). Strains also came from a diverse geographic background, with isolate representation seen for most continents (Fig. 6). While some strains in the collection are outbreak associated (e.g. Thailand and Bolivia isolates), these represent less than half of the collection (32 of 82), and most strains were sporadic in nature. The majority of isolates were healthcare associated, and varied in their colonisation site (respiratory, wound and catheter associated infections). Two isolates (IHIT38008, BAuABod-3) were obtained from animal sources. The variety of strains gathered suggests that this collection is appropriate for further detailed epidemiological exploration, and that it is possible to draw a conclusion on the viability of the K and OC locus as an epidemiological marker in this clonal lineage.
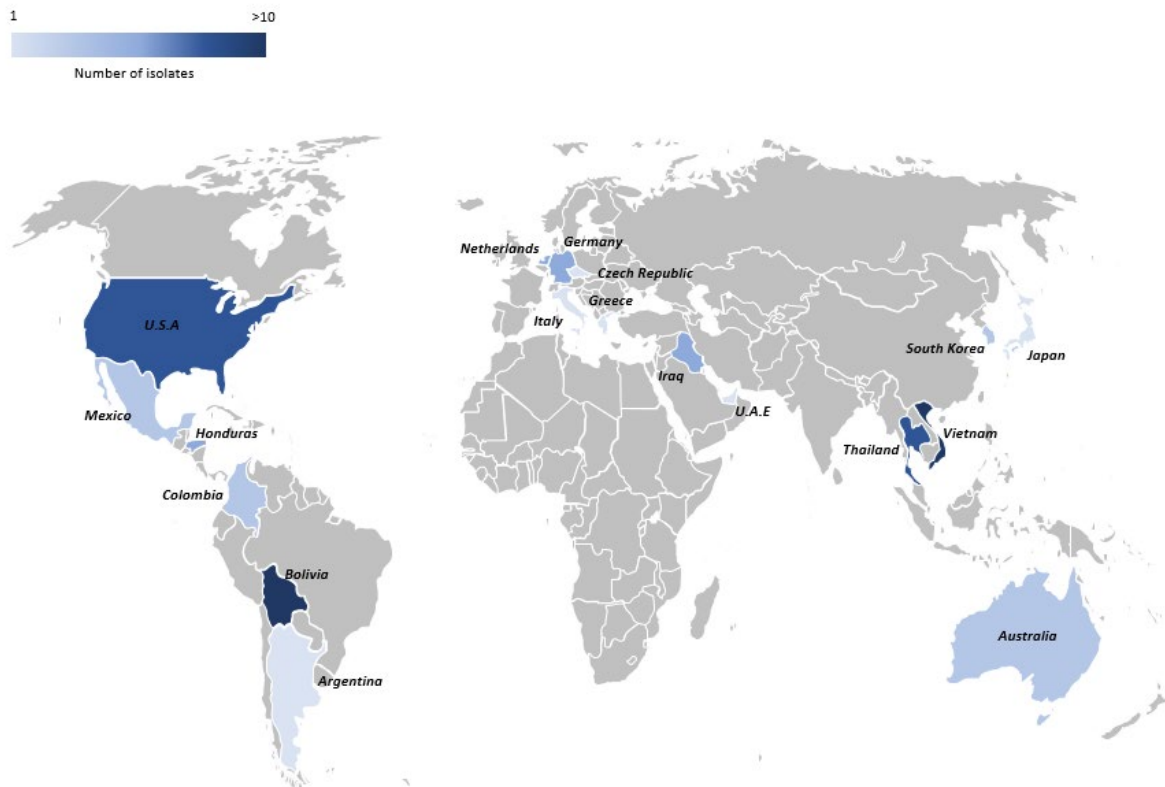
Figure 6. Geographical distribution of ST25 isolates identified in the bulk genome download (n=82). Countries are coloured by number of isolates present, with darkest blue indicating ten or more isolates (e.g. Bolivia) and the lightest blue indicating only one isolate originating from this country (e.g. Japan)

2.2.2 Kaptive and QUAST Analysis Indicates Possible Poor Genome Assembly Quality of 25 Genomes in the Pool

For an isolate to be included in this study, both the KL and OCL regions in the genome assembly must be confidently assigned. Previously, the quality of a genome assembly has been shown to be a critical factor for confident typing of the K and OC loci using *Kaptive* (Wyres et al., 2020). It was suggested that *Kaptive* assignments with a confidence level of 'Good' or lower may indicate a poor-quality sequence query. Thus, sequence quality was initially gauged via preliminary analysis of the KL and OCL regions within the 82 confirmed ST25 genome assemblies using *Kaptive v. 0.7*. Indeed, *Kaptive* analysis of the K locus revealed the following confidence levels: 22 (perfect), 20 (very high), 1 (high), 13 (good), 4 (low) and 22 (none), with

a total of 39 assemblies receiving a match confidence level of 'Good' or lower (see Appendix 2). Similarly, OCL analysis for the same set of genomes revealed confidence levels of: 3 (perfect), 39 (very high), 2 (high), 14 (good), 2 (low) and 21 (none), with 37 receiving a match confidence level of 'Good' or lower (see Appendix 3). This preliminary analysis showed that ~33% of genome assemblies could not be assigned a match with a confidence level of Perfect, Very high or High for both KL and OCL regions, indicating that a proportion of the query set of 82 ST25 genome assemblies may be of poor-quality sequence. However, poor genome quality does not necessarily prevent typing of K or OC loci. As this study is concerned with these genomic regions, poorer quality sequence elsewhere in the genome should not impact the typing outcome. Therefore, if genomes are assembled using sequence of poor quality it can still be possible to characterise the KL and OCL regions, and the KL and OCL were manually checked in each specific example examined.

Firstly, all 82 confirmed ST25 genome assemblies were subjected to QUAST analysis for sequence quality checking. A total of 19 assemblies (Table 2) failed to pass the quality inclusion criteria outlined in *Section 1.4*. Of these assemblies, 15 had received a *Kaptive* OCL assignment with a confidence match of 'none' and a KL assignment of either 'low' (2) or 'none' (13). Further manual inspection of the KL and OCL regions could determine the locus types. Thus, these 15 isolates were excluded from further analyses.

Three of the remaining four assemblies (Table 2) had *Kaptive* KL assignments of 'Good' (1), 'Low' (1) or 'None (1), and OCL assignments of 'Very high' (1) or 'Low' (2). However, typing of these loci was possible through manual inspection techniques outlined in *Section 1.4.* Therefore, these isolates were included in subsequent analysis. The final assembly, isolate 4190 (Assembly accession number ASM18971v2), had been created from short reads originally sequenced using 454 GS FLX sequencing technology. Poor quality sequence had been previously noted for this isolate (Shashkov et al., 2016), and the K locus had been re-

sequenced in this study (see GenBank accession number KT266827). Thus, the identity of these regions could be confirmed, and the isolate was included in this study.

Table 2. A summary of the 19 ST25 genome assemblies that did not meet sequence quality inclusion criteria. Match confidence scores and detected problems provided by *Kaptive* for K locus and OC locus analysis are included, as well as QUAST quality metrics for each isolate. Genome assemblies with confirmed locus regions that were subsequently recovered and included in the final pool of 61 isolates are shaded in grey

| Assembly | Strain | K Locus | | OC Locus | | |
		Match Confidence	Problem	Match Confidence	Problem	QUAST Metrics
ASM58105v1	984213	None	12/17 genes detected locus broken	None	6/9 genes detected locus broken	449 contigs, 4.1Mbps
ASM358353v1	MC103	None	12/20 genes detected locus broken	None	6/9 genes detected locus broken	489 contigs, 3.7Mbp
ASM358356v1	MC95	None	16/21 genes detected locus broken	None	6/9 genes detected locus broken	462 contigs, 4.0Mbps
ASM358357v1	MC94	None	8/20 genes detected locus broken	None	6/9 genes detected locus broken	474 contigs, 3.7Mbps
ASM358362v1	MC98	None	17/21 genes detected locus broken	None	6/9 genes detected locus broken	365 contigs, 4.0Mbps
ASM358430v1	MC101	None	11/20 genes detected locus broken	None	6/9 genes detected locus broken	527 contigs, 3.7Mbps
ASM358432v1	MC96	None	14/21 genes detected locus broken	None	6/9 genes detected locus broken	598 contigs, 3.6Mbps
ASM358436v1	MC91	None	12/20 genes detected locus broken	None	6/9 genes detected locus broken	555 contigs, 3.7Mbps
ASM358437v1	MC77	None	23/28 genes detected locus broken	None	6/9 genes detected locus broken	514 contigs, 4.0Mbps
ASM358444v1	MC90	None	14/21 genes detected locus broken	None	6/9 genes detected locus broken	332 contigs, 3.9Mbps
ASM359566v1	MC69	Low	18/20 genes detected locus broken	None	6/9 genes detected locus broken	311 contigs, 3.9Mbps
ASM359567v1	MC64	None	17/20 genes detected locus broken	None	6/9 genes detected locus broken	418 contigs 4.1Mbps
ASM359592v1	MC34	Low	19/20 genes detected locus broken	None	6/9 genes detected locus broken	355 contigs, 4.2Mbps

Assembly	Strain					
ASM359613v1	MC48	None	13/20 genes detected locus broken	None	6/9 genes detected locus broken	421 contigs, 3.9Mbps
ASM359617v1	MC27	None	23/28 genes detected locus broken	None	6/9 genes detected locus broken	334 contigs, 3.9Mbps
ASM257380v1	BAuABod-3	Good	15/17 genes detected locus broken	Very high	10/10 genes detected	329 contigs, 4.5Mbps
ASM359568v1	MC63	Low	17/20 genes detected locus broken	Low	7/9 genes detected locus broken	443 contigs, 4.0Mbps
ASM359593v1	MC32	None	16/21 genes detected locus broken 2 unexpected genes	Low	7/9 genes detected locus broken	400 contigs, 4.0Mbps
ASM18971v2	4190	None	13/24 genes detected locus broken	High	8/9 genes detected locus broken	396 contigs, 4.0Mbps

Of the remaining genome assemblies that passed the sequence quality criteria for inclusion in this study (see Table 3), six received a KL or OCL match with a confidence level of 'Good' or lower. However, manual inspection of these assemblies could not determine the KL or OCL type. Hence, these assemblies were also excluded from the final pool.

Table 3. A summary table of isolates which passed sequence quality criteria for inclusion by QUAST but were ultimately excluded from the final study pool due to poor sequence quality at the OC or K locus.

Assembly	Strain	K Locus		OC Locus		QUAST Metrics
		Match Confidence	Problems	Match Confidence	Problems	
ASM358360v1	MC87	None	16/21 genes detected locus broken	None	6/9 genes detected locus broken	291 contigs, 3.8Mbps
ASM358438v1	MC78	Good	19/20 genes detected locus broken	None	6/9 genes detected locus broken	285 contigs, 4.1Mbps
ASM358445v1	MC71	None	17/28 genes detected locus broken	None	7/9 genes detected locus broken	264 contigs, 3.8Mbps
ASM359576v1	MC59	Good	19/20 genes detected locus broken	None	7/9 genes detected locus broken	244 contigs 3.9Mbps
ASM359595v1	MC31	None	18/21 genes detected locus broken	None	6/9 genes detected locus broken	177 contigs, 4.0Mbps
ASM359620v1	MC18	Good	18/20 genes detected locus broken	None	6/9 genes detected locus broken	185 contigs, 3.9Mbps

Therefore, a total of 21 isolates were ultimately excluded from the study (Table 4), providing 61 genome assemblies for further analysis. Most of the excluded genomes (20 of 21) came from the same sequencing project using Illumina MiSeq technology and included Bolivian isolates from a 2016 outbreak. The other isolate was 98421 (Assembly accession number ASM58105v1), which was also sequenced using Illumina sequencing technology and uploaded to the NCBI public whole genome database in February 2014.

Table 4. A summary table of the final 21 isolates excluded from further analysis.

Assembly	Strain	K Locus			OC Locus		
		KL	Match Confidence	Problems	OCL	Match Confidence	Problems
ASM58105v1	984213	KL14	None	12/17 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358353v1	MC103	KL3	None	12/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358356v1	MC95	KL22	None	16/21 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358357v1	MC94	KL65	None	8/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358360v1	MC87	KL22	None	16/21 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358362v1	MC98	KL22	None	17/21 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358430v1	MC101	KL3	None	11/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358432v1	MC96	KL22	None	14/21 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358436v1	MC91	KL3	None	12/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358437v1	MC77	KL49	None	23/28 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358438v1	MC78	KL33	Good	19/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM358444v1	MC90	KL22	None	14/21 genes detected locus broken	OCL5	None	6/9 genes detected locus broken

ASM358445v1	MC71	KL49	None	17/28 genes detected locus broken	OCL5	None	7/9 genes detected locus broken
ASM359566v1	MC69	KL33	Low	18/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM359567v1	MC64	KL33	None	17/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM359576v1	MC59	KL33	Good	20/20 genes detected locus broken	OCL5	None	7/9 genes detected locus broken
ASM359592v1	MC34	KL33	Low	19/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM359595v1	MC31	KL22	None	18/21 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM359613v1	MC48	KL33	None	13/20 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM359617v1	MC27	KL49	None	23/28 genes detected locus broken	OCL5	None	6/9 genes detected locus broken
ASM359620v1	MC18	KL33	Good	18/20 genes detected locus broken	OCL5	None	5/9 genes detected locus broken

2.2.3 Kaptive Analysis Reveals Extensive Genomic Variation at the K locus

For the 61 ST25 genome assemblies included in the final study pool, the confidence levels called by *Kaptive* during K locus assignment were: 22 (perfect), 20 (very high), 1 (high), 10 (good), 2 (low) and 6 (none). The output summary of KL identification in the initial 82 isolates is presented in Appendix 2 of this document. The 22 genomes that Kaptive assigned a 'perfect' match confidence score had a locus match found on a single contig, with 100% sequence identity and 100% coverage to an existing K locus type. The 22 'perfect' genomes were assigned a K locus match to 4 unique locus types, KL37 (7), KL23 (1), KL14 (13) and KL16 (1). The K locus type KL14 was the most prevalent amongst 'perfect' isolates, KL37 was the second most common and shares a similar genetic structure to KL14.

For those genomes that received a 'Very high' confidence score (20), the issues detected by *Kaptive* were limited to one or more of the expected genes in the locus being

detected with low nucleotide sequence identity (<95%) or coverage (>99%) to the best match locus. Nine sequences with one or more genes of low nucleotide sequence identity to the assigned match were manually inspected. However, the individual gene(s) of interest had a translated amino acid sequence identity of >85% to the translated sequence of the best match, indicating that the genes were not novel (Kenyon & Hall, 2013). Thus, the locus type could be confirmed as the assigned best match locus for each.

A further isolate, 107m (Assembly accession ABIBUN_1), with a 'Very high' KL confidence score was flagged for a length discrepancy of an additional 889 bp to the best match locus type, KL32. As all other KL32 genes were present with high nucleotide sequence identity (>95%) on a single contig, and no additional known genes were detected within the locus, this sequence was determined to be a novel variant of KL32 and subjected to further analysis (see Section 2.2.4).

Only one isolate, AB5256, received a *Kaptive* assignment of KL14 with a confidence score of 'High'. An issue was detected with the *wzc* gene missing from the capsule export operon of the K locus. Manual examination of ORFs within the K locus of this assembly confirmed the presence of a *wzc* gene that had replaced the original *wzc* gene in sequence. Hence, this isolate was confirmed to carry the KL14 type.

A total of 10 isolates were assigned KL matches with a 'Good' confidence score, and 8 of these had the match found in more than one contig. By using manual methods outlined in *Section 1.4*, the contigs with K locus sequence could be directly abutted to form a single contig for these 8 isolates, allowing confirmation of the best match locus type assigned by *Kaptive* for each assembly. The remaining two isolates, LUH_7841 and MCR10179, were marked for other problems including the presence of additional genes, the absence of expected genes, and/or genes with low nucleotide sequence identity to the best match locus type *Kaptive* identified KL95 as the best match locus type for isolate LUH_7841 (Assembly accession

ASM100768v1), though an additional gene as identified in the locus with 90% sequence identity to *qdtA*. However, manual inspection of ORFs within the LUH_7841 locus using ORFfinder did not identify *qdtA*. ACT sequence alignment of the LUH_7841 K locus and the sequence for KL95 revealed 100% coverage and 100% identity between the two sequences, confirming that LUH_7841 K locus types was KL95. For isolate MCR10179 (Assembly accession ASM292796v1), *Kaptive* assigned the KL10 type but identified numerous problems with this match, including low nucleotide sequence identity to reference KL10 sequence, and both the absence of expected genes (20/23 identified) and the presence of an additional gene. Moreover, 510 bp of sequence was missing from the K locus. These factors warranted further investigation into the possibility of this isolate containing a novel K locus type, and the assembly of this isolate was subjected to further analysis (see Section 2.2.4).

The 2 isolates that were assigned matches with a 'Low' confidence score were also found to contain K locus sequence across multiple contigs in the assemblies. *Kaptive* further marked both isolates for problems including the absence of expected genes, which could be a consequence of discontinuous assembly at the K locus. Manual examination and abutment of the contigs within the locus was sufficient in correcting the problems for both isolates, and the best match locus types assigned by *Kaptive* were confirmed.

Finally, nine isolates were assigned a match with a confidence score of 'None'. All nine were manually assessed and found to contain either additional genes within the locus, missing expected genes, and/or expected genes with low nucleotide sequence identity to the best match. For two isolates, PR388 and HEU3, novel K types were suspected due to these factors, and were further examined in Section 2.2.4. A further isolate was identified as 4190, which had the KL re-sequenced as described above and confirmed as KL27.

2.2.4 KL Characterisation in the 61 ST25 Isolates Uncovered Three Novel KL Types and a Variant Type.

As described in the previous section, three isolates were suspected of carrying novel KL sequences and one isolate was predicted to carry a variant of a known gene cluster.

The suspected variant gene cluster was carried by isolate 107m (Assembly accession number ABIBUN_1). The gene cluster consisted of the same genetic content as KL32 but included an additional IS between *atr9* and *orf* (Fig. 7a). This IS was identified as IS*Ab*a27 using ISFinder (https://isfinder.biotoul.fr). The IS insertion does not disrupt any genes in KL32; therefore, it is not expected to affect the synthesis of the capsular polysaccharide in this isolate, and the variant was named KL32a.
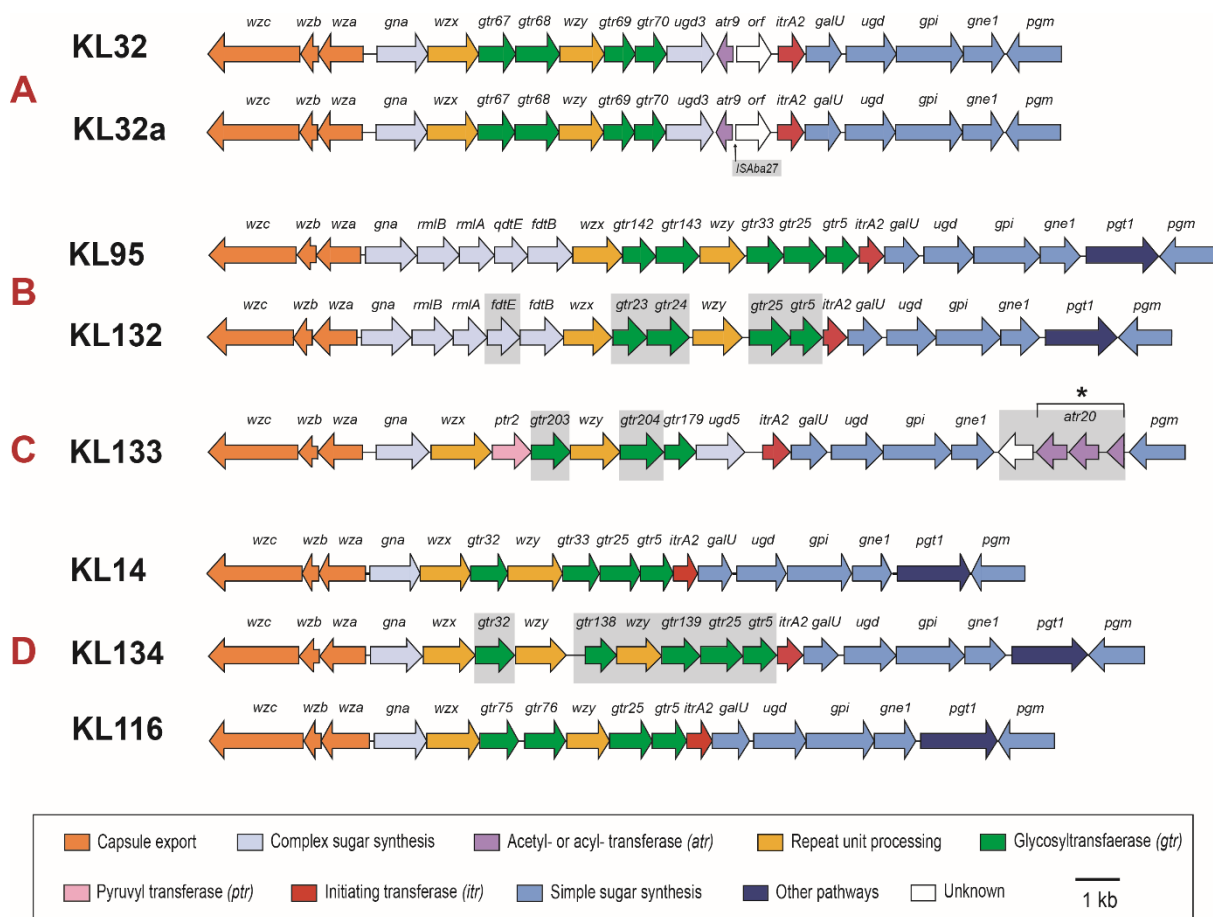
Figure 7. Novel and variant K locus types found in ST25 isolates, grouped (A, B, C, D) with the most related KL type. Grey shadow boxes highlight regions which define each KL as a novel or variant type. Arrows indicate direction of transcription. Asterix indicates a gene disrupted by breaks.

Isolate MCR10179 (Assembly accession number ASM292796v1) was found to carry a sequence closely related to the known KL95 locus sequence (Fig. 7b). However, the MCR10179 sequence contains the known *gtr23-gtr24-wzy* gene module that replaces *gtr142-gtr143-wzy-gtr33* in KL95. In addition, a *fdtE* gene replaces *qdtE* in the complex sugar synthesis pathway. This gene combination has not been seen previously, hence the sequence is considered novel and was designated KL132.

The K locus type in isolate PR388 (Accession assembly number ASM213810v1) had no close relationship to a known KL type, and as such was designated KL133 (Fig. 7c). The PR388 sequence includes two novel glycosyltransferase genes, assigned as *gtr203* and *gtr204*. Gtr203 was identified as belonging to glycosyltransferase family 2, clan GT-A (Pfam accession: CL0110) and had the closest match to Gtr152 (50% coverage) in the in-house database *(Kenyon, unpublished data)*. Gtr204 was identified as a belonging to glycosyltransferase family 1, clan GT-B (Pfam accession: CL0113) and had the closest match to Gtr172 (30% coverage). The PR388 sequence also includes a novel sequence segment in the module of genes for simple sugar synthesis, which includes an unknown gene encoding a hypothetical protein. Immediately adjacent to the unknown gene, is a broken *atr20* acetyltransferase gene. The presence of the gene in three pieces is due to this K locus being manually joined as described above.

The KL sequence from isolate HEU3 (Assembly accession number ASM292785v1) is most like previously identified KL types, KL116 and KL14 (Fig. 7d), which both appear in ST25 isolates in this collection. However, the sequence includes a *gtr138* glycosyltransferase gene, identified by standalone BLASTp using an in-house *gtr* database. The KL sequence also unusually includes two *wzy* genes. The first *wzy* gene encodes a protein that has 100% shared identity to the Wzy encoded by KL14 (GenPept accession number WP_000367709.1), while the second *wzy* encodes a protein that is 56% identical to a Wzy repeat-unit polymerase from

Acinetobacter pittii (GenPept accession number WP_130174168.1). Though the presence of two *wzy* genes is atypical, biochemical studies would be required to determine if one or both genes are functional. These features indicated that this sequence was also a novel KL type and thus the sequence was designated KL134.

2.2.5 Review of the Gene Clusters Found at the K Locus in ST25 Genomes Reveals Extensive Genetic Variation

Analysis performed on the 61 ST25 genome assemblies revealed a total of 19 different gene clusters at the K locus. The assignment of KL to each isolate in the collection can be found in Appendix 2 of this document. The genetic content of each KL type found amongst the ST25 collection is shown in Figure 8. The most prevalent KL type in the collection is KL14, present in 19 isolates, followed by KL37 present in 13 isolates. Notably, both KL14 and KL37 share much of the same genetic content. Both types lack gene modules that direct the synthesis of complex sugars, instead carrying the same simple sugar synthesis gene pathway: *galU, ugd, gpi, gne1, pgt1* and *pgm*.

