Genomic and Experimental Investigations into Pneumococcal Bacteriocins and their Role in Competition

Madeleine Ella Bowler Butler

Thesis submitted for examination for the degree of PhD

June 2022

Department of Infectious Disease

Faculty of Medicine

Imperial College, London

Primary supervisor: Professor Angela Brueggemann

Abstract

Streptococcus pneumoniae ('the pneumococcus') is a frequent asymptomatic coloniser of the nasopharynx, from where it may disseminate to cause life-threatening infections including pneumonia, bacteraemia, and meningitis. Pneumococcal disease remains a leading cause of global mortality despite the use of safe and effective pneumococcal conjugate vaccines (PCVs). Bacteriocins are antimicrobial peptides that are produced by bacteria to target competitor bacteria within the ecological niche. Twenty pneumococcal bacteriocins have been characterised *in silico*, but their role in competition within the nasopharynx is not yet understood.

In the first part of this project, I studied the distribution of bacteriocin genes in two large genomic datasets (>5,000 pneumococcal genomes in total) sampled from Iceland and Kenya. The distribution of some bacteriocins differed by location, between pneumococci recovered from carriage and disease, and between pneumococci recovered before and after the introduction of PCVs. These observations were largely explained by the association of bacteriocins with clonal complexes and suggested that there were different competition dynamics among pneumococci.

A functional model of the streptococcin bacteriocins was generated using structural predictions. This informed further genomic studies, which observed genetic heterogeneity in the streptococcins. A dataset of >1,800 genomes from non-pneumococcal streptococci was screened for streptococcins, which were commonly harboured by viridans streptococci. There was evidence that the streptococcin

diversification was driven by horizontal exchange between pneumococci and non-pneumococcal streptococci.

In the final part of the project, the streptococcins were studied experimentally. A streptococcin toxin was isolated for the first time using a recombinant expression and purification method. The streptococcin was used in susceptibility assays against a panel of pneumococci and non-pneumococcal streptococci. Preliminary results suggested that the streptococcin had activity against some of the test strains.

Results presented in this thesis expand our understanding of pneumococcal bacteriocins and will be used to inform further genomic and experimental studies.

Statement of Originality

All work presented in this thesis is my own and was performed under the supervision of Professor Angela Brueggemann, unless otherwise referenced. Work that was undertaken in collaboration with others is indicated. Results presented in some chapters have been included in posters at conferences. These are indicated at the start of each chapter.

Copyright Declaration

Acknowledgments

Above all I must thank Professor Angela Brueggemann for her excellent supervision throughout this PhD. Angela has offered her guidance and expertise throughout, and always encouraged me to I explore many aspects of the project. I am also grateful for the support of members of the Brueggemann group at the BDI in Oxford, and to past members whose research underlies my own. Special thanks to Dr Melissa Jansen van Rensburg for teaching me the ways of BIGSdb and curation, and for introducing me to the world of computational biology. Thanks also to Femke Ahlers and Dr David Shaw for sanity checks and productive discussions. I thank the Wellcome Trust for funding my research, and Imperial College London and The University of Oxford for providing excellent research facilities.

I am grateful to Professor Shiranee Sriskandan for allowing me to use her laboratory space, for her many useful suggestions, and for her constructive feedback, along with Professor Gad Frankel, on my Progress Review Panel. Thanks also to members of Professor Sriskandan's research group, particularly Dr Kristin Huse, for the warm welcome and productive suggestions, and to Emily Wood and Jacob Lee in the Department of Infectious Disease. Finally, I am grateful to Dr Erin Cutts, who taught me how to purify a protein back in 2015, and who provided productive suggestions for streptococcin purification (and an MBP vector).

Genomic datasets are collaborative efforts, and I am grateful to past members of the Brueggemann group and members of collaborating groups in Reykjavik and Kilifi who

contributed to the collection of isolates, DNA extractions and sequencing, and genome assembly and quality control. Thanks also to Dr Keith Jolley for developing and maintaining the excellent BIGSdb platform, and for assisting me whenever I had an issue. Similarly, I wish to thank the community of bioinformaticians and software developers who produced the various open-source programmes that I made use of throughout my PhD.

Finally, I would not have been able to complete this thesis without my family and friends, who have offered support and encouragement throughout. Beth and Clarissa, thank you for always listening to me ramble about science, whether it be face to face or down a phone. Florence, thank you for your patience (and for checking my spelling). Tom, thank you for never being rude about my code, for the endless supply of coffee, and for always finding a way to make me laugh.

Table of Contents

Table of Figures

Table of Tables

1 General Introduction

1.1 The Pneumococcus

1.1.1 Background

Streptococcus pneumoniae, also known as the pneumococcus, is a Gram-positive, facultative anaerobic bacterial species that was first identified in the late 19th century as a cause of lower respiratory tract infection.[1] The study of pneumococcus in the first half of the 20th century led to several key discoveries, including the first description of genetic transformation[2] and the subsequent identification of DNA as the hereditary material.[3] Today, the pneumococcus remains a major cause of mortality and morbidity worldwide, and the World Health Organisation (WHO) has recently designated the pneumococcus one of 12 antimicrobial-resistant priority pathogens.[4]

1.1.2 Pneumococcal biology

1.1.2.1 Taxonomy

The pneumococcus belongs to the *Streptococcus* genus, which includes a wide range of species inhabiting diverse ecological niches, several of which are pathogenic to humans or other mammalian species. Group A *Streptococcus* (*S. pyogenes*) and group B *Streptococcus* (*S. agalactiae*) each have the potential to cause severe disease in humans, with the former causing scarlet fever and systemic infections,[5] and the latter being a leading cause of neonatal invasive disease.[6] *S. suis* is a major pathogen of pigs that is also capable of causing severe disease in humans,[7] and *S. equi* is a pathogen restricted to horses, where it causes 'strangles'.[8] Viridans streptococci are commonly found in the

human oral and respiratory tract microbiomes.[9] Pneumococci are genetically most similar to viridans streptococci such as *S. mitis*, *S. oralis,* and *S. pseudopneumoniae*.[1,10–12] Mitis group viridans streptococci are only rarely pathogenic, although species including *S. oralis* and *S. mitis* can cause invasive diseases such as endocarditis, especially in immunocompromised patients.[9,13]

1.1.2.2 Cell wall

The pneumococcal cell envelope consists of a single cell membrane surrounded by a thick peptidoglycan cell wall, which plays important roles in maintaining cell morphology, evading host immune responses, and resistance to antimicrobials.[14] The cell wall consists of peptidoglycan covalently joined to teichoic acid molecules. Peptidoglycan is composed of a repeating unit of two sugar residues (one N-acetylglucosamine, one N-acetylmuramic acid) with a stem peptide of three to five amino acid residues. The stem peptide is branched, and the repeating units are covalently linked by peptide bonds between the branches.[15] The peptidoglycan structure is not fixed; for example, the structure of the stem peptides in antimicrobial resistant strains are known to differ from susceptible strains.[16] Teichoic acids are a family of glycopolymers that are found in all Gram-positive cell walls, although there are many variant teichoic acids particular to different species.[17] Pneumococcal teichoic acids are notable as they often possess a choline residue.[18]

1.1.2.3 Polysaccharide capsule

Pneumococci express a polysaccharide capsule around the outside of the cell wall. Around 100 distinct polysaccharide capsules (serotypes) have been identified to date.[19–22] Capsular polysaccharide synthesis proceeds by attachment of a monosaccharide to a lipid, followed by the sequential addition of further monosaccharides, forming the lipid-

linked repeat unit. Each serotype has a characteristic set of modified monosaccharides. A specialised flippase transporter exports the repeat unit across the cell membrane, where it is polymerised with other repeat units, detached from its lipid, and covalently linked to the cell wall peptidoglycan.[1,23]

The genes required for production of the different capsule polysaccharides are found in the *cps* locus, which ranges in size from 10-30 kilobases (Kb).[19,24] All serotypes except 3 and 37 are synthesised *via* the Wzx/Wzy pathway, and the corresponding *wzx* and *wzy* genes, encoding the polysaccharide polymerase and flippase, respectively, are conserved in the *cps* loci.[19] Other conserved genes include the regulatory and processing genes *wzg*, *wzk*, *wzd*, and *wze.* The variable genes are largely glycosyl transferases, acetyl transferases, and sugar phosphate transferases, responsible for modifying the monosaccharide components of the capsular polysaccharide.

1.1.2.4 Pneumococcal genome

The pneumococcal genome is organised as a single circular chromosome around 2.1 mega bases (Mb) in length, encoding over 2000 predicted genes.[1,25,26] The pneumococcal core genome, defined as the genes that are shared across all or nearly all pneumococci, has been estimated to include at least 500 genes, and the pneumococcal pangenome, defined as all the genes that are found in the whole pneumococcal species, is estimated to be much larger (up to 7,000 genes).[26–28] This indicates a large accessory genome of non-essential genes, and in a given pneumococcal genome around 20% of the genes are expected to be accessory genes.[29] A notable feature of the pneumococcal genome is its plasticity: pneumococci are highly recombinant and readily exchange genetic material, resulting in

high adaptability to environmental conditions while maintaining a relatively small overall genome size.[30]

1.1.2.5 Horizontal genetic exchange

The horizontal exchange of genetic material between bacterial strains of the same or different species is a well-documented driver of bacterial genomic diversity.[3,31] There are three mechanisms of horizontal genetic exchange in bacteria:[32]

- Transformation: internalisation of exogenous DNA by naturally competent bacteria, followed by the incorporation of this DNA into the genome by homologous recombination.
- Transduction: the exchange of genetic material *via* a bacteriophage that integrates into the host genome in the lysogenic phase before replicating and lysing the host cell to release phage particles in the lytic phase. Phage genetic material that has integrated into a host genome is called a prophage.[33]
- Conjugation: the exchange of DNA (often a plasmid) *via* the specialised conjugation machinery, which requires cell-to-cell contact.[34]

Genetic material exchanged *via* transduction or conjugation includes genes for the machinery used in the exchange (either for the phage lifestyle or for conjugation machinery), and therefore is a large and discrete section of DNA transferred from the donor cell. These elements often carry additional cargo genes that are advantageous either for the host or the mobile genetic element, such as virulence factors and antimicrobial resistance genes.[35]

Pneumococci are naturally competent, and entry into the transient competent state is tightly regulated by the competence stimulating peptide.[1,36,37] Following competence

signalling, pneumococci are capable of taking up large amounts of DNA (over 1 Mb), that is then stabilised in the cytoplasm by specialised DNA-binding proteins and used as a source of new genetic material for multiple recombination events.[38–40] The length of recombined fragments is highly variable. Transformation events observed *in vitro* are typically on a scale of a single gene,[29] but transformation of smaller sections of sequence can also result in mosaic genes,[41–44] and transformation events spanning many kilobases of DNA have also been observed.[39,45,46]

Integrative conjugative elements (ICEs) in pneumococcus are integrated into the chromosome rather than existing as discrete conjugative plasmids.[47] Diverse ICEs have been identified in pneumococci and appear to have played an important role in the evolution of the species.[48–50] The evolution and spread of ICEs are complex: composite ICEs that contain genetic material from multiple different ICEs have been observed, genetic material within ICEs can be exchanged *via* transformation,[49] and pneumococcal ICEs have been observed with degenerated conjugative machinery.[48] Many pneumococcal ICEs appear to have origins in other streptococcal species, particularly *S. mitis*.[51]

Finally, diverse prophages have been observed in pneumococcal genomes.[52–54] Prophage genes have been associated with contributions to various bacterial characteristics, including virulence,[55,56] and an assessment of a pneumococcal prophage with a putative virulence gene suggests that this is likely the case in pneumococcal prophages.[57] The same study revealed the high number and diversity of pneumococcal satellite prophages, which have lost their structural components and rely on other prophages for survival.[57]

1.1.3 Pneumococcal carriage and disease

1.1.3.1 Nasopharyngeal carriage and transmission

The primary ecological niche of pneumococcus is the nasopharynx, and disease progresses from pneumococcal acquisition and asymptomatic carriage.[58,59] Carriage is common in infants and young children, with the highest carriage rates among children less than 5 years of age.[60–62] Carriage rates vary by location: for example, in a study of Kenyan carriage, 79% of children had detectable pneumococcal carriage in the first year of life,[63] whereas childhood carriage rates in European countries are consistently much lower.[62] Carriage rates decrease with age, typically reaching less than 10% in adolescents and remain low through adulthood.[61,62,64,65]

Nasopharyngeal pneumococci are the reservoir for transmission to new hosts.[66] A key step in pneumococcal transmission is shedding: the more pneumococci are released from the host, the higher the chances of colonising a new host. The capsular polysaccharide contributes to shedding by reducing entrapment in host mucous, and this is influenced by the serotype.[67] Another pneumococcal factor affecting shedding is the toxin pneumolysin, which prompts a strong inflammatory response in the host (see Section 1.1.4).[68] Inflammation is associated with increased shedding, explaining why co-infection with influenza A has been associated with higher pneumococcal transmission.[69] Following shedding, pneumococci are transmitted in saliva and nasal secretions either through airborne methods (such as sneezing) or by physical contact between individuals.[66,70] In order to establish colonisation in a new host, pneumococci must escape the immune response and successfully compete with the pre-existing nasopharyngeal flora in the niche (discussed below, Sections 1.1.4 and 1.1.5).

Close physical proximity is a requirement for efficient transmission, and transmission is higher between young children than between adults.[71] Children who attend day-care centres, and those with siblings, therefore have a significantly higher risk of pneumococcal carriage.[65,72] Adults with young children are also at higher risk, which increases further with the number of children in the household. Poverty increases the risk of pneumococcal carriage in children and adults, and this is likely due to a number of factors associated with a disadvantaged socio-economic background, including household crowding (resulting in closer proximity between individuals), restricted access to healthcare, malnutrition, and exposure to pollution.[65,72] Finally, immunosuppression and chronic respiratory disease both increase the risk of carriage by reducing the host immune response, removing a barrier to colonisation.[73]

1.1.3.2 Pneumococcal disease

Pneumococci colonising the nasopharynx may migrate to other mucosal surfaces to cause a localised symptomatic infection. The most serious of these is pneumonia, where the nasopharyngeal pneumococci are aspirated to the lower respiratory tract.[66,74] Pneumococcal pneumonia has a high mortality rate, particularly in the elderly, and commonly leads to invasive disease (defined as an infection of a normally sterile site, such as the bloodstream (bacteraemia) or cerebrospinal fluid (meningitis)), particularly in infants.[75] In 2016, pneumococcus was the leading cause of pneumonia mortality and morbidity, causing an estimated 1.2 million deaths globally.[76] Other localised pneumococcal infections include otitis media, sinusitis, and conjunctivitis.[60,77–80] Although these are less severe than pneumococcal pneumonia, they occur at a higher rate and represent a major burden on healthcare systems.[81]

Pneumococcus is an important cause of invasive disease.[82–84] There were an estimated 400,000 cases and 50,000 deaths due to invasive pneumococcal disease globally in 2015 in children under five years of age,[85] with high estimated case fatality rates (44% for pneumococcal meningitis, rising to 60% in Africa). Invasive pneumococcal disease typically proceeds from a localised infection, such as a lower respiratory tract infection, before gaining access to the bloodstream *via* the respiratory epithelium, or by invading the cerebrospinal fluid from the bloodstream.[66] However, pneumococci can also invade these sterile sites from other localised infections or directly from nasopharyngeal carriage.

Pneumococcal diseases are most prevalent among children under five years of age and elderly people, as well as immunocompromised individuals and those with other risk factors for infection (chronic heart, liver, renal and respiratory diseases, diabetes, smoking and deprivation all increase risk).[86] The burden of pneumococcal disease is disproportionately high in low- and middle-income countries in Africa and Asia.[85]

1.1.4 Virulence and the host immune response

The clearance of pneumococci from both nasopharyngeal carriage and disease is reliant on the complement system.[87,88] The complement system is triggered by three separate pathways: the classical pathway, the lectin pathway, and the alternative pathway.[89] Each pathway is triggered by different signals: the classical pathway responds to immunoglobulins IgG and IgM, the lectin pathway is triggered by recognition of mannose or other sugar moieties on the bacterial cell surface, and the alternative pathway is continuously activated and is amplified by the activation of the other pathways. Once triggered, complement signalling proceeds *via* proteolytic cascades and results in the

deposition of opsonins on the bacterial cell surface. Opsonisation is recognised by host phagocytes, increasing the efficiency of phagocytosis. Complement cascades also trigger an inflammatory response, lead to the assembly of the membrane attack complex (which generates pores in the target cell membrane), and interact with the adaptive immune system to promote pathogen clearance.[90]

All three complement pathways can be triggered by a pneumococcal infection. The most important is the classical pathway, which responds to antibodies against cell wall phosphorylcholine and other specific cell surface targets, particularly the polysaccharide capsule.[88,91] Complement-mediated opsonophagocytosis by host neutrophils is the main mechanism of pneumococcal clearance,[92,93] and the general inflammatory response also plays a role.[88] Cell lysis by the membrane attack complex is minimal due to the pneumococcal cell wall. Deficiencies in complement pathways have been associated with higher susceptibility to pneumococcal infection and more severe disease in humans and in animal models.[91,94,95] Nasopharyngeal colonisation by a particular pneumococcal strain is protective against re-colonisation or symptomatic infection caused by the same strain due to specific antibodies and a helper T-cell response.[96,97]

1.1.4.1 Serotype and disease

The highly variable polysaccharide capsule (defined by its serotype) is the major pneumococcal virulence factor.[98–100] Each serotype is antigenically different and varies both in the ability to cause disease and in the duration of nasopharyngeal carriage, and these characteristics are inversely correlated.[59,101] For example, serotypes 1, 4 and 14 have particularly high invasive disease potential, and are rarely observed in carriage, whereas the less invasive serotypes 23F, 19F, 6A and 6B, and serogroup 15 have longer

carriage durations.[59] Serotypes are unevenly distributed in the pneumococcal population - different serotypes are predominant in both carriage and disease in different geographic locations.[102]

The role of the polysaccharide capsule is to evade the host immune response.[1,88] The capsule acts as a physical barrier, reducing complement activation by blocking access to recognisable surface antigens and by inhibiting opsonisation of the cell surface, resulting in reduced phagocytosis.[103–105] The extent of immune evasion varies with serotype. For example, serotype 3 expresses an unusually thick, mucoid capsule and is able to evade host immunity particularly well, resulting in prolonged carriage and a high mortality rate when it causes symptomatic infections.[106–108] The importance of the capsule in pneumococcal pathogenesis is emphasised by the low rates of pneumococcal disease caused by nontypable pneumococci, which do not express a polysaccharide capsule.[109–111] The lack of capsule expression is due to disruptions within the *cps* locus. Nontypable pneumococci are typically associated with carriage and only cause invasive disease in rare cases,[112] although a widely distributed lineage of nontypable pneumococci can cause pneumococcal conjunctivitis.[78,109,113]

1.1.4.2 Determination of serotype

As the capsule has such an important contribution to pneumococcal virulence and host interactions, it is important to identify the serotype of pneumococcal isolates *in vitro*. Serotypes can be detected directly using immunological techniques or indirectly *via* the sequence of the *cps* locus. The Quellung reaction and latex agglutination are two widely used immunological techniques.[114,115] The Quellung reaction is the gold standard serotyping technique, but it is also technically demanding and more expensive, so latex

agglutination is also widely used in many settings.[116] Alternatively, the sequence of the *cps* locus can be assessed using DNA microarrays,[117] real-time PCR,[118] and multiplex PCR.[119,120] These approaches are reliant on knowledge of *cps* locus sequences, and sequence diversity within the *cps* locus may give false results.[116] Phenotypic, microarray and PCR-based approaches are limited by the antisera or primer sets used - they cannot detect serotypes that are not included in these materials and are prone to false negatives when a pneumococcal isolate possesses a novel serotype (or possesses sequence diversity in the *cps* locus in the case of sequence-based approaches). This is exemplified by examples of isolates that were historically designated as nontypeable but were later found to have detectable capsules.[121]

The relationship between *cps* locus sequence and serotype allows the prediction of serotype from whole genome sequences by comparison to reference *cps* sequences of known serotypes.[122–124] Serotype prediction tools have been useful in the investigation of *cps* locus sequence diversity and distribution.[21,122] However, *cps* locus sequence variation does not correspond directly to changes in the polysaccharide capsule structure. Rather, the encoded capsular processing and transport proteins are altered, with unpredictable effects on the capsule. For example, serotype 6A and 6B *cps* loci differ by a single SNP in the gene *wciP.*[125] Sequence diversity within serotypes is also not uniform: serotypes 1 and 3 are noted for their low *cps* sequence diversity, whereas serotypes 6B and 6E(6Bii) *cps* locus sequences differ by 7% but produce identical capsular polysaccharide structures.[20,122,126] *In silico* serotype predictions must therefore always rely on experimental data.

1.1.4.3 Other virulence factors

A great number of virulence factors beyond the polysaccharide capsule have been identified in pneumococcus.[66,127,128] Although none have as large a contribution to virulence as the capsule, they do fulfil a variety of important roles. Modulation of the host immune response is essential for evasion of complement-mediated clearance: virulence factors may prevent complement activation, degrade components of the signalling cascade, or block opsonisation of the cell surface.[88] Other factors promote the invasion of sterile sites, either by contributing to niche adaptation or by causing inflammatory responses and tissue injury, improving invasion of epithelial barriers.[66,129,130] Finally, factors that promote nasopharyngeal colonisation can also be considered virulence factors, as symptomatic infections progress from carriage. These include factors that mediate adhesion *via* interactions with cell surface ligands, the host extracellular matrix, and glycans within the mucous.[66]

Pneumolysin, a pore-forming toxin, is an important virulence factor with a multi-faceted role.[130,131] Lysis of host cells triggers an inflammatory response, which not only contributes to transmission, but also facilitates invasion of epithelial barriers. Additionally, pneumolysin appears to play a role in evading complement-mediated immunity.[88,132] Other important virulence factors are part of the family of choline-binding proteins, which are presented on the pneumococcal cell surface. These fulfil a range of functions, including adherence to host cells during both colonisation and infection (CbpA), evasion of host immune responses (PspA, LytA), and promotion of invasion either into the bloodstream or across the blood-brain barrier (CbpA, CbpL).[66,88,133] A final important virulence factor is the P1 pilus, in particular the RrgA sub-unit, which both acts

as an adhesin to improve colonisation and promotes the invasion of the blood-brain barrier from the bloodstream.[134,135]

1.1.5 Role of the nasopharyngeal microbiome

The nasopharynx is part of the upper respiratory tract, adjacent to the nasal cavity and the oropharynx. Bacterial species from diverse genera colonise the nasopharynx and are influenced by both other colonising species and the host immune response,[136] resulting in a dynamic microbiome that is nevertheless distinct from both nasal and oropharyngeal microbiomes.[137,138] In young children, the most common genera are *Moraxella, Haemophilus, Streptococcus, Flavobacteria, Dolosigranulum, Corynebacterium*, and *Neisseria.*[139–141] Colonisation by streptococci decreases with age.[142] Competition in the nasopharynx is high, and commensal bacteria employ a range of strategies to compete against other members of the microbiome for limited space and resources.[138,143]

For a pneumococcal strain to colonise the nasopharynx, it may need to out-compete other pneumococci, non-pneumococcal streptococci, *Staphylococcus aureus, Moraxella catarrhalis,* and *Haemophilus influenzae.*[138,140,144,145] There have been reports of inverse correlations between pneumococci and other colonising species, suggesting inter-species competition.[58,146] As nasopharyngeal colonisation precedes symptomatic disease, the ability of a strain to colonise the nasopharynx must be considered in pneumococcal pathogenesis. If a single strain is highly competitive and dominates the niche, its subsequent expansion could increase the chance of spread to another site. Indeed, nasopharyngeal microbiomes that are dominated by a single species, such as pneumococcus, are associated with an increased risk of respiratory disease.[140]

1.1.6 Pneumococcal vaccination

1.1.6.1 Pneumococcal polysaccharide vaccines

As the polysaccharide capsule is the major antigen on the pneumococcal cell surface, it is the target of all currently licenced pneumococcal vaccines. Polyvalent vaccines have been developed to target the serotypes responsible for the majority of pneumococcal disease.[147] The first pneumococcal polysaccharide vaccines (PPVs) were developed in the 20th century, starting with formulations of six serotypes and eventually leading to a 14-valent vaccine in the 1970s,[5,148] then a 23-valent vaccine (PPV23, Pneumovax23, Merck).[147] Although they were effective in reducing pneumococcal disease,[149] a major drawback of the PPVs is that they do not mount an anamnestic immune response in children under 2 years of age.[147,150,151] PPV23 remains in use globally and still has value as part of national vaccination strategies.[152]

1.1.6.2 Pneumococcal conjugate vaccines

Protein-polysaccharide conjugate vaccines attach polysaccharide antigens to a carrier protein. The first successful conjugate vaccine against the *H. influenzae* type B (Hib) capsule was introduced in 1990 and significantly reduced childhood Hib infections in countries that implemented the Hib vaccine.[153] Multivalent pneumococcal conjugate vaccines (PCVs) have been developed that include serotypes that (pre-PCV) were responsible for a large proportion of pneumococcal disease.[147] The initial heptavalent PCV (PCV7, Prevnar, Wyeth Pharmaceuticals Ltd.) was licensed in 2000 and included serotypes 4, 6B, 9V, 14, 18C, 19F and 23F. PCV7 significantly reduced the incidence of pneumococcal disease caused by the vaccine serotypes.[154–157] Higher valency vaccines have since been developed: PCV10 (Synflorix, GlaxoSmithKline,) added serotypes 1, 5 and 7F; and PCV13 (Prevnar13, Pfizer) included the PCV7 serotypes plus serotypes 1, 3, 5, 6A,

7F and 19A. Both PCV10 and PCV13 have further reduced the global burden of pneumococcal disease.[158–160] Importantly, PCVs also reduce carriage of vaccine serotype pneumococci in vaccinated children, interrupting transmission to unvaccinated children and adults and resulting in significant herd protection.[157,161–163] However, despite the success of PCVs, global pneumococcal mortality remains high in regions with low vaccine coverage.[76]

1.1.6.3 Vaccine escape

PCVs have had great success in reducing both carriage and disease caused by vaccine types but there were always concerns that introducing a limited-valency vaccine would introduce a selection pressure for nonvaccine types, potentially resulting in an increase in disease caused by nonvaccine type pneumococci.[154,155,164] These concerns were indeed borne out in the years following PCV introduction: while the overall rates of pneumococcal disease decreased, the proportion of pneumococcal disease caused by nonvaccine types increased significantly.[165–168] In post-PCV populations, carriage of vaccine serotype pneumococci decreases and carriage of nonvaccine serotypes typically increases, while the overall rate of pneumococcal carriage is not significantly altered.[61,169]

Generally, the pneumococcal population is dynamic, and lineages and serotypes fluctuate over time.[170] Additionally, PCVs introduce a selection pressure against a subset of pneumococcal serotypes, which results in changes to the overall pneumococcal population structure, since genetic lineages tend to be associated with particular serotypes.[98,171] As the vaccine serotypes decrease in prevalence in a vaccinated population, nonvaccine serotype genetic lineages expand to fill the niche. This is referred to as serotype replacement,[166] and a clear example was the increase in serotype 19A

infection in multiple post-PCV7 areas, driven by the expansion of the multidrug resistant clonal complex (CC) 320.[172]

Capsular switching events, horizontal genetic exchange of the *cps* locus that results in a change of the phenotypic serotype of the transformed pneumococcus, are also common among pneumococci.[2,173,174] Vaccine escape recombinants occur where a lineage undergoes a capsular switch from vaccine type to nonvaccine type following PCV introduction, thus evading immune responses in vaccinated individuals.[39,46,174] As pneumococci are highly recombinant, capsular switch events appear to occur relatively frequently within the population, and some genetic lineages appear to be more likely than others to exchange *cps* loci. Capsular switching events have also been described where the serotype switch event occurred prior to the introduction of PCVs, but the expansion of the lineage was only observed in the post-PCV time period, presumably as a result of the selection pressures introduced with the PCVs.[28,174]

The major consequences of serotype replacement and capsular switching are increased incidence of disease caused by nonvaccine serotypes and reduced efficacy of PCVs over time in restructured populations. Surveillance of pneumococci to monitor changes in serotypes causing disease is therefore required in post-vaccine populations and has informed the development of higher valency PCVs.[175] For example, serotypes 3 and 19A, both of which increased in post-PCV7 areas, were included in PCV13. Even higher valency PCVs have recently been licensed (PCV15, Merck; PCV20, Pfizer), and it is hoped that these will further reduce the pneumococcal disease burden. However, it will not be possible to develop a PCV to target all pneumococcal serotypes simultaneously, so any PCV may be expected to cause population restructuring and an increase in nonvaccine

type disease. Pneumococcal protein vaccines targeting a universal protein surface antigen are therefore a long-term goal of pneumococcal vaccine development.[147,176]

1.1.7 Antimicrobials

The discovery of penicillin in 1928 revolutionised the treatment of bacterial infections and rapidly reduced global pneumococcal mortality. Since their discovery, antimicrobial therapeutics have been used effectively to treat both localised and invasive pneumococcal infections,[1] although their high usage resulted in the evolution of resistant pneumococcal strains.[177–180] Multi-drug resistant lineages, which are resistant to penicillin and at least two further classes of antimicrobial, have emerged and become widely distributed.[181] The Pneumococcal Molecular Epidemiology Network (PMEN) was established to catalogue important resistant strains.[182] The antibacterial activities of important antimicrobial families are described below, with a discussion of the mechanism and distribution of pneumococcal resistance.

1.1.7.1 Beta-lactams

Beta-lactam antimicrobials include the penicillins, carbapenems, and cephalosporins. These antimicrobials have a common beta-lactam ring structure that is usually joined to a second cyclic motif. Otherwise, the beta-lactams exhibit diverse structures, which dictate their specificity.[183] Beta-lactams kill cells by covalently inhibiting penicillin binding proteins (PBPs) that are essential to the synthesis of cell wall peptidoglycan, weakening the overall cell wall structure.[184]

Pneumococcal cell wall biosynthesis is complex.[14] Amino acid residues are added to the peptidoglycan precursor lipid II on the cytoplasmic face of the cell membrane. This

molecule is then transported to the periplasmic face, where it is polymerised and cross-linked with the cell wall peptidoglycan by glycosyltransferases and transpeptidases. The majority of the peptidoglycan stem peptides are trimmed to reduce the pentapeptide to a tripeptide, which is important for immune evasion and epithelial adhesion.[185] Six PBPs are used in the polymerisation and cross-linking of peptidoglycan in pneumococcal cell wall biosynthesis: PBP1a, PBP1b and PBP2A are bifunctional glycosyltransferase-transpeptidases, PBP2b and PBP2x are transpeptidases, and PBP3 is a carboxypeptidase.[14]

Alterations in three of the pneumococcal PBPs (PBP2b, PBP2x, and PBP1a) that reduce their beta-lactam binding affinity are responsible for resistance in pneumococci.[14,186] Other pneumococcal factors involved in cell wall biosynthesis have also been shown to contribute to beta-lactam resistance, particularly in highly resistant pneumococci.[186] These include *murM, ciaRH,* and *clpL.*[187–189] Pneumococcal resistance to beta-lactams was first recognised in 1967 in Australia[190] and since then has been extensively documented globally in response to widespread beta-lactam usage.[177,180]

1.1.7.2 Macrolides

The macrolide family includes erythromycin and its variants, which share a macrocyclic lactam ring structure. Macrolide antimicrobials prevent bacterial growth by inhibiting protein synthesis *via* an interaction with the bacterial 50s ribosomal RNA (rRNA).[191] Macrolides were first used to treat infections caused by penicillin-resistant pneumococci, but their widespread use resulted in the emergence of macrolide resistance.[177] Macrolide resistance is of greater concern in some regions than others: the highest incidence of

resistance is observed in China, South Africa, South Korea and the USA, and lower incidence is observed in Europe.[192–194]

There are two mechanisms of macrolide resistance in pneumococci, ribosomal modification and macrolide efflux.[195] Ribosomal modification involves a methylation of the rRNA by the ErmB methyltransferase to prevent macrolide binding.[196] Macrolide efflux in pneumococci requires the *mefE/mel* operon. The mechanism of resistance conferred by these genes is not fully understood but is believed to involve both ribosomal protection and macrolide efflux.[197] The relative contribution of each resistance system varies depending on geographic location,[195] and pneumococci possessing both macrolide resistance systems simultaneously are not uncommon.[193]

1.1.7.3 Other classes of antimicrobials

The fluoroquinolones are a family of broad-spectrum antimicrobials with a shared bicyclic core structure. Fluoroquinolones target type II topoisomerases, which are involved in modulating DNA supercoiling, by interacting with the enzymes at the DNA binding site. This interaction ultimately kills the cell by causing fragmentation of the DNA.[198] In pneumococcus, the major fluoroquinolone targets are GyrA and ParC, and variants of these proteins confer resistance by reducing binding affinity for the antimicrobial. Overall, fluoroquinolone resistance rates are lower than for beta-lactams and macrolides due to reduced use of molecules from this class.[178,180]

Tetracyclines are a large group of compounds with a common structure comprising four fused rings with variable chemical groups that kill susceptible cells by binding the 16s rRNA and preventing transfer RNA (tRNA) interaction.[199] Pneumococcal resistance to

tetracyclines is largely *via* the ribosomal protection proteins TetM and TetO, which prevent tetracycline interactions with the ribosome.[180,200]

The broad-spectrum antimicrobial chloramphenicol and its derivatives have historically been used to treat pneumococcal disease.[201] These antimicrobials have a p-nitrobenzene ring with a dichloroacetyl tail and inhibit protein synthesis by binding to the bacterial ribosome. Resistance to chloramphenicol in pneumococci is largely *via* enzymatic modification by chloramphenicol acetyltransferases such as CatQ[202,203] and is more widespread in low- and middle-income countries where chloramphenicol is more widely used due to its low cost.[204,205] The antimicrobials trimethoprim/sulfamethoxazole (TMP/SMX) are a final example that are also used to treat pneumococcal disease more commonly in low- and middle-income countries due to their low cost.[180] TMP/SMX inhibit folic acid synthesis with distinct but synergistic mechanisms: TMP targets dihydrofolate reductase (DHFR) and SMX targets dihydropteroate synthetase (DHPS). Resistance is conferred by acquisition of variants of these enzymes with reduced TMP or SMX binding and is widespread globally.[180]

1.1.7.4 Evolution of antimicrobial resistance in pneumococci

The evolution and distribution of antimicrobial resistance in pneumococci is largely driven by horizontal genetic exchange between pneumococcal strains and with other non-pneumococcal commensal streptococci, resulting in widely distributed resistant lineages.[51,182,206] Penicillin resistant PBPs appear to be mosaic genes with sections acquired from horizontal genetic exchange with commensal streptococci in the niche, notably *S. mitis* and *S. oralis*.[207–210] These alleles have been disseminated through the pneumococcal population by further homologous recombination events,[51] which in some

cases coincided with a capsular switch event.[44,46,211] Macrolide resistance factors are found on large mobile genetic elements. The macrolide efflux genetic assembly (Mega), containing the *mefE/mel* operon, and *ermB* are both found on Tn*916*-like ICEs.[49,197,212] The ICEs appear to be initially acquired *via* an inter-species exchange event, and then spread through the pneumococcal population *via* transformation.[49] Tetracycline and chloramphenicol resistance genes are also found in Tn916-like ICEs.[51,213] ICEs have been implicated in the evolution of multi-resistant pneumococci, particularly of the highly successful PMEN1 lineage.[44,48]

1.1.8 Molecular typing

Bacteria of the same species can exhibit extraordinary sequence diversity due to high rates of SNP accumulation and as a result of horizontal gene transfer between individuals.[31] To add to the complexity, bacteria from different species can also share genetic material by horizonal exchange, blurring species groupings. When working with large genomic datasets, it is important to impose some order onto the dataset by grouping similar genomes together.[214] There are many approaches to clustering whole genome data by similarity, each with their own advantages:[214]

- Single nucleotide polymorphism (SNP) approaches catalogue single base differences relative to a reference genome and are therefore highly precise but also computationally expensive.[215,216]
- K-mer approaches compare the sequences of short stretches of sequence (k-mers) without using a reference genome, alignments, or annotations, and are faster but less precise than SNP approaches.[217]
- Bayesian approaches to genome clustering use machine learning to group genomes based on sequence alignments, generating reliable clusters. This

approach requires high levels of bioinformatic expertise and is therefore not widely accessible.[21,218,219]

- Gene-by-gene approaches catalogue diversity of assembled genomes based on the sequences of selected genes within the genome. These approaches are accessible and reproducible, and provide variable levels of precision depending on the scheme used.[220]

1.1.8.1 Multi-locus sequence typing

Multi-locus sequence typing (MLST) is a widely used gene-by-gene approach to molecular typing that pre-dates next generation sequencing data.[220] MLST characterises bacterial isolates by the alleles present for a specified number (typically seven) of housekeeping genes.[171] Observed alleles at each gene are assigned by a curator, and the combination of seven alleles is allocated a numerical sequence type (ST). These data are assigned and catalogued in the public PubMLST database (pubmlst.org). MLST results in a standardised genotyping nomenclature that is reproducible, and the data are easily stored and shared. Pneumococci can be further grouped into clonal complexes (CCs) of closely-related STs, using programmes such as Phyloviz, which groups isolates using the goeBURST algorithm.[221,222]

An advantage of MLST in highly recombinogenic organisms such as pneumococcus is that all alleles are considered to be equally different. This means that a single large recombination event, resulting in many sequence differences, is given the same weight as a single SNP in the gene. This is an advantage over other techniques, which may overestimate the evolutionary distance between strains that have undergone

transformation.[220] MLST has been widely adopted due to its accessibility, reproducibility, and effectiveness in describing bacterial populations.

1.1.8.2 Extended MLST schemes

With the increase in availability of whole genome sequences, different MLST schemes have been developed to make use of the valuable genomic data. In ribosomal MLST (rMLST), the 53 genes encoding protein subunits of the ribosome, which are conserved across bacterial genera, are characterised.[223] rMLST successfully differentiates bacterial species and can resolve groups within species in some circumstances, however due to high conservation of the genes used it is not suitable for discriminating similar members of the same species. For example, rMLST can successfully differentiate pneumococci from other streptococcal species, but does not reliably resolve different pneumococcal genetic lineages. For higher precision than 7-locus MLST, extended schemes have been developed that make use of hundreds or thousands of genes from the core genome or the whole genome (cgMLST and wgMLST).[220] These schemes exhibit much higher discriminatory power than standard MLST and rMLST by characterising a larger proportion of the genome. cgMLST defines a core genome, which is a set of genes common to all (or nearly all) strains of a species. In wgMLST all genes are identified in principle, but this can be very sensitive to low quality assemblies as it includes the accessory genome, which is highly variable between strains. A cgMLST scheme for pneumococcus is currently under development.

1.2 Bacteriocins

1.2.1 Bacteriocin overview

Bacteriocins are broadly defined as a ribosomally-synthesised peptides or small proteins produced by bacteria with the purpose of killing or inhibiting competitors in an ecological niche.[224] The producing strain is protected from the action of its own bacteriocin. Many bacteriocins have a narrow spectrum of activity against strains closely related to the producer,[225] although this is not always the case. Bacteriocins are widely distributed among eubacteria and do not appear to share a common evolutionary history; rather, they exhibit diverse synthesis pathways, mechanisms of action, and regulatory systems.[224]

The first 'bacteriocin' to be described was colicin, which was identified in *E. coli* in 1925.[226] Early bacteriocin studies therefore focussed on the colicins and related bacteriocins of Gram-negative species. The colicins are large, multi-domain proteins that are released from the producing bacteria by cell lysis and exhibit diverse mechanisms of antibacterial activity, from depolarization of the target cell membrane through pore formation to specific activity against key cellular processes (such as protein synthesis).[227–230] The colicins are exclusively found in Gram-negative species and are considered as a separate group to other bacteriocins. Accordingly, they will not be considered further in this thesis. Diverse small bacteriocins have been discovered in both Gram-positive and Gram-negative species, although in the latter they are often called microcins.[224] Historically the bacteriocins of these two groups have been considered as entirely separate,[231] although recently this distinction has broken down somewhat as more bacteriocins are described in a wider range of species.

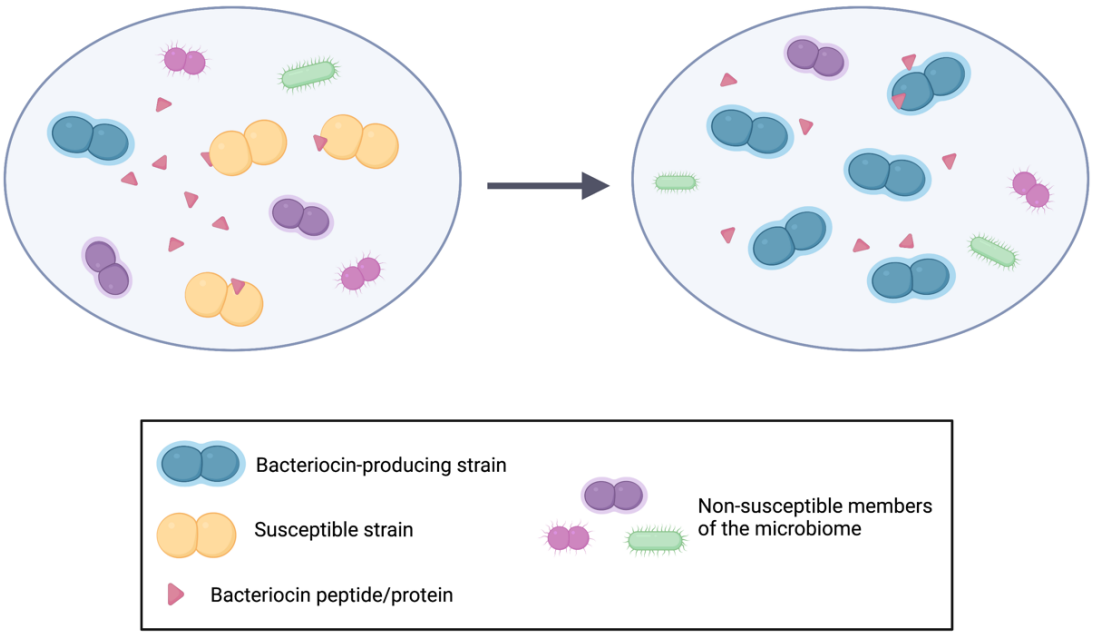

Figure 1.1: An illustration of a bacteriocin-producing strain (blue) competing with a susceptible strain (yellow) within an ecological niche.

1.2.1.1 Genetic organisation of bacteriocins

Bacteriocins are typically encoded on biosynthetic gene clusters, which encode the bacteriocin gene (or genes), as well as all the genes required for the modification and export of the bacteriocin.[224,232] Types of gene typically found in bacteriocin biosynthetic gene clusters include:

- Enzymes for the post-translational modification of the bacteriocin peptide
- Specialised export systems
- Immunity proteins that protect the producing strain from the antibacterial activity of its own bacteriocin
- Dedicated regulatory systems controlling the expression of other components of the cluster (often involving quorum sensing).[233]

1.2.1.2 Applications of bacteriocins

In light of increasing antimicrobial resistance, the development of novel antimicrobial compounds is urgently required, and bacteriocins are potentially valuable candidates for development. Common features of bacteriocins that make them promising targets include their heat stability and resistance to protease activity (often due to post-translational modifications) and the narrow target specificity observed in many bacteriocins, reducing 'collateral damage' to beneficial species in the microbiome.[234] Several bacteriocins have been identified as having clinical potential and some are in clinical trials.[235–237] Bacteriocins also have the potential to be engineered to alter target specificity or increase potency.[234,238] For example, the introduction of non-proteinogenic amino acids into lacticin 481 increased its inhibition of peptidoglycan synthesis *in vitro*.[239] Bacteriocins and bacteriocin-producing strains have been proposed as modulators of human microbiomes with the goal of preventing colonisation by potentially pathogenic bacteria.[233]

The second major application of bacteriocins is as food preservatives.[240] Many bacterial species involved in food production are known bacteriocin producers, in particular the lactic acid bacteria.[241] As these organisms are classified as generally regarded as safe, a bacteriocin-producing strain may be added to foods to prevent the growth of undesirable species. An alternative approach is to use an isolated bacteriocin with activity against an undesirable strain or species as a food additive: the lanthipeptide nisin has been successfully used as a food preservative since the 1960s.[242]

1.2.2 Bacteriocin classification

It has proven challenging to establish a precise definition of bacteriocins as their size, properties, mode of action, regulation and biosynthesis vary even between systems found in the same species.[232] Beyond this, recent genome mining investigations suggest that there is as yet unappreciated diversity among bacteriocins within bacteria.[243–245] The classification schemes discussed below exclusively classify ribosomally synthesised products, excluding the non-ribosomally synthesised antimicrobial peptides (such as lugdunin)[246] that are sometimes classed as bacteriocins.[233]

Various attempts have been made to develop a unified classification system for bacteriocins.[232,240,241] Broadly, bacteriocins have been grouped into Class I, which are post-translationally modified, and Class II, which are unmodified. Sub-groups within this scheme required revisions as more bacteriocins were discovered.[224,240] The most comprehensive scheme proposed to date by Arnison *et al.* places bacteriocins within the broader group of ribosomally synthesised and post-translationally modified peptide natural products (RiPPs), which are found across all Kingdoms of life and perform a broad range of functions.[232] Under this system, bacteriocin classes are defined based on shared biosynthetic pathways and modifications, rather than based on the organism or mode of antibacterial activity, and this classification was intended to be robust to future discoveries. Some relevant RiPP classes are outlined below.

1.2.2.1 Lanthipeptides

Lanthionine-containing peptides, or lanthipeptides, are a large and diverse group of RiPPs. Lanthipeptides contain lanthionine and methyllanthionine residues, which are introduced post-translationally.[232] The lanthionine residue comprises a thioether cross-

link between a cysteine residue and a non-adjacent serine or threonine residue,[247] resulting in two alanine residues joined into a ring within the peptide. A single lanthipeptide may contain multiple lanthionine modifications.[248] Lanthipeptides with antibacterial activity fall under the definition of bacteriocins and are termed lantibiotics.[249] The lantibiotics represent a large sub-group of lanthipeptides and tend to have broad-spectrum activity against Gram-positive species.[224,250,251]

The first lanthipeptide (and lantibiotic) to be identified was nisin, which was described in 1928 as an "inhibitory substance" produced by a *Lactococcus lactis* strain used in fermentation of dairy products.[252,253] Other notable lantibiotics include epidermin, which was identified in *Staphylococcus epidermis* in 1985,[249,254,255] and subtilin, which was discovered in *Bacillus subtilis* in 1973.[256] Both have sequence and structural similarities to nisin. The antibiotic activity of many lantibiotics, including nisin, involves disrupting cell wall biosynthesis by interacting specifically with lipid II *via* the lanthionine rings.[239,257–259] A subset of lipid II-binding lantibiotics, including nisin, epidermin and lacticin 3147, also form transient pores in the cell membrane, dissipating the membrane potential and leading to cell death.[260,261] Other lantibiotics target different molecules, for example cinnamycin and duramycin, which disrupt the cell membrane through an interaction with phosphatidylethanolamine.[262,263]

Lanthipeptide biosynthetic gene clusters encode at least one lanthipeptide precursor, biosynthetic enzymes, dedicated transport systems, immunity proteins, regulatory systems, and other accessory genes involved in lanthipeptide production.[264] In some lanthipeptides, such as the two-component lantibiotic lacticin 3147, the biosynthetic cluster contains multiple precursor genes.[265] Lanthipeptide export proceeds *via* an ABC

transporter,[266–268] and, in some systems, a distinct ABC transporter mediates lantibiotic immunity by removing the lantibiotic from its site of action.[269,270] Other lantibiotics use a separate binding protein to sequester the lantibiotic from its site of action.[271,272]

The modification enzymes are the basis for lanthipeptide classification:[232,273]

- Class I: A separate dehydratase (*lanB*) and zinc-dependent cyclase (*lanC*). Members include nisin, epidermin and subtilin.[249,274,275]
- Class II: A single bifunctional enzyme, *lanM*, with a dehydratase domain and a *lanC*-like cyclase domain. Members include cinnamycin and lacticin 481.[276–278]
- Class III: A single bifunctional enzyme, *lanKC*, with a lyase domain, a kinase domain, and a *lanC*-like cyclase domain.[279,280]
- Class IV: A single bifunctional enzyme, *lanL*, which exhibits the same domain organisation as *lanKC.*[281]

These biosynthetic enzymes are restricted to lanthipeptides, making them valuable tools for identification of putative lanthipeptides through genome mining.[282–284] Putative lanthipeptide clusters have been identified in an increasingly broad range of bacterial species, indicating that this group of RiPPs may not be restricted to Gram-positive species as previously assumed.[281,282,285]

1.2.2.2 Head-to-tail circularised peptides

Head-to-tail circularised peptides are joined at the N- and C-terminal amino acids *via* a peptide bond, resulting in a circular product with no exposed termini.[232] This modification results in peptides that are resistant to peptidase activity and are relatively heat and pH stable. There is little sequence homology between peptides of this group, but the circularised peptides appear to have similar structures of four to five α-helices

forming a saposin fold.[286–289] The circular bacteriocins also appear to share a mechanism of bactericidal activity: insertion in the target cell membrane resulting in pore formation.

Arnison *et al.* distinguish the group of head-to-tail circularised peptides from other groups of circular peptides by their relatively large size (typically 35-70 residues).[232] Enterocin AS-48, a bacteriocin produced by various species of *Enterococcus* with activity against both Gram-positive and Gram-negative species, including multi-drug resistant *Staphylococcus aureus,*[290] was the first member of this group to be identified.[289,291] Several other head-to-tail circular bacteriocins have also been identified, including uberolysin from *Streptococcus uberis,*[292] circularin A from *Clostridium beijerinck,*[293] and subtilosin A from *Bacillus subtilis*.[294] Circularised peptides have been sub-divided into three classes, all of which are found in Gram-positive species:

- Type I, the AS-48-like peptides, are generally cationic. This is the largest group and includes AS-48, uberolysin, and circularin A.
- Type II, the gassericin-like peptides,[295] are generally anionic or neutral.
- Type III is a divergent group of smaller circular bacteriocins with sactipeptide modifications. The best-studied member of this group is the *Bacillus subtilis* bacteriocin subtilisin A, which also falls into the class of sactipeptides (see Section 1.2.4.4).[241]

The mechanism by which the head-to-tail circularisation is achieved has not been full elucidated, although the proteins responsible are encoded on the biosynthetic gene clusters.[296] Identified clusters exhibit diverse gene content, but always include a single gene encoding the peptide precursor and one or more ABC transporters, membrane proteins of unknown function, and small immunity proteins.[297,298] The transporter genes

have been exploited in a genome mining study to identify putative circularised peptide biosynthetic gene clusters.[298]

1.2.2.3 Lasso peptides

The lasso peptides are characterised by an unusual 'slipknot' structure comprising an N-terminal macrolactam ring encircling the C-terminal region of the lasso peptide precursor.[299–302] The ring is joined by an isopeptide bond between the N-terminal residue and the side chain of an acidic residue typically in position 8 or 9.[303] The structure is stabilised by disulphide bonds or by steric hinderance from bulky side chains adjacent to the threaded region, resulting in heat stable and protease resistant peptides. Since its discovery in *E. coli* in 1992, microcin J25 (MccJ25) has become the most widely studied lasso peptide,[304] although the lasso peptide architecture was not elucidated until 2003.[299–301] Since then, the structures of many other lasso peptides have been solved.[305]

Lasso peptides that have been studied experimentally largely originated in actinobacteria and proteobacteria,[303] and the majority of these function as antimicrobials and therefore class as bacteriocins. MccJ25 has activity against Gram-negative species including clinical strains of *Salmonella* and *Shigella*, as well as other strains of *E. coli.*[304] Lassomycin and lariatins A and B (isolated from *Rhodococcus* sp. K01-B0171) both show specific antibacterial activity against mycobacterial species, including *Mycobacterium tuberculosis.*[306–308] The mechanisms of antibacterial lasso peptides vary. MccJ25 is a specific inhibitor of DNA-dependent RNA polymerase, and therefore shuts down transcription in target cells,[309] although it also interferes with the integrity of the cell membrane.[310] In contrast, lassomycin targets the ClpC1 ATPase, disrupting the essential ClpC1/ClpP1/P2 complex in mycobacterial species.[308]

Lasso peptide biosynthetic gene clusters typically encode the short peptide precursor (15-26 amino acids), specialised modification enzymes, and a dedicated transport system for export of the mature peptide.[303,311] Post-translational modification is mediated by two enzymes: a cysteine protease with sequence similarity to asparagine synthetase B, which cleaves the leader sequence of the precursor peptide and is unique to this class of RiPPs,[232] and an ATP-dependent lactam synthetase responsible for the formation of the macrolactam ring.[312,313] While the modification genes are conserved, there is variation in the gene composition of lasso peptide clusters.[303] As the availability of genomic data has increased, specialised genome mining tools have been developed for lasso peptide identification.[314,315] The number of putative lasso peptide biosynthetic clusters is now upwards of 1300,[315] and these putative clusters show a much wider taxonomic distribution than the experimentally studied systems.

1.2.2.4 Sactipeptides

The sactipeptides are a small group of RiPPs with one or more thioether cross-links between the sulphur group of a cysteine residue and the alpha carbon of a separate residue within the same peptide.[232] The characteristic cross-links are formed by S-adenosylmethionine (SAM) synthase enzymes, which utilise unusual iron-sulphur clusters to achieve the modification *via* a radical intermediate species.[316] Sactipeptide biosynthetic gene clusters encode one or, in the case of thurincin CD, two sactipeptide precursor genes with a leader peptide that is required for the modification and is cleaved during processing.[232] The clusters also encode SAM synthase enzymes, an ABC transporter complex believed to be involved in the export of processed sactipeptides, and small immunity genes.[317–319]

Sactipeptides that have antibacterial activity (which is most sactipeptides identified to date) are termed sactibiotics and class as bacteriocins. Early sactipeptides, including subtilosin A (see also section 1.2.2.2), thuricin H, and two-component thurincin CD,[294,319,320] were discovered in *Bacillus* species, but recent genome mining approaches are expanding the group into other Gram-positive genera including *Clostridium* and *Streptococcus.*[321–323] Sactibiotics tend to show high specificity in their target strains, leading Arnison *et al.* to speculate that their mechanism may involve an interaction with a cell surface receptor, although subtilosin A has been shown to interact with cell membranes leading to permeabilisation.[324]

1.2.2.5 Unmodified peptides

The Arnison *et al.* classification system does not account for peptides that are ribosomally synthesised but not post-translationally modified (beyond cleavage of a signal peptide). This highlights a flaw in the application of the RiPP classification system to bacteriocins: many known bacteriocins are unmodified, and this is even used as a category in some bacteriocin classification systems (Class II bacteriocins).[224,240] As these peptides are grouped based on a lack of modification, they do not necessarily share close evolutionary relationships, similar mechanisms of activity, or comparable bacterial species distribution.

1.2.3 Pneumococcal bacteriocins

1.2.3.1 Bacteriocin-like peptides

The Blp (bacteriocin-like peptide) systems are the most comprehensively studied pneumococcal bacteriocins. They were discovered in the early 2000s as putative

bacteriocin biosynthetic clusters under the control of a quorum sensing regulatory system.[325,326] Competition assays have been used to demonstrate the antimicrobial properties of various Blp peptides against a broad spectrum of bacteria, including some pneumococci, non-pneumococcal streptococci, and other Gram-positive species. BlpMN and BlpIJ are unmodified two-component bacteriocins,[327–329] and BlpK is also unmodified but is functional as a single peptide.[329] Studies in mouse models of carriage have suggested a role for Blp bacteriocins in nasopharyngeal colonisation.[327,328,330]

Studies in large genomic datasets have found that *blp* gene clusters are ubiquitous in pneumococcal populations, and that each isolate possesses a single *blp* cluster encoding the following genes:[329,331,332]

- Bacteriocin precursor genes, sometimes called pneumocins (such as *blpM* and *blpN*)
- Immunity genes for protection from the bacteriocin peptides
- Regulatory genes including a peptide pheromone (*blpC*), a histidine kinase (*blpH*), and a response regulator (*blpR*)
- A dedicated ABC transporter for the export of both the peptide pheromone and bacteriocins (*blpA* and *blpB*)

The bacteriocin precursor and immunity genes are found together in a region of the cluster termed the BIR (bacteriocin immunity region). Expression of the entire cluster is regulated by the sensing of extracellular BlpC, which triggers bacteriocin production upon reaching a threshold concentration.[325,326]

As more detailed genomic analyses have become possible, it has become clear that *blp* gene clusters show remarkable diversity in gene content and organisation. 16 putative

bacteriocin peptides have been identified in total, and clusters have been observed with up to six bacteriocin genes within the BIR. The immunity genes found within the BIR vary according to the bacteriocin genes.[332] The pheromone/receptor pair (BlpC and BlpH) shows additional diversity and varies independently of the BIR.[333] BlpH is a highly specific receptor, resulting in multiple distinct 'pherotypes' that respond exclusively to specific BlpC alleles.[334] The combination of specific signalling systems and diverse bacteriocin repertoires may result in highly complex competition dynamics between pneumococci within the nasopharynx.

Adding to this complexity, disrupted *blp* clusters have been observed,[333] which carry a conserved frameshift mutation in the dedicated bacteriocin transporter gene *blpA*, resulting in a truncated, and presumably non-functional, product. This was thought to prevent the export of the BlpC pheromone and of the bacteriocin peptides encoded in the BIR but allow strains to respond to exogenous BlpC by expressing the immunity genes from the BIR. Therefore, these strains were hypothesised 'cheaters': they avoid the cost of pheromone and bacteriocin export while taking advantage of the neighbouring strains that are exporting the bacteriocin. In mixed populations, the presence of cheater strains adds further complexity to the competition dynamics.[335] However, this has been complicated by a subsequent study: there appears to be substrate redundancy between the Blp exporters and the competence pheromone exporters, so pneumococci with disrupted *blpA* genes may in fact still export functional bacteriocin toxins.[336]

1.2.3.2 Other experimentally confirmed bacteriocins

The competence-induced bacteriocin, Cib, is an unmodified bacteriocin involved in inter-strain predation in pneumococcus.[337] Expression of Cib is triggered by the competence

system. The bacteriocin lyses neighbouring pneumococci, releasing their genomic DNA for internalisation by the competent producing cell, which can then acquire advantageous genes by homologous recombination.

The first lanthipeptide identified experimentally in pneumococcus, originally called pneumolancidin (but also referred to as streptolancidin A), was detected in competition assays in a pneumococcal strain with a notable ability to out-compete other pneumococci.[338,339] The biosynthetic gene cluster encodes multiple (four or five) lanthipeptide precursors with a *lanM* family lanthionine modification enzyme, making pneumolancidin/streptolancidin A a class II lanthipeptide. This lanthipeptide has also been identified in *Streptococcus salivarius*, where it is named salivaricin E.[340]

Other experimentally studied pneumococcal lanthipeptide gene clusters were identified as sites of regulation by the TrpA/PhrA quorum sensing system. Genes of a putative lanthipeptide cluster, later named streptolancidin G,[339,341] were found to be expressed using RNA sequencing. Similarly, a second lanthipeptide biosynthetic gene cluster, later named streptolancidin B, has been identified by regulation by TrpA2/PhrA2 and by its presence on a pneumococcal ICE.[48,339,342] The function of these lanthipeptide products have not yet been confirmed experimentally, although both are putatively classed as bacteriocins.

1.2.3.3 Identification of pneumococcal bacteriocins by genome mining

An early genome mining study identified a putative type II lanthipeptide gene cluster, later named streptolancidin E.[282,339] This study identified clusters based on the presence of a *lanM* gene and was therefore limited to discovery of class II lanthipeptides. The

second cluster discovered through genome mining was streptocyclicin (previously named pneumocyclicin), which was discovered during a thorough genomic characterisation of the Blp clusters by the Brueggemann group.[331] The most comprehensive genome mining study to date was performed by the Brueggemann group and published in 2018 by Rezaei Javan *et al*. This study screened a diverse, historical genomic dataset for any bacteriocin biosynthetic gene clusters using bacteriocin databases and the identification tools antiSMASH, BACTIBASE and BAGEL.[343–345] This study identified:

- 11 lanthipeptide gene clusters, four of which had been described previously
- Five unmodified bacteriocin clusters with homology to the *Lactococcus lactis* unmodified bacteriocin lactococcin 972
- One novel lasso peptide cluster
- One novel sactipeptide cluster
- The previously described head-to-tail circularised bacteriocin gene cluster

These findings represented a huge increase in the number of recognised bacteriocin clusters in pneumococcus, so a new naming system was proposed to handle the complexity and to prevent confusion in the field. As many of the clusters had homologs in other streptococcal species, the 'strepto' prefix was used in the new naming system.

In total, 21 putative and confirmed bacteriocin biosynthetic gene clusters have been identified in pneumococcus to date, using a combination of genome mining and experimental approaches (summarised in Table 1.1). In this thesis, I have studied 20 of these, excluding Blp because of the more advanced state of research into this bacteriocin compared to the others,[327,329] and due to the high complexity of the Blp biosynthetic gene clusters in pneumococcal genomes.[331]

Table 1.1: Complete list of bacteriocin biosynthetic gene clusters identified in pneumococcus experimentally and through genome mining.

Bacteriocin cluster	Aliases	Class	Method of identification	Cluster size (kb)[a]	Number of genes
Streptococcin A	-	Unmodified (Lactococcin 972-like)	Genome mining[339]	3.1	3
Streptococcin B	-	Unmodified (Lactococcin 972-like)	Genome mining[339]	3.0	3
Streptococcin C	-	Unmodified (Lactococcin 972-like)	Genome mining[339]	3.1	3
Streptococcin D	-	Unmodified (Lactococcin 972-like)	Genome mining[339]	3.0	3
Streptococcin E	-	Unmodified (Lactococcin 972-like)	Genome mining[339]	3.0	3
Streptocyclicin	Pneumocyclicin	Head-to-tail circularised	Genome mining[331]	3.0	5
Streptolancidin A	Pneumolancidin Salivaricin E	Lanthipeptide (class II)	*In vitro* competition assays[338,340]	11.6	*11*
Streptolancidin B	IcpAMT ICESp23FST81	Lanthipeptide (class II)	TrpA/PhrA regulation[342] ICE cargo[48]	8.2	6
Streptolancidin C	-	Lanthipeptide (class IV)	Genome mining[339]	3.9	4
Streptolancidin D	-	Lanthipeptide (class I)	Genome mining[339]	5.6	4
Streptolancidin E	SP23-BS72 lantibiotic	Lanthipeptide (class II)	Genome mining[282]	10.9	10
Streptolancidin F	-	Lanthipeptide (class IV)	Genome mining[339]	2.6	2
Streptolancidin G	Phr lantibiotic	Lanthipeptide (class II)	TrpA/PhrA regulation[341]	9.4	7
Streptolancidin H	-	Lanthipeptide (class I)	Genome mining[339]	15.5	13
Streptolancidin I	-	Lanthipeptide (class I)	Genome mining[339]	13.2	11

Streptolancidin J	-	Lanthipeptide (class IV)	Genome mining[339]	8.9	7
Streptolancidin K	-	Lanthipeptide (class IV)	Genome mining[339]	4.4	3
Streptolassin	-	Lasso peptide	Genome mining[339]	8.5	9
Streptosactin	-	Sactipeptide	Genome mining[339]	4.3	6
Cib	-	Unmodified	*In vitro* competence-mediated[337]	0.5	3
Blp	Pneumocins	Unmodified	*In vitro* competition assays[327] Genome mining[331]	Variable	Variable

Note: Bacteriocins names according to the nomenclature proposed in Rezaei Javan et al., aliases given where bacteriocins were previously identified under a different name. Cib: competence induce bacteriocin, Blp: bacteriocin-like peptide.

a. Cluster size includes non-coding intergenic regions and is stated to the nearest 0.1 kilobase (Kb).

1.3 Whole genome sequencing

1.3.1 History of sequencing

DNA sequencing technologies were developed following the discovery of the structure of DNA and the genetic code.[346] Sanger sequencing (also known as chain termination or dideoxy sequencing) was a Nobel-prize winning technique developed by Frederick Sanger in 1977.[347] It uses labelled dideoxynucleotides (ddNTPs) in DNA polymerisation reactions to detect the order in which bases are added to the synthesised strand. Originally, the ddNTPs were radiolabelled and the sequence read-out used polyacrylamide gel electrophoresis. More recently, fluorescent ddNTP labels and capillary electrophoresis improved the efficiency of Sanger sequencing.[346] The maximum length of DNA that can be sequenced by Sanger sequencing is around 1 Kb, so whole genome sequences must be divided into small enough sections and sequenced separately (shotgun sequencing), before being reassembled using overlapping sequence.[348]

Sanger sequencing was successfully used to obtain the first whole genome sequences, starting with the *Haemophilus influenzae* genome in 1995,[349] the first draft human genome and the first pneumococcal genome in 2001.[25,350] Although undoubtedly a major step forwards, early whole genome sequencing was limited by the short read length and low throughput capabilities of sequencing platforms.

1.3.2 Next generation sequencing

Next generation sequencing (NGS) refers to the set of technologies that superseded classic Sanger sequencing, offering much higher throughput and lower per-base cost of whole genome sequencing.[351] The most widely used technologies to date are short-read

platforms, but in recent years long-read sequencing technologies have become more practical and are increasingly used.[346] As of May 2022, over 33,000 assembled pneumococcal whole genome sequences were publicly available in the pneumococcal genome library (pubmlst.org/organisms/streptococcus-pneumoniae/pgl), the vast majority of which were sequenced using Illumina technology. All short read sequencing data must be assembled *in silico*, as in Sanger shotgun sequencing, but the high throughput capacity of next generation sequencing makes it more practical to generate the great number of short reads required to get coverage across a whole genome sequence of several Mb.

1.3.2.1 Short-read sequencing platforms

Short-read NGS technology was developed in the 2000s, and many platforms have been developed, starting with pyrosequencing (later called 454 sequencing).[346,352] This method indirectly detects pyrophosphate generation from the incorporation of a dNTP into a synthesised DNA strand using the luminescent enzyme luciferase.[353] In the 454 platform, clonal DNA fragments are immobilised on microbeads and each base is supplied in turn; when light is detected, this indicates that the base was incorporated. This approach relies on the light generated by luciferase being proportional to the amount of pyrophosphate present, and therefore number of bases incorporated. However, this becomes unreliable in homopolymeric tracts of more than four base pairs.[346]

In recent years Illumina technology (previously called Solexa) emerged as the dominant platform, offering improved accuracy and reduced costs over other technologies.[351,354] Illumina has achieved higher throughput by immobilising DNA in a flow cell. Each individual fragment of DNA is cloned using bridge amplification, resulting in 100-200

million clusters of clonal DNA, which are sequenced simultaneously on the same flow cell.[352] Illumina technology differs from 454 sequencing by using fluorescently labelled dNTPs that are detected using total internal reflection fluorescence microscopy.[355] The fluorophore blocks sequencing of the complementary strand, and removal of the fluorophore allows the next labelled dNTP to be incorporated and detected.

1.3.2.2 Genome assembly

A challenge in short-read sequencing technologies is computationally resolving the original genome sequence from the multitude of short reads.[356–358] This process is referred to as genome assembly. If a high-quality genome of the organism is available, genomes can be assembled by mapping short reads onto the reference genome. This is a straightforward concept and is computationally relatively simple, it is therefore quick and accessible. However, mapping to a reference genome introduces biases into the assembly. By definition, sequences can only be detected if they are present in the reference genome, and the order of sequences in the final genome will depend on their order in the reference. The approach is therefore more suitable in species with conserved genome content and is less well suited to species such as pneumococcus that exhibit high genome plasticity and a diverse accessory genome.