Closely related KL12 and KL13 gene clusters were found in 4 and 2 genome assemblies, respectively. These two gene clusters are closely related, differing only by the *wzy* gene, and contain a module of genes that encode for the non-2-ulosonic acid, 5,7-di-N-acetyl-acinetaminic acid (Aci), only so far found in *A. baumannii* (Kenyon, Marzaioli, De Castro, et al., 2015). Three KL types containing the same complex sugar synthesis pathway, KL27, KL7 and KL130 were found in 1 isolate each. This complex sugar synthesis pathway is responsible for producing another non-2-ulosonic acid, 5,7-di-N-acetyl-legionaminic acid (Leg). A third gene module for another non-2-ulosonic acid, 5,7-di-N-acetyl-pseudaminic acid (Pse), was found in 5 isolates carrying the K locus types, KL33 (3), KL16 (1) and KL23 (1). The presence of Leg and Pse in bacterial sugar structures have been correlated with enhanced virulence and pathogenicity in a variety of multi-drug resistant bacterial species (Glaze et al., 2008; Y. J. Lee

et al., 2011; Schoenhofen et al., 2006; Schoenhofen et al., 2009; Zunk & Kiefel, 2014), though the exact role of these sugars and their derivatives are yet to be fully understood. It can be speculated that the Aci sugar may also have a similar role in enhancing virulence in *A. baumanii*, as it belongs to the same family of non-2-ulosonic acids.
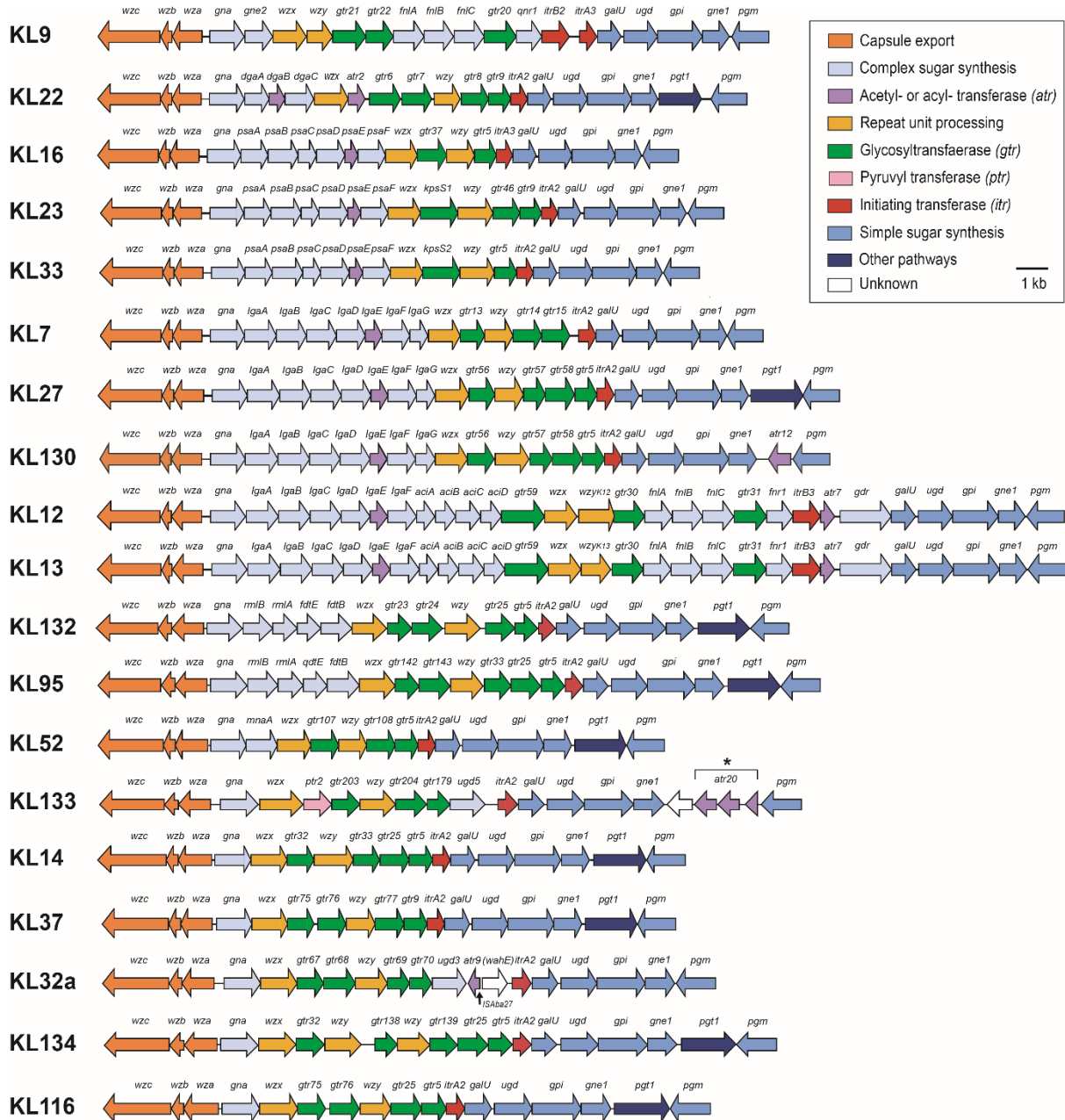
Figure 8. Total diversity at the K locus responsible for CPS biosynthesis in all 61 isolates representing the ST25 clonal lineage. Individual genes are shown as arrows, the direction of which indicates direction of transcription.

2.2.6 Kaptive Analysis of the OC Locus in ST25 Genomes Reveals a Range of OCL Types

The OCL assignments of the 61 ST25 genome assemblies had the following confidence levels called by *Kaptive*: 3 (perfect), 39 (very high), 3 (high), 14 (good) and 2 (low). The OCL identification output table generated by *Kaptive* is available as Appendix 3 in this document. The 42 isolates assigned a confidence score of 'Perfect' or 'Very high' were 100% coverage with >99% sequence identity to the best match locus assigned. The three isolates with a 'Perfect' identity score were all matched to OCL6, while the 39 isolates with a 'Very high' confidence matched were assigned to a variety of OCL types, including OCL5, OCL6, OCL7 and OCL10.

Three isolates assigned a match with a confidence score of 'High' included 741019 (Assembly accession number ASM100777v1) and MC57 (Assembly accession number ASM359612v1), which were assigned to OCL10 as the best match locus and included an identical length discrepancy of an additional 1048 bp. Given this information, these isolates were speculated to be variations of OCL10 due to a sequence insertion, and were further investigated for confirmation (see Section 2.2.7). The third isolate assigned a match confidence of 'High' was 4190 (Assembly accession number ASM18971v2). This isolate was best matched to OCL5 with 100% coverage and 99% identity but was flagged as missing a gene (8/9 genes), *atrOC1*. Manual examination using ORFfinder identified 9 ORFs in the sequence, and BLASTp was able to identify *atrOC1* (189aa) with 99% sequence identity to an *atrOC1* in the NR database (protein ID: AHK10232.1). BLASTp noted a single amino acid change of glutamine (Q) for lysine (K) in the sequence, which is likely the cause of less than perfect confidence match by Kaptive. In summary, 4190 was confirmed as carrying the OC type OCL5.

The 14 isolates assigned a match with a confidence score of 'Good' had various issues preventing a perfect match to the best match locus type. Of these, 13 were detected as having

the locus present across multiple contigs, with 10 of these missing expected genes. 8 of the 10 flagged as missing expected genes were missing only one gene in the cluster, which was accounted for following reverse-complementation of reverse strands and abutment of contigs. The 8 isolates were found to contain three OCL types, including OCL5, OCL6 and OCL10. The remaining 2 sequences (Assembly accession numbers ASM358425v1 and ASM358431v1) were best matched to OCL5 by Kaptive, and both had only 7 of 9 genes accounted for, with the OCL region separated across 5 contigs. Manual manipulation of these contigs using reverse-complementation and abutment was successful in accounting for the 2 missing genes in both sequences, and so both were confidently assigned the OCL type, OCL5. In summary, 13 of 14 isolates assigned a 'Good' match by Kaptive were able to be confidently assigned to the best match locus types assigned by Kaptive. However, one isolate in the 14, OCU_Ac2 (Assembly accession number ASM356944v1), was a notable outlier. *Kaptive* revealed that only 6 of 9 expected genes were present, with an additional gene present and genes present at low nucleotide sequence identity to the best match locus. Furthermore, this isolate had an OC locus length discrepancy of an additional 1743 bp. Considering this information, the isolate was subjected to further investigation to determine the presence of a novel locus type (see Section 2.2.7).

The two isolates (Assembly accession number ASM359568v1 and ASM359593v1) assigned a match with 'Low' confidence were both found by *Kaptive* to be discontinuous and were missing expected genes (7 of the 9 detected). Each sequence was spread across three contigs, and reverse complementation of reverse strands followed by abutment of contigs guided by ACT alignment was successful in typing these OCL as OCL5.

2.2.7 One Novel and One Variant OC locus Type were Discovered Amongst the ST25 Genomes

A total of three isolates were suspected of carrying novel OCL types, due to notable length discrepancies to the closest OCL match, and so these were manually examined. Using manual inspection methods outlined in Section 2.1.3, isolate OCU_Ac2 (Assembly accession number ASM356944v1) harboured an additional 1743 bp to the closest locus match and was found to carry a novel type (Fig. 9a). This novel type was designated OCL15 and belongs to the Group B OCL family. OCL15 appears to be a partial hybrid of OCL6 and OCL7, including *gtrOC1-rmlC* from OCL6 and *gtrOC24-gtrOC25* from OCL7. However, OCL15 includes an additional novel gene, *gtrOC23*, inserted between *gtrOC19a* and *rmlB*.
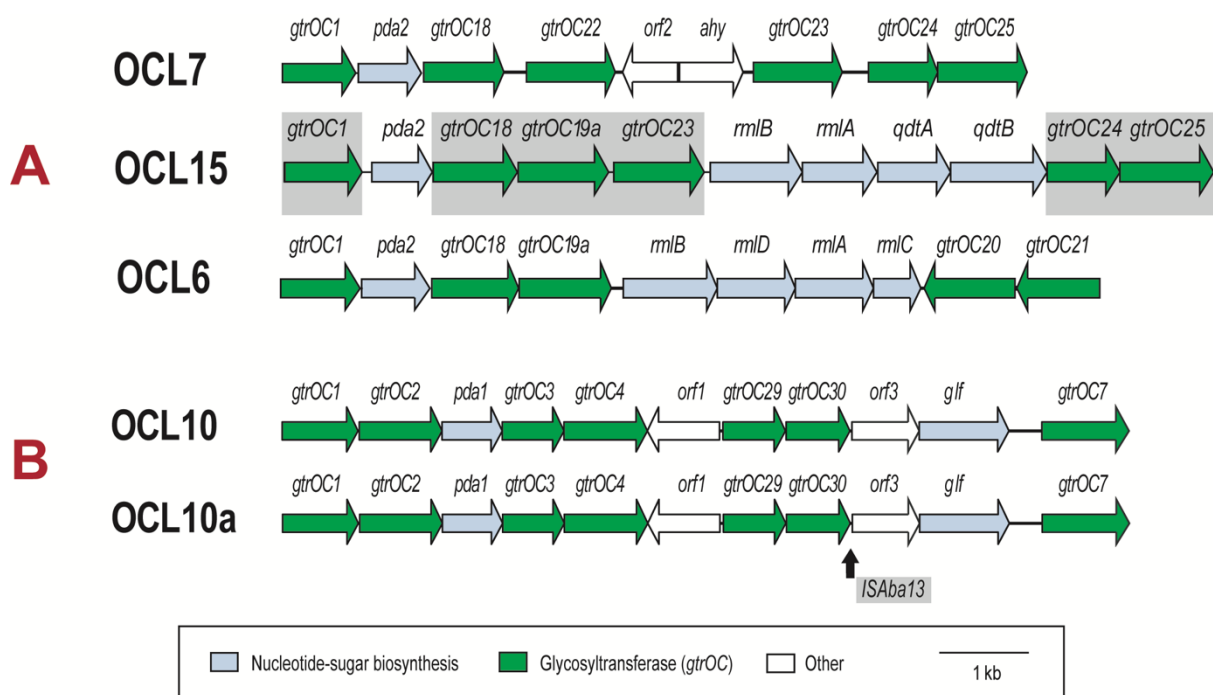
Figure 9. Novel and variant OC locus types found in ST25 isolates compared to their closest OCL reference. Grey shadow boxes highlight regions which define each OCL as a novel or variant type. Arrows indicate direction of transcription.

The same variant OCL type was found in isolates 741019 and MC57 (Fig. 9b). These two isolates were found to contain a length discrepancy of an additional 1048 bp compared to best locus match, OCL10. The additional sequence was located between *gtrOC30* and *orf3* and

was identified as the insertion sequence, IS*Aba*13. The insertion of IS*Aba*13 did not appear to disrupt either gene. While it is likely this intergenic insertion would not disrupt outer-core LOS production, as evidenced by previous studies, this would need to be confirmed by observing the produced structure from isolates 741019 and MC57.

2.2.8 Variation at the OC Locus in ST25 Genomes Accounts for 5 Distinct OCL Types

Analysis of the OC locus in the 61 ST25 genome assemblies revealed 5 distinct OCL types, with one being a novel OCL gene cluster, OCL15 (Fig. 10). A single variant of OCL10, named OCL10a, was also detected. It was shown that ST25 isolates include both OCL families (Group A and Group B) and contain a variety of genes that encode proteins for sugar synthesis, glycosyltransferase, acetyltransferase and of unknown function. The most common OCL type was OCL5, which was found in 57% of the isolates in this collection. The second most common OCL type was OCL6, which was in 25% of the isolates. The assignment of OCL to the 61 isolates in the collection can be found in Appendix 1 of this document.
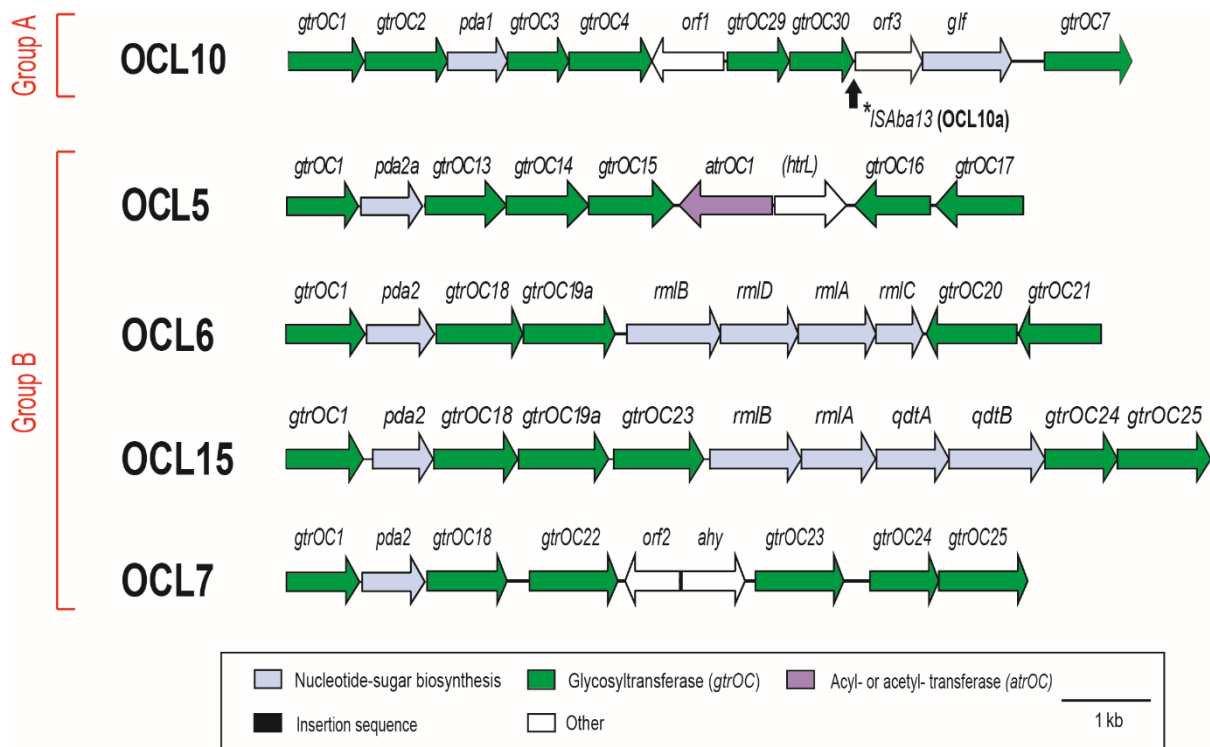
Figure 10. Total diversity at the OC locus responsible for outer-core lipooligosaccaride biosynthesis in ST25 clonal lineage. Loci are divided into their groups via the presence of *pda1* or *pda2* genes. Arrows indicate direction of transcription.

2.2.9 Comparison of KL and OCL Variation Amongst Other Clonal Lineages Indicates Extensive Genomic Diversity in ST25

Understanding the extent of KL and OCL diversity in ST25 requires a comparison to other major clonal lineages. Previous studies have assessed diversity amongst the major multi-drug resistant clonal groups, GC1 and GC2 (ST1 and ST2 in the Pasteur MLST scheme, respectively) (Holt et al., 2016; Schultz et al., 2016). However, these studies involved limited isolate collections that were mostly collected from the same hospital outbreaks or regions. Thus, isolates from the initial bulk download in this study that were found to belong to GC1, GC2 and ST10 by MLST analysis were also subjected to *Kaptive* analysis to elucidate KL and OCL types in these clones. Only KL and OCL locus matches with confidence levels above and including the 'Good' confidence threshold were included in the analysis. For the K locus analysis, this included 124 ST1, 1910 ST2, 47 ST10 and the 61 ST25 isolates identified

previously as part of this study. The OC locus analysis also contained a similar number of isolates for each lineage: 134 ST1, 1760 ST2, 49 ST10 and 61 ST25 isolates.

As the clone with the largest isolate representation, ST2 (n=1910) was found to have the most diversity at the K locus, with 30 KL types identified (Fig. 11a). The most common KL type was KL2, which accounted for 34% of KL types identified in ST2. However, only 4 OCL types were identified in the clone, with ~90% (n=1584) carrying the OCL1 gene cluster (Fig. 11b). ST1 had the second highest isolate count (n=134), with 12 different KL types found. KL1 was the most prominent, found in 33% of ST1 isolates. Notably, ST1 had more diversity at the OC locus than ST2, with 6 OCL types detected. The most common OCL type in ST1 being OCL1 (40% of isolates). ST10 had a similar isolate representation as ST25, with 49 isolates included in the comparative analysis. However, it demonstrated less diversity at both the KL and OCL region than ST25, with only 4 KL types and one OC locus type, OCL2. Comparatively, the ST25 collection included 19 KL and 5 OCL types. 35 of the 49 (~71%) ST10 collection had the KL type, KL49.
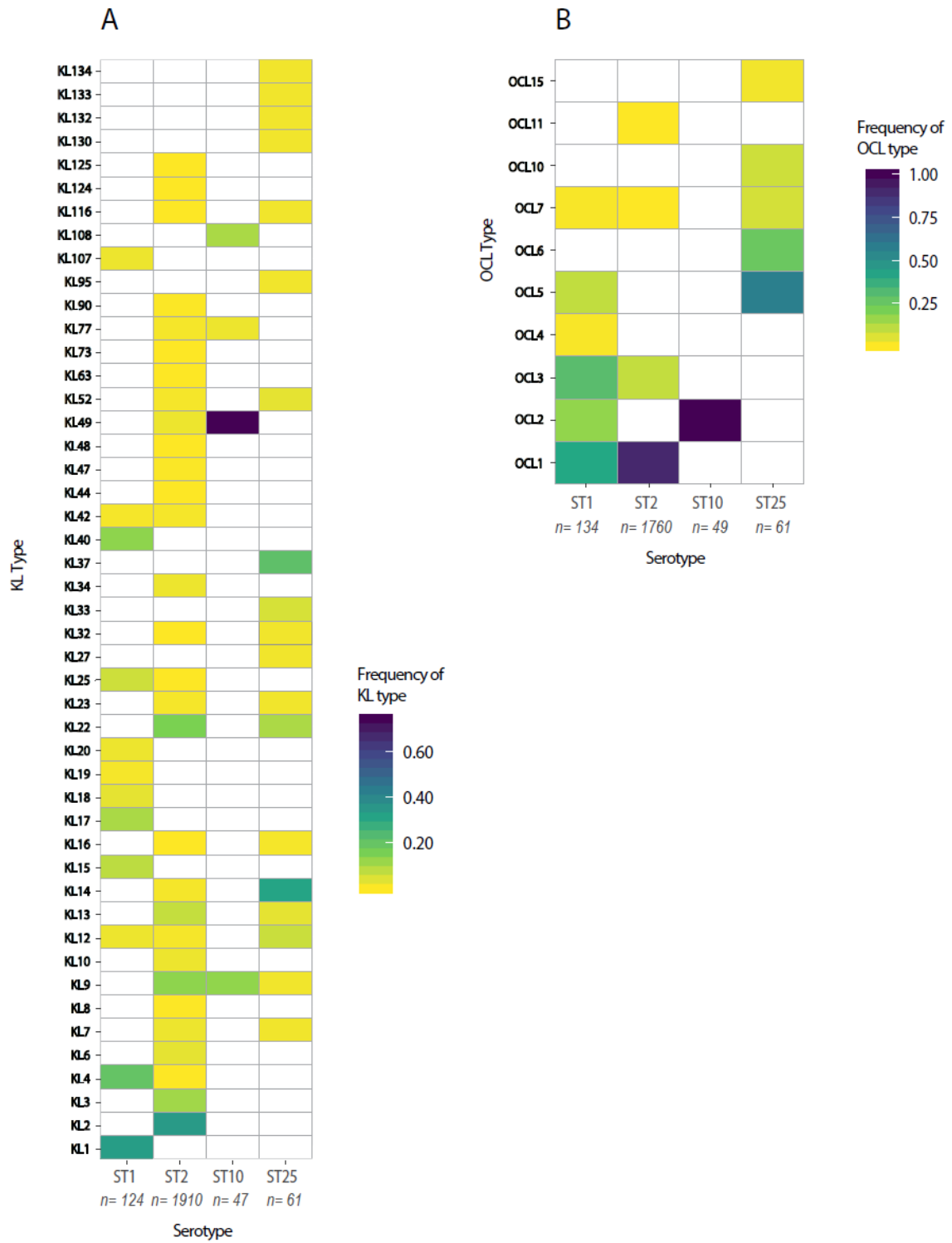
Figure 11: Frequency heatmap depicting distribution of KL **(A)** and OCL **(B)** locus types across four major clonal lineages, ST1, ST2, ST10 and ST25.

Chapter 3: Structural Determination of the CPS Produced by the K-Locus Identified in ST25

This chapter describes the approach taken to correlate newly available CPS structural information (determined by multi-institutional team of collaborators lead by Prof. Yuriy Knirel in Moscow, Russia) with KL sequences from known *A. baumannii* ST25 isolates. More specifically, the genetic content of the KL32 gene cluster identified in ST25 isolates from the previous aim (Chapter 2) was correlated to the elucidated K32 CPS structure. Direct comparison of the CPS structure to the genes found at the K locus not only allows the functions of encoded proteins to be predicted but ensures that all genes for the biosynthesis of this structure have been identified and not otherwise located in genomic regions or prophage outside the K locus. While *Kaptive* is foremost a tool for epidemiological typing, correlating the genes within a locus and the produced CPS structure of a given isolate will bolster the future potential for *Kaptive* to work as a tool for structure prediction.

Furthermore, this chapter collates all existing data on CPS structures identified in ST25 isolates to draw linkages between shared structural features and common genetic elements found at the K locus. Many studies have already elucidated the CPS structures produced by various *A. baumannii* isolates, including some belonging to ST25. However, the information on ST25 CPS structures and biosynthesis gene clusters has never been collated to build a comprehensive picture of the structural or genetic elements that are common within the ST25 lineage, thus identifying shared epitopes that may be used to inform therapeutic approaches.

The methodology and results presented in Sections 3.1.1 and 3.2.1, respectively, formed part of a manuscript that is published in Biochemistry (Moscow), accepted December 30th 2019 (Cahill et al., 2020), which is attached in *Appendix 5*. This publication presents the

structure and KL genetics of K32 produced by *A. baumannii* isolate LUH 5549. Cultivation of LUH 5549, subsequent isolation of the produced CPS, and the elucidation of the chemical structure by NMR and sugar analysis was completed by collaborators lead by Prof. Yuriy Knirel at the Russian Academy of Sciences. Bioinformatics analysis including the assembly of the LUH 5549 genome sequence, and the identification and annotation of the KL32 gene cluster, as well as the assignment of encoded proteins to their respective biosynthesis roles was completed by the MPhil candidate at Queensland University of Technology as part of this thesis.

3.1 METHODOLOGY

3.1.1 Elucidation of the capsule polysaccharide structure from *A. baumannii* strain LUH5549 via Smith degradation and NMR spectroscopy

Assigning encoded proteins in the KL to their respective biosynthesis roles in the CPS requires collaboration between biochemists and bioinformaticians. Subsequently, the elucidation of the CPS in the LUH5549 isolate was conducted by collaborators in Moscow, Russia across multiple institutes, the methods of which are detailed in the original publication (Cahill et al., 2020) and summarised here. The isolate used in the study, LUH5549, belongs to a collection of pre-2000s isolates from Leiden University in The Netherlands and has been a part of influential early genomic sequencing schemes (Traub & Bauer, 2000).