The alternative approach is *de novo* assembly, where the whole genome sequence is assembled using only information contained within the short reads.[359] There are various computational assembly programmes, such as Velvet and SPAdes,[360,361] which use the overlapping regions between adjacent reads to unambiguously resolve their order. This relies on high depth of sequencing at each nucleotide position. Some short-read NGS technologies (including Illumina) use paired-end reads.[351] Following the sequencing of a

template strand, its reverse complement is sequenced from the opposite end, generating a pair of reads whose positions are known relative to each other. Paired end read information can be used to improve assemblies using scaffolding software such as SSPACE and GapFiller.[362,363]

1.3.2.3 Limitations of short read sequencing

Even after scaffolding, *de novo* genome assembly generates draft genomes comprising several long contigs with an unknown order. Further sequencing is required to resolve these gaps in sequencing and generate a 'complete' whole genome sequence, which is not practical in large genomic datasets. When annotating genes in draft genomes, some genes will inevitably be interrupted by contig breaks, resulting in loss of sequence data. Compounding this issue, the presence of repetitive genes results in issues during *de novo* genome assembly: where the same gene is present more than once, as is often the case in large mobile genetic elements such as ICEs and prophages, the short reads cannot be unambiguously assembled into a contiguous sequence, resulting in further contig breaks.

1.3.2.4 Long read sequencing

Sometimes referred to as 'third generation sequencing', long read platforms offer further advantages over the short read platforms discussed above.[346] These technologies aim to retain the efficiency and accuracy of Illumina-like technologies but remove the need to fragment genomes into short sections. This reduces the potential for errors introduced in the *de novo* assembly procedure, particularly in the assembly of repetitive genes within large mobile elements. Long sequence reads can be expected to span the whole length of any repeated genes, allowing a full, unambiguous sequence of the region. The major long read technologies are Pacific Biosciences (PacBio, also known as SMRT sequencing) and

Oxford Nanopore, both of which sequence a single molecule of DNA.[352] In PacBio sequencing, DNA is circularised using hairpin adaptors, and replicated using an immobilised DNA polymerase.[364] As in Illumina sequencing, sequence detection utilises fluorescently labelled dNTPs. Oxford Nanopore takes a different approach: sequencing is not dependent on synthesis of a complementary strand, and instead makes use of electrical charges generated when a DNA strand is transported through a biological pore.[365]

Long read sequencing has typically had a much higher per-base error rate compared to short read technologies. To overcome this, a combined long-read and short-read approach has been used successfully in bacterial genomics.[366] The two technologies are complimentary, with long reads overcoming the issues of assembly over repetitive regions, and short reads overcoming low accuracy in long reads. More recently, long read technologies have achieved higher accuracy and higher throughput, and it is becoming practical to generate large genomic datasets using long-read sequencing alone.

1.4 Thesis outline and aims

In this thesis, I investigated the pneumococcal bacteriocins previously identified by genome mining, in order to gain a better understanding of their role in pneumococcal competition. This was achieved using a combination of genomic and experimental approaches.

In Chapter 3, the distribution of previously discovered bacteriocin gene clusters was investigated in both carriage and disease-causing pneumococci recovered from Iceland

and Kenya, in the context of PCV introduction in the populations. Significant differences in bacteriocin distribution were observed, suggesting that different subsets of the pneumococcal population exhibit distinct competition dynamics.

In Chapter 4, I used structural and functional predictions to generate a model of streptococcin function. This model predicts that streptococcin gene clusters encode a single toxin and an ABC transporter with a putative role in immunity. This model informed further genomic analysis of the streptococcins presented in Chapter 5. Diversity was observed in the composition of streptococcin gene clusters, and this diversity likely has phenotypic consequences. Streptococcin distribution in pneumococcal and streptococcal genomes was indicative of horizontal exchange of individual streptococcin genes and whole clusters.

Chapter 6 describes a new procedure for the recombinant expression and subsequent purification of a streptococcin, and a protocol for screening strains for susceptibility to the streptococcin toxin. Preliminary assay results are presented, indicating that the isolated streptococcin may inhibit some pneumococci and commensal streptococci.

2 General Methods

2.1 Genomic datasets

The Icelandic dataset analysed in this thesis was generated in collaboration with Professors Helga Erlendsdóttir, Ásgeir Haraldsson, and Karl Kristinsson at the University of Iceland. Microbiology, DNA extractions and serotyping were performed by Dr Sigríður Quirk. Sequencing was performed at the Sanger Institute and quality control of the assemblies was performed by Dr Andries van Tonder as part of his doctoral work in the Brueggemann group.[27] The Kenyan dataset was generated in collaboration with Professor Anthony Scott's research group in Kilifi. Microbiology was performed by Dr Angela Karani, Benedict Mvera, and Donald Akech. DNA extraction was performed in the Brueggemann laboratory at Imperial College London by Dr Asma Aktar and Dr Calum Forest, and whole genome sequencing was performed at the Sanger Institute. Quality control of genome assemblies was performed by Dr Melissa Jansen van Rensburg and me. The non-pneumococcal streptococcal dataset was compiled by Dr Melissa Jansen van Rensburg and Femke Ahlers. European nucleotide archive (ENA) accession numbers and/or BioSample accession numbers for genomes included in all datasets are provided at github.com/mebbutler/thesiscode. Assembled genomes are currently stored in a private database, access to which can be provided on request.

2.1.1 BIGSdb

Whole genome sequences were stored in the Brueggemann group private Streptococcal database using Bacterial Isolate Genome Sequence Database (BIGSdb) software.[367]

BIGSdb provides a web-hosted platform for the storage of complete or draft whole genome sequences generated by any sequencing platform. Each genome is stored as a record in an isolate database with associated metadata (such as the source of the genome and associated microbiological data including the serotype and antimicrobial susceptibility). The databases are highly customisable for use with any bacterial species. Genomic data can be analysed within the database environment using a range of analysis tools and plugins.

BIGSdb software was designed to support gene-by-gene typing schemes such as MLST by facilitating the annotation of genomic loci within the database framework.[220,368] Annotations are associated to isolate records as an allele identification number (describing the sequence of the locus) and a set of sequence coordinates (describing the position of the annotation). Allele sequences are stored in a separate sequence definitions database. Multiple isolate databases can therefore be linked to the same sequence database, allowing the annotation of the same loci in multiple separate databases. The procedure to annotate a genomic locus is described in more detail in Section 2.2.

2.1.2 The Kenyan genomic dataset

A Kenyan pneumococcal whole genome sequence dataset was generated in collaboration with Professor Anthony Scott's research group at the KEMRI-Wellcome Trust Research Programme in Kilifi using isolates collected from residents of the Kilifi Health and Demographic Surveillance System (KHDSS) study area.[159] All genomes in the Kenyan dataset have not been published.

2.1.2.1 Pneumococcal sampling

Kenyan carriage pneumococci were collected from all ages in multiple studies between 2004 and 2017 (excluding 2011, when PCV10 was introduced). Invasive pneumococci were collected from patients presenting at Kilifi County Hospital with IPD between 2003 and 2017. The pneumococci to be sequenced were sampled randomly from the age groups represented by the carriage studies: 3-5, 6-11, 12-23 and 24-50 months, 5-14, 15-64 and 65+ years. Sampling within each age stratum was weighted to reflect the observed population structure of residents of the KHDSS study area. All Kenyan pneumococci recovered from invasive disease in the study period were selected for sequencing.

2.1.2.2 DNA extraction, sequencing, and assembly

Pneumococci for sequencing were recovered from freezer stocks, cultured to standard blood agar plates, and incubated overnight at 37°C with 5% CO_2. Isolates were checked for purity, and viridans streptococci were excluded using optochin disk susceptibility. Suitable isolates were grown overnight in brain-heart infusion broth at 37°C with 5% CO_2. DNA extraction was performed using the Maxwell 16 Buccal Swab LEV DNA Purification kit and the Maxwell 16 instrument in LEV Research Mode (Promega), following manufacturer's instructions, and eluting in 50 µL of extraction buffer. DNA quantification was performed using a Qubit fluorometer and Quant-iT dsDNA Broad-Range Assay Kit (Thermo Fisher Scientific).

Library preparations and sequencing were performed at the Sanger Institute using the Illumina HiSeq2000 platform. Draft genomes were assembled *de novo* into contigs from the short paired-end reads using a computational pipeline that utilised the Velvet assembler and SSPACE and GapFiller for scaffolding.[359,360,362,363]

2.1.2.3 Quality control

Genomes assemblies were assessed using a quality control approach developed by Dr Melissa Jansen van Rensburg. This approach made use of sequence assembly statistics such as overall length of the assembled genome, the number of contigs, the overall GC content, and the N50 value (the length of the contig at 50% of the overall genome length, describes the distribution of the contig lengths). rMLST locus annotations were used to identify and exclude genomes derived from non-pneumococcal species.[223] Genomes were flagged for manual investigation if any of the assembly statistics were more than two standard deviations from the mean of the dataset. Genomes derived from multiple isolates were identified by mixed alleles at MLST and rMLST loci.

2.1.2.4 Molecular typing

The seven pneumococcal MLST loci were curated and new alleles and STs were uploaded to PubMLST.[368] CCs were defined using the goeBURST algorithm in Phyloviz[221,222] as groups of STs that were single locus variants (SLVs) from each other. Each CC was named for its founder ST. The two largest CCs were separated into smaller CCs due to distinct CCs becoming joined to one another by long chains of SLVs. Singletons were defined as STs that were not SLVs to any other ST. All STs in PubMLST as of 17/10/2019 were assigned to a CC named for its founder ST (the ST with the most SLVs within the CC), as determined by Phyloviz.

2.1.2.5 Duplicate removal

90 pairs of invasive isolates that had been sequenced twice were identified during database metadata uploads. These isolates were recovered from the same case of invasive disease, on the same date, but from different sources, most commonly blood and

cerebrospinal fluid, and less commonly blood and pleural fluid. In all cases, the genome derived from blood was excluded.

2.1.3 The Icelandic genomic dataset

A dataset of whole genome sequences from carriage and disease pneumococci recovered in Iceland was generated as part of an earlier collaborative project.[27,163,169] Genomes in this dataset were recovered from carriage, lower respiratory tract infections, and invasive disease. Genomes of pneumococci recovered from carriage, otitis media and lower respiratory tract infections in the Icelandic dataset have been published previously.[27,163,169] Genomes from invasive pneumococci were sequenced as described for the non-invasive genomes. The quality of all the Icelandic genome sequences was assessed previously.[27] Following the definition of CCs in the Kenyan dataset, STs observed in the Icelandic dataset were cross-checked with the updated CCs, and if the ST was now assigned to a different CC the record was updated within BIGSdb.

2.1.4 The non-pneumococcal streptococcal genomic dataset

A dataset of genomes from non-pneumococcal streptococci (NPS) was generated in the Brueggemann group using whole genome sequences in the publicly available rMLST database. Genomes were included from a species list including viridans streptococci (which are members of the oral and nasopharyngeal microbiomes) and important streptococcal pathogens (including human pathogens *S. pyogenes* and *S. agalactiae* and animal pathogens such as *S. suis* and *S. equi*). Genome quality was assessed using a quality control procedure as described above (Section 2.1.2) and low-quality genomes were not considered for inclusion in the dataset. To prevent skewing the dataset heavily towards any single species, an upper cap on the number of genomes to include from each species

was set at 180. Where more than 180 genomes were available, genomes were selected to maximise genetic diversity using rMLST.

2.2 Bacteriocin gene annotation

The bacteriocin gene annotation approach described below was developed with Dr Melissa Jansen van Rensburg.

2.2.1 Standard procedure

2.2.1.1 BIGSdb for gene annotation

BIGSdb supports annotation of sequences within the database environment. Whole genome sequences in a BIGSdb isolate database can be screened for genetic loci of interest using the basic local alignment search tool (BLAST, see Section 2.3.1) against a set of previously observed sequences (alleles) stored in the linked sequence database. Every unique sequence of the locus is stored as a separate allele, and alleles may represent pseudogenes (*i.e.* genes with disruptions to the coding sequence). Annotations within a BIGSdb database comprise an allele designation, referring to the unique identification number of the detected allele, and a set of sequence coordinates, describing where in the genome the designated allele was detected. When no sequence is detected for a locus, the allele designation is given as '0' to indicate that the genome has been screened and that the locus was not detected, distinguishing it from genomes that have not been screened.

Prior to this project, bacteriocin loci were created in a BIGSdb sequence database for the loci associated with bacteriocin biosynthetic gene clusters identified in the previous genome mining studies (Table 2.1).[339] Allele sequences from the previously published reference bacteriocin gene clusters were used as the first allele recorded at each locus (allele 1).

2.2.1.2 Automated and manual scanning

Genomes in a BIGSdb database are scanned for genes of interest using BLAST within the database software, returning partial or exact hits to previously defined alleles (Figure 2.1). Exact hits are identical to an allele and partial hits are not identical but are within thresholds of similarity (length of hit, overall percentage identity) that are set for each scan. When annotating the bacteriocin loci, the initial thresholds used were 70% sequence identity over 50% of the length of the locus. When partial hits are observed, the sequences of these hits are automatically extracted within BIGSdb. New alleles can be batch added to the sequence database with filters to prevent duplicate sequences or sequences containing Ns from being added to the database. Additionally, any sequences that do not represent complete coding sequences can be filtered out at this stage. This allows for the manual investigation of atypical sequences and incomplete coding sequences (described below).

When all the matches to existing alleles have been processed, the scan thresholds can be reduced to allow the detection of shorter matches, or of matches with a lower percentage identity. It is convenient to start with higher thresholds so that the less ambiguous matches can be processed before any more complicated examples are investigated. During bacteriocin annotation, the thresholds were lowered to 50% sequence identity

over 30% of the locus length. Any genomes without a good match at these thresholds were assigned an allele designation of '0', as the locus was not detected in the genome.

2.2.2 Atypical sequences

2.2.2.1 Contig breaks

Draft genomes are comprised of multiple assembled contigs, and loci of interest can be interrupted by a sequence break and thus be incompletely assembled at the end of a contig. Even if the rest of the locus is detectable on a separate contig, we cannot necessarily conclude that the whole sequence of the locus has been assembled across the two contigs. When a bacteriocin locus was interrupted by a contig break, the largest contiguous section of the locus was tagged, but no allele was designated, and the sequence was labelled as incomplete.

2.2.2.2 Ambiguous reads

Ambiguous reads, indicated by an N, indicate a position in the sequence that could not be assigned a base. Ambiguous reads can occur as a single isolated N, but more commonly in scaffolded assemblies they occur as long stretches and are used when the relative position of two contigs is known but the sequence joining them could not be unambiguously assembled ('gaps'). When a bacteriocin locus was identified with one or more ambiguous reads, the whole sequence, including the ambiguous reads, was tagged, but no allele was designated. The sequence was labelled as incomplete and as containing ambiguous reads.

2.2.2.3 Incomplete coding sequences

In cases where the locus was detectable above the scan thresholds but did not represent a complete coding sequence, the sequence was investigated manually by aligning it with typical alleles of the locus and by investigating flanking sequences for evidence of the disruption. Prokka was used to annotate flanking sequences where required.[369] In cases where issues with the sequence assembly were suspected to cause an apparent disruption to the locus, the raw sequencing reads were aligned with the assembled contig to assess the quality of the assembly in this region.

Disruptions to the coding sequence could result from SNPs leading to internal stop codons or the loss of cognate stop or start codons, single base pair insertions/deletions resulting in frameshifts, or from large sequence insertions or deletions. The whole region with sequence homology to previously defined alleles was used to define the new allele. Where a different sequence had been inserted into the locus, if homologous sequence from the locus was detected on both sides of the inserted sequence, the whole region was defined as an allele, including the inserted sequence. Disrupted sequences (pseudogenes) were given flags to describe the disruption.

2.2.2.4 Type alleles

The definition of some disrupted sequences as alleles can lead to issues with future annotation of the locus. This is particularly the case where alleles are truncated or when a whole inserted sequence is defined with the allele. In the case of truncated alleles, future scans may result in hits to the truncated allele that mask hits to novel full-length alleles (where the truncated allele is a sub-sequence of the novel full-length one). Where an allele contains a different inserted sequence, future hits may be to the inserted sequence

rather than to the parts of the allele representing the locus to be annotated. These issues are avoided by the definition of type alleles. Scans can be performed where only hits to the previously designated type alleles are returned. Any problematic alleles, as described above, are not defined as type alleles.

2.2.3 Paralogous loci

The majority of annotated bacteriocin loci were detected no more than once per genome. However, streptolancidin A and J clusters each contain genes that are indistinguishable on a sequence level (*slaA1* and *slaA2*, *sljA1* and *sljA2*). Additionally, two genes from streptolancidin E clusters (*sleT* and *sleX1*) were found as a small fragment at a different location within the pneumococcal genome that was sometimes found in addition to a typical streptolancidin E cluster. Because the standard annotation method does not take genomic location into account, and therefore cannot distinguish loci with highly similar or identical sequences in different locations, these loci required an adapted protocol that utilised an *in silico* 'hybridisation' approach to annotate the genes according to proximity to a known ('probe') sequence. Both streptolancidin A and J genes were annotated, and the streptolancidin E genes were annotated only if they were found as part of a typical streptolancidin E cluster.

2.2.4 Assessment of whole clusters

The annotated loci from each bacteriocin biosynthetic gene cluster were assessed together. If fewer than half of the expected loci were detected, the cluster was categorised as a fragment cluster. If the loci were not located adjacent to one another within the genome, the cluster was categorised as non-contiguous, as described in section 2.3.2.2. Fragment and non-contiguous clusters were excluded from analysis.

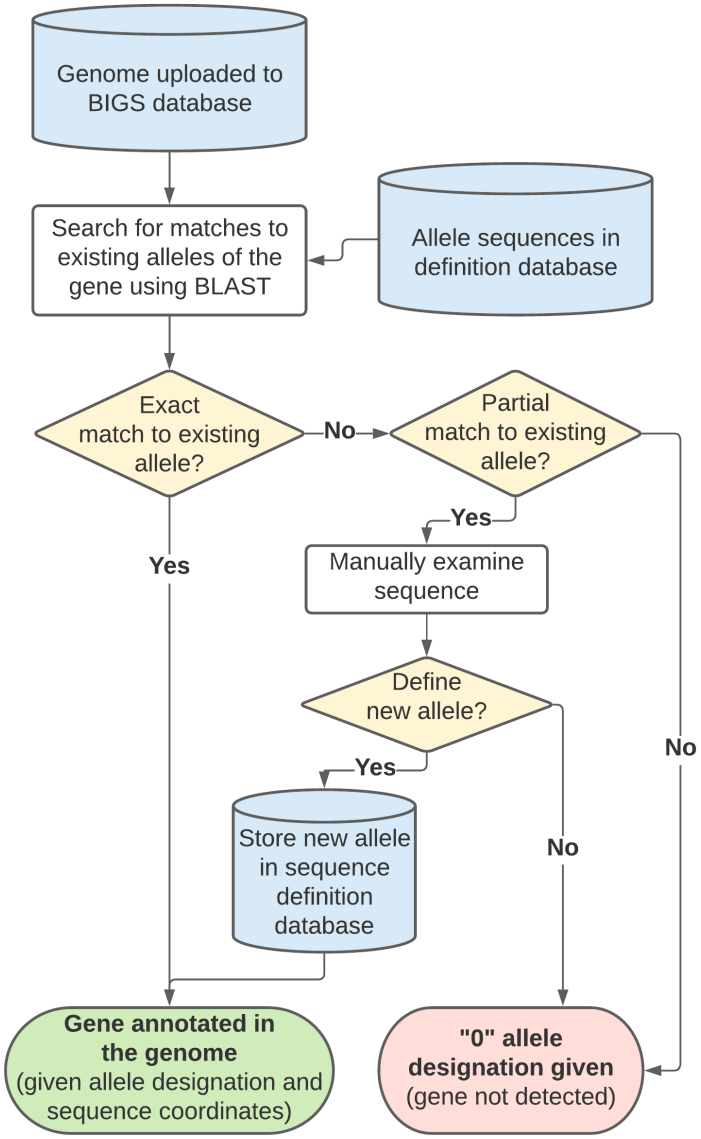

Figure 2.1: Flow chart illustrating the bacteriocin gene annotation procedure.

Table 2.1: List of all bacteriocin genes annotated in the Icelandic and Kenyan datasets, including the predicted functions of the gene products.

Bacteriocin	Gene	Typical length (bp)	Predicted function
Cib	*cibA*	186	Bacteriocin precursor
	cibB	153	Bacteriocin precursor
	cibC	198	Immunity
Streptococcin A	*scaA*	285	Bacteriocin precursor
	scaB	2109	Immunity
	scaC	642	Immunity
Streptococcin B	*scbA*	297	Bacteriocin precursor
	scbB	2031	Immunity
	scbC	642	Immunity
Streptococcin C	*sccA*	348	Bacteriocin precursor
	sccB	2022	Immunity
	sccC	642	Immunity
Streptococcin D	*scdA*	297	Bacteriocin precursor
	scdB	2010	Immunity
	scdC	633	Immunity
Streptococcin E	*sceA*	297	Bacteriocin precursor
	sceB	2016	Immunity
	sceC	633	Immunity
Streptocyclicin	*scyA*	297	Bacteriocin precursor
	scyB	1137	Bacteriocin biosynthesis
	scyC	483	Bacteriocin biosynthesis
	scyD	597	Bacteriocin biosynthesis
	scyE	492	Bacteriocin biosynthesis
Streptolancidin A	*slaA1*	183	Bacteriocin precursor
	slaA2	183	Bacteriocin precursor
	slaA3	183	Bacteriocin precursor
	slaA4	177	Bacteriocin precursor
	slaA5	108	Bacteriocin precursor
	slaF	738	Immunity
	slaE	2016	Immunity
	slaK	1575	Histidine kinase/response regulator
	slaR	597	Histidine kinase/response regulator
	slaM	2955	Bacteriocin biosynthesis
	slaT	2124	Transporter
Streptolancidin B	*slbF*	936	Immunity

	slbG	741	Immunity
	slbE	729	Immunity
	slbA	216	Bacteriocin precursor
	slbM	3252	Bacteriocin biosynthesis
	slbT	2067	Transporter
Streptolancidin C	*slcA*	105	Bacteriocin precursor
	slcX	993	Unknown
	slcL	1437	Bacteriocin biosynthesis
	slcT	1239	Transporter
Streptolancidin D	*sldA*	105	Bacteriocin precursor
	sldB	2793	Bacteriocin biosynthesis
	sldC	1278	Bacteriocin biosynthesis
	sldT	1257	Transporter
Streptolancidin E	*sleM1*	3036	Bacteriocin biosynthesis
	sleA1	171	Bacteriocin precursor
	sleA2	192	Bacteriocin precursor
	sleM2	2016	Bacteriocin biosynthesis
	sleM3	747	Bacteriocin biosynthesis
	sleT	2142	Transporter
	sleX1	171	Unknown
	sleF	729	Immunity
	sleG	738	Immunity
	sleX2	711	Unknown
Streptolancidin F	*slfA*	99	Bacteriocin precursor
	slfL	2496	Bacteriocin biosynthesis
Streptolancidin G	*slgA1*	225	Bacteriocin precursor
	slgA2	189	Bacteriocin precursor
	slgM	2991	Bacteriocin biosynthesis
	slgD	705	Bacteriocin biosynthesis
	slgP1	930	Bacteriocin biosynthesis
	slgT	2109	Transporter
	slgP2	1740	Bacteriocin biosynthesis
Streptolancidin H	*slhP*	1368	Bacteriocin biosynthesis
	slhR	684	Histidine kinase/response regulator
	slhK	1341	Histidine kinase/response regulator
	slhF	687	Immunity
	slhE	741	Immunity
	slhG	678	Immunity
	slhX1	645	Unknown

	slhX2	285	Unknown
	slhA	177	Bacteriocin precursor
	slhB	2964	Bacteriocin biosynthesis
	slhT	1776	Transporter
	slhC	1272	Bacteriocin biosynthesis
	slhI	663	Immunity
Streptolancidin I	*sliP*	1374	Bacteriocin biosynthesis
	sliR	699	Histidine kinase/response regulator
	sliK	1344	Histidine kinase/response regulator
	sliF	702	Immunity
	sliE	738	Immunity
	sliG	687	Immunity
	sliA	168	Bacteriocin precursor
	sliB	2976	Bacteriocin biosynthesis
	sliT	1809	Transporter
	sliC	1278	Bacteriocin biosynthesis
	sliI	717	Immunity
Streptolancidin J	*sljA1*	138	Bacteriocin precursor
	sljL	2610	Bacteriocin biosynthesis
	sljP	1941	Bacteriocin biosynthesis
	sljT1	1608	Transporter
	sljT2	741	Transporter
	sljT3	1320	Transporter
	sljA2	138	Bacteriocin precursor
Streptolancidin K	*slkA*	99	Bacteriocin precursor
	slkL	2517	Bacteriocin biosynthesis
	slkT	1221	Transporter
Streptolassin	*slsA*	129	Bacteriocin precursor
	slsC	1725	Bacteriocin biosynthesis
	slsB1	252	Bacteriocin biosynthesis
	slsB2	2232	Bacteriocin biosynthesis
	slsF	717	Immunity
	slsE	792	Immunity
	slsG	711	Immunity
	slsR	774	Histidine kinase/response regulator
	slsK	1098	Histidine kinase/response regulator
Streptosactin	*ssaA*	177	Bacteriocin precursor

	ssaCD	1338	Bacteriocin biosynthesis
	ssaX1	153	Unknown
	ssaX2	996	Unknown
	ssaP	861	Bacteriocin biosynthesis
	ssaX3	696	Unknown

Note: 'Typical' length refers to the length of the gene in the previously published reference clusters.[339]

2.3 Computational data analysis

2.3.1 Sequence comparisons

2.3.1.1 Basic local alignment search tool

BLAST[370,371] is an algorithm developed in 1990 that searches for the best match to a query in sequence data and is still widely used due to its adaptability to a range of applications.[372,373] BLAST works by splitting the query sequence into short segments ('words') and uses matches to these in the seed alignments against the target sequence (hence 'local alignment'). The best match is returned if it exceeds thresholds set by the user. BLAST outputs can be assessed by the length and the percentage identity of the alignment relative to the query sequence, by the overall alignment score (also called the bit score), where higher bit scores are better matches to the query, and the E-value, which describes the probability of the alignment arising by chance.

The BLAST plugin in the BIGSdb isolate database was used for exploratory analyses in the genomic databases,[367] including initial screens for sequences of interest and to investigate the genomic context of annotated sequences using the option to extract hits to the query sequence with the flanking genomic sequence. The National Center for Biotechnology Information (NCBI) BLAST databases were used to identify unannotated sequences.[372]

2.3.1.2 Multiple sequence alignments

Multiple sequence alignments are used to compare multiple sequences simultaneously by aligning regions that share sequence similarity and introducing gaps where required.[374] Consensus sequences can be generated from alignments by taking the most commonly observed value at each position in the sequence. Pairwise percentage identities describe the percentage of positions in two sequences that have identical values, and in a multiple sequence alignment the mean pairwise percentage identity of all the pairs of sequences in the alignment is used as an overall assessment of sequence similarity. Percentage identity can also be given by position in an alignment, as the percentage of sequences in which that position is the same as the consensus sequence. Finally, the number of identical sites in an alignment can be used to assess sequence similarity.

A number of algorithms for multiple sequence alignments have been developed. In this thesis, the MUSCLE (multiple sequence comparison by log-expectation) algorithm was used for all alignments.[375] MUSCLE works by assessing how many k-mers (short stretches of sequence of length k) each pair of sequences has in common, using this to quickly cluster sequences by similarity. Clustered sequences are used to generate a multiple sequence alignment that is improved iteratively to minimise distances between sequences.

Alignments were performed within analysis code using biopython (Section 2.3.2) or within Geneious. When required, duplicate sequences were removed prior to alignments within python code or using the ElimDupes webtool (www.hiv.lanl.gov/content/sequence/elimdupesv2/elimdupes.html). When amino acid sequences were

aligned, nucleotide sequences were translated using the standard genetic code (Appendix Table 9.1). Alignments were visualised and annotated in Geneious Prime (Biomatters Ltd., V. 11.0.12+7). Image files and distance matrices (based on either pairwise percentage identities or number of differences in sequences) of alignments were exported from Geneious.

2.3.1.3 Phylogenetic trees

Phylogenetic trees are constructed to describe the evolutionary relationships between sequences, where sequence similarity can be used to infer close evolutionary history.[376] Trees may be used to compare variants of an individual gene or protein, the concatenated sequences of several genes or proteins, or even whole genome sequences. Each sequence is represented as a 'leaf' (or external node) separated from the others by 'branches' with lengths corresponding to the evolutionary distance. The internal nodes (branch points) represent hypothetical ancestral sequences inferred from the sequences and distances. Trees can be estimated with a sequence that is known to be less closely related to all the other sequences in the tree (an 'outgroup'), allowing the tree to be 'rooted'. In the absence of a root, trees only show relationships between groups, and cannot be used to infer directionality (*i.e.* whether one sequence evolved from another).

Trees are estimated based on sequence alignments using one of several methods, which are either algorithmic or tree-searching. Algorithmic approaches make use of a distance matrix to group sequences based on similarity, producing a single tree. Neighbour-joining is a widely used algorithmic approach. Tree-searching approaches search many trees and return the tree (or trees) that is assessed as the 'best' estimate. The criteria by which the 'best' tree is selected varies in different tree-searching methods: in parsimony

approaches, the trees with the minimum number of sequence changes are returned, while maximum likelihood approaches return the trees that maximise the likelihood of observing the input data. Both algorithmic and tree-searching approaches are widely used: generally algorithmic approaches are faster, but tree-searching methods can achieve better accuracy.

Unless otherwise stated, phylogenetic trees in this thesis were constructed in Geneious using the Tree Builder tool, which makes use of the neighbour joining method. Tree files were exported in Newick format and visualised and annotated using the Interactive Tree of Life software (iTOL).[377] All trees are unrooted and are therefore displayed in an unrooted format. Scale bars shown on trees represent the number of substitutions per site.

2.3.2 Python tools developed for bacteriocin analysis

I used Python (Python Software Foundation, www.python.org/, V. 3.8.8) to develop code for the analysis of annotated BIGSdb dataset exports. Command line programmes were written using Visual Studio Code (Microsoft, V. 1.65.2). Code for analysis of datasets was developed within Jupyter notebooks (jupyter.org/, V. 1.0.0).[378] scripts and documented notebooks are available at github.com/ mebbutler/thesiscode. Various open-source packages were utilised during python code development: pandas[379] (V. 1.2.3) was used for handling datasets, biopython[380] (V. 1.78) was used to handle sequence data and alignments, matplotlib[381] (V. 3.3.4) and seaborn[382] (V. 0.11.1) were used for data visualisation, and click (click.palletsprojects.com, V. 7.1.2) was used for the command line interface in python scripts. Conda (docs.anaconda.com, V. 4.10.3) was used to install and manage open-source programmes and packages.

Version control of code was performed using git (V. 2.32.0).[383] Files were stored in private git repositories hosted by GitHub. Version control of code in Jupyter notebooks was achieved using JupyText (jupytext.readthedocs.io/en/latest/, V. 1.10.3), which converts a Jupyter notebook to a plain text file suitable for version control with git.

2.3.2.1 BIGSgenbankerator.py

A script was developed to export the sequence coordinates of annotated genes in the private BIGSdb database *via* the BIGSdb REST API (bigsdb.readthedocs.io). As the relevant datasets are password protected, the OAuth1 system from the rauth library (github.com/litl/rauth, V. 0.7.3) was used to generate an access token. The BIGSdb API returns the annotated contigs of a single isolate in json format, which biopython can convert to a SeqRecord object before exporting as a genbank file. This annotated contig record can be parsed using biopython for other applications.

2.3.2.2 contiguity_cat.py

A second script was written to evaluate the sequence coordinates generated by BIGSgenbankerator.py to assess whether the loci of each bacteriocin cluster were located adjacent to one another in each genome, as would be expected for contiguous clusters of biosynthetic genes. This was necessary as the loci were annotated independent of genomic context, so their organisation into biosynthetic gene clusters required validation. For each genome, the script uses an annotated dataset export to determine which loci of each bacteriocin cluster were detected. It then parses the annotated genbank format file for that isolate (generated as described above by BIGSgenbankerator.py) and looks up the coordinates of each locus in the order that they

occur in the reference clusters. The coordinates are used to assesses the relative positions of each locus.

If the loci of a cluster are separated by less than 2.5 Kb, the cluster is classed as contiguous. When a cluster is spread across more than one contig, it is classed as contiguous if the loci nearest the ends of the contigs are within 2.5 Kb of the contig break. Any clusters that do not meet this condition are classed as non-contiguous (Figure 2.2). The script outputs are used to exclude non-contiguous clusters from analysis and to report the number of excluded clusters.

2.3.2.3 narwhal.py and manatee.py

The bulk of the processing, analysis and data visualisation of genomic datasets was performed using python code developed for this project. All code is available with documentation at github.com/mebbutler/thesiscode. Code used in Chapter 3 can be found in the Narwhal directory, and comprises two text files (processing.py and analysis.py) and an accompanying Jupyter notebook (narwhal.ipynb). The code used to generate results presented in Chapters 4 and 5 can be found in the Jupyter notebook manatee.ipynb.

2.3.3 Other software

Microsoft Excel (V. 16.61.1) was used for exploratory data analysis and formatting of data tables. Flow charts were generated using Lucidchart (lucid.app/) and other figures were generated using BioRender (biorender.com). Geneious Prime was used to examine sequences and generate figures with annotated sequence features. Figures exported from

visualisation software were arranged, and modified if necessary, using Affinity Designer (Affinity, V. 1.10.4).

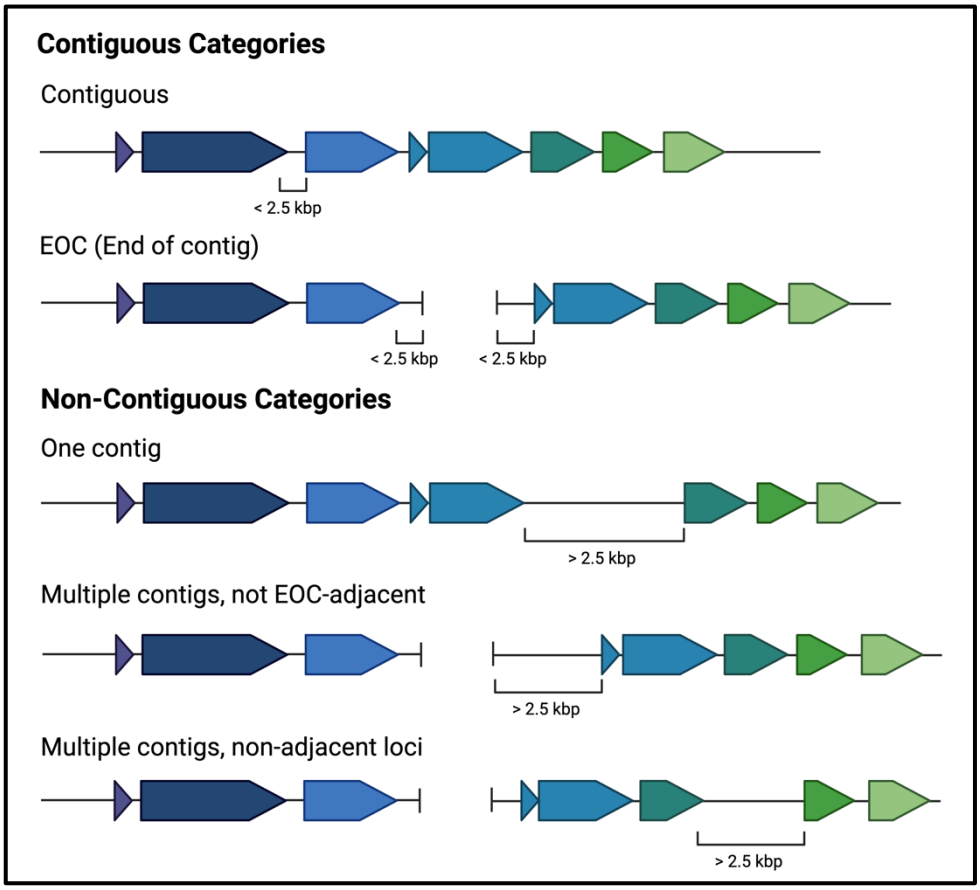

Figure 2.2: The bacteriocin cluster contiguity categories. Illustrated with a hypothetical gene cluster. Contiguous categories were included in downstream analysis, the non-contiguous categories were excluded.

3 Variation in Bacteriocin Distribution in Icelandic and Kenyan Pneumococci

The genomic datasets analysed in this chapter were generated by others as described in Section 2.1. Findings from this chapter were presented at two conferences: some results were included on a poster at the 14th Meeting on the Molecular Biology of the Pneumococcus (EuroPneumo 2019) in Greifswald, Germany, and others within a poster at the 31st European Congress of Clinical Microbiology and Infectious Diseases (ECCMID 2021), which took place online. The conference abstracts are provided in Appendix Section 9.5.

3.1 Introduction

3.1.1 Pneumococcal population biology

The global pneumococcal population is complex, with around 100 phenotypically distinct serotypes and almost 17,500 recorded MLST sequence types (as of May 2022, pubmlst.org/organisms/streptococcus-pneumoniae). These groups of pneumococci exhibit different behaviours and patterns: for example, some are far more likely to cause invasive disease than others, and some are more likely to be resistant to antimicrobials.[98,122,206] Moreover, when the population structure is perturbed, as has been observed following the introduction of PCVs, the distribution of pneumococci responsible for disease also changes, *i.e.* vaccine serotypes decrease in prevalence and

nonvaccine serotypes typically increase in prevalence.[159,166,169] Studying the pneumococcal population as a whole can therefore contribute to the understanding of pneumococcal biology and disease.

3.1.2 Bacteriocins in the pneumococcal population

The previous genome mining study identified a diverse range of bacteriocin clusters in a large global dataset of pneumococcal genomes sampled to maximise genetic diversity from 39 countries between 1916 and 2009 (Table 1.1).[339] The bacteriocins were harboured by different proportions of the dataset: some were ubiquitous while others were very rare, and each genome possessed between 5 and 11 different bacteriocins. Because the dataset was sampled to maximise diversity, the numbers of pneumococci from each country and clonal complex (CC) were relatively low. This meant that it was not possible to determine whether bacteriocin distribution varies in the global pneumococcal population, *i.e.* whether pneumococci from different locations possess different bacteriocins, nor whether bacteriocin distribution was consistent in genetically similar pneumococci (from the same CC).

An additional aspect of pneumococcal bacteriocin distribution that could not be studied in the global dataset was whether bacteriocin distribution differs in pneumococci that were recovered from carriage and those that caused disease. Bacteriocins would be expected to influence the competitiveness of pneumococci, and it is not known whether this impacts the ability of a pneumococcus to cause disease. Furthermore, bacteriocin distribution in the context of population restructuring following the introduction of PCVs has not been studied. In restructured populations, nonvaccine serotypes are more frequently carried, and can also cause an increased proportion of pneumococcal

disease.[61,165,166,169] It is not known whether bacteriocin distribution is altered in restructured populations, and if so, whether this affects competition dynamics in post-PCV populations.

3.1.3 Aims

In this chapter, I built on the previous genome mining study using two large datasets sampled from carriage and disease-causing pneumococci in Iceland and Kenya over time periods spanning PCV introduction to investigate:

- bacteriocin distribution in different geographic locations,
- changes to bacteriocin distribution following PCV introduction and differences in pneumococci recovered from carriage and disease,
- combinations of bacteriocins found in individual genomes.

3.2 Materials and methods

3.2.1 Genomic datasets

Results presented in this chapter used two large genomic datasets of pneumococci recovered from Iceland and Kenya. Descriptions of the generation of these datasets can be found in Section 2.1. Genes associated with bacteriocin biosynthetic gene clusters were annotated as described in Section 2.2 (Table 2.1).

3.2.2 Serotyping

Icelandic pneumococci were serotyped using latex agglutination and confirmed using multiplex PCR where required. Additionally, the sequence-based serotyping programme

seqSerotyper was used to determine serotypes based on the sequence of the capsular polysaccharide loci in the assembled genomes. Kenyan pneumococci were serotyped using the Quellung reaction and predicted *in silico* based on sequences of the *cps* loci using seroBA,[124] which was run as part of the genome assembly pipeline at the Sanger Institute. These were compared to the listed phenotypic serotypes of each genome and where there were discrepancies the genome was investigated manually. The BIGSdb BLAST plugin was used to query the genomes against reference *cps* locus sequences to validate serotype designations. If the apparent phenotypic serotype could be unambiguously excluded, and the seroBA serotype was supported, the seroBA serotype was assigned. Otherwise, the phenotypic serotype was retained. The validated serotype was recorded in the BIGSdb isolate database.

3.2.3 Chi-square test

Differences in bacteriocin prevalence were assessed using the Chi-square test for independence implemented in the Narwhal.ipynb Jupyter notebook. This test is suitable for assessing whether distributions of categorical variables differ from each other. The null hypothesis for all tests was: 'distribution of the bacteriocin cluster does not differ between the subsets of data'. Bacteriocin distributions were compared between the two datasets overall (Icelandic and Kenyan pneumococci), and in subsets of each dataset (pneumococci recovered before and after PCV10 introduction, and carriage and disease-causing pneumococci). Each bacteriocin cluster was considered independently. All Chi-square tests were therefore set up in 2x2 contingency tables (Table 3.1).

The expected values for each condition are calculated as the values if the null hypothesis is true; that is, if the bacteriocin distribution is the same in the subsets. The expected values are used to calculate the overall Chi-square statistic (χ^2) for the contingency table:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The χ^2 value was used to get the p-value from the Chi-square tables of p-values for a contingency table with one degree of freedom. Separate functions were used to generate contingency tables for each Chi-square test and a general function to calculate the p-value. If any observed value in the contingency table was < 5, the test was not performed due to insufficient sample size. If the p-value was > 0.05, the difference in frequency was classed as non-significant. Otherwise, the function returned the p-value interval (< 0.05, < 0.01, < 0.001). These were incorporated into plotting functions to indicate whether plotted bacteriocin frequencies were significantly different.

Table 3.1: Example contingency table used in Chi-square tests to assess the difference in frequency of bacteriocins.

	Subset 1	**Subset 2**	**Row marginals (RMs)**
Number of genomes with bacteriocin cluster	O1 E1 = (RM1/N) x CM1	O2 E2 = (RM1/N) x CM2	RM1 = O1 + O2
Number of genomes without bacteriocin cluster	O3 E3 = (RM2/N) x CM1	O4 E4 = (RM2/N) x CM2)	RM2 = O3 + O4
Column marginals (CMs)	CM1 = O1 + O3	CM2 = O2 + O4	$N = \sum O_i$

Note: Values O1-O4 are the observed number of genomes in each category, values E1-E4 are the expected values for each category if the null hypothesis is true.

3.3 Results

3.3.1 Genomic datasets

3.3.1.1 Summary of genomic dataset generation

A total of 1,912 genomes recovered in Iceland between 2009 and 2014 were analysed (Table 3.2). This total includes previously published genomes recovered from carriage and non-invasive disease pneumococci (recovered from otitis media or lower respiratory tract infection)[20,21,122,163,169] and an additional 183 genomes recovered from invasive pneumococci. Four genome assemblies (BIGSdb ID numbers 1586, 1594, 1644, 2540) from carriage pneumococci were removed because the low quality of the assemblies interfered with the identification of bacteriocin gene clusters, which were fragmented across contigs or interrupted by scaffolded regions.

3,372 pneumococcal isolates recovered between 2003 and 2017 were selected for inclusion in the Kenyan dataset (2,507 from carriage, 865 from invasive disease, Table 3.2). Of these, 3,280 were recovered from freezer stocks and 3,258 whole genome assemblies were obtained. Quality control found that one genome assembly was likely recovered from a *Streptococcus pseudopneumoniae* isolate and was therefore excluded. All other genome assemblies were of sufficiently high quality. The Kenyan dataset contained 90 pairs of duplicate invasive pneumococcal genomes that were isolated from the same patient and disease episode from two different specimen source. As these were inadvertently sequenced, a single genome from each pair was retained (Table 3.2). In all duplicate pairs, both genomes exhibited identical MLST and rMLST profiles. The final number of genomes in the Kenyan dataset was 3,159 (Table 3.3).

Table 3.2: Whole genome sequencing of pneumococci recovered from the isolate collections in the KEMRI-Wellcome Trust Research Programme (KWTRP).

Count of genomes			
	Carriage	**Invasive**	**Grand Total**
Included in final dataset	**2387**	**772**	**3159**
Excluded from final dataset	**26**	**99**	**125**
Recovered pre-2003	0	6	6
No gDNA extracted	19	0	19
No assembly received	4	3	7
Sequencing duplicate	2	0	2
Invasive sampling duplicate	0	90	90
QC fail	1	0	1
Grand Total	**2413**	**871**	**3284**

Note: The grand total is 3,284 because gDNA extractions and sequencing were attempted twice for four carriage pneumococci. In one case, two genomes from different pneumococci were received and included in the final dataset, in two cases a pair of genomes from the same pneumococcus were received, one of which was excluded from the final project (noted in table as 'sequencing duplicate'), and in the fourth instance only one assembly was received and included in the final dataset. Assemblies from six pneumococci recovered prior to 2003 were excluded, as were genomes from 90 pairs of duplicate invasive pneumococci. One genome failed quality control (QC) and was excluded from the final dataset.

Table 3.3: Pneumococci in the Icelandic and Kenyan study datasets.

		Number of genomes recovered from:	
		Iceland	**Kenya**
Total		1,912	3,159
Carriage		983	2,387
Disease	**Total**	929	772
	IPD	183	772
	LRTI	283	0
	OM	463	0
PCV time period	**Pre**	1,039	1,660
	Post	873	1,499

Note: IPD: invasive pneumococcal disease; LRTI: lower respiratory tract infection; OM: otitis media.

3.3.1.2 Differences in pneumococci recovered from each location

Pneumococci were collected over six years (2009-2014) in Iceland and 14 years (2003-2017) in Kenya (Figure 3.1A). In both countries, these time periods spanned the introduction of an infant PCV10 regimen in 2011. The pre-vaccine time period was defined as the study years up to and including 2011, and the post-vaccine time period as 2012 onwards. Disease sampling differed in the two datasets: both included invasive pneumococci, but Iceland additionally sampled pneumococci causing non-invasive disease (Figure 3.1A). Carriage pneumococci were recovered from patients of all ages in Kenya, and from children under 7 years of age in Iceland. Disease-causing pneumococci were recovered from patients of all ages in both datasets (Figure 3.1B).

40 and 56 serotypes were represented among the Icelandic and Kenyan pneumococci, respectively. Serotype 19F was the most common in both datasets (Figure 3.2) and was particularly prevalent in the Icelandic dataset due to the high incidence of otitis media caused by serotype 19F CC236/271/320 pneumococci in the country during the study period.[169] Other serotypes commonly found in both datasets include 6A, 19A and 23F. Other serotypes differed between the two populations, such as the highly invasive serotype 1, which was a major cause of invasive disease in the Kenyan dataset but rare in the Icelandic dataset (Figure 3.2). The Icelandic dataset included pneumococci from 59 unique CCs (and 5 singletons), and the Kenyan dataset included pneumococci from 116 unique CCs (and 62 singletons) (Table 3.4). 18 CCs were represented in both datasets, indicating that different pneumococci were circulating in the two locations during the study periods.

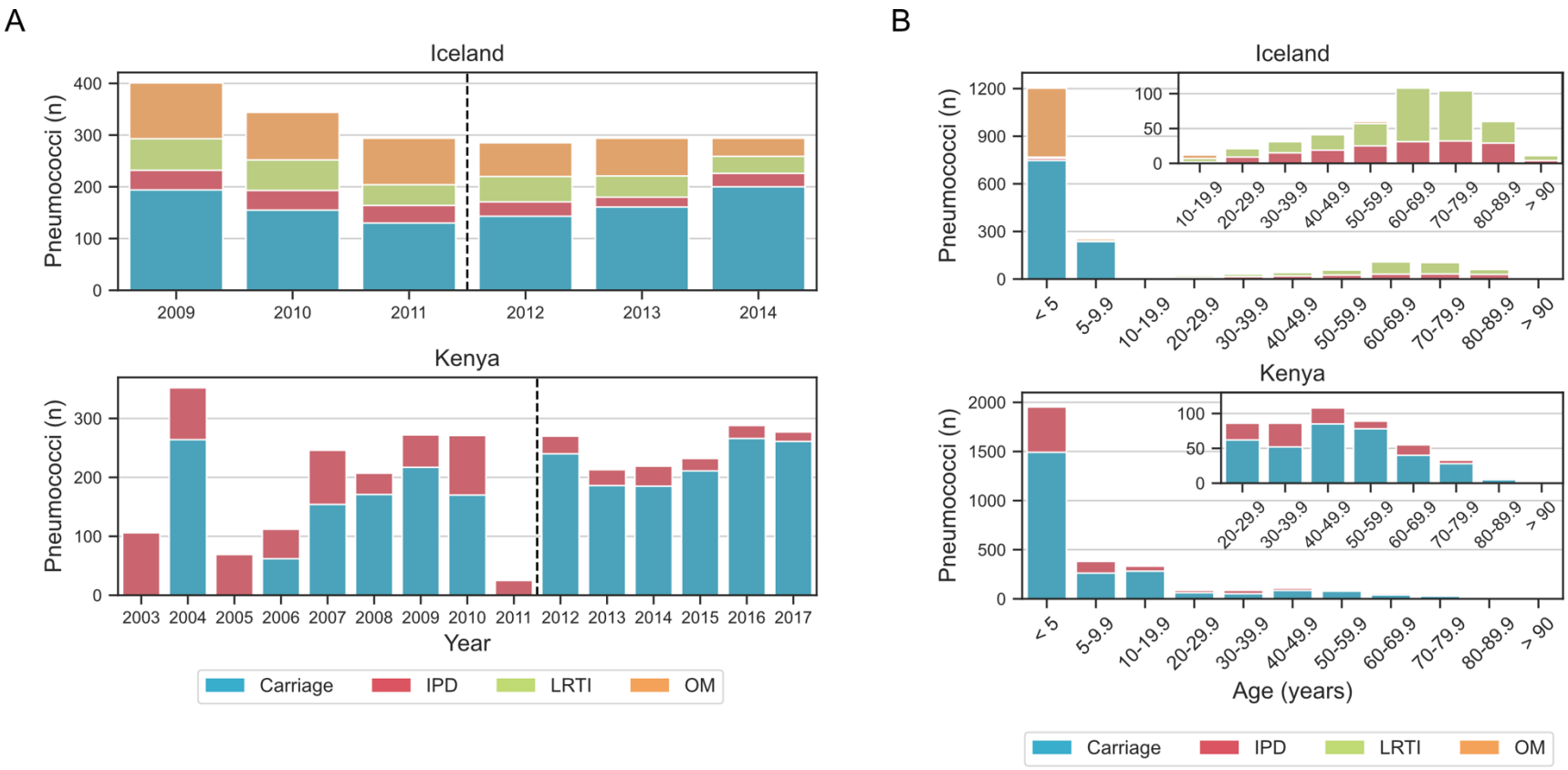

Figure 3.1: Description of the Icelandic and Kenyan datasets. Panel A: Number of pneumococci in each dataset by year of isolation, coloured by carriage (blue) or disease category (red, green, or orange). The dashed line separates pre- and post-PCV10 periods. Panel B: Number of pneumococci by age group of study subjects. Inset plots increase data resolution for the older age groups. IPD, invasive pneumococcal disease; LRTI, lower respiratory tract infection; OM, otitis media.

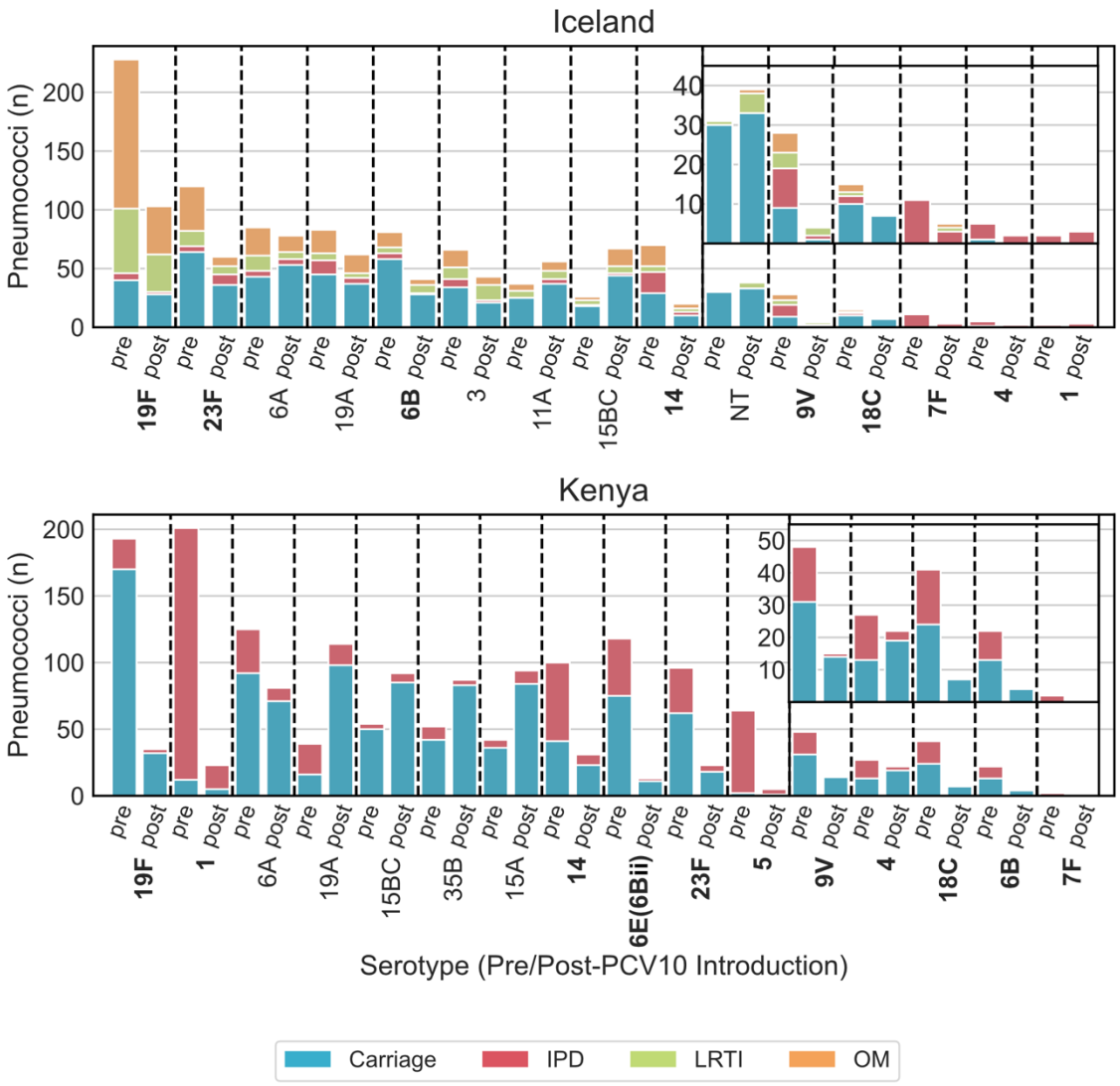

Figure 3.2: Serotype distribution in the Icelandic and Kenyan datasets. The 10 most common serotypes in each country are included, plus any additional PCV10 serotypes not in the top ten by rank order. Inset plots increase data resolution for those serotypes observed fewer than 50 times in both pre- and post-vaccine periods. PCV10 serotypes are marked in bold; note that serotype 5 was not observed in the Icelandic dataset. An additional 25 serotypes in Iceland (516 pneumococci) and 40 serotypes in Kenya (1,546 pneumococci) are not included in this figure. IPD, invasive pneumococcal disease; LRTI, lower respiratory tract infection; OM, otitis media.

Table 3.4: The 20 most prevalent clonal complexes (CCs) in the Icelandic and Kenyan datasets.

Iceland		Kenya	
CC	**n (%)**	**CC**	**n (%)**
236/271/320	293 (15.3%)	5902	239 (7.6%)
439	217 (11.3%)	217	223 (7.1%)
199	179 (9.4%)	701	163 (5.2%)
138/176	122 (6.4%)	5339	142 (4.5%)
180	107 (5.6%)	1146	139 (4.4%)
62	94 (4.9%)	138/176	133 (4.2%)
97	87 (4.6%)	156/162	131 (4.1%)
490	74 (3.9%)	991	104 (3.3%)
124	62 (3.2%)	230	92 (2.9%)
433	61 (3.2%)	852	78 (2.5%)
30	60 (3.1%)	5258	77 (2.4%)
392	47 (2.5%)	63	70 (2.2%)
156/162	46 (2.4%)	289	69 (2.2%)
344	37 (1.9%)	347	62 (2.0%)
15	36 (1.9%)	914	61 (1.9%)
1262	35 (1.8%)	7053	58 (1.8%)
448	29 (1.5%)	702	58 (1.8%)
193	28 (1.5%)	854	57 (1.8%)
100	25 (1.3%)	499	55 (1.7%)
90	22 (1.2%)	1381	49 (1.6%)
Other CCs[a]	235 (12.3%)	Other CCs[a]	954 (30.2%)
Singletons	16 (0.8%)	Singletons	145 (4.6%)

Note: ' Other CCs' represent 44 CCs in Iceland and 158 CCs in Kenya.

3.3.2 Bacteriocin cluster distribution in Icelandic and Kenyan pneumococci

116 genes associated with 20 putative bacteriocin biosynthetic gene clusters were annotated in the two datasets (Table 1.1, Table 2.1). Six bacteriocin clusters (streptococcins B and E, streptolancidins B, C, E and J) were observed as partial clusters, lacking at least one gene. Fragment clusters and clusters with non-contiguous genes were rare and were excluded from analyses (0-5% of observed clusters, Appendix Tables 9.3 and 9.4).

3.3.2.1 Bacteriocin distribution differed in the two datasets

Overall, cib and streptococcin B, C and E clusters were ubiquitous (or nearly so) among pneumococci from both countries, and streptolancidins H and I were never observed (Figure 3.3A). The remaining bacteriocins were observed in between 0.1% and 81% of genomes. The prevalence of twelve bacteriocins differed significantly ($p < 0.05$) between the two datasets. The largest significant differences in prevalence were observed for the streptolancidins. For example, streptolancidin B was observed in 340 Kenyan pneumococci (10.7%), and in only two Icelandic pneumococci (0.1%, Figure 3.3A). Streptolancidin B was harboured by pneumococci from 41 CCs and Singletons, all but two of which were represented in the Kenyan dataset but not the Icelandic dataset (Table 3.5). Similar patterns were observed for the other streptolancidins (Appendix Table 9.5). Therefore, the presence or absence of the different streptolancidins within each dataset was primarily determined by the pneumococcal CCs that were circulating within each country.

3.3.2.2 Bacteriocin prevalence differed following PCV10 introduction

Statistically significant differences in the prevalence of eight bacteriocins were observed among pneumococcal genomes in the post-PCV10 period relative to the pre-PCV10 period (Figure 3.3B). In the Icelandic dataset, streptocyclicin and streptolancidin G were more common in the post-PCV10 time period, and streptolancidin A and C were less common. In the Kenyan dataset, streptocyclicin and streptolancidins E and G increased in prevalence, and streptococcins D and E, streptolancidin C and streptolassin decreased.

3.3.2.3 Bacteriocin prevalence differed among carriage and disease pneumococci

Significant differences in bacteriocin prevalence were also observed between carriage and disease pneumococci in each dataset. In the Kenyan dataset, five bacteriocins were significantly more prevalent in invasive than in carriage pneumococci, and five were more prevalent in carriage than invasive pneumococci (Figure 3.3C, left panel). In the Icelandic dataset, bacteriocin prevalence in pneumococci recovered from carriage was compared separately to prevalence in pneumococci from invasive disease, lower respiratory tract infections and otitis media (Figure 3.3C, all panels). Significant differences were observed when carriage pneumococci were compared to pneumococci from otitis media (11 bacteriocins), lower respiratory tract infection (seven bacteriocins) and invasive disease (six bacteriocins). Streptococcins A and E were significantly more common among Icelandic pneumococci recovered from all three disease processes relative to carriage pneumococci. In both datasets, a range of serotypes were observed among pneumococci recovered from carriage and disease (Appendix Table 9.6).

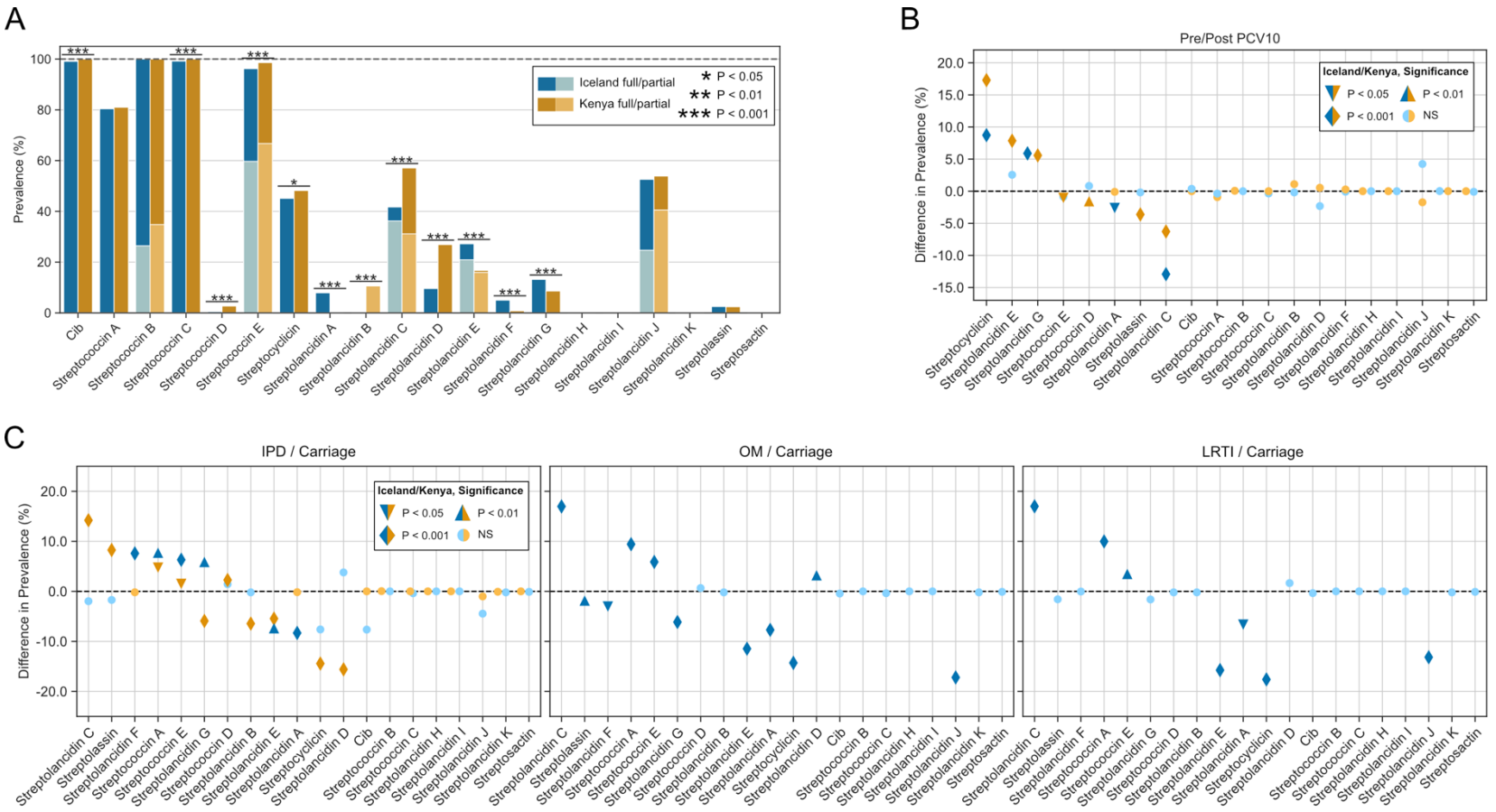

Figure 3.3: Prevalence of 19 different bacteriocin gene clusters in the Icelandic and Kenyan datasets. Icelandic data are displayed with blue bars and symbols, and Kenyan data are displayed with tan bars and symbols. Significant differences were assessed using a Chi-square test. Panel A: Overall prevalence of each bacteriocin in Iceland and Kenya. Panel B: Differences in prevalence of bacteriocins in the post-PCV10 vs pre-PCV10 time periods. Panel C: Differences in the prevalence of bacteriocins among invasive pneumococci (IPD) vs carriage pneumococci in Iceland and Kenya (left panel), pneumococci causing otitis media (OM) vs carriage pneumococci in Iceland (middle panel), and pneumococci causing lower respiratory tract infection (LRTI) vs carriage pneumococci in Iceland (right panel).

Table 3.5: Distribution of 340 streptolancidin B clusters within the Icelandic and Kenyan datasets.

Clonal complex	Pneumococci harbouring streptolancidin B n (% of CC representatives in each country with streptolancidin B)	
	Iceland	Kenya
CC702	0	57 (98.3)
CC499	0	55 (100)
CC5902	0	32 (13.4)
Sing11162	0	23 (100)
CC347	0	18 (29.0)
CC5250/5947/15006	0	18 (100)
CC703	0	16 (100)
CC385	0	13 (41.9)
CC1264	0	11 (100)
CC6446/14764	0	11 (100)
Other CCs	2 (100)	62 (34.6)
Other Singletons	0	22 (100)

Note: CC, clonal complex; Sing, Singleton. The 10 CCs with the highest prevalence of streptolancidin B clusters are shown.

3.3.3 Differences in bacteriocin prevalence could be explained by differences in population structure

The differences in bacteriocin prevalence were investigated relative to changes in the frequency of CCs pre- and post-PCV10, and also to differences in CCs causing disease relative to those recovered from carriage. Streptolancidin C was significantly associated with pre-PCV10 pneumococci in both datasets (Figure 3.3B), and with invasive pneumococci in the Kenyan dataset, and pneumococci recovered from otitis media and lower respiratory tract infections in the Icelandic dataset (Figure 3.3C). In both datasets, streptolancidin C was harboured by pneumococci from CCs with PCV10 serotypes (Figure 3.4A). Many of these CCs, such as CC236/270/320 (serotype 19F) in Iceland and CC217 (serotype 1) in Kenya, were also more highly represented in disease-causing pneumococci than carriage pneumococci. In contrast, streptocyclicin was significantly associated with the post-PCV10 time period (Figure 3.3B) and with carriage rather than disease pneumococci (Figure 3.3C) in both datasets. Streptocyclicin was found in CCs with both PCV10 and non-PCV10 serotypes, and in pneumococci recovered from carriage and disease (Figure 3.4B). All bacteriocins with significantly different pre- or post-PCV10 frequencies (Figure 3.3B) were inspected, and corresponding changes in the prevalence of CCs pre-/post-PCV10 introduction were generally found to explain the differences in bacteriocin prevalence (Appendix Tables 9.7 and 9.8). Vaccine-induced population restructuring may therefore lead to changes in bacteriocin distribution among pneumococci in post-PCV time periods.

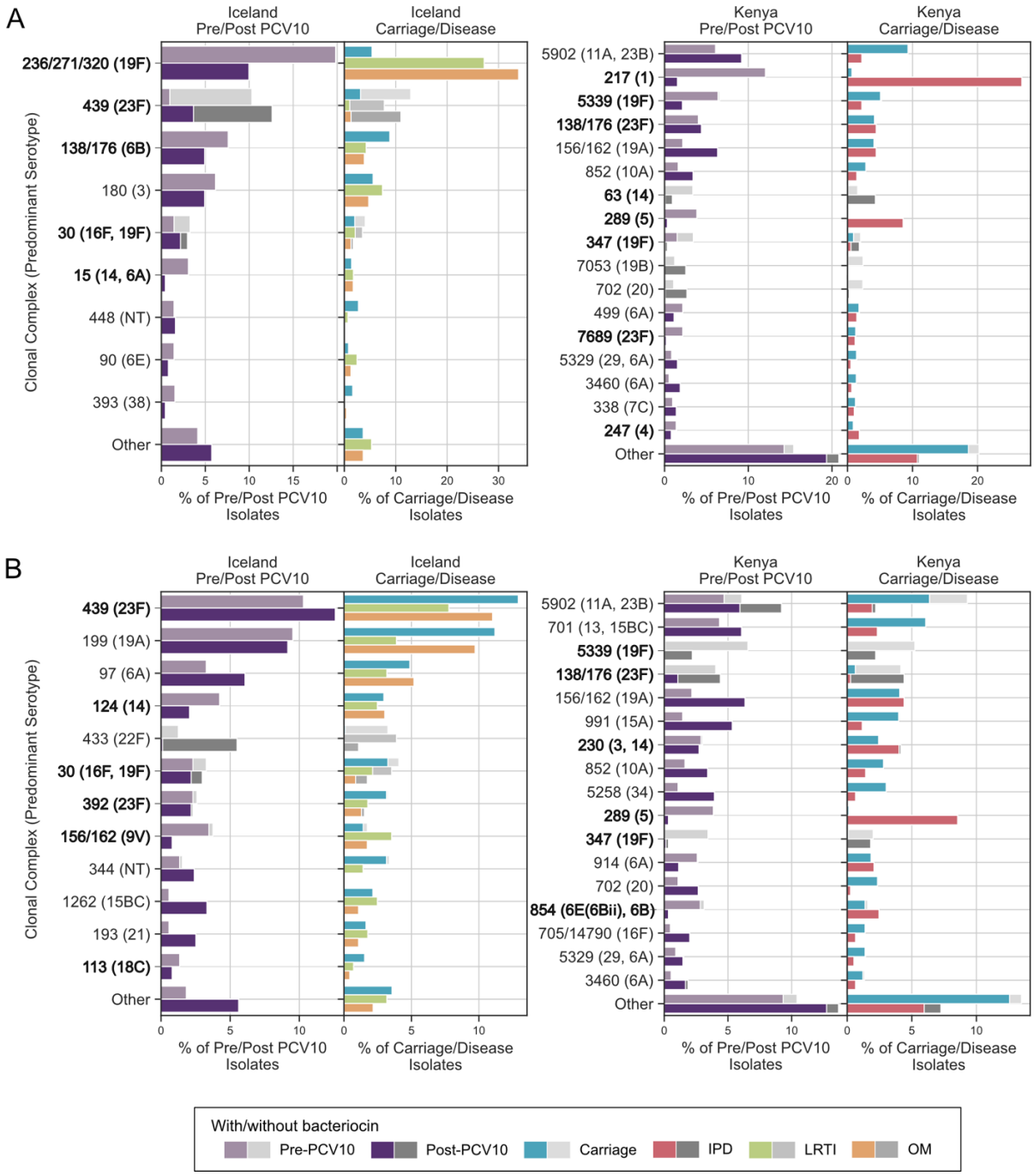

Figure 3.4: Prevalence of streptolancidin C (panel A) and streptocyclicin (panel B) bacteriocins among pneumococci in two groups, carriage vs disease and pre- vs post-PCV10, and stratified by clonal complex (CC). Each plot shows all CCs in which the bacteriocin was detected. Any CC representing <1% of the overall dataset was placed in the 'Other' category. Each bar represents the percentage of pneumococci from that CC, the coloured section of the bar represents pneumococci with the bacteriocin, and the grey section represents those without the bacteriocin. For Icelandic disease-causing pneumococci, only the disease in which the bacteriocin was significantly altered relative to carriage are shown.

Figure 3.4 (continued): The dominant serotype(s) of each CC is shown in brackets. Serotypes targeted by PCV10 are in bold text. IPD, invasive pneumococcal disease; LRTI, lower respiratory tract infection; OM, otitis media.

3.3.4 Bacteriocin repertoires

The combination of different bacteriocin clusters detected in each genome, or the bacteriocin 'repertoire', was investigated. Overall, each pneumococcal genome harboured between 4 and 11 bacteriocin clusters, and in both datasets the mode was seven (Figure 3.5A). The majority of genomes in the Icelandic and Kenyan datasets (89% and 81% respectively) had between 6 and 8 bacteriocin clusters. A total of 134 different bacteriocin repertoires were observed, 43 of which were present in both datasets. More repertoires were observed in the Kenyan dataset (n=103) than the Icelandic dataset (n=74). Although the most common repertoire was consistent in both datasets (cib, streptococcins A, B, C, E, and streptolancidin C), other common repertoires were restricted to one dataset (Figure 3.5B).

3.3.4.1 Bacteriocin repertoire varied within CCs and STs

The bacteriocin repertoire was largely consistent among pneumococci from the same CC; however, there were examples of CCs (Iceland, n=29; Kenya, n=61) in which more than one bacteriocin repertoire was observed. Some of the STs that comprised the CC also showed variable repertoires (Appendix Tables 9.9 and 9.10). Generally, individual CCs and STs were dominated by a single repertoire with few examples of divergent repertoires. For example, CC439 in Iceland (Figure 3.5C) and CC5902 (Figure 3.5D) in Kenya demonstrated minor differences in the presence or absence of bacteriocins, and these differences were maintained even within some STs.

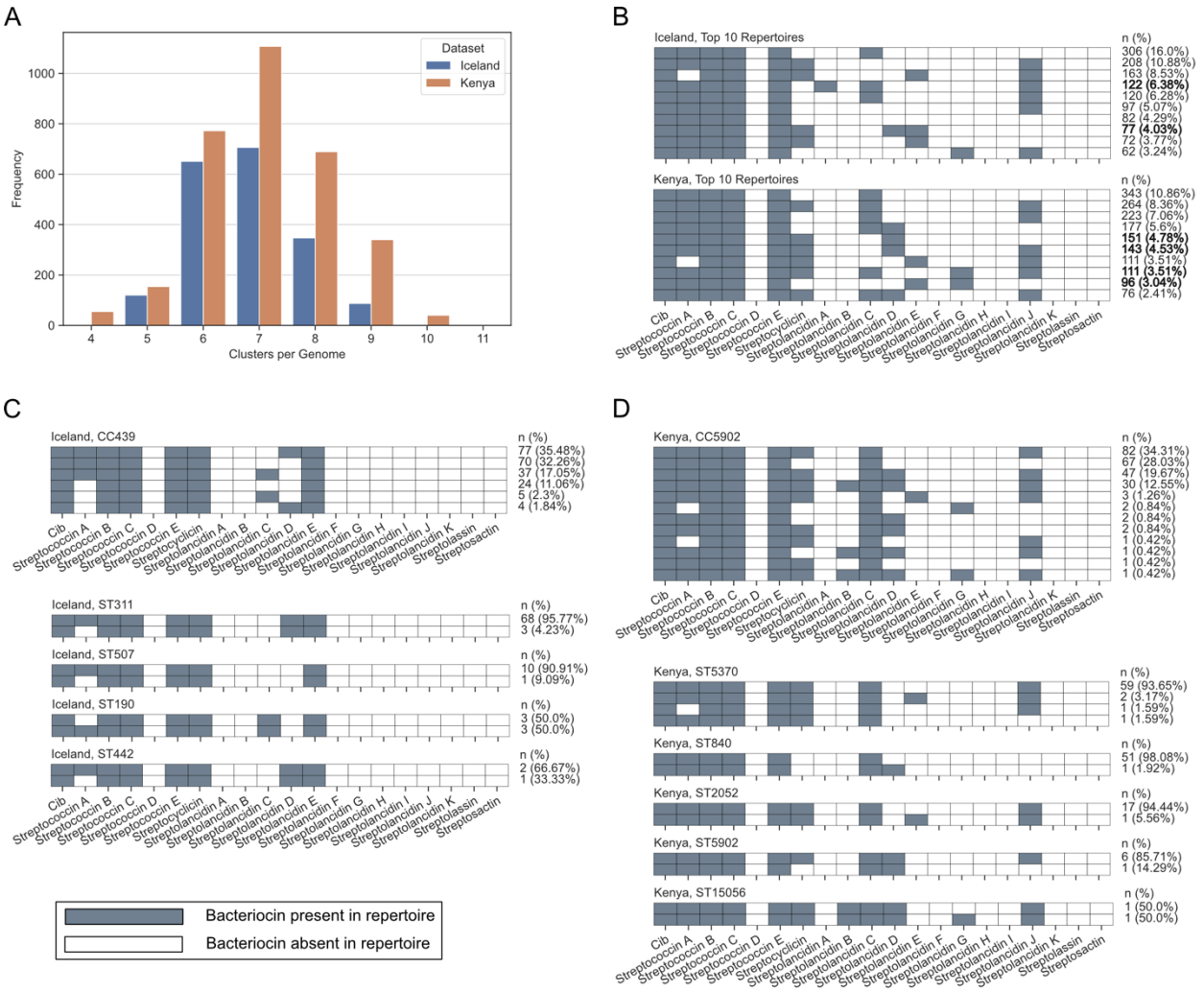

Figure 3.5: Bacteriocin repertoires observed among Icelandic and Kenyan pneumococci. Panel A: Number of bacteriocin clusters detected per genome. Panel B: The composition of the 10 most frequently observed repertoires in Icelandic and Kenyan pneumococci. Bold text indicates that the repertoire was restricted to that country. Panels C and D: Bacteriocin repertoires observed in genomes from CC439 in the Icelandic dataset (C), and CC5902 in the Kenyan dataset (D), including a more detailed breakdown of the CC by sequence type (ST) below the CC summary. STs with no differences in repertoire are not shown. Note: the number of genomes with each repertoire within the Icelandic or Kenyan dataset, respectively, is given at the far right of each diagram.

3.4 Discussion

3.4.1 Bacteriocin distribution varied with population structure

Bacteriocin gene clusters were found in pneumococci from both Iceland and Kenya and their distribution across subgroups of pneumococci was not uniform. Significant differences in bacteriocin prevalence were found between geographic location, between pneumococci recovered from carriage and disease, and in post-PCV time periods. These observations could largely be explained by different underlying pneumococcal population structures in each country, as bacteriocins tended to be associated with specific CCs.

3.4.1.1 Bacteriocins and pneumococcal disease

It is not clear whether the increased prevalence of some bacteriocins among disease-causing pneumococci is an indirect effect of the differing pathogenicity of pneumococcal CCs (driven in part by association to serotype),[59,98] or if bacteriocins themselves contribute to pneumococcal pathogenicity. A direct contribution to invasive disease seems unlikely: invasive disease is defined as the infection of a normally sterile site, so there should not be other bacteria in the niche that an invasive pneumococcus needs to out-compete. However, colonisation of the lower respiratory tract or the middle ear may be influenced by bacteriocins, as these sites do have characteristic microbiomes in which bacteriocins may provide a competitive advantage.[137,138] Bacteriocin contribution to pathogenicity could also be indirect: by promoting nasopharyngeal colonisation, a bacteriocin-producing pneumococcus may have more opportunity to infect different sites in the host, resulting in disease.

3.4.1.2 PCVs alter bacteriocin distribution

It is important to understand pneumococcal competition dynamics in the nasopharynx as colonisation of this niche is a precursor to disease. The effect of population restructuring following PCV introduction on competition dynamics in the nasopharynx is not well understood, although the observation that bacteriocin distribution is affected in pneumococci recovered from post-PCV time periods is suggestive of altered competition dynamics. Further work is required to confirm this and to determine the effect, if any, on pneumococcal disease.

3.4.2 Bacteriocin repertoires varied in size and content

Pneumococci have extensive repertoires of bacteriocins (up to 11 different bacteriocins per genome), and the combination of bacteriocins is not fixed within CCs. This could be suggestive of highly complex competition dynamics within the nasopharynx, where different combinations of bacteriocins prove advantageous in different circumstances. The rarer clusters may only give a competitive advantage over an unusual competitor or set of competitors. Alternatively, they may have been acquired by pneumococci relatively recently, and not yet become distributed throughout the pneumococcal population.

3.4.2.1 Horizontal exchange of bacteriocin gene clusters

The observation of pneumococci from the same CC, and even with the same ST, with variable bacteriocin repertoires is suggestive of a high rate of gain or loss of bacteriocin gene clusters. This would be consistent with horizontal gene transfer of bacteriocins among pneumococci, either by homologous recombination, or as genes within integrative conjugative elements (ICEs), as is observed for bacteriocins of other streptococcal species.[384] Moreover, the horizontal exchange of bacteriocins may not be restricted to

pneumococci: previous work has shown that some of the pneumococcal bacteriocin gene clusters included in this study are also detectable in genomes of non-pneumococcal streptococci such as *S. mitis, S. oralis* and *S. pseuodopneumoniae*,[339] and examples of genetic exchanges between pneumococci and these species have been described.[385,386] If bacteriocin clusters are exchanged between pneumococci and other streptococci, genetic lineages may be able to adapt to altered competition dynamics, such as in remodelled post-PCV populations, by acquiring a bacteriocin repertoire that improves competitiveness. It would be beneficial to study bacteriocin repertoires in co-colonising pneumococci and commensal streptococci from the same ecological niche.

3.4.3 Limitations

3.4.3.1 Sampling differences between Icelandic and Kenyan datasets

While the Icelandic and Kenyan datasets are useful as comparisons of pneumococci before and after PCV10 introduction in two different geographic locations, the differences in sampling must be considered. The greatest difference is in the sampling of disease-causing pneumococci: the Icelandic dataset includes pneumococci recovered from lower respiratory tract and middle ear infections as well as invasive infections, whereas the Kenyan dataset only includes invasive pneumococci. This prevents the comparison of pneumococci causing non-invasive disease in the two locations. An additional difference was in the sampling of carriage pneumococci: in Kenya, carriage pneumococci were recovered from patients of all ages in the study area, whereas Icelandic carriage pneumococci were recovered exclusively from children under 7 years of age. The strategy used to choose which isolates should be sequenced also varied: alternate Icelandic carriage and non-invasive disease pneumococci were selected, while Kenyan carriage pneumococci were sampled randomly and weighted to reflect the human population in

the study area. Both datasets used robust sampling methods and are likely good representations of the pneumococci that were circulating in the human populations during the study periods. However, no sampling strategy is perfect and both studies are likely to be biased to some extent, and the inconsistent methodologies may undermine comparisons between the two datasets.

3.4.3.2 Putative bacteriocins

This study used a large set of bacteriocins that were identified in a genome mining study, of which only a subset have been studied experimentally (Section 1.2.3, Table 1.1). The putative functions of the bacteriocins that have only been identified *in silico* are based on homology to other bacteriocin systems, but experimental work is required to confirm their role in pneumococcal competition and to establish their mechanism of action and target species. Moreover, many bacteriocin clusters appear to be under the control of dedicated regulatory systems that are not yet fully characterised. A fuller understanding of bacteriocin function and mechanism would be useful in interpreting the prevalence data presented here, as a bacteriocin can only provide a competitive advantage if it is expressed in an environment with a competing susceptible bacterial strain.

3.4.3.3 Multi-locus sequence typing

MLST is a well-established technique for molecular typing of pneumococci. However, it is limited by using only seven loci from the pneumococcal genome, and as whole genomes were sequenced it would have been possible to use higher resolution typing schemes such as cgMLST. A cgMLST scheme for pneumococcus is currently under development but fell outside of the scope of this study. Future investigations will make use of the

cgMLST scheme to further characterise the distribution of bacteriocins in the pneumococcal population.

3.4.4 Conclusions

Results presented in this chapter build on previous genomic analyses of putative bacteriocin gene clusters by using two large genomic datasets representative of pneumococci circulating in two geographic locations (Iceland and Kenya) to describe bacteriocin distribution in the pneumococcal population structure. These results show:

- Bacteriocin distribution is affected by the underlying population structure due to associations of bacteriocins with pneumococcal CCs.
- PVC-induced population restructuring alters the distribution of some bacteriocins, with unknown effects on competition dynamics.
- Bacteriocin repertoires can vary within CC, suggesting that whole bacteriocin clusters may be exchanged between pneumococci.

Further work will be required to fully understand the role of bacteriocins in nasopharyngeal competition, and the consequences of their altered distribution in post-PCV populations.

4 A Model of Streptococcin Function

Results presented in this chapter contributed to a poster that will be presented at ISPPD-12 in June 2022. This abstract can be found in Appendix Section 9.5.

4.1 Introduction

4.1.1 Lactococcin 972

The five streptococcin gene clusters were identified previously based on sequence homology to lactococcin 972,[339] a well-studied bacteriocin produced by *Lactococcus lactis*.[387] Lactococcin 972 biosynthetic gene clusters comprise one small gene encoding the bacteriocin toxin, and two larger genes encoding the transmembrane domain and nucleotide binding domain of an ABC transporter.

4.1.1.1 Lactococcin 972 mechanism

Lactococcin 972 interferes with cell wall synthesis of dividing cells by interacting specifically with extracellular lipid II at the septum.[388,389] Lipid II is an intermediate in peptidoglycan synthesis and is a common target among bacteriocins produced by Gram-positive species, including nisin.[14,257] Interestingly, lactococcin 972 appears to have a distinct mode of interaction with lipid II compared to nisin, as the presence of excess lactococcin 972 does not fully antagonise the nisin-lipid II interaction.

Nuclear magnetic resonance spectroscopy has been used to determine the structure of lactococcin 972 as a soluble, monomeric protein with a β-sandwich fold.[390] The lipid II

binding site has not been characterised, but the authors propose a patch of hydrophobic, aromatic amino acid residues on the surface of the structure as a potential site of interaction. As lactococcin 972 binds lipid II only at the septum of dividing cells, it is likely that is also interacts with another unidentified target to confer specificity, although this target and its binding site have not been identified.

4.1.1.2 Lactococcin 972 export

The SecYEG translocon is a highly conserved system used in all domains of life for the co-translational export of proteins and for the insertion of transmembrane segments into the plasma membrane.[391] Proteins that are processed by this pathway have characteristic N-terminal signal peptides that target them to the translocon and are typically cleaved following export. Lactococcin 972 is considered unusual among bacteriocins as it possesses an N-terminal signal peptide and is believed to be exported *via* the SecYEG machinery rather than *via* a specialised export system.[392] In native lactococcin 972 purified from *L. lactis*, the signal peptide had been cleaved.[390]

4.1.1.3 Lactococcin 972 immunity and resistance

The immunity mechanism of lactococcin 972-producing strains has not been determined, but it is proposed to be conferred by the two genes that encode an ABC transporter.[392] Early publications of lactococcin 972 activity note that it was not possible to generate strains in which the immunity genes were knocked out while the toxin gene was retained. This result is consistent with the proposed function of immunity. More recent studies have characterised *Lactococcus lactis* strains that have evolved resistance to lactococcin 972 without the need for the putative immunity genes.[393] This is achieved by the resistant

L. lactis strain sensing lactococcin 972-induced cell envelope stress and modulating the cell wall composition to evade lactococcin 972 activity.

4.1.2 Predicting protein structure

4.1.2.1 Relationship between amino acid sequence and protein function

The three-dimensional structure, or 'fold', determines the functionality of a protein and is in turn determined by its amino acid sequence.[394] The prediction of protein structure from an amino acid sequence is desirable because experimental structural biology is slow and resource-intensive relative to the identification of novel genes of interest from whole genome sequences. Although it is theoretically possible, this problem has presented computational challenges to structural bioinformaticians due to the sheer number of interactions and variables that must be considered.[395]

4.1.2.2 Protein folding

Protein folding refers to the process by which an extended polypeptide (the primary structure) transitions to a functional three-dimensional structure.[396] The first stage is the formation of the secondary structure, which is mediated by hydrogen bonds between amino acid residues. The secondary structure comprises α-helices, which are coiled helical elements, and β-sheets, which are formed by hydrogen bonds between extended β-strands (arranged either parallel or anti-parallel to one another). The tertiary structure refers to the three-dimensional arrangement of the secondary structural elements relative to one another, and in a globular protein is driven by the hydrophobic collapse, where hydrophobic residues form the core and hydrophilic residues remain on the surface of the protein. Quaternary structure refers to the arrangement of multiple

discrete subunits into a larger protein complex, sometimes with additional non-polypeptide co-factors.

In integral membrane proteins, which fully span cell membranes, the transmembrane regions are hydrophobic or aliphatic α-helices with hydrophobic side chains that interact with the lipid core of the membrane bilayer.[391] These proteins are inserted into the membrane using the SecYEG export machinery. One contiguous polypeptide can possess a single transmembrane helix, or many helices joined by loops on either side of the membrane. Long loop regions can fold into discrete domains on either side of the membrane joined by transmembrane helices. Membrane proteins can also include β-strands, but this has only been observed in the outer membrane of Gram-negative bacteria.[397]

4.1.2.3 The protein folding problem

As the amino acid sequence of a protein determines its folded structure, the prediction of structure from sequence data is theoretically possible.[394] However, in practice, the 'protein folding problem' has been challenging to solve.[398] Various approaches have been adopted to predict the structure, and therefore the function, of proteins without the need for intensive experimental structural biology approaches. Many of these approaches rely on sequence homology between the protein of interest and a protein with an experimentally determined structure. This may be used to infer the structure of the whole protein (homology modelling),[399,400] or simply to identify functionally important features, such as residues known to be involved in catalytic mechanisms, binding sites facilitating an interaction with another protein, nucleotide, or other ligand, or sites that are recognised by other proteins, such as sites of post-translational modifications.

Structural features can also be identified based on the chemical properties of the amino acids. For example, transmembrane helices can reliably be identified based on the hydrophobic properties of the constituent residues and by their characteristic length to match the thickness of the membrane (typically 15-30 amino acids).[401] Likewise, the properties of SecYEG signal peptides are sufficiently understood to permit the reliable detection of these regions from protein sequences.[402]

Proteins are organised into large families that have similar sequences and structural features, and similar functionality, and novel proteins can be assigned to a protein family based on sequence similarity. InterPro is a database that integrates many of these protein family databases, including the pfam protein family database[403] and the NCBI conserved domain database (CDD),[404] as well as tools for predicting functional and sequence motifs, such as phobius, a signal peptide prediction tool.[401] The database can be queried using a protein sequence (InterProScan), returning an InterPro entry that the protein is assigned to and a summary of any functional or structural motifs that were identified.[405]

4.1.2.4 AlphaFold

The machine learning approach AlphaFold (alphafold.ebi.ac.uk)[406,407] represents a remarkable breakthrough in the prediction of protein structure: the algorithm utilises a neural network trained on experimental structures deposited in the protein data bank (PDB, rcsb.org)[408] and achieves structural predictions that approach the accuracy of experimental structures. AlphaFold has been applied to UniProt reference proteomes of a range of species of interest, including the WHO priority pathogens,[4] which includes pneumococcus. As of March 2022, this represents almost 1,000,000 structural predictions that are freely available for use by the community.

4.1.3 Aims

Overall, the streptococcins represent good candidates for functional studies. Their biosynthetic gene clusters are relatively simple, comprising only three genes and lacking any enzymes for post-translational modifications. In this thesis, I have shown that four of the streptococcins are ubiquitous or very common in two pneumococcal genomic datasets (Figure 3.3A). In this chapter, I aimed to generate a unified model for the functionality of the streptococcin clusters that would inform and contextualise future work. I achieved this by:

- Comparing the streptococcin clusters to each other, and to the homologous bacteriocin lactococcin 972, identifying regions of conservation in amino acid sequence
- Investigating the predicted structure and functionality of each gene from the streptococcin clusters, making use of large databases of protein families, domains, and functional motifs
- Making use of structural prediction tools.

4.2 Materials and Methods

4.2.1 Streptococcin amino acid sequences

Protein sequences used for structural and functional predictions were taken from the Icelandic and Kenyan pneumococcal whole genome sequences stored and annotated in the private BIGSdb database (Sections 2.1 and 2.2). Only coding sequences (hereafter referred to as genes for simplicity) with a full-length sequence were included in this chapter: pseudogenes (which do not constitute a complete coding sequence), and cases

where a gene was detected but no allele could be designated (due to an interruption by a contig break or ambiguous reads), were excluded. Amino acid sequences of genes from the lactococcin 972 biosynthetic gene clusters were taken from the UniProt protein database.[409]

Nucleotide sequences were translated using BioPython or within Geneious. All translations used the standard genetic code (Appendix Table 9.1). Where a single representative sequence was required, the reference allele of each gene was used (BIGSdb allele ID 1). Otherwise, the nucleotide sequences were translated, and then any replicate amino acid sequences were removed to generate a set of unique amino acid sequences observed at that gene.

4.2.2 Streptococcin structural and functional predictions

4.2.2.1 Assigning protein families

The InterPro database integrates various tools for identifying experimentally studied structural homologues of query protein sequences and it was used to identify protein families for the translated products of the streptococcin reference alleles.[405]

4.2.2.2 Identification of conserved motifs, domains, and transmembrane topology

InterPro also returns annotations of features identified on a query protein, including conserved sequence motifs and predictions of transmembrane helices, signal peptides and topology (using phobius).[401] These predictions were used to annotate sequences with features of interest.

4.2.2.3 AlphaFold structural predictions

Publicly available structural predictions of streptococcin genes from a reference pneumococcal genome (R6, ATCC BAA-255) were downloaded from the AlphaFold protein structure database (alphafold.ebi.ac.uk/) on 18th March 2022.[406,410] The predicted structures were visualised and annotated using PyMol (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC, version 2.5.2). The quality of the structural models is given as a per-residue local distance difference test value (pLDDT), which describes the confidence in the modelled position of each residue. pLDDT values ≥90 indicate high confidence, ≥70 indicate moderate confidence, <70 indicate low confidence, and <50 indicate very low confidence.[410,411] Predicted structures are also assessed with the predicted aligned error, which is used to describe the error in the positions of each residue relative to the other residues of the protein. This is shown as a matrix and allows the modelled positions of discrete domains and regions in the structure to be assessed relative to one another. A high expected position error between two residues indicates that there is uncertainty in their relative positions. The pLDDT values and predicted aligned error matrices for structural predictions can be found within the AlphaFold protein structure database.

4.2.3 Streptococcin sequence comparisons

4.2.3.1 Multiple sequence alignments and phylogenetic trees

Protein sequences were compared using multiple sequence alignments and phylogenetic trees and sequences were handled using Python and Geneious, as described in Section 2.3. Code used to generate results in this chapter can be found in manatee.ipynb.

4.2.3.2 Generating a plot of the percentage identity of the B genes

The unique amino acid alleles of each B gene from streptococcins A-C and E were used to generate multiple sequence alignments in Geneious. The mean percentage identity at each position in the alignment was used to generate a line plot of the mean percentage identity across each sequence. The predicted structural features of each gene were overlaid on the axes.

4.2.4 Generating a model of streptococcin function

Structural and functional predictions of the proteins encoded by the streptococcin biosynthetic gene clusters were used to develop a model for the overall functionality of the streptococcins. A cartoon of this model was generated using BioRender (biorender.com).

4.3 Results

4.3.1 Streptococcin toxins

4.3.1.1 Streptococcin toxins are homologues of the unmodified bacteriocin lactococcin 972

Streptococcin clusters contain a single, small (288-348 bp) toxin gene. The translated toxin genes from all five pneumococcal streptococcin gene clusters had structural homology to lactococcin 972 (Table 4.1). Many bacteriocins are post-translationally modified, and the biosynthetic gene clusters encode the highly specialised modification enzymes, for example the genes for the lanthionine modifications found in the streptolancidin gene clusters (Section 1.2.2). The streptococcin clusters do not encode

any modification enzymes, and lactococcin 972 is not believed to be post-translationally modified, so it is likely that the streptococcin toxins are also unmodified.

4.3.1.2 Streptococcin toxin export is predicted to be SecYEG-dependent

The toxin gene products were predicted to possess an N-terminal signal peptide for targeting the translated product to the SecYEG translocon (Table 4.1). Therefore, it is likely that streptococcin toxins are exported, which is consistent with their predicted function as bacteriocins, and that their export is SecYEG-dependent like lactococcin 972.

Table 4.1 Summary of functional predictions of the streptococcin toxins using Phobius for signal peptide and transmembrane region predictions and InterPro for assignment to protein families.

Streptococcin	Phobius-predicted signal peptide location	InterPro predicted protein family
A	1 - 23	Bacteriocin, lactococcin 972 (IPR006540)
B	1 - 27	
C	1 - 41	
D	1 - 25	
E	1 - 25	

4.3.1.3 Streptococcin toxin sequence diversity

The amino acid sequences of the streptococcin A, B, C and E toxin alleles observed in the Icelandic and Kenyan pneumococcal genomic datasets were compared. (Note that only a single streptococcin D toxin allele has been observed in pneumococcal genomes.) Streptococcin A and C toxins were slightly more diverse than B and E toxins based on mean pairwise identity and proportion of identical sites, but among all streptococcins the overall variation was low (Table 4.2). The similarity of the toxin sequences was assessed

by constructing an unrooted phylogenetic tree using each allele of the streptococcin A - E toxins and a representative lactococcin 972 sequence (UniProt accession number O86283). Each streptococcin formed a distinct phylogenetic cluster (Figure 4.1A), validating their groupings.

Table 4.2: Sequence diversity of streptococcin toxin genes observed in pneumococci.

Streptococcin	Number of unique amino acid alleles	Mean pairwise identity (%)	Identical sites (%)
A	11	89.9	75.8
B[a]	10	96.3	84.7
C	29	94.1	73.3
D	1	-	-
E	5	94.9	90.8

a - One rare, atypical streptococcin B allele with a 10 amino acid insertion was excluded from this analysis.

4.3.1.4 Conserved amino acid motifs in the streptococcin toxins

When representative streptococcin and lactococcin 972 toxin sequences were aligned, some amino acid motifs were highly conserved across the sequences (Figure 4.1B). The signal peptide regions were excluded from the alignments as they are likely to be cleaved following export and are therefore not expected to be involved in the mechanism of streptococcin toxicity. Alignments revealed three regions of highly conserved sequence in the toxins: a glycine-rich region, a histidine-containing region, and an aromatic-rich C-terminal region. Across all streptococcins, the majority of conserved amino acids possessed non-polar side chains (Table 4.3), and many of these were aromatics (in particular tyrosine).

The glycine-rich region found at the N-terminal of the protein contains a highly conserved double-glycine motif (Gly_2), which is associated with other streptococcal bacteriocins and is the site of cleavage of 'leader peptides' by specialised C39 peptidases in the final processing step.[412,413] The pro-peptide is typically exported by a specialised export system, but as the streptococcins are not expected to utilise such a system, it remains to be seen whether this conserved region represents a true Gly_2 motif or whether these conserved amino acids have a different role in the streptococcins.

The histidine-containing region is the longest region of interest and has two universally conserved histidine residues separated by four less-conserved residues. These are notable as the only histidine residues found in the toxin sequences and it is possible that these residues are particularly important to streptococcin function. Finally, the C-terminal regions of the streptococcins and lactococcin 972 contained several conserved aromatic residues. According to the published structure of lactococcin 972, the C-terminal aromatics were found near to conserved N-terminal aromatics on the surface of the protein and thus were proposed as the binding site for lipid II.[390] If the streptococcin toxins do have a similar structure to lactococcin 972, the conserved aromatics are likely to play an important role in their mechanism of activity. Nonetheless, the consistency of these residues in all five streptococcin toxin groups suggests that the streptococcins have a similar mechanism of toxicity.

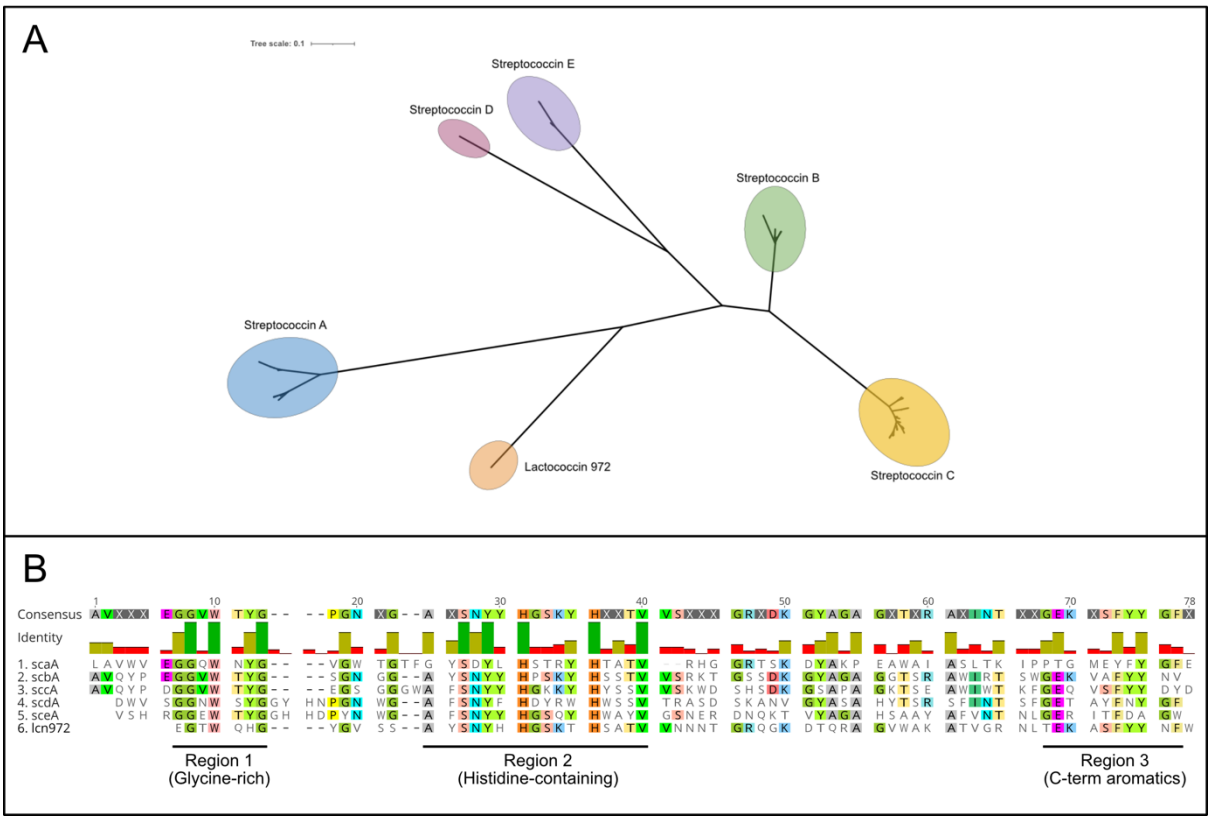

Figure 4.1: Comparisons of amino acid sequences of toxin genes from the five pneumococcal streptococcins plus lactococcin 972. Panel A: Phylogenetic tree of unique amino acid streptococcin gene sequences and the reference lactococcin 972 sequence (UniProt accession O86283). Panel B: Toxin gene amino acid sequence alignments for each streptococcin and the reference lactococcin 972 sequence. Predicted signal peptide regions were removed prior to alignment. Amino acids were indicated if there was a consensus residue at that position and coloured according to the residue. Overall identity at each position is indicated by the vertical bars (green: 100% identity, yellow: 30 - 99.9% identity, red: 0 - 29.9% identity).

Table 4.3: Frequency of conserved amino acids in streptococcins A-E.

Chemical group	Amino acid	Conserved residues (n)
Hydrophobic (aliphatic and aromatic)	Tyrosine	5
	Glycine	5
	Alanine	4
	Phenylalanine	1
	Tryptophan	1
	Valine	1
Polar, positively charged	Histidine	2
Polar, uncharged	Serine	1
	Threonine	1
	Asparagine	1
Polar, negatively charged	Glutamate	1

Note: Amino acid residues were considered to be conserved if four or more of the reference streptococcins had the same residue in the same position. The chemical group refers to the side chain.

4.3.2 Immunity genes

4.3.2.1 Streptococcin cluster B and C genes encode ABC transporters

The B and C genes of each streptococcin gene cluster were predicted to encode the transmembrane and nucleotide binding domains of an ABC transporter, respectively. InterPro predicted that the proteins are part of the bacteriocin-associated membrane protein (IPR006541) and putative bacteriocin-export ABC transporter, lactococcin 972 group (IPR019895) families, both of which were identified as part of the lactococcin 972 biosynthetic gene cluster. This prediction was supported by various features identified in both the putative transmembrane and nucleotide binding domains. The transmembrane domains were predicted to have seven transmembrane helices, which is consistent with other ABC transporter transmembrane domains (Figure 4.2A).[412]

All the transmembrane domains exhibited the same predicted topology: two large extracellular regions, the first immediately after the predicted signal peptides and the second between the transmembrane helices 4 and 5, which are likely to fold into globular extracellular domains. The amino acid sequences of the reference streptococcin cluster B genes were used to query the InterPro database. This did not return any conserved domains or recognised motifs that could be used to predict the function of the protein. Notably, they did not possess a C39 cysteine protease domain (InterPro record IPR005074) for cleavage of leader peptides with Gly_2 motifs,[413] and none of the reference sequences encode a single cysteine residue. This suggests that the streptococcin-associated ABC transporters are not specialised bacteriocin exporters, as is the case in other bacteriocin systems such as the pneumococcal Blp clusters.[331]

The streptococcin cluster C genes possessed the Walker A, B, and C motifs essential for the ATPase activity of nucleotide binding domains (Figure 4.2B, Table 4.4).[414] These regions were highly conserved across all the streptococcin-associated and lactococcin 972-associated nucleotide binding domains.

Table 4.4: Motifs in the nucleotide binding domains of ABC transporters and putative roles in activity.[414]

Motif name	Typical amino acid sequence	Function
Walker A motif (P loop)	GxxGxxGKST	ATP interaction, Mg^{2+} coordination
Walker B motif	4 aliphatic residues followed by 2 negatively charged residues	Mg^{2+} coordination, H_2O interaction
Walker C motif (ABC transporter signature)	LSGGEQQRIA	ATP interaction

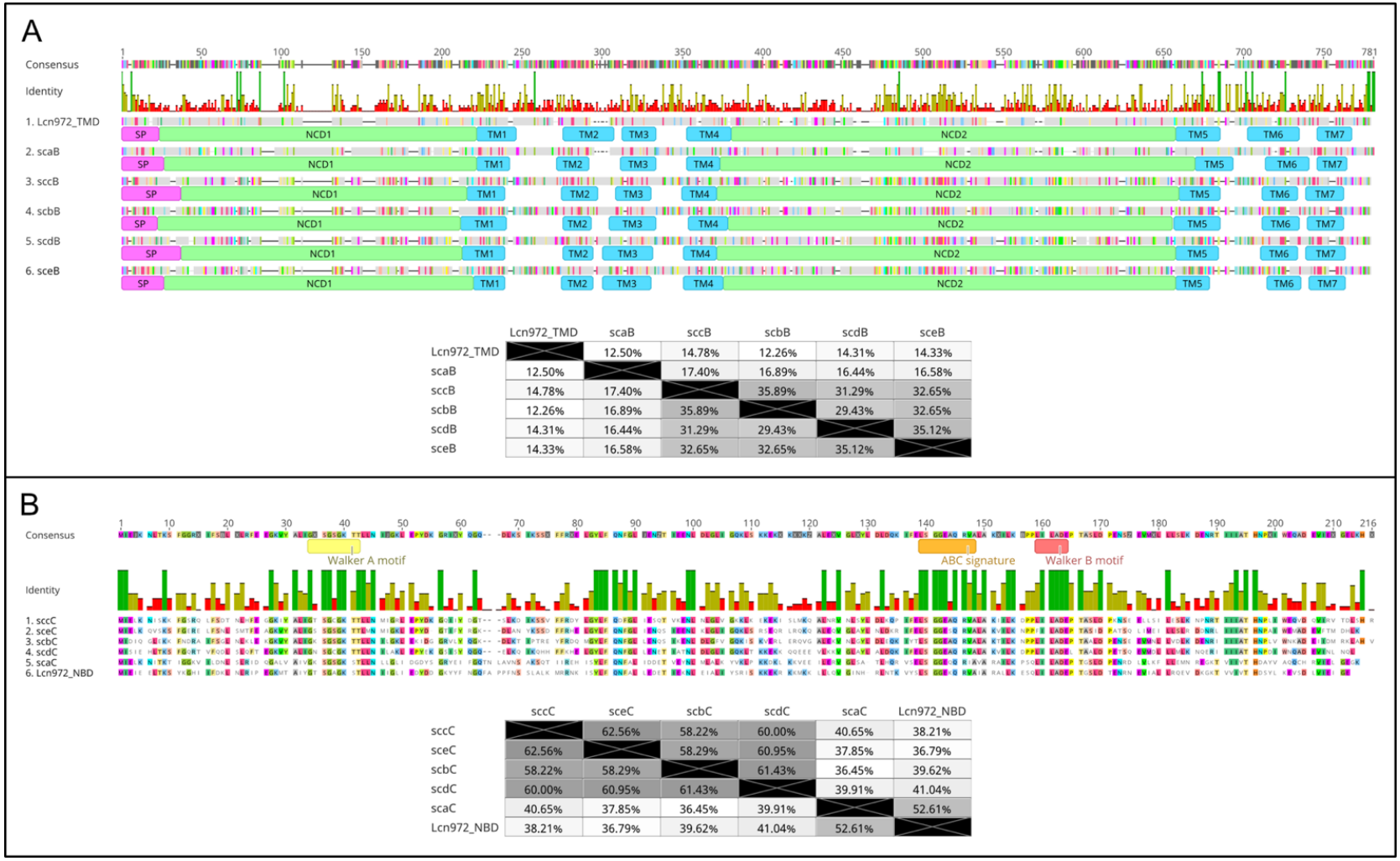

Figure 4.2: Amino acid sequence alignments of the transmembrane domains (panel A) and nucleotide binding domains (panel B) of the reference streptococcin- and lactococcin 972-associated ABC transporter genes. Panel A: Annotation blocks delineate phobius-predicted signal peptides (SP, pink), non-cytoplasmic domains (NCDs, green) and transmembrane helices (TMs, blue). Panel B: Annotation blocks on the consensus sequence indicate the positions of conserved motifs that are characteristic of ABC transporter nucleotide binding domains.

Figure 4.2 (continued): Amino acids are indicated when there was a consensus residue at that position and coloured by the amino acid at that position. Mean percentage identity at each site shown by the vertical bars (green: 100%, yellow: 30-99%, red: 0-29.9%). Pairwise percentage identity is shown by the matrices and coloured with grey scale (black: 100% identity, white: minimum percentage identity in the matrix).

4.3.2.2 Sequence diversity of the streptococcin-associated transmembrane and nucleotide binding domains

When the reference alleles of the streptococcin A-E and lactococcin 972 transmembrane domain genes were aligned, only 16 amino acid positions were identical (2.1% of the total alignment length, Fig 4.2A). There were no clearly conserved sequence motifs visible in the alignment, unlike those observed for the reference toxin alleles (Figure 4.1B).

When the reference sequences of the streptococcin nucleotide binding domains (C genes) were compared, there was much higher sequence conservation (Figure 4.2B). The Walker motifs in the C genes were notable as sites of particularly high conservation, indicating a low tolerance for SNPs within these positions. As these motifs are essential for the ATPase activity of the nucleotide binding domains, the conservation suggests that the ATPase functionality is important to the overall function of streptococcin-associated ABC transporters, and that mutations in these motifs are selected against.

Streptococcin A-associated transmembrane domain amino acid sequences had the lowest percentage of identical sites (67.6%), and streptococcin C-associated transmembrane sequences had the lowest mean pairwise identity (96.6%, Table 4.5). Variable amino acids were unevenly distributed throughout the transmembrane domain sequences and were clustered in the regions predicted to form non-cytoplasmic domains (Figure 4.3). With the exception of streptococcin E, the streptococcin-associated nucleotide binding domains (C genes) showed lower amino acid sequence diversity than the transmembrane domains (B genes) (Table 4.5). This was particularly clear when the percentage of identical sites were compared: the nucleotide binding domains consistently had a higher percentage of sites with identical amino acids.

Table 4.5: Diversity of the amino acid sequences of the streptococcin-associated ABC transporter transmembrane domain genes (B genes) and nucleotide binding domain genes (C genes).

Streptococcin	Gene	Unique amino acid alleles (n)	Mean pairwise identity (%)	Identical sites (%)
A	*scaB*	99	97.4	67.6
	scaC	31	98.3	87.4
B	*scbB*	187	98.8	79.0
	scbC	37	98.8	86.9
C	*sccB*	172	96.6	76.3
	sccC	41	98.5	85.5
D	*scdB*	2	-	-
	scdC	1	-	-
E	*sceB*	166	98.6	82.7
	sceC	58	98.3	79.7

4.3.2.3 Streptococcin-associated ABC transporters have a putative role in immunity

Streptococcin toxins appear to be targeted to the general secretion pathway (*i.e.*, using the SecYEG translocon) by an N-terminal signal peptide. It is therefore unlikely that the streptococcin-associated ABC transporters have a role in toxin export. Other classes of genes typically found in bacteriocin biosynthetic gene clusters are post-translational modification enzymes and genes to confer immunity against the bacteriocin on the producing strain. The streptococcin-associated ABC transporters are unlikely to be involved in modification as they do not possess any catalytic motifs associated with post-translational protein modification, and the streptococcins are predicted to be unmodified. It is therefore most likely that the ABC transporters have a role in immunity, which is also the case in multiple other bacteriocins and is the proposed role of the homologous lactococcin 972-associated ABC transporter.[392]

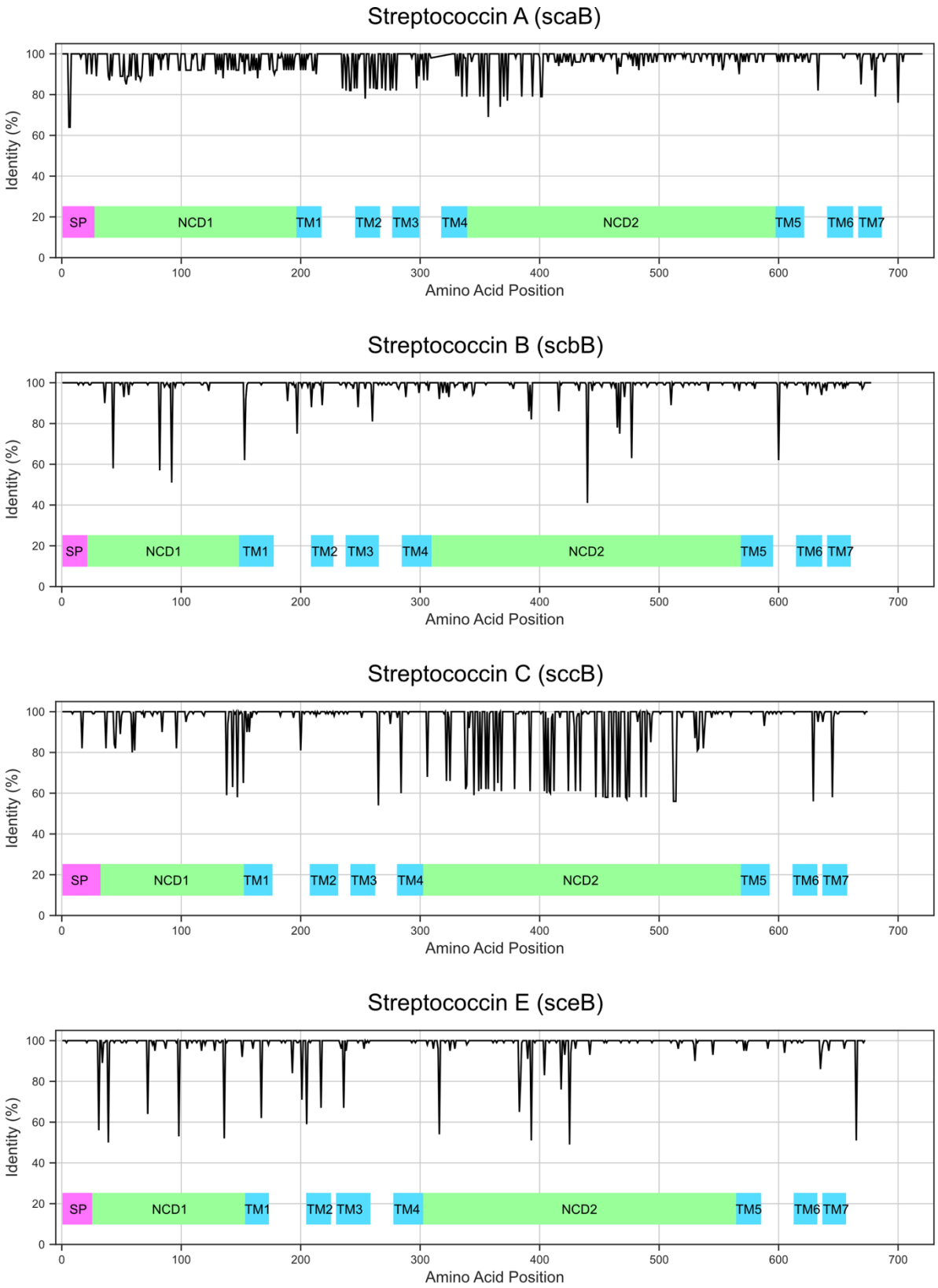

Figure 4.3: Amino acid sequence identity at each position of the B genes from each unique streptococcin cluster. The coloured bars indicate the location of the predicted signal peptides (pink), non-cytoplasmic domains (green) and transmembrane helices (blue), based upon phobius predictions for each reference gene sequence.

4.3.3 AlphaFold structural predictions of streptococcin-associated genes

AlphaFold structural predictions have been made publicly available for the predicted proteomes of a number of important bacteria, including pneumococcus. The structural predictions used sequences from the R6 pneumococcal reference genome (ATCC accession BAA-255, Table 4.6). Eight streptococcin-associated genes from four different streptococcin clusters are represented in the R6 genome, but streptococcin C was the only complete streptococcin cluster (with all three expected genes).

Table 4.6: AlphaFold structural predictions for streptococcin genes from the pneumococcal R6 genome sequence.

Streptococcin	Protein	Predicted function	UniProt accession
A	ScaB	ABC transporter transmembrane domain	Q8DRI8
	ScaC	ABC transporter nucleotide binding domain	Q8DRI7
B	ScbB	ABC transporter transmembrane domain	Q8DND1
C	SccA	Bacteriocin toxin	Q8DQM5
	SccB	ABC transporter transmembrane domain	Q8DQM4
	SccC	ABC transporter nucleotide binding domain	Q8DQM3
E	SceB	Truncated ABC transporter transmembrane domain	Q8CY97
	SceC	ABC transporter nucleotide binding domain	Q8DNF2

4.3.3.1 Streptococcin toxin predicted structure

The streptococcin C toxin (SccA) structural prediction is similar to the previously published structure of lactococcin 972 (Figure 4.4). The structure has an overall β-sandwich architecture, which was predicted with a high degree of confidence according to pLDDT values and the predicted aligned error. The signal peptide structure was predicted with a lower degree of confidence to form an extended tail (Figure 4.4A). This is typical of flexible or disordered regions of proteins due to the greater range of stable conformations such regions could take. As the signal peptide is likely to be cleaved from the exported streptococcin, the low certainty of its structure should not detract from the prediction of the remainder of the streptococcin structure. Residues of interest (as identified in Section 4.3.1) were annotated on the AlphaFold structure. The highly conserved histidine residues (labelled in Figure 4.4 as H68 and H75) were predicted to be close to one another and largely buried within the structure. The C-terminal aromatic residues are located on the final strand of the β-sandwich and are likely exposed on the surface of the protein. The other aromatic residues are largely found on the opposite face of the structure to the conserved C-terminal aromatics (Figure 4.4C).

4.3.3.2 Immunity complex predicted structure

The AlphaFold prediction of the pneumococcal R6 SccB protein shows a bundle of eight α-helices (including the predicted signal peptide) with two large globular domains (Figure 4.5A). This is consistent with the predicted function of this gene as an ABC transporter transmembrane domain. The phobius-predicted topology was overlaid on the predicted structure: the predicted helices coincided with the transmembrane regions of the protein, and the globular domains coincided with the non-cytoplasmic domains. The AlphaFold prediction of the R6 SccC protein shows a globular protein consisting

largely of α-helical elements with some β-strands (Figure 4.5B). The structure of the nucleotide binding domain was predicted with a high degree of confidence, and the confidence in the transmembrane domain prediction was lower, although the pLDDT values were still > 70 for the majority of the structure. The predicted structures of the *sccB* and *sccC* gene products are consistent with experimentally determined structures of ABC transporters.[412]

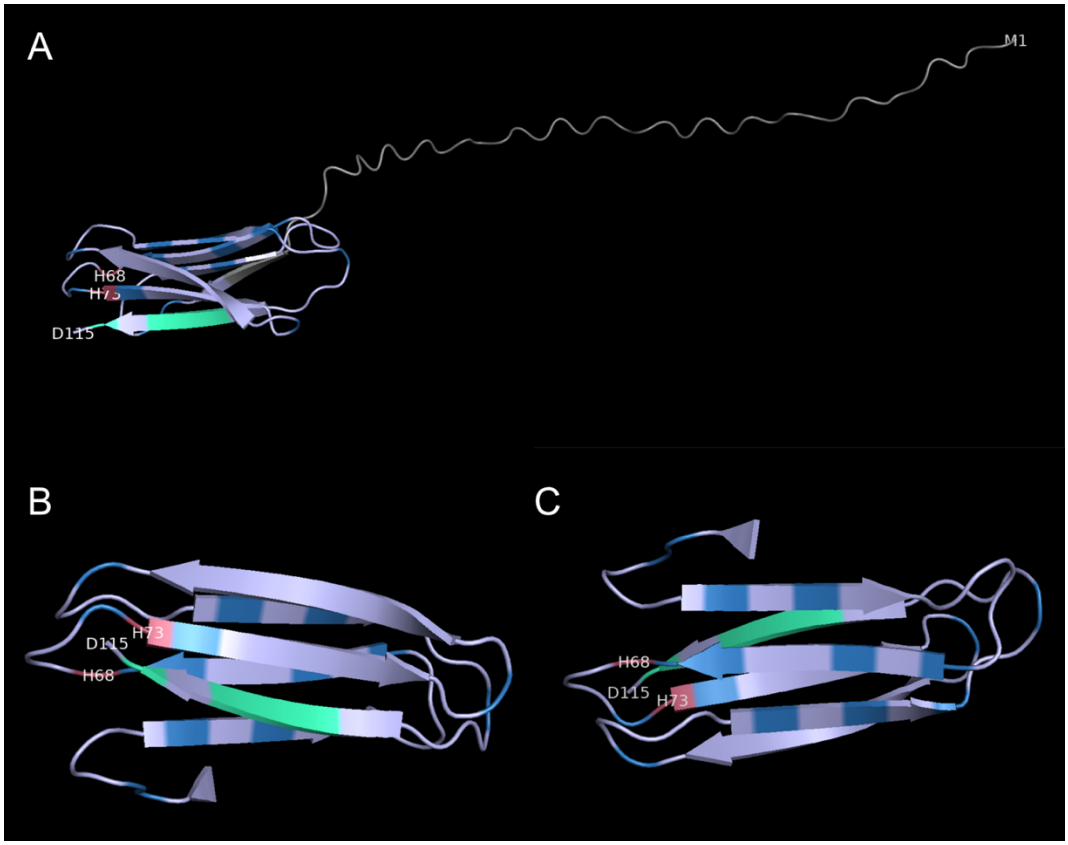

Figure 4.4: Annotated cartoon of the AlphaFold structural prediction of SccA from pneumococcal strain R6. Panel A: The full structure of SccA (UniProt accession Q8DQM5) including the signal peptide. Panel B: A view of the structure without the predicted signal peptide. Panel C: Structural prediction as in panel B but rotated 180° to show the opposite face of the β-sandwich fold. The signal peptide (as predicted by phobius) is coloured grey, histidine residues are red, aromatic residues are blue, conserved C-terminal residues are cyan, and other residues are coloured lilac. N- and C-terminal residues (M1 and D115) and conserved histidine residues (H68 and H73) have text labels.

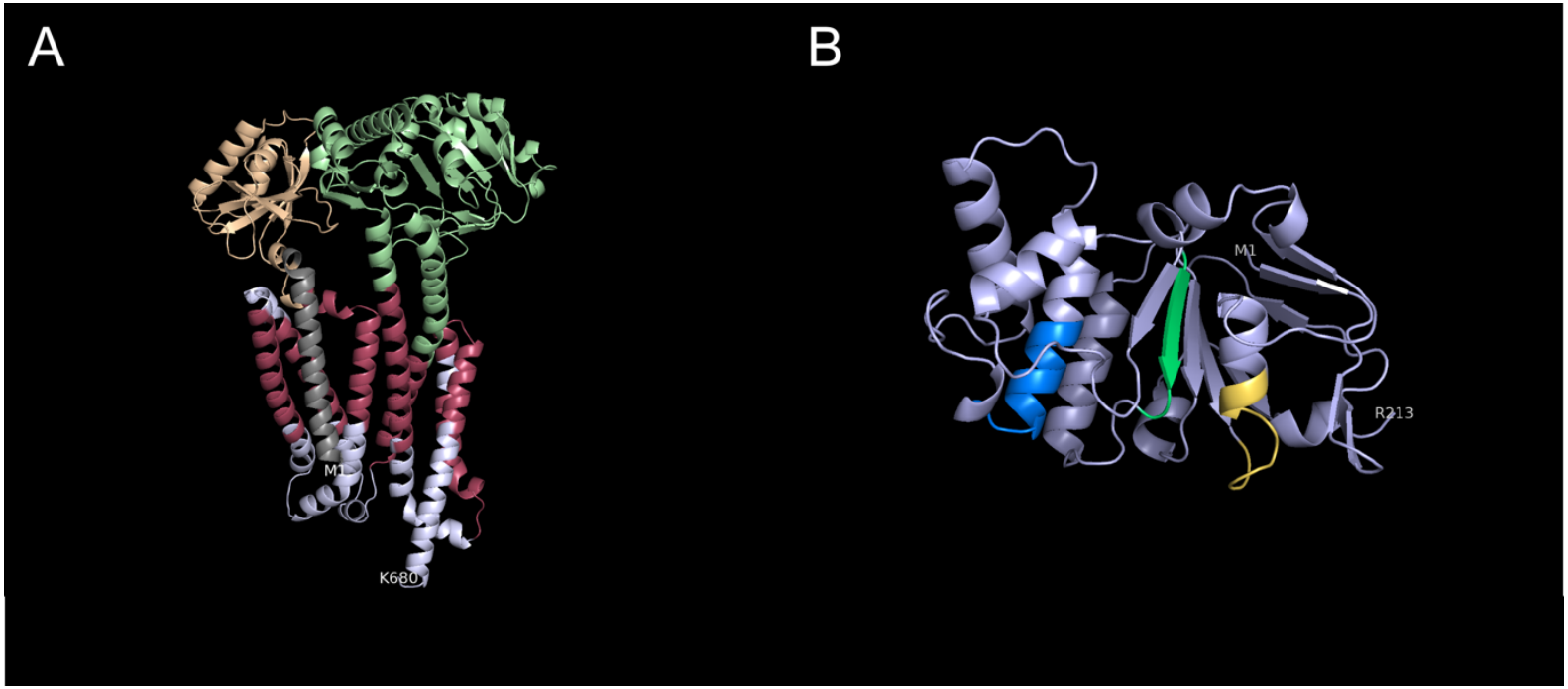

Figure 4.5: Annotated AlphaFold structural predictions of SccB and SccC from pneumococcal strain R6. Panel A: SccB (UniProt accession Q8DQM4), transmembrane domain of an ABC transporter. Seven phobius-predicted transmembrane helices are coloured red, the phobius-predicted signal peptide is grey, non-cytoplasmic domain 1 is yellow, and non-cytoplasmic domain 2 is coloured green. Panel B: SccC (UniProt accession Q8DQM3) nucleotide-binding domain of an ABC transporter. The Walker A motif is shown in yellow, Walker B motif in green and the Walker C motif is shown in blue. N- and C-terminal residues are labelled and numbered according to their positions in the sequence.

4.3.4 A model of streptococcin function

By integrating the structural and functional predictions of the streptococcin-associated genes, I propose a generalised model for the function of the streptococcin gene clusters (Figure 4.6). The toxin genes encode unmodified extracellular proteins that are secreted *via* the SecYEG machinery. The toxins are expected to be globular and possess conserved aromatic residues that mediate an interaction with lipid II, as proposed for lactococcin 972. The streptococcin B and C genes encode the transmembrane domain and nucleotide binding domain of an ABC transporter with a putative role in streptococcin immunity. Functional ABC transporters typically comprise two transmembrane domains and two nucleotide binding domains. The model presented here assumes that this is the case and that both subunits are encoded by the same gene, although in the absence of experimental data the possibility of association with a transmembrane domain and nucleotide binding domain encoded by genes separate from the streptococcin cluster cannot be excluded.

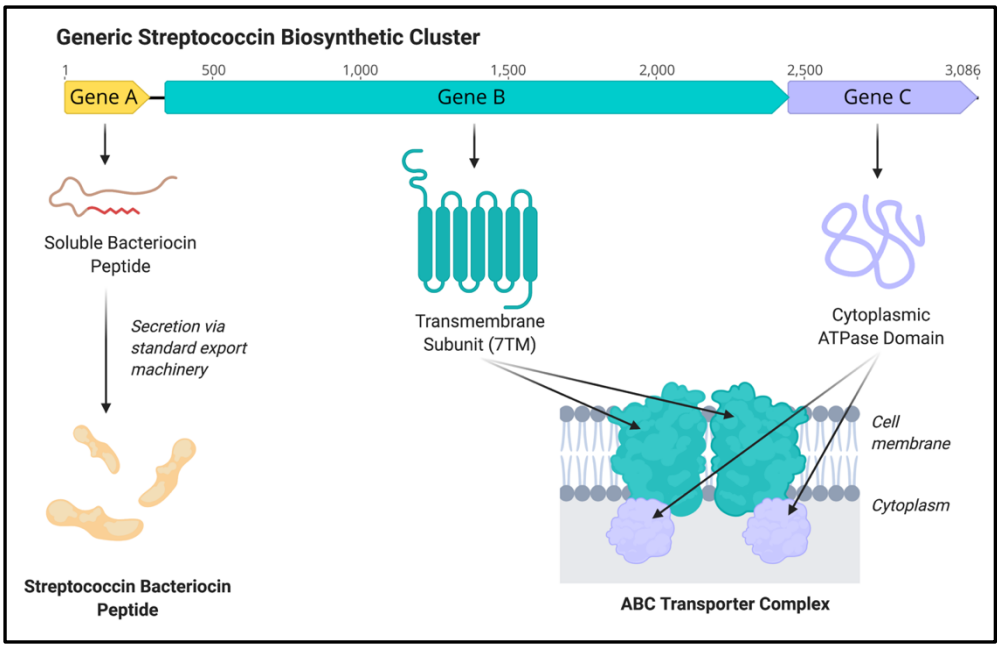

Figure 4.6: A proposed model for the function of streptococcin cluster gene products. Gene cluster based on the streptococcin A reference cluster[1] and shown to scale. Predicted extracellular domains of the ABC transporter complex not shown.

4.4 Discussion

4.4.1 Mechanism of streptococcin toxicity

These analyses of amino acid sequences have shown that all pneumococcal streptococcins have sequence motifs in common with lactococcin 972, and aromatic residues were highly conserved. Therefore, it is likely that the streptococcins have a similar mechanism of action as lactococcin 972, *i.e.* an interaction with lipid II to prevent cell wall synthesis, which weakens the cell envelope and results in morphological changes and eventual cell death. This is similar to the mechanism of the beta-lactam family of antimicrobials, which also target peptidoglycan synthesis (Section 1.1.7). Cell wall disruption is also utilised by the nisin bacteriocin, which also interacts with lipid II.[257]

Many bacteriocins show high target specificity, often only killing other strains of the same species.[234] Lipid II is a crucial intermediate in cell wall peptidoglycan synthesis across all bacterial genera,[14] and therefore an interaction with lipid II alone would be expected to kill bacteria indiscriminately. In the streptococcins (and in lactococcin 972) it is likely that a second, non-lipid II binding partner in proximity to lipid II determines the specificity of any interaction. As yet, both the mechanism and the specificity of streptococcin-mediated killing is unknown.

Amino acid sequence conservation among the streptococcin toxins has been observed in two large genomic datasets sampled from Iceland and Kenya (Figure 4.1, Table 4.2). Although these datasets represent large and diverse pneumococcal genetic lineages and serotypes, they are not entirely representative of the global diversity of the pneumococcal population. It is possible that the sequence conservation and functional predictions reported in this chapter are not universal across these loci. In order to

address this, the genes should be annotated in a more widely sampled pneumococcal dataset, such as the pneumococcal genome library (pubmlst.org/organisms/streptococcus-pneumoniae/pgl).

4.4.2 Streptococcin immunity

The model of streptococcin immunity where the B and C genes together form a multimeric immunity complex advances the previous functional predictions for these genes. When the streptococcin gene clusters were first identified, the B and C genes were annotated as an immunity gene and a transporter gene, respectively, based on automated annotations and bacteriocin database screens.[339] The updated model is in agreement with experimentally-determined features of the homologous bacteriocin, lactococcin 972: the streptococcins do not appear to be post-translationally modified, they are larger than most bacteriocin peptides and of a similar size to lactococcin 972 (9-10kDa), and they are exported *via* the general secretion pathway, rather than by a dedicated transporter.

4.4.2.1 Mechanism of immunity

Each streptococcin toxin gene is associated with a pair of genes encoding an ABC transporter, which is likely to function as an immunity complex, protecting the producing strain from the activity of its own streptococcin. ABC transporters are widespread among all domains of life and play a major role in antimicrobial resistance by functioning as efflux pumps to remove intracellular antimicrobials from the cytoplasm, for example the multidrug pump Sav1866 from *Staphylococcus aureus* and the macrolide efflux pump encoded on the *mefB/mel* operon in pneumococcus.[195,415] Moreover, ABC transporters are commonly associated with bacteriocins both as dedicated export systems and as

immunity genes.[412] Assuming that the streptococcin bacteriocin functions extracellularly, a role as an efflux pump seems unlikely. The transporter could instead import the toxin into the cell to remove it from its site of action (presumably the cell wall). Alternatively, the role of the transporter could be not to transport the streptococcin at all, rather to present an extracellular binding or protease domain, and sequester or degrade any toxin threatening to harm the producer.

A final suggestion for the involvement of the ABC transporter in streptococcin immunity is an indirect role in cellular signalling. ABC transporters have been shown to play a role in cellular signalling by detecting an extracellular ligand and transducing the signal to an intracellular one- or two-component signalling system to generate a transcriptional response.[416] Such a mechanism has been found in bacteriocin resistance systems harboured by strains that do not produce the bacteriocin and do not rely on the specific immunity proteins associated with the bacteriocin. The most well-studied of these is the BceRS-BceAB family of transporters, which confer bacitracin resistance in *Bacillus subtilis* and have homologues in bacteriocin resistance systems in a number of firmicutes.[417] One BceRS homologue, CesSR, senses cell envelope stress in lactococcin 972 and is implicated in lactococcin 972-resistance.[393,418]

4.4.3 Limitations of functional predictions

The model presented in this chapter has limitations that should be considered. All functional and structural predictions that rely on homology to existing experimental data will be biased by the biological systems that have been studied and corresponding data deposited in the PDB and InterPro databases. Such databases are more likely to be skewed towards proteins that are amenable to experimental structural biology

techniques. Secondly, the analyses described here assume that the streptococcin gene products have a single stable structure, rather than multiple, functionally relevant conformations, and that they are not found in larger multi-protein complexes that affect their final structure. Finally, the predictions do not consider the presence of post-translational modifications or the association of co-factors, both of which could also have an impact on protein structure.

The only way to address these limitations is to confirm structural and functional predictions of the streptococcin genes experimentally. The streptococcin toxins are attractive candidates for experimental work: they appear to be unmodified and soluble, so a heterologous expression and purification protocol would be an appropriate approach for their isolation. Following their isolation, a standard antimicrobial susceptibility assay would be an appropriate way to test activity against a wide panel of strains. This approach is addressed in Chapter 6. The immunity genes are expected to encode a membrane-associated transporter complex, which presents a greater challenge to experimental design and study. An alternative approach would be genetic manipulations of resistant and susceptible strains to determine the specific role of these genes.