Briefly, the *A. baumannii* strain LUH5549 from the W. H. Traub collection Institut fur Medizinische Mikrobiologie und Hygiene, Universitat des Saarlandes (provided by Prof. Peter Reeves) was cultivated overnight and then harvested and dried. The CPS was then isolated and purified using standardised methodology including gel-permeation chromatography. Monosaccharide analysis was conducted to obtain alditol acetates by CPS hydrolysis. These

acetates were analysed using gas-liquid chromatography, and glucuronic acid in the CPS was identified using anion-exchange chromatography. Next, Smith degradation was used to yield the modified polysaccharide product (MPS). The final CPS product was elucidated and characterised using established NMR spectroscopy methodology as outlined in the original publication (Cahill et al., 2020).

3.1.2 Genome assembly and sequence typing of isolate LUH5549 from public short read data

The LUH 5549 genome was assembled from short reads obtained from the Sequence Read Archive (SRA number DRS005644) using SPAdes *v. 3.14* (Bankevich et al., 2012). To identify the sequence type of LUH 5549, the assembled genome sequence was subjected to MLST analysis using command-line MLST script and the PubMLST Institut Pasteur typing scheme for *A. baumannii* (abaumannii_2_scheme). The KL32 gene cluster was identified between conserved KL flanking genes, *fkpA* and *lldP* genes and annotated using the established nomenclature previously described (Kenyon & Hall, 2013).

3.1.3 Assignment of encoded proteins in the KL32 gene cluster to their biosynthesis roles in the production of the CPS from LUH 5549

The elucidated K32 CPS structure was used to correlate the presence of genes in KL32 by assigning encoded products to specific structural features including sugars in the K-unit, acetylation, pyruvylation and glycosylation patterns, as well as the linkages between sugars formed by glycosyltransferases and the linkages between K-units in the CPS chain formed by the Wzy polymerase.

This approach involved first identifying each ORF within KL32 using ORFFinder (https://www.ncbi.nlm.nih.gov/orffinder/) and submitting the translated protein sequences to BLASTp to identify possible homologues of known or predicted function. For all *gtr* genes

(*gtr67, gtr68, gtr69, gtr70*), the encoded products were further searched against the Carbohydrate-Active enZYmes database (CAZy) (http://www.cazy.org/) (Lombard et al., 2014) and Pfam (https://pfam.xfam.org/) (Finn et al., 2014) databases to confirm the assignment to glycosyltransferase family and to determine mechanism of glycosyltranfer (inverting or retaining) for each glycosyltransferase enzyme.

3.1.4 Summarising Literature Predicting the CPS Biosynthesis Roles of KL Found in ST25 isolates

Published research articles on the CPS structures and corresponding KL gene clusters in *A. baumannii* ST25 were collated, and the available structural and genetic information was compiled for direct comparison in the collection of ST25 isolates. CPS structures produced by ST25 isolates were redrawn following the standardised Symbol Nomenclature for Glycans (SNFG) format (Neelamegham et al., 2019; Varki et al., 2015) and linkages between K-units were correlated to the causative glycosyltransferase where possible.

3.2 RESULTS

3.2.1 Assembly of the LUH5549 Genome and Reannotation of the KL32 Gene Cluster

The LUH5549 genome was originally sequenced as part of an earlier study which charactered the KL gene clusters in an effort to develop a serotyping scheme, using a traditional nomenclature system (Hu et al., 2013). At this time, the KL gene cluster was described as the polysaccharide gene cluster (PSgc), and genes within were typed according to this previously used nomenclature system. Since then, many isolates included in that original study have had their produced CPS structures correlated with their KL gene cluster region. Because of this, many of these isolates have been reannotated and have had their KL gene clusters revised in

accordance with the more transparent and widely adopted K locus nomenclature system for *A. baumannii* used currently.

In the original 2013 study, LUH5549 was found to contain PSgc21 (GenBank accession number KC526897.1) (Hu et al., 2013). However, this gene cluster was found to be missing the capsule export operon which included genes *wza-wzb-wzc*, the absence of which indicates potential assembly problems. Therefore, the short reads of LUH5549 (SRA number DRS005644) were obtained from the SRA database and reassembled using SPAdes *v. 3.14*. The KL gene cluster of LUH5549 was located between conserved K locus flanking genes *fkpA* and *lldP*, and was found to be 99.56% identical to the KL32 gene cluster described for a Vietnamese *A. baumannii* isolate, BAL_058 (GenBank accession number KT359615.1) (Schultz et al., 2016). The KL gene cluster of LUH5549 was renamed KL32 and genes contained within were reannotated according to the current nomenclature used for KL gene clusters (Kenyon & Hall, 2013).

The KL32 gene cluster includes the conserved *A. baumannii* capsule export operon containing genes *wzc-wzb-wza*, as well as genes responsible for K-unit processing, both *wzy* and *wzx*. The *galU-ugd-gpi-gne1-pgm* simple sugar synthesis module indicates the likely presence of simple sugars in the K-units, such as UDP-D-Glcp, UDP-D-GlcpNAc, and UDP-D-GalpNAc. Upstream of this simple sugar synthesis module is *itrA2*, which is known to code for an initiating transferase which adds D-GalpNac as the first residue of the K-unit, binding it to the lipid carrier in the inner membrane (Kenyon, Marzaioli, et al., 2014). The variable central region of the KL gene cluster was found to contain four genes encoding for four glycosyltransferases, *gtr67, gtr68, gtr69* and *gtr70.* The presence of four *gtr* genes indicates that four internal KL unit sugar linkages are likely found in the produced CPS. The central region additionally contains a gene encoding an acetyltransferase (*atr9*), which indicates the potential acetylation of a sugar in the K-unit and an unidentified protein product (*orf*).

Interestingly, a second *ugd* gene exists in this central region, designated *ugd3* (GenPept accession number AHB32286.1). This *ugd3* was found to be only 20% identical (with 74% coverage) to the first *ugd* (GenPept accession number AHB32291.1) found in the simple sugar synthesis cluster of KL32. This *ugd* gene was first predicted to be responsible for conversion of UDP-D-Glc*p* to UDP-D-glucuronic acid (D-Glc*p*A) (Kenyon & Hall, 2013). While the role of *ugd* has been predicted before, and it is a feature common to the simple sugar synthesis cluster in all *A. baumannii* KL gene clusters, its role in CPS synthesis has never been established. In KL32, the *ugd3* gene was found to be 27% identical (71% coverage) to the second ugd gene type (*ugd2*) identified in the species, found in the central region of the KL20 and KL21 gene clusters. In both KL20 and KL21, the presence of Ugd2 (GenPept accession numbers AUG44319.1 and AIT56461.1) was found to be responsible for D-Glc*p*A in the produced CPS of both. Because of the similarity between *ugd2* and *ugd3*, and the established role of ugd*2* in other KL gene clusters, it could be speculated that the K32 CPS also contains D-Glc*p*A.

Figure 12. The KL32 capsule biosynthesis cluster from *A. baumannii* LUH5549. Arrows indicate direction of transcription. Genes encoding for glycosyltransferases are coloured in green, genes encoding for initiating transferase is coloured red and the gene responsible for repeat unit processing is yellow. Genes that have no established role in synthesis are coloured in a dark grey.

3.2.2 The KL32 gene cluster content is responsible for the produced CPS structure in *A. baumannii* isolate LUH5549

The CPS structure of LUH5549 was determined by Knirel and colleagues and details of this work is reported in Cahill et al. 2019. The K32 CPS structure was found to consist of branched

pentasaccharide repeats (K-units) that include two N-Acetyl-D-Galactosamine (D-Gal*p*NAc)

residues, one N-Acetyl-D-Glucosamine (D-Glc*p*NAc), one D-Glucose (D-Glc*p*), and one D-

Glucuronic acid (D-Glc*p*A) sugar (Fig.13a).

A

→3)-β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*NAc-(1→
4
↑
1
β-D-Glc*p*-(1→6)-α-Glc*p*NAc

B

Gtr67
β-D-Glc*p*-(1→6)-α-Glc*p*NAc
1
Gtr68 ↓   **Wzy**$_{K32}$
4   **Gtr69**   **Gtr70**   **[ItrA2]**
→3)-β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*NAc-(1→

Figure 13. Structure of the K32 CPS from *A. baumannii* LUH5549 as determined through NMR spectroscopy **(A)**. The K32 CPS structure starting from the first sugar in the K-unit, glycosyltransferase and initiating transferase products have been assigned to their predicted linkages **(B)**

As predicted above, the D-Gal*p*NAc, D-Glc*p*NAc, and D-Glc*p* sugars are synthesised by

proteins encoded in the simple sugar synthesis gene module in KL32 (on the right in Figure

12), and the presence of D-GlcA indicates the involvement of Ugd3 in the synthesis of this

sugar as discussed in the previous section.

As the K32 structure includes two D-Gal*p*NAc residues, it is unknown which sugar is

the first sugar added by ItrA2. Thus, the identity of the first sugar in the pentassacharide was

determined by examining the linkage between K-units in the CPS polymer formed by the

Wzy$_{K32}$ polymerase (Fig. 13b). A BLASTp query search revealed that Wzy$_{K32}$ (GenPept

accession number AHB32283.1) was 28% identical to Wzy$_{K116}$ (GenBank accession number

MK399425.1) encoded by *A. baumannii* KL116, which has been shown to form the β-D-

Gal*p*NAc-(1→3)-D-Gal*p*NAc linkage between units in the K116 CPS (Shashkov et al., 2019). In the structure of K32, there is an identical linkage and therefore Wzy$_{K32}$ could be assigned to it, establishing which D-Gal*p*NAc is the first sugar residue in the chain. The confirmed topology of the K-unit with the first sugar drawn on the right is shown in Figure 13a.

3.2.3 Determination of the Glycosyltransferase Functions in the K32 CPS via Assessment of Four *gtr* Genes in KL32

Pfam (https://pfam.xfam.org/) and CAZy (http://www.cazy.org/) databases were used to confirm the family assignment of all four Gtr in the gene cluster of KL32. The assignment of the four Gtr produced by the KL32 gene cluster are summarised in Table 5 and include GenPept accession number. This analysis showed that Gtr68$_{K32}$ was the only enzyme predicted to have a retaining mechanism, which retains the stereochemistry of the donor sugars anomeric bond (where α → α and β → β).

Table 5. Summary of Gtr classification and assignment in the KL32 gene cluster, including family assignment, GenPept accession number and linkage mechanism.

Glycosyltransferases	GenPept accession no.	Family assignment	Linkage mechanism
Gtr67$_{K32}$	AHB32281.1	glycos_transf_2_family (Pfam PF00535)	Inverting
Gtr68$_{K32}$	AHB32282.2	glycos_transf_4_family (Pfam PF00953)	Retaining
Gtr69$_{K32}$	AHB32284.1	glycos_transf_2_family (Pfam PF00535)	Inverting
Gtr70$_{K32}$	AHB32285.1	glycos_transf_2_family (Pfam PF00535)	Inverting

The assembly of the K-unit structure in Figure 12 requires three inverting glycosyltransferases to form the three β-linkages, and one retaining enzyme for the only α-linkage present. Therefore, Gtr68$_{K32}$ was assigned to the α-linkage, α-D-Glc*p*NAc-(1→4)-D-

Gal$p$NAc. The remaining glycosyltransferases with inverting mechanisms were assigned to linkages based on their homology to enzymes with known or predicted activities in other structures in bacterial species as follows.

Gtr70$_{K32}$ (Table 5) was identified as belonging to family 2 and was classed as an inverting enzyme. A BLASTp query search found no match to predicted enzymes in *A. baumannii*, but did show 54% amino acid sequence identity to WdbN (GenPept accession number STM86374.1) from the *E. coli* O143 O-antigen gene cluster. The structure of O143 has been determined elsewhere and contains a β-D-Glc$p$A-(1→3)-D-Glc$p$NAc linkage (Landersjö et al., 1996). A similar inverting linkage, β-D-GlcpA-(1→3)-D-GalpNAc, is found in K32 and therefore Gtr70$_{K32}$ was assigned to this linkage (Fig, 13b).

Gtr67$_{K32}$ was found to belong to glycosyltransferase family 2 and was classed as an inverting enzyme (Table 5). A BLASTp search found that Gtr67$_{K32}$ was 27% identical to Gtr75$_{K37}$ (GenBank accession number KX712115.1) encoded by the KL37 gene cluster of *A. baumannii,* while 27% similarity between genes within the same species is low, this similarity can still provide some insight into potential linkages. The K37 CPS structure has been elucidated and characterised previously (Arbatsky et al., 2015; Shashkov et al., 2019). For this structure, it was determined that Gtr75$_{K37}$ catalysed formation of the β-D-Glc$p$-(1→6)-D-Gal$p$NAc linkage. A similar linkage, β-D-Glc$p$-(1→6)-D-Glc$p$NAc, is also found in K32 as a side-branch, and hence Gtr67$_{K32}$ was assigned to this linkage (Fig. 13b). As these linkages are similar yet not identical, the low sequence identity between them (27%) can be contextualised.

The final glycosyltransferase, Gtr69$_{K32}$, was found to belong to glycosyltransferase family 2 and is classed as an inverting enzyme (Table 5). However, a BLASTp search found no significant match to existing proteins in the NCBI NR database. Because only one linkage was left in the pentasaccharide, and as this linkage required an inverting enzyme, it was

possible to predict that Gtr69$_{K32}$ would catalyse this linkage. Therefore, the final linkage, β-D-Gal$p$NAc-(1→4)-D-Glc$p$A, is predicted to be formed by Gtr69 (Fig. 13b).

3.2.4 Determination of other genes coding for products in the KL32 gene cluster of LUH5549

Earlier work theorised that the initial *ugd* present in the simple sugar synthesis module of *A. baumannii* isolates is redundant, and that a second *ugd* gene would be required for the synthesis of D-Glc$p$A in the resulting CPS (Kenyon & Hall, 2013). It was therefore hypothesised here following characterisation of the KL32 gene cluster that the second ugd gene in the locus, *ugd3*, would be responsible for synthesis of D-Glc$p$A in the CPS. Indeed, we saw in the K32 structure a D-GlcpA residue and therefore *ugd3* could be predicted to synthesise this particular sugar residue.

All structural features of the K32 CPS could therefore be assigned to products encoded by genes in the KL32 gene cluster. However, the KL32 gene cluster includes an additional 3 genes that could not be attributed to the CPS structure produced by LUH5549 (Fig. 12). The *gna* gene at the start of the gene cluster encodes a dehydrogenase that converts UDP-D-Glc$p$NAc to UDP-D-Glc$p$NAcA or UDP-D-Gal$p$NAc to UDP-D-Gal$p$NAcA and is not required for the synthesis of K32. Similarly, K32 does not contain acetyl or acyl groups, and therefore the role of Atr9 (GenPept accession number AHB32287.2) in the gene cluster could not be determined (Fig. 12). Furthermore, the product of *orf* (GenPept accession number AHB32288.2) found directly upstream of *atr9* had no match to predicted proteins in the GenPept database and could not be assigned a function in CPS synthesis. It is likely that these two genes are remnants of a past evolutionary event.

3.2.5 Comparison of the CPS structures produced by ST25 isolates and correlation of structural similarities to shared genetic content of KL types found in the lineage.

A review of the literature on *A. baumannii* CPS revealed that structures had been elucidated

for 13 of the 19 KL types found in the ST25 clonal lineage (Aim 1 of this study), including the

K32 structure reported in Cahill et al. 2019 described above (see Table 6 for the list of

references reporting these structures). These structures had been determined from strains

belonging to a variety of sequence types, not just ST25. In all cases, the produced CPS structure

correlated with the genes present at the KL region, and no additional genes outside the locus

were noted to affect CPS synthesis. For this comparison it is assumed that the same CPS

structures are produced by these KL types in ST25, and therefore additional genes outside the

K locus that may affect CPS synthesis are not present.

Table 6. Literature reporting structures derived from KL types found in the ST25 lineage.

K116	Shashkov, A. S., Cahill, S. M., Arbatsky, N. P., Westacott, A. C., Kasimova, A. A., Shneider, M. M., ... & Yanushevich, Y. G. (2019). Acinetobacter baumannii K116 capsular polysaccharide structure is a hybrid of the K14 and revised K37 structures. *Carbohydrate research*, *484*, 107774.
K37	Arbatsky, N. P., Shneider, M. M., Kenyon, J. J., Shashkov, A. S., Popova, A. V., Miroshnikov, K. A., ... & Knirel, Y. A. (2015). Structure of the neutral capsular polysaccharide of Acinetobacter baumannii NIPH146 that carries the KL37 capsule gene cluster. *Carbohydrate research*, *413*, 12-15.
K14	Kenyon, J. J., Hall, R. M., & De Castro, C. (2015). Structural determination of the K14 capsular polysaccharide from an ST25 Acinetobacter baumannii isolate, D46. *Carbohydrate research*, *417*, 52-56.
K27	Shashkov, A. S., Kenyon, J. J., Senchenkova, S. Y. N., Shneider, M. M., Popova, A. V., Arbatsky, N. P., ... & Knirel, Y. A. (2016). Acinetobacter baumannii K27 and K44 capsular polysaccharides have the same K-unit but different structures due to the presence of distinct wzy genes in otherwise closely related K gene clusters. Glycobiology, 26(5), 501-508.
K7	Arbatsky, N. P., Kenyon, J. J., Shashkov, A. S., Shneider, M. M., Popova, A. V., Kalinchuk, N. A., ... & Knirel, Y. A. (2019). The K5 capsular polysaccharide of the bacterium Acinetobacter baumannii SDF with the same K-unit containing Leg5Ac7Ac as the K7 capsular polysaccharide but a different linkage between the K-units. Russian Chemical Bulletin, 68(1), 163-167.

K33	Arbatsky, N. P., Shneider, M. M., Shashkov, A. S., Popova, A. V., Miroshnikov, K. A., Volozhantsev, N. V., & Knirel, Y. A. (2016). Structure of the N-acetylpseudaminic acid-containing capsular polysaccharide of Acinetobacter baumannii NIPH67. Russian Chemical Bulletin, 65(2), 588-591.
K16	Kenyon, J. J., Arbatsky, N. P., Sweeney, E. L., Shashkov, A. S., Shneider, M. M., Popova, A. V., ... & Knirel, Y. A. (2019). Production of the K16 capsular polysaccharide by Acinetobacter baumannii ST25 isolate D4 involves a novel glycosyltransferase encoded in the KL16 gene cluster. International journal of biological macromolecules, 128, 101-106.
K12	Kenyon, J. J., Marzaioli, A. M., Hall, R. M., & De Castro, C. (2015). Structure of the K12 capsule containing 5, 7-di-N-acetylacinetaminic acid from Acinetobacter baumannii isolate D36. Glycobiology, 25(8), 881-887.
K13	Kenyon, J. J., Kasimova, A. A., Notaro, A., Arbatsky, N. P., Speciale, I., Shashkov, A. S., ... & Knirel, Y. A. (2017). Acinetobacter baumannii K13 and K73 capsular polysaccharides differ only in K-unit side branches of novel non-2-ulosonic acids: di-N-acetylated forms of either acinetaminic acid or 8-epiacinetaminic acid. Carbohydrate research, 452, 149-155.
K52	Hu, D., Liu, B., Dijkshoorn, L., Wang, L., & Reeves, P. R. (2013). Diversity in the major polysaccharide antigen of Acinetobacter baumannii assessed by DNA sequencing, and development of a molecular serotyping scheme. PloS one, 8(7), e70329.
K9	Hu, D., Liu, B., Dijkshoorn, L., Wang, L., & Reeves, P. R. (2013). Diversity in the major polysaccharide antigen of Acinetobacter baumannii assessed by DNA sequencing, and development of a molecular serotyping scheme. PloS one, 8(7), e70329.
K32	Cahill, S. M., Arbatsky, N. P., Shashkov, A. S., Shneider, M. M., Popova, A. V., Hall, R. M., ... & Knirel, Y. A. (2020). Elucidation of the K32 Capsular Polysaccharide Structure and Characterization of the KL32 Gene Cluster of Acinetobacter baumannii LUH5549. Biochemistry (Moscow), 85(2), 241-247.
K22	Arbatsky, N. P., Shneider, M. M., Kenyon, J. J., Shashkov, A. S., Popova, A. V., Miroshnikov, K. A., Volozhantsev, N. V., & Knirel, Y. A. (2015). Structure of the neutral capsular polysaccharide of Acinetobacter baumannii NIPH146 that carries the KL37 capsule gene cluster. Carbohydrate research, *413*, 12–15.

These 13 structures were compiled into the same figure to draw direct and novel comparisons between the structures, along with the corresponding gene cluster. In all cases bar K22, a correlation between the structure and the corresponding gene cluster had been made previously and the functions of the encoded Gtrs, Itr and Wzy proteins had been assigned (shown in Figure 14). However, the following comparisons have been made between diverse CPS structures and their corresponding KL in this collection relating to ST25, rather than to the closest known CPS or KL otherwise found in the species as published in the literature. Shared sugars, linkages between sugars and linkages between K-unit structures correlate perfectly with shared sequence portions in the corresponding gene clusters. A direct

comparison of the KL gene clusters responsible for the synthesis of these structures is shown in Figure 14.
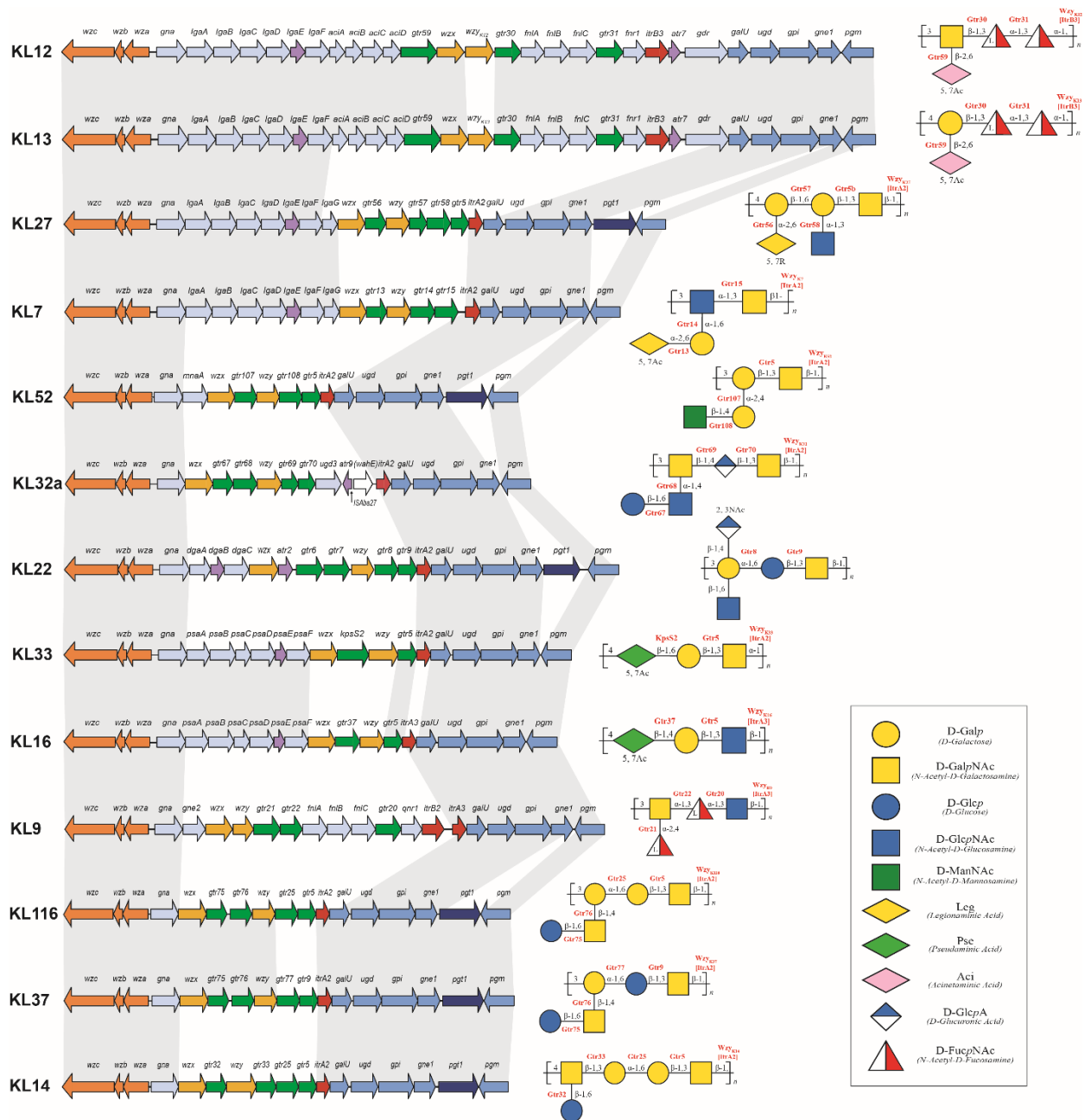
Figure 14. Summary of the 13 known CPS structures and their corresponding KL gene cluster in the ST25 isolates analysed in this study. CPS structures are presented in structure nomenclature for glycans (SNFG) and glycosyltransferase responsible for K-unit linkages are noted where known. Shared genes between adjacent gene clusters are indicated by grey shadow boxes.

This collection of K-unit structures have diverse topologies and structural features. Seven structures are pentasaccharides with differing side-branches of 1 or 2 sugar residues (Fig. 14). Four structures are tetrasaccharides and two are trisaccharides representing the K-units with the smallest number of sugar residues. Interestingly, complex non-2-ulosonic acid sugars (Aci, Leg, and Pse) are found in six different CPS structures, while three structures contain only simple monosaccharides.

The smallest trisaccharide K-units in the collection, K33 and K16, both contain Pse and the simple sugar, D-Galactose (D-Gal*p*), but differ in the third sugar and one of the two internal linkages. The linkages between these K-units are also different from each other. This corresponds to the shared presence of *gtr5* and *psa* genes for Pse synthesis in the KL33 and KL16 gene clusters, and presence of one different *gtr* gene and a different *wzy* gene (Fig. 14).

The CPS structures, K7 and K27 both contain the non-2-ulosonic acid, Leg5Ac7Ac in their K-unit. Similarly, the KL7 and KL27 gene clusters both include *lga* genes for Leg5Ac7Ac synthesis. However, K27 is a pentasaccharide unit with four glycosyltransferases responsible for the glycosidic bonds between each sugar residue, whereas K7 is a shorter tetrasaccharide with only four sugar residues glycosylated by three glycosyltransferase enzymes, and this is reflected by a difference of *gtr* genes in the corresponding gene clusters. The KL12 and KL13 gene clusters also contain *lga* genes, though they include the additional *aciABCD* genes for the extended synthesis pathway to produce Aci from a Leg precursor. KL12 and KL13 are closely related, and have been previously shown to differ only in the *wzy* gene sequence responsible for forming the linkage between K-units (Kenyon, Kasimova, et al., 2017).

The KL116, KL14 and KL37 gene clusters are also closely related, and have previously been shown to produce related CPS (Shashkov et al., 2019). However, none of these structures is similar to the ten other structures produced by KL found in the ST25 lineage.

The final 4 KL, KL52, KL9, KL32 and KL22 have largely differing genetic content and therefore produce unique CPS structures in this collection (Fig. 14). The CPS structure of K52 and K9 were amongst the first CPS structures of *A. baumannii* to be isolated and correlated with the genomic content at the K locus (Hu et al., 2013), though the nomenclature for gene clusters at this time differed from the nominated method (Kenyon & Hall, 2013) and were reannotated in this study. However, when comparing all CPS in this set, it is observed that K32 is the only structure to include D-GlcA due to the presence of *ugd3* in the KL32 gene cluster. Similarly, K52 is the only structure in this set to include N-Acetyl-D-mannosamine (ManNAc) due to the presence of *mnaA* gene in KL52. The K9 structure is also unique in that it includes an L-FucNAc sidebranch, though this sugar is also found in the mainchains of K12 and K13. Currently, the structure for K22 is published and discussed in various studies (Arbatsky et al., 2015; Yang et al., 2017). Yet, the assignment of glycosyltransferases to K-unit linkages is for the most part unpublished (Arbatsky et al., 2015). The structure of K22 is identical to another known structure, K3 and produces a side branch unique to the 13 K structures, UDP-D-Glc*p*NAc3NAcA (Kenyon & Hall, 2013). This side-branch is an *O*-acetylated derivative of 2,3-diacetamido-2,3-dideoxy-α-D-glucuronic acid, due to the presence of the complex sugar synthesis operon, *gna-dgaA-dgaB-dgaC* (Kenyon & Hall, 2013).

While the CPS structures of 13 of the 19 identified KL types in ST25 had been previously characterised in literature, the structures of six KL types (K23, K130, K132, K95, K133, K134) are yet to be determined to explore how the genetic variation identified in ST25 affects the CPS structure. The CPS structures for these seven types will be elucidated in future studies once the appropriate strains have been acquired for experimental analysis.

Chapter 4: Discussion

The following chapter provides an in-depth discussion of the results obtained from the *in-silico* exploration of both aims addressed in this project. Firstly, the extensive variation discovered at the genomic K and OC loci in the ST25 clonal lineage in Aim 1 will be discussed with respect to existing literature on the variation observed in other clonal lineages. Furthermore, the implications these results have on current epidemiological typing schemes will be explored in detail, outlining potential future directions of the project and possible areas of expansion regarding surface polysaccharide research. Secondly, the results from Aim 2 will be discussed in the context of how these findings add value to existing knowledge and guide future studies informing therapeutics research. Finally, the limitations of this study will be addressed, and future directions will be proposed for expanding this research to explore the evolution of the ST25 lineage.

4.1   GENERAL DISCUSSION

The first aim of this study was to identify the level of variation at the K and OC loci in the genomes of isolates belonging to the ST25 clonal lineage of *A. baumannii*. The decision to focus on the ST25 clonal lineage is multifaceted but is driven by a major knowledge gap concerning the surface polysaccharide variation within the emerging ST25 global clone. Though ST25 isolates have been recovered in many countries, studies are often limited to outbreaks in a single geographical location (da Silva et al., 2018; Di Popolo et al., 2011; Stietz et al., 2013; Zarrilli et al., 2011), and therefore a complete understanding of global dissemination is yet to be ascertained. As a lineage that has been shown to harbour extensive AMR, with a significant number of isolates showing resistance to last-line drugs, tracking the

spread of ST25 globally is a priority. Previous studies have investigated the efficacy of the KL and OCL genomic regions as epidemiological markers in other clonal groups. Specifically, a 2016 study on the evolution of GC1 proved that multiple successful subclones of GC1 existed globally, and that GC1 could be delineated into sublineages through distinctions in their surface polysaccharide gene loci (Holt et al., 2016). Since this time, a similar finding has been made for other *A. baumannii* clonal lineages, including GC2 and emerging clones such as ST10 (Hamidian & Nigro, 2019; Leungtongkam et al., 2018; Meumann et al., 2019). While WGS technologies allow for precise epidemiological tracing of isolates using schemes such as MLST, the high degree of genome plasticity and frequent recombination seen in the A. baumannii species stresses the requirement for a combination of distinct genomic regions as epidemiological markers. As KL and OCL are distinct regions, which experience frequent recombination non-tangentially, they are attractive targets for a highly specific typing scheme (Wyres et al., 2020).

The first step was to gather existing whole genome sequences of *A. baumannii* from public databases and to identify which of these genomes belong to the ST25 clonal lineage. Previous studies had taken an *in-silico* approach to obtaining genome sequences of ST25 with some success. Specifically, a 2015 study assessed genomic diversity in the ST25 lineage by taking a dual *in-silico* and *in-vitro* approach, sequencing 7 known ST25 genomes, and downloading a further 12 publicly deposited ST25 genomes (Sahl et al., 2015). This approach allowed for an assessment of phenotypic diversity, in correlation with genotypic traits. However, the overall genome pool in this study was limited to 19 ST25 sequences. While this could be reflective of the limited number of publicly available ST25 genomes at that time, the number of genomes restricted the understanding gained on the global dissemination and genotypic diversity of ST25. Since 2015, further ST25 genomes have been publicly released

providing the current study with a sizeable genome pool to better understand the genomic diversity at the K and OC locus in ST25.

In this study, 82 ST25 genome sequences were gathered from various public databases, including NCBI WGS and NR databases as well as ENA SRA, many of which were sequenced in or after 2015 (42/82 genomes). A large portion of the newly sequenced 42 isolates (32) were from multiple hospital associated *A. baumannii* outbreaks in Bolivia (Cerezales et al., 2019; Cerezales et al., 2020). The ST25 clonal lineage is often colloquially referred to as the South American clone, due to its prevalence on the continent, and this trend was seen in the 82 genome sequences. The study pool contained 34 genomes from Argentina, Bolivia and Colombia, with 3 isolates seen in nearby Central American countries including Honduras and Mexico. However, it is important to note that the genome collection in this study identified outbreaks of ST25 outside of South America, in particular within Asian countries, which included 6 Thailand and 10 Vietnam isolates. Furthermore, numerous sporadic isolates were noted worldwide, including in geographically isolated locations like Australia. Thus, it can be deduced that ST25 is globally disseminated and has been observed on nearly every continent.

The original 82 ST25 genomes were not only geographically dispersed, but had been isolated over four decades, with the oldest isolate in the collection being from 1985 and the most recent sequence isolated in 2018. This extensive range likely illustrates that ST25 is not a newly emerged clone, but rather had delineated from an ancestral genetic pool. The same 2015 study that assessed genomic diversity in a smaller collection of ST25 genomes concluded through their phylogenomic analysis, that the SNP (single nucleotide polymorphism) density seen in the ST25 lineage was likely due to homoplasy driven by homologous recombination events (Sahl et al., 2015). This is also similar to what has been observed for GC1 genomes (Hamidian et al., 2019; Holt et al., 2016). Further, while many notable STs belong to a clonal complex (e.g. CC1 including ST1, ST7, ST8, ST19 and ST20), ST25 is genetically distinct

from any clonal complex, with studies indicating a likely delineation from CC3 sometime in the distant past, due to the high number of allelic mismatches shared between (Diancourt et al., 2010). The genetic distinction from other clonal lineages indicates a pattern of true clonal spread of ST25 globally, as opposed to possible spontaneous generation events. The collection of 82 ST25 genomes presents an opportunity to assess the genomic evolution of the ST25 clone in further detail in a future study expanding on the information gained in this project on the KL and OCL diversity.

Unfortunately, the present study found that the complete collection of 82 ST25 genomes were not all of good sequence quality. In-house quality control of these genomes revealed that 21 assemblies that had untypable K and OC loci also had a large number of contigs indicating potential poor genome sequence quality. However, other isolates that were also found to have poor quality genome assemblies were included in this study if the KL and OCL regions could be recovered and typed. Therefore, it is also possible that the 61 genomes with typable KL and OCL regions may have issues in other areas of the genome, and an in-depth quality control with specific exclusion criteria will need to be implemented on these ST25 genomes for any future analysis.

The hypothesis of the KL and OCL regions representing viable epidemiological markers to track ST25 clonal isolates was demonstrated by the finding of diverse KL and OCL regions in *A. baumannii* ST25 genomes studied. Trends could also be deduced from a combination of distinct KL and OCL types. The most common combination of KL and OCL type was KL14 with OCL6, which was seen in five countries: U.S.A, Germany, Australia, United Arab Emirates, and the Netherlands. Several countries were also found to carry unique combinations or singleton KL or OCL types in the collection suggesting a single outbreak strain circulating in these regions. For example, Japan was the only country with the combination of KL14 and OCL15, and similarly Mexico with KL7 and OCL7, Italy with KL27 and OCL5,

and Iraq with KL12 and OCL5, though these were all countries with only one ST25 isolate noted. However, KL types were paired with an array of OCL types and vice versa, indicating that these genomic regions evolve independently from each other and may also be replaced via homologous recombination as seen in GC1 (Holt et al., 2016).

It is important to note that certain countries, particularly those with endemic ST25 strains such as Bolivia and Vietnam, presented multiple locus type combinations. Bolivia isolates included the KL types: KL37, KL33 and KL22 as well as the OCL types: OCL10a, OCL10 and OCL5. These could indicate multiple specific strains of *A. baumannii* in circulation through the healthcare of Bolivia, as the Bolivian isolates in this pool were known to be collected from more than one hospital location.

Interestingly, the comparison of KL and OCL variation in multiple lineages, including GC1, GC2, ST10 and ST25 indicated that ST25 has notably more diverse genomic content at the KL and OCL regions in consideration of the number of genomes available. While GC2 had the most KL types found (30), this variation can be explained by the overwhelming number of publicly available GC2 genomes included in this study (n= 1910). As there were similar isolate counts for ST10 and ST25, with 47 and 61 respectively, it was expected that similar diversity at the KL and OCL would be seen. Another reason to suspect similar diversity in both lineages, is because both are considered globally prevalent MDR clades. ST10 has shown enhanced virulence potential, and has been implicated in high mortality rate outbreaks in a number of hospitals globally (Abhari et al., 2019; Jones et al., 2015). Yet, the results of this analysis revealed that the 47 ST10 isolates carried only 4 KL types and 1 OCL type, which is dramatically less than the 19 KL and 6 OCL types found amongst the 61 ST25 genomes. This may be due to a bias in publicly available sequences if the isolates were recovered from a homogenous sample group including similar geographical and clinical backgrounds, which would have some effect on how varied these surface polysaccharide structures would be.

However, brief investigation of the metadata associated with 38 of the 49 ST10 isolates revealed similar country distribution and collection dates. The 36 ST10 isolates were recovered from human and 2 were isolated from companion dogs in Germany, with the oldest isolate collected in 1994 and the latest isolate collected in 2016. Unlike ST25, 5 of the ST10 isolates were from providences across China, though this is the most notable difference in metadata between both collections. Interestingly, ~71% of ST10 isolates had the KL49 capsule type. The CPS produced by this KL type includes a complex sugar synthesis module for the production of legionaminic acid. Recent studies have alluded to the enhanced virulence associated with strains possessing KL49 (Deng et al., 2020), as the addition of legionaminic acid aids the bacterium in mimicking the host cell surface, preventing detection from the host immune system (Jones et al., 2015). Ultimately, further investigation is needed to understand why the KL and OCL in ST25 are so extensively varied, and furthermore, why ST10 demonstrates so little KL and OCL variation.

It is likely that there are external factors here influencing capsule divergence and driving whole capsule switching events in ST25, though the reason behind capsule switching in *A. baumannii* is currently unclear (Holt et al., 2016). It is speculated that host-immune factors, pressure from antimicrobials or bacteriophage could be drivers in genomic diversity at the KL and OCL region (Geisinger & Isberg, 2015; Holt et al., 2016). However, to date, no study has investigated the exact reasons why bacteria produce a wide variety of structures of the same surface polysaccharide. Therefore, a future direction of this project would be to assess the reason for KL or OCL phenotypes *in vitro* using an isogenic background for direct comparison of types in a variety of biological assays assessing specific components of virulence, such as biofilm formation, adherence and invasion of host cells potential and desiccation resistance.

The second aim of this project was complementary to the first aim and provided a correlation of the KL content with the corresponding CPS structure. However, the K32 structure was the only CPS structure for ST25 that had been made available for analysis from our collaborators, and many of the most common KL types found in this lineage had been correlated with a structure previously. The study of the K32 structure and KL32 gene cluster added to the pool of KL gene clusters found in ST25 that had corresponding CPS structural data available in the literature. The analysis of KL32 in conjunction with the structural data allowed the function of various enzymes, such as glycosyltransferases and initiating transferases to be inferred, strengthening our understanding of enzyme specificity in CPS synthesis. The comparison of the structures produced by 13 of the 19 KL types identified amongst the 61 ST25 isolates indicates that the vast genomic diversity seen at the K locus translates to incredibly varied CPS structures. This CPS diversity is not only limited to the simple and complex sugars which make up the CPS K-units, but also the linkages which bind the K-units together and has many implications for future research. The diversity CPS complicates the design of phage therapies, the most current of which utilise structural depolymerase tail spikes to recognise and digest specific CPS (Knirel et al., 2020; Oliveira et al., 2019). The specificity of these phage combined with the noted diversity of CPS in *A. baumannii* highlights the necessity of phage libraries in treating *A. baumannii* infections. Furthermore, the dramatic variation seen in a single sequence type of *A. baumannii* (ST25) in this study further indicates that phage libraries may be the most appropriate option for treatment of infections caused by *A. baumannii*.

Recently, studies have considered the value of specific sugars identified in capsules as an epitope in glycoconjugate vaccine development. Namely, the use of pseudaminic acid as an epitope in a prophylactic glycoconjugate vaccine was able to induce antibodies that could recognise surface polysaccharide from ~40% of 250 other clinical *A. baumannii* strains (Lee et

al., 2018). Indeed, glycoconjugate vaccines are a promising therapeutic for option for treating a range of bacterial infections, including *A. baumannii*. However, the complexity observed at the CPS in *A. baumannii* again presents as an issue in development of a vaccine. Conversely, the variability in genomic content at the K locus has proved advantageous in other recent studies concerning both treatment and identification of hypervirulent isolates (Deng et al., 2020). Using schemes such as MLST genotyping to identify hypervirulent CRAB isolates is not optimal due to the spread of virulence elements across multiple genotypes. However, identifying genes conserved across hypervirulent isolates allows for unique strategies into tracking and treatment of these isolate, as proven in recent studies (Deng et al., 2020; Hua et al., 2021). Specifically, *gtr100* found only in KL49, a KL type noted to coincide with a hypervirulence phenotype, proved a valuable target for rapid identification. Moreover, *in vitro* studies indicated that the linkage established by *gtr100* specifically is closely related to the virulence potential of these KL49 *A. baumannii* isolates (Deng et al., 2020). These recent discoveries emphasise the need for elucidation of new CPS structures, and the importance of correlating structures with genomic content with at the KL region, as this underpins research into future therapeutics and methods of rapid identification.

In summary, the extensive variation seen at the KL and OCL in ST25 isolates establishes these regions as ideal candidates for epidemiological markers, as seen in other clinically relevant *A. baumannii* lineages (Hamidian & Nigro, 2019; Holt et al., 2016). The advantages of *in-silico* typing schemes utilising WGS and the K and OC locus, such as *Kaptive*, have been affirmed. Importantly, the insight gained in this study on the global dissemination and genomic diversity of ST25 warrants further investigation into the genomic evolution of this unique clade, with specific emphasis on capsule switching and evolution. Furthermore, the ease of *in-silico* typing using *Kaptive*, and the potential knowledge on clade dissemination and diversity, opens the door for future studies on other notable *A. baumannii* clades, such as ST10.

This study has also stressed the importance of combined typing schemes, especially in bacterial species with notably high recombination potential, such as that seen here in *A. baumannii*. Recent studies have stressed the importance of utilising multiple genomic sequence typing schemes, (Kongthai et al., 2020; Wyres et al., 2020), due to the high frequency of recombination occurring within the species. By combining schemes like MLST with the surface polysaccharide loci gene typing explored here, it is possible to generate a high degree of distinction between otherwise similar isolates as determined using only one scheme. Further, the specificity achieved through typing these gene loci has implications for epidemiological tracking in not only *A. baumannii*, but potentially other clinically relevant Gram-negative bacterial pathogens, such as *E. coli*. The comparisons of known ST25 CPS structures also informs the design of future therapeutic studies, specifically those involving phage therapies. Importantly, the CPS diversity noted in ST25 further solidifies the current thought that a phage library is necessary for treating *A. baumannii* infections.

Chapter 5: Bibliography

Abhari, S. S., Badmasti, F., Modiri, L., Aslani, M. M., & Asmar, M. (2019). Circulation of imipenem-resistant Acinetobacter baumannii ST10, ST2 and ST3 in a university teaching hospital from Tehran, Iran. *Journal of Medical Microbiology, 68*(6), 860-865. https://doi.org/https://doi.org/10.1099/jmm.0.000987

Adams, M., Wright, M., Karichu, J., Venepally, P., Fouts, D., Chan, A., Richter, S., Jacobs, M., & Bonomo, R. (2019). Rapid replacement of *Acinetobacter baumannii* strains accompanied by changes in lipooligosaccharide loci and resistance gene repertoire. *mBio, 10*(2), e00356-00319.

Ahmed, S. S., & Alp, E. (2015, Apr 15). Genotyping methods for monitoring the epidemic evolution of A. baumannii strains. *J Infect Dev Ctries, 9*(4), 347-354. https://doi.org/10.3855/jidc.6201

Antunes, N. T., & Fisher, J. F. (2014, Aug 21). Acquired Class D β-Lactamases. *Antibiotics (Basel), 3*(3), 398-434. https://doi.org/10.3390/antibiotics3030398

Arbatsky, N., Kenyon, J., Shashkov, A., Shneider, M., Popova, A., Kalinchuk, N., Hall, R., & Knirel, Y. (2019). The K5 capsular polysaccharide of the bacterium *Acinetobacter baumannii* SDF with the same K unit containing Leg5Ac7Ac as the K7 capsular polysaccharide but a different linkage between the K units. . *Russ. Chem. Bull., Int. Ed. 68*(1), 163-167.

Arbatsky, N., Shneider, M., Kenyon, J., Shashkov, A., Popova, A., Miroshinikov, K., Volozhantsev, N., & Knirel, Y. (2015). Structure of the neutral capsular polysaccharide of *Acinetobacter baumannii* NIPH146 that carries the KL37 capsule gene cluster. *Carbohydr Res, 413*, 12-15.

Arbatsky, N. P., Kenyon, J. J., Kasimova, A. A., Shashkov, A. S., Shneider, M. M., Popova, A. V., Knirel, Y. A., & Hall, R. M. (2019, 2019/09/01/). K units of the K8 and K54 capsular polysaccharides produced by Acinetobacter baumannii BAL 097 and RCH52 have the same structure but contain different di-N-acyl derivatives of legionaminic acid and are linked differently. *Carbohydr Res, 483*, 107745. https://doi.org/https://doi.org/10.1016/j.carres.2019.107745

Ayoub Moubareck, C., & Hammoudi Halat, D. (2020, Mar 12). Insights into Acinetobacter baumannii: A Review of Microbiological, Virulence, and Resistance Traits in a Threatening Nosocomial Pathogen. *Antibiotics (Basel), 9*(3). https://doi.org/10.3390/antibiotics9030119

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A., Lesin, V., Nikolenko, S., Pham, S., Prjibelski, A., Pyshkin, A., Sirotkin, A., Vyahhi, N., Tesler, G., Alekseyev, M., & Pevzner, P. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol., 19*(5), 455–477.

Bartual, S., Seifert, H., Hippler, C., Luzon, M., Wisplinghoff, H., & Rodríguez-Valera, F. (2005). Development of a multilocus sequence typing scheme for characterization of clinical isolates of *Acinetobacter baumannii. Journal of Clinical Microbiology, 43*(9), 4382-4390.

Bergogne-Bérézin, E., & Towner, K. J. (1996, Apr). Acinetobacter spp. as nosocomial pathogens: microbiological, clinical, and epidemiological features. *Clin Microbiol Rev, 9*(2), 148-165.

Blossom, D. B., & Srinivasan, A. (2008). Drug-Resistant Acinetobacter baumannii-calcoaceticus Complex: An Emerging Nosocomial Pathogen With Few Treatment Options. *Infectious Diseases in Clinical Practice, 16*(1), 1-3. https://doi.org/10.1097/ipc.0b013e3181635def

Bouvet, P. J. M., & Grimont, P. A. D. (1986). Taxonomy of the Genus Acinetobacter with the Recognition of Acinetobacter baumannii sp. nov., Acinetobacter haemolyticus sp. nov., Acinetobacter johnsonii sp. nov., and Acinetobacter junii sp. nov. and Emended Descriptions of Acinetobacter calcoaceticus and Acinetobacter lwoffii. *International Journal of Systematic and Evolutionary Microbiology, 36*(2), 228-240. https://doi.org/https://doi.org/10.1099/00207713-36-2-228

Brisse, S., Issenhuth-Jeanjean, S., & Grimont, P. A. (2004, Aug). Molecular serotyping of Klebsiella species isolates by restriction of the amplified capsular antigen gene cluster. *J Clin Microbiol, 42*(8), 3388-3398. https://doi.org/10.1128/jcm.42.8.3388-3398.2004

Cahill, S. M., Arbatsky, N. P., Shashkov, A. S., Shneider, M. M., Popova, A. V., Hall, R. M., Kenyon, J. J., & Knirel, Y. A. (2020, 2020/02/01). Elucidation of the K32 Capsular Polysaccharide Structure and Characterization of the KL32 Gene Cluster of Acinetobacter baumannii LUH5549. *Biochemistry (Moscow), 85*(2), 241-247. https://doi.org/10.1134/S000629792002011X

Cai, Y., Chai, D., Wang, R., Liang, B., & Bai, N. (2012, Jul). Colistin resistance of Acinetobacter baumannii: clinical reports, mechanisms and antimicrobial strategies. *J Antimicrob Chemother, 67*(7), 1607-1615. https://doi.org/10.1093/jac/dks084

Carretto, E., Barbarini, D., Dijkshoorn, L., van der Reijden, T. J., Brisse, S., Passet, V., & Farina, C. (2011, Aug). Widespread carbapenem resistant Acinetobacter baumannii clones in Italian hospitals revealed by a multicenter study. *Infect Genet Evol, 11*(6), 1319-1326. https://doi.org/10.1016/j.meegid.2011.04.024

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012, 2012/02//). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics (Oxford, England), 28*(4), 464-469. https://doi.org/10.1093/bioinformatics/btr703

Cerezales, M., Xanthopoulou, K., Wille, J., Bustamante, Z., Seifert, H., Gallego, L., & Higgins, P. G. (2019, 2019/06/01/). Acinetobacter baumannii analysis by core genome multi-locus sequence typing in two hospitals in Bolivia: endemicity of international clone 7 isolates (CC25). *International Journal of Antimicrobial Agents, 53*(6), 844-849. https://doi.org/https://doi.org/10.1016/j.ijantimicag.2019.03.019

Cerezales, M., Xanthopoulou, K., Wille, J., Krut, O., Seifert, H., Gallego, L., & Higgins, P. G. (2020). Mobile Genetic Elements Harboring Antibiotic Resistance Determinants in Acinetobacter baumannii Isolates From Bolivia. *Frontiers in Microbiology, 11*(919). https://doi.org/10.3389/fmicb.2020.00919

Centres for Disease, Control and Prevention. (CDC). (2019). Antibiotic resistance threats in the United States, 2019 [Report]. https://doi.org/http://dx.doi.org/10.15620/cdc:82532

Corbett, D., & Roberts, I. (2008). Capsular polysaccharides in *Escherichia coli*. In A. Laskin, S. Sariaslani, & G. Gadd (Eds.), *Advances in Applied Microbiology* (Vol. 65). Academic Press.