4.4.4 Conclusions

In this chapter, I have presented a functional model for the streptococcin biosynthetic gene clusters, informed by experimental work in a homologous system and structural and functional predictions of gene sequences. This model represents an advancement in our understanding of the streptococcins as bacteriocins and can be used to inform future study design.

A general model for the functionality of the streptococcin clusters is useful in order to contextualise genomic analyses. Results presented in Chapter 3 showed that streptococcin biosynthetic gene clusters are not always complete: streptococcins B and E were often observed as partial clusters, with one gene entirely absent (Figure 3.3A). With rational predictions for the role of each gene, the potential consequences of these partial clusters can be considered in the context of bacterial populations. This is addressed in the next chapter.

The model can also inform the generation of specific hypotheses to be tested experimentally. Beyond simply confirming the activity of the streptococcin toxins (explored in Chapter 6), experimental work could examine the role of the highly conserved residues in the toxin genes, potential interactions between toxin and immunity proteins, and the specificity of the immunity systems to their toxins (discussed in Chapter 7). Finally, a clear understanding of the structure and function of the streptococcins could facilitate the rational design of novel therapeutics to overcome the increasing problem of antimicrobial resistance.

5 Streptococcin Clusters are Widespread and Heterogeneous in Pneumococci and in Oral Streptococci

The non-pneumococcal streptococcal genomic dataset used in this chapter was generated by Dr Melissa Jansen van Rensburg and Femke Ahlers as described in Section 2.1. Pneumococcal bacteriocins have been studied previously in a different non-pneumococcal streptococcal database in two unpublished master's degree theses, the first by me in 2017 and the second by Hannes Hagson in 2020. Some of the results in this chapter were included in a poster I presented at EuroPneumo 2019. Other findings in this chapter will be presented as a poster at the 12th International Symposium on Pneumococci and Pneumococcal Disease (ISSPD-12) in Toronto, Canada, in June 2022. The conference abstracts are provided in Appendix Section 9.5.

5.1 Introduction

5.1.1 Streptococcins in pneumococcus

Results presented in Chapter 3 showed that streptococcins B, C and E were ubiquitous (or nearly so) in pneumococcal genomes sampled from Iceland and Kenya, and streptococcin A was also present in more than 80% of these (Figure 3.3A). Previous work

has found that streptococcins B and E are often detected as partial clusters (lacking a single gene from the reference gene cluster) (Figure 3.3A).[339] This might indicate that there are differences in the functionality of the gene clusters, which should be considered in the context of the generalised model of streptococcin function presented in Chapter 3. Moreover, previous work on the pneumococcal bacteriocins has suggested that gene clusters may be exchanged horizontally within the pneumococcal population,[338,339] and there is evidence that the homologous lactococcin 972 gene cluster from *Lactococcus lactis* is mobile.[419] The role of horizontal genetic exchange in driving the diversity of the streptococcins has not yet been investigated.

5.1.2 Streptococcins in non-pneumococcal streptococci

Pneumococcal bacteriocin gene clusters have been studied in genomes of non-pneumococcal streptococci (NPS) previously.[339] The streptococcins were commonly harboured by viridans streptococci, which are closely related to pneumococci and are often found in the respiratory tract microbiome. The streptococcins from NPS species have not yet been directly compared to their pneumococcal counterparts in terms of cluster composition, sequence diversity, and predicted functionality. Such studies would improve our understanding of the evolutionary origin of the streptococcins and of the role that they may be playing a role in inter-species competition in the respiratory tract.

5.1.2.1 Inter-species genetic exchange between pneumococci and non-pneumococcal streptococci

Genetic exchange between pneumococci and NPS species, especially *S. mitis* and *S. pseudopneumoniae*, has influenced the pneumococcal genome.[29] Notably, the PBP alleles that confer penicillin resistance are mosaic genes that include sequence that originated

from *S. mitis* and *S. oralis* (Section 1.1.7).[209,210] Exchanges with other NPS species have also been relevant to pneumococcal disease: a non-encapsulated pneumococcal lineage has acquired an adhesin encoded on an ICE originating in *S. suis* that facilitates attachment to ocular epithelium (*sspB*), resulting in conjunctivitis.[420] Finally, these exchanges are not unidirectional: the acquisition of the pneumococcal *cps* locus has been observed in examples of *S. mitis, S. oralis* and *S. infantis* (although there is evidence suggesting that pneumococci more commonly receive sequences from, rather than donate to, *S. mitis*).[10,385,386] It is therefore possible that bacteriocin clusters or individual genes are also exchanged between *Streptococcus* spp. in the nasopharynx, and that this may drive the diversification of pneumococcal bacteriocins.

The composition of the nasopharyngeal microbiome is influenced by inter-species competition. For example, commensal species from the genera *Corynebacterium* and *Dolosigranulum* are believed to exclude pneumococci from the nasopharynx, suggesting that species from these genera can out-compete colonising pneumococci.[138] Viridans streptococci exhibit relatively low abundance (1-15%) in the nasopharyngeal microbiome when they are present, in contrast to pneumococci, which tend to dominate.[137] This observation is indicative of differing competitive strategies. The contribution of bacteriocins, including the streptococcins, to inter-species competitive interactions in the nasopharynx has not yet been studied.

5.1.3 Aims

Previously, I showed that four out of five of the streptococcins are very widely distributed among pneumococcal genomes recovered from Iceland and Kenya, and that two streptococcins (B and E) were commonly observed as partial gene clusters (Figure 3.3A).

I also found that streptococcin A and E presence was variable within some CCs. In Chapter 4, I used functional predictions of the individual streptococcin genes to propose a generalised model of streptococcin functionality. Here, I used the two pneumococcal genomic datasets and a new NPS genomic dataset to investigate the following:

- The diversity of the streptococcin gene clusters in pneumococci, including both sequence diversity and their predicted functionality (informed by the general model developed in Chapter 4),
- The distribution of streptococcin gene clusters both within pneumococci and in the broader *Streptococcus* genus,
- The evidence for the horizontal exchange of streptococcin clusters and individual genes.

5.2 Materials and methods

5.2.1 Dataset compilation and quality control

The previously described Icelandic and Kenyan pneumococcal genomic datasets and the NPS genomic dataset were analysed further (Section 2.1).

5.2.2 Streptococcin gene annotations and cluster categorisation

Streptococcin genes were annotated in the NPS dataset as described previously (Section 2.2). The streptococcin gene annotations in the pneumococcal genomes were generated previously. All observed alleles of each streptococcin gene were categorised: if the allele sequence contained a disruption to the coding sequence that would be expected to

prevent the expression of a typical product (such as an internal stop codon or frameshift) the allele was classed as a pseudogene. Allele categories were recorded in manatee.ipynb. The combination of typical genes and pseudogenes was used to categorise the observed streptococcin gene clusters (Figure 5.1). If all the genes were present, the cluster was 'full'. If any gene was missing entirely, and the remaining genes of the cluster were complete coding sequences, the cluster was 'partial'. If the cluster had a mix of typical genes and pseudogenes, it was classed as 'disrupted toxin' (if the toxin gene was a pseudogene and both immunity genes were typical), 'disrupted immunity' (if the toxin gene was typical and one or both immunity genes were disrupted), or 'degrading' (if both the toxin and the immunity genes were disrupted or absent). Degrading clusters encompassed all variations where the cluster was reliably detected, but the assumption was that they do not encode a functional toxin nor a functional immunity system.

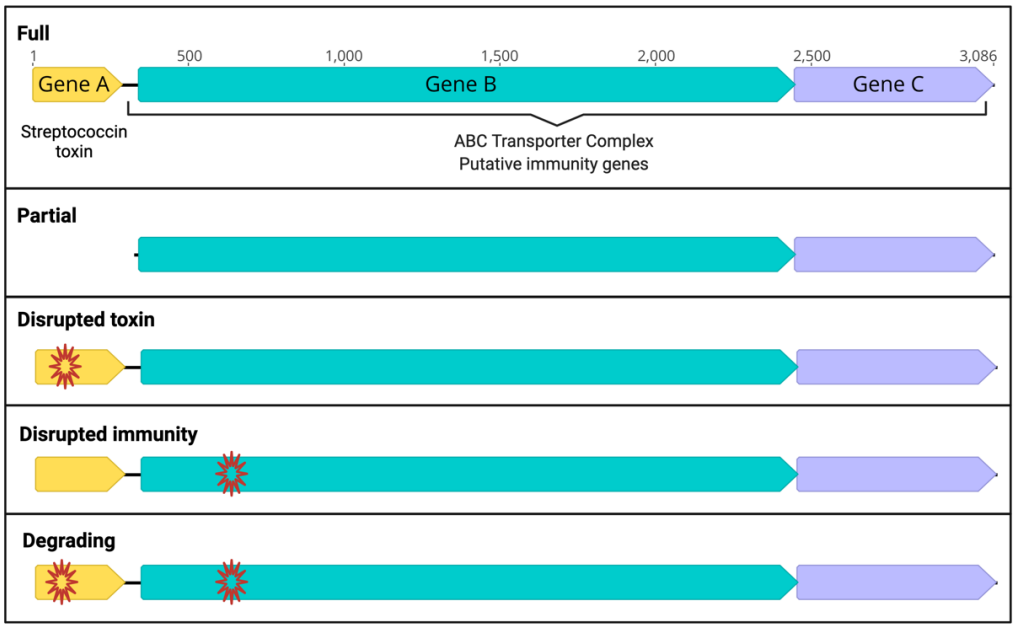

Figure 5.1: Illustration of the functional categories of the streptococcin gene clusters. Red stars represent a disruption to the coding sequence such as an internal stop codon caused by a single nucleotide polymorphism or a frameshift. Disruptions in the 'disrupted immunity' and 'degrading' categories could be in the B or C gene.

Fragment clusters and non-contiguous clusters were identified as described previously and excluded from analysis (Section 2.2).

5.2.3 Sequence comparisons

5.2.3.1 Streptococcin cluster sequences

Streptococcin gene clusters were described by their allelic profiles, *i.e.* the combination of alleles for the three constituent genes. Whole cluster sequences were generated by concatenating the sequences of each allele in the allelic profiles. If any genes were annotated but not given an allele designation (*e.g.* interrupted by a contig break, or possessed a string of ambiguous reads), the cluster sequence was not generated and the cluster was excluded from any analyses requiring sequence data. Cluster sequences were also generated including intergenic and flanking sequences using the BLAST plugin within BIGSdb to query the required genomes with the reference cluster sequence. Hits were exported with flanking sequence, and duplicate sequences (*i.e.*, identical streptococcin genes and identical flanking sequences) were removed.

5.2.3.2 Multiple sequence alignments and phylogenetic trees

Multiple sequence alignments were performed, and phylogenetic trees were estimated, to compare the streptococcin whole cluster sequences as described previously (Section 2.3.1). A neighbour joining phylogenetic tree of all genomes from the NPS datasets was estimated based on rMLST allele sequences using the phylogenetic tree building tool within BIGSdb.

5.2.4 Streptococcin diversity

Diversity was assessed using sequence alignments and using Simpson's index of diversity. Streptococcin D was excluded from diversity calculations as it was not well represented in the pneumococcal datasets.

5.2.4.1 Simpson's index of diversity

Simpson's index of diversity is a widely used statistic that describes the diversity of a categorical variable and can be applied to alleles or allelic profiles within a dataset.[421] The diversity (D) of a dataset is described as follows:

$$D = 1 - \frac{\sum n_j(n_j - 1)}{N(N-1)}$$

Where:

$$N = total\ size\ of\ dataset$$
$$n_j = size\ of\ the\ jth\ category$$

D values range between 1 and 0, where values closer to 1 reflect higher diversity. Approximate 95% confidence intervals on D can be given as two standard deviations away from D, using the approach published by Grundmann *et al.*[421] Standard deviation (σ) calculated as follows:

$$\sigma = \sqrt{\frac{4}{n}\left(\sum \pi_j^3 - \left(\sum \pi_j^2\right)^2\right)}$$

Where:

$$\pi_j = \frac{n_j}{n}$$

$$n_j = size\ of\ the\ jth\ category$$
$$n = total\ size\ of\ the\ dataset$$

5.3 Results

5.3.1 Streptococcin species distribution

5.3.1.1 Streptococcin gene annotations

The NPS dataset contained 1,825 genomes recovered from 55 NPS species (Figure 5.2, Appendix Table 9.11). The 15 genes associated with streptococcin clusters were annotated previously in the pneumococcal datasets. The genes were also annotated in the NPS dataset as described in Section 2.2. Fragment and non-contiguous streptococcin clusters were excluded from analysis in the NPS dataset as described previously for the pneumococcal dataset (Section 2.3.2, Appendix Table 9.12).

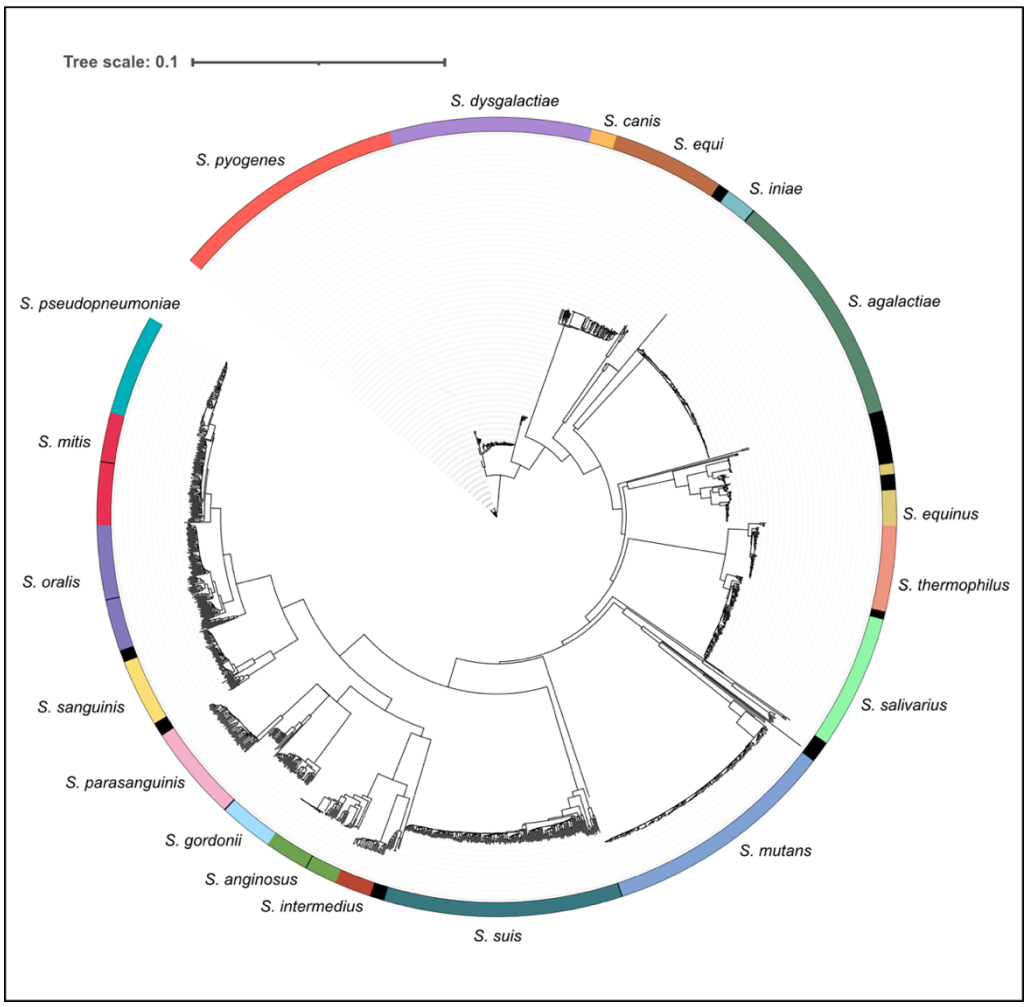

Figure 5.2: Neighbour joining tree based on rMLST gene annotations of the non-pneumococcal Streptococcus database. The 20 most common species are labelled. Full species list can be found in Appendix Table 9.11.

5.3.1.2 Streptococcins are widespread among pneumococci and some viridans streptococci

As shown in Chapter 3, among pneumococci, streptococcins B and C were ubiquitous, streptococcins A and E were very common (streptococcin A prevalence: 80%, streptococcin E prevalence: 96% - 99%), and streptococcin D was rarely observed (prevalence of 0.5% - 2.7%). All five streptococcins were also observed in the NPS dataset. Streptococcins A, B and C were restricted to species of mitis subgroup streptococci (*S. mitis, S. oralis* and *S. pseudopneumoniae,* Figure 5.3A). Streptococcins B and C were ubiquitous in *S. pseudopneumoniae.* Otherwise, streptococcin A, B and C prevalence in the NPS genomes was lower than in pneumococci.

Streptococcins D and E were observed in a greater range of NPS species. Streptococcin D was found in species of the mitis, sanguinis and anginosus subgroup viridans streptococci (*S. oralis, S. parasanguinis, S. gordonii, S. sanguinis* and *S. anginosus*) and were ubiquitous among *S. cristatus* genomes (n = 19). Streptococcin E was found in *S. mitis* and *S. pseudopneumoniae*, but was also found in *S. cristatus, S. parasanguinis* and *S. anginosus*, and was the only streptococcin observed in *S. suis* and *S. equi,* albeit rarely.

5.3.1.3 Multiple streptococcin clusters were present in some NPS genomes

The number of streptococcin clusters per genome was assessed (Figure 5.3B). Each pneumococcal genome had between two and five different streptococcin clusters, and the modal value was four. The range and mode number of streptococcins per genome varied by species. *S. pseudopneumoniae* was most similar to pneumococcus: genomes possessed between two and four streptococcins, and the mode was three. *S. mitis* and *S. oralis* genomes possessed between zero and three (*S. mitis*) or four (*S. oralis*) streptococcin

clusters. *S. sanguinis* and *S. anginosus* most commonly had no detectable streptococcin cluster, and rarely possessed a single one. *S. cristatus* usually possessed a single streptococcin cluster and rarely possessed two. *S. pseudopneumoniae* and *S. cristatus* were the only NPS species that always possessed at least one streptococcin cluster.

5.3.2 Heterogeneous composition of streptococcin clusters

Streptococcins were sometimes observed as partial clusters and pseudogenes were observed. A higher percentage of the toxin genes were pseudogenes than the putative immunity genes, and the streptococcin A toxin had the highest percentage of pseudogene sequences (56%, Table 5.1). The allelic profile of each streptococcin cluster was categorised by gene presence and whether they encoded complete coding sequences or pseudogenes (Figure 5.1).

5.3.2.1 Differences in streptococcin cluster composition

Among pneumococci, all five streptococcins were observed as full clusters in the Icelandic and Kenyan datasets (Figure 5.3C). Partial clusters, lacking the toxin gene, were observed for streptococcins B and E, and the disrupted toxin (*i.e.* pseudogene) profile was observed among streptococcins A, B, C and E. Among the NPS genomes, all streptococcins were detected as full clusters, although the prevalence of clusters in the full category was generally lower than in pneumococci (Figure 5.3C). Unlike in pneumococci, partial streptococcin A, C and D clusters were observed in NPS genomes.

Other functional categories of streptococcins were rare in genomes of all species. Degrading clusters, which do not encode a functional toxin or immunity complex, were most commonly observed for streptococcin E. In pneumococcal genomes, the majority of

these were caused by the insertion of streptolancidin G across the *sceB* gene.[339] The disrupted immunity category was rare among all streptococcins except for streptococcin C, where it was restricted to certain CCs (Table 5.2).

Table 5.1: Number of unique alleles observed at each streptococcin locus in the pneumococcal and NPS datasets.

Streptococcin	Locus	Number of alleles observed (% of total unique alleles)		
		Genes	Pseudogenes	Total
A	*scaA*	25 (44%)	32 (56%)	57
	scaB	243 (96%)	10 (4%)	253
	scaC	130 (96%)	5 (4%)	135
B	*scbA*	28 (82%)	6 (18%)	34
	scbB	328 (96%)	15 (4%)	343
	scbC	134 (99%)	2 (1%)	136
C	*sccA*	43 (50%)	43 (50%)	86
	sccB	340 (84%)	65 (16%)	405
	sccC	184 (97%)	6 (3%)	190
D	*scdA*	6 (55%)	5 (45%)	11
	scdB	53 (95%)	3 (5%)	56
	scdC	48 (100%)	0 (0%)	48
E	*sceA*	17 (74%)	6 (26%)	23
	sceB	246 (87%)	38 (13%)	284
	sceC	125 (98%)	3 (2%)	128

Table 5.2: The most common disrupted immunity profiles from streptococcin C in the pneumococcal datasets, including the distribution of these profiles in clonal complexes (CCs) and datasets.

Streptococcin C profile	n	CCs	Dataset(s)
2-156-2	217	CC217	Kenya
7-229-24	65	CC97	Iceland
8-130-13	65	CC138/176, CC338	Iceland, Kenya
8-223-13	55	CC138/176	Iceland
Other (22 profiles)	34	-	-

Note: the 'other' category 2 represents 2 profiles that were observed less than 5 times in total.

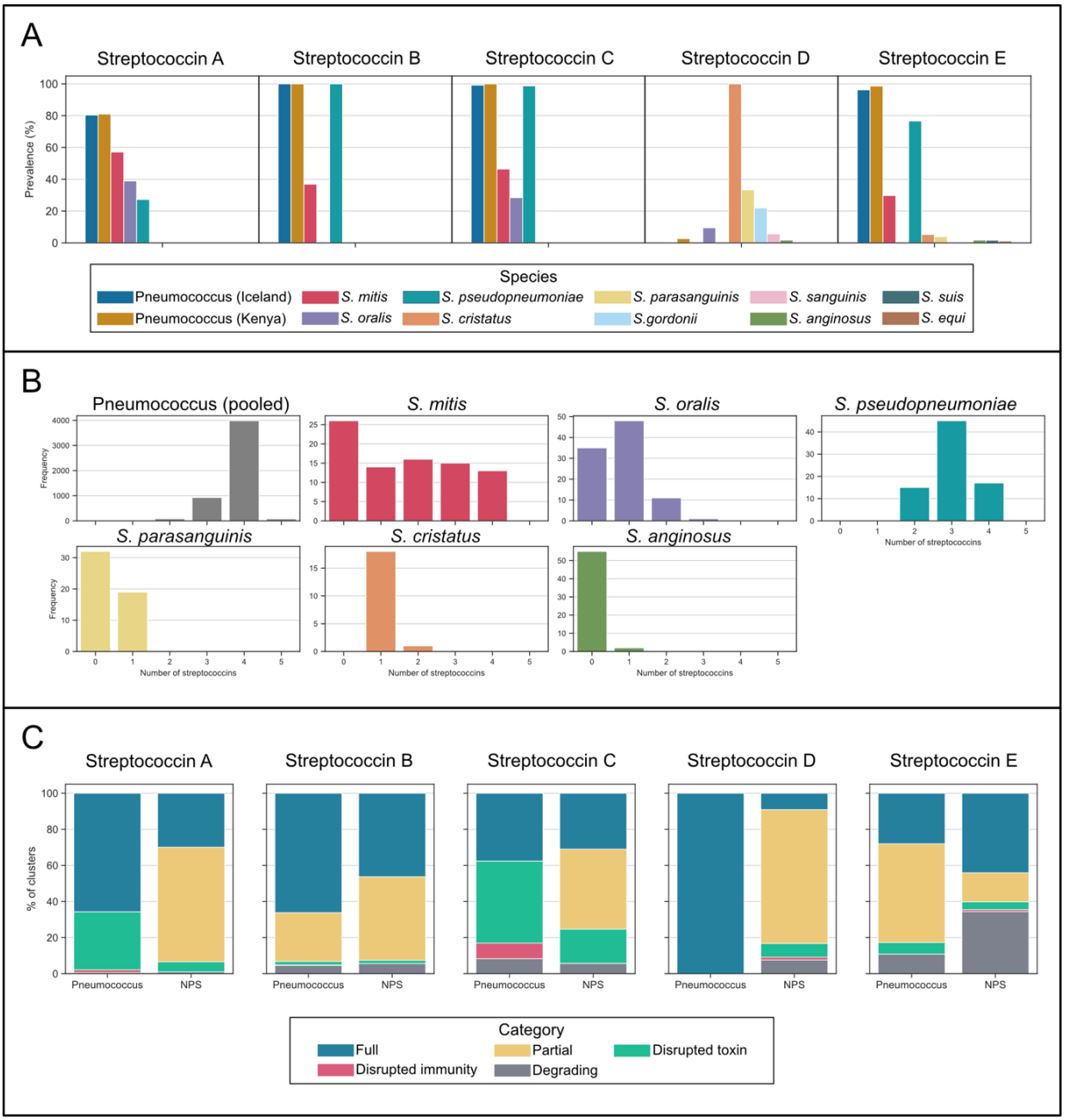

Figure 5.3: The prevalence and functional categories of streptococcin clusters in the pneumococcal and non-pneumococcal Streptococcus (NPS) genomic datasets. Panel A: The overall prevalence of the streptococcins by species. Species represented by fewer than 10 genomes in the NPS dataset were excluded, full data can be found in Appendix Table 9.13. Panel B: The number of different streptococcin clusters found per genome among pneumococci (Icelandic and Kenyan datasets pooled) and NPS species. NPS species shown were represented more than 10 times in the NPS dataset. NPS species in which only a single streptococcin type was observed were not included. Panel C: Functional categories of streptococcin clusters.

5.3.3 Streptococcin cluster sequence diversity and distribution within pneumococci

5.3.3.1 Diverse streptococcin allelic profiles observed in pneumococci

The number of unique allelic profiles observed in the pneumococcal datasets for each streptococcin is shown in Table 5.3. As assessed by the Simpson's diversity indices, each streptococcin was significantly more diverse among Kenyan versus Icelandic pneumococci, and streptococcin A was significantly less diverse overall than the other streptococcin clusters.

5.3.3.2 Identical allelic profiles in different clonal complexes in Icelandic and Kenyan pneumococci

Identical (*i.e.*, same allelic profile) streptococcin clusters were present in unrelated pneumococci from different CCs, and both the Icelandic and Kenyan datasets (Table 5.4). The distribution of the most common shared allelic profiles in CCs of Icelandic and Kenyan pneumococci was investigated, and it was found that they were harboured by different sets of CCs in the two datasets (Figure 5.4, Appendix Table 9.14). The high conservation of streptococcin cluster sequences between CCs suggests either that the rate of diversification of the streptococcin genes is much lower than the seven MLST housekeeping genes (which underly the CC definitions), or that the streptococcin clusters are commonly exchanged between pneumococcal genomes. The observation that streptococcin allelic profile is not fixed within genomes from the same CC supports the latter hypothesis (Appendix Table 9.15).

Table 5.3: Diversity of streptococcin allelic profiles in the Icelandic and Kenyan datasets.

Streptococcin	Iceland		Kenya	
	n	D	n	D
A	75	0.933 (0.927-0.939)	183	0.959 (0.957-0.961)
B	95	0.948 (0.944-0.952)	256	0.973 (0.971-0.975)
C	121	0.954 (0.949-0.959)	267	0.972 (0.97-0.974)
D	1	-	2	-
E	98	0.948 (0.943-0.953)	216	0.971 (0.969-0.973)

The number (n) and Simpson's index of diversity with 95% confidence intervals (D) of streptococcin allelic profiles.

Table 5.4: Distribution of allelic profiles in clonal complexes (CCs) and datasets.

Streptococcin	Individual allelic profiles (n)		
	Total	> 1 CC[a]	Iceland and Kenya[b]
A	247	45	11
B	335	52	16
C	375	54	13
D	2	1	1
E	295	59	19

a. The number of identical allelic profiles observed in more than one CC.
b. The number of identical allelic profiles found in both Icelandic and Kenyan pneumococci.

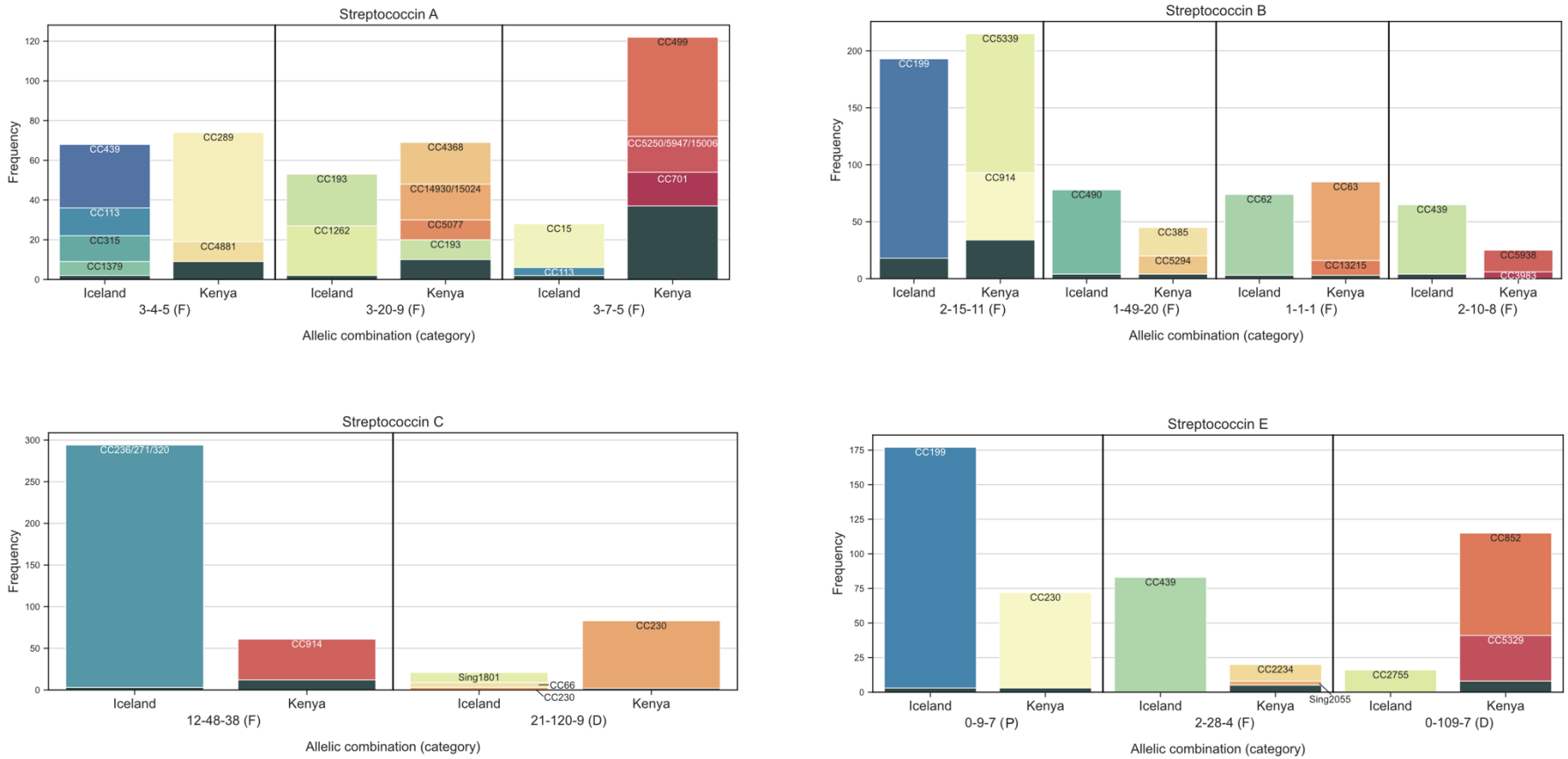

Figure 5.4: Frequency and clonal complex (CC) distribution of allelic profiles of streptococcins A, B, C and E that are commonly found in the Icelandic and Kenyan pneumococcal datasets. Profiles shown if they were observed more than 15 times in both datasets. Bars coloured by the CCs (or singletons [Sing]) that the allelic profile was found in, colour schemes are independent for each streptococcin panel. The 'Other' category represents CCs that represented less than 10% of the examples of the allelic profile and is shown in dark grey. Functional categories of the allelic profiles shown in brackets: F - full, P - partial, D- degrading. Full data can be found in Appendix Table 9.14.

5.3.4 Streptococcin sequence diversity across streptococcal species

Each streptococcin allelic profile was restricted to one species, with very few exceptions: one streptococcin B allelic profile was observed in 14 pneumococcal and two *S. pseudopneumoniae* genomes, and one streptococcin D allelic profile was observed in a single genome each from *S. oralis* and *S. anginosus*. It should be noted that, because the sampling of pneumococci was far denser than for the NPS species, it is possible that identical clusters are found in multiple species but that they were not observed here. To assess the similarity of streptococcins found in different species, sequences of the observed allelic profiles were used to construct neighbour-joining phylogenetic trees (Figure 5.5). Most streptococcin sequences did group with other sequences from the same bacterial species, and pneumococcal streptococcins A, B, C and E formed large groups. Smaller groups were observed among NPS genomes, particularly streptococcins A and B in *S. mitis* and streptococcins A and C in *S. oralis*.

There were exceptions to the species grouping. Subsets of streptococcin A, B and C sequences from different species showed high levels of sequence similarity (Figure 5.5A-C). The groups were dominated by sequences from *S. mitis* and *S. pseudopneumoniae*, with fewer examples from *S. oralis* and pneumococci. Streptococcin E clusters formed clearer species groups, with the most similarity observed between the *S. mitis* group and some sequences from *S. pseudopneumoniae*. Streptococcin E sequences from *S. cristatus, S. anginosus, S. suis* and *S. equi* formed a divergent branch (Figure 5.5E). Overall, streptococcin sequences from *S. pseudopneumoniae* showed less species grouping than those from other viridans streptococci; they tended to be found in groups with sequences from *S. mitis* and pneumococcus. The patterns of streptococcin D distribution were less clear due to the relatively low prevalence of this bacteriocin, and although some species

grouping was apparent, cluster sequences from *S. cristatus* and *S. gordonii* showed similarities (Figure 5.5D). Pneumococcal streptococcin D clusters were most similar to examples from *S. sanguinis* and *S. parasanguinis*.

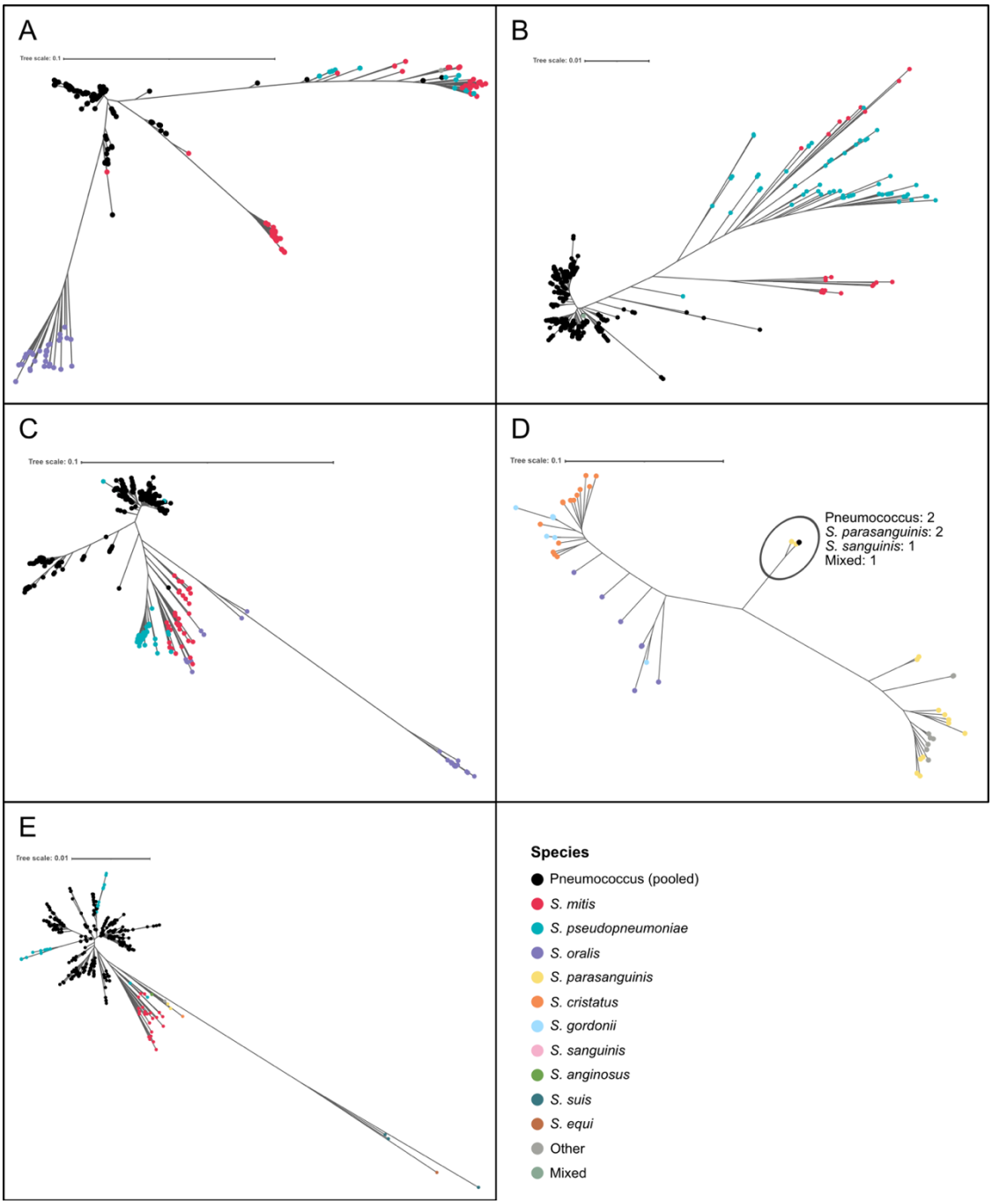

Figure 5.5: Unrooted neighbour-joining trees of all streptococcin cluster sequences observed in pneumococcal and non-pneumococcal Streptococcus genomes. Panels A-E: Streptococcins A-E, respectively. Branch tips are annotated according to the bacterial species that streptococcin cluster was observed in. All trees shown with a scale bar representing nucleotide substitutions per site.

5.3.5 Distribution of toxin alleles within pneumococcal streptococcin clusters

5.3.5.1 Identical toxin genes are observed with divergent immunity genes

Among pneumococcal streptococcins A, B, C and E, identical toxin alleles were observed with different sets of immunity genes. The most common toxin allele of each streptococcin was observed with between 51 and 110 different sets of immunity gene alleles (Table 5.5). The sequences of these clusters were compared using multiple sequence alignments and neighbour-joining phylogenetic trees, and the immunity genes exhibited high sequence diversity despite being associated with the same toxin allele (Figure 5.6). This is indicative of the horizontal exchange of individual toxin genes or sets of immunity genes between streptococcin clusters, although it could also be explained by a very low rate of mutation in the toxin genes relative to the immunity genes.

Table 5.5: Summary of streptococcin toxin alleles.

Streptococcin	Most common toxin allele ID[a]	Frequency	Allelic profiles
A	3	1950	105
B	1	1728	110
C	24	921	55
D	1	94	2
E	1	1014	51

Note: The most common toxin alleles for streptococcins A, B, D and E are complete coding sequences, streptococcin C toxin allele 24 is a pseudogene.

a. Allele ID refers to the identification number the allele is assigned in the BIGSdb database.

Multiple sequence alignments of streptococcin immunity genes showed that nucleotide sequence diversity was not evenly distributed (Figure 5.6). Instead, there were clear patches of divergent sequence within the B (transmembrane domain) genes, particularly in streptococcins A and C. A similar observation was made in previous comparisons of the translated B gene sequences (Figure 4.3). The variation in sequence despite the genes

being associated with the same streptococcin toxin is unexpected and suggests that the mechanism of immunity may not be specific to the toxin allele.

5.3.5.2 Identical immunity genes are observed in full and partial streptococcin clusters in pneumococcus

Pneumococcal streptococcins B and E were often observed as partial clusters lacking the toxin gene, and in some cases, identical immunity genes were found as both full and partial clusters (Appendix Table 9.16). The sequences of these clusters, including the intergenic regions and up to 2 Kb of flanking genomic sequence, were compared with the aim of understanding the relationship between full and partial streptococcin clusters. The alignments of the most prevalent examples of streptococcins B and E are shown in Figure 5.7.

Streptococcin B clusters with *scbB* allele 77 and *scbC* allele 29 were observed as full clusters 63 times and as partial clusters twice. The multiple sequence alignment of these sequences showed that the flanking sequences were highly conserved. In the alignment of partial clusters, there was a 424 bp gap and 79 bp of poorly aligned sequence in the between the upstream flanking gene, *plcR*, and the immunity genes (Figure 5.7A), which spanned the location of the toxin gene in the full clusters. Streptococcin E clusters with *sceB* allele 177 and *sceC* allele 11 were observed as full clusters 128 times and as partial clusters 82 times. Additionally, there were 58 examples of clusters with those immunity genes that were interrupted by the putative insertion sequence IS1515. Most of these clusters were interrupted with contig breaks, and all were excluded from the comparison. In the remaining full and partial clusters, the flanking genes were reliably detectable and relatively conserved (Figure 5.7B). As above, there were gaps in the alignment of the

partial clusters in the expected location of the toxin gene. In this case, there was variation in the length of the gaps (between 264 and 728 bp).

Other examples of streptococcin clusters with identical immunity genes and variable toxin gene presence showed the same pattern: high sequence conservation in flanking regions, and a small region of 'missing' sequence spanning the expected location of the toxin gene. This finding is consistent with the horizontal transfer of a small section of sequence resulting in the change from a full to a partial cluster, or vice versa. The observation of variable gap size in the streptococcin E example suggests that the alignment represents multiple individual exchange events. The small length of exchanged sequence and the high conservation of the flanking sequences is consistent with exchange by homologous recombination, although it is not possible to say from these data whether the toxin or the immunity genes are exchanged.

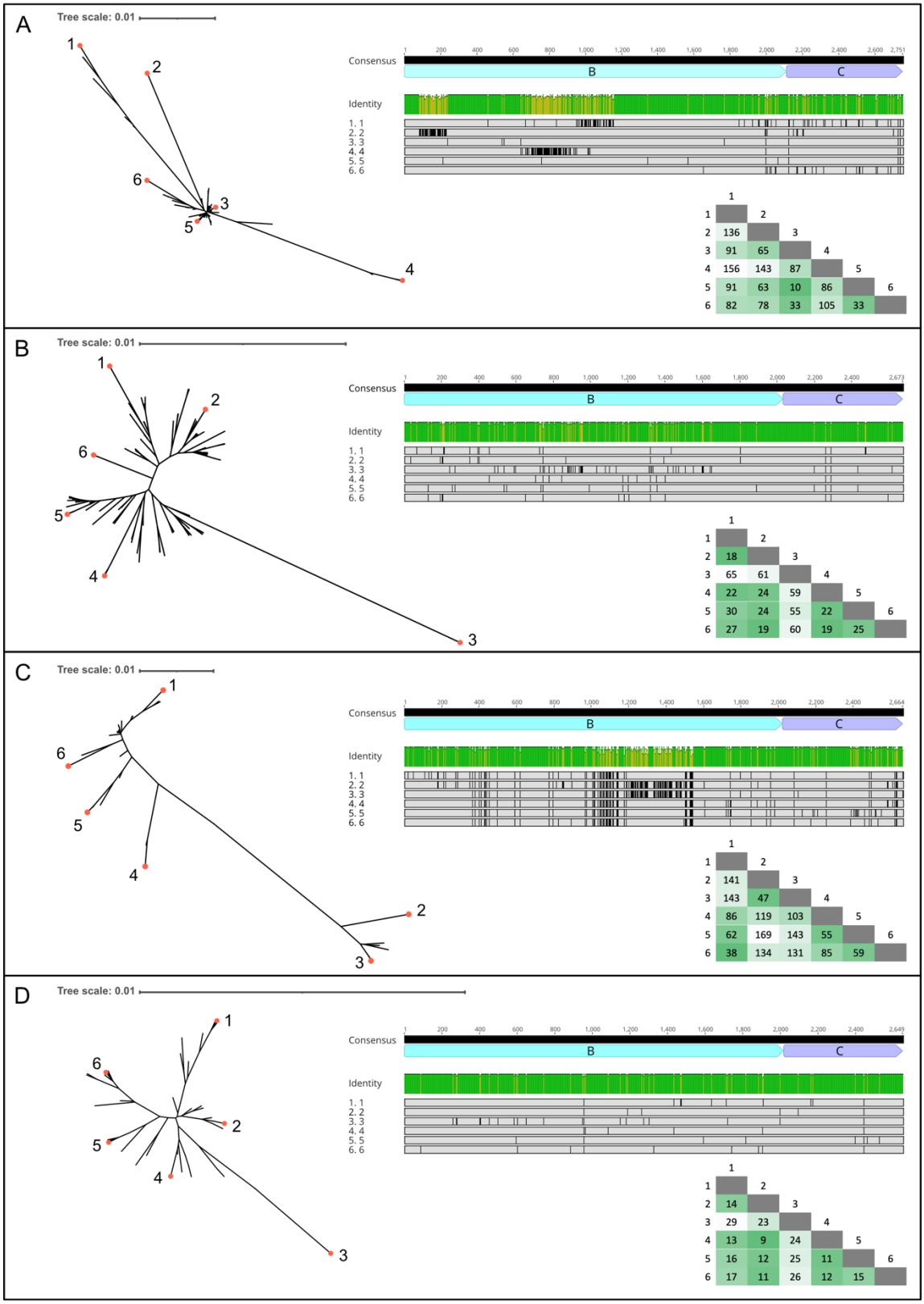

Figure 5.6: Similarity of streptococcin gene cluster sequences associated with the most common toxin allele observed in the pooled pneumococcal datasets. Panel A: Streptococcin A, toxin allele 3; Panel B: Streptococcin B, toxin allele 1; Panel C: Streptococcin C, toxin allele 24 (pseudogene); Panel D: Streptococcin E, toxin allele 1. (Continued)

Figure 5.6 (continued): All trees shown with a scale bar representing 0.01 nucleotide substitutions per site. Alignments show immunity gene sequences from selected clusters (indicated on the tree by red circles and numbers). On the aligned sequences, shaded nucleotide positions indicate differences from the consensus sequence. Pairwise distance matrices indicate the number of differences between the representative cluster sequences, darker shading indicates higher sequence similarity. All sequences included in the alignments are complete coding sequences. Streptococcin D excluded as only two allelic profiles were observed in pneumococcal genomes.

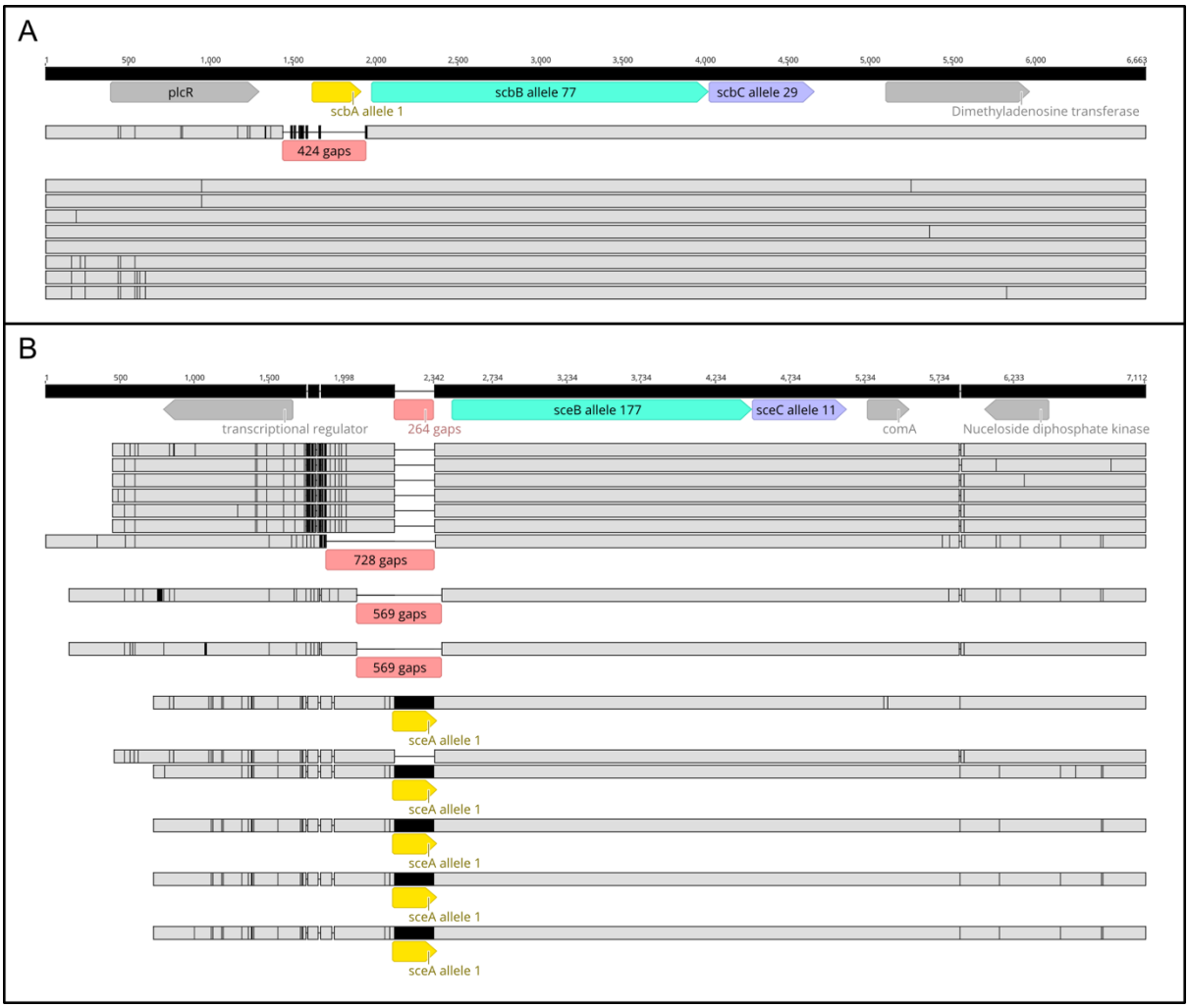

Figure 5.7: Multiple sequence alignments of examples of streptococcin B and E clusters with the same immunity gene alleles found as both full and partial clusters. The most common combination of immunity gene alleles that were found as both full and cheater clusters are shown with 2 Kb flanking sequence. The consensus sequences of the alignments are shown at the top of the alignment and shaded in black. Conserved genes are annotated on the consensus sequence. In the aligned sequences, bases that differ from the consensus sequences are shaded. A: Streptococcin B, scbB allele 77 and scbC allele 29. The alignment represents 63 full clusters (with scbA allele 1) and 2 partial clusters. B: Streptococcin E, sceB allele 177 and sceC allele 11, excluding cases where IS1515 was present upstream of sceB (58 examples). The alignment represents 82 full clusters (with sceA allele 1) and 128 partial clusters.

5.4 Discussion

5.4.1 Streptococcin cluster heterogeneity

The streptococcin gene clusters, with the exception of streptococcin D, are highly prevalent among pneumococci, sometimes as partial gene clusters (Chapter 3).[339] Analyses presented in this chapter revealed heterogeneity even among clusters where all three genes are present. Disruptions to the coding sequences were observed at varying frequencies for all the streptococcin-associated genes. The functional model of streptococcin clusters developed previously (Figure 4.6) was used to contextualise these disruptions, and it was found that a disrupted toxin gene was often found with typical immunity genes. These 'disrupted toxin' clusters are expected to be functionally similar to the partial clusters, which lack the toxin gene entirely. In both cases, pneumococci with these profiles would not be expected to produce a functional toxin but would be protected from the toxins produced by other pneumococci by the immunity complex. Pneumococci with partial and disrupted immunity streptococcin clusters may represent 'cheaters', which are hypothesised to take advantage of other bacteriocin-producing bacteria in the niche while avoiding the bioenergetic cost of bacteriocin production and export.[333,335] It is not clear why some streptococcins are commonly found as partial clusters (B, E), while others are found as disrupted toxin clusters (A, C) in pneumococcus.

5.4.2 Streptococcins in non-pneumococcal streptococci

Among the NPS genomic dataset, the streptococcins were most prevalent in species from the mitis group of streptococci, notably *S. mitis, S. oralis,* and *S. pseudopneumoniae*. Genomes from these species often possessed multiple streptococcins simultaneously, especially *S. pseudopneumoniae*. Streptococcins were only found in a single pyogenic

species (*S. equi*), albeit very rarely (streptococcin E, n = 1), and were not found in species from the salivarius or mutans groups of viridans streptococci. These species distributions may be informative in the question of streptococcin target specificity, as the possession and maintenance of immunity genes would only be advantageous in a strain that would otherwise be susceptible to the corresponding streptococcin toxin. The patterns of streptococcin cluster composition differed in NPS genomes: streptococcins A and C were commonly observed as partial clusters, which was never the case in pneumococcal genomes (including in the previous study of pneumococcal bacteriocins in a diverse, global dataset).[339]

The frequency of streptococcin clusters in NPS genomes indicates that they confer an advantage on the streptococci that possess them, as they would otherwise not be maintained in the genomes. However, the observation of NPS genomes that lack any detectable streptococcins suggests that they are not essential for the survival of these species. It is therefore likely that in NPS species, streptococcins confer an advantage only under certain conditions, for example in the presence of a target competitor, or in a particular ecological niche. We do not yet know enough about competition dynamics of streptococci in the nasopharyngeal or oropharyngeal microbiomes to determine how exactly each streptococcin might contribute to survival. Moreover, as discovery of novel streptococcin clusters (beyond the five previously described) was not within the scope of this study, it is possible that unrecognised streptococcin diversity exists in *Streptococcus* spp.

5.4.2.1 Streptococcin D may be a recent addition to the pneumococcal accessory genome

Streptococcin D was a notable outlier: it was rare among pneumococci and among the streptococcal species that commonly possessed the other streptococcins. It had the highest prevalence in species that typically did not possess other streptococcins *(S. cristatus, S. parasanguinis* and *S. gordonii*). The streptococcin D clusters observed in pneumococcus showed remarkably low diversity: two alleles were observed for *scdB,* and only a single allele each for *scdA* and *scdB.* Either there is a very low tolerance of mutations at these genes, or they have not been present in pneumococci for long and so have not yet diverged as the other streptococcins have. The higher diversity observed in the streptococcin sequences from NPS genomes (Figure 5.5D), and the restriction of streptococcin D to a limited set of CCs in the pneumococcal datasets (CC63, CC13215, Sing14766), both support the recent acquisition of the cluster in pneumococcus.

5.4.3 Horizontal transfer of streptococcin genes

5.4.3.1 Exchange of individual genes and whole clusters between pneumococci

Diverse streptococcin cluster sequences were observed in genomes from the same CC, and identical sequences were observed in genomes from different CCs. The weak association between CC and streptococcin cluster sequence is suggestive of the horizontal exchange of the streptococcins between pneumococci with different genetic backgrounds (approximated by CC). Moreover, identical toxin alleles were observed with diverse immunity gene sequences, which is strong evidence for the horizontal exchange of individual genes. This observation has consequences for the mechanism of immunity: if immunity were highly specific to particular toxin sequences, we would not expect to see this apparent lack of association between toxin and immunity gene alleles. It is not possible to determine whether the exchange of streptococcins is driven by

transformation and homologous recombination, by the movements of larger mobile genetic elements such as ICEs and prophages, or by a combination of both, although generally in pneumococci transformation is believed to be the most important mechanism of horizontal genetic exchange.[26]

The exchange of both whole clusters and of individual genes appears to be contributing to the heterogeneity and sequence diversity of the streptococcin clusters. For example, recombination with another cluster is a mechanism by which a full cluster could switch toxin gene alleles or lose the toxin entirely (if the donor genome possesses a partial cluster). Altogether, there is potential for pneumococci to adapt to altered circumstances by changing to a more advantageous set of streptococcin clusters, as has been observed at the *cps* locus in vaccine escape recombinants in the post-PCV time period (Section 1.1.6). Further work will be required to investigate this possibility.

5.4.3.2 Inter-species exchange

The high sequence similarity between streptococcin cluster sequences found in different species suggests that, rather than diverging from gene clusters that were present in the common ancestor, streptococcins are exchanged horizontally between some streptococcal species. Although the pneumococcal and NPS datasets cannot be directly compared due to differing sampling strategies (discussed below), the lack of identical alleles of each gene being observed in more than one species suggests that inter-species exchange is less common than within pneumococci. This would be expected if exchange were by homologous recombination: it is established that horizontal exchange by transformation is less efficient between less genetically similar strains.[422] The phylogenetic trees of streptococcin A, B, C, and E cluster sequences are suggestive of

multiple inter-species horizontal genetic exchange events. Notably, streptococcins from *S. pseudopneumoniae* showed the least species grouping, suggesting a higher rate of inter-species horizontal exchange in *S. pseudopneumoniae* than in the other species.

5.4.4 Limitations

5.4.4.1 Use of the functional model

Results presented in this chapter were informed by the model of streptococcin function developed in Chapter 4. Therefore, the conclusions drawn here rely on that model being accurate and require experimental confirmation of this model to be validated (as discussed in Section 4.4). However, should the model prove to be inaccurate, results presented here remain useful and informative. The streptococcin gene clusters are clearly maintained in pneumococcal and NPS populations, and the patterns of streptococcin cluster composition and disruption are reliable due to the large size of the genomic datasets used. The observed heterogeneity of clusters will be an avenue for investigation whatever the function of the individual components.

In the majority of cases, patterns of cluster disruption were consistent with the model of streptococcin functionality: toxin genes were not observed without immunity genes, and the disrupted immunity profile, which would be suicidal according to the model, was largely absent. The exception was streptococcin C, which exhibited a notably higher frequency of disrupted immunity clusters in pneumococci (8.6%). The disrupted immunity clusters were restricted to certain CCs (Table 5.2), with the largest contribution in CC217 in the Kenyan dataset. Further work will be required to investigate these profiles - it may be that the toxin is not expressed, that the toxin is expressed but pneumococci from these CCs are tolerant, that there is redundancy among streptococcin

immunity genes and that these pneumococci are protected by a different streptococcin, or, finally, that the functional model is inaccurate.

5.4.4.2 Sampling differences between datasets

The pneumococcal datasets were densely sampled from both carriage and disease-causing pneumococci in two well-defined regions. These datasets were intended to be representative of the pneumococci circulating during the study time periods in Iceland and Kenya, and as such, they can be used to assess the distribution and diversity of the streptococcins circulating at this time. The NPS dataset was designed with a different goal: it used publicly available whole genome sequences of multiple species and aimed to maximise within-species diversity while avoiding a skew towards any one species. There are therefore far fewer examples of each NPS species relative to the pneumococcal datasets, and this limits comparisons of streptococcin diversity that can be drawn between the datasets. Additionally, there was less information available regarding the origin of the NPS genomes relative to the pneumococcal genomes, and CCs were not defined for NPS species, which prevented the study of streptococcin gene prevalence in different subsets of the NPS populations, as in Chapter 3.

5.4.5 Conclusions

In this chapter, the diversity and distribution of the streptococcins were investigated in more detail than in prior studies, revealing complexity despite their high prevalence or ubiquity.

- Streptococcin clusters exhibited diversity in both gene composition and sequence, and this diversity likely has phenotypic consequences.

- The distribution of streptococcin sequences in the pneumococcal datasets strongly suggests that both the clusters and individual genes are exchanged horizontally between species, either *via* transformation or as part of larger mobile genetic elements.
- Streptococcin species distribution was also investigated in an NPS dataset and, not only were streptococcin clusters common in some mitis group viridans streptococci, but they also appeared to be horizontally exchanged between species.

Overall, results presented here improve our understanding of the streptococcins and raise questions regarding their proposed role in competition dynamics in the nasopharynx. In particular, their presence in NPS species suggests that the streptococcins may influence competition in the broader microbiome, not just between pneumococci. In order to investigate the role of the streptococcins further, their function as bacteriocins must be validated experimentally.

6 Streptococcin Isolation and Susceptibility Testing

Experimental work described in this chapter was performed in Professor Shiranee Sriskandan's laboratory in the Department of Infectious Diseases at Imperial College, London. Professor Sriskandan made helpful suggestions regarding the refolding procedure and control assays during susceptibility testing. *Streptococcus pyogenes* M1 hyper-variable region protein was generated by Dr Kristin Huse and Lucy Reeves. Mass spectrometry analysis was performed by Dr Rod Chalk at the Centre for Medicines Discovery at the University of Oxford, who also assisted in the interpretation of results.

6.1 Introduction

6.1.1 Investigating streptococcin function

The data presented in the preceding chapters of this thesis were gathered through *in silico* analyses of putative bacteriocin gene clusters. The majority of these, including the streptococcins, were identified using genome mining. The large amounts of available sequence data have allowed detailed study of the streptococcins, and the existence of a well-characterised homologues has facilitated the development of a general model for the function of streptococcin clusters (Chapter 4). However, models require experimental validation, and many aspects of streptococcin functionality cannot be anticipated from sequence data alone.

Experimental work on the streptococcins to date has been restricted to transcriptomic analyses. RNA sequencing was used to demonstrate the transcription of some of the putative streptococcin genes as a response to stress.[339] While it is important to establish that these genes are expressed, and to what extent they are expressed, such studies are not sufficient to understand the role of the gene products. Experiments using bacteriocin peptides are required to determine the target species, mechanism, and potency of the streptococcins.

6.1.2 Competition assays

6.1.2.1 Competition assay overview

The study of pneumococcal bacteriocin function, in particular of the Blp bacteriocins, has often focussed on competition assays between different strains.[327,338,340,423] These agar plate-based assays involve the growth of a bacteriocin-producing strain overlaid with a susceptible strain on top of, or suspended in, the agar. Bacteriocin activity is inferred when clear zones of inhibition are observed in the overlaid strain.[424] These studies are strengthened when genetic manipulations to knock out the studied bacteriocin result in the loss of the observed competitive advantage.[327]

Competition assays have the advantage of not requiring the synthesis or isolation of the bacteriocin peptide, which can be challenging for many bacteriocins, particularly those with post-translational modifications (such as lasso peptides and lanthipeptides, Section 1.2.2). Isolation of a bacteriocin from its native source can overcome these difficulties, as the native strain can be expected to correctly modify the peptides, but this approach comes with its own challenges of obtaining sufficient yield and purity for experiments.

6.1.2.2 Limitations of competition assays

Recent studies of pneumococcal bacteriocins, including work presented in this thesis, have found that the range of bacteriocins potentially expressed by a single pneumococcal isolate is far higher than previously anticipated: in some cases, 11 distinct bacteriocin biosynthetic gene clusters have been identified in a single genome (Figure 3.5A). Additionally, in a previous RNA sequencing experiment, a single pneumococcus transcribed multiple bacteriocins simultaneously.[339] Competition assays using strains with undefined repertoires of bacteriocins are therefore likely to have been confounded. Moreover, as the expression of pneumococcal bacteriocins is thought to be tightly regulated,[325,329,334] there is no way to know whether the assay conditions induce expression of the bacteriocin of interest without additional experiments. These assays also do not consider the potential complexity of interactions between strains with multiple bacteriocin systems,[339] nor the presence of cheater bacteriocin clusters (Chapter 5).[333] Attributing results of competition assays to the activity of a single bacteriocin is therefore flawed, as the assays do not adequately account for the complexity of pneumococcal bacteriocins.

Overall, competition assays between two strains are useful in establishing the potential for a single strain to out-compete another single strain in lab conditions, but this is a poor replicate of the complex microbial community of the nasopharynx. They are also unreliable when used to infer the functionality of a particular bacteriocin in isolation. The uncertainty in competition assays could be reduced if the number of variables were reduced. A more reliable approach might be to isolate the bacteriocin of interest and test its activity against a panel of potential target species, as traditional antimicrobial compounds are studied. This approach would be suitable for assessing the activity of a

variety of bacteriocins in both isolation and in combination and would potentially provide more certainty about bacteriocin inhibition activity.

6.1.3 Antimicrobial susceptibility testing

The aim of antimicrobial susceptibility testing is to determine whether a compound effectively kills, or inhibits the growth of, a test bacterium. These assays are widely used both to investigate novel antimicrobials and to monitor the acquisition of antimicrobial resistance in clinically important species.[425] Standardised protocols for antimicrobial susceptibility testing have been developed by the Clinical Laboratory Standards Institute (CLSI),[426] and the European Committee on Antimicrobial Susceptibility Testing (EUCAST).[427]

The most widely used methods are agar and broth dilutions.[425,426] A major advantage of dilution protocols is the ability to estimate a minimum inhibitory concentration for the antimicrobial. This is the lowest concentration at which the test compound prevents growth of the test strains. Dilution can be performed in liquid media (broth dilution) or on agar plates (agar dilution). A defined number of bacterial cells are applied to the plate or used to inoculate the media, and the growth of the bacteria in the presence of the antimicrobial is monitored over a set time period. Broth dilution usually allows for testing a larger number of distinct concentrations of the antimicrobial as the compound can be conveniently tested in serial dilution. It has the additional advantage of using relatively small quantities of the antimicrobial when performed in a microtiter plate (broth microdilution).[425]

Antimicrobial peptides (and proteins), including bacteriocins, present challenges during susceptibility testing.[428] Not only must assay conditions allow for the growth of the bacterial isolates to be tested, but they must also be conditions at which the antimicrobial peptide is soluble and active. Performing assays at unfavourable conditions can result in inaccurate minimum inhibitory concentration values and, while adaptations to the standard testing protocols have been published, the optimum conditions for different peptides and proteins must be determined individually. As antimicrobial peptides tend to be cationic, agar-based methodologies are not appropriate due to peptide interactions with anionic components in agar that substantially reduce activity.[428] Broth microdilutions are more suitable, and standard protocols with adaptations for use with peptides have been published.[425,428] Recommendations include the use of low protein-binding plasticware and careful consideration of media composition. Parameters that may be adjusted include salt concentration, overall ionic strength, temperature, oxygenation, and the presence of any co-factors required by the peptide or protein for its native function. In practice, when no functional data are available, optimum conditions must be obtained by trial and error within conditions tolerated by the test organism.

6.1.4 Planned isolation and susceptibility testing of streptococcins

6.1.4.1 Strategy for recombinant expression and protein purification

In this chapter, I aimed to address the hypothetical function of the streptococcins experimentally by isolating them and designing an assay to test their activity against a panel of pneumococcal and NPS strains. Strain selection for susceptibility testing was informed by my earlier results, including strains with both full and partial streptococcin biosynthetic gene clusters, and using NPS species in which the streptococcins were also detected (Figure 5.3A).

Previously, lactococcin 972 has been isolated from native expression in *Lactococcus lactis*. In my work I chose to recombinantly over-express the streptococcins in a heterologous expression system (*E. coli*), which offers several advantages over native expression:

- The availability of well-developed genetic and molecular biology tools for handling *E. coli*, including commercially available competent cell lines and a range of readily available inducible vector systems, and the highly active T7 RNA polymerase and promoter system.[429,430]
- Recombinant expression facilitates tagging of the target protein for purification by affinity chromatography,[430] reducing the chance of co-purifying a protein with similar properties to the target streptococcin (or even a second type of streptococcin if the pneumococcal strain used possesses multiple streptococcin clusters).
- Reduced likelihood of self-toxicity in a Gram-negative species such as *E. coli* than if expressed in pneumococci (due to their proposed mechanism of inhibiting cell wall synthesis).

6.1.4.2 A note on streptococcin classification

The streptococcins are composed of 95-115 amino acid residues. They are expected to be 9-10 kDa and to have secondary structural elements and a stable folded structure without post-translational modifications. Under most definitions, this would class the streptococcins as 'small proteins' rather than 'peptides'. However, lactococcin 972 is consistently referred to as a peptide in the pre-existing literature. The experimental

approach described in this chapter assumes that the expressed products behave as folded proteins.

6.1.5 Research aims

The overall aim of this chapter was to validate the putative function of streptococcins as bacteriocins by investigating their antibacterial activity. This was addressed with the following specific aims:

- To clone and recombinantly express streptococcins in in *E. coli,*
- To isolate recombinantly expressed streptococcins in a functional state,
- To design and use a broth microdilution assay to investigate the antimicrobial function of one or more streptococcins against a panel of streptococci.

6.2 Materials and Methods

6.2.1 General experimental methods

6.2.1.1 General microbiology methods

E. coli for both cloning and recombinant expression were grown in LB broth (Sigma Aldrich, L3522) at 37°C with shaking at 220 rpm. LB-agar (Sigma Aldrich, L3147) was used for growth of colonies on plates, which were grown aerobically at 37°C. Kanamycin was used for selection at 50 μg/mL (Gibco, cat. number 11815024). Optical density of liquid cultures was monitored using scatter of light at 600 nm wavelength (OD_{600}). Where the OD_{600} value was greater than 1.0, samples were diluted in LB broth until a value less than 1.0 was obtained. A high efficiency cloning cell line (NEB 5-alpha, NEB, cat. number C2987H) and an expression cell line (NiCo21(DE3), NEB, cat. number C2529H) were

transformed using the standard heat shock method as recommended by New England Biolabs.[431]

Streptococcal strains were grown on Colombia sheep blood agar plates (BAPs, Oxoid labs, cat. number 12947128). BAPs were incubated at 37˚C with 5% CO_2 for 18-20 hours. When required, growth from plates was resuspended in liquid media using either brain heart infusion broth (BHI, Sigma Aldrich, Cat. number 53286) or Mueller Hinton broth (MHB, Merck, cat. number 70192).

6.2.1.2 General buffer methods

Buffers were made in distilled water (dH_2O) filtered with the Millipore Milli-Q E-pod filtration system. For cloning, ultra-pure water (ddH_2O) filtered with the Millipore Milli-Q Q-pod filtration system was used. Unless otherwise stated, all buffers used in cell lysis and protein purification were based on phosphate buffered saline (PBS), which was prepared by diluting a 10X stock (Life Technologies, cat. number AM9625) in dH_2O. Buffers were supplemented with various additives that were fully dissolved before the buffer was volumetrically made up to the correct volume. A pH meter was used to measure buffer pH, which was adjusted using concentrated hydrochloric acid (HCl, Honeywell, cat. number 258148) or sodium hydroxide (NaOH, Honeywell, Cat. number 30620). Buffers were filtered using 0.2 μm sterile vacuum filtration units (Nalgene, cat. number 514-0027, Millipore, cat. number SCGVU05RE).

6.2.1.3 Centrifugation

Centrifugation of small volumes (up to 1.5 mL) was performed using an Eppendorf 5242 microcentrifuge. Centrifugation of larger volumes was performed using a Heraeus

Megafuge 40R benchtop centrifuge with a BIOLiner swinging bucker rotor (Thermo Fisher Scientific, 75003667). For higher speeds, a Sorvall WX 80+ ultracentrifuge with an AH-650 swinging bucket rotor was used (Thermo Fisher Scientific).

6.2.1.4 SDS-PAGE

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to monitor expression and purification trials using pre-cast Bolt 4-12% bis-tris mini protein gels (Thermo Fisher Scientific, cat. number NW04122BOX, NW04125BOX). 30 µL SDS-PAGE samples were made with 20 µL of protein solution, 7 µL lithium dodecyl sulphate (LDS) loading buffer (Invitrogen, NP0007) and 3 µL 1 M dithiothreitol (DTT, Merck, 10197777001), then heated to 70°C for 10 minutes. Samples were run at 165V for 35 minutes using the Thermo Fisher Scientific mini gel tank electrophoresis system (A25977) in 2-(*N*-morpholino) ethanesulfonic acid (MES) buffer (Invitrogen, cat. number NP0002) with SeeBlue Plus2 pre-stained ladder (Invitrogen, cat. number LC5925).

6.2.2 Cloning expression vectors

6.2.2.1 Streptococcin gene selection and sourcing

The streptococcin genes were chosen from the observed alleles in the Icelandic and Kenyan pneumococcal datasets (Section 2.1). The gene sequences, excluding the regions encoding the phobius-predicted N-terminal signal peptides (Table 4.1), were produced synthetically by Genewiz, using the codon optimisation service for expression in *E. coli* (Table 6.1). Synthetic genes were supplied on a pUC57 plasmid with a kanamycin resistance cassette. The empty expression vector pET-47b, which encodes an N-terminal 6-histidine (6His) tag with a 3c protease cleavage site expressed from a T7 promoter, was

sourced from Novagen (cat. number 71461). Geneious was used to predict the molecular weight (Mw), isoelectric point (pI) and extinction co-efficient for absorption of 280 nm light of the 6His-tagged streptococcin fusion proteins.

6.2.2.2 Polymerase chain reaction

Polymerase chain reaction (PCR) primers were designed to amplify the streptococcin genes with an additional region of overlap with the template vector, facilitating ligase-free assembly.[432] Additional primers were designed to linearise the pET-47b vector backbone at the site of insertion (in-frame with the vector-encoded start codon). Primers were designed using the NEBuilder primer design tool (nebuilder.neb.com) and verified manually by aligning with template DNA sequences in Geneious (Table 6.2). Primers were sourced from Sigma Aldrich as solid pellets, re-suspended according to manufacturer's instructions in dH_2O to generate 100 μM solutions and 10 μM working stocks, and stored at -20°C.

PCRs were performed using a Bio-Rad T100 using the Phusion DNA polymerase high fidelity PCR kit (Thermo Fisher Scientific, cat. Number F553S, Tables 6.3 and 6.4). The positive control reaction supplied with the Phusion polymerase was performed per manufacturer's instructions. Negative control reactions for streptococcin amplification were set up without template DNA. PCR products were purified using the QIAquick PCR purification kit (Qiagen, cat. number 28104). A 2-hour incubation at 37°C with the DpnI restriction enzyme was used to selectively digest any remaining template DNA (Thermo Fisher Scientific, cat. number ER1701).

Table 6.1: Sequences of codon optimised synthetic genes ordered from Genewiz.

Name	Gene	Sequence
scaA2	streptococcin A allele 2	GTTTGGGTGGACGGCGGTCAGTGGAACTATGGTGTTGGTTGGTCCGGCAACTTCGGCTATTCTGACTACCTGCACTCTAC CCGTTCTCACACCGCGACCGTCAAAGACGGCAACAAATTCTCTAAAGACCGCGCGGAAGCTGAAGCGTGGGCGCGCGCGT CTATCTTCAAATTCCCGCCAACTGGTATGGAGTACTTCTACGGCTTCTGA
scaA3	streptococcin A allele 3	CTGGCGGTCTGGGTGGAGGGCGGTCAGTGGAATTATGGTGTGGGCTGGACCGGTACCTTTGGTTACTCCGACTATCTGC ACTCCACTCGCTACCACACCGCAACGGTACGCCACGGCGGTCGTACTAGCAAAGACTACGCAAAGCCAGAAGCGTGGGCA CGTGCTAGCCTGACCAAAATCCCGCCTACGGGTATGGAATACTTCTACGGTTTTGAATAA
scbA1	streptococcin B allele 1	GCTGTGCAGTACCCTGAAGGTGGCGTTTGGACTTATGGCTCCGGCAACGGCGGTGCTTATAGCAACTACTATCACCCGAG CAAATATCACTCCTCTACCGTCGTTAGCCGCAAAACCGGTTCTAGCGACAAAGGCTACGCTGGCGCGGGTGGCACCTCTC GTGCATGGATCCGTACCTCCTGGGGTGAGAAAGTTGCGTTCTACTATAACGTTTGA
sccA2	streptococcin C allele 2	GCTGTGCAGTACCCTGACGGTGGCGTTTGGACTTATGGCGAAGGCTCCGGCGGTGGCTGGGCTTTCAGCAACTACTATCA CGGTAAAAAGTACCATTACTCCTCTCTGGTCTCTCGCTGGAACAGCCACTCCGATAAAGGCGAAGCGAGCGCTGGCAAAA CGAGCTATGCTTGGATCTGGACCAAATGGGGCGAACAGGTTGCGTTTTACTGCGACTACGATTGA
sccA3	streptococcin C allele 3	GCTGTGCAGTACCCTGACGGTGGCGTTTGGACTTATGGCGAAGGCTCCGGCGGTGGCTGGGCTTTCAGCAACTACTATCA CGGTAAAAAGTACCATTACTCCTCTGTGGTATCTAAATGGGACAGCCACTCCGATAAAGGCAGCGCGCCGGCTGGCAAA ACGAGCGAAGCTTGGATCTGGACCAAATTCGGCGAACAGGTTTCTTTCTACTATGACTACGATTGA
scdA1	streptococcin D allele 1	GATTGGGTGAGCGGCGGTAATTGGAGCTATGGTGGCTATCATAACCCGGGCAACTGGGGCGCTTTCAGCAACTACTTCC ATGACTACCGTTGGCACTGGTCCTCTGTGACCCGTGCAAGCGACAGCAAAGCTAACGTGGGCTACGCATCTGCACACTAT ACCTCTCGTAGCTTCATCAACACCTCCTTTGGTGAGACCGCGTACTTCAACTATGGCTTCTAA
sceA1	streptococcin E allele 1	GTTTCCCACCGCGGTGGCGAATGGACTTATGGTGGCCATCACGACCCGTACAACTGGGGCGCTTTCAGCAACTACTATCA CGGTAGCCAGTATCACTGGGCGTACGTGGGCTCTAACGAACGCGACAACCAGAAAACCGTCTACGCAGGTGCACACTCTG CAGCTTATGCATTCGTGAACACCAACCTGGGTGAGCGTATCACCTTCGATGCTGGCTGGTAA

sceA2	streptococcin E allele 2	GTTTCCCACCGCGGTGGCGAATGGACTTATGGTGGCCATCACGACCCGAACAATTGGGGTGCATTTTCTAACTACTATCA CGGTTCTCAATACCATTGGGCATACGTGGGTAGCAACGGCCGCAACAATCAGAAAACTGTCTACGCAGGCGCACGTAGCG CGGCTTACGCGTTTGTCAACACCAACTTCGGTGAGCAGGTTACCTTCGACGCCGGTTGGTGA

Table 6.2: Primers used in PCR for amplification of streptococcin genes and for the linearisation of the pET-47b vector.

Primer	Name	Direction	Target	Sequence[a]
p01	pET47b_lin_for	Forward	pET-47b	GGGTACCAGGATCCGAATTCTG
p02	pET47b_lin_rev	Reverse	pET-47b	GGGTCCCTGAAAGAGGACTTC
p03	aA2_synth_F	Forward	scaA2 synthetic gene	aagtcctctttcagggacccGTTTGGGTGGACGGCGGTC
p04	aA2_synth_R	Reverse	scaA2 synthetic gene	gaattcggatcctggtacccTCAGAAGCCGTAGAAGTACTCCATAC
p05	aA3_synth_F	Forward	scaA3 synthetic gene	aagtcctctttcagggacccCTGGCGGTCTGGGTGGAG
p06	aA3_synth_R	Reverse	scaA3 synthetic gene	gaattcggatcctggtacccTTATTCAAAACCGTAGAAGTATTCCATAC
p07	bA1_synth_F	Forward	scbA1 synthetic gene	aagtcctctttcagggacccGCTGTGCAGTACCCTGAAG
p08	bA1_synth_R	Reverse	scbA1 synthetic gene	gaattcggatcctggtacccTCAAACGTTATAGTAGAACGC
p09	cA2_synth_F	Forward	sccA2 synthetic gene	aagtcctctttcagggacccGCTGTGCAGTACCCTGAC
p12[b]	cA2_synth_R	Reverse	sccA2 synthetic gene	gaattcggatcctggtacccTCAATCGTAGTCGCAGTAAAAC
p11	cA3_synth_F	Forward	sccA3 synthetic gene	aagtcctctttcagggacccGCTGTGCAGTACCCTGAC
p10[b]	cA3_synth_R	Reverse	sccA3 synthetic gene	gaattcggatcctggtacccTCAATCGTAGTCATAGTAGAAAGAAAC
p13	dA1_synth_F	Forward	scdA1 synthetic gene	aagtcctctttcagggacccGATTGGGTGAGCGGCGGTAATTG

p14	dA1_synth_R	Reverse	scdA1 synthetic gene	gaattcggatcctggtacccTTAGAAGCCATAGTTGAAGTACGCG
p15	eA1_synth_F	Forward	sceA1 synthetic gene	aagtcctctttcagggacccGTTTCCCACCGCGGTGGC
p16	eA1_synth_R	Reverse	sceA1 synthetic gene	gaattcggatcctggtacccTTACCAGCCAGCATCGAAGG
p17	eA2_synth_F	Forward	sceA2 synthetic gene	aagtcctctttcagggacccGTTTCCCACCGCGGTGGC
p18	eA2_synth_R	Reverse	sceA2 synthetic gene	gaattcggatcctggtacccTCACCAACCGGCGTCGAAG

a. Sequences shown in upper case anneal to the template DNA, and sequences in lower case show the additional regions of overlap with the vector insertion site.

b. Primers p10 and p12 names were switched, primer p12 targets *sccA2* and primer p10 targets *sccA3*

Table 6.3: 50 µL reaction mixtures for PCR performed using Phusion DNA polymerase to amplify streptococcin genes and to amplify and linearise pET-47b.

Component	Volume (µL)
ddH_2O	32.5
5X HiFi buffer (5x stock)	10
dNTPs (1 mM stock)	1
Forward primer (10 µM stock)	2.5
Reverse primer (10 µM stock)	2.5
Template DNA (1 ng/µL)	1
Phusion DNA polymerase	0.5

Note: dNTPs: deoxynucleoside triphosphates, ddH_2O: sterile, nuclease-free water.

Table 6.4: PCR thermocycler steps used in the amplification of streptococcin genes and the amplification and linearisation of pET-47b.

Step		Temperature (°C)	Duration (seconds)
Initial denaturation		98	180
35 cycles	Denaturation	98	5
	Annealing	60	20
	Extension	72	45
Final extension		72	600
Storage		4	-

6.2.2.3 Expression vector assembly and storage

Amplified inserts were assembled into the linearised vector backbone using the NEBuilder HiFi DNA assembly cloning kit (New England Biolabs, cat. number E5520S). Reaction mixtures were set up with 50-100 ng of linearised vector with a 2-fold molar excess of insert DNA, and a total amount of DNA in reaction mixture of 0.03 – 0.2 pmol (Table 6.5). NEB 5-α competent *E. coli* (DH5α-derived) were transformed with the vector and plated on kanamycin to select successful transformants (New England Biolabs, cat. number C2987H). Three single colonies were picked and grown up overnight in 10 mL LB. Plasmids were extracted from each clone using a QIAprep spin miniprep kit, eluting in ddH_2O (Qiagen, cat. Number 27104). Concentrations of the recovered plasmids were

measured using absorbance of 260 nm light (A_{260}) using a NanoDrop One (Thermo Fisher Scientific, ND-ONE-W). Each plasmid was sequenced by Genewiz using Sanger sequencing with the standard T7 promoter and T7 terminator primers (supplied by Genewiz). Sequences were compared to the expected sequence of each expression vector and a single correct version of each vector was stored at -20°C. Additional expression vectors encoding a maltose binding protein (MBP) tag with a 10x asparagine linker (Asn10) and a Tobacco Etch Virus (TEV) protease cleavage site with a streptococcin fusion (His6-MBP-Asn10-TEV-Streptococcin) were ordered using the Genscript custom gene synthesis service with a modified pET vector (Addgene reference: 29654, sourced from Dr. Erin Cutts). Vectors were verified by sequencing as above.

Table 6.5: HiFi assembly 20 µL reaction mixtures.

Component	Amount
HiFi Master Mix (2x stock)	10 µL
Linearised vector (pET-47b)	~ 100 ng
Insert DNA (streptococcin gene)	~ 117 ng (2x molar excess of vector)
ddH_2O	To 20 µL

Note: Volumes of insert DNA varied since PCR amplification yielded variable concentrations. 100 ng of vector with a 2-fold molar excess of insert DNA corresponds to approximately 117 ng of each amplified streptococcin gene (approximately 300 bp). ddH_2O: sterile, nuclease-free water.

All expression vectors were stored as purified plasmids (minipreps) in dH_2O at -20°C. Additionally, glycerol stocks were generated for the long-term storage of cloning and expression cell lines transformed with each expression vector. Glycerol stocks were generated using 500 µL of culture grown overnight in LB with 500 µL 50% glycerol (Honeywell, G7757, diluted v/v in dH_2O) and stored at -80°C.

6.2.2.4 Agarose gel electrophoresis

Agarose gel electrophoresis was used to assess the products of PCR and assembly. Agarose gels were made with 1-2% agarose in TAE buffer (Tris base, acetic acid and EDTA, Thermo Fisher Scientific, cat. number B49) and stained with SYBR safe DNA gel stain (Thermo Fisher Scientific, cat. number S33102). Gels were run at 80V for 1 hour with the Invitrogen 1 Kb plus ladder (Invitrogen, cat. number 10787018).

6.2.3 Recombinant expression of streptococcins

6.2.3.1 Recombinant protein expression

NiCo21 competent *E. coli* transformed with streptococcin expression vectors were recovered from glycerol stocks and used to inoculate LB media. Cultures were grown to an OD of 0.6 before induction of expression using 0.5-1 mM Isopropyl ß-D-1-thiogalactopyranoside (ITPG, Molecular Dimensions, cat. number PAL-IPTG-1000-5). Induced cultures were grown at 37°C for four hours or at room temperature (18-25°C) overnight (12-18 hours). Expressions were initially performed in small volume trials with non-induced controls to determine the optimum conditions for induction of each streptococcin construct (Appendix Figure 9.1). Scaled up expressions were performed in 500 mL batches. Cells were harvested by centrifugation (small volume trials: 1 mL samples at 3,000 rpm for 3 minutes, large volume cultures: 3,000 rpm for 15 minutes), supernatants were discarded, and pellets were stored at -80°C (small volume trial samples at -20°C).

6.2.3.2 Cell lysis

Thawed cell pellets were chemically lysed using the BugBuster proprietary detergent mixture. BugBuster was either diluted from a 10x master mix (Merck, cat. number 70921-

4) into lysis buffer and supplemented with benzonase (500 U/mL, Merck, cat. number E1014-5KU) and, in some cases, lysozyme (1 KU/mL, Sigma Aldrich, cat. number 71110-3) or used as a pre-made mixture in a tris-based buffer including lysozyme and benzonase (Merck, cat. number 71456-3, Table 6.6, buffers L1-L5). In all cases, lysis buffers were supplemented with ethylenediaminetetraacetic acid (EDTA)-free protease inhibitors (Calbiochem, cat. number 539134). All lysis buffers contained small concentrations of imidazole (Sigma Aldrich, 56750). Cell pellets from small volume expression trials were resuspended in 200 μL of lysis buffer (Table 6.6, buffer L1), pellets from scaled up expression cultures were resuspended in 50 mL lysis buffer per 1L of original growth culture (Table 6.6, buffer L3). Resuspended pellets were incubated at room temperature with agitation for 20 minutes.

Cell lysates from small volume expression trials were ultracentrifuged at 3,000 rpm for 2 minutes to separate the soluble fractions (supernatants) and insoluble fractions (pellets). Proteins in the insoluble cell fraction were re-solubilised with urea, either by resuspending the pellet with 8M urea buffer (Table 6.6, buffer L4) or by including 6M urea in the initial lysis buffer (Table 6.6, buffer L2). Lysates from scaled up expressions were ultracentrifuged at 10,000 rpm for 45 minutes. The pellet (insoluble fraction) was resuspended in buffer containing 8 M urea (Table 6.6, buffer L4) and incubated at room temperature for 1 hour, then centrifuged at 2500 rpm for 15 minutes to pellet any remaining cell debris. Pellets were discarded.

6.2.4 Purification and refolding of streptococcins

6.2.4.1 Immobilised metal ion affinity chromatography

Immobilised metal ion affinity chromatography (IMAC) was used to separate the His6-streptococcins from the native *E. coli* proteins. His GraviTrap pre-packed gravity flow columns (Cytiva, cat. number 11-0033-99) were equilibrated using 10 mL of the same buffer as the streptococcin load sample (Table 6.6, buffer L4). The re-solubilised cell lysate supernatant was loaded onto the column and the flow through was collected. The column was washed with 10 mL wash buffer (Table 6.6, buffer W1), and then bound proteins were eluted with 3 mL of elution buffer containing 500 mM imidazole (Table 6.6, buffer E1). Flow through, washes and elutions were all collected, and samples from each fraction were used to assess the purification by SDS-PAGE.

6.2.4.2 6His-tagged streptococcin refolding by dialysis

Proteins were refolded by dialysis: the denaturing buffer of the eluted IMAC product was exchanged to a native buffer, lacking both urea and imidazole, in a single step (Table 6.6, buffers E1 and D1). Dialysis was performed using SnakeSkin dialysis tubing with a 3.5 kDa molecular weight cut-off (Thermo Fisher Scientific, cat. number 11552541). An excess volume of target buffer was used (at least 200x the total volume of samples in dialysis) and dialysis proceeded overnight at 4°C. Dialysed products were recovered from the tubing and centrifuged at 2,500 rpm for 15 minutes to pellet aggregates. A modified dialysis procedure was also used, where the 8 M urea was reduced in 1 M increments until a concentration of 2 M was reached, then decreased in 0.5 M increments until 0 M was reached (Table 6.6, buffers D1-D10). When the difference in urea concentration was 1 M, the dialysis proceeded for 8-24 hours, and when the difference was 0.5 M, dialysis proceeded for 18-24 hours (summarised in Figure 6.1). At each stage, 200 µL samples

were taken and the soluble fraction was separated by centrifugation at 2,500 rpm for 3 minutes and stored at -20°C.

6.2.4.3 Buffer exchange, concentration, and long-term storage of purified streptococcins

Refolded His6-streptococcins were dialysed further to exchange the buffer to standard 1x PBS at pH 7.4. This proceeded in two steps, first to 1x PBS with an intermediate NaCl concentration (250 mM, Table 6.6, buffer D12) and then to 1x PBS (137 mM NaCl, Table 6.6, buffer D13). Dialysed products were concentrated using Amicon ultra-15 3 KDa MWCO spin concentrators (Millipore, cat. number UFC900324) that were centrifuged at 4000xg for 30 – 60 minutes. Concentration of the protein was monitored with a Nanodrop using absorbance of 280 nm light and the estimated extinction co-efficients. Purified and concentrated proteins were stored at -80°C. The whole purification procedure is summarised in Figure 6.2.

6.2.4.4 Validation of purified proteins by mass spectrometry

Intact electrospray ionisation mass spectrometry was used to validate the purified product. Mass spectrometry relies on the mass-to-charge (m/z) ratio of the products generated by electrospray ionisation. Highly accurate molecular weights of the species in the sample are generated by deconvoluting the m/z spectra. These data can be used to confirm that expected product has been purified and also to detect post-translational modifications.[433] The UniMod database lists previously observed post-translational modifications and their corresponding molecular weights (accessed May 2022, www.unimod.org).

Table 6.6: Buffers used in the purification of 6His-tagged streptococcins.

Buffer use	Buffer name	Notes	Base	pH	[NaCl] (mM)	[Imidazole] (mM)	[Urea] (M)	Additional components/notes
Lysis	L1	Expression trial cell lysis, scaled up cell lysis	PBS	8	500	20	0	BugBuster (1x), benzonase (500 U/mL), protease inhibitors
	L2	Expression trial cell lysis	PBS	8	500	20	6	BugBuster (1x), benzonase (500 U/mL), protease inhibitors
	L3	Scaled up cell lysis	PBS	8	500	5	0	BugBuster (1x), benzonase (500 U/mL), protease inhibitors, lysozyme (1 KU/mL)
	L4*	Scaled up cell lysis, solubilisation of proteins	PBS	8	500	5	8	-
	L5*	BugBuster master mix for scaled up expression cell lysis	Tris	Pre-made, proprietary composition (Merck, cat. number 71456-3)			0	Contains benzonase and lysozyme, supplemented with protease inhibitors
IMAC wash	W1*	IMAC wash buffer, denaturing	PBS	8	500	5	8	Same as buffer L4
	W2	IMAC wash buffer, native, for on-column refolding	PBS	8	500	5	0	-

	W3	IMAC wash buffer, denaturing, lower pH	PBS	7	500	5	8	-
IMAC elution	E1*	IMAC elution buffer, denaturing	PBS	8	500	500	8	-
	E2	IMAC elution buffer, native	PBS	8	500	500		-
	E3	IMAC elution buffer, denaturing, lower pH	PBS	7	500	500	8	-
Refolding (dialysis)	D1*	Dialysis buffer	PBS	8	500	0	0	-
	D2*	Dialysis buffer (incremental)	PBS	8	500	0	7	-
	D3*	Dialysis buffer (incremental)	PBS	8	500	0	6	-
	D4*	Dialysis buffer (incremental)	PBS	8	500	0	5	-
	D5*	Dialysis buffer (incremental)	PBS	8	500	0	4	-
	D6*	Dialysis buffer (incremental)	PBS	8	500	0	3	-
	D7*	Dialysis buffer (incremental)	PBS	8	500	0	2	-

	D8*	Dialysis buffer (incremental)	PBS	8	500	0	1.5	-
	D9*	Dialysis buffer (incremental)	PBS	8	500	0	1	-
	D10*	Dialysis buffer (incremental)	PBS	8	500	0	0.5	-
	D11	Dialysis buffer (low pH)	PBS	6	500	0	0	-
Buffer exchange (dialysis)	D12*	Dialysis buffer (intermediate [NaCl])	PBS	7.4	250	0	0	Intermediate salt buffer
	D13*	Dialysis buffer, low [NaCl]	PBS	7.4	137	0	0	Long term storage buffer

Note: Buffers used in the final protocol for purification of 6His-tagged streptococcin B allele 1 (Appendix Section 9.4.4) indicated with an asterisk.

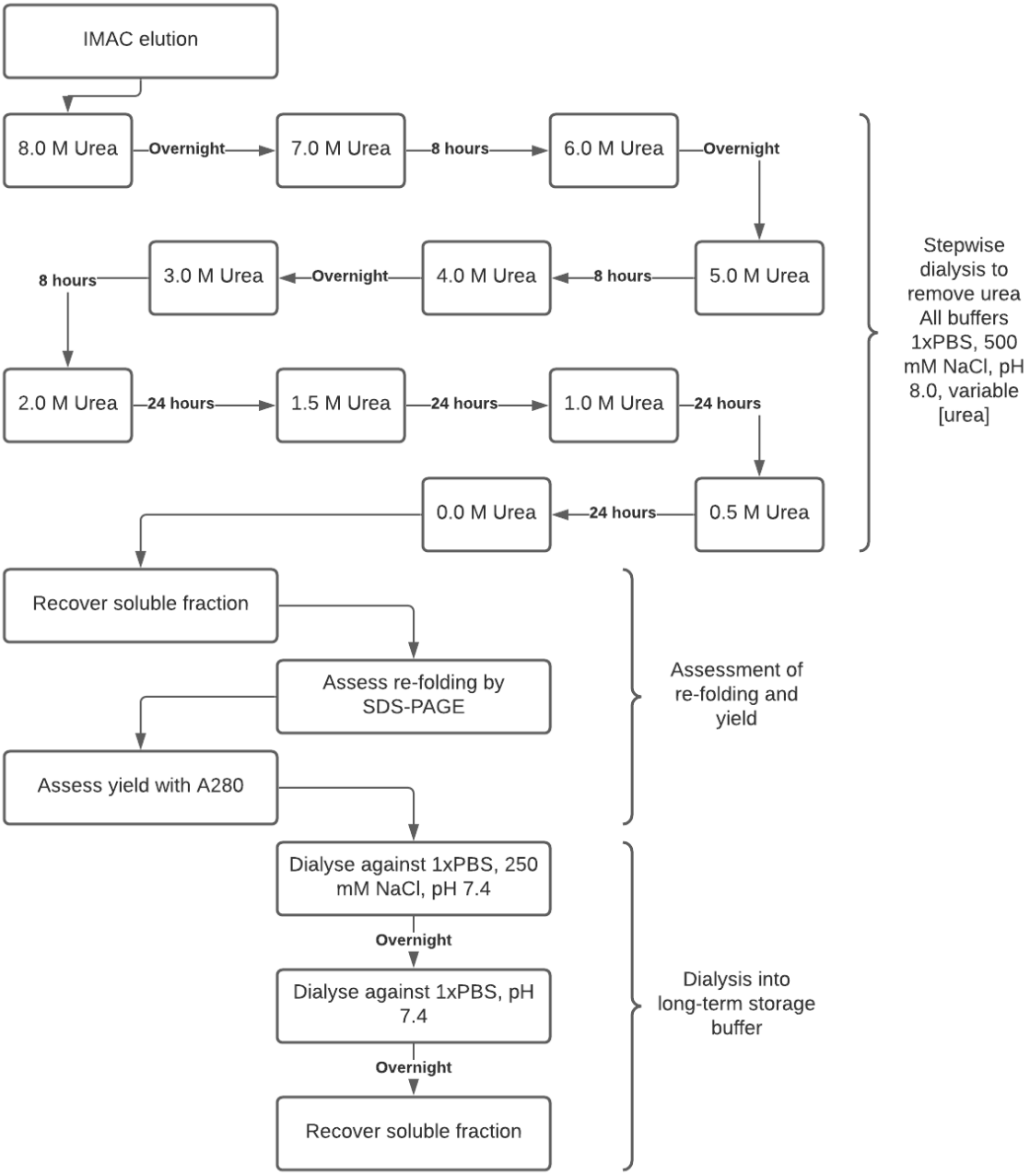

Figure 6.1: Flow chart summarising the procedure for incremental dialysis to re-fold streptococcins and to change to the long-term storage buffer. Timings for the steps to reduce urea are approximate.

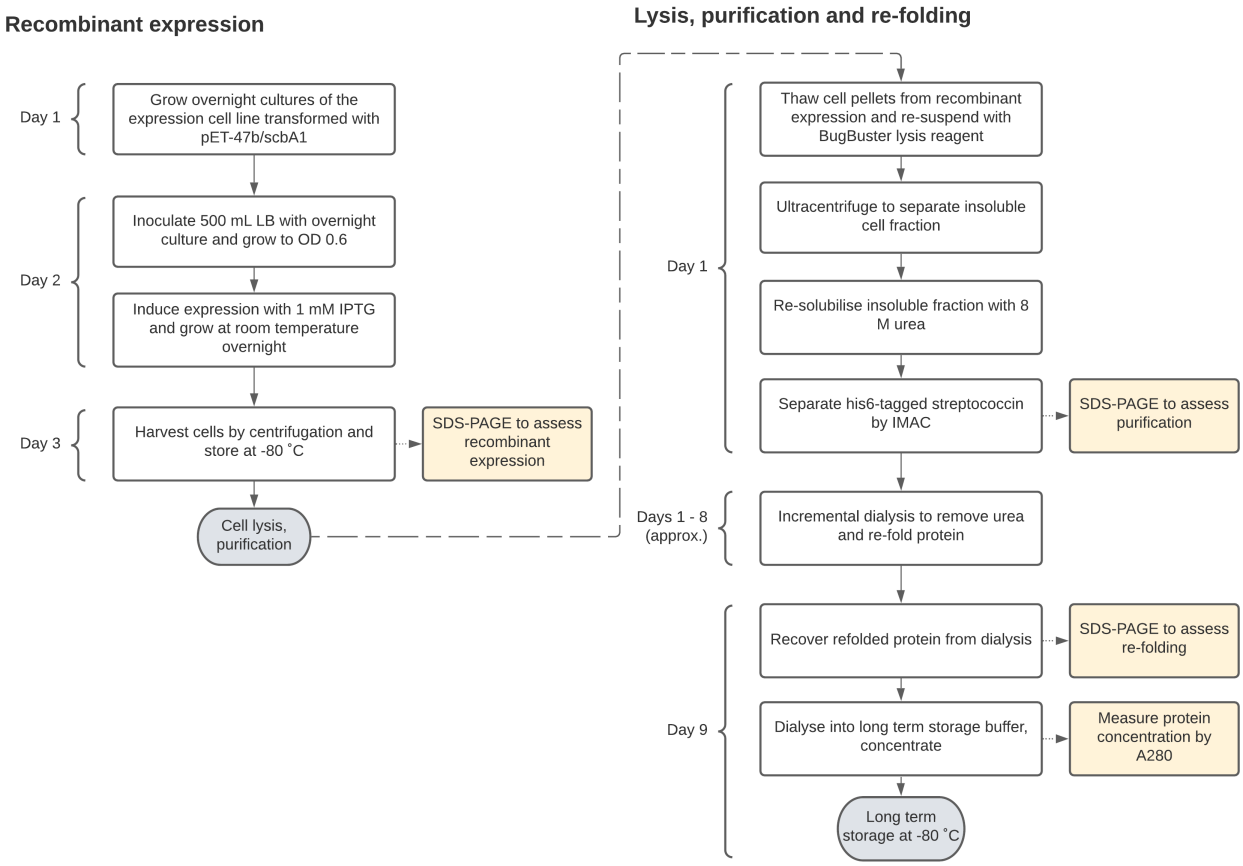

Figure 6.2: Summary of the 6His-tagged streptococcin expression and purification protocol. Stopping points shown in grey, analysis points shown in yellow.

6.2.5 Streptococcin susceptibility assays

6.2.5.1 Streptococcal strain selection, recovery, and stock generation

Strains of *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis* with available whole genome sequences were selected from the National Culture Type Collection (NCTC) and ordered as lyophilised bacteria (Table 6.7). Five pneumococci were selected with long read PacBio genomes available from the Sanger Institute/Public Health England reference collections NCTC3000 project as of January 2022. One *S. pseudopneumoniae* strain and two strains each of *S. mitis* and *S. oralis* from the NCTC had available full

genomes and were included in the assays. If the genome of the strain was already included in my study datasets, prior annotations of the streptococcin cluster genes were retrieved. If not, I identified and characterised the streptococcin clusters using BLAST.

Lyophilised strains were rehydrated in 500 µL BHI and plated onto BAPs. Recovered strains were subcultured from single colonies onto fresh BAPs. The total growth from each BAP was resuspended in 500 µL BHI with 15% v/v glycerol and stored at -80˚C. Strains were recovered from freezer stocks by culturing onto a BAP and incubating overnight.

A broth microdilution assay was designed to test whether purified streptococcins inhibited the growth of streptococci. This assay was adapted from previously published protocols of antimicrobial susceptibility testing and adapted protocols for testing antimicrobial peptides.[425,428] Assays were set up in 96-well round-bottom polypropylene microtiter plates (Sigma Aldrich, cat. number M8060). Assays included 0.01% v/v acetic acid (Sigma Aldrich, cat. number A6283) and 0.2% w/v bovine serum albumin (BSA, Sigma Aldrich, cat. number A0336).

Table 6.7: Streptococci included in streptococcin B susceptibility assays.

Strain name	Species	Aliases	Streptococcin B cluster description[a]
NCTC11904	*S. pneumoniae*	-	Full cluster
NCTC07465	*S. pneumoniae*	PMEN USA1-29	Degrading cluster
NCTC07466	*S. pneumoniae*	D39	Partial cluster
NCTC11886	*S. pneumoniae*	-	Partial cluster
NCTC12495	*S. pneumoniae*	-	Disrupted immunity cluster
NCTC13806	*S. pseudopneumoniae*	NPS dataset ID: 10570	Full cluster
NCTC12261	*S. mitis*	NPS dataset ID: 10931	No detectable streptococcin B cluster
NCTC11189	*S. mitis*	NPS dataset ID: 11256	No detectable streptococcin B cluster
NCTC11427	*S. oralis*	NPS dataset ID: 10892	No detectable streptococcin B cluster
NCTC10232	*S. oralis*	NPS dataset ID: 11261	No detectable streptococcin B cluster

a. Streptococcin B cluster descriptions correspond to the cluster categories as described in (Chapter 5).

6.2.5.2 Susceptibility assays

Aliquots of purified streptococcin were thawed, pooled, and sterilised by filter sterilisation using a 0.2 µm pore syringe filter with a low protein-binding membrane (Millipore, cat. number SLGC004SL). The final concentration of the was measured using a Nanodrop. The protein was diluted to the maximum assay concentration in MHB supplemented with 2.5% v/v lysed horse blood (LHB, Oxoid, cat. number 11464149). This was used to generate a 2-fold serial dilution in 10 columns of the 96-well plate with a volume of 50 µL/well. Column 11 contained 50 µL of MHB + 2.5 % LHB with sterile PBS and no test protein to act as a growth control following inoculation. The ratio of PBS to MHB + 2.5% LHB was held constant in columns 1-11.

Bacterial test strains were cultured from glycerol stocks onto fresh BAPs and grown overnight. Colonies were suspended in MHB + 2.5% LHB to an OD_{625} value of 0.07-0.13. The suspensions were used to inoculate assay plates within 30 minutes. Columns 1 - 11 were inoculated with 50 µL of bacterial suspension, resulting in an assay volume of 100 µL/well. A sterility control was included in column 12, which consisted of 100 µL of MHB + 2.5% LHB without inoculum. In some cases, an additional sterility control was set up in at least triplicate in spare wells using the maximum streptococcin concentration with MHB + 2.5% LHB (also without inoculum).

The assay plate was incubated at 37˚C with 5% CO_2 for 18-20 hours. Bacterial growth after this time was assessed by eye and confirmed by gently resuspending pellets from selected wells and plating 20 µL onto a BAP. All assays were performed as technical duplicates for each bacterial strain. The assay tested either 6His-tagged streptococcin B allele 1 purified as described above (Section 6.2.5) or *S. pyogenes* 10His-tagged M1 protein hyper-variable region (M1-HVR) from a recombinant over-expression in *E. coli* using a similar affinity chromatography strategy.

6.2.5.3 PBS tolerance assay

An assay was designed to test the tolerance of PBS in the growth media. One strain from each species was grown with a variable ratio of sterile PBS to MHB with 2.5% LHB. The ratios tested were (in µL) 50:50, 40:60, 30:70, 20:80, 10:90, and 5:95. Strains were also grown in 100 µL of MHB + 2.5% LHB as a growth control. Two sterility controls were included: one with a 50:50 ratio of PBS:MHB + 2.5% LHB, and one of 100 µL MHB + 2.5% LHB. Neither sterility control was inoculated with bacteria. The plate was incubated at 37˚C with 5% CO_2 for 18-20 hours and bacterial growth was assessed by eye.

6.3 Results

6.3.1 Streptococcin expression vector design and cloning

6.3.1.1 Streptococcin gene selection and synthesis

Streptococcin toxin allele sequences were selected from the annotated genes in the Icelandic and Kenyan genomic datasets (Section 2.2). The most frequently observed allele was chosen for streptococcins B (allele 1) and D (allele 1). Two alleles were selected for streptococcins A (alleles 2 and 3), C (alleles 2 and 3) and E (alleles 1 and 2) as there was no clearly dominant allele for each of these genes. In all cases, the alleles were chosen to maximise diversity in the predicted amino acid sequences of the products. Start codons and the sequence encoding the phobius-predicted signal peptides were excluded from the streptococcin amino acid sequences.

6.3.1.2 Cloning and assembly of tagged streptococcin expression vectors

The streptococcin gene sequences were cloned into the pET-47b vector, which encoded an N-terminal His6 tag with a 3c protease cleavage site. PCR was used to amplify the genes from the synthesised sequences and to linearise the vector at the insert site, and PCR products were verified by agarose gel electrophoresis (Figure 6.3). The expression vectors were assembled using the NEB HiFi assembly reaction and verified by Sanger sequencing. All eight expression vectors were obtained with the correct sequence.

Two additional expression vectors were designed encoding streptococcins A (allele 3) and B (allele 1) with an N-terminal MBP tag, which acts as a solubility tag to prevent the expression of insoluble proteins in inclusion bodies.[434] Expression vectors were designed

with a His6-MBP-Asn10-TEV vector and ordered from Genscript using the custom gene synthesis service.

6.3.1.3 Predicted properties of the tagged streptococcins

The amino acid sequences of the expected products, including the N-terminal 6His tags, were used to predict the Mw, pI and extinction co-efficients of each final product (Table 6.8). These were used to inform the expression and purification of the tagged streptococcins. Mws ranged between 9.82 and 10.54 kDa, and predicted pI values varied between 6.51 and 9.49.

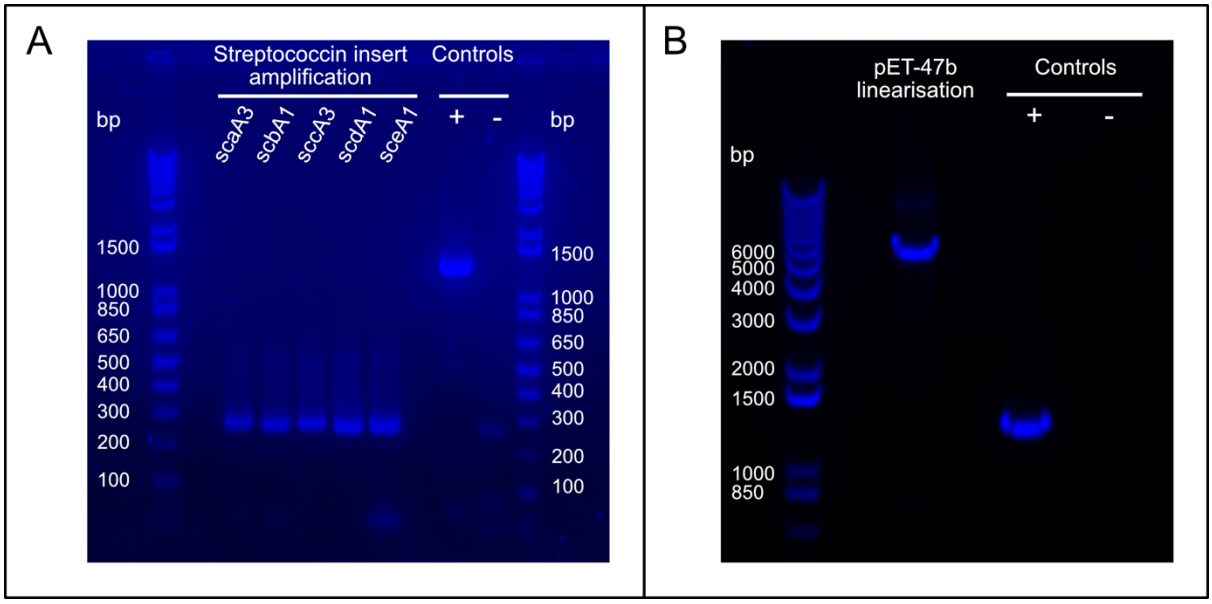

Figure 6.3: Representative agarose gels showing the products of the polymerase chain reaction (PCR) experiments used to generate streptococcin expression vectors. Panel A: Amplification of five streptococcin genes from synthetic gene templates. Panel B: Amplification of the pET-47b vector linearised at the insert site.

Table 6.8: Predicted properties of tagged streptococcins.

Streptococcin	Tag	Construct name	Predicted Mw (kDa)	Predicted pI	Predicted extinction co-efficient
Streptococcin A	6His	6His-scaA2	9.99	6.86	29,450
		6His-scaA3	10.32	7.54	32,430
	MBP	MBP-scaA3	52.73	5.52	100,270
Streptococcin B	6His	6His-scbA1	9.82	9.49	29,910
	MBP	MBP-scbA1	52.23	5.78	97,750
Streptococcin C	6His	6His-sccA2	10.54	6.57	46,410
		6His-sccA3	10.47	6.34	40,910
Streptococcin D	6His	6His-scdA1	10.50	6.96	39,420
Streptococcin E	6His	6His-sceA1	10.44	6.51	33,920
		6His-sceA2	10.33	7.00	32,430

Note: Precited extinction co-efficients given for the absorption of light with a wavelength of 280 nm. Mw: molecular weight, pI: isoelectric point. 6His: 6-histidine tag, MBP: maltose binding protein tag.

6.3.2 Streptococcin expression in *E. coli*

6.3.2.1 Tagged streptococcins are expressed in the insoluble fraction of E. coli cell lysates

Induction from the streptococcin expression vectors was investigated in small volume trials to compare levels of expression and to determine which conditions give optimum yield of product. The overall growth of induced cells was consistently lower than the uninduced controls. Cultures expressing 6His-tagged streptococcin D allele 1 and streptococcin E alleles 1 and 2 showed very little growth following induction, which may be indicative of a toxic effect of these streptococcins on *E. coli* (Table 6.9). Expression of all 6His-tagged streptococcins was induced to varying extents, as determined by eye from the intensity of bands on SDS-PAGE (Figure 6.4A, Appendix Figure 9.2). Both 6His-tagged streptococcin A alleles showed higher yield from the fast induction, 6His-tagged streptococcin B, C and E alleles showed a higher yield from the slower induction, and 6His-tagged streptococcin D was similarly induced in both conditions. Expression from the uninduced controls was also observed in most trials. In all cases, the expressed 6His-tagged streptococcins were located entirely in the insoluble fraction of cell lysates (Appendix Figure 9.2). Following separation of the insoluble fraction, the insoluble product could be resolubilised effectively by incubation in an 8 M urea buffer (Appendix Figure 9.3).

Small volume expression trials of MBP-tagged streptococcin A allele 3 and streptococcin B allele 1 were also performed. While high expression of both MBP-tagged streptococcins was induced using 0.5 mM IPTG at the higher temperature induction (37 °C, four hours), both MBP-streptococcins were consistently observed only in the insoluble fraction of the cell lysate, indicating that the proteins were not correctly folded (Appendix Figure 9.4).

A control trial using the MBP vector without a streptococcin fusion showed good induction of MBP in the soluble cell fraction.

Table 6.9: Growth of NiCo21 *E. coli* expressing 6His-tagged streptococcins in small volume expression trials.

Streptococcin	[IPTG] (mM)	Trial 1, OD_{600}		Trial 2, OD_{600}	
		37 °C, 4 hours	18-25 °C, O/N	37 °C, 4 hours	18-25 °C, O/N
Empty pET-47b	0	2.00	3.07		
	1	1.30	1.86		
His6-scaA2	0	2.24	2.42		
	1	1.72	1.88		
His6-scaA3	0	2.48	3.08	3.12	3.08
	1	2.18	2.08	2.24	2.56
His6-scbA1	0	2.18	2.92	1.84	3.12
	1	1.28	2.64	1.36	3.12
His6-sccA2	0	1.84	2.92		
	1	1.04	2.72		
His6-sccA3	0	2.00	3.04	2.24	2.88
	1	1.20	2.56	1.16	2.88
His6-scdA1	0	2.30	2.76	2.52	3.08
	1	1.06	1.84	0.72	1.68
His6-sceA1	0	1.72	2.00	1.72	2.80
	1	1.06	1.28	0.52	1.12
His6-sceA2	0	2.24	1.86		
	1	1.04	1.20		

Note: Includes uninduced controls assessed using optical density at 600 nm light (OD_{600}). Streptococcins that were only included in a single trial are indicated by shading.

6.3.3 Purification and refolding of 6His-tagged streptococcins

6His-tagged streptococcins A (allele 3) and B (allele 1) were selected for scaled-up expression and purification. Following cell lysis, the streptococcins were successfully separated from *E. coli* proteins using IMAC (Figure 6.4A). As this was performed at denaturing conditions (with 8 M urea), the proteins needed to be refolded if they were to be used in functional assays. All attempts to refold streptococcin A were unsuccessful (Appendix Section 9.4.3).

6.3.3.1 Streptococcin B can be refolded using an extended dialysis procedure

A one-step dialysis to exchange the denaturing buffer to a buffer without urea was unsuccessful for 6His-tagged streptococcin B refolding, and all the product was lost to precipitation (Figure 6.4A). An incremental dialysis procedure was more successful: a proportion of the streptococcin remained in solution, indicating successful refolding to a soluble conformation (Figure 6.4B). Refolded streptococcin B was switched to a lower salt buffer (Table 6.6, buffer D13) without further precipitation.

6.3.3.2 Streptococcin B concentration and storage

6His-tagged streptococcin B was purified independently from two independent expression cultures and concentrated. Visible precipitation was observed following an approximately 2-fold increase in concentration. Precipitant was removed by centrifugation. The final yield of 6His-tagged streptococcin B from purifications can be found in Table 6.10. Each batch was split in two during concentration and in both cases one fraction was of a higher concentration than the other. The fractions were kept separate to maintain the higher concentration. Concentrated streptococcin B was stored at -80°C. Final purity is shown in Figure 6.4C.

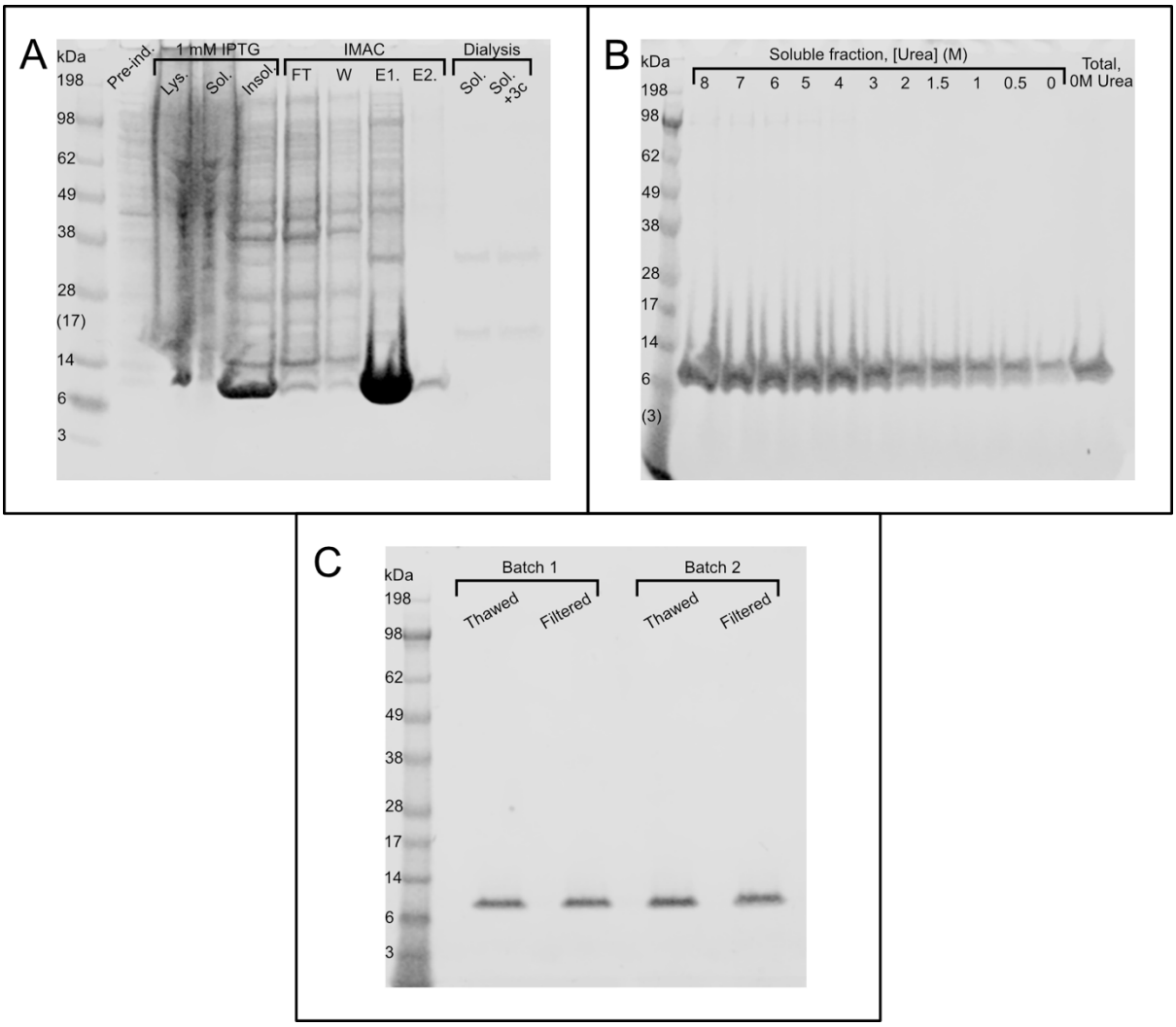

Figure 6.4: Purification and refolding of 6His-tagged streptococcin B. Panel A: expression and purification of 6His-tagged streptococcin B by immobilised metal affinity chromatography (IMAC) and attempted refolding with a single step dialysis. Pre-ind.: pre-induction, Lys: total cell lysate, Sol.: soluble cell fraction, Insol.: insoluble cell fraction, FT: flow through, W: wash, E1: elution 1, E2: elution 2, Dialysis Sol.: soluble fraction of dialysed product, 3c: 3c protease. Panel B: Refolding of 6His-tagged streptococcin B by incremental dialysis. Soluble fraction of the dialysed product shown at each concentration of urea, final lane shows the total product at 0M urea including precipitated product. Panel C: Purity of thawed 6His-tagged streptococcin B following storage at -80˚C and following filter sterilisation.

Table 6.10: The overall yield of 6His-tagged streptococcin B purified independently from two batches of 500 mL *E. coli* expression culture.

Batch	Concentration fraction	Final concentration (µg/mL)	Volume (mL)	Approx. total yield (µg)
1	1	95	8.0	2100
	2	112	12.0	
2	1	104	10.75	2300
	2	138	8.5	

Note: Each batch of purified protein was further split in two during concentration.

6.3.4 Mass spectrometry of streptococcin B allele 1

Intact mass spectrometry was performed on samples of streptococcin B from both independent purifications. The m/z and mass spectra were very similar for both batches, indicating consistent results from the purification procedure (Figure 6.5). The pET-47b expression vector was purified from the streptococcin B expression cell line, sequenced using Sanger sequencing, and found to have the expected sequence. This validates the predicted molecular weight of streptococcin B and confirms that the gene did not acquire any mutations prior to expression.

The m/z spectra of streptococcin B suggested that it retained a fully or partially folded structure in the usually denaturing conditions of electrospray ionisation (Figure 6.5A). This was inferred from the relatively small number of charge states: when proteins are in a denatured and extended conformation there are more exposed amino acid residues that may carry a charge, resulting in a higher number of charge states. This is often observed for proteins with disulphide bridges (covalent cross-links between cysteine

residue side chains), but streptococcin B does not encode any cysteine residues. Streptococcin B may therefore possess an unusually rigid structure.

The deconvoluted mass spectra showed a range of molecular weights that indicated that the proteins in the sample were not uniform (Figure 6.5B). The molecular weights ranged between 9,696.8 and 9,885.8 Da, and the most abundant species was 9,725.1 Da. The expected molecular weight of streptococcin B, 9,820.0 Da, was not observed. Since the observed values were close to the expected value, it is likely that the purified product was streptococcin B with various post-translational modifications. The Unimod database was used to search for modifications of the correct molecular weight to account for the abundant species, but no modifications corresponding to the various mass discrepancies were found. It is not clear whether streptococcin B was modified during the recombinant expression in *E. coli* or whether modifications were acquired during the prolonged purification procedure.

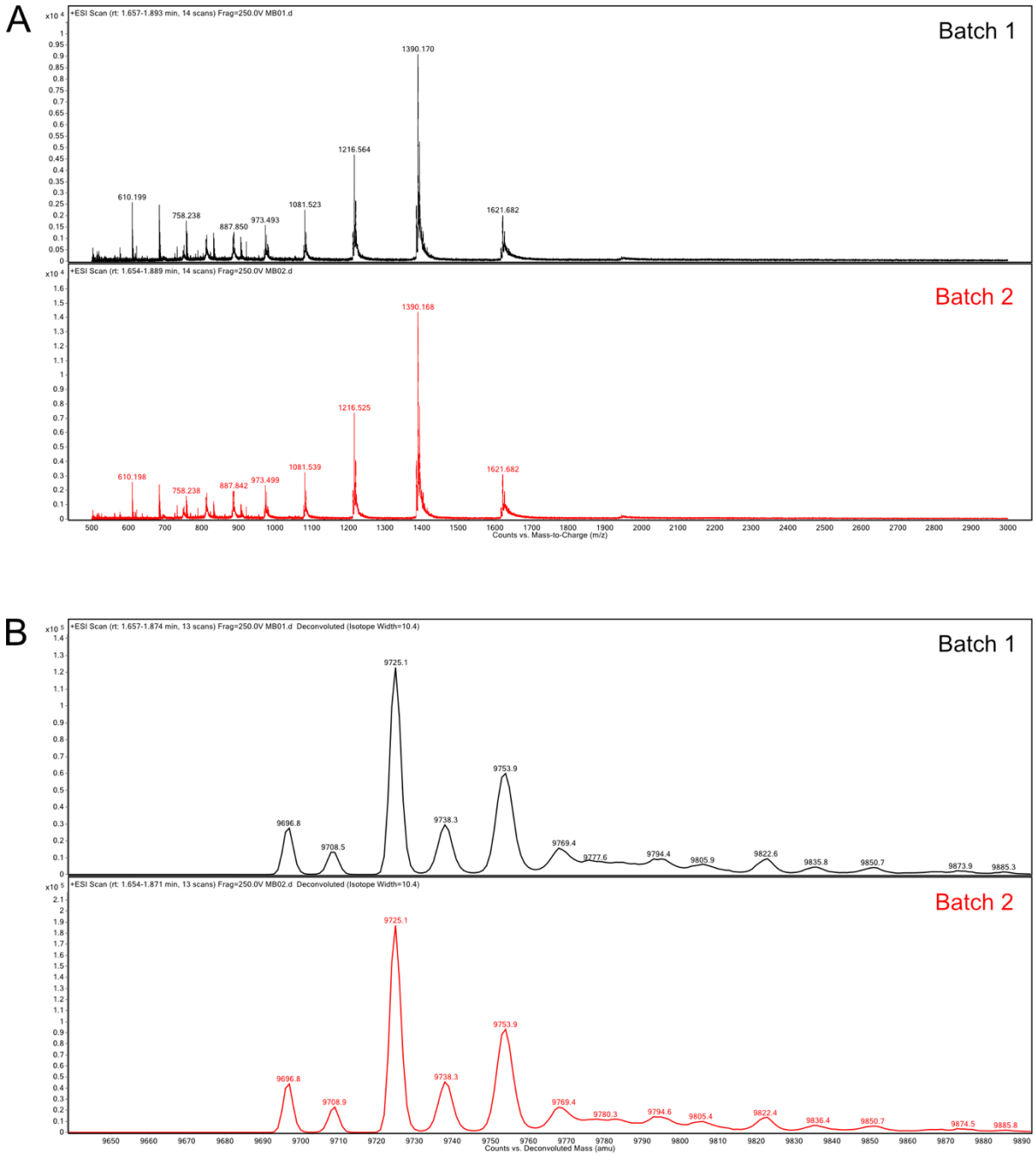

Figure 6.5: Intact electrospray ionisation mass spectrometry of two batches of 6His-tagged streptococcin B. Panel A: the mass-to-charge (m/z) ratio spectra. Panel B: the deconvoluted mass spectra.

6.3.5 Preliminary susceptibility assays

A susceptibility assay was designed to test inhibition of streptococci by purified streptococcin B. The assay results presented below represent preliminary results from an initial set of experiments. It became clear during these assays that putative bacterial growth must be validated, and this validation was not completed on the first set of assays. Results should be treated as preliminary until further replicates with growth validation confirm the findings; these replicates were not possible within the timeframe of the project.

6.3.5.1 Susceptibility assay strain selection

Susceptibility assays were planned using pneumococci and viridans streptococci species that commonly harbour streptococcins: *S. pseudopneumoniae, S. mitis* and *S. oralis* (Figure 5.3). Streptococcin B was ubiquitous in *S. pseudopneumoniae* and common in *S. mitis* (detected in 37% of genomes). Among *S. oralis* genomes streptococcin B was absent, although three other streptococcins were detected (streptococcin A, 39%; C, 28%; and D, 9%).

6.3.5.2 Susceptibility assay optimisation and setup

The PBS tolerance screen demonstrated that the maximum ratio of PBS to MHB + 2.5% LHB in which the strains could successfully grow was 40 µL PBS to 60 µL MHB. This ratio allowed for a maximum concentration of streptococcin B of 40 µg/mL from stocks of 100 µg/mL. Four independent susceptibility assays were performed for each test strain, and each was setup as a technical duplicate. Each replicate used strains independently recovered from freezer stocks. Streptococcin B from each independent purification was used in two replicate assays each (Figure 6.4C). The concentration of streptococcin B

recovered after thawing and filter sterilisation varied, and therefore the maximum concentration of protein used in each susceptibility assay also varied between 42 and 35 µg/mL (Table 6.11, rounded to nearest whole number, approximate values due to variability in Nanodrop concentration measurements).

Table 6.11: Concentrations of streptococcin B used in each replicate susceptibility assay.

Assay replicate	Streptococcin B batch	Concentration of thawed and sterilised streptococcin B (µg/mL)	Maximum assay concentration of streptococcin B (µg/mL)
1	2	106 (90 - 129)	42
2	2	105 (103 - 109)	42
3	1	91 (82 - 101)	36
4	1	87 (75 - 99)	35

Note: Concentration of the thawed protein samples is the mean value of four Nanodrop measurements. All values rounded to the nearest whole number.

6.3.5.3 Bacterial growth in assays must be validated

Following incubation of assay plates, bacterial growth in wells was assessed by eye, as is standard procedure in antimicrobial susceptibility tests.[425] Plates were only analysed if the sterility controls were clear and there was clear growth in the corresponding growth control wells. In assay repeats 2, 3, and 4, bacterial growth was confirmed by plating the two highest streptococcin B concentration wells and the growth control well onto fresh BAPs and growing overnight.

A sterility control performed in assay replicates 2 and 3 using the same conditions as the maximum streptococcin B concentration well without any inoculum found that even in the absence of bacteria, a visible pellet resembling bacterial growth did form on the

bottom of the well. No growth was recovered from these wells on BAPs (Figure 6.6A). It is therefore likely that this pellet was precipitant from the assay mixture, most likely of streptococcin B. This highlights the importance of validation of apparent bacterial growth on BAPs where there was a visible pellet in the assay well.

6.3.5.4 Recombinant S. pyogenes M1 hyper-variable region does not inhibit streptococcal growth

In control assays, all strains exhibited growth at 40 μg/mL M1-HVR, suggesting that the presence of M1-HVR is not inhibitory. Bacterial survival was confirmed by growth on BAPs for 8 out of 10 strains (Figure 6.6B). Two strains, NCTC11886 (pneumococcus) and NCTC13806 (*S. pseudopneumoniae*), did not grow well in either control assay.

6.3.5.5 Four strains were inhibited by streptococcin B in preliminary assays

S. pneumoniae NCTC07466, *S. mitis* NCTC12261 and NCTC11189, and *S. oralis* NCTC10232 were uninhibited by the highest concentrations of streptococcin B in each assay (Table 6.12), and the bacterial growth was recovered on BAPs. The results of the assays were inconclusive for a further two strains, *S. pneumoniae* NCTC11886 and *S. pseudopneumoniae* NCTC13806, which were the same strains that had inconclusive results for the M1-HVR control assay. In all assay replicates, *S. pneumoniae* NCTC11886 growth appeared to be reduced in the presence of 40 μg/mL streptococcin B, but growth from the growth controls wells was not consistently recovered on BAPs. Scant visible growth was observed for *S. pseudopneumoniae* NCTC13806 across the susceptibility assays, and recovery of bacterial growth on BAPs was also inconsistent. NCTC11886 and NCTC13806 do not appear to grow well under assay conditions, and this assay was therefore not suitable for assessing inhibition of these strains.

Three *S. pneumoniae* (NCTC11904, NCTC07465, NCTC12495) and *S. oralis* NCTC11427 appeared to be susceptible to the highest concentrations of streptococcin B (Figure 6.7). Growth was visibly reduced at approximately 40 µg/mL streptococcin B in assays 1, 2 and 3, although often a small pellet was visible in the highest concentration well. However, growth from these wells was not recovered on BAPs, suggesting that the pellet was precipitated streptococcin B, as observed in the sterility controls (Figure 6.6A).

The results of the final replicate assay deviated from the other assays: less inhibition was observed for all four previously inhibited strains (Table 6.12). The approximate concentration of streptococcin B in this replicate was lowest of all the assays (Table 6.11), which would be consistent with the outlying result, and it is possible that the unreliability of nanodrop measurements or human error in the assay setup resulted in an actual streptococcin B concentration lower than measured.

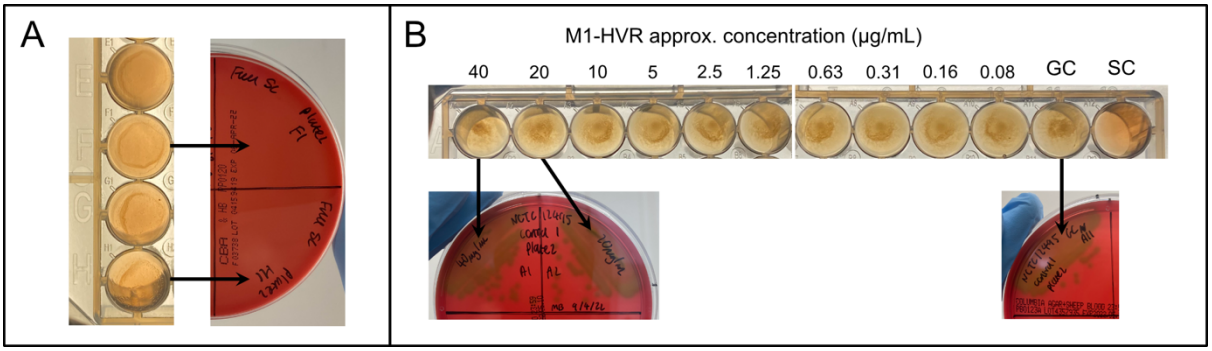

Figure 6.6: Representative results of susceptibility testing control assays. Panel A: Sterility controls using thawed and sterilised streptococcin B diluted in Mueller-Hinton broth with 2.5% lysed horse blood, and attempted recovery of bacterial growth from two wells with visible pellets onto a blood agar plate (BAP). Panel B: Testing the effect of histidine-tagged S. pyogenes M1 protein hyper-variable region (M1-HVR) on streptococci (pneumococcal strain NCTC12495), showing growth recovery on BAPs from indicated wells on the assay plate.