Cornejo-Juárez, P., Cevallos, M. A., Castro-Jaimes, S., Castillo-Ramírez, S., Velázquez-Acosta, C., Martínez-Oliva, D., Pérez-Oseguera, A., Rivera-Buendía, F., & Volkow-Fernández, P. (2020). High mortality in an outbreak of multidrug resistant Acinetobacter baumannii infection introduced to an oncological hospital by a patient transferred from a general hospital. *PLOS ONE, 15*(7), e0234684. https://doi.org/10.1371/journal.pone.0234684

Coyne, S., Courvalin, P., & Périchon, B. (2011, Mar). Efflux-mediated antibiotic resistance in Acinetobacter spp. *Antimicrob Agents Chemother, 55*(3), 947-953. https://doi.org/10.1128/aac.01388-10

da Silva, K. E., Maciel, W. G., Croda, J., Cayô, R., Ramos, A. C., de Sales, R. O., Kurihara, M. N. L., Vasconcelos, N. G., Gales, A. C., & Simionatto, S. (2018). A high mortality rate associated with multidrug-resistant Acinetobacter baumannii ST79 and ST25 carrying OXA-23 in a Brazilian intensive care unit. *PLOS ONE, 13*(12), e0209367. https://doi.org/10.1371/journal.pone.0209367

Dallo, S. F., & Weitao, T. (2010, Apr). Insights into acinetobacter war-wound infections, biofilms, and control. *Adv Skin Wound Care, 23*(4), 169-174. https://doi.org/10.1097/01.ASW.0000363527.08501.a3

Daniels, C. C., Rogers, P. D., & Shelton, C. M. (2016, Jan-Feb). A Review of Pneumococcal Vaccines: Current Polysaccharide Vaccine Recommendations and Future Protein Antigens. *J Pediatr Pharmacol Ther, 21*(1), 27-35. https://doi.org/10.5863/1551-6776-21.1.27

De Oliveira, D. M. P., Forde, B. M., Kidd, T. J., Harris, P. N. A., Schembri, M. A., Beatson, S. A., Paterson, D. L., & Walker, M. J. (2020). Antimicrobial Resistance in ESKAPE Pathogens. *Clinical Microbiology Reviews, 33*(3), e00181-00119. https://doi.org/10.1128/cmr.00181-19

Deng, Q., Zhang, J., Zhang, M., Liu, Z., Zhong, Y., Liu, S., Cui, R., Shi, Y., Zeng, H., Yang, X., Lin, C., Luo, Y., Chen, H., Wu, W., Wu, J., Zhang, T., Lu, Y., Liu, X., Zou, Q., & Huang, W. (2020). Rapid Identification of KL49 Acinetobacter baumannii Associated with Clinical Mortality. *Infect Drug Resist, 13*, 4125-4132. https://doi.org/10.2147/idr.S278891

Di Popolo, A., Giannouli, M., Triassi, M., Brisse, S., & Zarrilli, R. (2011, Feb). Molecular epidemiological investigation of multidrug-resistant Acinetobacter baumannii strains in four Mediterranean countries with a multilocus sequence typing scheme. *Clin Microbiol Infect, 17*(2), 197-201. https://doi.org/10.1111/j.1469-0691.2010.03254.x

Diancourt, L., Passet, V., Nemec, A., Dijkshoorn, L., & Brisse, S. (2010). The Population Structure of Acinetobacter baumannii: Expanding Multiresistant Clones from an Ancestral Susceptible Genetic Pool. *PLOS ONE, 5*(4), e10034. https://doi.org/10.1371/journal.pone.0010034

Dijkshoorn, L., Aucken, H., Gerner-Smidt, P., Janssen, P., Kaufmann, M. E., Garaizar, J., Ursing, J., & Pitt, T. L. (1996, Jun). Comparison of outbreak and nonoutbreak

Acinetobacter baumannii strains by genotypic and phenotypic methods. *J Clin Microbiol, 34*(6), 1519-1525.

Evans, B. A., & Amyes, S. G. (2014, Apr). OXA β-lactamases. *Clin Microbiol Rev, 27*(2), 241-263. https://doi.org/10.1128/cmr.00117-13

Finn, R., Bateman, A., Clements, J., Coggill, P., Eberhardt, R., Eddy, S., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E., Tate, J., & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res, 42*, D222-D230.

Gao, L., Lyu, Y., & Li, Y. (2017). Trends in Drug Resistance of Acinetobacter baumannii over a 10-year Period: Nationwide Data from the China Surveillance of Antimicrobial Resistance Program. *Chinese medical journal, 130*(6), 659-664. https://doi.org/10.4103/0366-6999.201601

Geisinger, E., & Isberg, R. (2015). Antibiotic modulation of capsular exopolysaccharide and virulence in *Acinetobacter baumannii*. *PLoS Pathogens, 11*(2), e1004691.

Ghajavand, H., Esfahani, B., Havaei, A., Fazeli, H., Jafari, R., & Moghim, S. (2017). Isolation of bacteriophages against multidrug resistant *Acinetobacter baumannii*. *Res Pharm Sci, 12*(5), 373–380.

Giannouli, M., Antunes, L. C., Marchetti, V., Triassi, M., Visca, P., & Zarrilli, R. (2013, Jun 20). Virulence-related traits of epidemic Acinetobacter baumannii strains belonging to the international clonal lineages I-III and to the emerging genotypes ST25 and ST78. *BMC Infect Dis, 13*, 282. https://doi.org/10.1186/1471-2334-13-282

Glaze, P., Watson, D., Young, N., & Tanner, M. (2008). Biosynthesis of CMP-N,N'-diacetyllegionaminic acid from UDP-N,N'-diacetylbacillosamine in Legionella pneumophila. *Biochemistry, 47*(10), 3272-3282.

Gogou, V., Pournaras, S., Giannouli, M., Voulgari, E., Piperaki, E. T., Zarrilli, R., & Tsakris, A. (2011, Dec). Evolution of multidrug-resistant Acinetobacter baumannii clonal lineages: a 10 year study in Greece (2000-09). *J Antimicrob Chemother, 66*(12), 2767-2772. https://doi.org/10.1093/jac/dkr390

Goon, S., Kelly, J. F., Logan, S. M., Ewing, C. P., & Guerry, P. (2003). Pseudaminic acid, the major modification on Campylobacter flagellin, is synthesized via the Cj1293 gene. *Molecular microbiology, 50*(2), 659-671.

Gordillo Altamirano, F. L., & Barr, J. J. (2019, Apr). Phage Therapy in the Postantibiotic Era. *Clin Microbiol Rev, 32*(2). https://doi.org/10.1128/cmr.00066-18

Gordon, N. C., & Wareham, D. W. (2010, Mar). Multidrug-resistant Acinetobacter baumannii: mechanisms of virulence and resistance. *Int J Antimicrob Agents, 35*(3), 219-226. https://doi.org/10.1016/j.ijantimicag.2009.10.024

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013, Apr 15). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England), 29*(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086

Hamidian, M., & Hall, R. (2013). ISAba1 targets a specific position upstream of the intrinsic *ampC* gene of *Acinetobacter baumannii* leading to cephalosporin resistance. *J Antimicrob Chemother, 68*, 2682-2683.

Hamidian, M., & Hall, R. (2016). The resistance gene complement of D4, a multiply antibiotic-resistant ST25 *Acinetobacter baumannii* isolate, resides in two genomic islands and a plasmid. *J Antimicrob Chemother, 71*, 1730-1841.

Hamidian, M., & Hall, R. M. (2011, Nov). AbaR4 replaces AbaR3 in a carbapenem-resistant *Acinetobacter baumannii* isolate belonging to global clone 1 from an Australian hospital. *J Antimicrob Chemother, 66*(11), 2484-2491. https://doi.org/10.1093/jac/dkr356

Hamidian, M., Hawkey, J., Wick, R., Holt, K. E., & Hall, R. M. (2019). Evolution of a clade of *Acinetobacter baumannii* global clone 1, lineage 1 via acquisition of carbapenem- and aminoglycoside-resistance genes and dispersion of ISAba1. *Microbial Genomics, 5*(1), e000242.

Hamidian, M., Holt, K., & Hall, R. (2015). Genomic resistance island AGI1 carrying a complex class 1 integron in a multiply antibiotic-resistant ST25 *Acinetobacter baumannii* isolate. *J Antimicrob Chemother, 70*, 2519-2523.

Hamidian, M., Holt, K. E., & Hall, R. M. (2015). Genomic resistance island AGI1 carrying a complex class 1 integron in a multiply antibiotic-resistant ST25 Acinetobacter baumannii isolate. *Journal of Antimicrobial Chemotherapy, 70*(9), 2519-2523. https://doi.org/10.1093/jac/dkv137

Hamidian, M., & Nigro, S. J. (2019, Oct). Emergence, molecular mechanisms and global spread of carbapenem-resistant Acinetobacter baumannii. *Microb Genom, 5*(10). https://doi.org/10.1099/mgen.0.000306

Harrison, L. H., Granoff, D. M., & Pollard, A. J. (2018). 38 - Meningococcal Capsular Group A, C, W, and Y Conjugate Vaccines. In S. A. Plotkin, W. A. Orenstein, P. A. Offit, & K. M. Edwards (Eds.), *Plotkin's Vaccines (Seventh Edition)* (pp. 619-643.e611). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-323-35761-6.00038-9

Holt, K., Kenyon, J., Hamidian, M., Schultz, M., Pickard, D., Dougan, G., & Hall, R. (2016). Five decades of genome evolution in the globally distributed, extensively antibiotic resistant *Acinetobacter baumannii* global clone 1. *Microbial Genomics, 2*(2), e000052.

Howard, A., O'Donoghue, M., Feeney, A., & Sleator, R. D. (2012, May 1). Acinetobacter baumannii: an emerging opportunistic pathogen. *Virulence, 3*(3), 243-250. https://doi.org/10.4161/viru.19700

Hu, D., Liu, B., Dijkshoorn, L., Wang, L., & Reeves, P. R. (2013). Diversity in the major polysaccharide antigen of *Acinetobacter baumannii* assessed by DNA sequencing, and development of a molecular serotyping scheme. *PLOS ONE, 8*(7), e70329.

Hua, M., Liu, J., Du, P., Liu, X., Li, M., Wang, H., Chen, C., Xu, X., Jiang, Y., Wang, Y., Zeng, H., & Li, A. (2021, 2021/01/01). The novel outer membrane protein from OprD/Occ family is associated with hypervirulence of carbapenem resistant Acinetobacter baumannii ST2/KL22. *Virulence, 12*(1), 1-11. https://doi.org/10.1080/21505594.2020.1856560

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., & Scheutz, F. (2015, Aug). Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol, 53*(8), 2410-2426. https://doi.org/10.1128/jcm.00008-15

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 Res, 3*, 124. https://doi.org/10.12688/wellcomeopenres.14826.1

Joly-Guillou, M. L. (2005, 2005/11/01/). Clinical impact and pathogenicity of Acinetobacter. *Clinical Microbiology and Infection, 11*(11), 868-873. https://doi.org/https://doi.org/10.1111/j.1469-0691.2005.01227.x

Jones, C. L., Clancy, M., Honnold, C., Singh, S., Snesrud, E., Onmus-Leone, F., McGann, P., Ong, A. C., Kwak, Y., Waterman, P., Zurawski, D. V., Clifford, R. J., & Lesho, E. (2015). Fatal Outbreak of an Emerging Clone of Extensively Drug-Resistant

Acinetobacter baumannii With Enhanced Virulence. *Clinical Infectious Diseases, 61*(2), 145-154. https://doi.org/10.1093/cid/civ225

Karah, N., Giske, C. G., Sundsfjord, A., & Samuelsen, Ø. (2011, Dec). A diversity of OXA-carbapenemases and class 1 integrons among carbapenem-resistant Acinetobacter baumannii clinical isolates from Sweden belonging to different international clonal lineages. *Microb Drug Resist, 17*(4), 545-549. https://doi.org/10.1089/mdr.2011.0089

Kasimova, A., Shneider, M., Arbatsky, N., Popova, A., Shashkov, A., Miroshnikov, K., Balaji, V., Biswas, I., & Knirel, Y. (2017). Structure and gene cluster of the K93 capsular polysaccharide of *Acinetobacter baumannii* B11911 containing 5-N-Acetyl-7-N-[(R)-3-hydroxybutanoyl]pseudaminic acid. *Biochemistry (Moscow), 82*(4), 483-489.

Kenyon, J., & Hall, R. (2013). Variation in the complex carbohydrate biosynthesis loci of *Acinetobacter baumannii* genomes. *PLOS ONE, 8*(4), e62160.

Kenyon, J., Hall, R., & De Castro, C. (2015). Structural determination of the K14 capsular polysaccharide from an ST25 *Acinetobacter baumannii* isolate, D46. *Carbohydr Res, 417*, 52-56.

Kenyon, J., Holt, K., Pickard, D., Dougan, G., & Hall, R. (2014). Insertions in the OCL1 locus of *Acinetobacter baumannii* lead to shortened lipooligosaccharides. *Research in Microbiology, 165*(6), 472-475. https://doi.org/10.1016/j.resmic.2014.05.034

Kenyon, J., Kasimova, A., Notaro, A., Arbatsky, N., Speciale, I., Shashkov, A., De Castro, C., Hall, R., & Knirel, Y. (2017). *Acinetobacter baumannii* K13 and K73 capsular polysaccharides differ only in K-unit side branches of novel non-2-ulosonic acids: di-N-acetylated forms of either acinetaminic acid or 8-epiacinetaminic acid. *Carbohydr Res, 452*, 149-155.

Kenyon, J., Marzaioli, A., De Castro, C., & Hall, R. (2015). 5,7-Di-*N*-acetylacinetaminic acid - a novel non-2-ulosonic acid found in the capsule of an *Acinetobacter baumannii* isolate. *Glycobiology, 25*(6), 644-654.

Kenyon, J., Marzaioli, A., Hall, R., & De Castro, C. (2014). Structure of the K2 capsule associated with the KL2 gene cluster of *Acinetobacter baumannii*. *Glycobiology, 24*(6), 554-563.

Kenyon, J., Marzaioli, A., Hall, R., & De Castro, C. (2015). Structure of the K12 capsule containing 5,7-di-*N*-acetylacinetaminic acid from A*cinetobacter baumannii* isolate D36. *Glycobiology, 25*(8), 881-887. https://doi.org/10.1093/glycob/cwv028

Kenyon, J., Nigro, S., & Hall, R. (2014). Variation in the OC locus of *Acinetobacter baumannii* genomes predicts extensive structural diversity in the lipoligosaccharide. *PLOS ONE, 9*(9), e107833.

Kenyon, J., Notaro, A., Hsu, L. Y., De Castro, C., & Hall, R. (2017). 5,7-Di-N-acetyl-8-epiacinetaminic acid: A new non-2-ulosonic acid found in the K73 capsule produced by an *Acinetobacter baumannii* isolate from Singapore. *Scientific Reports, 7*, 11357.

Kenyon, J., Shashkov, A., Senchenkova, S., Shneider, M., Liu, B., Popova, A., Arbatsky, N., Miroshnikov, K., Wang, L., Knirel, Y., & Hall, R. (2017). *Acinetobacter baumannii* K11 and K83 capsular polysaccharides have the same 6-deoxy-l-talose-containing pentasaccharide K units but different linkages between the K units. *International Journal of Biological Macromolecules, 103*, 648-655.

Kenyon, J. J., Arbatsky, N. P., Sweeney, E. L., Shashkov, A. S., Shneider, M. M., Popova, A. V., Hall, R. M., & Knirel, Y. A. (2019). Production of the K16 capsular polysaccharide by *Acinetobacter baumannii* ST25 isolate D4 involves a novel glycosyltransferase encoded in the KL16 gene cluster. *International Journal of Biological Macromolecules, 128*, 101-106.

Knirel, Y., Shevelev, S., & Perepelov, A. (2011). Higher aldulosonic acids: components of bacterial glycans. *Mendeleev Commun., 21*, 173-182.

Knirel, Y. A., Shneider, M. M., Popova, A. V., Kasimova, A. A., Senchenkova, S. N., Shashkov, A. S., & Chizhov, A. O. (2020, 2020/05/01). Mechanisms of Acinetobacter baumannii Capsular Polysaccharide Cleavage by Phage Depolymerases. *Biochemistry (Moscow), 85*(5), 567-574. https://doi.org/10.1134/S0006297920050053

Kongthai, P., Thummeepak, R., Leungtongkam, U., Pooarlai, R., Kitti, T., Thanwisai, A., Chantratita, N., Millard, A. D., & Sitthisak, S. (2020, Jul 23). Insight into Molecular Epidemiology, Antimicrobial Resistance, and Virulence Genes of Extensively Drug-Resistant Acinetobacter baumannii in Thailand. *Microb Drug Resist*. https://doi.org/10.1089/mdr.2020.0064

Landersjö, C., Weintraub, A., & Widmalm, G. J. C. r. (1996). Structure determination of the O-antigen polysaccharide from the enteroinvasive Escherichia coli (EIEC) O143 by component analysis and NMR spectroscopy. *291*, 209-216.

Lee, I. M., Yang, F.-L., Chen, T.-L., Liao, K.-S., Ren, C.-T., Lin, N.-T., Chang, Y.-P., Wu, C.-Y., & Wu, S.-H. (2018, 2018/07/18). Pseudaminic Acid on Exopolysaccharide of Acinetobacter baumannii Plays a Critical Role in Phage-Assisted Preparation of Glycoconjugate Vaccine with High Antigenicity. *Journal of the American Chemical Society, 140*(28), 8639-8643. https://doi.org/10.1021/jacs.8b04078

Lee, Y., Kim, C.-K., Lee, H., Jeong, S. H., Yong, D., & Lee, K. (2011). A Novel Insertion Sequence, ISAba10, Inserted into ISAba1 Adjacent to the bla _{OXA-23} Gene and Disrupting the Outer Membrane Protein Gene carO in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy, 55*(1), 361-363. https://doi.org/10.1128/aac.01672-09

Lee, Y. J., Kubota, A., Ishiwata, A., & Ito, Y. (2011, 2011/01/19/). Synthesis of pseudaminic acid, a unique nonulopyranoside derived from pathogenic bacteria through 6-deoxy-AltdiNAc. *Tetrahedron Letters, 52*(3), 418-421. https://doi.org/https://doi.org/10.1016/j.tetlet.2010.11.078

Leungtongkam, U., Thummeepak, R., Wongprachan, S., Thongsuk, P., Kitti, T., Ketwong, K., Runcharoen, C., Chantratita, N., & Sitthisak, S. (2018, Jan/Feb). Dissemination of bla(OXA-23), bla(OXA-24), bla(OXA-58), and bla(NDM-1) Genes of Acinetobacter baumannii Isolates from Four Tertiary Hospitals in Thailand. *Microb Drug Resist, 24*(1), 55-62. https://doi.org/10.1089/mdr.2016.0248

Lin, M. F., & Lan, C. Y. (2014, Dec 16). Antimicrobial resistance in Acinetobacter baumannii: From bench to bedside. *World J Clin Cases, 2*(12), 787-814. https://doi.org/10.12998/wjcc.v2.i12.787

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P., & Henrissat, B. (2014). The Carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res, 42*, D490-D495.

Luke, N., Sauberan, S., Russo, T., Beanan, J., Olson, R., Loehfelm, T., Cox, A., St. Michael, F., Vinogradov, E., & Campagnari, A. (2010). Identification and characterization of a glycosyltransferase involved in *Acinetobacter baumannii* lipopolysaccharide core biosynthesis. *Infect. Immun., 78*(5), 2017.

Madsen, J. S., Burmølle, M., Hansen, L. H., & Sørensen, S. J. (2012). The interconnection between biofilm formation and horizontal gene transfer. *FEMS Immunology & Medical Microbiology, 65*(2), 183-195.

McConnell, M. J., Actis, L., & Pachón, J. (2013, Mar). Acinetobacter baumannii: human infections, factors contributing to pathogenesis and animal models. *FEMS Microbiol Rev, 37*(2), 130-155. https://doi.org/10.1111/j.1574-6976.2012.00344.x

McQueary, C., & Actis, L. (2011). *Acinetobacter baumannii* biofilms: variations among strains and correlations with other cell properties. *J. Microbiol., 49*(2), 243-250.

McQueary, C., Kirkup, B., Si, Y., Barlow, M., Actis, L., Craft, D., & Zurawsk, D. (2012). Extracellular stress and lipopolysaccharide modulate Acinetobacter baumannii surface-associated motility. *The Journal of Microbiology, 50*(3), 434-443.

Meumann, E., Anstey, N., Currie, B., Piera, K., Kenyon, J., Hall, R., Davis, J., & Sarovich, D. (2019). Genomic epidemiology of severe community-onset *Acinetobacter baumannii* infection. *Microbial Genomics, 5*. https://doi.org/10.1099/mgen.0.000258

Neelamegham, S., Aoki-Kinoshita, K., Bolton, E., Frank, M., Lisacek, F., Lütteke, T., O'Boyle, N., Packer, N. H., Stanley, P., Toukach, P., Varki, A., & Woods, R. J. (2019, Aug 20). Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology, 29*(9), 620-624. https://doi.org/10.1093/glycob/cwz045

Nigro, S., Holt, K., Pickard, D., & Hall, R. (2015). Carbapenem and amikacin resistance on a large conjugative Acinetobacter baumannii plasmid *J Antimicrob Chemother, 70*(4), 1259-1261.

Nigro, S. J., & Hall, R. M. (2018, Dec 1). Does the intrinsic oxaAb (blaOXA-51-like) gene of Acinetobacter baumannii confer resistance to carbapenems when activated by ISAba1? *J Antimicrob Chemother, 73*(12), 3518-3520. https://doi.org/10.1093/jac/dky334

O'Shea, M. K. (2012, May). Acinetobacter in modern warfare. *Int J Antimicrob Agents, 39*(5), 363-375. https://doi.org/10.1016/j.ijantimicag.2012.01.018

Oliveira, H., Costa, A., Ferreira, A., Konstantinides, N., Santos, S., Boon, M., Noben, J., Lavigne, R., & Azeredo, J. (2019). Functional analysis and antivirulence properties of a new depolymerase from a Myovirus that infects *Acinetobacter baumannii* capsule K45. *J Virol, 93*(4), e01163-01118.

World Health Organisation (WHO). (2017). Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. https://www.who.int/medicines/publications/WHO-PPL-Short_Summary_25Feb-ET_NM_WHO.pdf

---

Perepelov, A., Wang, Q., Liu, B., Senchenkova, S., Feng, L., Shashkov, A., Wang, L., & Knirel, Y. (2009). Structure of the O-polysaccharide of *Escherichia coli* O61, another *E. coli* O-antigen antigen that contains 5,7-diacetamido-3,5,7,9-tetradeoxy-L-*glycero*-D-*galacto*-non-2-ulosonic (Di-*N*-acetyl-8-epilegionaminic) acid. *Journal of Carbohydrate Chemistry, 28*(7-8), 463-472.

Perilli, M., Felici, A., Oratore, A., Cornaglia, G., Bonfiglio, G., Rossolini, G. M., & Amicosante, G. (1996). Characterization of the chromosomal cephalosporinases produced by Acinetobacter lwoffii and Acinetobacter baumannii clinical isolates. *Antimicrobial Agents and Chemotherapy, 40*(3), 715-719. https://doi.org/10.1128/aac.40.3.715

Preston, A., Mandrell, R. E., Gibson, B. W., & Apicella, M. A. (1996). The lipooligosaccharides of pathogenic gram-negative bacteria. *Crit Rev Microbiol, 22*(3), 139-180. https://doi.org/10.3109/10408419609106458

Principi, N., Silvestri, E., & Esposito, S. (2019). Advantages and Limitations of Bacteriophages for the Treatment of Bacterial Infections. *Front Pharmacol, 10*, 513. https://doi.org/10.3389/fphar.2019.00513

Qureshi, Z. A., Hittle, L. E., O'Hara, J. A., Rivera, J. I., Syed, A., Shields, R. K., Pasculle, A. W., Ernst, R. K., & Doi, Y. (2015). Colistin-Resistant Acinetobacter baumannii: Beyond Carbapenem Resistance. *Clinical Infectious Diseases, 60*(9), 1295-1303. https://doi.org/10.1093/cid/civ048

Raetz, C. R., & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annu Rev Biochem, 71*, 635-700. https://doi.org/10.1146/annurev.biochem.71.110601.135414

Rice, L. B. (2008). Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. *The Journal of Infectious Diseases, 197*(8), 1079-1081. https://doi.org/10.1086/533452

Rice, L. B. (2010, Nov). Progress and challenges in implementing the research on ESKAPE pathogens. *Infect Control Hosp Epidemiol, 31 Suppl 1*, S7-10. https://doi.org/10.1086/655995

Roca, I., Espinal, P., Vila-Farrés, X., & Vila, J. (2012). The Acinetobacter baumannii Oxymoron: Commensal Hospital Dweller Turned Pan-Drug-Resistant Menace. *Front Microbiol, 3*, 148. https://doi.org/10.3389/fmicb.2012.00148

Rouphael, N. G., Zimmer, S. M., & Stephens, D. S. (2009). Chapter 53 - Neisseria meningitidis. In A. D. T. Barrett & L. R. Stanberry (Eds.), *Vaccines for Biodefense and Emerging and Neglected Diseases* (pp. 1061-1079). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-369408-9.00053-6

Russo, T., Beanan, J., Olson, R., MacDonald, U., Cox, A., St. Michael, F., Vinogradov, E., Spellberg, B., Luke-Marshall, N., & Campagnari, A. (2013). The K1 capsular polysaccharide from *Acinetobacter baumannii* is a potential therapeutic target via passive immunization. *Infect. Immun., 81*(3), 915-922. https://doi.org/doi: 10.1128/IAI.01184-12

Russo, T., Luke, N., Beanan, J., Olson, R., Sauberan, S., MacDonald, U., Schultz, L., Umland, T., & Campagnari, A. (2010). The K1 capsular polysaccharide of *Acinetobacter baumannii* strain 307-0294 is a major virulence factor. *Infect. Immun., 78*(9), 3993-4000.