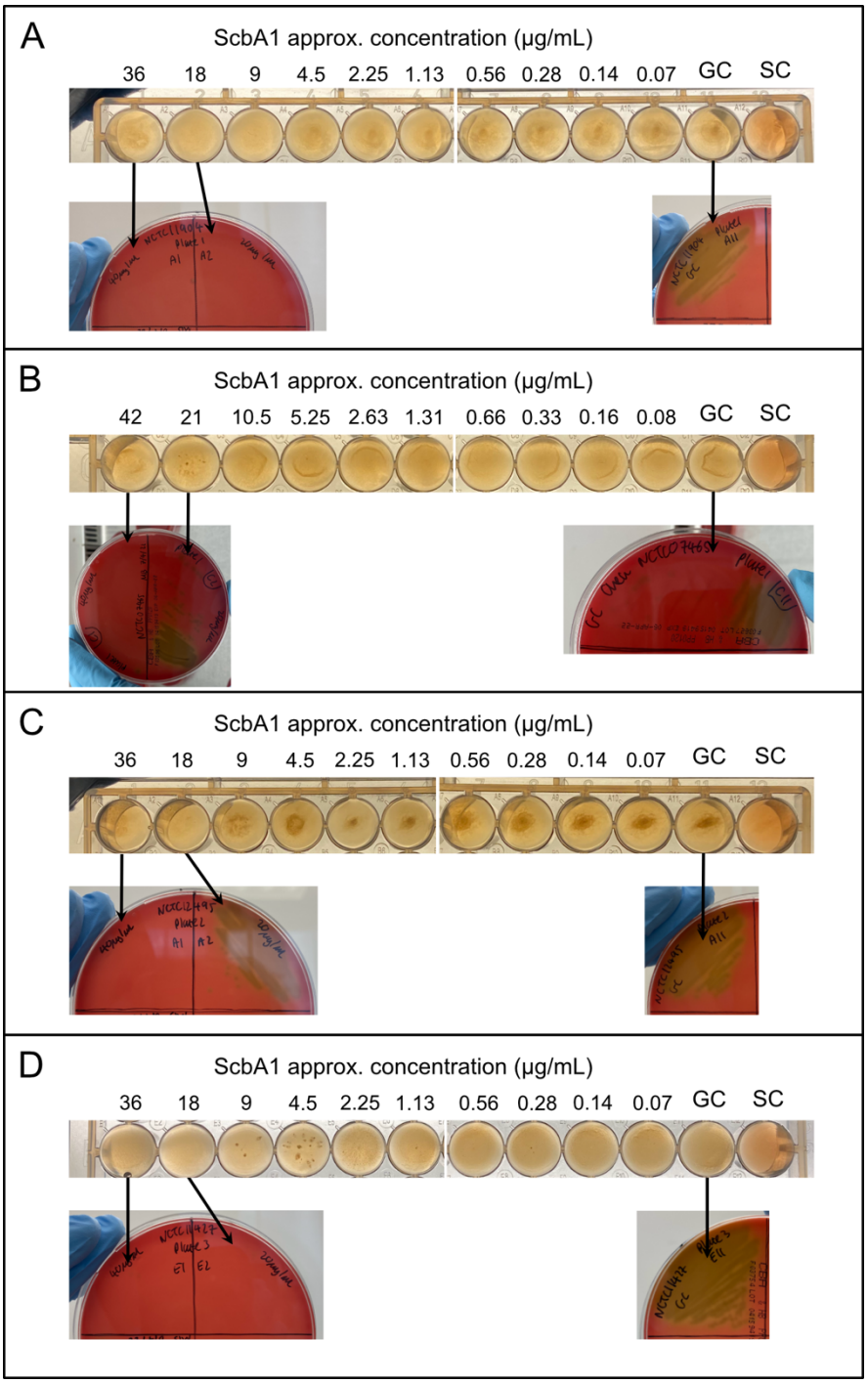

Figure 6.7: Inhibition of four streptococcal strains by streptococcin B. Each panel shows a row from a susceptibility assay plate and growth recovery on blood agar plates from indicated wells of the four inhibited strains. GC: growth control, inoculated well consisting of 40 µL sterilised PBS and 60 µL Mueller Hinton broth (MHB) with 2.5% lysed horse blood (LHB), SC: sterility control, 100 µL of MHB + 2.5% LHB. ScbA1: 6His-tagged streptococcin B allele 1. Panel A: S. pneumoniae NCTC11904, replicate 3. Panel B: S. pneumoniae NCTC07465, replicate 2. Panel C: S. pneumoniae NCTC12495, replicate 3. Panel D: S. oralis NCTC11427, replicate 3.

Table 6.12: Summary of susceptibility assay results.

Strain	Assay replicate[a]	Visible growth reduction	Growth on BAP			SCs	Notes	Preliminary result
			GC	~40 µg/mL[b]	~20 µg/mL[b]			
***S. pneumoniae* NCTC11904**	1	Yes	NA			Clear	No growth validation	Inhibited by ~40 µg/mL streptococcin B
	2	Yes	No	No	No	Clear	GC could not be recovered	
	3	Yes	Yes	No	No	Clear	-	
	4	Yes	Yes	No	Scant	Clear	-	
***S. pneumoniae* NCTC07465**	1	Yes	NA			Clear	No growth validation	Inhibited by ~40 µg/mL streptococcin B
	2	Yes	Yes	No	Yes	Clear	-	
	3	Yes	Yes	No	Yes	Clear	-	
	4	No	Yes	Yes	Yes	Clear	Outlying result	
***S. pneumoniae* NCTC07466**	1	No	NA			Clear	No growth validation	Not inhibited by streptococcin B in assay conditions
	2	No	Yes	Yes	Yes	Clear	-	
	3	No	Yes	Yes	Yes	Clear	-	
	4	No	Yes	Yes	Yes	Clear	-	
***S. pneumoniae* NCTC11886**	1	Yes	NA			Clear	No growth validation	Inconclusive
	2	Yes	No	No	No	Clear	GC could not be recovered	
	3	Yes	Yes	No	No	Clear		

	4	Yes	No	No	Scant	Clear	GC could not be recovered	
***S. pneumoniae* NCTC12495**	1	Yes	NA			Clear	No growth validation	Inhibited by ~40 µg/mL streptococcin B
	2	Yes	Yes	No	Yes	Clear	-	
	3	Yes	Yes	No	Yes	Clear	-	
	4	Yes	Yes	Scant	Yes	Clear	Outlying result	
***S. pseudopneumoniae* NCTC13806**	1	No	NA			Clear	No growth validation	Inconclusive
	2	No (low growth at all conditions)	No	Yes	No	Clear	GC could not be recovered	
	3	No (low growth at all conditions)	Yes	Yes	Yes	Clear	-	
	4	No (low growth at all conditions)	No	Yes	No	Clear	GC could not be recovered	
***S. mitis* NCTC12261**	1	No	NA			Clear	No growth validation	Not inhibited by streptococcin B in assay conditions
	2	No	Yes	Yes	Yes	One of two contaminated	Contamination of SC confirmed by BAP growth recovery	
	3	No	Yes	Yes	Yes	Clear	-	

	4	No	Yes	Yes	Yes	Clear	-	
***S. mitis* NCTC11189**	1	No	NA			Clear	No growth validation	Not inhibited by streptococcin B in assay conditions
	2	No	Yes	Yes	Yes	One of two contaminated	Contamination of SC confirmed by BAP growth recovery	
	3	No	Yes	Yes	Yes	Clear	-	
	4	No	Yes	Yes	Yes	Clear	-	
***S. oralis* NCTC11427**	1	Yes	NA			Clear	No growth validation	Inhibited by ~40 µg/mL streptococcin B
	2	Yes	Yes	No	No	Clear	-	
	3	Yes	Yes	No	No	Clear	-	
	4	Yes	Yes	Scant	Scant	Clear	Outlying result	
***S. oralis* NCTC10232**	1	No	NA			Clear	No growth validation	Not inhibited by streptococcin B in assay conditions
	2	No	Yes	Yes	Yes	Clear	-	
	3	No	Yes	Yes	Yes	Clear	-	
	4	No	Yes	Yes	Yes	Clear	-	

Note: BAP - blood agar plate, SC - sterility control, GC - growth control. Includes an assessment of visible growth on the 96-well assay plate, BAP growth validation, any notes on the individual assays, and the overall result of the replicates. Rows shaded in grey either had growth in the SC or no growth recovery from the GC, and therefore should not be interpreted.

a. Acetic acid was omitted in error from assay replicate 3, all other conditions were identical across assay replicates.
b. Validation of growth on BAPs was not attempted for assay replicate 1. Streptococcin B concentrations listed are approximate and varied in each replicate according to Table 6.11.

6.4 Discussion

6.4.1 First isolation of a streptococcin

Results presented in this chapter represent the first isolation of a streptococcin bacteriocin that has previously only been studied *in silico* as a putative bacteriocin and in RNAseq studies.[339] The recombinant over-expression in *E. coli* followed by purification by affinity chromatography differs from previous approaches used to isolate lactococcin 972.[390,418] Although time consuming, it generated a reproducibly high yield of re-folded product (~2mg/500 mL expression culture). There is scope to optimise the protocol, for example by trialling shorter dialysis times or wider steps between urea concentrations. The purification procedure was not attempted for streptococcins C, D or E. Expression trials demonstrated that these streptococcins are also located in the insoluble cell fraction, so, like streptococcin B, their purification would require a denaturing and re-folding procedure. The successful purification of streptococcin B demonstrates that this approach can generate sufficient yields of re-folded protein, and the protocol would be a good starting point for future attempts to isolate these streptococcins.

Mass spectrometry of purified streptococcin B found that the purified product comprises multiple species with similar but distinct molecular weights, none of which are the expected molecular weight, and which cannot be explained by individual post-translational modifications. The streptococcins are not expected to be modified, and because a heterologous expression system was used, any modifications on the purified product are more likely to represent non-native modifications from expression in *E. coli* or artefacts from the prolonged purification procedure than native modifications. Additionally, streptococcin B appeared to retain either a full or partial fold in mass

spectrometry indicating an unusually rigid structure. This is consistent with previously observed structural rigidity in lactococcin 972.[390] Further experimental work will be required to account for these results.

6.4.2 Preliminary experimental confirmation of streptococcin antibacterial activity

6.4.2.1 Preliminary evidence for antimicrobial activity of streptococcin B

A susceptibility assay assessed inhibition of streptococcal growth by streptococcin B. Results from the assay suggested that the highest concentrations of streptococcin B inhibited growth of three pneumococcal strains and one *S. oralis* strain, which supported the proposed bacteriocin function of streptococcin B and, despite the need for further validation, represents the first functional data for any streptococcin.

The apparent antibacterial activity was not restricted to a particular species, nor was it uniform within each species. Inhibition was observed for four out of five of the screened pneumococcal strains, and one out of two *S. oralis* strains. Neither *S. mitis* strain was inhibited at assay conditions, but the mixed results for pneumococcus and *S. oralis* show that resistance in one strain is not predictive of resistance across the whole species. In order to investigate the specificity of streptococcin B inhibition further, the assay should be applied to a wider panel of genetically diverse strains.

6.4.2.2 Lack of correlation with immunity gene presence

The results of the susceptibility assays did not correlate with predictions of susceptibility based on the streptococcin B clusters identified in the test strain whole genome sequences. *S. pneumoniae* NCTC11904 possesses a full streptococcin B cluster including

putative immunity genes and yet it was inhibited by purified streptococcin B in assay conditions. Conversely, neither of the two *S. oralis* test strains possess a streptococcin B cluster, but while one (NCTC11427) was inhibited by streptococcin B, the other (NCTC10232) was not. It seems that the presence of a putative streptococcin B immunity system in a genome is not predictive of its susceptibility to purified streptococcin B. Superficially, this result undermines the functional model for the streptococcins presented in Chapter 4. RNA sequencing of strains following exposure to streptococcin B would be valuable, both to detect whether the putative immunity genes are expressed in strains that possess them, and to investigate the broader transcriptional response to streptococcin B exposure. In *L. lactis* strains that do not possess lactococcin 972 toxin or immunity genes, but that are resistant to lactococcin 972 activity, there is a transcriptional response to the bacteriocin presence,[393] and a similar effect may be observable in the strains that were not inhibited by streptococcin B despite their lack of putative immunity genes.

6.4.3 Limitations of the experimental approach

6.4.3.1 Recombinantly expressed protein

There is inherent uncertainty in any experiment that uses a recombinantly expressed and purified protein: streptococcin B produced natively by a pneumococcal strain may be structurally distinct to re-folded streptococcin B produced by *E. coli*. The assessment of this would require a comparison to natively expressed streptococcin B. Various biophysical techniques could be used to compare and characterise the proteins, including mass spectrometry to detect deviations from expected molecular weights, circular dichroism to compare secondary structural elements, and nuclear magnetic resonance

spectroscopy to determine the overall structure of the proteins, as used successfully in determining the structure of lactococcin 972.[390]

6.4.3.2 Susceptibility assays

The susceptibility assay as described in this chapter does not quantify bacterial growth at each condition. This could be addressed using a plate reader to quantify the turbidity of each well at the assay end point. This would require additional sets of uninoculated control wells at each streptococcin concentration, so that the impact of precipitation on the turbidity of the solution could be fully assessed. An advantage of this modification would be the ability to reliably detect a dose-dependent reduction in growth, rather than simply the inhibition of the test strains.

The susceptibility assay described in this chapter has the potential to generate a minimum inhibitory concentration value for each test strain; however, it is clear from the preliminary assays that a visible pellet in a well of the 96-well plate is not a reliable indicator of bacterial survival. Some strains appeared to grow in the growth control well, but were not consistently viable on BAPs, suggesting that the strains grew poorly in these assay conditions. Additionally, when the highest concentration of streptococcin B (~40 µg/mL) was incubated without bacterial inoculation, a pellet resembling bacterial growth formed that was likely to be precipitated streptococcin B. Therefore, in order to reliably obtain an MIC value, growth recovery on BAPs should be attempted from all wells in the assay plate, to be certain of the point at which growth is inhibited, and to confirm the apparent dose-dependent inhibition of growth. Moreover, a narrower range of streptococcin B concentrations should be tested.

6.4.4 Summary

In this chapter, I have described the first isolation of a streptococcin, streptococcin B, the development of an assay to test susceptibility of streptococcal strains to streptococcin B, and preliminary results of that assay that demonstrated inhibition of growth of three *S. pneumoniae* strains and an *S. oralis* strain. While further development of the assay is needed to validate these results, the data nevertheless advance our understanding of pneumococcal bacteriocins, and will be used to inform future experimental and genomic investigations of the streptococcins.

7 Summary and Future Work

7.1 Summary of results

7.1.1 Chapter 3

The distribution of 20 bacteriocins was studied in Icelandic and Kenyan pneumococci, and the results showed that the bacteriocin distribution differed by geographical location, in pneumococci recovered from carriage or disease, and in restructured pneumococcal populations following PCV introduction. These differences could be explained by the association of bacteriocins with different pneumococcal genetic lineages, so that differences in population structure resulted in differences in bacteriocin composition. What effect this has on the nasopharyngeal competition dynamics remains to be determined. Despite the overall association of bacteriocins to lineages, some minor differences in bacteriocin repertoire were observed between closely related pneumococci, suggesting that bacteriocin clusters are exchanged horizontally among pneumococci and potentially facilitating adaptation to altered competition dynamics.

7.1.2 Chapter 4

In order to study the streptococcins in more detail, a functional model was developed using the observed gene sequences. This chapter made use of previous experimental work in a homologous bacteriocin system, as well as a range of approaches to functional and structural prediction to determine that streptococcin clusters likely encode a small toxin with an ABC transporter that has a role in immunity. The streptococcin toxin is

expected to interfere with cell wall synthesis *via* an interaction with lipid II, a peptidoglycan precursor molecule.

7.1.3 Chapter 5

The widespread streptococcin clusters were investigated in more detail in genomes recovered from pneumococci and non-pneumococcal streptococci. Despite their high prevalence, streptococcin clusters exhibit a heterogeneous composition, and the functional significance of this was assessed using the model developed in Chapter 4. Clusters were observed that encoded a typical set of immunity genes either without a detectable toxin gene or with a toxin pseudogene, both of which may represent a cheater strategy. The distribution of streptococcin sequences found evidence for the horizontal exchange of whole clusters between pneumococci and other viridans streptococci, and for the exchange of individual genes between different pneumococcal strains.

7.1.4 Chapter 6

A recombinant expression and purification approach was used to isolate a streptococcin toxin (streptococcin B) for the first time. The purified product was used to test the susceptibility of a panel of pneumococci and mitis group streptococci, and preliminary results suggested that streptococcin B is active against some pneumococci and an *S. oralis* strain. Mass spectrometry found that the molecular weight of the purified streptococcin B deviated from the expected value, indicating that it may have been modified in the expression or purification procedure. Further experimental work will be required to both confirm the preliminary assay results and to account for the observed discrepancy in molecular weight.

7.2 Future work

Results presented in this thesis represent significant advances in our understanding of pneumococcal bacteriocins both in terms of their distribution in populations and from a mechanistic perspective. However, there are still many open questions that were beyond the scope of this thesis. Some of these are discussed below along with suggested further work that could be used to address them.

7.2.1 How do the bacteriocins function?

7.2.1.1 Further in vitro testing of streptococcin toxins

In Chapter 4, highly conserved residues within streptococcin toxins, including multiple aromatic residues, were identified. These are likely to be important to streptococcin function and are therefore excellent targets for site-directed mutagenesis, where an individual amino acid is specifically substituted with a different residue, commonly alanine or glycine.[435,436] Mutated streptococcin toxins could be isolated and used in susceptibility assays to assess whether antibacterial activity is altered, implicating the mutated residue in the mechanism of antibacterial activity.

Another approach for understanding the streptococcin mechanism would be to identify the site of action and binding partners on target cells. Purified streptococcins could be fluorescently labelled using a small inorganic fluorophore or a fluorescent protein fusion, and then applied to a susceptible strain. The fluorophore would allow the localisation of the streptococcin either within or on the surface of the target cell by microscopy. There are many ways to identify interactions between proteins and binding partners. A priority for the streptococcins should be the interaction with lipid II, which has been conclusively

demonstrated in lactococcin 972.[389] The proposed second binding partner may be more challenging, as there are many multi-protein complexes on the pneumococcal cell surface that could feasibly be candidates for an interaction, but a pull-down experiment using inhibited cells would be a good starting point to identify potential binding partners for further characterisation.[437]

7.2.1.2 Mechanism of streptococcin immunity

In order to confirm the hypothesised protective function of the immunity genes, immunity gene knock-in and knock-out strains could be generated to determine the effect these genes have on susceptible and non-susceptible strains, respectively. CRISPR-Cas genome modification technology has been applied in pneumococcus and could be used to manipulate wild-type strains that have been used in susceptibility assays such as the ones presented in this chapter.[438,439] Similarly, site-directed mutagenesis could be used to identify amino acids that are important to immunity complex function, and a recombinant expression and purification approach could be developed to isolate the immunity complex for mechanistic and structural studies. A potentially valuable line of enquiry for the mechanism of immunity would be characterising any interaction between the toxin and immunity proteins.

7.2.1.3 Non-streptococcin bacteriocins

Among the pneumococcal bacteriocins studied in Chapter 3, only streptolancidin A and Cib have been studied experimentally. It will therefore be important to confirm the function of the other bacteriocins *in vitro*. As discussed in section 6.1, a procedure to isolate the bacteriocin peptides for susceptibility testing would be optimal, but as these peptides are all anticipated to be subject to post-translational modification, the isolation

may be more challenging than for the unmodified streptococcins. One way to achieve the native modifications would be to isolate the bacteriocin from a producing pneumococcal strain, although this would likely generate a low yield and may encounter issues with self-toxicity and contamination by other expressed peptides.

An alternative would be a recombinant expression in a heterologous system, as used successfully in Chapter 6. In this case, the modification enzymes encoded in the biosynthetic gene clusters would also be required by the expression strain. Additional validation would be required to confirm the expression and function of the modification enzymes, and it is possible that a Gram-positive expression system (such as *Lactococcus lactis*)[440,441] would have more success in replicating native modifications. A final approach would be to generate the peptides synthetically. Some relevant peptide modifications have been achieved synthetically, but some, such as the slipknot structure of lasso peptides, have not been obtained synthetically to date.[302]

7.2.2 How do bacteriocins influence the nasopharyngeal microbiome?

7.2.2.1 Co-colonisation studies

Pneumococcal genomic datasets used in this thesis sampled a single isolate from each patient, and the non-pneumococcal genomes were taken from the rMLST isolate database. All genomes were therefore considered in isolation, away from the complex microbiome where the bacteriocins are expected to function. It is known that pneumococcal strains can co-colonise the nasopharynx, so a different approach would be to sample multiple isolates of both pneumococci and other nasopharyngeal microbes from the same patient and study what combination of bacteriocins were harboured by each organism. The design of these studies would be complex: the choice of multiple

isolates from the same patient risks introducing biases, and classic microbiological techniques will not give any indication of the relative abundance of each genome in the environment. A different approach is to sample a single pneumococcus from the same patient at multiple time points, and to track the bacteriocin repertoires over time. This approach is currently underway using pneumococcal genomes from the South African Drakenstein dataset.[442]

7.2.2.2 Metagenomics

Rather than isolating and sequencing individual colonising strains, a metagenomic approach could be taken to identify the combinations of bacteriocins found in the nasopharynx simultaneously without the need to culture individual isolates and assemble whole genome sequences.[443] While this would not allow the combinations of bacteriocins within individual genomes to be assessed, nor the species that possessed each bacteriocin, it would provide a more comprehensive understanding of the overall bacteriocin landscape within the nasopharynx. The data could also be used to characterise the microbiome composition of the niche.[139] There may be a correlation between the bacteriocins present in the niche and the microbiome composition, although a large sample size would be required for such a study.

7.2.2.3 Mobility

Studying the distribution of bacteriocin clusters has found evidence for the horizontal transfer of bacteriocins both between pneumococci and between non-pneumococcal streptococci. The best way to further investigate the mechanism of horizontal transfer of bacteriocin clusters would be by utilising long read sequencing platforms. It is challenging to study large mobile elements such as integrative conjugate elements in

short read draft genomes both because of the large size of the elements, and because they often contain repeat regions that sequence assembly algorithms struggle to resolve unambiguously. It is therefore difficult to establish whether a bacteriocin cluster is associated with a mobile genetic element. Long read sequencing platforms are much more likely to sequence across a whole mobile genetic element, and so studying mechanisms of bacteriocin mobility would be more achievable using long read genomes.

7.2.2.4 In vivo functionality

A limitation of any *in vitro* assay such as the susceptibility assay developed in Chapter 6 is that it cannot mimic *in vivo* conditions. For example, bacterial growth in isolation in nutrient-rich broth is not a good replicate for the competitive conditions of the nasopharynx in which the bacteriocins are proposed to function. Results from susceptibility assays therefore may not accurately reflect the advantage (or lack thereof) conferred by a bacteriocin *in vivo*. Mouse models of nasopharyngeal carriage have been developed and can be used to assess the effect of an isolated bacteriocin in an environment that more closely resembles the human nasopharynx.[128,327] To investigate bacteriocin function in this environment, an isolated bacteriocin could be applied directly to observe the effect it has on the colonising strains, or carefully selected, potentially genetically manipulated, strains could be used to assess how the presence of bacteriocin gene clusters influence colonisation in the animal model.

7.2.3 How are bacteriocins regulated?

When the pneumococcal bacteriocins were first described, it was noted that many of the biosynthetic gene clusters are flanked by genes with putative functions in transcriptional regulation, in particular genes with homology to quorum sensing systems were

identified.[334,339,341] It is not known what signals stimulate bacteriocin production, nor whether some genes within clusters are regulated differently to others.

7.2.3.1 Genomic studies

It would be valuable to examine whether bacteriocin flanking genes in the Icelandic and Kenyan datasets are consistent with those observed identified in the previous genome mining study.[339] Additionally, it is not known whether the bacteriocin clusters in non-pneumococcal streptococci are in consistent genomic locations, nor whether they possess the same flanking regulatory elements as the homologous clusters in pneumococcal genomes. Although it is challenging to predict transcriptional start sites and regulatory element binding sites in bacterial genomes, some tools have been developed for detecting these sequence elements and predicting transcripts.[444] These approaches could be applied to the genomic datasets used in this thesis to further investigate the transcriptional regulation of the bacteriocin gene clusters.

7.2.3.2 Transcriptomics

A transcriptomic approach has been used previously to establish that transcription of some pneumococcal bacteriocin gene clusters is induced under stress conditions and during *in vitro* competition.[339] Transcriptomics have also been used in pneumococcus to map transcriptional start and stop sites,[445] revealing complexity in patterns of operon expression and extensive regulation by non-coding RNAs. Further transcriptomic studies could therefore be used to identify the signals that up- or down-regulate the expression of bacteriocin gene clusters, and to assess which components of each cluster are co-expressed.

7.2.4 Why do pneumococci possess so many bacteriocins?

7.2.4.1 Up to 11 putative bacteriocin clusters per genome

Pneumococcal genomes possessed up to 11 different bacteriocin clusters, and the most common number per genome was seven. Presumably it is advantageous for pneumococci to possess multiple bacteriocins, but the variation observed in pneumococcal repertoires suggests that the advantage is not attached to a particular combination of bacteriocins. A simple explanation is that each bacteriocin targets different strains or species, and that a wide range of bacteriocins are required to effectively compete in the diverse nasopharyngeal microbiome. Functional studies to determine the target specificity of each bacteriocin could address this, as could competition assays using pneumococci with a variety of bacteriocin repertoires. Another possibility is that some of these 'bacteriocin' gene clusters have functions unrelated to competition (*e.g.* virulence) that are not yet known.

7.2.4.2 Five homologous streptococcin clusters

There are five distinct streptococcins in pneumococcus that appear to have diverged from a common ancestor (Figure 4.1), each with characteristic genomic locations.[339] The toxins have highly conserved amino acid motifs, suggesting a similar mechanism of activity. Why have the streptococcins diversified from a single common ancestor into five distinct clusters in pneumococcus? One possibility is that the five streptococcins have different specificities. While they are all likely to target lipid II to interfere with cell wall synthesis, it is possible that they do this under different circumstances, for example in characteristic cellular locations, at particular stages of the cell cycle, or in particular target species. Alternatively, the streptococcins could share a mechanism but be under the control of distinct regulatory systems, and therefore be expressed in different circumstances.

7.2.4.3 Further bacteriocin discovery

The discovery of novel bacteriocins was beyond the scope of this thesis. However, it is possible that bacteriocins beyond the 20 studied here are present in the global pneumococcal population. For example, streptosactin was discovered in a single genome from the Icelandic pneumococcal dataset and has not been observed in any other pneumococcal genomes since, and streptolancidins A and B were both almost entirely restricted to the Icelandic or Kenyan datasets, respectively. It therefore seems likely that other rare or geographically restricted bacteriocins are awaiting discovery.

Many other streptococcal species are known to possess bacteriocins that have not been observed in pneumococci,[384,446,447] and their influence on the respiratory tract microbiome, particularly in respect to co-colonising pneumococci, has not been considered here. The non-pneumococcal streptococcus genomic dataset was used in this thesis to identify bacteriocins found in pneumococcus but could also be used in a genome mining study for further bacteriocin identification.

7.3 Conclusions

Overall, in this thesis I have used a range of genomic and experimental techniques to characterise pneumococcal bacteriocins. 20 bacteriocins were widely distributed in pneumococci from two distinct geographic locations, and the introduction of PCVs altered the distribution of some bacteriocins. The five streptococcins were studied in more detail, and diversity was observed in the sequences and composition of their gene clusters. There is evidence for the horizontal exchange of the bacteriocins, and the streptococcins were observed in other streptococcal species from the nasopharyngeal microbiome.

Finally, structural predictions were used to develop a general model for streptococcin function. The model was used to inform further genomic studies and experimental work, resulting in the first isolation of a streptococcin and preliminary assay results that suggested antimicrobial activity. Further experimental work will be required to fully understand the role of bacteriocins in pneumococci.

8 References

1. Tuomanen EI, Mitchell TJ, Morrison DA, Spratt BG. *The Pneumococcus*. ASM Press; 2004. doi:10.1128/9781555816537.

2. Griffith F. The Significance of Pneumococcal Types. *J Hyg (Lond)*. 1928;27(2):113-159. doi:10.1017/s0022172400031879

3. Avery OT, Macleod CM, McCarty M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J Exp Med*. 1944;79(2):137-158. doi:10.1084/jem.79.2.137

4. Tacconelli E, Carrara E, Savoldi A, et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis*. 2018;18(3):318-327. doi:10.1016/S1473-3099(17)30753-3

5. Walker MJ, Barnett TC, McArthur JD, et al. Disease Manifestations and Pathogenic Mechanisms of Group A *Streptococcu*s. *Clin Microbiol Rev*. 2014;27(2):264-301. doi:10.1128/CMR.00101-13

6. Raabe VN, Shane AL. Group B *Streptococcus* (*Streptococcus agalactiae*). Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Rood JI, eds. *Microbiol Spectr*. 2019;7(2):7.2.17. doi:10.1128/microbiolspec.GPP3-0007-2018

7. Haas B, Grenier D. Understanding the virulence of *Streptococcus suis*: A veterinary, medical, and economic challenge. *Médecine Mal Infect*. 2018;48(3):159-166. doi:10.1016/j.medmal.2017.10.001

8. Waller AS, Paillot R, Timoney JF. *Streptococcus equi:* a pathogen restricted to one host. *J Med Microbiol*. 2011;60(9):1231-1240. doi:10.1099/jmm.0.028233-0

9. Hakenbeck R, Chhaatwal GS. *Molecular Biology of Streptococci*. Horizon Bioscience; 2007.

10. Kilian M, Poulsen K, Blomqvist T, et al. Evolution of *Streptococcus pneumoniae* and Its Close Commensal Relatives. Ahmed N, ed. *PLoS ONE*. 2008;3(7):e2683. doi:10.1371/journal.pone.0002683

11. Arbique JC, Poyart C, Trieu-Cuot P, et al. Accuracy of Phenotypic and Genotypic Testing for Identification of *Streptococcus pneumoniae* and Description of *Streptococcus pseudopneumoniae* sp. nov. *J Clin Microbiol*. 2004;42(10):4686-4696. doi:10.1128/JCM.42.10.4686-4696.2004

12. Gao XY, Zhi XY, Li HW, Klenk HP, Li WJ. Comparative Genomics of the Bacterial Genus Streptococcus Illuminates Evolutionary Implications of Species Groups. Reid SD, ed. *PLoS ONE*. 2014;9(6):e101229. doi:10.1371/journal.pone.0101229

13. Mitchell J. *Streptococcus mitis*: walking the line between commensalism and pathogenesis. *Mol Oral Microbiol*. 2011;26(2):89-98. doi:10.1111/j.2041-1014.2010.00601.x

14. Vollmer W, Massidda O, Tomasz A. The Cell Wall of *Streptococcus pneumoniae*. Fischetti VA, Novick RP, Ferretti JJ, Portnoy DA, Braunstein M, Rood JI, eds. *Microbiol Spectr*. 2019;7(3):7.3.19. doi:10.1128/microbiolspec.GPP3-0018-2018

15. Vollmer W, Blanot D, de Pedro MA. Peptidoglycan structure and architecture. *FEMS Microbiol Rev*. Published online 2008:19.

16. Bustos JG, TOMASZt A. A biological price of antibiotic resistance: Major changes in the peptidoglycan structure of penicillin-resistant pneumococci. :5.

17. Brown S, Santa Maria JP, Walker S. Wall Teichoic Acids of Gram-Positive Bacteria. *Annu Rev Microbiol*. 2013;67(1):313-336. doi:10.1146/annurev-micro-092412-155620

18. Vollmer W, Tomasz A. Identification of the teichoic acid phosphorylcholine esterase in *Streptococcus pneumoniae*. *Mol Microbiol*. 2001;39(6):1610-1622. doi:10.1046/j.1365-2958.2001.02349.x

19. Bentley SD, Aanensen DM, Mavroidi A, et al. Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. Frasier CM, ed. *PLoS Genet*. 2006;2(3):e31. doi:10.1371/journal.pgen.0020031

20. van Tonder AJ, Bray JE, Quirk SJ, et al. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. *Microb Genomics*. 2016;2(10). doi:10.1099/mgen.0.000090

21. van Tonder AJ, Gladstone RA, Lo SW, et al. Putative novel cps loci in a large global collection of pneumococci. *Microb Genomics*. 2019;5(7). doi:10.1099/mgen.0.000274

22. Ganaie F, Saad JS, McGee L, et al. A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large *cps* Fragment from an Oral Streptococcus. McDaniel LS, ed. *mBio*. 2020;11(3):e00937-20. doi:10.1128/mBio.00937-20

23. Sørensen UB, Henrichsen J, Chen HC, Szu SC. Covalent linkage between the capsular polysaccharide and the cell wall peptidoglycan of Streptococcus pneumoniae revealed by immunochemical methods. *Microb Pathog*. 1990;8(5):325-334. doi:10.1016/0882-4010(90)90091-4

24. Kolkman MAB, van der Zeijst BAM, Nuijten PJM. Diversity of Capsular Polysaccharide Synthesis Gene Clusters in *Streptococcus pneumoniae*. *J Biochem (Tokyo)*. 1998;123(5):937-945. doi:10.1093/oxfordjournals.jbchem.a022028

25. Tettelin H, Nelson KE, Paulsen IT, et al. Complete Genome Sequence of a Virulent Isolate of *Streptococcus pneumoniae*. *Science*. 2001;293(5529):498-506. doi:10.1126/science.1061217

26. Hiller NL, Sá-Leão R. Puzzling Over the Pneumococcal Pangenome. *Front Microbiol*. 2018;9:2580. doi:10.3389/fmicb.2018.02580

27. van Tonder AJ, Bray JE, Jolley KA, et al. Genomic Analyses of >3,100 Nasopharyngeal Pneumococci Revealed Significant Differences Between Pneumococci Recovered in Four Different Geographical Regions. *Front Microbiol*. 2019;10:317. doi:10.3389/fmicb.2019.00317

28. Croucher NJ, Finkelstein JA, Pelton SI, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet*. 2013;45(6):656-663. doi:10.1038/ng.2625

29. Donati C, Hiller NL, Tettelin H, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol*. 2010;11(10):R107. doi:10.1186/gb-2010-11-10-r107

30. Claverys JP, Prudhomme M, Mortier-Barriere I, Martin B. Adaptation to the environment: *Streptococcus pneumoniae*, a paradigm for recombination-mediated genetic plasticity? *Mol Microbiol*. 2000;35(2):251-259. doi:10.1046/j.1365-2958.2000.01718.x

31. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. *Science*. 2009;323(5915):741-746. doi:10.1126/science.1159388

32. Thomas CM, Nielsen KM. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat Rev Microbiol*. 2005;3(9):711-721. doi:10.1038/nrmicro1234

33. Casjens S. Prophages and bacterial genomics: what have we learned so far?: Prophage genomics. *Mol Microbiol*. 2003;49(2):277-300. doi:10.1046/j.1365-2958.2003.03580.x

34. Iannelli F, Santoro F, Fox V, Pozzi G. A Mating Procedure for Genetic Transfer of Integrative and Conjugative Elements (ICEs) of Streptococci and Enterococci. Published online 2021:9.

35. Rankin DJ, Rocha EPC, Brown SP. What traits are carried on mobile genetic elements, and why? *Heredity*. 2011;106(1):1-10. doi:10.1038/hdy.2010.24

36. Håvarstein LS, Coomaraswamy G, Morrison DA. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proc Natl Acad Sci*. 1995;92(24):11140-11144. doi:10.1073/pnas.92.24.11140

37. Claverys JP, Havarstein LS. EXTRACELLULAR-PEPTIDE CONTROL OF COMPETENCE FOR GENETIC TRANSFORMATION IN. :17.

38. Attaiech L, Olivier A, Mortier-Barrière I, et al. Role of the Single-Stranded DNA–Binding Protein SsbB in Pneumococcal Transformation: Maintenance of a Reservoir for Genetic Plasticity. Matic I, ed. *PLoS Genet*. 2011;7(6):e1002156. doi:10.1371/journal.pgen.1002156

39. Golubchik T, Brueggemann AB, Street T, et al. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat Genet*. 2012;44(3):352-355. doi:10.1038/ng.1072

40. Hiller NL, Ahmed A, Powell E, et al. Generation of Genic Diversity among *Streptococcus pneumoniae* Strains via Horizontal Gene Transfer during a Chronic Polyclonal Pediatric Infection. Bessen DE, ed. *PLoS Pathog*. 2010;6(9):e1001108. doi:10.1371/journal.ppat.1001108

41. Hakenbeck R, Balmelle N, Weber B, Gardès C, Keck W, de Saizieu A. Mosaic Genes and Mosaic Chromosomes: Intra- and Interspecies Genomic Variation of *Streptococcus pneumoniae*. Tuomanen EI, ed. *Infect Immun*. 2001;69(4):2477-2486. doi:10.1128/IAI.69.4.2477-2486.2001

42. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-Recombination, Diversity, and Antibiotic Resistance in Pneumococcus. *Science*. 2009;324(5933):1454-1457. doi:10.1126/science.1171908

43. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. A High-Resolution View of Genome-Wide Pneumococcal Transformation. Didelot X, ed. *PLoS Pathog*. 2012;8(6):e1002745. doi:10.1371/journal.ppat.1002745

44. Wyres KL, Lambertsen LM, Croucher NJ, et al. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. *Genome Biol*. 2012;13(11):R103. doi:10.1186/gb-2012-13-11-r103

45. Cowley LA, Petersen FC, Junges R, Jimson D. Jimenez M, Morrison DA, Hanage WP. Evolution via recombination: Cell-to-cell contact facilitates larger recombination events in *Streptococcus pneumoniae*. Matic I, ed. *PLOS Genet*. 2018;14(6):e1007410. doi:10.1371/journal.pgen.1007410

46. Brueggemann AB, Pai R, Crook DW, Beall B. Vaccine Escape Recombinants Emerge after Pneumococcal Vaccination in the United States. Wessels MR, ed. *PLoS Pathog*. 2007;3(11):e168. doi:10.1371/journal.ppat.0030168

47. Wozniak RAF, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Microbiol*. 2010;8(8):552-563. doi:10.1038/nrmicro2382

48. Croucher NJ, Walker D, Romero P, et al. Role of Conjugative Elements in the Evolution of the Multidrug-Resistant Pandemic Clone *Streptococcus pneumoniae* ^{Spain23F} ST81. *J Bacteriol*. 2009;191(5):1480-1489. doi:10.1128/JB.01343-08

49. Chancey ST, Agrawal S, Schroeder MR, Farley MM, Tettelin H, Stephens DS. Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. *Front Microbiol*. 2015;6. doi:10.3389/fmicb.2015.00026

50. Ambroset C, Coluzzi C, Guédon G, et al. New Insights into the Classification and Integration Specificity of Streptococcus Integrative Conjugative Elements through Extensive Genome Exploration. *Front Microbiol*. 2016;6. doi:10.3389/fmicb.2015.01483

51. D'Aeth JC, van der Linden MP, McGee L, et al. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. *eLife*. 2021;10:e67113. doi:10.7554/eLife.67113

52. Ramirez M, Severina E, Tomasz A. A High Incidence of Prophage Carriage among Natural Isolates of *Streptococcus pneumoniae*. *J Bacteriol*. 1999;181(12):3618-3625. doi:10.1128/JB.181.12.3618-3625.1999

53. Croucher NJ, Coupland PG, Stevenson AE, Callendrello A, Bentley SD, Hanage WP. Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun*. 2014;5(1):5471. doi:10.1038/ncomms6471

54. Brueggemann AB, Harrold CL, Rezaei Javan R, van Tonder AJ, McDonnell AJ, Edwards BA. Pneumococcal prophages are diverse, but not without structure or history. *Sci Rep*. 2017;7(1):42976. doi:10.1038/srep42976

55. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Mol Microbiol*. 2001;39(2):260-272. doi:10.1046/j.1365-2958.2001.02234.x

56. Boyd EF, Brüssow H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol*. 2002;10(11):521-529. doi:10.1016/S0966-842X(02)02459-9

57. Rezaei Javan R, Ramos-Sevillano E, Akter A, Brown J, Brueggemann AB. Prophages and satellite prophages are widespread in Streptococcus and may play a role in pneumococcal pathogenesis. *Nat Commun*. 2019;10(1):4852. doi:10.1038/s41467-019-12825-y

58. Bogaert D, de Groot R, Hermans P. *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *Lancet Infect Dis*. 2004;4(3):144-154. doi:10.1016/S1473-3099(04)00938-7

59. Sleeman KL, Griffiths D, Shackley F, et al. Capsular Serotype–Specific Attack Rates and Duration of Carriage of *Streptococcus pneumoniae* in a Population of Children. *J Infect Dis*. 2006;194(5):682-688. doi:10.1086/505710

60. Faden H, Duffy L, Wasielewski R, et al. Relationship between Nasopharyngeal Colonization and the Development of Otitis Media in Children. *J Infect Dis*. 1997;175(6):1440-1445. doi:10.1086/516477

61. Usuf E, Bottomley C, Adegbola RA, Hall A. Pneumococcal Carriage in Sub-Saharan Africa—A Systematic Review. Trotter CL, ed. *PLoS ONE*. 2014;9(1):e85001. doi:10.1371/journal.pone.0085001

62. Yahiaoui RY, den Heijer CD, van Bijnen EM, et al. Prevalence and antibiotic resistance of commensal *Streptococcus pneumoniae* in nine European countries. *Future Microbiol*. 2016;11(6):737-744. doi:10.2217/fmb-2015-0011

63. Abdullahi O, Karani A, Tigoi CC, et al. The Prevalence and Risk Factors for Pneumococcal Colonization of the Nasopharynx among Children in Kilifi District, Kenya. Ratner AJ, ed. *PLoS ONE*. 2012;7(2):e30787. doi:10.1371/journal.pone.0030787

64. Regev-Yochay G, Raz M, Dagan R, et al. Nasopharyngeal Carriage of *Streptococcus pneumoniae* by Adults and Children in Community and Family Settings. *Clin Infect Dis*. 2004;38(5):632-639. doi:10.1086/381547

65. Neal EFG, Chan J, Nguyen CD, Russell FM. Factors associated with pneumococcal nasopharyngeal carriage: A systematic review. Homaira N, ed. *PLOS Glob Public Health*. 2022;2(4):e0000327. doi:10.1371/journal.pgph.0000327

66. Weiser JN, Ferreira DM, Paton JC. *Streptococcus pneumoniae*: transmission, colonization and invasion. *Nat Rev Microbiol*. 2018;16(6):355-367. doi:10.1038/s41579-018-0001-8

67. Zafar MA, Hamaguchi S, Zangari T, Cammer M, Weiser JN. Capsule Type and Amount Affect Shedding and Transmission of *Streptococcus pneumoniae*. Adam Thornton J, ed. *mBio*. 2017;8(4):e00989-17. doi:10.1128/mBio.00989-17

68. Zafar MA, Wang Y, Hamaguchi S, Weiser JN. Host-to-Host Transmission of *Streptococcus pneumoniae* Is Driven by Its Inflammatory Toxin, Pneumolysin. *Cell Host Microbe*. 2017;21(1):73-83. doi:10.1016/j.chom.2016.12.005

69. Diavatopoulos DA, Short KR, Price JT, et al. Influenza A virus facilitates *Streptococcus pneumoniae* transmission and disease. *FASEB J*. 2010;24(6):1789-1798. doi:10.1096/fj.09-146779

70. Morimura A, Hamaguchi S, Akeda Y, Tomono K. Mechanisms Underlying Pneumococcal Transmission and Factors Influencing Host-Pneumococcus Interaction: A Review. *Front Cell Infect Microbiol*. 2021;11:639450. doi:10.3389/fcimb.2021.639450

71. Musher DM. How Contagious Are Common Respiratory Tract Infections? *N Engl J Med*. Published online 2003:11.

72. Huang SS. Community-Level Predictors of Pneumococcal Carriage and Resistance in Young Children. *Am J Epidemiol*. 2004;159(7):645-654. doi:10.1093/aje/kwh088

73. García-Rodríguez JÁ, Fresnadillo Martínez MJ. Dynamics of nasopharyngeal colonization by potential respiratory pathogens. *J Antimicrob Chemother*. 2002;50(suppl_3):59-74. doi:10.1093/jac/dkf506

74. Drijkoningen JJC. Pneumococcal infection in adults: burden of disease. :7.

75. Asner SA, Agyeman PKA, Gradoux E, et al. Burden of *Streptococcus pneumoniae* Sepsis in Children After Introduction of Pneumococcal Conjugate Vaccines: A Prospective Population-based Cohort Study. *Clin Infect Dis*. 2019;69(9):1574-1580. doi:10.1093/cid/ciy1139

76. Troeger C, Blacker B, Khalil IA, et al. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Infect Dis*. 2018;18(11):1191-1210. doi:10.1016/S1473-3099(18)30310-4

77. Bluestone CD. Pathogenesis of otitis media: role of eustachian tube. *Pediatr Infect Dis J*. 1996;15(4):281-291. Accessed April 22, 2022. https://journals.lww.com/pidj/fulltext/1996/04000/pathogenesis_of_otitis_media__role_of_eustachian.2.aspx

78. Hanage WP, Kaijalainen T, Saukkoriipi A, Rickcord JL, Spratt BG. A Successful, Diverse Disease-Associated Lineage of Nontypeable Pneumococci That Has Lost the Capsular Biosynthesis Locus. *J Clin Microbiol*. 2006;44(3):743-749. doi:10.1128/JCM.44.3.743-749.2006

79. Buck JM, Lexau C, Shapiro M, et al. A community outbreak of conjunctivitis caused by nontypeable *Streptococcus pneumoniae* in Minnesota. *Pediatr Infect Dis J*. 2006;25(10):906-911. doi:10.1097/01.inf.0000238143.96607.ec

80. Olarte L, Hulten KG, Lamberth L, Mason EO, Kaplan SL. Impact of the 13-valent pneumococcal conjugate vaccine on chronic sinusitis associated with *Streptococcus pneumoniae* in children. *Pediatr Infect Dis J*. 2014;33(10):1033-1036. doi:10.1097/INF.0000000000000387

81. Grijalva CG. Antibiotic Prescription Rates for Acute Respiratory Tract Infections in US Ambulatory Settings. *JAMA*. 2009;302(7):758. doi:10.1001/jama.2009.1163

82. Brown AO, Mann B, Gao G, et al. *Streptococcus pneumoniae* Translocates into the Myocardium and Forms Unique Microlesions That Disrupt Cardiac Function. Wessels MR, ed. *PLoS Pathog*. 2014;10(9):e1004383. doi:10.1371/journal.ppat.1004383

83. Fletcher MA, Schmitt HJ, Syrochkina M, Sylvester G. Pneumococcal empyema and complicated pneumonias: global trends in incidence, prevalence, and serotype epidemiology. *Eur J Clin Microbiol Infect Dis*. 2014;33(6):879-910. doi:10.1007/s10096-014-2062-6

84. Mook-Kanamori BB, Geldhoff M, van der Poll T, van de Beek D. Pathogenesis and Pathophysiology of Pneumococcal Meningitis. *Clin Microbiol Rev*. 2011;24(3):557-591. doi:10.1128/CMR.00008-11

85. Wahl B, O'Brien KL, Greenbaum A, et al. Burden of *Streptococcus pneumoniae* and *Haemophilus influenzae* type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000–15. *Lancet Glob Health*. 2018;6(7):e744-e757. doi:10.1016/S2214-109X(18)30247-X

86. Vinogradova Y, Hippisley-Cox J, Coupland C. Identification of new risk factors for pneumonia: population-based case-control study. *Br J Gen Pract*. 2009;59(567):e329-e338. doi:10.3399/bjgp09X472629

87. Brown EJ, Joiner KA, Cole RM, Berger M. Localization of complement component 3 on *Streptococcus pneumoniae*: anti-capsular antibody causes complement deposition on the pneumococcal capsule. *Infect Immun*. 1983;39(1):403-409. doi:10.1128/iai.39.1.403-409.1983

88. Andre GO, Converso TR, Politano WR, et al. Role of *Streptococcus pneumoniae* Proteins in Evasion of Complement-Mediated Immunity. *Front Microbiol*. 2017;8. doi:10.3389/fmicb.2017.00224

89. Merle NS, Church SE, Fremeaux-Bacchi V, Roumenina LT. Complement System Part I - Molecular Mechanisms of Activation and Regulation. *Front Immunol*. 2015;6. doi:10.3389/fimmu.2015.00262

90. Merle NS, Noe R, Halbwachs-Mecarelli L, Fremeaux-Bacchi V, Roumenina LT. Complement System Part II: Role in Immunity. *Front Immunol*. 2015;6. doi:10.3389/fimmu.2015.00257

91. Brown JS, Hussell T, Gilliland SM, et al. The classical pathway is the dominant complement pathway required for innate immunity to *Streptococcus pneumoniae* infection in mice. *Proc Natl Acad Sci*. 2002;99(26):16969-16974. doi:10.1073/pnas.012669199

92. Janoff EN, Fasching C, Orenstein JM, Rubins JB, Opstad NL, Dalmasso AP. Killing of *Streptococcus pneumoniae* by capsular polysaccharide–specific polymeric IgA, complement, and phagocytes. *J Clin Invest*. 1999;104(8):1139-1147. doi:10.1172/JCI6310

93. Standish AJ, Weiser JN. Human Neutrophils Kill *Streptococcus pneumoniae* via Serine Proteases. *J Immunol*. 2009;183(4):2602-2609. doi:10.4049/jimmunol.0900688

94. Brouwer MC, Baas F, van der Ende A, van de Beek D. Genetic Variation and Cerebrospinal Fluid Levels of Mannose Binding Lectin in Pneumococcal Meningitis Patients. Manganelli R, ed. *PLoS ONE*. 2013;8(5):e65151. doi:10.1371/journal.pone.0065151

95. Yuste J, Sen A, Truedsson L, et al. Impaired Opsonization with C3b and Phagocytosis of *Streptococcus pneumoniae* in Sera from Subjects with Defects in the Classical Complement Pathway. *Infect Immun*. 2008;76(8):3761-3770. doi:10.1128/IAI.00291-08

96. Ferreira DM, Neill DR, Bangert M, et al. Controlled Human Infection and Rechallenge with *Streptococcus pneumoniae* Reveals the Protective Efficacy of Carriage in Healthy Adults. *Am J Respir Crit Care Med*. 2013;187(8):855-864. doi:10.1164/rccm.201212-2277OC

97. Cohen JM, Khandavilli S, Camberlein E, Hyams C, Baxendale HE, Brown JS. Protective Contributions against Invasive *Streptococcus pneumoniae* Pneumonia of Antibody and Th17-Cell Responses to Nasopharyngeal Colonisation. Metzger DW, ed. *PLoS ONE*. 2011;6(10):e25558. doi:10.1371/journal.pone.0025558

98. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal Relationships between Invasive and Carriage *Streptococcus pneumoniae* and Serotype- and Clone-Specific Differences in Invasive Disease Potential. *J Infect Dis*. 2003;187(9):1424-1432. doi:10.1086/374624

99. Sjostrom K, Spindler C, Ortqvist A, et al. Clonal and Capsular Types Decide Whether Pneumococci Will Act as a Primary or Opportunistic Pathogen. *Clin Infect Dis*. 2006;42(4):451-459. doi:10.1086/499242

100. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. *The Lancet*. 2011;378(9807):1962-1973. doi:10.1016/S0140-6736(10)62225-8

101. Brueggemann AB, Peto TEA, Crook DW, Butler JC, Kristinsson KG, Spratt BG. Temporal and Geographic Stability of the Serogroup-Specific Invasive Disease Potential of *Streptococcus pneumoniae* in Children. *J Infect Dis*. 2004;190(7):1203-1211. doi:10.1086/423820

102. Balsells E, Guillot L, Nair H, Kyaw MH. Serotype distribution of *Streptococcus pneumoniae* causing invasive disease in children in the post-PCV era: A systematic review and meta-analysis. Borrow R, ed. *PLOS ONE*. 2017;12(5):e0177113. doi:10.1371/journal.pone.0177113

103. Li Y, Weinberger DM, Thompson CM, Trzciński K, Lipsitch M. Surface Charge of *Streptococcus pneumoniae* Predicts Serotype Distribution. Pirofski L, ed. *Infect Immun*. 2013;81(12):4519-4524. doi:10.1128/IAI.00724-13

104. Hyams C, Trzcinski K, Camberlein E, et al. *Streptococcus pneumoniae* Capsular Serotype Invasiveness Correlates with the Degree of Factor H Binding and Opsonization with C3b/iC3b. Camilli A, ed. *Infect Immun*. 2013;81(1):354-363. doi:10.1128/IAI.00862-12

105. Nelson AL, Roche AM, Gould JM, Chim K, Ratner AJ, Weiser JN. Capsule Enhances Pneumococcal Colonization by Limiting Mucus-Mediated Clearance. *Infect Immun*. 2007;75(1):83-90. doi:10.1128/IAI.01475-06

106. Wood WB, Smith MR. The inhibition of surface phagocytosis by the capsular slime layer of pneumococcus type III. *J Exp Med*. 1949;90(1):85-96. doi:10.1084/jem.90.1.85

107. Harboe ZB, Thomsen RW, Riis A, et al. Pneumococcal Serotypes and Mortality following Invasive Pneumococcal Disease: A Population-Based Cohort Study. Klugman KP, ed. *PLoS Med*. 2009;6(5):e1000081. doi:10.1371/journal.pmed.1000081

108. Choi EH, Zhang F, Lu YJ, Malley R. Capsular Polysaccharide (CPS) Release by Serotype 3 Pneumococcal Strains Reduces the Protective Effect of Anti-Type 3 CPS Antibodies. Burns DL, ed. *Clin Vaccine Immunol*. 2016;23(2):162-167. doi:10.1128/CVI.00591-15

109. Carvalho MGS, Steigerwalt AG, Thompson T, Jackson D, Facklam RR. Confirmation of Nontypeable *Streptococcus pneumoniae*- Like Organisms Isolated from Outbreaks of Epidemic Conjunctivitis as *Streptococcus pneumoniae*. *J Clin Microbiol*. 2003;41(9):4415-4417. doi:10.1128/JCM.41.9.4415-4417.2003

110. Mohale T, Wolter N, Allam M, et al. Genomic differences among carriage and invasive nontypeable pneumococci circulating in South Africa. *Microb Genomics*. 2019;5(10). doi:10.1099/mgen.0.000299

111. Keller LE, Robinson DA, McDaniel LS. Nonencapsulated *Streptococcus pneumoniae*: Emergence and Pathogenesis. Camilli A, Collier RJ, eds. *mBio*. 2016;7(2):e01792-15. doi:10.1128/mBio.01792-15

112. Park IH, Geno KA, Sherwood LK, Nahm MH, Beall B. Population-Based Analysis of Invasive Nontypeable Pneumococci Reveals That Most Have Defective Capsule Synthesis Genes. Miyaji EN, ed. *PLoS ONE*. 2014;9(5):e97825. doi:10.1371/journal.pone.0097825

113. Zegans ME, Sanchez PA, Likosky DS, et al. Clinical Features, Outcomes, and Costs of a Conjunctivitis Outbreak Caused by the ST448 Strain of *Streptococcus pneumoniae*. *Cornea*. 2009;28(5):503-509. doi:10.1097/ICO.0b013e3181909362

114. Habib M, Porter BD, Satzke C. Capsular Serotyping of *Streptococcus pneumoniae* Using the Quellung Reaction. *J Vis Exp*. 2014;(84):51208. doi:10.3791/51208

115. Lalitha MK, Pai R, John TJ, et al. Serotyping of *Streptococcus pneumoniae* by agglutination assays: a cost-effective technique for developing countries. 1996;74:4.

116. Satzke C, Dunne EM, Porter BD, Klugman KP, Mulholland EK, PneuCarriage project group. The PneuCarriage Project: A Multi-Centre Comparative Study to Identify the Best Serotyping Methods for Examining Pneumococcal Carriage in Vaccine Evaluation Studies. Bell D, ed. *PLOS Med*. 2015;12(11):e1001903. doi:10.1371/journal.pmed.1001903

117. Tomita Y, Okamoto A, Yamada K, Yagi T, Hasegawa Y, Ohta M. A New Microarray System to Detect *Streptococcus pneumoniae* Serotypes. *J Biomed Biotechnol*. 2011;2011:1-21. doi:10.1155/2011/352736

118. Pimenta FC, Roundtree A, Soysal A, et al. Sequential Triplex Real-Time PCR Assay for Detecting 21 Pneumococcal Capsular Serotypes That Account for a High Global Disease Burden. *J Clin Microbiol*. 2013;51(2):647-652. doi:10.1128/JCM.02927-12

119. Pai R, Gertz RE, Beall B. Sequential Multiplex PCR Approach for Determining Capsular Serotypes of *Streptococcus pneumoniae* Isolates. *J CLIN MICROBIOL*. 2006;44:8.

120. Saha SK, Darmstadt GL, Baqui AH, et al. Identification of Serotype in Culture Negative Pneumococcal Meningitis Using Sequential Multiplex PCR: Implication for Surveillance and Vaccine Design. Ratner AJ, ed. *PLoS ONE*. 2008;3(10):e3576. doi:10.1371/journal.pone.0003576

121. Ing J, Mason EO, Kaplan SL, et al. Characterization of Nontypeable and Atypical *Streptococcus pneumoniae* Pediatric Isolates from 1994 to 2010. *J Clin Microbiol*. 2012;50(4):1326-1330. doi:10.1128/JCM.05182-11

122. van Tonder AJ, Bray JE, Roalfe L, et al. Genomics Reveals the Worldwide Distribution of Multidrug-Resistant Serotype 6E Pneumococci. Munson E, ed. *J Clin Microbiol*. 2015;53(7):2271-2285. doi:10.1128/JCM.00744-15

123. Kapatai G, Sheppard CL, Al-Shahib A, et al. Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ*. 2016;4:e2477. doi:10.7717/peerj.2477

124. Epping L, van Tonder AJ, Gladstone RA, et al. SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data. *Microb Genomics*. 2018;4(7). doi:10.1099/mgen.0.000186

125. Mavroidi A, Godoy D, Aanensen DM, Robinson DA, Hollingshead SK, Spratt BG. Evolutionary Genetics of the Capsular Locus of Serogroup 6 Pneumococci. *J Bacteriol*. 2004;186(24):8181-8192. doi:10.1128/JB.186.24.8181-8192.2004

126. Burton RL, Geno KA, Saad JS, Nahm MH. Pneumococcus with the "6E" *cps* Locus Produces Serotype 6B Capsular Polysaccharide. Diekema DJ, ed. *J Clin Microbiol*. 2016;54(4):967-971. doi:10.1128/JCM.03194-15

127. Lau GW, Haataja S, Lonetto M, et al. A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence: Functional genomics of Streptococcus pneumoniae. *Mol Microbiol*. 2001;40(3):555-571. doi:10.1046/j.1365-2958.2001.02335.x

128. Green AE, Howarth D, Chaguza C, et al. Pneumococcal Colonization and Virulence Factors Identified Via Experimental Evolution in Infection Models. Agashe D, ed. *Mol Biol Evol*. 2021;38(6):2209-2226. doi:10.1093/molbev/msab018

129. Orihuela CJ, Radin JN, Sublett JE, Gao G, Kaushal D, Tuomanen EI. Microarray Analysis of Pneumococcal Gene Expression during Invasive Disease. *Infect Immun*. 2004;72(10):5582-5596. doi:10.1128/IAI.72.10.5582-5596.2004

130. Nishimoto AT, Rosch JW, Tuomanen EI. Pneumolysin: Pathogenesis and Therapeutic Target. *Front Microbiol*. 2020;11:1543. doi:10.3389/fmicb.2020.01543

131. Hirst RA, Kadioglu A, O'Callaghan C, Andrew PW. The role of pneumolysin in pneumococcal pneumonia and meningitis. *Clin Exp Immunol*. 2004;138(2):195-201. doi:10.1111/j.1365-2249.2004.02611.x

132. Yuste J, Botto M, Paton JC, Holden DW, Brown JS. Additive Inhibition of Complement Deposition by Pneumolysin and PspA Facilitates *Streptococcus pneumoniae* Septicemia. *J Immunol*. 2005;175(3):1813-1819. doi:10.4049/jimmunol.175.3.1813

133. Rosenow C, Ryan P, Weiser JN, et al. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of *Streptococcus pneumoniae*. *Mol Microbiol*. 1997;25(5):819-829. doi:10.1111/j.1365-2958.1997.mmi494.x

134. Barocchi MA, Ries J, Zogaj X, et al. A pneumococcal pilus influences virulence and host inflammatory responses. *Proc Natl Acad Sci*. 2006;103(8):2857-2862. doi:10.1073/pnas.0511017103

135. Ness S, Hilleringmann M. *Streptococcus pneumoniae* Type 1 Pilus – A Multifunctional Tool for Optimized Host Interaction. *Front Microbiol*. 2021;12:615924. doi:10.3389/fmicb.2021.615924

136. Shak JR, Vidal JE, Klugman KP. Influence of bacterial interactions on pneumococcal colonization of the nasopharynx. *Trends Microbiol*. 2013;21(3):129-135. doi:10.1016/j.tim.2012.11.005

137. De Boeck I, Wittouck S, Wuyts S, et al. Comparing the Healthy Nose and Nasopharynx Microbiota Reveals Continuity As Well As Niche-Specificity. *Front Microbiol*. 2017;8:2372. doi:10.3389/fmicb.2017.02372

138. Man WH, de Steenhuijsen Piters WAA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat Rev Microbiol*. 2017;15(5):259-270. doi:10.1038/nrmicro.2017.14

139. Bogaert D, Keijser B, Huse S, et al. Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. Semple M, ed. *PLoS ONE*. 2011;6(2):e17035. doi:10.1371/journal.pone.0017035

140. Biesbroek G, Tsivtsivadze E, Sanders EAM, et al. Early Respiratory Microbiota Composition Determines Bacterial Succession Patterns and Respiratory Health in Children. *Am J Respir Crit Care Med*. 2014;190(11):1283-1292. doi:10.1164/rccm.201407-1240OC

141. Flynn M, Dooley J. The microbiome of the nasopharynx. *J Med Microbiol*. 2021;70(6). doi:10.1099/jmm.0.001368

142. Man WH, Clerc M, de Steenhuijsen Piters WAA, et al. Loss of Microbial Topography between Oral and Nasopharyngeal Microbiota and Development of Respiratory Infections Early in Life. *Am J Respir Crit Care Med*. 2019;200(6):760-770. doi:10.1164/rccm.201810-1993OC

143. Ghoul M, Mitri S. The Ecology and Evolution of Microbial Competition. *Trends Microbiol*. 2016;24(10):833-845. doi:10.1016/j.tim.2016.06.011

144. Auranen K, Mehtälä J, Tanskanen A, S. Kaltoft M. Between-Strain Competition in Acquisition and Clearance of Pneumococcal Carriage—Epidemiologic Evidence From a Longitudinal Study of Day-Care Children. *Am J Epidemiol*. 2010;171(2):169-176. doi:10.1093/aje/kwp351

145. Kono M, Zafar MA, Zuniga M, Roche AM, Hamaguchi S, Weiser JN. Single Cell Bottlenecks in the Pathogenesis of *Streptococcus pneumoniae*. Wessels MR, ed. *PLOS Pathog*. 2016;12(10):e1005887. doi:10.1371/journal.ppat.1005887

146. Pericone CD, Overweg K, Hermans PWM, Weiser JN. Inhibitory and Bactericidal Effects of Hydrogen Peroxide Production by *Streptococcus pneumoniae* on Other Inhabitants of the Upper Respiratory Tract. Tuomanen EI, ed. *Infect Immun*. 2000;68(7):3990-3997. doi:10.1128/IAI.68.7.3990-3997.2000

147. Siber G, Klugman KP, Mäkelä H. *Pneumococcal Vaccines: The Impact of Conjugate Vaccines*. ASM Press; 2008.

148. Austrian R. A Brief History of Pneumococcal Vaccines: *Drugs Aging*. 1999;15(Supplement 1):1-10. doi:10.2165/00002512-199915001-00001

149. Shapiro ED, Berg AT, Austrian R, et al. The protective efficacy of polyvalent pneumococcal polysaccharide vaccine. *N Engl J Med*. 1991;325(21):1453-1460. doi:10.1056/NEJM199111213252101

150. Ahonkhai VI, Landesman SH, Fikrig SM, et al. Failure of Pneumococcal Vaccine in Children with Sickle-Cell Disease. *N Engl J Med*. 1979;301(1):26-27. doi:10.1056/NEJM197907053010106

151. Sloyer JL, Ploussard JH, Howie VM. Efficacy of pneumococcal polysaccharide vaccine in preventing acute otitis media in infants in Huntsville, Alabama. *Rev Infect Dis*. 1981;3 Suppl:S119-123. doi:10.1093/clinids/3.supplement_1.s119

152. Falkenhorst G, Remschmidt C, Harder T, Hummers-Pradier E, Wichmann O, Bogdan C. Effectiveness of the 23-Valent Pneumococcal Polysaccharide Vaccine (PPV23) against Pneumococcal Disease in the Elderly: Systematic Review and Meta-Analysis. Ho PL, ed. *PLOS ONE*. 2017;12(1):e0169368. doi:10.1371/journal.pone.0169368

153. Adams WG, Deaver KA, Cochi SL, et al. Decline of childhood *Haemophilus influenzae* type b (Hib) disease in the Hib vaccine era. *JAMA*. 1993;269(2):221-226.

154. Black S, Shinefield H, Fireman B, et al. Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children: *Pediatr Infect Dis J*. 2000;19(3):187-195. doi:10.1097/00006454-200003000-00003

155. Whitney CG, Farley MM, Hadler J, et al. Decline in Invasive Pneumococcal Disease after the Introduction of Protein–Polysaccharide Conjugate Vaccine. *N Engl J Med*. 2003;348(18):1737-1746. doi:10.1056/NEJMoa022823

156. Whitney CG, Pilishvili T, Farley MM, et al. Effectiveness of seven-valent pneumococcal conjugate vaccine against invasive pneumococcal disease: a matched case-control study. *The Lancet*. 2006;368(9546):1495-1502. doi:10.1016/S0140-6736(06)69637-2

157. Pilishvili T, Lexau C, Farley MM, et al. Sustained Reductions in Invasive Pneumococcal Disease in the Era of Conjugate Vaccine. *J Infect Dis*. 2010;201(1):32-41. doi:10.1086/648593

158. Moore MR, Link-Gelles R, Schaffner W, et al. Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance. *Lancet Infect Dis*. 2015;15(3):301-309. doi:10.1016/S1473-3099(14)71081-3

159. Hammitt LL, Etyang AO, Morpeth SC, et al. Effect of ten-valent pneumococcal conjugate vaccine on invasive pneumococcal disease and nasopharyngeal carriage in Kenya: a longitudinal surveillance study. *The Lancet*. 2019;393(10186):2146-2154. doi:10.1016/S0140-6736(18)33005-8

160. Peckeu L, van der Ende A, de Melker HE, Sanders EAM, Knol MJ. Impact and effectiveness of the 10-valent pneumococcal conjugate vaccine on invasive pneumococcal disease among children under 5 years of age in the Netherlands. *Vaccine*. 2021;39(2):431-437. doi:10.1016/j.vaccine.2020.11.018

161. Miller E, Andrews NJ, Waight PA, Slack MP, George RC. Herd immunity and serotype replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and Wales: an observational cohort study. *Lancet Infect Dis*. 2011;11(10):760-768. doi:10.1016/S1473-3099(11)70090-1

162. Rodrigo C, Bewick T, Sheppard C, et al. Impact of infant 13-valent pneumococcal conjugate vaccine on serotypes in adult pneumonia. *Eur Respir J*. 2015;45(6):1632-1641. doi:10.1183/09031936.00183614

163. Quirk SJ, Haraldsson G, Hjálmarsdóttir MÁ, et al. Vaccination of Icelandic Children with the 10-Valent Pneumococcal Vaccine Leads to a Significant Herd Effect among Adults in Iceland. Diekema DJ, ed. *J Clin Microbiol*. 2019;57(4):e01766-18. doi:10.1128/JCM.01766-18

164. Spratt BG, Greenwood BM. Prevention of pneumococcal disease by vaccination: does serotype replacement matter? *The Lancet*. 2000;356(9237):1210-1211. doi:10.1016/S0140-6736(00)02779-3

165. Feikin DR, Kagucia EW, Loo JD, et al. Serotype-Specific Changes in Invasive Pneumococcal Disease after Pneumococcal Conjugate Vaccine Introduction: A Pooled Analysis of Multiple Surveillance Sites. Viboud C, ed. *PLoS Med*. 2013;10(9):e1001517. doi:10.1371/journal.pmed.1001517

166. Beall B, McEllistrem MC, Gertz RE, et al. Pre- and Postvaccination Clonal Compositions of Invasive Pneumococcal Serotypes for Isolates Collected in the United States in 1999, 2001, and 2002. *J Clin Microbiol*. 2006;44(3):999-1017. doi:10.1128/JCM.44.3.999-1017.2006

167. Gladstone RA, Jefferies JM, Tocheva AS, et al. Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. *Vaccine*. 2015;33(17):2015-2021. doi:10.1016/j.vaccine.2015.03.012

168. Brandileone MCC, Almeida SCG, Minamisava R, Andrade AL. Distribution of invasive *Streptococcus pneumoniae* serotypes before and 5 years after the introduction of 10-valent pneumococcal conjugate vaccine in Brazil. *Vaccine*. 2018;36(19):2559-2566. doi:10.1016/j.vaccine.2018.04.010

169. Quirk SJ, Haraldsson G, Erlendsdóttir H, et al. Effect of Vaccination on Pneumococci Isolated from the Nasopharynx of Healthy Children and the Middle Ear of Children with Otitis Media in Iceland. Diekema DJ, ed. *J Clin Microbiol*. 2018;56(12):e01046-18. doi:10.1128/JCM.01046-18

170. Jefferies JM, Smith AJ, Edwards GFS, McMenamin J, Mitchell TJ, Clarke SC. Temporal Analysis of Invasive Pneumococcal Clones from Scotland Illustrates Fluctuations in Diversity of Serotype and Genotype in the Absence of Pneumococcal Conjugate Vaccine. *J Clin Microbiol*. 2010;48(1):87-96. doi:10.1128/JCM.01485-09

171. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. *Microbiology*. 1998;144(11):3049-3060. doi:10.1099/00221287-144-11-3049

172. Beall BW, Gertz RE, Hulkower RL, Whitney CG, Moore MR, Brueggemann AB. Shifting Genetic Structure of Invasive Serotype 19A Pneumococci in the United States. *J Infect Dis*. 2011;203(10):1360-1368. doi:10.1093/infdis/jir052

173. Coffey TJ, Enright MC, Daniels M, et al. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural

isolates of *Streptococcus pneumoniae*. *Mol Microbiol*. 1998;27(1):73-83. doi:10.1046/j.1365-2958.1998.00658.x

174. Wyres KL, Lambertsen LM, Croucher NJ, et al. Pneumococcal Capsular Switching: A Historical Perspective. *J Infect Dis*. 2013;207(3):439-449. doi:10.1093/infdis/jis703

175. WHO Publication. Pneumococcal vaccines WHO position paper – 2012 – Recommendations. *Vaccine*. 2012;30(32):4717-4718. doi:10.1016/j.vaccine.2012.04.093

176. Daniels CC, Rogers PD, Shelton CM. A Review of Pneumococcal Vaccines: Current Polysaccharide Vaccine Recommendations and Future Protein Antigens. *J Pediatr Pharmacol Ther*. 2016;21(1):27-35. doi:10.5863/1551-6776-21.1.27

177. Bergman M, Huikko S, Huovinen P, Paakkari P, Seppälä H, Finnish Study Group for Antimicrobial Resistance (FiRe Network). Macrolide and Azithromycin Use Are Linked to Increased Macrolide Resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother*. 2006;50(11):3646-3650. doi:10.1128/AAC.00234-06

178. Simoens S, Verhaegen J, van Bleyenbergh P, Peetermans WE, Decramer M. Consumption Patterns and In Vitro Resistance of *Streptococcus pneumoniae* to Fluoroquinolones. *Antimicrob Agents Chemother*. 2011;55(6):3051-3053. doi:10.1128/AAC.00019-11

179. Hicks LA, Bartoces MG, Roberts RM, et al. US Outpatient Antibiotic Prescribing Variation According to Geography, Patient Population, and Provider Specialty in 2011. *Clin Infect Dis*. Published online March 5, 2015:civ076. doi:10.1093/cid/civ076

180. Cherazard R, Epstein M, Doan TL, Salim T, Bharti S, Smith MA. Antimicrobial Resistant *Streptococcus pneumoniae*: Prevalence, Mechanisms, and Clinical Implications. *Am J Ther*. 2017;24(3):e361-e369. doi:10.1097/MJT.0000000000000551

181. Jacobs MR, Koornhof HJ, Robins-Browne RM, et al. Emergence of Multiply Resistant Pneumococci. *N Engl J Med*. 1978;299(14):735-740. doi:10.1056/NEJM197810052991402

182. McGee L, McDougal L, Zhou J, et al. Nomenclature of Major Antimicrobial-Resistant Clones of *Streptococcus pneumoniae* Defined by the Pneumococcal Molecular Epidemiology Network. *J Clin Microbiol*. 2001;39(7):2565-2571. doi:10.1128/JCM.39.7.2565-2571.2001

183. Mora-Ochomogo M, Lohans CT. β-Lactam antibiotic targets and resistance mechanisms: from covalent inhibitors to substrates. *RSC Med Chem*. 2021;12(10):1623-1639. doi:10.1039/D1MD00200G

184. Waxman DJ, Strominger JL. PENICILLIN-BINDING PROTEINS AND THE MECHANISM OF ACTION OF BETA-LACTAM ANTIBIOTICS. *Annu Rev Biochem*. 1983;52(1):825-869. doi:10.1146/annurev.bi.52.070183.004141

185. Abdullah MR, Gutiérrez-Fernández J, Pribyl T, et al. Structure of the pneumococcal L , D -carboxypeptidase DacB and pathophysiological effects of disabled cell wall hydrolases DacA and DacB: Structure of pneumococcal DacB and pathophysiological

role. *Mol Microbiol*. Published online September 2014:n/a-n/a. doi:10.1111/mmi.12729

186. Hakenbeck R, Brückner R, Denapaite D, Maurer P. Molecular mechanisms of β-lactam resistance in *Streptococcus pneumoniae*. *Future Microbiol*. 2012;7(3):395-410. doi:10.2217/fmb.12.2

187. Smith AM, Klugman KP. Alterations in MurM, a Cell Wall Muropeptide Branching Enzyme, Increase High-Level Penicillin and Cephalosporin Resistance in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother*. 2001;45(8):2393-2396. doi:10.1128/AAC.45.8.2393-2396.2001

188. Mascher T, Heintz M, Zahner D, Merai M, Hakenbeck R. The CiaRH System of *Streptococcus pneumoniae* Prevents Lysis during Stress Induced by Treatment with Cell Wall Inhibitors and by Mutations in pbp2x Involved in ␣-Lactam Resistance. *J BACTERIOL*. 2006;188:10.

189. Tran TDH, Kwon HY, Kim EH, et al. Decrease in Penicillin Susceptibility Due to Heat Shock Protein ClpL in *Streptococcus pneumoniae*. *Antimicrob Agents Chemother*. 2011;55(6):2714-2728. doi:10.1128/AAC.01383-10

190. Hansman D, Bullen MM. A RESISTANT PNEUMOCOCCUS. *The Lancet*. 1967;290(7509):264-265. doi:10.1016/S0140-6736(67)92346-X

191. Kannan K, Mankin AS. Macrolide antibiotics in the ribosome exit tunnel: species-specific binding and action: Species-specific binding and action of macrolides. *Ann N Y Acad Sci*. 2011;1241(1):33-47. doi:10.1111/j.1749-6632.2011.06315.x

192. Xiao Y, Wei Z, Shen P, et al. Bacterial-resistance among outpatients of county hospitals in China: significant geographic distinctions and minor differences between central cities. *Microbes Infect*. 2015;17(6):417-425. doi:10.1016/j.micinf.2015.02.001

193. Farrell DJ, Couturier C, Hryniewicz W. Distribution and antibacterial susceptibility of macrolide resistance genotypes in *Streptococcus pneumoniae*: PROTEKT Year 5 (2003–2004). *Int J Antimicrob Agents*. 2008;31(3):245-249. doi:10.1016/j.ijantimicag.2007.10.022

194. Halpern MT, Schmier JK, Snyder LM, et al. Meta-analysis of bacterial resistance to macrolides. *J Antimicrob Chemother*. 2005;55(5):748-757. doi:10.1093/jac/dki060

195. Schroeder MR, Stephens DS. Macrolide Resistance in *Streptococcus pneumoniae*. *Front Cell Infect Microbiol*. 2016;6. doi:10.3389/fcimb.2016.00098

196. Johnston NJ, de Azavedo JC, Kellner JD, Low DE. Prevalence and Characterization of the Mechanisms of Macrolide, Lincosamide, and Streptogramin Resistance in Isolates of *Streptococcus pneumoniae*. *Antimicrob Agents Chemother*. 1998;42(9):2425-2426. doi:10.1128/AAC.42.9.2425

197. Schroeder MR, Lohsen S, Chancey ST, Stephens DS. High-Level Macrolide Resistance Due to the Mega Element [mef(E)/mel] in *Streptococcus pneumoniae*. *Front Microbiol*. 2019;10:868. doi:10.3389/fmicb.2019.00868

198. Redgrave LS. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. 2014;22(8):8.

199. Nguyen F, Starosta AL, Arenz S, Sohmen D, Dönhöfer A, Wilson DN. Tetracycline antibiotics and resistance mechanisms. *Biol Chem*. 2014;395(5):559-575. doi:10.1515/hsz-2013-0292

200. Widdowson CA, Klugman KP. The Molecular Mechanisms of Tetracycline Resistance in the Pneumococcus. *Microb Drug Resist*. 1998;4(1):79-84. doi:10.1089/mdr.1998.4.79

201. Dinos G, Athanassopoulos C, Missiri D, et al. Chloramphenicol Derivatives as Antibacterial and Anticancer Agents: Historic Problems and Current Solutions. *Antibiotics*. 2016;5(2):20. doi:10.3390/antibiotics5020020

202. Schwarz S, Kehrenberg C, Doublet B, Cloeckaert A. Molecular basis of bacterial resistance to chloramphenicol and florfenicol. *FEMS Microbiol Rev*. 2004;28(5):519-542. doi:10.1016/j.femsre.2004.04.001

203. Mingoia M, Morici E, Morroni G, et al. Tn *5253* Family Integrative and Conjugative Elements Carrying *mef* (I) and *catQ* Determinants in *Streptococcus pneumoniae* and Streptococcus pyogenes. *Antimicrob Agents Chemother*. 2014;58(10):5886-5893. doi:10.1128/AAC.03638-14

204. Manning L, Laman M, Greenhill AR, et al. Increasing Chloramphenicol Resistance in *Streptococcus pneumoniae* Isolates from Papua New Guinean Children with Acute Bacterial Meningitis. *Antimicrob Agents Chemother*. 2011;55(9):4454-4456. doi:10.1128/AAC.00526-11

205. Nitzan O, Suponitzky U, Kennes Y, Chazan B, Raz R, Colodner R. Is Chloramphenicol Making a Comeback? 2010;12:4.

206. Hiller NL, Eutsey RA, Powell E, et al. Differences in Genotype and Virulence among Four Multidrug-Resistant *Streptococcus pneumoniae* Isolates Belonging to the PMEN1 Clone. Tse H, ed. *PLoS ONE*. 2011;6(12):e28850. doi:10.1371/journal.pone.0028850

207. Laible G, Spratt BG, Hakenbeck R. Interspecies recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Mol Microbiol*. 1991;5(8):1993-2002. doi:10.1111/j.1365-2958.1991.tb00821.x

208. Dowson CG, Hutchison A, Spratt BG. Extensive re-modelling of the transpeptidase domain of penicillin-binding protein 2B of a penicillin-resistant South African isolate of *Streptococcus pneumoniae*. *Mol Microbiol*. 1989;3(1):95-102. doi:10.1111/j.1365-2958.1989.tb00108.x

209. Sibold C, Henrichsen J, König A, Martin C, Chalkley L, Hakenbeck R. Mosaic pbpX genes of major clones of penicillin-resistant *Streptococcus pneumoniae* have evolved from pbpX genes of a penicillin-sensitive *Streptococcus oralis*. *Mol Microbiol*. 1994;12(6):1013-1023. doi:10.1111/j.1365-2958.1994.tb01089.x

210. Chi F, Nolte O, Bergmann C, Ip M, Hakenbeck R. Crossing the barrier: Evolution and spread of a major class of mosaic pbp2x in *Streptococcus pneumoniae, S. mitis* and *S. oralis. Int J Med Microbiol*. 2007;297(7-8):503-512. doi:10.1016/j.ijmm.2007.02.009

211. Farrell DJ, Klugman KP, Pichichero M. Increased Antimicrobial Resistance Among Nonvaccine Serotypes of *Streptococcus pneumoniae* in the Pediatric Population After the Introduction of 7-Valent Pneumococcal Vaccine in the United States: *Pediatr Infect Dis J*. 2007;26(2):123-128. doi:10.1097/01.inf.0000253059.84602.c3

212. Gay K, Stephens DS. Structure and Dissemination of a Chromosomal Insertion Element Encoding Macrolide Efflux in *Streptococcus pneumoniae*. :10.

213. Mingoia M, Morici E, Brenciani A, Giovanetti E, Varaldo PE. Genetic basis of the association of resistance genes mef(I) (macrolides) and catQ (chloramphenicol) in streptococci. *Front Microbiol*. 2015;5. doi:10.3389/fmicb.2014.00747

214. Simar SR, Hanson BM, Arias CA. Techniques in bacterial strain typing: past, present, and future. *Curr Opin Infect Dis*. 2021;Publish Ahead of Print. doi:10.1097/QCO.0000000000000743

215. Eyre DW, Cule ML, Wilson DJ, et al. Diverse Sources of *C. difficile* Infection Identified on Whole-Genome Sequencing. *N Engl J Med*. 2013;369(13):1195-1205. doi:10.1056/NEJMoa1216064

216. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clin Microbiol Infect*. 2018;24(4):350-354. doi:10.1016/j.cmi.2017.12.016

217. Lees JA, Ferwerda B, Kremer PHC, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun*. 2019;10(1):2176. doi:10.1038/s41467-019-09976-3

218. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software. *Mol Biol Evol*. 2013;30(5):1224-1228. doi:10.1093/molbev/mst028

219. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res*. 2019;47(11):5539-5549. doi:10.1093/nar/gkz361

220. Maiden MCJ, van Rensburg MJJ, Bray JE, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol*. 2013;11(10):728-736. doi:10.1038/nrmicro3093

221. Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*. 2009;10(1):152. doi:10.1186/1471-2105-10-152

222. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*. 2012;13(1):87. doi:10.1186/1471-2105-13-87

223. Jolley KA, Bliss CM, Bennett JS, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*. 2012;158(4):1005-1015. doi:10.1099/mic.0.055459-0

224. Drider D, Rebuffat S. *Prokaryotic Antimicrobial Peptides*. Springer

225. Riley MA, Wertz JE. Bacteriocins: Evolution, Ecology, and Application. *Annu Rev Microbiol*. 2002;56(1):117-137. doi:10.1146/annurev.micro.56.012302.161024

226. Cascales E, Buchanan SK, Duché D, et al. Colicin Biology. *Microbiol Mol Biol Rev*. 2007;71(1):158-229. doi:10.1128/MMBR.00036-06

227. Nomura M. MECHANISM OF ACTION OF COLICINES. *Proc Natl Acad Sci U S A*. 1964;52:1514-1521. doi:10.1073/pnas.52.6.1514

228. Gould JM, Cramer WA. Studies on the depolarization of the *Escherichia coli* cell membrane by colicin E1. *J Biol Chem*. 1977;252(15):5491-5497.

229. Schein SJ, Kagan BL, Finkelstein A. Colicin K acts by forming voltage-dependent channels in phospholipid bilayer membranes. *Nature*. 1978;276(5684):159-163. doi:10.1038/276159a0

230. Cramer WA, Heymann JB, Schendel SL, et al. Structure-Function of the Channel-Forming Colicins. Published online 1995:33.

231. Tagg JR, Dajani AS, Wannamaker LW. Bacteriocins of Gram-Positive Bacteria. *BACTERIOL REV*. Published online 1976:35.

232. Arnison PG, Bibb MJ, Bierbaum G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep*. 2013;30(1):108-160. doi:10.1039/C2NP20085F

233. Heilbronner S, Krismer B, Brötz-Oesterhelt H, Peschel A. The microbiome-shaping roles of bacteriocins. *Nat Rev Microbiol*. 2021;19(11):726-739. doi:10.1038/s41579-021-00569-w

234. Cotter PD, Ross RP, Hill C. Bacteriocins — a viable alternative to antibiotics? *Nat Rev Microbiol*. 2013;11(2):95-105. doi:10.1038/nrmicro2937

235. Upert G, Luther A, Obrecht D, Ermert P. Emerging peptide antibiotics with therapeutic potential. *Med Drug Discov*. 2021;9:100078. doi:10.1016/j.medidd.2020.100078

236. Imai Y, Meyer KJ, Iinishi A, et al. A new antibiotic selectively kills Gram-negative pathogens. *Nature*. 2019;576(7787):459-464. doi:10.1038/s41586-019-1791-1

237. Cebrián R, Rodríguez-Cabezas ME, Martín-Escolano R, et al. Preclinical studies of toxicity and safety of the AS-48 bacteriocin. *J Adv Res*. 2019;20:129-139. doi:10.1016/j.jare.2019.06.003

238. Knappe TA, Manzenrieder F, Mas-Moruno C, et al. Introducing Lasso Peptides as Molecular Scaffolds for Drug Design: Engineering of an Integrin Antagonist. *Angew Chem*. 2011;123(37):8873-8876. doi:10.1002/ange.201102190

239. Knerr PJ, Oman TJ, Garcia De Gonzalo CV, Lupoli TJ, Walker S, van der Donk WA. Non-proteinogenic Amino Acids in Lacticin 481 Analogues Result in More Potent Inhibition of Peptidoglycan Transglycosylation. *ACS Chem Biol*. 2012;7(11):1791-1795. doi:10.1021/cb300372b

240. Cotter PD, Hill C, Ross RP. Bacteriocins: developing innate immunity for food. *Nat Rev Microbiol*. 2005;3(10):777-788. doi:10.1038/nrmicro1273

241. Alvarez-Sieiro P, Montalbán-López M, Mu D, Kuipers OP. Bacteriocins of lactic acid bacteria: extending the family. *Appl Microbiol Biotechnol*. 2016;100(7):2939-2951. doi:10.1007/s00253-016-7343-9

242. Delves-Broughton J, Blackburn P, Evans RJ, Hugenholtz J. Applications of the bacteriocin, nisin. *Antonie Van Leeuwenhoek*. 1996;69(2):193-202. doi:10.1007/BF00399424

243. Letzel AC, Pidot SJ, Hertweck C. Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC Genomics*. 2014;15(1):983. doi:10.1186/1471-2164-15-983

244. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nat Chem Biol*. 2015;11(9):639-648. doi:10.1038/nchembio.1884

245. Azevedo AC, Bento CBP, Ruiz JC, Queiroz MV, Mantovani HC. Distribution and Genetic Diversity of Bacteriocin Gene Clusters in Rumen Microbial Genomes. Nojiri H, ed. *Appl Environ Microbiol*. 2015;81(20):7290-7304. doi:10.1128/AEM.01223-15

246. Zipperer A, Konnerth MC, Laux C, et al. Human commensals producing a novel antibiotic impair pathogen colonization. *Nature*. 2016;535(7613):511-516. doi:10.1038/nature18634

247. Vigneaud V du, Brown GB. THE SYNTHESIS OF THE NEW SULFUR-CONTAINING AMINO ACID (LANTHIONINE) ISOLATED FROM SODIUM CARBONATE-TREATED WOOL. *J Biol Chem*. 1941;138(1):151-154. doi:10.1016/S0021-9258(18)51420-4

248. Ven FJM, Hooven HW, Konings RNH, Hilbers CW. NMR studies of lantibiotics. The structure of nisin in aqueous solution. *Eur J Biochem*. 1991;202(3):1181-1188. doi:10.1111/j.1432-1033.1991.tb16488.x

249. Schnell N, Entian KD, Schneider U, et al. Prepeptide sequence of epidermin, a ribosomally synthesized antibiotic with four sulphide-rings. *Nature*. 1988;333(6170):276-278. doi:10.1038/333276a0

250. Ryan MP, Rea MC, Hill C, Ross RP. An application in cheddar cheese manufacture for a strain of *Lactococcus lactis* producing a novel broad-spectrum bacteriocin, lacticin 3147. *Appl Environ Microbiol*. 1996;62(2):612-619. doi:10.1128/aem.62.2.612-619.1996

251. Hasper HE, Kramer NE, Smith JL, et al. An Alternative Bactericidal Mechanism of Action for Lantibiotic Peptides That Target Lipid II. *Science*. 2006;313(5793):1636-1637. doi:10.1126/science.1129818

252. Rogers LA. THE INHIBITING EFFECT OF *STREPTOCOCCUS LACTIS* ON *LACTOBACILLUS BULGARICUS*. *J Bacteriol*. 1928;16(5):321-325. doi:10.1128/jb.16.5.321-325.1928

253. Mattick ATR, Hirsch A. A Powerful Inhibitory Substance Produced by Group N Streptococci. *Nature*. 1944;154(3913):551-551. doi:10.1038/154551a0

254. Allgaier H, Jung G, Werner RG, Schneider U, Zähner H. Elucidation of the Structure of Epidermin, a Ribosomally Synthesized, Tetracyclic Heterodetic Polypeptide Antibiotic. *Angew Chem Int Ed Engl*. 1985;24(12):1051-1053. doi:10.1002/anie.198510511

255. Allgaier H, Jung G, Werner RG, Schneider U, Zahner H. Epidermin: sequencing of a heterodet tetracyclic 21-peptide amide antibiotic. *Eur J Biochem*. 1986;160(1):9-22. doi:10.1111/j.1432-1033.1986.tb09933.x

256. Banjeree S, Hansen JN. Structure and expression of a gene encoding the precursos of subtilin, a small protein antibiotic. *J Biol Chem*. 1988;263(19):9508-9541.

257. Breukink E, Wiedemann I, Kraaij C van, Kuipers OP, Sahl HG, de Kruijff B. Use of the Cell Wall Precursor Lipid II by a Pore-Forming Peptide Antibiotic. *Science*. 1999;286(5448):2361-2364. doi:10.1126/science.286.5448.2361

258. Hsu STD, Breukink E, Tischenko E, et al. The nisin–lipid II complex reveals a pyrophosphate cage that provides a blueprint for novel antibiotics. *Nat Struct Mol Biol*. 2004;11(10):963-967. doi:10.1038/nsmb830

259. Böttiger T, Schneider T, Martínez B, Sahl HG, Wiedemann I. Influence of Ca^{2+} Ions on the Activity of Lantibiotics Containing a Mersacidin-Like Lipid II Binding Motif. *Appl Environ Microbiol*. 2009;75(13):4427-4434. doi:10.1128/AEM.00262-09

260. Hasper HE, de Kruijff B, Breukink E. Assembly and Stability of Nisin−Lipid II Pores. *Biochemistry*. 2004;43(36):11567-11575. doi:10.1021/bi049476b

261. Wiedemann I, Bottiger T, Bonelli RR, et al. The mode of action of the lantibiotic lacticin 3147 - a complex mechanism involving specific interaction of two peptides and the cell wall precursor lipid II. *Mol Microbiol*. 2006;61(2):285-296. doi:10.1111/j.1365-2958.2006.05223.x

262. Wakamatsu K, Choung SY, Kobayashi T, Inoue K, Higashijima T, Miyazawa T. Complex formation of peptide antibiotic Ro09-0198 with lysophosphatidylethanolamine: proton NMR analyses in dimethyl sulfoxide solution. *Biochemistry*. 1990;29(1):113-118. doi:10.1021/bi00453a013

263. Makino A, Baba T, Fujimoto K, et al. Cinnamycin (Ro 09-0198) Promotes Cell Binding and Toxicity by Inducing Transbilayer Lipid Movement. *J Biol Chem*. 2003;278(5):3204-3209. doi:10.1074/jbc.M210347200

264. Repka LM, Chekan JR, Nair SK, van der Donk WA. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem Rev*. 2017;117(8):5457-5520. doi:10.1021/acs.chemrev.6b00591

265. McAuliffe O, Ryan MP, Ross RP, Hill C, Breeuwer P, Abee T. Lacticin 3147, a Broad-Spectrum Bacteriocin Which Selectively Dissipates the Membrane Potential. Vol 64.; 1998. doi:10.1128/aem.64.2.439-445.1998

266. Qiao M, Saris PEJ. Evidence for a role of NisT in transport of the lantibiotic nisin produced by *Lactococcus lactis* N8. *FEMS Microbiol Lett*. 1996;144(1):89-93. doi:10.1111/j.1574-6968.1996.tb08513.x

267. Kuipers A, de Boef E, Rink R, et al. NisT, the Transporter of the Lantibiotic Nisin, Can Transport Fully Modified, Dehydrated, and Unmodified Prenisin and Fusions of the Leader Peptide with Non-lantibiotic Peptides. *J Biol Chem*. 2004;279(21):22176-22182. doi:10.1074/jbc.M312789200

268. Izaguirre G, Hansen JN. Use of alkaline phosphatase as a reporter polypeptide to study the role of the subtilin leader segment and the SpaT transporter in the posttranslational modifications and secretion of subtilin in *Bacillus subtilis* 168. *Appl Environ Microbiol*. 1997;63(10):3965-3971. doi:10.1128/aem.63.10.3965-3971.1997

269. Klein C, Entian KD. Genes involved in self-protection against the lantibiotic subtilin produced by *Bacillus subtilis* ATCC 6633. *Appl Environ Microbiol*. 1994;60(8):2793-2801. doi:10.1128/aem.60.8.2793-2801.1994

270. Siegers K, Entian KD. Genes involved in immunity to the lantibiotic nisin produced by *Lactococcus lactis* 6F3. *Appl Environ Microbiol*. 1995;61(3):1082-1089. doi:10.1128/aem.61.3.1082-1089.1995

271. Kuipers OP, Beerthuyzen MM, Siezen RJ, Vos WM. Characterization of the nisin gene cluster nisABTCIPR of *Lactococcus lactis*. Requirement of expression of the nisA and nisIgenes for development of immunity. *Eur J Biochem*. 1993;216(1):281-291. doi:10.1111/j.1432-1033.1993.tb18143.x

272. McAuliffe O, Hill C, Ross RP. Identification and overexpression of ltnI, a novel gene which confers immunity to the two-component lantibiotic lacticin 3147. *Microbiology*. 2000;146(1):129-138. doi:10.1099/00221287-146-1-129

273. Chatterjee C, Paul M, Xie L, van der Donk WA. Biosynthesis and Mode of Action of Lantibiotics. *Chem Rev*. 2005;105(2):633-684. doi:10.1021/cr030105v

274. Li B, Yu JPJ, Brunzelle JS, Moll GN, van der Donk WA, Nair SK. Structure and Mechanism of the Lantibiotic Cyclase Involved in Nisin Biosynthesis. *Science*. 2006;311(5766):1464-1467. doi:10.1126/science.1121422

275. Ortega MA, Hao Y, Zhang Q, Walker MC, van der Donk WA, Nair SK. Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature*. 2015;517(7535):509-512. doi:10.1038/nature13888

276. Siezen RJ, Kuipers OP, de Vos WM. Comparison of lantibiotic gene clusters and encoded proteins. *Antonie Van Leeuwenhoek*. 1996;69(2):171-184. doi:10.1007/BF00399422

277. Widdick DA, Dodd HM, Barraille P, et al. Cloning and engineering of the cinnamycin biosynthetic gene cluster from *Streptomyces cinnamoneus cinnamoneus* DSM 40005. *Proc Natl Acad Sci*. 2003;100(7):4316-4321. doi:10.1073/pnas.0230516100

278. Xie L, Miller LM, Chatterjee C, Averin O, Kelleher NL, van der Donk WA. Lacticin 481: In Vitro Reconstitution of Lantibiotic Synthetase Activity. *Science*. 2004;303(5658):679-681. doi:10.1126/science.1092600

279. Müller WM, Schmiederer T, Ensle P, Süssmuth RD. In Vitro Biosynthesis of the Prepeptide of Type-III Lantibiotic Labyrinthopeptin A2 Including Formation of a C□C Bond as a Post-Translational Modification. *Angew Chem Int Ed*. 2010;49(13):2436-2440. doi:10.1002/anie.200905909

280. Wang H, van der Donk WA. Biosynthesis of the Class III Lantipeptide Catenulipeptin. *ACS Chem Biol*. 2012;7(9):1529-1535. doi:10.1021/cb3002446

281. Goto Y, Li B, Claesen J, Shi Y, Bibb MJ, van der Donk WA. Discovery of Unique Lanthionine Synthetases Reveals New Mechanistic and Evolutionary Insights. Herschlag D, ed. *PLoS Biol*. 2010;8(3):e1000339. doi:10.1371/journal.pbio.1000339

282. Begley M, Cotter PD, Hill C, Ross RP. Identification of a Novel Two-Peptide Lantibiotic, Lichenicidin, following Rational Genome Mining for LanM Proteins. *Appl Environ Microbiol*. 2009;75(17):5451-5460. doi:10.1128/AEM.00730-09

283. Singh M, Sareen D. Novel LanT Associated Lantibiotic Clusters Identified by Genome Database Mining. Biswas I, ed. *PLoS ONE*. 2014;9(3):e91352. doi:10.1371/journal.pone.0091352

284. Sandiford SK. Genome database mining for the discovery of novel lantibiotics. *Expert Opin Drug Discov*. 2017;12(5):489-495. doi:10.1080/17460441.2017.1303475

285. Zhang Q, Yu Y, Vélasquez JE, van der Donk WA. Evolution of lanthipeptide synthetases. *Proc Natl Acad Sci*. 2012;109(45):18361-18366. doi:10.1073/pnas.1210393109

286. Acedo JZ, van Belkum MJ, Lohans CT, McKay RT, Miskolzie M, Vederas JC. Solution Structure of Acidocin B, a Circular Bacteriocin Produced by *Lactobacillus acidophilus* M46. Elkins CA, ed. *Appl Environ Microbiol*. 2015;81(8):2910-2918. doi:10.1128/AEM.04265-14

287. Himeno K, Rosengren KJ, Inoue T, et al. Identification, Characterization, and Three-Dimensional Structure of the Novel Circular Bacteriocin, Enterocin NKR-5-3B, from *Enterococcus faecium*. *Biochemistry*. 2015;54(31):4863-4876. doi:10.1021/acs.biochem.5b00196

288. Martin-Visscher LA, Gong X, Duszyk M, Vederas JC. The Three-dimensional Structure of Carnocyclin A Reveals That Many Circular Bacteriocins Share a Common Structural Motif. *J Biol Chem*. 2009;284(42):28674-28681. doi:10.1074/jbc.M109.036459

289. González C, Langdon GM, Bruix M, et al. Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. *Proc Natl Acad Sci*. 2000;97(21):11221-11226. doi:10.1073/pnas.210301097

290. Velázquez-Suárez C, Cebrián R, Gasca-Capote C, et al. Antimicrobial Activity of the Circular Bacteriocin AS-48 against Clinical Multidrug-Resistant *Staphylococcus aureus*. *Antibiotics*. 2021;10(8):925. doi:10.3390/antibiotics10080925

291. Gálvez A, Giménez-Gallego G, Maqueda M, Valdivia E. Purification and amino acid composition of peptide antibiotic AS-48 produced by *Streptococcus (Enterococcus) faecalis* subsp. *liquefaciens* S-48. *Antimicrob Agents Chemother*. 1989;33(4):437-441. doi:10.1128/AAC.33.4.437

292. Wirawan RE, Swanson KM, Kleffmann T, Jack RW, Tagg JR. Uberolysin: a novel cyclic bacteriocin produced by *Streptococcus uberis*. *Microbiology*. 2007;153(5):1619-1630. doi:10.1099/mic.0.2006/005967-0

293. Kemperman R, Jonker M, Nauta A, Kuipers OP, Kok J. Functional Analysis of the Gene Cluster Involved in Production of the Bacteriocin Circularin A by *Clostridium beijerinckii* ATCC 25752. *Appl Environ Microbiol*. 2003;69(10):5839-5848. doi:10.1128/AEM.69.10.5839-5848.2003

294. Kawulka K, Sprules T, McKay RT, et al. Structure of Subtilosin A, an Antimicrobial Peptide from *Bacillus subtilis* with Unusual Posttranslational Modifications Linking Cysteine Sulfurs to α-Carbons of Phenylalanine and Threonine. *J Am Chem Soc*. 2003;125(16):4726-4727. doi:10.1021/ja029654t

295. Kawai Y, Saito T, Kitazawa H, Itoh T. Gassericin A; an uncommon cyclic bacteriocin produced by *Lactobacillus gasseri* LA39 linked at N- and C-terminal ends. *Biosci Biotechnol Biochem*. 1998;62(12):2438-2440. doi:10.1271/bbb.62.2438

296. Martinez-Bueno M, Valdivia E, Galvez A, Coyette J, Maqueda M. Analysis of the gene cluster involved in production and immunity of the peptide antibiotic AS-48 in *Enterococcus faecalis*. *Mol Microbiol*. 1998;27(2):347-358. doi:10.1046/j.1365-2958.1998.00682.x

297. Maqueda M, Sánchez-Hidalgo M, Fernández M, Montalbán-López M, Valdivia E, Martínez-Bueno M. Genetic features of circular bacteriocins produced by Gram-positive bacteria. *FEMS Microbiol Rev*. 2008;32(1):2-22. doi:10.1111/j.1574-6976.2007.00087.x

298. Major D, Flanzbaum L, Lussier L, Davies C, Caldo KMP, Acedo JZ. Transporter Protein-Guided Genome Mining for Head-to-Tail Cyclized Bacteriocins. *Molecules*. 2021;26(23):7218. doi:10.3390/molecules26237218

299. Bayro MJ, Mukhopadhyay J, Swapna GVT, et al. Structure of Antibacterial Peptide Microcin J25: A 21-Residue Lariat Protoknot. *J Am Chem Soc*. 2003;125(41):12382-12383. doi:10.1021/ja036677e

300. Rosengren KJ, Clark RJ, Daly NL, Göransson U, Jones A, Craik DJ. Microcin J25 Has a Threaded Sidechain-to-Backbone Ring Structure and Not a Head-to-Tail Cyclized Backbone. *J Am Chem Soc*. 2003;125(41):12464-12474. doi:10.1021/ja0367703

301. Wilson KA, Kalkum M, Ottesen J, et al. Structure of Microcin J25, a Peptide Inhibitor of Bacterial RNA Polymerase, is a Lassoed Tail. *J Am Chem Soc*. 2003;125(41):12475-12483. doi:10.1021/ja036756q

302. Martin-Gómez H, Tulla-Puche J. Lasso peptides: chemical approaches and structural elucidation. *Org Biomol Chem*. 2018;16(28):5065-5080. doi:10.1039/C8OB01304G

303. Hegemann JD, Zimmermann M, Xie X, Marahiel MA. Lasso Peptides: An Intriguing Class of Bacterial Natural Products. *Acc Chem Res*. 2015;48(7):1909-1919. doi:10.1021/acs.accounts.5b00156

304. Salomón RA, Farías RN. Microcin 25, a novel antimicrobial peptide produced by *Escherichia coli*. *J Bacteriol*. 1992;174(22):7428-7435. doi:10.1128/jb.174.22.7428-7435.1992

305. Xie X, Marahiel MA. NMR as an Effective Tool for the Structure Determination of Lasso Peptides. *ChemBioChem*. 2012;13(5):621-625. doi:10.1002/cbic.201100754

306. Iwatsuki M, Tomoda H, Uchida R, Gouda H, Hirono S, Ōmura S. Lariatins, Antimycobacterial Peptides Produced by *Rhodococcus* sp. K01−B0171, Have a Lasso Structure. *J Am Chem Soc*. 2006;128(23):7486-7491. doi:10.1021/ja056780z

307. Zhu S, Su Y, Shams S, Feng Y, Tong Y, Zheng G. Lassomycin and lariatin lasso peptides as suitable antibiotics for combating mycobacterial infections: current state of biosynthesis and perspectives for production. *Appl Microbiol Biotechnol*. 2019;103(10):3931-3940. doi:10.1007/s00253-019-09771-6

308. Gavrish E, Sit CS, Cao S, et al. Lassomycin, a Ribosomally Synthesized Cyclic Peptide, Kills *Mycobacterium tuberculosis* by Targeting the ATP-Dependent Protease ClpC1P1P2. *Chem Biol*. 2014;21(4):509-518. doi:10.1016/j.chembiol.2014.01.014

309. Yuzenkova J, Delgado M, Nechaev S, et al. Mutations of Bacterial RNA Polymerase Leading to Resistance to Microcin J25. *J Biol Chem*. 2002;277(52):50867-50875. doi:10.1074/jbc.M209425200

310. Bellomio A, Vincent PA, de Arcuri BF, Farías RN, Morero RD. Microcin J25 Has Dual and Independent Mechanisms of Action in *Escherichia coli*: RNA Polymerase Inhibition and Increased Superoxide Production. *J Bacteriol*. 2007;189(11):4180-4186. doi:10.1128/JB.00206-07

311. Solbiati JO, Ciaccio M, Farías RN, González-Pastor JE, Moreno F, Salomón RA. Sequence Analysis of the Four Plasmid Genes Required To Produce the Circular Peptide Antibiotic Microcin J25. *J Bacteriol*. 1999;181(8):2659-2662. doi:10.1128/JB.181.8.2659-2662.1999

312. Pan SJ, Rajniak J, Cheung WL, Link AJ. Construction of a Single Polypeptide that Matures and Exports the Lasso Peptide Microcin J25. *ChemBioChem*. 2012;13(3):367-370. doi:10.1002/cbic.201100596

313. Clarke DJ, Campopiano DJ. Maturation of McjA precursor peptide into active microcin MccJ25. *Org Biomol Chem*. 2007;5(16):2564. doi:10.1039/b708478a

314. Skinnider MA, Johnston CW, Edgar RE, et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc Natl Acad Sci*. 2016;113(42). doi:10.1073/pnas.1609014113

315. Tietz JI, Schwalen CJ, Patel PS, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat Chem Biol*. 2017;13(5):470-478. doi:10.1038/nchembio.2319

316. Flühe L, Knappe TA, Gattner MJ, et al. The radical SAM enzyme AlbA catalyzes thioether bond formation in subtilosin A. *Nat Chem Biol*. 2012;8(4):350-357. doi:10.1038/nchembio.798

317. Lee H, Churey JJ, Worobo RW. Biosynthesis and transcriptional analysis of thurincin H, a tandem repeated bacteriocin genetic locus, produced by *Bacillus thuringiensis* SF361. *FEMS Microbiol Lett*. 2009;299(2):205-213. doi:10.1111/j.1574-6968.2009.01749.x

318. Zheng G, Hehn R, Zuber P. Mutational Analysis of the *sbo-alb* Locus of *Bacillus subtilis*: Identification of Genes Required for Subtilosin Production and Immunity. *J Bacteriol*. 2000;182(11):3266-3273. doi:10.1128/JB.182.11.3266-3273.2000

319. Rea MC, Sit CS, Clayton E, et al. Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against *Clostridium difficile*. *Proc Natl Acad Sci*. 2010;107(20):9352-9357. doi:10.1073/pnas.0913554107

320. Sit CS, van Belkum MJ, McKay RT, Worobo RW, Vederas JC. The 3D Solution Structure of Thurincin H, a Bacteriocin with Four Sulfur to α-Carbon Crosslinks. *Angew Chem Int Ed*. 2011;50(37):8718-8721. doi:10.1002/anie.201102527

321. Wambui J, Stevens MJA, Sieber S, Cernela N, Perreten V, Stephan R. Targeted Genome Mining Reveals the Psychrophilic *Clostridium estertheticum* Complex as a Potential Source for Novel Bacteriocins, Including Cesin A and Estercticin A. *Front Microbiol*. 2022;12:801467. doi:10.3389/fmicb.2021.801467

322. Bushin LB, Covington BC, Rued BE, Federle MJ, Seyedsayamdost MR. Discovery and Biosynthesis of Streptosactin, a Sactipeptide with an Alternative Topology Encoded by Commensal Bacteria in the Human Microbiome. *J Am Chem Soc*. Published online August 26, 2020. doi:10.1021/jacs.0c05546

323. Roblin C, Chiumento S, Bornet O, et al. The unusual structure of Ruminococcin C1 antimicrobial peptide confers clinical properties. *Proc Natl Acad Sci*. 2020;117(32):19168-19177. doi:10.1073/pnas.2004045117

324. Thennarasu S, Lee DK, Poon A, Kawulka KE, Vederas JC, Ramamoorthy A. Membrane permeabilization, orientation, and antimicrobial mechanism of subtilosin A. *Chem Phys Lipids*. 2005;137(1-2):38-51. doi:10.1016/j.chemphyslip.2005.06.003

325. de Saizieu A, Gardès C, Flint N, et al. Microarray-Based Identification of a Novel *Streptococcus pneumoniae* Regulon Controlled by an Autoinduced Peptide. *J Bacteriol*. 2000;182(17):4696-4703. doi:10.1128/JB.182.17.4696-4703.2000

326. Reichmann P, Hakenbeck R. Allelic variation in a peptide-inducible two-component system of *Streptococcus pneumoniae*. *FEMS Microbiol Lett*. 2000;190(2):231-236. doi:10.1111/j.1574-6968.2000.tb09291.x

327. Dawid S, Roche AM, Weiser JN. The *blp* Bacteriocins of *Streptococcus pneumoniae* Mediate Intraspecies Competition both In Vitro and In Vivo. *Infect Immun*. 2007;75(1):443-451. doi:10.1128/IAI.01775-05

328. Lux T, Nuhn M, Hakenbeck R, Reichmann P. Diversity of Bacteriocins and Activity Spectrum in *Streptococcus pneumoniae*. *J Bacteriol*. 2007;189(21):7741-7751. doi:10.1128/JB.00474-07

329. Wholey WY, Abu-Khdeir M, Yu EA, Siddiqui S, Esimai O, Dawid S. Characterization of the Competitive Pneumocin Peptides of *Streptococcus pneumoniae*. *Front Cell Infect Microbiol*. 2019;9:55. doi:10.3389/fcimb.2019.00055

330. Valente C, Dawid S, Pinto FR, et al. The *blp* Locus of Streptococcus pneumoniae Plays a Limited Role in the Selection of Strains That Can Cocolonize the Human Nasopharynx. Elkins CA, ed. *Appl Environ Microbiol*. 2016;82(17):5206-5215. doi:10.1128/AEM.01048-16

331. Bogaardt C, van Tonder AJ, Brueggemann AB. Genomic analyses of pneumococci reveal a wide diversity of bacteriocins – including pneumocyclicin, a novel circular bacteriocin. *BMC Genomics*. 2015;16(1):554. doi:10.1186/s12864-015-1729-4

332. Miller EL, Abrudan MI, Roberts IS, Rozen DE. Diverse Ecological Strategies Are Encoded by *Streptococcus pneumoniae* Bacteriocin-Like Peptides. *Genome Biol Evol*. 2016;8(4):1072-1090. doi:10.1093/gbe/evw055

333. Son MR, Shchepetov M, Adrian PV, et al. Conserved Mutations in the Pneumococcal Bacteriocin Transporter Gene, *blpA* , Result in a Complex Population Consisting of Producers and Cheaters. McDaniel LS, ed. *mBio*. 2011;2(5):e00179-11. doi:10.1128/mBio.00179-11

334. Pinchas MD, LaCross NC, Dawid S. An Electrostatic Interaction between BlpC and BlpH Dictates Pheromone Specificity in the Control of Bacteriocin Production and Immunity in *Streptococcus pneumoniae*. O'Toole GA, ed. *J Bacteriol*. 2015;197(7):1236-1248. doi:10.1128/JB.02432-14

335. Biernaskie JM, Gardner A, West SA. Multicoloured greenbeards, bacteriocin diversity and the rock-paper-scissors game. *J Evol Biol*. 2013;26(10):2081-2094. doi:10.1111/jeb.12222

336. Wang CY, Patel N, Wholey WY, Dawid S. ABC transporter content diversity in *Streptococcus pneumoniae* impacts competence regulation and bacteriocin production. *Proc Natl Acad Sci*. 2018;115(25). doi:10.1073/pnas.1804668115

337. Guiral S, Mitchell TJ, Martin B, Claverys JP. Competence-programmed predation of noncompetent cells in the human pathogen *Streptococcus pneumoniae*: Genetic requirements. *Proc Natl Acad Sci*. 2005;102(24):8710-8715. doi:10.1073/pnas.0500879102

338. Maricic N, Anderson ES, Opipari AE, Yu EA, Dawid S. Characterization of a Multipeptide Lantibiotic Locus in *Streptococcus pneumoniae*. Thornton JA, McDaniel LS, eds. *mBio*. 2016;7(1):e01656-15. doi:10.1128/mBio.01656-15

339. Rezaei Javan R, van Tonder AJ, King JP, Harrold CL, Brueggemann AB. Genome Sequencing Reveals a Large and Diverse Repertoire of Antimicrobial Peptides. *Front Microbiol*. 2018;9:2012. doi:10.3389/fmicb.2018.02012

340. Walker GV, Heng NCK, Carne A, Tagg JR, Wescombe PA. Salivaricin E and abundant dextranase activity may contribute to the anti-cariogenic potential of the probiotic candidate *Streptococcus salivarius* JH. *Microbiology*. 2016;162(3):476-486. doi:10.1099/mic.0.000237

341. Hoover SE, Perez AJ, Tsui HCT, et al. A new quorum-sensing system (TprA/PhrA) for *Streptococcus pneumoniae* D39 that regulates a lantibiotic biosynthesis gene cluster: Phr-peptide signaling by *S. pneumoniae*. *Mol Microbiol*. 2015;97(2):229-243. doi:10.1111/mmi.13029

342. Kadam A, Eutsey RA, Rosch J, et al. Promiscuous signaling by a regulatory system unique to the pandemic PMEN1 pneumococcal lineage. Orihuela CJ, ed. *PLOS Pathog*. 2017;13(5):e1006339. doi:10.1371/journal.ppat.1006339

343. Hammami R, Zouhir A, Le Lay C, Ben Hamida J, Fliss I. BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol*. 2010;10:22. doi:10.1186/1471-2180-10-22

344. Weber T, Blin K, Duddela S, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res*. 2015;43(W1):W237-W243. doi:10.1093/nar/gkv437

345. de Jong A, van Hijum SAFT, Bijlsma JJE, Kok J, Kuipers OP. BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Res*. 2006;34(Web Server):W273-W279. doi:10.1093/nar/gkl237

346. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107(1):1-8. doi:10.1016/j.ygeno.2015.11.003

347. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463

348. Anderson S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res*. 1981;9(13):3015-3027. Accessed May 17, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC327328/

349. Fleischmann RD, Adams MD, White O, et al. Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd. *Science*. 1995;269(5223):496-512. doi:10.1126/science.7542800

350. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. *Hum GENOME*. 2001;291:50.

351. Forde BM, O'Toole PW. Next-generation sequencing technologies and their impact on microbial genomics. *Brief Funct Genomics*. 2013;12(5):440-453. doi:10.1093/bfgp/els062

352. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi:10.1038/nrg.2016.49

353. Hyman ED. A new method of sequencing DNA. *Anal Biochem*. 1988;174(2):423-436. doi:10.1016/0003-2697(88)90041-3

354. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. Rodriguez-Valera F, ed. *PLoS ONE*. 2012;7(2):e30087. doi:10.1371/journal.pone.0030087

355. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-59. doi:10.1038/nature07517

356. Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*. 1979;6(7):2601-2610. Accessed May 17, 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC327874/

357. Loman NJ, Constantinidou C, Chan JZM, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol*. 2012;10(9):599-606. doi:10.1038/nrmicro2850

358. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol J*. 2020;18:9-19. doi:10.1016/j.csbj.2019.11.002

359. Page AJ, De Silva N, Hunt M, et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb Genomics*. 2016;2(8). doi:10.1099/mgen.0.000083

360. Zerbino DR. Using the Velvet *de novo* Assembler for Short-Read Sequencing Technologies. *Curr Protoc Bioinforma*. 2010;31(1). doi:10.1002/0471250953.bi1105s31

361. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021

362. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*. 2011;27(4):578-579. doi:10.1093/bioinformatics/btq683

363. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. *Genome Biol*. 2012;13(6):R56. doi:10.1186/gb-2012-13-6-r56

364. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*. 2015;13(5):278-289. doi:10.1016/j.gpb.2015.08.002

365. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics*. 2016;14(5):265-279. doi:10.1016/j.gpb.2016.05.004

366. Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P. Scaffolding of a bacterial genome using MinION nanopore sequencing. *Sci Rep*. 2015;5(1):11996. doi:10.1038/srep11996

367. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. 2010;11(1):595. doi:10.1186/1471-2105-11-595

368. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3:124. doi:10.12688/wellcomeopenres.14826.1

369. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30(14):2068-2069. doi:10.1093/bioinformatics/btu153

370. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

371. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10(1):421. doi:10.1186/1471-2105-10-421

372. Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. :8.

373. Pearson WR. An Introduction to Sequence Similarity ("Homology") Searching. Published online 2014:9.

374. Edgar RC, Batzoglou S. Multiple sequence alignment. *Curr Opin Struct Biol*. 2006;16(3):368-373. doi:10.1016/j.sbi.2006.04.004

375. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340

376. Hall BG. *Phylogenetic Trees Made Easy: A How-To Manual*. Fifth Edition. Oxford University Press; 2017.

377. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49(W1):W293-W296. doi:10.1093/nar/gkab301

378. Kluyver T, Ragan-Kelley B, Pérez F, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. :4.

379. McKinney W. Data Structures for Statistical Computing in Python. In: 2010:56-61. doi:10.25080/Majora-92bf1922-00a

380. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422-1423. doi:10.1093/bioinformatics/btp163

381. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng*. 2007;9(3):90-95. doi:10.1109/MCSE.2007.55

382. Waskom M. seaborn: statistical data visualization. *J Open Source Softw*. 2021;6(60):3021. doi:10.21105/joss.03021

383. Scott Chacon, Straub B. *Pro Git*. APress; 2014.

384. Sun Y, Veseli IA, Vaillancourt K, Frenette M, Grenier D, Pombert JF. The bacteriocin from the prophylactic candidate *Streptococcus suis* 90-1330 is widely distributed across *S. suis* isolates and appears encoded in an integrative and conjugative element. Cloeckaert A, ed. *PLOS ONE*. 2019;14(4):e0216002. doi:10.1371/journal.pone.0216002

385. Lessa FC, Milucky J, Rouphael NG, et al. *Streptococcus mitis* Expressing Pneumococcal Serotype 1 Capsule. *Sci Rep*. 2018;8(1):17959. doi:10.1038/s41598-018-35921-3

386. Pimenta F, Gertz RE, Park SH, et al. *Streptococcus infantis, Streptococcus mitis*, and *Streptococcus oralis* Strains With Highly Similar cps5 Loci and Antigenic Relatedness to Serotype 5 Pneumococci. *Front Microbiol*. 2019;9:3199. doi:10.3389/fmicb.2018.03199

387. Martínez B, Suárez JE, Rodríguez A 1996. Lactococcin 972: a homodimeric lactococcal bacteriocin whose primary target is not the plasma membrane. *Microbiology*. 142(9):2393-2398. doi:10.1099/00221287-142-9-2393

388. Martínez B, Rodríguez A, Suárez JE. Lactococcin 972, a bacteriocin that inhibits septum formation in lactococci. *Microbiology*. 2000;146(4):949-955. doi:10.1099/00221287-146-4-949

389. Martínez B, Böttiger T, Schneider T, Rodríguez A, Sahl HG, Wiedemann I. Specific Interaction of the Unmodified Bacteriocin Lactococcin 972 with the Cell Wall Precursor Lipid II. *Appl Environ Microbiol*. 2008;74(15):4666-4670. doi:10.1128/AEM.00092-08

390. Turner DL, Lamosa P, Rodríguez A, Martínez B. Structure and properties of the metastable bacteriocin Lcn972 from *Lactococcus lactis*. *J Mol Struct*. 2013;1031:207-210. doi:10.1016/j.molstruc.2012.09.076

391. Voet D, Voet JG. Chapter 12: Lipids and Membranes. In: *Biochemsitry*. 4th ed. Wiley; 2011:386-466.

392. Martínez B, Fernandez M, Suárez JE, Rodríguez A. Synthesis of lactococcin 972, a bacteriocin produced by *Lactococcus lactis* IPLA 972, depends on the expression of a plasmid- encoded bicistronic operon. *Microbiology*. 1999;(145):3155-3161.

393. Campelo AB, López-González MJ, Escobedo S, et al. Mutations Selected After Exposure to Bacteriocin Lcn972 Activate a Bce-Like Bacitracin Resistance Module in *Lactococcus lactis*. *Front Microbiol*. 2020;11:1805. doi:10.3389/fmicb.2020.01805

394. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181(4096):223-230. doi:10.1126/science.181.4096.223

395. Kuhlman B, Bradley P. Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol*. 2019;20(11):681-697. doi:10.1038/s41580-019-0163-x

396. Voet D, Voet JG. Chapter 8: Three-Dimensional Structures of Proteins. In: *Biochemistry*. 4th ed. Wiley; 2011:221-277.

397. Tomasek D, Kahne D. The assembly of β-barrel outer membrane proteins. *Curr Opin Microbiol*. 2021;60:16-23. doi:10.1016/j.mib.2021.01.009

398. Dill KA, Ozkan SB, Shell MS, Weikl TR. The Protein Folding Problem. *Annu Rev Biophys*. 2008;37(1):289-316. doi:10.1146/annurev.biophys.37.092707.153558

399. Ginalski K. Comparative modeling for protein structure prediction. *Curr Opin Struct Biol*. 2006;16(2):172-177. doi:10.1016/j.sbi.2006.02.003

400. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. 2015;10(6):845-858. doi:10.1038/nprot.2015.053

401. Käll L, Krogh A, Sonnhammer ELL. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J Mol Biol*. 2004;338(5):1027-1036. doi:10.1016/j.jmb.2004.03.016

402. von Heijne G. Signal sequences. The limits of variation. *J Mol Biol*. 1985;184(1):99-105. doi:10.1016/0022-2836(85)90046-4

403. Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res*. 2021;49(D1):D412-D419. doi:10.1093/nar/gkaa913

404. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48(D1):D265-D268. doi:10.1093/nar/gkz991

405. Blum M, Chang HY, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res*. 2021;49(D1):D344-D354. doi:10.1093/nar/gkaa977

406. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

407. Cramer P. AlphaFold2 and the future of structural biology. *Nat Struct Mol Biol*. 2021;28(9):704-705. doi:10.1038/s41594-021-00650-1

408. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

409. The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*. 2010;38(suppl_1):D142-D148. doi:10.1093/nar/gkp846

410. Varadi M, Anyango S, Deshpande M, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*. 2022;50(D1):D439-D444. doi:10.1093/nar/gkab1061

411. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013;29(21):2722-2728. doi:10.1093/bioinformatics/btt473

412. Beis K, Rebuffat S. Multifaceted ABC transporters associated to microcin and bacteriocin export. *Res Microbiol*. 2019;170(8):399-406. doi:10.1016/j.resmic.2019.07.002

413. Aggarwal SD, Yesilkaya H, Dawid S, Hiller NL. The pneumococcal social network. Blumenthal A, ed. *PLOS Pathog*. 2020;16(10):e1008931. doi:10.1371/journal.ppat.1008931

414. Ambudkar SV, Kim IW, Xia D, Sauna ZE. The A-loop, a novel conserved aromatic acid subdomain upstream of the Walker A motif in ABC transporters, is critical for ATP binding. *FEBS Lett*. 2006;580(4):1049-1055. doi:10.1016/j.febslet.2005.12.051

415. Velamakanni S, Yao Y, Gutmann DAP, van Veen HW. Multidrug Transport by the ABC Transporter Sav1866 from *Staphylococcus aureus* †. *Biochemistry*. 2008;47(35):9300-9308. doi:10.1021/bi8006737

416. Piepenbreier H, Fritz G, Gebhard S. Transporters as information processors in bacterial signalling pathways: Transporters in signalling pathways. *Mol Microbiol*. 2017;104(1):1-15. doi:10.1111/mmi.13633

417. Dintner S, Staroń A, Berchtold E, Petri T, Mascher T, Gebhard S. Coevolution of ABC Transporters and Two-Component Regulatory Systems as Resistance Modules against Antimicrobial Peptides in Firmicutes Bacteria. *J Bacteriol*. 2011;193(15):3851-3862. doi:10.1128/JB.05175-11

418. Martínez B, Zomer AL, Rodríguez A, Kok J, Kuipers OP. Cell envelope stress induced by the bacteriocin Lcn972 is sensed by the lactococcal two-component system CesSR: CesSR senses cell envelope stress in Lactococcus lactis. *Mol Microbiol*. 2007;64(2):473-486. doi:10.1111/j.1365-2958.2007.05668.x

419. Sánchez C, Hernández de Rojas A, Martínez B, et al. Nucleotide Sequence and Analysis of pBL1, a Bacteriocin-Producing Plasmid from *Lactococcus lactis* IPLA 972. *Plasmid*. 2000;44(3):239-249. doi:10.1006/plas.2000.1482

420. Antic I, Brothers KM, Stolzer M, et al. Gene Acquisition by a Distinct Phyletic Group within *Streptococcus pneumoniae* Promotes Adhesion to the Ocular Epithelium. Limbago BM, ed. *mSphere*. 2017;2(5):e00213-17. doi:10.1128/mSphere.00213-17

421. Grundmann H, Hori S, Tanner G. Determining Confidence Intervals When Measuring Genetic Diversity and the Discriminatory Abilities of Typing Methods for

Microorganisms. *J Clin Microbiol*. 2001;39(11):4190-4192. doi:10.1128/JCM.39.11.4190-4192.2001

422. Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG. Barriers to Genetic Exchange between Bacterial Species: *Streptococcus pneumoniae* Transformation. *J Bacteriol*. 2000;182(4):1016-1023. doi:10.1128/JB.182.4.1016-1023.2000

423. Ikryannikova LN, Malakhova MV, Lominadze GG, et al. Inhibitory effect of streptococci on the growth of *M. catarrhalis* strains and the diversity of putative bacteriocin-like gene loci in the genomes of *S. pneumoniae* and its relatives. *AMB Express*. 2017;7(1):218. doi:10.1186/s13568-017-0521-z

424. Maricic N, Dawid S. Using the Overlay Assay to Qualitatively Measure Bacterial Production of and Sensitivity to Pneumococcal Bacteriocins. *J Vis Exp*. 2014;(91):51876. doi:10.3791/51876

425. Wiegand I, Hilpert K, Hancock REW. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc*. 2008;3(2):163-175. doi:10.1038/nprot.2007.521

426. CLSI, Wayne P. CLSI Performance Standards for Antimicrobial Susceptibility Testing. Published online 2020. http://em100.edaptivedocs.net/GetDoc.aspx?doc=CLSI%20M100%20ED32:2022&sbsok=CLSI%20M100%20ED32:2022%20TABLE%202G&format=HTML#CLSI%20M100%20ED32:2022%20TABLE%202G

427. Schön T, Werngren J, Machado D, et al. Antimicrobial susceptibility testing of *Mycobacterium tuberculosis* complex isolates – the EUCAST broth microdilution reference method for MIC determination. *Clin Microbiol Infect*. 2020;26(11):1488-1492. doi:10.1016/j.cmi.2020.07.036

428. Mercer DK, Torres MDT, Duay SS, et al. Antimicrobial Susceptibility Testing of Antimicrobial Peptides to Better Predict Efficacy. *Front Cell Infect Microbiol*. 2020;10:326. doi:10.3389/fcimb.2020.00326

429. Tabor S. Expression using the T7 RNA polymerase/promoter system. *Curr Protoc Mol Biol*. 2001;Chapter 16:Unit16.2. doi:10.1002/0471142727.mb1602s11

430. Wingfield PT. Overview of the purification of recombinant proteins. *Curr Protoc Protein Sci*. 2015;80:6.1.1-6.1.35. doi:10.1002/0471140864.ps0601s80

431. Biolabs NE. High Efficiency Transformation Protocol (C2987H). protocols.io. Published February 15, 2022. Accessed May 21, 2022. https://www.protocols.io/view/high-efficiency-transformation-protocol-c2987h-bddti26n

432. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345. doi:10.1038/nmeth.1318

433. Chalk R, Berridge G, Shrestha L, et al. High-Throughput Mass Spectrometry Applied to Structural Genomics. *Chromatography*. 2014;1(4):159-175. doi:10.3390/chromatography1040159

434. Sun P, Tropea JE, Waugh DS. Enhancing the Solubility of Recombinant Proteins in *Escherichia coli* by Using Hexahistidine-Tagged Maltose-Binding Protein as a Fusion Partner. In: Evans Jr Thomas C, Xu MQ, eds. *Heterologous Gene Expression in E.Coli: Methods and Protocols*. Methods in Molecular Biology. Humana Press; 2011:259-274. doi:10.1007/978-1-61737-967-3_16

435. Strain-Damerell C, Burgess-Brown NA. High-Throughput Site-Directed Mutagenesis. In: Vincentelli R, ed. *High-Throughput Protein Production and Purification: Methods and Protocols*. Methods in Molecular Biology. Springer; 2019:281-296. doi:10.1007/978-1-4939-9624-7_13

436. Kunkel TA. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proc Natl Acad Sci U S A*. 1985;82(2):488-492. doi:10.1073/pnas.82.2.488

437. Brymora A, Valova VA, Robinson PJ. Protein-protein interactions identified by pull-down experiments and mass spectrometry. *Curr Protoc Cell Biol*. 2004;Chapter 17:Unit 17.5. doi:10.1002/0471143030.cb1705s22

438. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol*. 2013;31(3):233-239. doi:10.1038/nbt.2508

439. Synefiaridou D, Veening JW. Harnessing CRISPR-Cas9 for Genome Editing in *Streptococcus pneumoniae* D39V. Kivisaar M, ed. *Appl Environ Microbiol*. 2021;87(6):e02762-20. doi:10.1128/AEM.02762-20

440. Song AAL, In LLA, Lim SHE, Rahim RA. A review on *Lactococcus lactis*: from food to factory. *Microb Cell Factories*. 2017;16(1):55. doi:10.1186/s12934-017-0669-x

441. Morello E, Bermúdez-Humarán LG, Llull D, et al. *Lactococcus lactis,* an Efficient Cell Factory for Recombinant Protein Production and Secretion. *J Mol Microbiol Biotechnol*. 2008;14(1-3):48-58. doi:10.1159/000106082

442. le Roux DM, Myer L, Nicol MP, Zar HJ. Incidence and severity of childhood pneumonia in the first year of life in a South African birth cohort: the Drakenstein Child Health Study. *Lancet Glob Health*. 2015;3(2):e95-e103. doi:10.1016/S2214-109X(14)70360-2

443. Garmendia L, Hernandez A, Sanchez MB, Martinez JL. Metagenomics and antibiotics. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. 2012;18 Suppl 4:27-31. doi:10.1111/j.1469-0691.2012.03868.x

444. de Jong A, Pietersma H, Cordes M, Kuipers OP, Kok J. PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics*. 2012;13(1):299. doi:10.1186/1471-2164-13-299

445. Warrier I, Ram-Mohan N, Zhu Z, et al. The Transcriptional landscape of *Streptococcus pneumoniae* TIGR4 reveals a complex operon architecture and abundant riboregulation

critical for growth and virulence. Orihuela CJ, ed. *PLOS Pathog*. 2018;14(12):e1007461. doi:10.1371/journal.ppat.1007461

446. Vogel V, Spellerberg B. Bacteriocin Production by Beta-Hemolytic *Streptococci. Pathogens*. 2021;10(7):867. doi:10.3390/pathogens10070867

447. Soto C, Padilla C, Lobos O. Mutacins and bacteriocins like genes in *Streptococcus mutans* isolated from participants with high, moderate, and low salivary count. *Arch Oral Biol*. 2017;74:1-4. doi:10.1016/j.archoralbio.2016.10.036

448. Voet D, Voet JG. Chapter 32: Translation. In: *Biochemistry*. 4th ed. Wiley; 2011:1338-1428.

9 Appendices

9.1 General Appendices

9.1.1 Standard genetic code

Table 9.1: The standard genetic code.

First position	Second position								Third position
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACG	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Note: Codon sequences refer to mRNA sequence. On the DNA coding strand, uracil (U) residues are replaced by thymine (T) residues. Adapted from Voet and Voet, Biochemistry.[448]

9.2 Chapter 3 Appendices

9.2.1 Serotypes in the Icelandic and Kenyan pneumococcal datasets

Table 9.2: The 20 most prevalent serotypes in the Icelandic and Kenyan datasets.

Iceland		Kenya	
Serotype	**n (%)**	**Serotype**	**n (%)**
19F	331 (17.3%)	19F	228 (7.2%)
23F	180 (9.4%)	1	224 (7.1%)
6A	163 (8.5%)	6A	206 (6.5%)
19A	145 (7.6%)	19A	153 (4.8%)
6B	122 (6.4%)	15BC	146 (4.6%)
3	109 (5.7%)	35B	139 (4.4%)
11A	93 (4.9%)	15A	136 (4.3%)
15BC	93 (4.9%)	14	131 (4.1%)
14	90 (4.7%)	6E(6Bii)	131 (4.1%)
nontypable	70 (3.7%)	23F	119 (3.8%)
22F	61 (3.2%)	11A	117 (3.7%)
23A	51 (2.7%)	13	98 (3.1%)
23B	47 (2.5%)	16F	96 (3.0%)
35B	33 (1.7%)	23B	90 (2.8%)
9V	32 (1.7%)	3	90 (2.8%)
21	29 (1.5%)	34	89 (2.8%)
6C	29 (1.5%)	10A	82 (2.6%)
33F	29 (1.5%)	5	69 (2.2%)
6E(6Bii)	26 (1.4%)	9V	63 (2.0%)
16F	26 (1.4%)	21	62 (2.0%)
Other serotypes[a]	153 (8.0%)	Other serotypes	690 (21.8%)

'Other serotypes' represents additional serotypes in Iceland (n=20) and Kenya (n=36).

9.2.2 Contiguity and composition of annotated bacteriocin gene clusters in the Icelandic and Kenyan datasets

Table 9.3: Contiguity of observed full and partial bacteriocin gene clusters among Icelandic and Kenyan pneumococci.

Iceland			
Cluster	**Category**	**Frequency**	**% of total clusters**
Cib	Contiguous	1895	100
Streptococcin A	Contiguous	1537	100
Streptococcin B	Contiguous	1912	100
Streptococcin C	Contiguous	1811	94.72
	EOC	82	4.29
	Non-contiguous (multiple contigs, not EOC-adjacent)	16	0.84
	Contiguous with Ns	3	0.16
Streptococcin D	Contiguous	9	100
Streptococcin E	Contiguous	1840	99.95
	Non-contiguous (multiple contigs, not EOC-adjacent)	1	0.05
Streptocyclicin	Contiguous	860	99.77
	EOC	2	0.23
Streptolancidin A	Contiguous	152	100
Streptolancidin B	Contiguous	2	100
Streptolancidin C	Contiguous	798	100
Streptolancidin D	Contiguous	183	99.46
	EOC	1	0.54
Streptolancidin E	Contiguous	461	85.06
	EOC	47	8.67
	Non-contiguous (multiple contigs, not EOC-adjacent)	15	2.77
	Contiguous with Ns	13	2.40
	Non-contiguous (multiple contigs, non-adjacent loci)	6	1.11
Streptolancidin F	Contiguous	95	100
Streptolancidin G	Contiguous	252	100
Streptolancidin J	Contiguous	1005	99.80

	Non-contiguous (one contig)	2	0.20
Streptolancidin K	Contiguous	2	100
Streptolassin	Contiguous	48	100
Streptosactin	Contiguous	1	100
Kenya			
Cluster	**Category**	**Frequency**	**% of total clusters**
Cib	Contiguous	3159	100
Streptococcin A	Contiguous	2559	99.88
	Non-contiguous (multiple contigs, non-adjacent loci)	3	0.12
Streptococcin B	Contiguous	3157	99.97
	EOC	1	0.03
Streptococcin C	Contiguous	3152	99.78
	EOC	7	0.22
Streptococcin D	Contiguous	85	100
Streptococcin E	Contiguous	3099	98.23
	Non-contiguous (multiple contigs, non-adjacent loci)	26	0.82
	EOC	16	0.51
	Non-contiguous (one contig)	8	0.25
	Non-contiguous (multiple contigs, not EOC-adjacent)	6	0.19
Streptocyclicin	Contiguous	1516	99.48
	EOC	7	0.46
	Non-contiguous (multiple contigs, not EOC-adjacent)	1	0.07
Streptolancidin A	Contiguous	8	100
Streptolancidin B	Contiguous	338	100
Streptolancidin C	Contiguous	1767	97.68
	EOC	38	2.10
	Non-contiguous (multiple contigs, non-adjacent loci)	3	0.17
	Non-contiguous (multiple contigs, not EOC-adjacent)	1	0.06
Streptolancidin D	Contiguous	850	99.65
	Non-contiguous (multiple contigs, non-adjacent loci)	1	0.12
	EOC	1	0.12

	Non-contiguous (multiple contigs, not EOC-adjacent)	1	0.12
Streptolancidin E	Contiguous	518	97.37
	EOC	8	1.50
	Non-contiguous (multiple contigs, not EOC-adjacent)	4	0.75
	Non-contiguous (multiple contigs, non-adjacent loci)	1	0.19
	Contiguous with Ns	1	0.19
Streptolancidin F	Contiguous	25	100
Streptolancidin G	Contiguous	272	100
Streptolancidin J	Contiguous	1691	99.12
	EOC	6	0.35
	Contiguous with Ns	5	0.29
	Non-contiguous (multiple contigs, not EOC-adjacent)	3	0.18
	Non-contiguous (multiple contigs, non-adjacent loci)	1	0.06
Streptolancidin K	Contiguous	2	100
Streptolassin	Contiguous	77	100

Note: Clusters were categorised according to the proximity of the constituent genes to one another and any clusters with an intergenic region >2.5 Kb were categorised as non-contiguous. Bacteriocin clusters with genes on multiple contigs were categorised as 'end of contig' (EOC) if the genes were found within 2.5 Kb of each other and the end of the contig, otherwise the clusters were categorised as non-contiguous (multiple contigs). Each category is shown by count and percentage of the observed clusters in the dataset. Rows shown in grey represent non-contiguous clusters, which were excluded from further analysis.