Sahl, J. W., Del Franco, M., Pournaras, S., Colman, R. E., Karah, N., Dijkshoorn, L., & Zarrilli, R. (2015, Oct 14). Phylogenetic and genomic diversity in isolates from the globally distributed Acinetobacter baumannii ST25 lineage. *Sci Rep, 5*, 15188. https://doi.org/10.1038/srep15188

Schirm, M., Soo, E., Aubry, A., Austin, J., Thibault, P., & Logan, S. (2003). Structural, genetic and functional characterization of the flagellin glycosylation process in *Helicobacter pylori. Molecular microbiology, 48*(6), 1579-1592.

Schoenhofen, I., McNally, D., Brisson, J., & Logan, S. (2006). Elucidation of the CMP-pseudaminic acid pathway in *Helicobacter pylori*: synthesis from UDP-*N*-acetylglucosamine by a single enzymatic reaction. *Glycobiology, 16*(9), 8C-14C.

Schoenhofen, I., Vinogradov, E. V., Whitfield, D., Brisson, J., & Logan, S. (2009). The CMP-legionaminic acid pathway in *Campylobacter*: Biosynthesis involving novel GDP-linked precursors. *Glycobiology, 19*(7), 715-725.

Schultz, M., Thanh, D., Hoan, N., Wick, R., Ingle, D., Hawkey, J., Edwards, D., Kenyon, J., Lan, N., Campbell, J., Thwaites, G., Nhu, N., Hall, R., Fournier-Level, A., Baker, S., & Holt, K. (2016). Repeated local emergence of carbapenem-resistant Acinetobacter baumannii in a single hospital ward. *Microbial Genomics, 2*(3), e000050.

Scott, P., Deye, G., Srinivasan, A., Murray, C., Moran, K., Hulten, E., Fishbain, J., Craft, D., Riddell, S., Lindler, L., Mancuso, J., Milstrey, E., Bautista, C. T., Patel, J., Ewell, A., Hamilton, T., Gaddy, C., Tenney, M., Christopher, G., Petersen, K., Endy, T., & Petruccelli, B. (2007, Jun 15). An outbreak of multidrug-resistant Acinetobacter

baumannii-calcoaceticus complex infection in the US military health care system associated with military operations in Iraq. *Clin Infect Dis, 44*(12), 1577-1584. https://doi.org/10.1086/518170

Senchenkova, S., Popova, A., Shashkov, A., Shneider, M., Mei, Z., Arbatsky, N., Liu, B., Miroshinikov, K., Volozhantsev, N., & Knirel, Y. (2015). Structure of a new pseudaminic acid-containing capsular polysaccharide of *Acinetobacter baumannii* LUH5550 having the KL42 capsule biosynthesis locus. *Carbohydr Res, 407*, 154-157.

Shashkov, A., Kenyon, J., Arbatsky, N., Shneider, M., Popova, A., Knirel, Y., & Hall, R. (2018). Genetics of biosynthesis and structure of the K53 capsular polysaccharide of *Acinetobacter baumannii* D23 made up of a disaccharide K unit. *Microbiology, 164*, 1289-1292.

Shashkov, A., Kenyon, J., Senchenkova, S., Shneider, M., Popova, A., Arbatsky, N., Miroshinikov, K., Volozhantsev, N., Hall, R., & Knirel, Y. (2016). *Acinetobacter baumannii* K27 and K44 capsular polysaccharides have the same K unit but different structures due to the presence of distinct *wzy* genes in otherwise closely related K gene clusters. *Glycobiology, 26*(5), 501-508.

Shashkov, A. S., Cahill, S. M., Arbatsky, N. P., Westacott, A. C., Kasimova, A. A., Shneider, M. M., Popova, A. V., Shagin, D. A., Shelenkov, A. A., Mikhailova, Y. V., Yanushevich, Y. G., Edelstein, M. V., Kenyon, J. J., & Knirel, Y. A. (2019, Oct 1). Acinetobacter baumannii K116 capsular polysaccharide structure is a hybrid of the K14 and revised K37 structures. *Carbohydr Res, 484*, 107774. https://doi.org/10.1016/j.carres.2019.107774

Shashkov, A. S., Senchenkova, S. N., Popova, A. V., Mei, Z., Shneider, M. M., Liu, B., Miroshnikov, K. A., Volozhantsev, N. V., & Knirel, Y. (2015). Revised structure of the capsular polysaccharide of *Acinetobacter baumannii* LUH5533 (serogroup O1) containing di-*N*-acetyllegionaminic acid. *Russian Chemical Bulletin, 64*(5), 1196–1199.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006, Jan 1). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res, 34*(Database issue), D32-36. https://doi.org/10.1093/nar/gkj014

Singh, H., Thangaraj, P., & Chakrabarti, A. (2013, Nov). Acinetobacter baumannii: A Brief Account of Mechanisms of Multidrug Resistance and Current and Future Therapeutic Management. *J Clin Diagn Res, 7*(11), 2602-2605. https://doi.org/10.7860/jcdr/2013/6337.3626

Soon, R., Nation, R., Harper, M., Adler, B., Boyce, J., Tan, C.-H., Li, J., & Larson, I. (2011). Effect of colistin exposure and growth phase on the surface properties of live *Acinetobacter baumannii* cells examined by atomic force microscopy. *International Journal of Antimicrobial Agents, 38*(493-501).

Spellberg, B., & Bonomo, R. A. (2014, May). The deadly impact of extreme drug resistance in Acinetobacter baumannii. *Crit Care Med, 42*(5), 1289-1291. https://doi.org/10.1097/ccm.0000000000000181

Stenutz, R., Weintraub, A., & Widmalm, G. (2006). The structures of Escherichia coli O-polysaccharide antigens. *FEMS microbiology reviews, 30*(3), 382-403.

Stietz, M. S., Ramírez, M. S., Vilacoba, E., Merkier, A. K., Limansky, A. S., Centrón, D., & Catalano, M. (2013, Mar). Acinetobacter baumannii extensively drug resistant lineages in Buenos Aires hospitals differ from the international clones I-III. *Infect Genet Evol, 14*, 294-301. https://doi.org/10.1016/j.meegid.2012.12.020

Thi Khanh Nhu, N., Riordan, D. W., Do Hoang Nhu, T., Thanh, D. P., Thwaites, G., Huong Lan, N. P., Wren, B. W., Baker, S., & Stabler, R. A. (2016, Jun 22). The induction and identification of novel Colistin resistance mutations in Acinetobacter baumannii and their implications. *Sci Rep, 6*, 28291. https://doi.org/10.1038/srep28291

Thrane, S. W., Taylor, V. L., Lund, O., Lam, J. S., & Jelsbak, L. (2016, Jul). Application of Whole-Genome Sequencing Data for O-Specific Antigen Analysis and In Silico Serotyping of Pseudomonas aeruginosa Isolates. *J Clin Microbiol, 54*(7), 1782-1788. https://doi.org/10.1128/jcm.00349-16

Tipton, K. A., Chin, C., Farokhyfar, M., Weiss, D. S., & Rather, P. N. (2018). Role of capsule in resistance to disinfectants, host antimicrobials, and desiccation in *Acinetobacter baumannii. Antimicrob. Agents Chemother.* https://doi.org/10.1128/AAC.01188-18

Traub, W. (1989). *Acinetobacter baumannii* serotyping for deliniation of outbreaks of nosocomial cross-infection. *Journal of Clinical Microbiology, 27*(12), 2713-2716.

Traub, W. (1999). Examination of polyclonal rabbit immune sera against serovars of *Acinetobacter baumannii* and Genospecies 3 for cross-reactions with reference strains of other named/unnamed Genospecies of *Acinetobacter. Zent.bl. Bakteriol., 289*, 787-795.

Traub, W., & Bauer, D. (2000). Surveillance of nosocomial cross-infections due to three *Acinetobacter* genospecies (*Acinetobacter baumannii*, genospecies 3 and genospecies 13) during a 10-year observation period: serotyping, macrorestriction analysis of genomic DNA and antibiotic susceptibilities. *Chemotherapy, 46*, 282-292.

Varki, A., Cummings, R. D., Aebi, M., Packer, N. H., Seeberger, P. H., Esko, J. D., Stanley, P., Hart, G., Darvill, A., Kinoshita, T., Prestegard, J. J., Schnaar, R. L., Freeze, H. H., Marth, J. D., Bertozzi, C. R., Etzler, M. E., Frank, M., Vliegenthart, J. F., Lütteke, T., Perez, S., Bolton, E., Rudd, P., Paulson, J., Kanehisa, M., Toukach, P., Aoki-Kinoshita, K. F., Dell, A., Narimatsu, H., York, W., Taniguchi, N., & Kornfeld, S. (2015, Dec). Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology, 25*(12), 1323-1324. https://doi.org/10.1093/glycob/cwv091

Villegas, M. V., Kattan, J. N., Correa, A., Lolans, K., Guzman, A. M., Woodford, N., Livermore, D., & Quinn, J. P. (2007, Jun). Dissemination of Acinetobacter baumannii clones with OXA-23 Carbapenemase in Colombian hospitals. *Antimicrob Agents Chemother, 51*(6), 2001-2004. https://doi.org/10.1128/aac.00226-07

Vinogradov, E., MacLean, L., Xu, H. H., & Chen, W. (2014). The structure of the polysaccharide isolated from *Acinetobacter baumannii* strain LAC-4. *Carbohydr Res, 390*, 42-45.

Wang-Lin, S. X., Olson, R., Beanan, J. M., MacDonald, U., Balthasar, J. P., & Russo, T. A. (2017, Dec). The Capsular Polysaccharide of Acinetobacter baumannii Is an Obstacle for Therapeutic Passive Immunization Strategies. *Infect Immun, 85*(12). https://doi.org/10.1128/iai.00591-17

Wick, R., Heinz, E., Holt, K., & Wyres, K. (2018). Kaptive Web: User-Friendly Capsule and Lipopolysaccharide Serotype Prediction for Klebsiella Genomes. *J Clin Microbiol, 56*(6), e00197-00118.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wilharm, G., Piesker, J., Laue, M., & Skiebe, E. (2013, Sep). DNA uptake by the nosocomial pathogen Acinetobacter baumannii occurs during movement along wet surfaces. *J Bacteriol, 195*(18), 4146-4153. https://doi.org/10.1128/jb.00754-13

Wright, M., Haft, D., Harkins, D., Perez, F., Hujer, K., Bajaksouzian, S., Benard, M., Jacobs, M., Bonomo, R., & Adams, M. (2014). New insights into dissemination and variation of the health care-associated pathogen *Acinetobacter baumannii* from genomic analysis *mBio, 5*(1), e00963-00913

Wyres, K., Wick, R., Gorrie, C., Jenney, A., Follador, R., Thomson, N., & Holt, K. (2016). Identification of *Klebsiella* capsule synthesis loci from whole genome data *Microbial Genomics, 2*. https://doi.org/10.1099/mgen.0.000102

Wyres, K. L., Cahill, S. M., Holt, K. E., Hall, R. M., & Kenyon, J. J. (2020). Identification of Acinetobacter baumannii loci for capsular polysaccharide (KL) and lipooligosaccharide outer core (OCL) synthesis in genome assemblies using curated reference databases compatible with Kaptive. *Microbial Genomics, 6*(3). https://doi.org/https://doi.org/10.1099/mgen.0.000339

Xie, R., Zhang, X. D., Zhao, Q., Peng, B., & Zheng, J. (2018, 2018/12/01). Analysis of global prevalence of antibiotic resistance in Acinetobacter baumannii infections disclosed a faster increase in OECD countries. *Emerging Microbes & Infections, 7*(1), 1-10. https://doi.org/10.1038/s41426-018-0038-9

Yachison, C. A., Yoshida, C., Robertson, J., Nash, J. H. E., Kruczkiewicz, P., Taboada, E. N., Walker, M., Reimer, A., Christianson, S., Nichani, A., & Nadon, C. (2017). The Validation and Implications of Using Whole Genome Sequencing as a Replacement for Traditional Serotyping for a National Salmonella Reference Laboratory. *Front Microbiol, 8*, 1044. https://doi.org/10.3389/fmicb.2017.01044

Yang, F.-L., Lou, T.-C., Kuo, S.-C., Wu, W.-L., Chern, J., Lee, Y.-T., Chen, S.-T., Zou, W., Lin, N.-T., & Wu, S.-H. (2017). A medically relevant capsular polysaccharide in Acinetobacter baumannii is a potential vaccine candidate. *Vaccine, 35*(10), 1440-1447. https://doi.org/https://doi.org/10.1016/j.vaccine.2017.01.060

Zarrilli, R., Giannouli, M., Rocco, F., Loman, N., Haines, A., Constantinidou, C., Pallen, M., Triassi, M., & Di Nocera, P. (2011). Genome sequences of three Acinetobacter baumannii strains assigned to the multilocus sequence typing genotypes ST2, ST25, and ST78. *J Bacteriol. , 193*(9), 2359-2360.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000, Feb-Apr). A greedy algorithm for aligning DNA sequences. *J Comput Biol, 7*(1-2), 203-214. https://doi.org/10.1089/10665270050081478

Zhou, H., Yao, Y., Zhu, B., Ren, D., Yang, Q., Fu, Y., Yu, Y., & Zhou, J. (2019). Risk factors for acquisition and mortality of multidrug-resistant Acinetobacter baumannii bacteremia: A retrospective study from a Chinese hospital. *Medicine, 98*(13), e14937. https://doi.org/10.1097/md.0000000000014937

Zunk, M., & Kiefel, M. (2014). The occurence and biological significance of the a-keto-sugars pseudaminic acid and legionaminic acid within pathogenic bacteria. *RSC Advances, 4*, 3413-3421.

Chapter 6: Appendices

Appendix 1. Summary of KL and OCL type for 61 isolates. Country data and year of isolate collection has been included to highlight the extent of locus variability across geographically confined strains.

Strain	Assembly	Year	Country	KL	OCL
741019	ASM100777v1	2011	Argentina	KL9	OCL10a
D46	7468_2_59	2010	Australia	KL14	OCL6
D4	7521_8_34	2006	Australia	KL16	OCL5
MC104	ASM358425v1	2016	Bolivia	KL22	OCL5
MC102	ASM358427v1	2016	Bolivia	KL22	OCL5
MC100	ASM358431v1	2016	Bolivia	KL22	OCL5
MC89	ASM358442v1	2016	Bolivia	KL22	OCL5
MC29	ASM358448v1	2016	Bolivia	KL33	OCL5
MC93	ASM358451v1	2016	Bolivia	KL22	OCL5
MC63	ASM359568v1	2016	Bolivia	KL33	OCL5
MC51	ASM359581v1	2016	Bolivia	KL37	OCL10
MC39	ASM359586v1	2016	Bolivia	KL37	OCL10
MC32	ASM359593v1	2016	Bolivia	KL22	OCL5
MC57	ASM359612v1	2016	Bolivia	KL37	OCL10a
MC14	ASM359623v1	2016	Bolivia	KL33	OCL5
107m	ABIBUN_1	N/A	Colombia	KL32a	OCL5
NIPH146	Acin_baum_NIPH_146_V1	1993	Czech Republic	KL37	OCL6
161/07	ASM100781v1	2007	Germany	KL130	OCL5
BAuABod-3	ASM257380v1	2015	Germany	KL14	OCL6
IHIT38008	ASM415359v1	2018	Germany	KL14	OCL6
4390	ASM100776v1	2003	Greece	KL37	OCL7
HEU3	ASM292785v1	2016	Honduras	KL134	OCL6
MCR10179	ASM292796v1	2015	Honduras	KL132	OCL10
CI86	ASM51657v2	2005	Iraq	KL12	OCL5
CI79	ASM51663v2	2005	Iraq	KL12	OCL5
4190	ASM18971v2	2009	Italy	KL27	OCL5
OCU_Ac2	ASM356944v1	2014	Japan	KL14	OCL15
7804	ASM343138v1	2006	Mexico	KL7	OCL7
AB_2008-15-69	ASM30133v1	N/A	N/A	KL14	OCL5
PR388	ASM213810v1	N/A	N/A	KL133	OCL5
ARLG1317	ASM214397v1	N/A	N/A	KL12	OCL5
AR_0088	ASM300603v1	N/A	N/A	KL14	OCL6
R348	ASM395535v1	N/A	N/A	KL14	OCL5
LUH_7841	ASM100768v1	2002	Netherlands	KL95	OCL7
RUH1486	ASM100770v1	1985	Netherlands	KL14	OCL6
LUH_6220	ASM100774v1	2000	Netherlands	KL37	OCL7
HWBA8	ASM208278v1	2013	South Korea	KL14	OCL5
4300STDY7045700	24276_2_166	2016	Thailand	KL116	OCL5
4300STDY7045786	24276_2_251	2016	Thailand	KL14	OCL6

Appendix 1. *Summary of KL and OCL type for 61 isolates. Country data and year of isolate collection has been included to highlight the extent of locus variability across geographically confined strains.*

4300STDY7045802	24276_2_267	2016	Thailand	KL14	OCL6
4300STDY7045824	24276_2_289	2016	Thailand	KL14	OCL6
4300STDY7045840	24276_2_305	2016	Thailand	KL14	OCL6
4300STDY7045893	24276_3_50	2016	Thailand	KL14	OCL6
NM3	ASM100772v1	2008	United Arab Emirates	KL14	OCL6
AB5256	ASM24170v2	2009	USA	KL14	OCL6
OIFC143	AcbauOIFC143v1.0	2003	USA	KL23	OCL5
1429530	ASM58113v1	2014	USA	KL14	OCL6
ABBL018	ASM143254v1	2005	USA	KL14	OCL5
AB2828	ASM161199v1	2006	USA	KL12	OCL5
AB3638	ASM161209v1	2007	USA	KL13	OCL5
AB3806	ASM161210v1	2007	USA	KL13	OCL5
UV_1036	SAMEA1569559	2003	Vietnam	KL37	OCL5
UV_964	SAMEA1569507	2003	Vietnam	KL52	OCL5
UV_973	SAMEA1569410	2003	Vietnam	KL52	OCL5
238_an	SAMEA1569453	2005	Vietnam	KL14	OCL5
259_an	SAMEA1569518	2005	Vietnam	KL37	OCL5
266_an	SAMEA1569452	2005	Vietnam	KL37	OCL5
276_ax	SAMEA1569532	2005	Vietnam	KL37	OCL5
295_c	SAMEA1569392	2005	Vietnam	KL37	OCL5
338_an	ERR263723	2006	Vietnam	KL37	OCL5
338_ax	ERR263724	2006	Vietnam	KL37	OCL5

Appendix 2. *Kaptive KL analysis output for the original 82 ST25 isolates including key output details. Isolates which were excluded from the final genome pool are highlighted in grey*

Assembly	Strain	Best match	Match confidence	Problems	Coverage	Identity	Length discrepancy	Expected genes in locus	Other genes in locus, details
8346_4#10	UV_1036	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
8346_4#80	259_an	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
8346_4#83	266_an	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
8346_4#86	276_ax	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
8346_4#94	295_c	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
9262_1#33	338_an	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
9262_1#34	338_ax	KL37	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
AcbauOIFC143v1.0	OIFC143	KL23	Perfect		100.00%	100.00%	n/a	21 / 21 (100%)	
ASM30133v1	AB_2008-15-69	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM100772v1	NM3	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM143254v1	ABBL018	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM208278v1	HWBA8	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM300603v1	AR_0088	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM395535v1	R348	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM415359v1	IHIT38008	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
7468_2_59	D46	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
7521_8_34	D4	KL16	Perfect		100.00%	100.00%	n/a	20 / 20 (100%)	
24276_2_251	4300STDY7045786	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
24276_2_267	4300STDY7045802	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
24276_2_289	4300STDY7045824	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
24276_2_305	4300STDY7045840	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
24276_3_50	4300STDY7045893	KL14	Perfect		100.00%	100.00%	n/a	17 / 17 (100%)	
8346_4#3	UV_964	KL52	Very high	*	100.00%	97.28%	-2 bp	17 / 17 (100%)	
8346_4#6	UV_973	KL52	Very high	*	100.00%	97.28%	-2 bp	17 / 17 (100%)	
8346_4#76	238_an	KL14	Very high		100.00%	99.96%	n/a	17 / 17 (100%)	
Acin_baum_NIPH_146_V1	NIPH146	KL37	Very high	*	100.00%	97.15%	n/a	17 / 17 (100%)	

Appendix 2. Kaptive KL analysis output for the original 82 ST25 isolates including key output details. Isolates which were excluded from the final genome pool are highlighted in grey

ABIBUN_1	107m	KL32	Very high		100.00%	97.67%	+889 bp	19 / 19 (100%)	
ASM51657v2	CI86	KL12	Very high	*	100.00%	95.32%	-5 bp	31 / 31 (100%)	
ASM51663v2	CI79	KL12	Very high	*	100.00%	95.32%	-5 bp	31 / 31 (100%)	
ASM58113v1	1429530	KL14	Very high		100.00%	100.00%	n/a	17 / 17 (100%)	
ASM100770v1	RUH1486	KL14	Very high		100.00%	98.98%	n/a	17 / 17 (100%)	
ASM100774v1	LUH_6220	KL37	Very high		100.00%	98.50%	+6 bp	17 / 17 (100%)	
ASM100777v1	741019	KL9	Very high	*	100.00%	96.84%	+3 bp	21 / 21 (100%)	
ASM100781v1	161/07	KL130	Very high		100.00%	100.00%	n/a	24 / 24 (100%)	
ASM161199v1	AB2828	KL12	Very high	*	100.00%	95.32%	-5 bp	31 / 31 (100%)	
ASM161209v1	AB3638	KL13	Very high	*	100.00%	95.30%	-5 bp	31 / 31 (100%)	
ASM161210v1	AB3806	KL13	Very high	*	100.00%	95.30%	-5 bp	31 / 31 (100%)	
ASM214397v1	ARLG1317	KL12	Very high	*	100.00%	95.32%	-5 bp	31 / 31 (100%)	
ASM343138v1	7804	KL7	Very high		100.00%	99.99%	n/a	22 / 22 (100%)	
ASM358448v1	MC29	KL33	Very high		100.00%	98.80%	-4 bp	20 / 20 (100%)	
ASM359623v1	MC14	KL33	Very high		100.00%	98.80%	-4 bp	20 / 20 (100%)	
24276_2_166	4300STDY7045700	KL116	Very high	*	100.00%	96.15%	-1 bp	17 / 17 (100%)	
ASM24170v2	AB5256	KL14	High	-	100.00%	100.00%	-1 bp	16 / 17 (94.1%)	
ASM100768v1	LUH_7841	KL95	Good	+*	100.00%	100.00%	n/a	22 / 22 (100%)	KL80/KL7_*qdtA* (90%)
ASM100776v1	4390	KL37	Good	?	100.00%	98.51%	n/a	17 / 17 (100%)	
ASM257380v1	BAuABod-3	KL14	Good	?-	99.82%	99.14%	n/a	15 / 17 (88.2%)	
ASM292796v1	MCR10179	KL10	Good	-+*	96.91%	94.48%	-510 bp	20 / 23 (87.0%)	KL24/KL7_*fdtE* (98%)
ASM356944v1	OCU_Ac2	KL14	Good	?-	100.00%	100.00%	n/a	16 / 17 (94.1%)	
ASM358425v1	MC104	KL22	Good	?	100.00%	99.93%	n/a	21 / 21 (100%)	
ASM358427v1	MC102	KL22	Good	?-	99.60%	99.92%	n/a	19 / 21 (90.5%)	
ASM358431v1	MC100	KL22	Good	?-	98.82%	99.93%	n/a	20 / 21 (95.2%)	
ASM358438v1	MC78	KL33	Good	?-	100.00%	98.80%	n/a	19 / 20 (95%)	
ASM358442v1	MC89	KL22	Good	?	100.00%	99.93%	n/a	21 / 21 (100%)	
ASM358451v1	MC93	KL22	Good	?	100.00%	99.93%	n/a	21 / 21 (100%)	
ASM359576v1	MC59	KL33	Good	?	100.00%	98.80%	n/a	20 / 20 (100%)	

***Appendix 2.** Kaptive KL analysis output for the original 82 ST25 isolates including key output details. Isolates which were excluded from the final genome pool are highlighted in grey*

ASM359620v1	MC18	KL33	Good	?-	98.31%	98.77%	n/a	18 / 20 (90%)	
ASM359566v1	MC69	KL33	Low	?-	91.97%	98.68%	n/a	18 / 20 (90%)	
ASM359568v1	MC63	KL33	Low	?-	92.51%	98.68%	n/a	17 / 20 (85%)	
ASM359581v1	MC51	KL37	Low	?-+*	93.45%	96.28%	n/a	14 / 17 (82.4%)	KL101/KL14_*gtr8* (81%)
ASM359592v1	MC34	KL33	Low	?-	94.38%	98.72%	n/a	19 / 20 (95%)	
ASM58105v1	984213	KL14	None	?-	86.23%	100.00%	n/a	12 / 17 (70.6%)	
ASM213810v1	PR388	KL124	None	?-+*	86.85%	91.22%	n/a	11 / 18 (61.1%)	KL118/KL5_*wzx*(96%); KL103/KL6_*ptr6*(89%)
ASM292785v1	HEU3	KL67	None	-+*	100.00%	92.50%	-2 bp	16 / 19 (84.2%)	KL14/KL16_*pgt1* (100%); KL10/16_*gtr5* (99%); KL100/KL14_*itrA2* (95%)
ASM358353v1	MC103	KL3	None	?-	71.65%	99.23%	n/a	12 / 20 (60%)	
ASM358356v1	MC95	KL22	None	?-	86.40%	99.91%	n/a	16 / 21 (76.2%)	
ASM358357v1	MC94	KL65	None	?-	51.86%	97.22%	n/a	8 / 20 (40%)	
ASM358360v1	MC87	KL22	None	?-	80.84%	99.91%	n/a	16 / 21 (76.2%)	
ASM358362v1	MC98	KL22	None	?-	88.68%	99.92%	n/a	17 / 21 (81.0%)	
ASM358430v1	MC101	KL3	None	?-	62.86%	99.35%	n/a	11 / 20 (55%)	
ASM358432v1	MC96	KL22	None	?-+*	78.48%	99.90%	n/a	14 / 21 (66.7%)	KL101/KL14_*gtr8* (98%)
ASM358436v1	MC91	KL3	None	?-	77.06%	99.17%	n/a	12 / 20 (60%)	
ASM358437v1	MC77	KL49	None	?-	87.32%	99.19%	n/a	23 / 28 (82.1%)	
ASM358444v1	MC90	KL22	None	?-	79.97%	99.90%	n/a	14 / 21 (66.7%)	
ASM358445v1	MC71	KL49	None	?-	74.63%	99.03%	n/a	17 / 28 (60.7%)	
ASM359567v1	MC64	KL33	None	?-	86.92%	98.60%	n/a	17 / 20 (85%)	
ASM359586v1	MC39	KL37	None	?-*	86.96%	95.93%	n/a	13 / 17 (76.5%)	
ASM359593v1	MC32	KL22	None	?-+*	94.13%	99.92%	n/a	16 / 21 (76.2%)	KL93/KL22_*pgt1* (99%); KL93/KL24_*pgt1* (100.0%)
ASM359595v1	MC31	KL22	None	?-	89.35%	99.92%	n/a	18 / 21 (85.7%)	
ASM359612v1	MC57	KL37	None	?-*	75.83%	95.37%	n/a	11 / 17 (64.7%)	
ASM359613v1	MC48	KL33	None	?-	78.57%	98.40%	n/a	13 / 20 (65%)	
ASM359617v1	MC27	KL49	None	?-	86.44%	99.19%	n/a	23 / 28 (82.1%)	
ASM18971v2	4190	KL27	None	?-	95%	95%	n/a	16 / 24 (64.2%)	

Appendix 3. *Kaptive OCL analysis output for the original 82 ST25 isolates including key output details. Isolates which were excluded from the final genome pool are highlighted in grey*

Assembly	Strain	Best match	Match confidence	Problems	Coverage	Identity	Length discrepancy	Expected genes in locus	Other genes in locus
ASM58113v1	1429530	OCL6	Perfect		100.00%	100.00%	n/a	10 / 10 (100%)	
ASM415359v1	IHIT38008	OCL6	Perfect		100.00%	100.00%	n/a	10 / 10 (100%)	
7468_2_59	D46	OCL6	Perfect		100.00%	100.00%	n/a	10 / 10 (100%)	
8346_4#10	UV_1036	OCL5	Very high		100.00%	99.73%	+1 bp	9 / 9 (100%)	
8346_4#80	259_an	OCL5	Very high		100.00%	99.76%	+1 bp	9 / 9 (100%)	
8346_4#83	266_an	OCL5	Very High		100.00%	99.76%	+1 bp	9 / 9 (100%)	
8346_4#86	276_ax	OCL5	Very High		100.00%	99.76%	+1 bp	9 / 9 (100%)	
8346_4#94	295_c	OCL5	Very High		100.00%	99.73%	+1 bp	9 / 9 (100%)	
9262_1#33	338_an	OCL5	Very High		100.00%	99.73%	+1 bp	9 / 9 (100%)	
9262_1#34	338_ax	OCL5	Very High		100.00%	99.73%	+1 bp	9 / 9 (100%)	
8346_4#3	UV_964	OCL5	Very High		100.00%	99.73%	+1 bp	9 / 9 (100%)	
8346_4#6	UV_973	OCL5	Very High		100.00%	99.73%	+1 bp	9 / 9 (100%)	
8346_4#76	238_an	OCL5	Very High		100.00%	99.73%	+1 bp	9 / 9 (100%)	
ASM24170v2	AB5256	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
AcbauOIFC143v1.0	OIFC143	OCL5	Very high		100.00%	99.76%	+1 bp	9 / 9 (100%)	
Acin_baum_NIPH_146_V1	NIPH146	OCL6	Very high		100.00%	99.89%	n/a	10 / 10 (100%)	
ABIBUN_1	107m	OCL5	Very high		100.00%	99.76%	+1 bp	9 / 9 (100%)	
_ASM51657v2	CI86	OCL5	Very high		100.00%	99.73%	+1 bp	9 / 9 (100%)	
ASM51663v2	CI79	OCL5	Very high		100.00%	99.73%	+1 bp	9 / 9 (100%)	
ASM100768v1	LUH_7841	OCL7	Very high		100.00%	98.78%	+3 bp	9 / 9 (100%)	
ASM100770v1	RUH1486	OCL6	Very high		100.00%	99.89%	n/a	10 / 10 (100%)	
ASM100772v1	NM3	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
ASM100774v1	LUH_6220	OCL7	Very high		100.00%	98.78%	+3 bp	9 / 9 (100%)	
ASM143254v1	ABBL018	OCL5	Very high		100.00%	99.76%	+1 bp	9 / 9 (100%)	
ASM161199v1	AB2828	OCL5	Very high		100.00%	99.75%	+1 bp	9 / 9 (100%)	
ASM161209v1	AB3638	OCL5	Very high		100.00%	99.75%	+1 bp	9 / 9 (100%)	
ASM161210v1	AB3806	OCL5	Very high		100.00%	99.75%	+1 bp	9 / 9 (100%)	

Appendix 3. *Kaptive OCL analysis output for the original 82 ST25 isolates including key output details. Isolates which were excluded from the final genome pool are highlighted in grey*

|									
ASM208278v1	HWBA8	OCL5	Very high		100.00%	99.78%	+1 bp	9 / 9 (100%)	
ASM213810v1	PR388	OCL5	Very high		100.00%	99.75%	+1 bp	9 / 9 (100%)	
ASM214397v1	ARLG1317	OCL5	Very high		100.00%	99.75%	+1 bp	9 / 9 (100%)	
ASM257380v1	BAuABod-3	OCL6	Very high		100.00%	99.51%	n/a	10 / 10 (100%)	
ASM292796v1	MCR10179	OCL10	Very high		100.00%	99.91%	n/a	11 / 11 (100%)	
ASM300603v1	AR_0088	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
ASM343138v1	7804	OCL7	Very high		100.00%	98.78%	+3 bp	9 / 9 (100%)	
ASM358442v1	MC89	OCL5	Very high		100.00%	99.88%	+1 bp	9 / 9 (100%)	
ASM395535v1	R348	OCL5	Very high		100.00%	99.76%	+1 bp	9 / 9 (100%)	
7521_8_34	D4	OCL5	Very high		100.00%	99.76%	+1 bp	9 / 9 (100%)	
24276_2_166	4300STDY7045700	OCL5	Very high		100.00%	99.71%	n/a	9 / 9 (100%)	
24276_2_251	4300STDY7045786	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
24276_2_267	4300STDY7045802	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
24276_2_289	4300STDY7045824	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
24276_2_305	4300STDY7045840	OCL6	Very high		100.00%	99.99%	n/a	10 / 10 (100%)	
ASM100777v1	741019	OCL10	High	-	100.00%	99.87%	+1048 bp	10 / 11 (90.9%)	
ASM18971v2	4190	OCL5	High	?-	99.93%	99.84%	-1 bp	10 / 11 (90.9%)	
ASM359612v1	MC57	OCL10	High	-	99.93%	99.84%	+1048 bp	10 / 11 (90.9%)	
ASM30133v1	AB_2008-15-69	OCL5	Good	?-	100.00%	99.75%	n/a	8 / 9 (88.9%)	
ASM100776v1	4390	OCL7	Good	?	100.00%	98.78%	n/a	9 / 9 (100%)	
ASM100781v1	161/07	OCL5	Good	?-	100.00%	99.74%	n/a	8 / 9 (88.9%)	
ASM292785v1	HEU3	OCL6	Good	?-	100.00%	99.99%	n/a	9 / 10 (90%)	
ASM356944v1	OCU_Ac2	OCL7	Good	-+*	78.52%	92.65%	+1743 bp	6 / 9 (66.7%)	OCL6/OCL04 *gtrOC19a* (82%)
ASM358425v1	MC104	OCL5	Good	?-	96.05%	99.89%	n/a	7 / 9 (77.8%)	
ASM358427v1	MC102	OCL5	Good	?	100.00%	99.82%	n/a	9 / 9 (100%)	
ASM358431v1	MC100	OCL5	Good	?-	98.87%	99.85%	n/a	7 / 9 (77.8%)	
ASM358448v1	MC29	OCL5	Good	?-	99.76%	99.87%	n/a	8 / 9 (88.9%)	
ASM358451v1	MC93	OCL5	Good	?-	99.09%	99.88%	n/a	8 / 9 (88.9%)	

Appendix 3. Kaptive OCL analysis output for the original 82 ST25 isolates including key output details. Isolates which were excluded from the final genome pool are highlighted in grey

ASM359581v1	MC51	OCL10	Good	?-	100.00%	99.87%	n/a	10 / 11 (90.9%)	
ASM359586v1	MC39	OCL10	Good	?-	100.00%	99.87%	n/a	10 / 11 (90.9%)	
ASM359623v1	MC14	OCL5	Good	?	100.00%	99.88%	n/a	9 / 9 (100%)	
24276_3_50	4300STDY7045893	OCL6	Good	?-	100.00%	100.00%	n/a	9 / 10 (90%)	
ASM359568v1	MC63	OCL5	Low	?-	94.99%	99.88%	n/a	7 / 9 (77.8%)	
ASM359593v1	MC32	OCL5	Low	?-	90.73%	99.73%	n/a	7 / 9 (77.8%)	
ASM58105v1	984213	OCL5	None	?-	77.04%	99.74%	n/a	6 / 9 (66.7%)	
ASM358353v1	MC103	OCL5	None	?-	73.36%	99.92%	n/a	6 / 9 (66.7%)	
ASM358356v1	MC95	OCL5	None	?-	73.59%	99.92%	n/a	6 / 9 (66.7%)	
ASM358357v1	MC94	OCL5	None	?-	71.81%	99.92%	n/a	6 / 9 (66.7%)	
ASM358360v1	MC87	OCL5	None	?-	76.58%	99.92%	n/a	6 / 9 (66.7%)	
ASM358362v1	MC98	OCL5	None	?-	74.09%	99.92%	n/a	6 / 9 (66.7%)	
ASM358430v1	MC101	OCL5	None	?-	83.29%	99.03%	n/a	6 / 9 (66.7%)	
ASM358432v1	MC96	OCL5	None	?-	76.45%	99.92%	n/a	6 / 9 (66.7%)	
ASM358436v1	MC91	OCL5	None	?-	77.00%	99.92%	n/a	6 / 9 (66.7%)	
ASM358437v1	MC77	OCL5	None	?-	78.78%	99.88%	n/a	6 / 9 (66.7%)	
ASM358438v1	MC78	OCL5	None	?-	74.95%	99.92%	n/a	6 / 9 (66.7%)	
ASM358444v1	MC90	OCL5	None	?-	75.65%	99.94%	n/a	6 / 9 (66.7%)	
ASM358445v1	MC71	OCL5	None	?-	84.59%	99.92%	n/a	7 / 9 (77.8%)	
ASM359566v1	MC69	OCL5	None	?-	84.56%	99.92%	n/a	6 / 9 (66.7%)	
ASM359567v1	MC64	OCL5	None	?-	77.52%	99.92%	n/a	6 / 9 (66.7%)	
ASM359576v1	MC59	OCL5	None	?-	88.74%	99.92%	n/a	7 / 9 (77.8%)	
ASM359592v1	MC34	OCL5	None	?-	89.80%	99.92%	n/a	6 / 9 (66.7%)	
ASM359595v1	MC31	OCL5	None	?-	76.99%	99.92%	n/a	6 / 9 (66.7%)	
ASM359613v1	MC48	OCL5	None	?-	74.81%	99.90%	n/a	6 / 9 (66.7%)	
ASM359617v1	MC27	OCL5	None	?-	72.96%	99.92%	n/a	6 / 9 (66.7%)	
ASM359620v1	MC18	OCL5	None	?-	72.72%	99.95%	n/a	5 / 9 (55.6%)	

Appendix 4. *Workflow diagram of ST25 genome assembly collection and KL and OCL identification outlined in Aim 1.*
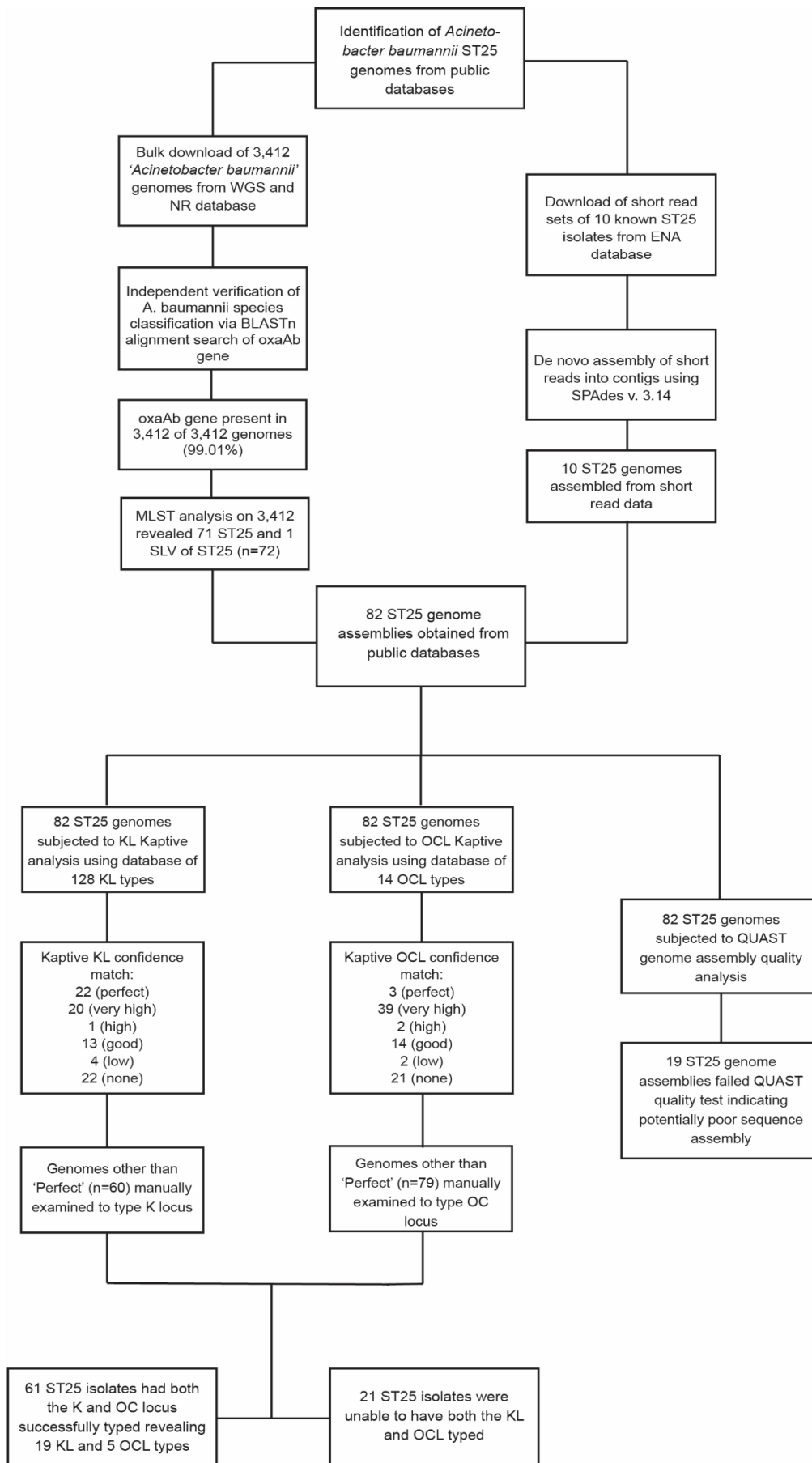
Elucidation of the K32 Capsular Polysaccharide Structure and Characterization of the KL32 Gene Cluster of *Acinetobacter baumannii* LUH5549

S. M. Cahill[1], N. P. Arbatsky[2], A. S. Shashkov[2], M. M. Shneider[3],
A. V. Popova[4,5], R. M. Hall[6#], J. J. Kenyon[1#], and Y. A. Knirel[2,a*#]

[1]*Institute of Health and Biomedical Innovation, School of Biomedical Sciences,*
Faculty of Health, Queensland University of Technology, QLD 4001 Brisbane, Australia
[2]*Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, 119991 Moscow, Russia*
[3]*Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 117997 Moscow, Russia*
[4]*Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Moscow Region, Russia*
[5]*State Research Center for Applied Microbiology and Biotechnology, 142279 Obolensk, Moscow Region, Russia*
[6]*School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia*
[a]*e-mail: yknirel@gmail.com*

Abstract—Capsular polysaccharide (CPS), isolated from *Acinetobacter baumannii* LUH5549 carrying the KL32 capsule biosynthesis gene cluster, was studied by sugar analysis, Smith degradation, and one- and two-dimensional ^{1}H and ^{13}C NMR spectroscopy. The K32 CPS was found to be composed of branched pentasaccharide repeats (K units) containing two residues of β-D-Gal*p*NAc and one residue of β-D-Glc*p*A (β-D-glucuronic acid) in the main chain and one residue each of β-D-Glc*p* and α-D-Glc*p*NAc in the disaccharide side chain. Consistent with the established CPS structure, the KL32 gene cluster includes genes for a UDP-glucose 6-dehydrogenase (Ugd3) responsible for D-GlcA synthesis and four glycosyltransferases that were assigned to specific linkages. Genes encoding an acetyltransferase and an unknown protein product were not involved in CPS biosynthesis. Whilst the KL32 gene cluster has previously been found in the global clone 2 (GC2) lineage, LUH5549 belongs to the sequence type ST354, thus demonstrating horizontal gene transfer between these lineages.

DOI: 10.1134/S000629792002011X

Keywords: *Acinetobacter baumannii*, capsule, K locus, polysaccharide structure, glucuronic acid

Capsular polysaccharide (CPS) is produced on the cell surface of many bacterial pathogens including the nosocomial species, *Acinetobacter baumannii*. It serves as a major pathogenicity determinant by providing a layer protecting against host immune factors [1, 2]. It also protects cells from desiccation and other environmental stressors [3, 4]. CPS is composed of a long chain of oligosaccharide repeats (K units), which consist of two to six sugar residues of various types joined by specific glycosidic linkages. The biosynthesis and export of CPS in *A. baumannii* is mostly determined by the genes at the chromosomal K locus (KL), which comes in diverse forms [5].

The K locus includes an operon of genes (*wza-wzb-wzc*) responsible for the CPS export, and a divergently transcribed region for the synthesis of specific K units. The latter includes genes for sugar synthesis, initiation of K-unit synthesis (*itr*), glycosyl transfer (*gtr*), K-unit translocation (*wzx*) and polymerization (*wzy*), and occasionally for addition of acetyl or pyruvyl groups (*atr/ptr*).

The arrangement of KL gene clusters carried by isolates from the Traub collection that were used to establish the original serotyping scheme for *A. baumannii* [6] had been examined previously [7]. However, these gene clus-

---

Abbreviations: CPS, capsular polysaccharide; KL, chromosomal K locus; K unit, oligosaccharide repeat; MPS, modified polysaccharide; PSgc, polysaccharide gene cluster.
[#] These authors contributed equally to this work.
[*] To whom correspondence should be addressed.

ters were originally described as polysaccharide gene clusters (PSgc) and annotated using a traditional nomenclature system. The structures of CPSs produced by many of these isolates have now been established and, for all of them, their sugar content and configuration correlate perfectly with the genes in the gene cluster found at the K locus [8-12]. Accordingly, the annotations of these gene clusters have been revised [8-14] in accordance with the more transparent and widely adopted K locus nomenclature system for *A. baumannii* CPS biosynthesis gene clusters [5].

In this work, we establish for the first time the CPS structure of another strain from this collection, LUH5549, which carries the KL32 gene cluster, previously designated PSgc21.

MATERIALS AND METHODS

Bacterial strain and cultivation of bacteria. *Acinetobacter baumannii* LUH5549 was originally from the W. H. Traub collection at the Institut fur Medizinische Mikrobiologie und Hygiene, Universitat des Saarlandes (Saarland, Germany) and was kindly provided by Prof. Peter Reeves. The bacteria were cultivated in 2TY media overnight; the cells were harvested by centrifugation ($10,000g$, 20 min), washed with phosphate buffered saline, resuspended in aqueous 70% acetone, precipitated, and dried.

Isolation of CPS. CPS preparation was obtained by phenol−water extraction [15] of bacterial cells (1.1 g); the extract was dialyzed without layer separation and clarified from insoluble contaminants by centrifugation. The resulting solution was treated with cold (4°C) 50% aqueous CCl_3COOH; after centrifugation, the supernatant was dialyzed against distilled water and freeze-dried to yield a crude CPS sample (370 mg). For CPS purification, this sample (120 mg) was heated with 2% aqueous AcOH at 100°C for 2 h, and CPS (27 mg) was purified by gel-permeation chromatography on a Sephadex G50 Superfine column 60×2.5 cm in 0.1% aqueous AcOH; eluted fractions were monitored with a differential refractometer (Knauer, Germany).

Monosaccharide analysis. Alditol acetates [16] were obtained by CPS hydrolysis with 2 M CF_3COOH (120°C, 2 h) and analyzed by gas-liquid chromatography on an HP-5 capillary column with a Maestro (Agilent 7820) chromatograph (Interlab, Russia) using a temperature gradient of 160°C (1 min) to 290°C at 7°C/min. Glucuronic acid was identified by anion-exchange chromatography on a DA×8 resin column (7×0.4 cm) in 5 mM potassium phosphate buffer (pH 3) at 70°C; elution was monitored using bicinchoninic acid assay.

Smith degradation. CPS sample (12 mg) was oxidized with $NaIO_4$ (29 mg in 1.5 ml water) at 20°C for 72 h in the dark and then reduced with $NaBH_4$ (30 mg) for 24 h; water was evaporated and boric acid was removed by evaporation with 10% aqueous AcOH in methanol (three times). The degradation product was desalted on a Sephadex G-25 column (108×1.2 cm) in water; elution was monitored as described above. The obtained polymer was hydrolyzed with 2% aqueous AcOH (100°C, 2 h), and the products of hydrolysis were fractionated by gel-permeation chromatography on Sephadex G-25 as described above to yield modified polysaccharide (MPS) (2.6 mg).

NMR spectroscopy. The samples were deuterium-exchanged by freeze-drying from 99.9% D_2O and then examined after dissolving in 99.95% D_2O. The NMR spectra were recorded on an Avance II 600 MHz spectrometer (Bruker, Germany) at 60°C. Sodium 3-trimethylsilylpropanoate-2,2,3,3-d_4 (δ_H 0, δ_C −1.6) was used as an internal reference for calibration. 2D NMR spectra were obtained using standard Bruker software; NMR data were acquired and processed with the Bruker TopSpin 2.1 program. A spin-lock time of 60 ms and a mixing time of 150 ms were used in TOCSY and ROESY experiments, respectively. A 60-ms delay was used for the evolution of long-range coupling to optimize $^1H,^{13}C$ HMBC experiments.

Bioinformatics analysis. Short reads for the LUH5549 genome sequence were obtained from the Sequence Read Archive (SRA number DRS005644) and assembled into contigs using SPAdes [17]. The KL32 gene cluster was identified between the *fkpA* and *lldP* genes and re-annotated using the nomenclature system described previously [5]. CAZy (http://www.cazy.org/) [18] and Pfam (https://pfam.xfam.org) [19] databases were used to assign encoded proteins to their respective biosynthesis roles. The assembled sequence and the newly assigned annotations are available in GenBank under accession number KC526897.2. The genome sequence was assigned to a sequence type (ST) using the Pasteur MLST scheme for *A. baumannii* (https://pubmlst.org/ bigsdb?db=pubmlst_abaumannii_pasteur_seqdef).

RESULTS AND DISCUSSION

The sequence of the PSgc21 CPS biosynthesis gene cluster from *A. baumannii* LUH5549 (GenBank accession number KC526897.1) was found to be incomplete as it lacked the *wza-wzb-wzc* export genes, indicating a potential assembly issue. As the sequence had been previously assembled from short reads using Velvet 2.0 [7], these short reads (SRA number DRS005644) were reassembled into contigs using SPAdes. The complete sequence of the gene cluster between the conserved *fkpA* and *lldP* genes that flank the K locus was identified and found to be 99.56% identical to the KL32 gene cluster previously described for the *A. baumannii* Vietnamese isolate, BAL_058 (GenBank accession number
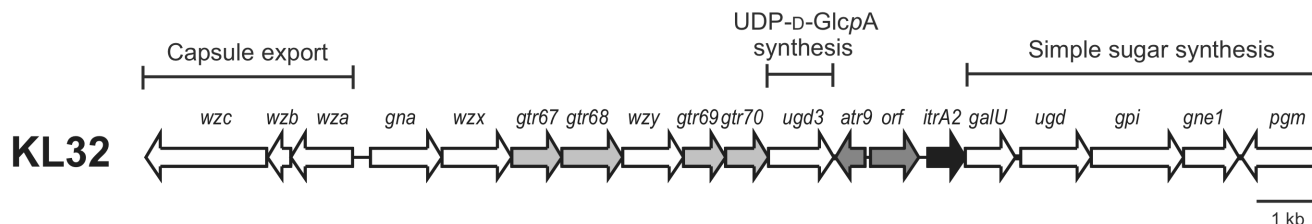
Fig. 1. The KL32 capsule biosynthesis gene cluster from *A. baumannii* LUH5549. Arrows indicate the direction of transcription and are drawn to scale using the sequence from GenBank accession number KC526879.2. Genes for glycosyltransferases (*gtr*) are depicted in light grey; initiating transferase gene (*itrA2*) is shown in black. Genes shown in dark grey (*atr9*, *orf*) have no established role in the synthesis of the K32 CPS.
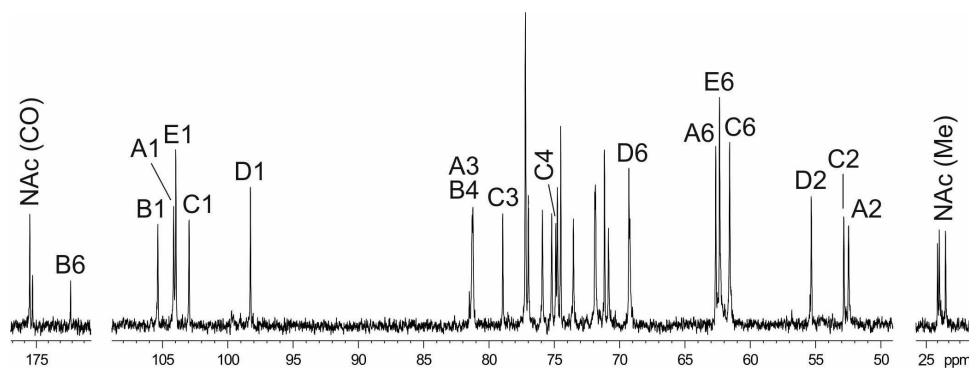
Fig. 2. ^{13}C NMR spectrum of the K32 CPS from *A. baumannii* LUH5549. Numbers refer to carbons in sugar residues denoted by letters as indicated in the table and Fig. 3.

KT359615.1) [20]. Hence, the LUH5549 gene cluster was renamed KL32 and its genes were reannotated according to the established nomenclature system [5]. The updated sequence and the annotations were deposited in the GenBank under accession number KC526897.2.

KL32 (Fig. 1) includes genes responsible for the K unit processing (*wzy* and *wzx*) and a *galU-ugd-gpi-gne1-pgm* gene module responsible for the synthesis of simple sugars, such as UDP-D-Glc*p*, UDP-D-Glc*p*NAc, and UDP-D-Gal*p*NAc. Adjacent to *galU*, there is an *itrA2* gene known to be responsible for initiating K unit synthesis by adding a D-Gal*p*NAc residue as the first sugar of the K unit to the lipid carrier in the inner membrane [21]. The central region of KL32 also includes four predicted glycosyltransferase genes (*gtr67*, *gtr68*, *gtr69*, *gtr70*) for four internal K unit sugar linkages, an acetyltransferase gene (*atr9*), and a gene encoding an unidentified protein product (Orf).

In the central region, another *ugd* gene, designated *ugd3*, is also present. Ugd3 (GenPept accession number AHB32286.1) is only 20% identical (74% coverage) to the product of the *ugd* gene (GenPept accession number AHB32291.1) located in the simple sugar synthesis gene module, which was previously predicted to be responsible for the conversion of UDP-D-Glc*p* to UDP-D-glucuronic acid (D-Glc*p*A) [5]. However, though *ugd* is a feature common to all *A. baumannii* KL gene clusters, its role in the CPS synthesis has never been established. Ugd3 is also 27% identical (71% coverage) to the second Ugd type (Ugd2) identified in the species, encoded in the central region of the KL20 and KL21 gene clusters. K20 and K21 CPSs contain D-Glc*p*A, which correlates with the presence of Ugd2 (GenPept accession numbers AUG44319.1 and AIT56461.1) [22]. The presence of *ugd3* also in the central region of KL32 specific to the CPS synthesis suggests that the K32 CPS may also contain D-Glc*p*A.

For the structure elucidation, CPS was isolated from the bacterial mass by phenol/water extraction followed by heating with 2% AcOH to remove contaminating short-chain lipopolysaccharides and then purified by gel-permeation chromatography on Sephadex G-50 Superfine. Sugar analysis using gas-liquid chromatography of alditol acetates derived after full acidic hydrolysis of CPS revealed the presence of Glc, GlcNAc, and GalNAc in the 1 : 1.6 : 1.8 ratio (detector response). In addition, GlcA was identified by anion-exchange chromatography. The absolute configuration of the monosaccharides was not determined chemically but inferred from the genetic data (see above).

The ^{1}H NMR and ^{13}C NMR (Fig. 2) spectra of the CPS showed signals for five monosaccharide residues and

^{1}H and ^{13}C NMR chemical shifts (ppm) of the K32 CPS from *A. baumannii* LUH5549

Sugar residue		C-1 *H-1*	C-2 *H-2*	C-3 *H-3*	C-4 *H-4*	C-5 *H-5*	C-6 *H-6*
CPS							
→3)-β-D-Gal*p*NAc-(1→	A	104.1 *4.54*	52.5 *3.83*	81.3 *3.89*	69.2 *4.06*	75.9 *3.59*	62.6 *3.74; 3.82*
→4)-β-D-Glc*p*A-(1→	B	105.3 *4.58*	73.5 *3.40*	75.2 *3.64*	81.2 *3.84*	74.8 *4.01*	172.1
→3,4)-β-D-Gal*p*NAc-(1→	C	102.9 *4.53*	52.8 *4.06*	79.0 *3.84*	74.9 *4.22*	77.0 *3.73*	61.6 *3.64; 3.68*
→6)-α-D-Glc*p*NAc-(1→	D	98.3 *4.86*	55.3 *3.90*	71.9 *3.87*	70.9 *3.71*	72.0 *4.34*	69.3 *4.06; 4.32*
β-D-Glc*p*-(1→	E	104.0 *4.47*	74.5 *3.35*	77.2 *3.49*	71.2 *3.40*	77.2 *3.43*	62.4 *3.72; 3.91*
MPS							
→3)-β-D-Gal*p*NAc-(1→	A	103.7 *4.58*	52.4 *3.97*	81.1 *3.85*	69.1 *4.12*	76.0 *3.64*	62.3 *3.76; 3.78*
→4)-β-D-Glc*p*A-(1→	B	105.4 *4.50*	73.9 *3.36*	75.1 *3.57*	81.1 *3.76*	77.9 *3.68*	
→3)-β-D-Gal*p*NAc-(1→	C	102.3 *4.51*	52.4 *3.95*	80.7 *3.80*	69.3 *4.16*	76.0 *3.70*	62.3 *3.77; 3.81*

Notes: ^{1}H NMR chemical shifts are italicized. Chemical shifts for the *N*-acetyl groups are δ_H 1.96-2.06, δ_C 23.3-24.0 (Me) and 175.3-175.5 (CO).

three *N*-acetyl groups. Assignment of the spectra using two-dimensional ^{1}H,^{1}H COSY, ^{1}H,^{1}H TOCSY, ^{1}H,^{1}H ROESY, ^{1}H,^{13}C HSQC, and ^{1}H,^{13}C HMBC experiments revealed spin systems for one residue each of β-Glc (unit **E**), β-GlcA (unit **B**), α-GlcNAc (unit **D**), and two residues of β-GalNAc (units **A** and **C**), all being in the pyranose form (table). The α-*gluco* configuration of unit **D**, β-*gluco* configuration of units **B** and **E**, and β-*galacto* configuration of units **A** and **C** were inferred from the $^3J_{H,H}$ coupling constants for the ring protons, which were estimated from the one- and two-dimensional NMR

spectra. The anomeric configuration of units **A**-**C** and **E** was confirmed by H-1/H-5 correlations in the ROESY spectrum.

Low-field positions of signals for C-3 of units **C** and **A** at δ 79.0 and 81.3, C-4 of units **C** and **B** at δ 74.9 and 81.2, respectively, as well as C-6 of unit **D** at δ 69.3 (table), as compared with their positions in the corresponding non-substituted monosaccharides [23, 24], showed that the CPS is branched and allowed determination of the glycosylation pattern in the K unit. Accordingly, the ^{13}C NMR chemical shifts for

MPS   →3)-β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*NAc-(1→
                              C   **B**   **A**

CPS
                E   **Gtr67**   **D**
         β-D-Glc*p*-(1→6)-α-D-Glc*p*NAc
                                        (1
         Gtr68   ↓   **Wzy$_{K32}$**
                                        4)   **Gtr69**   **Gtr70**   **[ItrA2]**
         →3)-β-D-Gal*p*NAc-(1→4)-β-D-Glc*p*A-(1→3)-β-D-Gal*p*NAc-(1→
                              C   **B**   **A**

Fig. 3. Structures of the K32 CPS from *A. baumannii* LUH5549 and MPS derived by Smith degradation of the CPS. Glycosyltransferases, Wzy polymerase, and ItrA2 initiating transferase are shown near the linkage they presumably form.

STRUCTURE OF *A. baumannii* K32 CAPSULAR POLYSACCHARIDE AND KL32 GENE CLUSTER   245

C-2,3,4,5,6 of the terminal unit **E** (table) were close to those of unsubstituted β-glucopyranose [23].

The ^{1}H,^{13}C HMBC spectrum of the CPS [Fig. S1; see Supplement to this paper on the web sites of the journal (http://protein.bio.msu.ru) and Springer (Link.springer.com)] showed correlations between the anomeric protons and linkage carbons at δ 4.47/69.3, 4.53/81.2, 4.54/79.0, 4.58/81.3, and 4.86/74.9, which were assigned to the **E** H-1/**D** C-6, **C** H-1/**B** C-4, **A** H-1/**C** C-3, **B** H-1/**A** C-3, and **D** H-1/**C** C-4 correlations, respectively. The glycosylation pattern of the CPS was confirmed by correlations of the anomeric protons and anomeric carbons with protons at the linkage carbons in the ^{1}H,^{1}H ROESY (see Fig. S2 in the Supplement) and ^{1}H,^{13}C HMBC spectra, respectively.

Therefore, the K32 CPS of *A. baumannii* LUH5549 has the structure shown in Fig. 3, which, to our knowledge, is unique among the known bacterial polysaccharide structures deposited in the Bacterial Carbohydrate Structure Database (BCSD; http://csdb.glycoscience.ru/bacterial/) [25]. This structure was confirmed by Smith degradation, which yielded a modified polysaccharide (MPS). Its structure (Fig. 3) was established by one- and two-dimensional NMR spectroscopy as described above for the CPS (see the table for assigned ^{1}H and ^{13}C NMR chemical shifts of the MPS).

The K32 CPS was found to be composed of pentasaccharide K units that include the common sugars, D-Gal*p*NAc, D-Glc*p*NAc, and D-Glc*p*, as well as D-Glc*p*A, as predicted. We previously suggested that the *ugd* gene present in the module for the simple sugar synthesis may be redundant with respect to the CPS production [5], and that D-Glc*p*A can only be a component of the *A. baumannii* CPS when a second *ugd* gene is present in the corresponding KL gene cluster [22]. Indeed, KL32 contains a *ugd3* gene in the central region, and as a result, K32 CPS contains D-Glc*p*A.

As the CPS structure determined includes two D-Gal*p*NAc residues (Fig. 3), it was unclear which residue was the first monosaccharide of the K32 unit that is added by ItrA2 (GenPept accession number AHB32289.2). However, the identity of the first sugar was determined by examining the linkage formed by the Wzy$_{K32}$ polymerase responsible for joining K units together into the CPS chain, as this linkage would connect the first and the last sugars in the chain of K units. Wzy$_{K32}$ (GenPept accession number AHB32283.1) was found to be 28% identical to Wzy$_{K116}$ encoded by *A. baumannii* KL116 (GenBank accession number MK399425.1), which forms the β-D-Gal*p*NAc-(1→3)-D-Gal*p*NAc linkage [26]. As an identical linkage is present in the K32 unit, Wzy$_{K32}$ was assigned to it. Therefore, the structure of the K32 unit begins with the monosubstituted D-Gal*p*NAc residue as drawn in Fig. 3.

The assembly of this K unit would therefore require three inverting glycosyltransferases to form the three β

linkages, and one glycosyltransferase with a retaining mechanism for the only α linkage. Only one glycosyltransferase encoded by KL32, Gtr68$_{K32}$ (GenPept accession number AHB32282.2), was identified as a retaining enzyme belonging to the GT4 family of retaining glycosyltransferases in the CAZy database [18]. Therefore, Gtr68$_{K32}$ was assigned to the α-D-Glc*p*NAc-(1→4)-D-Gal*p*NAc linkage.

The remaining glycosyltransferases were assigned to linkages in the K32 unit based on their homology to enzymes with known or predicted activities as follows. Gtr70$_{K32}$ (GenPept accession number AHB32285.1) is 54% identical to WdbN from the *Escherichia coli* O143 O-antigen gene cluster (GenPept accession number STM86374.1), and the O143 structure contains the β-D-Glc*p*A-(1→3)-D-Glc*p*NAc linkage [27]. A similar linkage, β-D-Glc*p*A-(1→3)-D-Gal*p*NAc, is present in K32 (Fig. 3), and therefore Gtr70$_{K32}$ was assigned to this linkage. Gtr67$_{K32}$ (GenPept accession number AHB32281.1) was found to be 27% identical to Gtr75$_{K37}$ encoded by *A. baumannii* KL37 gene cluster (GenBank accession number KX712115.1). The structure of K37 is known [26], and Gtr75$_{K37}$ was predicted to catalyze formation of the β-D-Glc*p*-(1→6)-D-Gal*p*NAc linkage. Hence, the β-D-Glc*p*-(1→6)-D-Glc*p*NAc side branch of K32 would be formed by Gtr67$_{K32}$. Finally, Gtr69$_{K32}$ (GenPept accession number AHB32284.1) that falls into the glycos_transf_2 family (Pfam PF00535) of glycosyltransferases, returns no significant hits to other proteins in BLASTp. However, as there was only one linkage left to be assigned in the K32 structure, Gtr69$_{K32}$ was assigned to the β-D-Gal*p*NAc-(1→4)-D-Glc*p*A linkage (Fig. 3).

Since the K32 unit contains no acetyl or other acyl groups, and all its structural features were attributed to the presence of other genes encoded by KL32, a role for Atr9 (GenPept accession number AHB32287.2) and Orf (GenPept accession number AHB32288.2) in the synthesis of the CPS could not be established.

Previously, the KL32 gene cluster had been identified in the widely disseminated clonal group, Global Clone 2 (GC2; equivalent to sequence type ST2 in the Pasteur MLST scheme) [20]. We found that *A. baumannii* LUH5549 belongs to another type, ST354, suggesting that the KL32 gene cluster is distributed amongst multiple distinct clonal lineages.

Funding. This work was supported by the Russian Science Foundation (project no. 19-14-00273) to YAK and Australian Research Council (ARC) DECRA Fellowship 180101563 to JJK.

Conflict of interest. The authors declare no conflict of interest in financial nor in any other area.

Ethical approval. This article does not contain any studies with human participants or animals performed by any of the authors.

BIOCHEMISTRY (Moscow) Vol. 85 No. 2 2020

REFERENCES

1. Russo, T., Luke N., Beanan, J., Olson, R., Sauberan, S., MacDonald, U., Schultz, L., Umland, T., and Campagnari, A. (2010) The K1 capsular polysaccharide of *Acinetobacter baumannii* strain 307-0294 is a major virulence factor, *Infect. Immun.*, **78**, 3993-4000.

2. Singh, J. K., Adams, F. G., and Brown, M. H. (2018) Diversity and function of capsular polysaccharide in *Acinetobacter baumannii*, *Front. Microbiol.*, **9**, 3301.

3. Tipton, K. A., Chin, C. Y., Farokhyfar, M., Weiss, D. S., and Rather, P. N. (2018) Role of capsule in resistance to disinfectants, host antimicrobials, and desiccation in *Acinetobacter baumannii*, *Antimicrob. Agents Chemother.*, **62**, e01188-18.

4. Geisinger, E., Huo, W., Hernandez-Bird, J., and Isberg, R. R. (2019) *Acinetobacter baumannii*: envelope determinants that control drug resistance, virulence, and surface variability, *Ann. Rev. Microbiol.*, **73**, 481-506, doi: 10.1146/annurev-micro-020518-115714.

5. Kenyon, J. J., and Hall, R. M. (2013) Variation in the complex carbohydrate biosynthesis loci of *Acinetobacter baumannii* genomes, *PLoS One*, **8**, e62160.

6. Traub, W. H. (1989) *Acinetobacter baumannii* serotyping for delineation of outbreaks of nosocomial cross-infection, *J. Clin. Microbiol.*, **27**, 2713-2716.

7. Hu, D., Liu, B., Dijkshoorn, L., Wang, L., and Reeves, P. R. (2013) Diversity in the major polysaccharide antigen of *Acinetobacter baumannii* assessed by DNA sequencing, and development of a molecular serotyping scheme, *PLoS One*, **8**, e70329, doi: 10.1371/journal.pone.0070329.

8. Senchenkova, S. N., Kenyon, J. J., Jia, T., Popova, A. V., Shneider, M. M., Kasimova, A. A., Shashkov, A. S., Liu, B., Hall, R. M., and Knirel, Y. A. (2019) The K90 capsular polysaccharide produced by *Acinetobacter baumannii* LUH5553 contains di-N-acetylpseudaminic acid and is structurally related to the K7 polysaccharide from *A. baumannii* LUH5533, *Carbohydr. Res.*, **479**, 1-5, doi: 10.1016/j.carres.2019.04.008.

9. Shashkov, A. S., Senchenkova, S. N., Popova, A. V., Mei, Z., Shneider, M. M., Liu, B., Miroshnikov, K. A., Volozhantsev, N. V., and Knirel, Y. A. (2015) Revised structure of the capsular polysaccharide of *Acinetobacter baumannii* LUH5533 (serogroup O1) containing di-*N*-acetyllegionaminic acid, *Russ. Chem. Bull. Int. Ed.*, **64**, 1196-1199, doi: 10.1007/s11172-015-1000-9.

10. Senchenkova, S. N., Popova, A. V., Shashkov, A. S., Shneider, M. M., Mei, Z., Arbatsky, N. P., Liu, B., Miroshinikov, K. A., Volozhantsev, N. V., and Knirel, Y. A. (2015) Structure of a new pseudaminic acid-containing capsular polysaccharide of *Acinetobacter baumannii* LUH5550 having the KL42 capsule biosynthesis locus, *Carbohydr. Res.*, **407**, 154-157, doi: 10.1016/j.carres. 2015.02.006.

11. Kasimova, A. A., Kenyon, J. J., Arbatsky, N. P., Shashkov, A. S., Popova, A. V., Knirel, Y. A., and Hall, R. M. (2018) Structure of the K82 capsular polysaccharide from *Acinetobacter baumannii* LUH5534 containing a D-galactose 4,6-pyruvic acid acetal, *Biochemistry (Moscow)*, **83**, 831-835, doi: 10.1134/S0006297918070064.

12. Shashkov, A. S., Liu, B., Kenyon, J. J., Popova, A. V., Shneider, M. M., Senchenkova, S. N., Arbatsky, N. P., Miroshnikov, K. A., Wang, L., and Knirel, Y. A. (2017) Structures of the K35 and K15 capsular polysaccharides of *Acinetobacter baumannii* LUH5535 and LUH5554 containing amino and diamino uronic acids, *Carbohydr. Res.*, **448**, 28-34, doi: 10.1016/j.carres.2017.05.017.

13. Kenyon, J. J., Shashkov, A. S., Senchenkova, S. N., Shneider, M. M., Liu, B., Popova, A. V., Arbatsky, N. P., Miroshnikov, K. A., Wang, L., Knirel, Y. A., and Hall, R. M. (2017) *Acinetobacter baumannii* K11 and K83 capsular polysaccharides have the same 6-deoxy-L-talose-containing pentasaccharide K units but different linkages between the K units, *Int. J. Biol. Macromol.*, **103**, 648-655, doi: 10.1016/j.ijbiomac.2017.05.082.

14. Shashkov, A. S., Kenyon, J. J., Arbatsky, N. P., Shneider, M. M., Popova, A. V., Miroshnikov, K. A., Hall, R. M., and Knirel, Y. A. (2016) Related structures of neutral capsular polysaccharides of *Acinetobacter baumannii* isolates that carry related capsule gene clusters KL43, KL47, and KL88, *Carbohydr. Res.*, **435**, 173-179, doi: 10.1016/j.carres. 2016.10.007.

15. Westphal, O., and Jann, K. (1965) Bacterial lipopolysaccharides: extraction with phenol–water and further applications of the procedure, in *Methods in Carbohydrate Chemistry* (Whistler, R., ed.) Academic Press, New York, pp. 83-91.

16. Sawardeker, J. S., Sloneker, J. H., and Jeanes, A. (1965) Quantitative determination of monosaccharides as their alditol acetates by gas liquid chromatography, *Anal. Chem.*, **37**, 1602-1604, doi: 10.1021/ac60231a048.

17. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, *J. Comput. Biol.*, **19**, 455-477, doi: 10.1089/cmb.2012. 0021.

18. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013, *Nucleic Acids Res.*, **42**, D490-D495, doi: 10.1093/nar/gkt1178.

19. Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016) The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Res.*, **44**, D279-D285, doi: 10.1093/nar/gkv1344.

20. Schultz, M. B., Thanh, D. P., Hoan, N. T. D., Wick, R. R., Ingle, D. J., Hawkey, J., Edwards, D. J., Kenyon, J. J., Lan, N. P. H., Campbell, J. I., Thwaites, G., Nhu, N. T. K., Hall, R. M., Fournier-Level, A., Baker, S., and Holt, K. E. (2016) Repeated local emergence of carbapenem-resistant *Acinetobacter baumannii* in a single hospital ward, *Microb. Genom.*, **2**, e000050, doi: 10.1099/mgen.0. 000050.

21. Kenyon, J. J., Marzaioli, A. M., Hall, R. M., and De Castro, C. (2014) Structure of the K2 capsule associated with the KL2 gene cluster of *Acinetobacter baumannii*, *Glycobiology*, **24**, 554-563, doi: 10.1093/glycob/cwu024.

22. Kasimova, A. A., Kenyon, J. J., Arbatsky, N. P., Shashkov, A. S., Popova, A. V., Shneider, M. M., Knirel, Y. A., and

Hall, R. M. (2018) *Acinetobacter baumannii* K20 and K21 capsular polysaccharide structures establish roles for UDP-glucose dehydrogenase Ugd2, pyruvyl transferase Ptr2 and two glycosyltransferases, *Glycobiology*, **28**, 876-884, doi: 10.1093/glycob/cwy074.

23. Lipkind, G. M., Shashkov, A. S., Knirel, Y. A., Vinogradov, E. V., and Kochetkov, N. K. (1988) A computer-assisted structural analysis of regular polysaccharides on the basis of ^{13}C NMR data, *Carbohydr. Res.*, **175**, 59-75, doi: 10.1016/0008-6215(88)80156-3.

24. Jansson, P. E., Kenne, L., and Schweda, E. (1987) Nuclear magnetic resonance and conformational studies on monoacetylated methyl D-*gluco-* and D-*galacto*-pyranosides, *J. Chem. Soc. Perkin Trans.*, **1**, 377-383, doi: 10.1039/P19870000377.

25. Toukach, P. (2011) Bacterial Carbohydrate Structure Database 3: principles and realization, *J. Chem. Inf. Model*, **51**, 159-170, doi: 10.1021/ci100150d.

26. Shashkov, A. S., Cahill, S. M., Arbatsky, N. P., Westacott, A. C., Kasimova, A. A., Shneider, M. M., Popova, A. V., Shagin, D. A., Shelenkov, A. A., Mikhailova, Y. V., Yanushevich, Y. G., Edelstein, M. V., Kenyon, J. J., and Knirel, Y. A. (2019) *Acinetobacter baumannii* K116 capsular polysaccharide structure is a hybrid of the K14 and revised K37 structures, *Carbohydr. Res.*, **484**, 107774, doi: 10.1016/j.carres.2019.107774.

27. Landersjo, C., Weintraub, A., and Widmalm, G. (1996) Structure determination of the O-antigen polysaccharide from the enteroinvasive *Escherichia coli* (EIEC) O143 by component analysis and NMR spectroscopy, *Carbohydr. Res.*, **291**, 209-216, doi: 10.1016/s0008-6215(96)00168-1.