Table 9.4: Compositions of observed bacteriocin clusters by category (full, partial or fragment) among Icelandic and Kenyan pneumococci.

Bacteriocin	Country	Profile	Category	Frequency
Cib	Kenya	A-B-C	Full	3159
	Iceland	A-B-C	Full	1895
		/-/-C	Fragment	17
Streptococcin A	Kenya	A-B-C	Full	2562
	Iceland	A-B-C	Full	1537
Streptococcin B	Kenya	A-B-C	Full	2058
		/-B-C	Partial	1100
		/-B-/	Fragment	1
	Iceland	A-B-C	Full	1408
		/-B-C	Partial	504
Streptococcin C	Kenya	A-B-C	Full	3159
	Iceland	A-B-C	Full	1912
Streptococcin D	Kenya	A-B-C	Full	85
	Iceland	A-B-C	Full	9
Streptococcin E	Kenya	A-B-C	Full	1048
		/-B-C	Partial	2107
	Iceland	A-B-C	Full	699
		/-B-C	Partial	1142
Streptocyclicin	Kenya	A-B-C-D-E	Full	1524
		/-/-/-D-/	Fragment	1
	Iceland	A-B-C-D-E	Full	862
Streptolancidin A	Kenya	A1-A2-A3-A4-A5-F-E-K-R-M-T	Full	8
		/-/-/-/-/-/-/-/-/-/-T	Fragment	2
	Iceland	A1-A2-A3-A4-A5-F-E-K-R-M-T	Full	152
		/-/-/-/-/-/-/-/-/-/-T	Fragment	38
Streptolancidin B	Kenya	F-G-E-A-M-T	Full	2
		F-G-E-/-/-/	Partial	336
	Iceland	F-G-E-/-/-/	Partial	2
Streptolancidin C	Kenya	A-X-L-T	Full	826
		A-X-/-/	Partial	983
		/-/-/-T	Fragment	1
	Iceland	A-X-L-T	Full	106
		A-X-/-/	Partial	692
Streptolancidin D	Kenya	A-B-C-T	Full	853
	Iceland	A-B-C-T	Full	184
Streptolancidin E	Kenya	M1-A1-A2-M2-M3-T-X1-F-G-X2	Full	29
		/-/-/-/-M3-T-X1-F-G-X2	Partial	493
		/-/-/-/-M3-T-/-F-G-X2	Partial	10

	Iceland	M1-A1-A2-M2-M3-T-X1-F-G-X2	Full	142
		/-/-/-/-M3-T-X1-F-G-X2	Partial	309
		/-/-/-/-M3-T-/-F-G-X2	Partial	91
Streptolancidin F	Kenya	A-L	Full	25
	Iceland	A-L	Full	95
Streptolancidin G	Kenya	A1-A2-M-D-P1-T-P2	Full	272
	Iceland	A1-A2-M-D-P1-T-P2	Full	252
Streptolancidin J	Kenya	A1-L-P-T1-T2-T3-A2	Full	426
		A1-L-P-T1-T2-T3-/	Partial	1
		A1-L-/-T1-T2-T3-A2	Partial	931
		A1-L-/-T1-T2-T3-/	Partial	340
		A1-/-/-/-/-/-/	Fragment	2
		/-L-P-T1-T2-T3-A2	Partial	1
		/-L-/-T1-T2-T3-A2	Partial	6
		/-L-/-T1-T2-T3-/	Partial	1
		/-/-/-/-/-T3-/	Fragment	1
	Iceland	A1-L-P-T1-T2-T3-A2	Full	535
		A1-L-P-T1-T2-T3-/	Partial	40
		A1-L-/-T1-T2-T3-A2	Partial	383
		A1-L-/-T1-T2-T3-/	Partial	30
		A1-L-/-/-T2-T3-A2	Partial	1
		/-L-P-T1-T2-T3-A2	Partial	12
		/-L-/-T1-T2-T3-A2	Partial	6
Streptolancidin K	Kenya	A-L-T	Full	2
		/-/-T	Fragment	3
	Iceland	A-L-T	Full	2
		/-/-T	Fragment	5
Streptolassin	Kenya	A-C-B1-B2-F-E-G-R-K	Full	77
	Iceland	A-C-B1-B2-F-E-G-R-K	Full	48
Streptosactin	Iceland	A-CD-X1-X2-P-X3	Full	1

Note: Rows shaded in grey indicate fragmented clusters, which were excluded from further analysis.

9.2.3 Streptolancidin association with clonal complexes

Table 9.5: Streptolancidin bacteriocins present in significantly different frequencies among Icelandic and Kenyan pneumococci, stratified by clonal complex (CC).

Number of pneumococci harbouring each bacteriocin n (% of CC representatives in each dataset with the bacteriocin)		
Streptolancidin A		
CC	**Iceland**	**Kenya**
CC138/176	122 (100)	1 (0.8)
CC448	29 (100)	2 (100)
CC802	0	5 (100)
CC338	1 (12.5)	0
Streptolancidin B		
CC	**Iceland**	**Kenya**
CC702	0	57 (98.3)
CC499	0	55 (100)
CC5902	0	32 (13.4)
Sing11162	0	23 (100)
CC347	0	18 (29.0)
CC5250/5947/15006	0	18 (100)
CC703	0	16 (100)
CC385	0	13 (41.9)
CC1264	0	11 (100)
CC6446/14764	0	11 (100)
Other CCs	2 (100)	62 (34.6)
Other Singletons	0	22 (100)
Streptolancidin C		
CC	**Iceland**	**Kenya**
CC236/271/320	293 (100)	4 (100)
CC138/176	122 (100)	133 (100)
CC5902	0	239 (100)
CC217	0	223 (100)
CC5339	0	138 (97.2)
CC156/162	0	131 (100)
CC180	107 (100)	6 (100)
CC852	0	78 (100)
CC289	0	69 (100)
CC499	0	53 (96.4)
Other CCs	262 (52.6)	652 (71.8)

Other Singletons	14 (100)	79 (92.9)
Streptolancidin D		
CC	**Iceland**	**Kenya**
CC701	0	161 (98.8)
CC5339	0	139 (97.9)
CC991	0	104 (100)
CC5902	0	83 (34.7)
CC439	81 (37.3)	0
CC854	0	57 (100)
CC706	0	37 (100)
CC15	36 (100)	0
CC14774	0	23 (100)
Sing11162	0	23 (100)
Other CCs	55 (25.0)	190 (42.6)
Other Singletons	12 (100)	34 (100)
Streptolancidin E		
CC	**Iceland**	**Kenya**
CC439	217 (100)	0
CC199	174 (97.2)	0
CC1146	0	99 (71.2)
CC230	3 (100)	88 (95.7)
CC5258	0	76 (98.7)
CC1381	0	49 (100)
CC344	37 (100)	1 (100)
CC705/14790	0	38 (100)
CC448	29 (100)	2 (100)
CC138/176	0	22 (16.5)
Other CCs	59 (43.4)	115 (29.0)
Other Singletons	2 (100)	37 (100)
Streptolancidin F		
CC	**Iceland**	**Kenya**
CC344	33 (89.2)	1 (100)
CC100	25 (100)	0
CC191	16 (100)	0
CC5560/6090/6103	0	14 (100)
CC433	9 (14.8)	0
CC5292	0	7 (100)
CC717	4 (100)	0
CC97	3 (3.4)	0
CC346	2 (100)	0
CC113	2 (9.5)	0

Other CCs	0	2 (2.2)
Other Singletons	1 (100)	1 (100)
Streptolancidin G		
CC	**Iceland**	**Kenya**
CC1146	0	134 (96.4)
CC852	0	78 (100)
CC433	61 (100)	0
CC392	47 (100)	0
CC5329	0	37 (97.4)
CC393	20 (100)	9 (100)
CC66	18 (94.7)	0
CC30	17 (28.3)	0
CC2755	16 (100)	0
CC315	13 (86.7)	0
Other CCs	48 (63.2)	12 (4.0)
Other Singletons	12 (100)	2 (100)

Note: The 10 CCs with the biggest contribution to the frequency of each streptolancidin are shown. Other CCs were pooled to the 'Other' categories.

9.2.4 Bacteriocin association with serotypes in the Icelandic and Kenyan datasets

Table 9.6: The association of bacteriocin clusters by pneumococcal serotype.

Streptococcin A			
Iceland		**Kenya**	
Significant in IPD, OM, LRTI		**Significant in IPD**	
Serotype	**n (%)**	**Serotype**	**n (%)**
19F	322 (97.3)	1	223 (99.6)
6A	163 (100)	19F	211 (92.5)
23F	151 (83.9)	6A	194 (94.2)
6B	121 (99.2)	19A	153 (100)
3	107 (98.2)	35B	139 (100)
11A	93 (100)	15A	136 (100)
14	81 (90.0)	15BC	130 (89.0)
22F	61 (100)	11A	116 (99.1)
23B	47 (100)	13	97 (99.0)
23A	45 (88.2)	14	95 (72.5)
Other serotypes	346 (61.3)	Other serotypes	1065 (68.9)
Streptococcin D			
Iceland		**Kenya**	
Not significant		**Significant in IPD**	
Serotype	**n (%)**	**Serotype**	**n (%)**
-	-	14	70 (53.4)
-	-	nontypable	15 (46.9)
Streptococcin E			
Iceland		**Kenya**	
Significant in IPD, OM, LRTI		**Significant in IPD**	
Serotype	**n (%)**	**Serotype**	**n (%)**
19F	328 (99.1)	19F	228 (100)
23F	180 (100)	1	224 (100)
6A	163 (100)	6A	206 (100)
19A	145 (100)	19A	153 (100)
6B	122 (100)	15BC	146 (100)
3	109 (100)	35B	139 (100)
11A	93 (100)	15A	136 (100)
15BC	93 (100)	14	131 (100)
14	90 (100)	6E(6Bii)	131 (100)
22F	61 (100)	23F	119 (100)
Other serotypes	456 (86.9)	Other serotypes	1502 (97.2)

Streptocyclicin			
Iceland		**Kenya**	
Significant in carriage		**Significant in carriage**	
Serotype	**n (%)**	**Serotype**	**n (%)**
23F	174 (96.7)	19A	147 (96.1)
19A	127 (87.6)	15A	126 (92.6)
15BC	93 (100)	6A	125 (60.7)
6A	69 (42.3)	13	98 (100)
14	66 (73.3)	11A	94 (80.3)
23A	50 (98.0)	16F	93 (96.9)
23B	43 (91.5)	23B	90 (100)
nontypable	36 (51.4)	34	85 (95.5)
9V	32 (100)	10A	82 (100)
16F	26 (100)	5	69 (100)
Other serotypes	146 (25.0)	Other serotypes	514 (29.8)
Streptolancidin A			
Iceland		**Kenya**	
Significant in carriage		**Not significant**	
Serotype	**n (%)**	**Serotype**	**n (%)**
6B	120 (98.4)	-	-
nontypable	29 (41.4)	-	-
6A	3 (1.8)	-	-
Streptolancidin B			
Iceland		**Kenya**	
Not significant		**Significant in carriage**	
Serotype	**n (%)**	**Serotype**	**n (%)**
-	-	6A	64 (31.1)
-	-	20	53 (98.1)
-	-	15BC	47 (32.2)
-	-	11A	33 (28.2)
-	-	16F	32 (33.3)
-	-	19F	23 (10.1)
-	-	6E(6Bii)	20 (15.3)
-	-	15A	12 (8.8)
-	-	24F	12 (100)
-	-	6C	10 (100)
-	-	Other serotypes	32 (5.1)
Streptolancidin C			
Iceland		**Kenya**	
Significant in OM, LRTI		**Significant in IPD**	
Serotype	**n (%)**	**Serotype**	**n (%)**

19F	325 (98.2)	1	224 (100)
6B	122 (100)	19F	180 (78.9)
3	109 (100)	19A	149 (97.4)
23A	40 (78.4)	6A	119 (57.8)
6A	29 (17.8)	23F	119 (100)
nontypable	29 (41.4)	11A	112 (95.7)
6E	24 (92.3)	15BC	88 (60.3)
14	21 (23.3)	23B	87 (96.7)
38	20 (100)	10A	81 (98.8)
7F	16 (100)	5	69 (100)
Other Serotypes	63 (10.8)	Other serotypes	577 (37.9)
Streptolancidin D			
Iceland		**Kenya**	
Significant in OM		**Significant in carriage**	
Serotype	**n (%)**	**Serotype**	**n (%)**
23F	83 (46.1)	19F	150 (65.8)
6A	22 (13.5)	15A	124 (91.2)
14	21 (23.3)	15BC	113 (77.4)
19F	17 (5.1)	13	97 (99.0)
35B	16 (48.5)	11A	75 (64.1)
19A	12 (8.3)	6E(6Bii)	57 (43.5)
6C	8 (27.6)	9V	51 (81.0)
18C	2 (9.1)	21	25 (40.3)
31	1 (33.3)	6A	22 (10.7)
6E	1 (3.8)	6B	21 (80.8)
Other serotypes	1 (12.5)	Other serotypes	116 (13.5)
Streptolancidin E			
Iceland		**Kenya**	
Significant in carriage		**Significant in carriage**	
Serotype	**n (%)**	**Serotype**	**n (%)**
23F	127 (70.6)	35B	96 (69.1)
19A	126 (86.9)	34	76 (85.4)
nontypable	67 (95.7)	16F	57 (59.4)
15BC	54 (58.1)	3	56 (62.2)
23A	50 (98.0)	18C	48 (100)
23B	42 (89.4)	14	34 (26.0)
18C	17 (77.3)	17F	20 (80.0)
9N	14 (77.8)	21	20 (32.3)
21	4 (13.8)	35F	17 (94.4)
1	3 (60.0)	15BC	14 (9.6)
Other serotypes	17 (2.6)	Other serotypes	89 (9.4)

Streptolancidin F			
Iceland		**Kenya**	
Significant in IPD, carriage (vs OM)		**Not significant**	
Serotype	**n (%)**	**Serotype**	**n (%)**
NT	34 (48.6)	-	-
33F	29 (100)	-	-
7F	16 (100)	-	-
22F	9 (14.8)	-	-
10A	3 (60.0)	-	-
19A	2 (1.4)	-	-
18C	2 (9.1)	-	-
Streptolancidin G			
Iceland		**Kenya**	
Significant in IPD, carriage (vs OM)		**Significant in carriage**	
Serotype	**n (%)**	**Serotype**	**n (%)**
22F	60 (98.4)	35B	134 (96.4)
23F	48 (26.7)	10A	82 (100)
35B	30 (90.9)	29	19 (67.9)
38	20 (100)	6A	12 (5.8)
6C	20 (69.0)	38	8 (34.8)
9N	18 (100)	6E(6Bii)	4 (3.1)
19F	17 (5.1)	10F	3 (33.3)
6A	14 (8.6)	15BC	2 (1.4)
19A	12 (8.3)	21	2 (3.2)
4	7 (100)	34	1 (1.1)
Other serotypes	6 (10.3)	Other serotypes	5 (1.5)
Streptolancidin J			
Iceland		**Kenya**	
Significant in carriage		**Not significant**	
Serotype	**n (%)**	**Serotype**	**n (%)**
6A	147 (90.2)	-	-
19A	139 (95.9)	-	-
6B	122 (100)	-	-
3	107 (98.2)	-	-
14	69 (76.7)	-	-
15BC	62 (66.7)	-	-
22F	61 (100)	-	-
23F	50 (27.8)	-	-
19F	34 (10.3)	-	-
9V	31 (96.9)	-	-
Other serotypes	183 (35.6)	-	-

Streptolassin			
Iceland		**Kenya**	
Significant in carriage (v OM)		**Significant in IPD**	
Serotype	**n (%)**	**Serotype**	**n (%)**
23F	48 (26.7)	5	69 (100)
-	-	37	3 (100)
-	-	7F	1 (50.0)
-	-	1	1 (0.4)
-	-	38	1 (4.3)
-	-	6A	1 (0.5)
-	-	8	1 (4.8)

Note: Up to 10 of the most common serotypes associated with each bacteriocin are listed separately, and the remainder were pooled as 'Other'. Bacteriocins that did not exhibit significantly altered prevalence in any subset of the data were excluded from this table. IPD, invasive pneumococcal disease; LRTI, lower respiratory tract infection; OM, otitis media.

9.2.5 Bacteriocin association with clonal complexes in subsets of Icelandic and Kenyan genomic datasets

Table 9.7: The association of bacteriocin clusters with clonal complexes (CCs) in the Icelandic dataset by vaccination time period (pre/post PCV), carriage, and disease.

Number of pneumococci harbouring each bacteriocin, stratified by CC n (% of CC representatives in each subset with the bacteriocin)						
Streptococcin A						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC236/271/320	201 (97.6)	85 (97.7)	53 (100)	6 (100)	72 (93.5)	155 (98.7)
CC439	86 (80.4)	98 (89.1)	106 (83.5)	16 (94.1)	21 (95.5)	41 (80.4)
CC138/176	79 (100)	42 (97.7)	87 (100)	5 (100)	11 (91.7)	18 (100)
CC180	64 (100)	42 (97.7)	55 (100)	9 (100)	21 (100)	21 (95.5)
CC62	37 (97.4)	56 (100)	62 (100)	5 (100)	13 (92.9)	13 (100)
CC490	40 (100)	34 (100)	46 (100)	5 (100)	9 (100)	14 (100)
CC433	13 (100)	48 (100)	32 (100)	13 (100)	11 (100)	5 (100)
CC30	34 (100)	26 (100)	40 (100)	2 (100)	10 (100)	8 (100)
CC97	30 (88.2)	30 (56.6)	32 (66.7)	4 (66.7)	4 (44.4)	20 (83.3)
CC124	36 (81.8)	17 (94.4)	23 (79.3)	12 (100)	5 (71.4)	13 (92.9)
Other CCs	213 (91.8)	211 (96.8)	200 (91.7)	75 (93.8)	65 (98.5)	84 (97.7)
Other Singletons	4 (100)	11 (100)	10 (100)	1 (100)	1 (100)	3 (100)
Streptococcin E						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC236/271/320	203 (98.5)	87 (100)	53 (100)	6 (100)	75 (97.4)	156 (99.4)
CC439	107 (100)	110 (100)	127 (100)	17 (100)	22 (100)	51 (100)
CC199	99 (100)	80 (100)	110 (100)	13 (100)	11 (100)	45 (100)
CC138/176	79 (100)	43 (100)	87 (100)	5 (100)	12 (100)	18 (100)
CC180	64 (100)	43 (100)	55 (100)	9 (100)	21 (100)	22 (100)
CC62	38 (100)	56 (100)	62 (100)	5 (100)	14 (100)	13 (100)
CC97	34 (100)	53 (100)	48 (100)	6 (100)	9 (100)	24 (100)
CC490	40 (100)	34 (100)	46 (100)	5 (100)	9 (100)	14 (100)
CC124	44 (100)	18 (100)	29 (100)	12 (100)	7 (100)	14 (100)
CC433	13 (100)	48 (100)	32 (100)	13 (100)	11 (100)	5 (100)
Other CCs	279 (99.6)	253 (100)	262 (99.6)	91 (100)	83 (100)	96 (100)
Other Singletons	4 (100)	11 (100)	10 (100)	1 (100)	1 (100)	3 (100)
Streptocyclicin						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC439	107 (100)	110 (100)	127 (100)	17 (100)	22 (100)	51 (100)

CC199	99 (100)	80 (100)	110 (100)	13 (100)	11 (100)	45 (100)
CC97	34 (100)	53 (100)	48 (100)	6 (100)	9 (100)	24 (100)
CC124	44 (100)	18 (100)	29 (100)	12 (100)	7 (100)	14 (100)
CC392	24 (88.9)	19 (95.0)	31 (96.9)	1 (33.3)	5 (100)	6 (85.7)
CC156/162	36 (92.3)	7 (100)	14 (82.4)	11 (100)	10 (100)	8 (100)
CC30	24 (70.6)	19 (73.1)	32 (80.0)	1 (50.0)	6 (60.0)	4 (50.0)
CC1262	6 (100)	29 (100)	21 (100)	2 (100)	7 (100)	5 (100)
CC344	14 (87.5)	21 (100)	31 (93.9)	0	4 (100)	0
CC193	6 (100)	22 (100)	16 (100)	2 (100)	5 (100)	5 (100)
Other CCs	33 (71.7)	55 (53.9)	50 (61.7)	16 (55.2)	11 (50.0)	11 (68.8)
Other Singletons	0	2 (100)	1 (100)	0	0	1 (100)
Streptolancidin A						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC138/176	79 (100)	43 (100)	87 (100)	5 (100)	12 (100)	18 (100)
CC448	15 (100)	14 (100)	27 (100)	0	2 (100)	0
CC338	1 (20.0)	0	0	1 (50.0)	0	0
Streptolancidin C						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC236/271/320	206 (100)	87 (100)	53 (100)	6 (100)	77 (100)	157 (100)
CC138/176	79 (100)	43 (100)	87 (100)	5 (100)	12 (100)	18 (100)
CC180	64 (100)	43 (100)	55 (100)	9 (100)	21 (100)	22 (100)
CC439	10 (9.3)	32 (29.1)	31 (24.4)	2 (11.8)	3 (13.6)	6 (11.8)
CC15	32 (100)	4 (100)	14 (100)	9 (100)	5 (100)	8 (100)
CC30	15 (44.1)	19 (73.1)	20 (50.0)	2 (100)	6 (60.0)	6 (75.0)
CC448	15 (100)	14 (100)	27 (100)	0	2 (100)	0
CC90	15 (100)	7 (100)	8 (100)	1 (100)	7 (100)	6 (100)
CC393	16 (100)	4 (100)	16 (100)	2 (100)	0	2 (100)
CC191	11 (100)	5 (100)	0	14 (100)	1 (100)	1 (100)
Other CCs	29 (100)	34 (100)	26 (100)	11 (100)	13 (100)	13 (100)
Other Singletons	3 (100)	11 (100)	10 (100)	0	1 (100)	3 (100)
Streptolancidin D						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC439	51 (47.7)	30 (27.3)	34 (26.8)	9 (52.9)	10 (45.5)	28 (54.9)
CC15	32 (100)	4 (100)	14 (100)	9 (100)	5 (100)	8 (100)
CC30	5 (14.7)	12 (46.2)	12 (30.0)	1 (50.0)	2 (20.0)	2 (25.0)
CC2755	8 (100)	8 (100)	4 (100)	3 (100)	5 (100)	4 (100)
CC177	6 (100)	8 (100)	5 (100)	0	3 (100)	6 (100)
Sing1801	2 (100)	10 (100)	9 (100)	0	1 (100)	2 (100)
CC473	1 (100)	1 (100)	1 (100)	0	0	1 (100)
CC102	2 (100)	0	2 (100)	0	0	0
CC338	2 (40.0)	0	0	0	1 (50.0)	1 (50.0)

CC1766	1 (100)	0	0	0	1 (100)	0
Other CCs	1 (100)	0	0	0	0	1 (25.0)
Other Singletons	0	0	0	0	0	0
Streptolancidin E						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC439	107 (100)	110 (100)	127 (100)	17 (100)	22 (100)	51 (100)
CC199	97 (98.0)	77 (96.2)	108 (98.2)	13 (100)	10 (90.9)	43 (95.6)
CC344	16 (100)	21 (100)	33 (100)	0	4 (100)	0
CC448	15 (100)	14 (100)	27 (100)	0	2 (100)	0
CC113	12 (85.7)	6 (85.7)	12 (80.0)	2 (100)	2 (100)	2 (100)
CC66	8 (88.9)	7 (70.0)	9 (90.0)	4 (66.7)	1 (50.0)	1 (100)
CC3017	6 (100)	2 (100)	2 (100)	3 (100)	3 (100)	0
CC432	0	4 (66.7)	3 (60.0)	0	0	1 (100)
CC306	1 (50.0)	2 (66.7)	0	3 (60.0)	0	0
CC230	1 (100)	2 (100)	0	1 (100)	0	2 (100)
Other CCs	7 (17.5)	4 (13.8)	3 (7.3)	3 (60.0)	5 (33.3)	0
Other Singletons	1 (100)	1 (100)	1 (100)	1 (100)	0	0
Streptolancidin F						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC344	14 (87.5)	19 (90.5)	29 (87.9)	0	4 (100)	0
CC100	15 (100)	10 (100)	10 (100)	7 (100)	4 (100)	4 (100)
CC191	11 (100)	5 (100)	0	14 (100)	1 (100)	1 (100)
CC433	7 (53.8)	2 (4.2)	2 (6.2)	2 (15.4)	5 (45.5)	0
CC717	1 (100)	3 (100)	1 (100)	0	0	3 (100)
CC97	2 (5.9)	1 (1.9)	3 (6.2)	0	0	0
CC113	2 (14.3)	0	2 (13.3)	0	0	0
CC346	0	2 (100)	1 (100)	0	0	1 (100)
Sing10346	0	1 (100)	1 (100)	0	0	0
Streptolancidin G						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC433	13 (100)	48 (100)	32 (100)	13 (100)	11 (100)	5 (100)
CC392	27 (100)	20 (100)	32 (100)	3 (100)	5 (100)	7 (100)
CC393	16 (100)	4 (100)	16 (100)	2 (100)	0	2 (100)
CC66	9 (100)	9 (90.0)	10 (100)	5 (83.3)	2 (100)	1 (100)
CC30	10 (29.4)	7 (26.9)	8 (20.0)	1 (50.0)	4 (40.0)	4 (50.0)
CC2755	8 (100)	8 (100)	4 (100)	3 (100)	5 (100)	4 (100)
CC315	4 (66.7)	9 (100)	4 (80.0)	1 (100)	2 (100)	6 (85.7)
CC15	12 (37.5)	1 (25.0)	8 (57.1)	0	3 (60.0)	2 (25.0)
CC198	0	13 (100)	11 (100)	0	1 (100)	1 (100)
Sing1801	2 (100)	10 (100)	9 (100)	0	1 (100)	2 (100)
Other CCs	8 (100)	14 (73.7)	7 (63.6)	9 (100)	2 (100)	4 (80.0)

Other Singletons	0	0	0	0	0	0
Streptolancidin J						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC199	86 (86.9)	67 (83.8)	97 (88.2)	11 (84.6)	7 (63.6)	38 (84.4)
CC138/176	79 (100)	43 (100)	87 (100)	5 (100)	12 (100)	18 (100)
CC180	64 (100)	43 (100)	55 (100)	9 (100)	21 (100)	22 (100)
CC97	34 (100)	52 (98.1)	48 (100)	6 (100)	8 (88.9)	24 (100)
CC490	39 (97.5)	32 (94.1)	44 (95.7)	5 (100)	9 (100)	13 (92.9)
CC124	44 (100)	18 (100)	29 (100)	12 (100)	7 (100)	14 (100)
CC433	13 (100)	48 (100)	32 (100)	13 (100)	11 (100)	5 (100)
CC30	34 (100)	26 (100)	40 (100)	2 (100)	10 (100)	8 (100)
CC392	27 (100)	20 (100)	32 (100)	3 (100)	5 (100)	7 (100)
CC156/162	38 (97.4)	7 (100)	17 (100)	10 (90.9)	10 (100)	8 (100)
Other CCs	66 (60.6)	113 (57.4)	91 (52.3)	24 (68.6)	29 (63.0)	35 (68.6)
Other Singletons	2 (100)	10 (100)	9 (100)	0	1 (100)	2 (100)
Streptolassin						
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**	**LRTI**	**OM**
CC392	27 (100)	20 (100)	32 (100)	3 (100)	5 (100)	7 (100)
CC433	0	1 (2.1)	1 (3.1)	0	0	0

Note: The 10 most common clonal complexes in which the bacteriocin was found are listed separately, and the remainder were pooled as 'Other'. Only bacteriocins with significant differences in prevalence in Figure 2B and 2C are included in this table. IPD, invasive pneumococcal disease; LRTI, lower respiratory tract infection; OM, otitis media.

Table 9.8: The association of bacteriocin clusters with clonal complexes in the Kenyan dataset, by vaccination time period (pre/post PCV), carriage, and invasive disease.

Number of pneumococci harbouring each bacteriocin, stratified by CC n (% of CC representatives in each subset with the bacteriocin)				
Streptococcin A				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC5902	101 (100)	135 (97.8)	219 (98.6)	17 (100)
CC217	199 (99.5)	23 (100)	16 (100)	206 (99.5)
CC701	67 (93.1)	80 (87.9)	130 (89.7)	17 (94.4)
CC1146	53 (100)	86 (100)	126 (100)	13 (100)
CC5339	106 (97.2)	33 (100)	122 (97.6)	17 (100)
CC156/162	36 (100)	95 (100)	97 (100)	34 (100)
CC138/176	66 (98.5)	64 (97.0)	97 (98.0)	33 (97.1)
CC991	24 (100)	80 (100)	95 (100)	9 (100)
CC852	26 (96.3)	51 (100)	66 (98.5)	11 (100)
CC63	56 (100)	14 (100)	37 (100)	33 (100)
Other CCs	576 (81.5)	472 (82.8)	801 (81.2)	247 (85.2)
Other Singletons	42 (100)	74 (98.7)	100 (99.0)	16 (100)
Streptococcin D				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC63	56 (100)	14 (100)	37 (100)	33 (100)
CC13215	0	14 (100)	14 (100)	0
Sing14766	1 (100)	0	0	1 (100)
Streptococcin E				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC5902	101 (100)	138 (100)	222 (100)	17 (100)
CC217	200 (100)	23 (100)	16 (100)	207 (100)
CC701	72 (100)	91 (100)	145 (100)	18 (100)
CC5339	109 (100)	33 (100)	125 (100)	17 (100)
CC1146	53 (100)	86 (100)	126 (100)	13 (100)
CC138/176	67 (100)	66 (100)	99 (100)	34 (100)
CC156/162	36 (100)	95 (100)	97 (100)	34 (100)
CC991	24 (100)	80 (100)	95 (100)	9 (100)
CC230	50 (100)	42 (100)	60 (100)	32 (100)
CC852	27 (100)	51 (100)	67 (100)	11 (100)
Other CCs	857 (98.6)	671 (96.1)	1166 (96.8)	362 (99.7)
Other Singletons	49 (100)	94 (100)	127 (100)	16 (100)
Streptocyclicin				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC5902	78 (77.2)	89 (64.5)	152 (68.5)	15 (88.2)

CC701	72 (100)	91 (100)	145 (100)	18 (100)
CC156/162	36 (100)	95 (100)	97 (100)	34 (100)
CC991	24 (100)	80 (100)	95 (100)	9 (100)
CC230	48 (96.0)	41 (97.6)	58 (96.7)	31 (96.9)
CC852	27 (100)	51 (100)	67 (100)	11 (100)
CC5258	18 (100)	59 (100)	72 (100)	5 (100)
CC289	64 (100)	5 (100)	3 (100)	66 (100)
CC914	43 (97.7)	17 (100)	44 (97.8)	16 (100)
CC702	18 (100)	40 (100)	56 (100)	2 (100)
Other CCs	213 (45.6)	257 (71.2)	395 (58.1)	75 (50.7)
Other Singletons	23 (100)	34 (100)	51 (100)	6 (100)
Streptolancidin B				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC702	17 (94.4)	40 (100)	55 (98.2)	2 (100)
CC499	36 (100)	19 (100)	44 (100)	11 (100)
CC5902	16 (15.8)	16 (11.6)	32 (14.4)	0
Sing11162	0	23 (100)	20 (100)	3 (100)
CC347	16 (28.1)	2 (40.0)	12 (25.0)	6 (42.9)
CC5250/5947/15006	10 (100)	8 (100)	16 (100)	2 (100)
CC703	9 (100)	7 (100)	14 (100)	2 (100)
CC385	10 (41.7)	3 (42.9)	6 (37.5)	7 (46.7)
CC1264	3 (100)	8 (100)	11 (100)	0
CC6446/14764	2 (100)	9 (100)	9 (100)	2 (100)
Other CCs	40 (40.0)	22 (27.8)	55 (36.7)	7 (24.1)
Other Singletons	10 (100)	12 (100)	19 (100)	3 (100)
Streptolancidin C				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC5902	101 (100)	138 (100)	222 (100)	17 (100)
CC217	200 (100)	23 (100)	16 (100)	207 (100)
CC5339	106 (97.2)	32 (97.0)	121 (96.8)	17 (100)
CC138/176	67 (100)	66 (100)	99 (100)	34 (100)
CC156/162	36 (100)	95 (100)	97 (100)	34 (100)
CC852	27 (100)	51 (100)	67 (100)	11 (100)
CC289	64 (100)	5 (100)	3 (100)	66 (100)
CC499	36 (100)	17 (89.5)	42 (95.5)	11 (100)
CC7689	36 (100)	3 (100)	30 (100)	9 (100)
CC338	16 (100)	21 (100)	29 (100)	8 (100)
Other CCs	285 (66.6)	301 (72.7)	485 (70.0)	101 (67.8)
Other Singletons	24 (88.9)	55 (94.8)	70 (92.1)	9 (100)
Streptolancidin D				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**

CC701	71 (98.6)	90 (98.9)	143 (98.6)	18 (100)
CC5339	107 (98.2)	32 (97.0)	122 (97.6)	17 (100)
CC991	24 (100)	80 (100)	95 (100)	9 (100)
CC5902	45 (44.6)	38 (27.5)	75 (33.8)	8 (47.1)
CC854	52 (100)	5 (100)	38 (100)	19 (100)
CC706	30 (100)	7 (100)	27 (100)	10 (100)
Sing11162	0	23 (100)	20 (100)	3 (100)
CC14774	6 (100)	17 (100)	21 (100)	2 (100)
CC4368	17 (94.4)	5 (100)	17 (94.4)	5 (100)
CC5938	9 (100)	9 (90.0)	16 (94.1)	2 (100)
Other CCs	68 (43.0)	82 (46.6)	131 (51.8)	19 (23.5)
Other Singletons	14 (100)	20 (100)	29 (100)	5 (100)
Streptolancidin E				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC1146	45 (84.9)	54 (62.8)	87 (69.0)	12 (92.3)
CC230	48 (96.0)	40 (95.2)	57 (95.0)	31 (96.9)
CC5258	18 (100)	58 (98.3)	71 (98.6)	5 (100)
CC1381	41 (100)	8 (100)	31 (100)	18 (100)
CC705/14790	8 (100)	30 (100)	33 (100)	5 (100)
CC138/176	5 (7.5)	17 (25.8)	17 (17.2)	5 (14.7)
CC5349	6 (100)	11 (100)	17 (100)	0
CC14858	5 (100)	11 (73.3)	14 (77.8)	2 (100)
CC14892	1 (16.7)	13 (81.2)	14 (63.6)	0
Sing14868	5 (100)	9 (100)	12 (100)	2 (100)
Other CCs	29 (20.3)	42 (21.2)	55 (18.2)	16 (42.1)
Other Singletons	4 (100)	19 (100)	22 (100)	1 (100)
Streptolancidin G				
CC	**Pre-PCV**	**Post-PCV**	**Carriage**	**IPD**
CC1146	48 (90.6)	86 (100)	121 (96.0)	13 (100)
CC852	27 (100)	51 (100)	67 (100)	11 (100)
CC5329	14 (93.3)	23 (100)	33 (97.1)	4 (100)
CC393	4 (100)	5 (100)	5 (100)	4 (100)
CC5902	0	3 (2.2)	3 (1.4)	0
CC5796	2 (15.4)	0	2 (14.3)	0
CC14774	0	2 (11.8)	2 (9.5)	0
CC909	1 (100)	1 (100)	2 (100)	0
Sing14823	1 (100)	0	1 (100)	0
CC473	1 (25.0)	0	1 (33.3)	0
Other CCs	1 (50.0)	1 (7.1)	2 (15.4)	0
Other Singletons	0	1 (100)	1 (100)	0
Streptolassin				

CC	Pre-PCV	Post-PCV	Carriage	IPD
CC289	64 (100)	5 (100)	3 (100)	66 (100)
CC13854/15057	0	3 (100)	3 (100)	0
CC404	1 (100)	0	1 (100)	0
CC5936/14865	1 (100)	0	1 (100)	0
Sing5359	1 (100)	0	1 (100)	0
Sing14840	1 (100)	0	1 (100)	0
CC5068	1 (33.3)	0	0	1 (50.0)

Note: The 10 most common clonal complexes in which the bacteriocin was found are listed separately, and the remainder were pooled as 'Other'. Only bacteriocins with significant differences in prevalence in Figure 2B and 2C are included in this table. IPD, invasive pneumococcal disease.

9.2.6 Bacteriocin repertoires within clonal complexes in the Icelandic and Kenyan datasets

Table 9.9: Clonal complexes (CCs) in the Icelandic dataset with multiple bacteriocin repertoires, including any constituent sequence types (STs) with mixed repertoires.

CC	Variable bacteriocins (CC)	Mixed STs	Variable bacteriocins (ST)
236/271/320	Streptococcin A, Streptococcin E	271	Streptococcin E
		1968	Streptococcin A
439	Streptococcin A, Streptolancidin C, Streptolancidin D	311	Streptococcin A
		507	Streptococcin A
		442	Streptococcin A
		190	Streptococcin A
199	Streptolancidin J, Streptosactin	199	Streptolancidin J, Streptosactin
138/176	Streptococcin A	176	Streptococcin A
180	Streptococcin A	180	Streptococcin A
62	Streptococcin A, Streptolancidin J	62	Streptolancidin J
97	Streptococcin A, Streptolancidin F, Streptolancidin J	1635	Streptolancidin J
490	Streptolancidin J	2221	Streptolancidin J
124	Streptococcin A	124	Streptococcin A
433	Streptocyclicin, Streptolancidin F, Streptolassin	433	Streptolancidin F
30	Streptocyclicin, Streptolancidin C, Streptolancidin D, Streptolancidin E, Streptolancidin G	30	Streptolancidin E
392	Streptocyclicin	440	Streptocyclicin
156/162	Streptocyclicin, Streptolancidin J	162	Streptolancidin J
344	Streptocyclicin, Streptolancidin F, Streptolancidin K	10371	Streptocyclicin, Streptolancidin F
		344	Streptolancidin F, Streptolancidin K
15	Streptolancidin G	None	NA
1262	Streptolancidin J	1262	Streptolancidin J
193	Streptolancidin J	1877	Streptolancidin J

113	Streptococcin A, Streptolancidin F	113	Streptococcin A, Streptolancidin F
113	Streptococcin A, Streptolancidin F	110	Streptococcin A
393	Streptococcin A	None	NA
66	Streptolancidin G	None	NA
315	Streptolancidin G, Streptolancidin J	386	Streptolancidin J
6524	Streptolancidin J	6524	Streptolancidin J
338	Streptolancidin A, Streptolancidin D	None	NA
63	Streptolancidin D	None	NA
432	Streptolancidin G	432	Streptolancidin G
205	Streptolancidin J	205	Streptolancidin J
306	Streptococcin A	None	NA
230	Streptolancidin J	None	NA
473	Streptolancidin G	None	NA

Note: Table includes bacteriocins that varied within each CC or ST.

Table 9.10: Clonal complexes (CCs) in the Kenyan dataset with multiple bacteriocin repertoires, including any constituent sequence types (STs) with mixed repertoires.

CC	Variable bacteriocins (CC)	Mixed STs	Variable bacteriocins (ST)
5902	Streptococcin A, Streptocyclicin, Streptolancidin B, Streptolancidin D, Streptolancidin E, Streptolancidin G, Streptolancidin J	5902	Streptocyclicin, Streptolancidin J
		5370	Streptococcin A, Streptolancidin E
		840	Streptolancidin D
		2052	Streptolancidin E
		15056	Streptolancidin G
217	Streptococcin A, Streptolancidin E	613	Streptococcin A, Streptolancidin E
701	Streptococcin A, Streptolancidin D, Streptolancidin J	701	Streptococcin A, Streptolancidin D
		5340	Streptococcin A
5339	Streptococcin A, Streptocyclicin, Streptolancidin C, Streptolancidin D, Streptolancidin J	5339	Streptococcin A, Streptolancidin D
		844	Streptococcin A, Streptocyclicin, Streptolancidin C, Streptolancidin D
		5367	Streptolancidin J
		5268	Streptococcin A, Streptolancidin C
1146	Streptolancidin E, Streptolancidin G, Streptolancidin J	5952	Streptolancidin E
		5396	Streptolancidin G
138/176	Streptococcin A, Streptocyclicin, Streptolancidin A, Streptolancidin E, Streptolancidin J	848	Streptococcin A, Streptolancidin E, Streptolancidin J
156/162	Streptolancidin E, Streptolancidin K	847	Streptolancidin E, Streptolancidin K
230	Streptocyclicin, Streptolancidin D, Streptolancidin E, Streptolancidin F, Streptolancidin J	230	Streptolancidin D, Streptolancidin J
		700	Streptocyclicin
		4351	Streptolancidin E
852	Streptococcin A	852	Streptococcin A
5258	Streptococcin A	5258	Streptococcin A
63	Streptolancidin C, Streptolancidin J	842	Streptolancidin J
		2716	Streptolancidin C

347	Streptococcin A, Streptocyclicin, Streptolancidin B, Streptolancidin C, Streptolancidin J	6088	Streptococcin A, Streptolancidin J
		2715	Streptolancidin B, Streptolancidin J
		5769	Streptococcin A, Streptolancidin J
		6095	Streptolancidin J
		14817	Streptococcin A, Streptolancidin J
914	Streptolancidin B	None	NA
7053	Streptococcin A, Streptolancidin C, Streptolancidin J	5368	Streptolancidin J
702	Streptolancidin B, Streptolancidin C	702	Streptolancidin B, Streptolancidin C
854	Streptococcin A, Streptocyclicin, Streptolancidin J	854	Streptococcin A, Streptocyclicin, Streptolancidin J
499	Streptolancidin C, Streptolancidin J	499	Streptolancidin C
		5907	Streptolancidin J
5329	Streptocyclicin, Streptolancidin C, Streptolancidin G, Streptolancidin J, Streptolancidin K	5329	Streptolancidin C, Streptolancidin J
338	Streptolancidin D, Streptolancidin E, Streptolancidin J	172	Streptolancidin E
		2054	Streptolancidin D, Streptolancidin J
3460	Streptococcin A, Streptocyclicin, Streptolancidin B	14886	Streptococcin A, Streptocyclicin
		3460	Streptococcin A
385	Streptococcin A, Streptolancidin B, Streptolancidin J	6097	Streptolancidin B
		2713	Streptococcin A
		3207	Streptolancidin J
989	Streptolancidin D	989	Streptolancidin D
14774	Streptolancidin B, Streptolancidin G, Streptolancidin J	6092	Streptolancidin B, Streptolancidin G
		14774	Streptolancidin J
14892	Streptococcin A, Streptolancidin C, Streptolancidin D, Streptolancidin E	None	NA

2386	Streptolancidin C, Streptolancidin D	5331	Streptolancidin D
14858	Streptocyclicin, Streptolancidin E	14858	Streptolancidin E
		14910	Streptocyclicin
4894	Streptococcin A	4894	Streptococcin A
5938	Streptolancidin D	None	NA
14930/15024	Streptocyclicin, Streptolancidin C, Streptolancidin J	14930	Streptocyclicin, Streptolancidin C, Streptolancidin J
5349	Streptococcin A	None	NA
5294	Streptococcin A, Streptocyclicin	5294	Streptococcin A, Streptocyclicin
703	Streptocyclicin, Streptolancidin C, Streptolancidin J	703	Streptolancidin C, Streptolancidin J
Sing5373	Streptolancidin C	5373	Streptolancidin C
5372/15025	Streptolancidin G	5372	Streptolancidin G
Sing14868	Streptococcin A	14868	Streptococcin A
5796	Streptococcin A, Streptolancidin G, Streptolancidin J	5796	Streptococcin A
5560/6090/6103	Streptococcin A	6103	Streptococcin A
193	Streptolancidin C, Streptolancidin J	None	NA
1766	Streptococcin A	None	NA
1264	Streptolancidin D	1264	Streptolancidin D
3735	Streptocyclicin	None	NA
14846/14876	Streptococcin A, Streptolancidin D	14846	Streptococcin A, Streptolancidin D
5798/5879	Streptococcin A, Streptolancidin B	5798	Streptococcin A, Streptolancidin B
5321/14966	Streptolancidin C	None	NA
393	Streptolancidin J	None	NA
547	Streptolancidin D, Streptolancidin J	None	NA
3518	Streptococcin A, Streptolancidin B	None	NA
3983	Streptococcin A	3983	Streptococcin A
5266	Streptolancidin J	None	NA
5901	Streptococcin A	14976	Streptococcin A
5398/14814	Streptolancidin C	None	NA
473	Streptolancidin G	None	NA

5839/14990	Streptolancidin J	None	NA
5068	Streptolancidin J, Streptolassin	None	NA
Sing5376	Streptolancidin J	5376	Streptolancidin J
849/5343/5351	Streptococcin A	None	NA

Note: Table includes bacteriocins that varied within each CC or ST.

9.3 Chapter 5 Appendices

9.3.1 Species breakdown of the non-pneumococcal streptococcal genomic dataset

Table 9.11: Composition of the non-pneumococcal Streptococcus genomic dataset.

Species	Number of genomes
Streptococcus agalactiae	180
Streptococcus pyogenes	180
Streptococcus suis	180
Streptococcus mutans	177
Streptococcus dysgalactiae	153
Streptococcus salivarius	102
Streptococcus oralis	95
Streptococcus equi	85
Streptococcus mitis	84
Streptococcus pseudopneumoniae	77
Streptococcus thermophilus	65
Streptococcus anginosus	57
Streptococcus sanguinis	53
Streptococcus parasanguinis	51
Streptococcus gordonii	41
Streptococcus equinus	35
Streptococcus intermedius	28
Streptococcus iniae	23
Streptococcus canis	19
Streptococcus cristatus	19
Streptococcus gallolyticus	13
Streptococcus lutetiensis	12
Streptococcus pasteurianus	12
Streptococcus constellatus	10
Streptococcus sobrinus	8
Streptococcus infantis	7
Streptococcus macedonicus	7
Streptococcus australis	6
Streptococcus vestibularis	6
Streptococcus porcinus	5
Streptococcus pseudoporcinus	4
Streptococcus hyovaginalis	3

Streptococcus ratti	2
Streptococcus infantarius	2
Streptococcus downei	2
Streptococcus rubneri	2
Streptococcus alactolyticus	2
Streptococcus massiliensis	1
Streptococcus peroris	1
Streptococcus devriesei	1
Streptococcus criceti	1
Streptococcus chosunense	1
Streptococcus koreensis	1
Streptococcus periodonticum	1
Streptococcus halitosis	1
Streptococcus xiaochunlingii	1
Streptococcus orisratti	1
Streptococcus hyointestinalis	1
Streptococcus ferus	1
Streptococcus urinalis	1
Streptococcus macacae	1
Streptococcus lactarius	1
Streptococcus sinensis	1
Streptococcus pharyngis	1
Streptococcus timonensis	1

9.3.2 Streptococcin cluster contiguity in the non-pneumococcal streptococcal genomic dataset

Table 9.12 Contiguity of observed full and partial bacteriocin gene clusters among genomes of the non-pneumococcal streptococcal database.

Cluster	Category	Frequency	% of total clusters
Streptococcin A	Contiguous	106	99.07
	EOC	1	0.93
Streptococcin B	Contiguous	107	99.07
	Contiguous with Ns	1	0.93
Streptococcin C	Contiguous	140	97.90
	Non-contiguous (one contig)	2	1.40
	EOC	1	0.70
Streptococcin D	Contiguous	66	97.06
	Non-contiguous (one contig)	2	2.94
Streptococcin E	Contiguous	66	66.00
	EOC	27	27.00
	Non-contiguous (one contig)	5	5.00
	Non-contiguous (multiple contigs, not EOC-adjacent)	2	2.00

Note: Clusters were categorised according to the proximity of the constituent genes to one another and any clusters with an intergenic region >2.5kbp were categorised as non-contiguous. Bacteriocin clusters with genes on multiple contigs were categorised as 'end of contig' (EOC) if the genes were found within 2.5kbp of each other and the end of the contig, otherwise the clusters were categorised as non-contiguous (multiple contigs). Each category is shown by count and percentage of the observed clusters in the dataset. Rows shown in grey represent non-contiguous clusters, which were excluded from further analysis.

9.3.3 Streptococcin prevalence in the non-pneumococcal streptococcal genomic dataset

Table 9.13: Prevalence of each streptococcin in the pneumococcal and non-pneumococcal streptococcal datasets.

Streptococcin A	
Species	**n with streptococcin (% of species)**
S pneumoniae (Iceland)	1537 (80.4%)
S pneumoniae (Kenya)	2559 (81.0%)
Streptococcus mitis	48 (57.1%)
Streptococcus oralis	37 (38.9%)
Streptococcus pseudopneumoniae	21 (27.3%)
Streptococcus chosunense	1 (100.0%)
Streptococcin B	
Species	**n with streptococcin (% of species)**
S pneumoniae (Iceland)	1912 (100.0%)
S pneumoniae (Kenya)	3158 (100.0%)
Streptococcus pseudopneumoniae	77 (100.0%)
Streptococcus mitis	31 (36.9%)
Streptococcin C	
Species	**n with streptococcin (% of species)**
S pneumoniae (Iceland)	1896 (99.2%)
S pneumoniae (Kenya)	3159 (100.0%)
Streptococcus mitis	39 (46.4%)
Streptococcus pseudopneumoniae	76 (98.7%)
Streptococcus oralis	27 (28.4%)
Streptococcin D	
Species	**n with streptococcin (% of species)**
S pneumoniae (Iceland)	9 (0.5%)
S pneumoniae (Kenya)	85 (2.7%)
Streptococcus gordonii	9 (22.0%)
Streptococcus cristatus	19 (100.0%)
Streptococcus parasanguinis	17 (33.3%)
Streptococcus oralis	9 (9.5%)
Streptococcus sanguinis	3 (5.7%)
Streptococcus anginosus	1 (1.8%)
Streptococcus australis	4 (66.7%)
Streptococcus koreensis	1 (100.0%)
Streptococcus rubneri	2 (100.0%)

Streptococcus xiaochunlingii	1 (100.0%)
Streptococcin E	
Species	**n with streptococcin (% of species)**
S pneumoniae (Iceland)	1840 (96.2%)
S pneumoniae (Kenya)	3115 (98.6%)
Streptococcus anginosus	1 (1.8%)
Streptococcus mitis	25 (29.8%)
Streptococcus pseudopneumoniae	59 (76.6%)
Streptococcus parasanguinis	2 (3.9%)
Streptococcus suis	3 (1.7%)
Streptococcus xiaochunlingii	1 (100.0%)
Streptococcus cristatus	1 (5.3%)
Streptococcus equi	1 (1.2%)

9.3.4 Streptococcin allelic profile distribution in the Icelandic and Kenyan datasets

Table 9.14: The clonal complex (CC) distribution of streptococcin allelic profiles that were commonly observed (>15 times) in both the Icelandic and Kenyan datasets.

Streptococcin A			
Profile	**CC**	**Iceland**	**Kenya**
3-4-5	289	0	55 (79.7%)
	439	32 (14.7%)	0
	113	14 (66.7%)	0
	315	13 (86.7%)	0
	4881	0	10 (100.0%)
	1379	7 (100.0%)	0
	5902	0	2 (0.8%)
	102	2 (100.0%)	0
	2386	0	2 (9.1%)
	Sing5345	0	1 (100.0%)
	4088	0	1 (100.0%)
	Sing5359	0	1 (100.0%)
	1765	0	1 (100.0%)
	Sing14840	0	1 (100.0%)
3-20-9	193	26 (92.9%)	10 (83.3%)
	1262	25 (71.4%)	0
	4368	0	21 (91.3%)
	14930/15024	0	18 (94.7%)
	5077	0	10 (100.0%)
	5902	0	3 (1.3%)
	Sing14961	0	2 (100.0%)
	Sing15022	0	1 (100.0%)
	7053	0	1 (1.7%)
	Sing6102	0	1 (100.0%)
	124	1 (1.6%)	0
	Sing10356	1 (100.0%)	0
	338	0	1 (2.7%)
	Sing14785	0	1 (100.0%)
3-7-5	499	0	50 (90.9%)
	15	22 (61.1%)	0
	5250/5947/15006	0	18 (100.0%)
	701	0	17 (10.4%)
	8397	0	7 (87.5%)
	703	0	6 (37.5%)

	113	4 (19.0%)	0
	Sing14793	0	3 (100.0%)
	1748/14916	0	3 (100.0%)
	Sing2055	0	3 (100.0%)
	177	2 (14.3%)	0
	Sing14924	0	2 (100.0%)
	Sing14779	0	2 (66.7%)
	Sing14862	0	1 (100.0%)
	1146	0	1 (0.7%)
	5936/14865	0	1 (100.0%)
	14977	0	1 (100.0%)
	2213	0	1 (8.3%)
	5329	0	1 (2.6%)
	5339	0	1 (0.7%)
	854	0	1 (1.8%)
	458	0	1 (50.0%)
	4368	0	1 (4.3%)
	138/176	0	1 (0.8%)
Streptococcin B			
Profile	**CC**	**Iceland**	**Kenya**
2-15-11	199	175 (97.8%)	0
	5339	0	122 (85.9%)
	914	0	59 (96.7%)
	2386	0	21 (95.5%)
	66	18 (94.7%)	0
	338	0	7 (18.9%)
	845/14754	0	2 (100.0%)
	Sing14796	0	1 (100.0%)
	14977	0	1 (100.0%)
	Sing14824	0	1 (100.0%)
	Sing14877	0	1 (100.0%)
1-49-20	490	74 (100.0%)	0
	385	1 (100.0%)	25 (80.6%)
	5294	0	16 (100.0%)
	346	2 (100.0%)	0
	14888/14922	0	2 (100.0%)
	Sing14936	0	1 (100.0%)
	3691	1 (100.0%)	0
	5953	0	1 (100.0%)
1-1-1	62	71 (75.5%)	1 (100.0%)
	63	0	69 (98.6%)
	13215	2 (100.0%)	13 (92.9%)

	1012	1 (16.7%)	0
	289	0	1 (1.4%)
	Sing14766	0	1 (100.0%)
2-10-8	439	61 (28.1%)	0
	5938	0	19 (100.0%)
	3983	0	6 (100.0%)
	1379	4 (57.1%)	0
Streptococcin C			
Profile	**CC**	**Iceland**	**Kenya**
12-48-38	236/271/320	291 (99.3%)	4 (100.0%)
	914	0	49 (80.3%)
	338	3 (37.5%)	5 (13.5%)
	172	0	2 (100.0%)
	Sing14914	0	1 (100.0%)
21-120-9	230	3 (100.0%)	81 (88.0%)
	Sing1801	12 (100.0%)	0
	66	6 (31.6%)	0
	5258	0	1 (1.3%)
	Sing14796	0	1 (100.0%)
Streptococcin E			
Profile	**CC**	**Iceland**	**Kenya**
0-9-7	199	174 (97.2%)	0
	230	3 (100.0%)	69 (75.0%)
	2386	0	3 (13.6%)
2-28-4	439	83 (38.2%)	0
	2234	0	12 (85.7%)
	Sing2055	0	3 (100.0%)
	5326	0	2 (100.0%)
	5068	0	1 (33.3%)
	4088	0	1 (100.0%)
	404	0	1 (100.0%)
0-109-7	852	0	74 (94.9%)
	5329	0	33 (86.8%)
	2755	16 (100.0%)	0
	5902	0	3 (1.3%)
	14774	0	2 (8.7%)
	5796	0	2 (14.3%)
	473	0	1 (25.0%)

Note: Showing frequency of allelic profiles within each CC and the percentage of genomes of that CC in each dataset with the allelic profile.

Table 9.15: The number (n) of different allelic profiles of each streptococcin observed in the ten most common clonal complexes (CCs) of the Icelandic and Kenyan pneumococcal datasets.

Iceland								
CC	**Streptococcin A**		**Streptococcin B**		**Streptococcin C**		**Streptococcin E**	
	n	**categories**	**n**	**categories**	**n**	**categories**	**n**	**categories**
236/271/320	1	[Dt]	3	[P]	3	[F]	2	[Dt]
439	7	[F, Dt]	8	[F, Dt]	15	[Dt, D]	11	[F, P]
199	0	-	4	[F]	8	[Dt, F]	5	[P]
138/176	3	[Dt, D]	3	[F]	3	[Di]	3	[F]
180	2	[Dt, F]	5	[P, F]	5	[F, Dt]	2	[P]
62	2	[F]	3	[F]	5	[D, Di]	4	[P]
97	2	[Dt]	6	[F]	5	[Di]	4	[P]
490	1	[Dt]	1	[F]	2	[F]	3	[P]
124	1	[Di, F]	2	[F]	1	[F]	2	[P]
433	5	[Dt, F]	1	[F]	3	[D, Dt]	2	[D]
Kenya								
CC	**Streptococcin A**		**Streptococcin B**		**Streptococcin C**		**Streptococcin E**	
	n	**categories**	**n**	**categories**	**n**	**categories**	**n**	**categories**
5902	16	[F, Dt]	19	[F, P]	23	[Dt, D, F]	27	[P, F, D]
217	3	[F]	3	[D]	5	[Di]	1	[P]
701	7	[Dt, F]	10	[P]	8	[F, Dt]	11	[F, P, Dt]
5339	7	[F, Dt]	9	[F, Di, P]	20	[Dt, F, D]	9	[F, P]
1146	5	[F]	10	[F]	6	[F, Di]	6	[D, P]
138/176	12	[F, Di, Dt]	13	[F]	10	[Dt, F, Di]	9	[F, P]
156/162	6	[F, Dt, Di]	11	[P, F]	8	[Dt, F]	4	[P]
991	4	[F]	4	[F]	5	[Dt]	4	[P]
230	0	-	4	[F]	9	[D, Dt, F]	7	[P, F, D]
852	5	[F]	6	[F]	2	[Dt]	4	[D]

Note: Includes the functional categories represented in each CC: F - full, P - partial, Dt - disrupted toxin, Di - disrupted immunity, D - degrading. Streptococcin D excluded due to low prevalence in pneumococci.

9.3.5 Full and partial streptococcin B and E clusters

Table 9.16: Streptococcin B and E immunity gene alleles found in both full and partial clusters, and their frequency, in pneumococcal genomes.

Streptococcin B			
Immunity gene alleles (B-C)	**Allelic profile**	**Category**	**Frequency**
77-29	1-77-29	Full	64
	0-77-29	Partial	1
146-1	14-146-1	Full	2
	0-146-1	Partial	13
244-104	2-244-104	Full	3
	0-244-104	Partial	147
Streptococcin E			
Immunity gene alleles (B-C)	**Allelic profile**	**Category**	**Frequency**
177-11	1-177-11	Full	140
	0-177-11	Partial	128
40-11	2-40-11	Full	24
	0-40-11	Partial	80
288-7	1-288-7	Full	36
	0-288-7	Partial	66
45-11	1-45-11	Full	53
	0-45-11	Partial	8
28-4	2-28-4	Full	103
	11-28-4	Full	21
	12-28-4	Full	3
	16-28-4	Full	1
	0-28-4	Partial	1

9.4 Chapter 6 Appendices

9.4.1 Recombinant expression trials

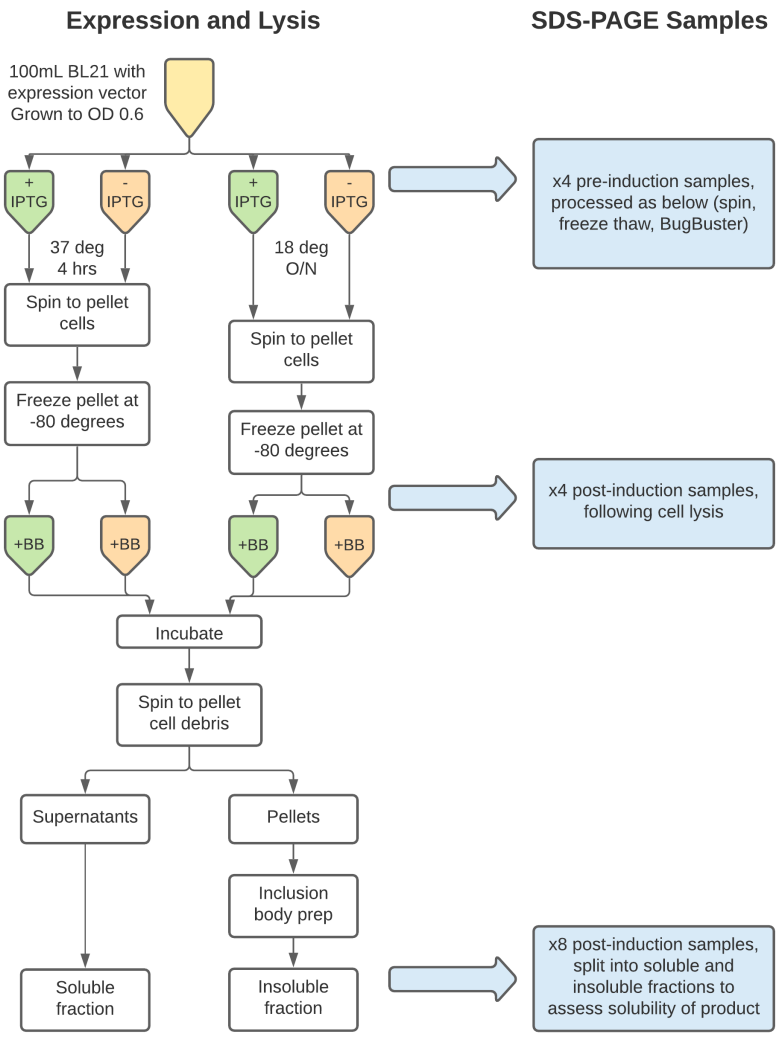

Figure 9.1: Flow chart showing the procedure for trialling expression from the 6His-tagged streptococcin expression vectors. Blue arrows indicate points at which samples were taken for SDS-PAGE.

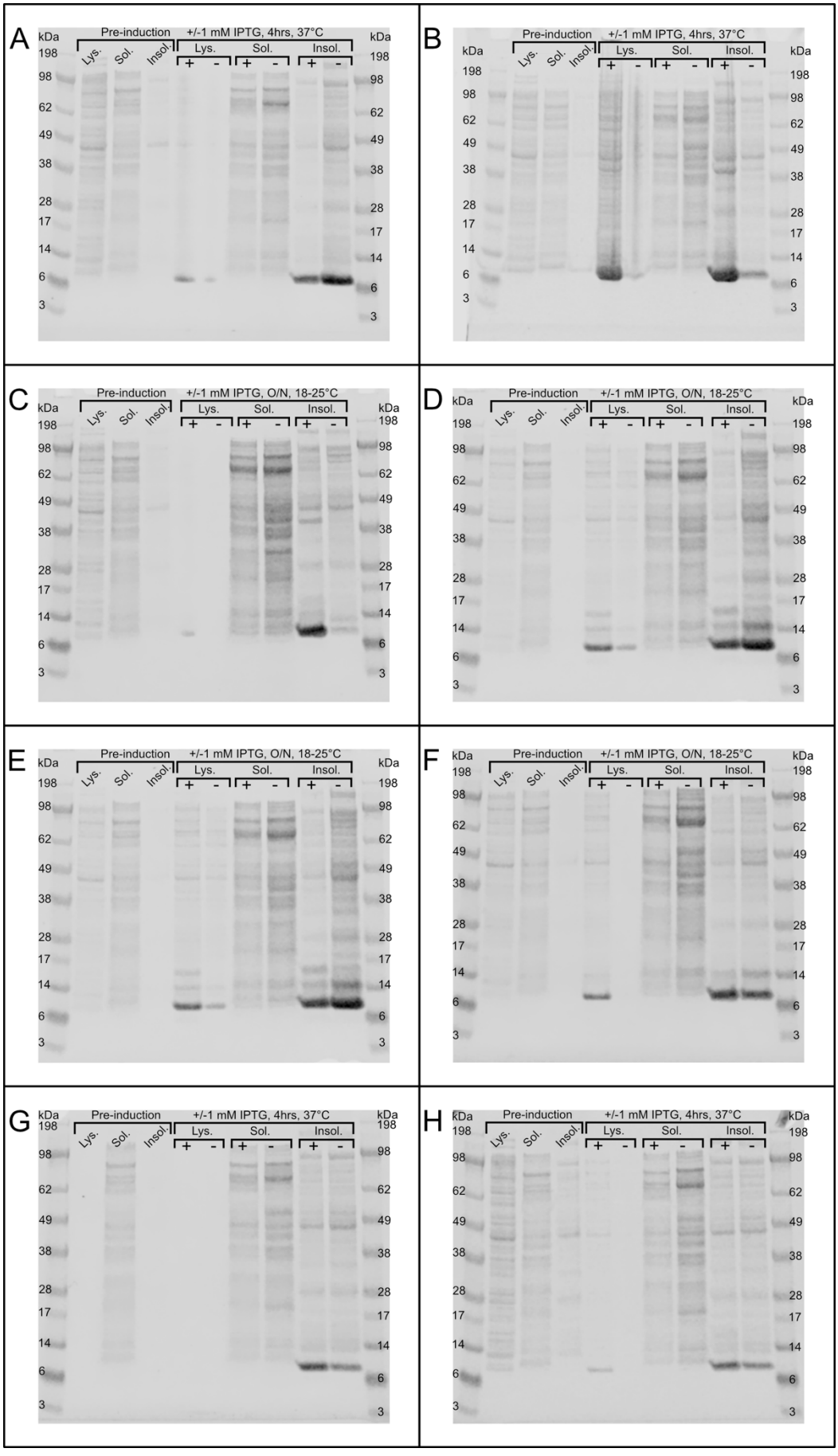
A
B
C
D
E
F
G
H
Pre-induction
+/-1 mM IPTG, 4hrs, 37°C
+/-1 mM IPTG, O/N, 18-25°C
Lys.
Sol.
Insol.
kDa
198
98
62
49
38
28
17
14
6
3

Figure 9.2: Results of small volume expression trials of 6His-tagged streptococcins. One SDS-PAGE gel per construct shows the induction condition that generated the highest yield of product (4 hours at 37 °C or overnight at 18-25 °C). Lys.: total cell lysate, Sol.: soluble fraction of lysate, Insol.: insoluble fraction of lysate. Total cell lysate samples, particularly those from post-induction cultures, were viscous, and in some cases could not be loaded onto SDS-PAGE, resulting in empty lanes. Panel A: streptococcin A allele 2, 4 hours at 37 °C. Panel B: streptococcin A allele 3, 4 hours at 37 °C. Panel C: streptococcin B allele 1, overnight at 18-25 °C. Panel D: streptococcin C allele 2 overnight at 18-25 °C. Panel E: streptococcin C allele 3, overnight at 18-25 °C. Panel F: streptococcin D allele 1, overnight at 18-25 °C. Panel G: streptococcin E allele 1, 4 hours at 37 °C. Panel H: streptococcin E allele 2, 4 hours at 37 °C.

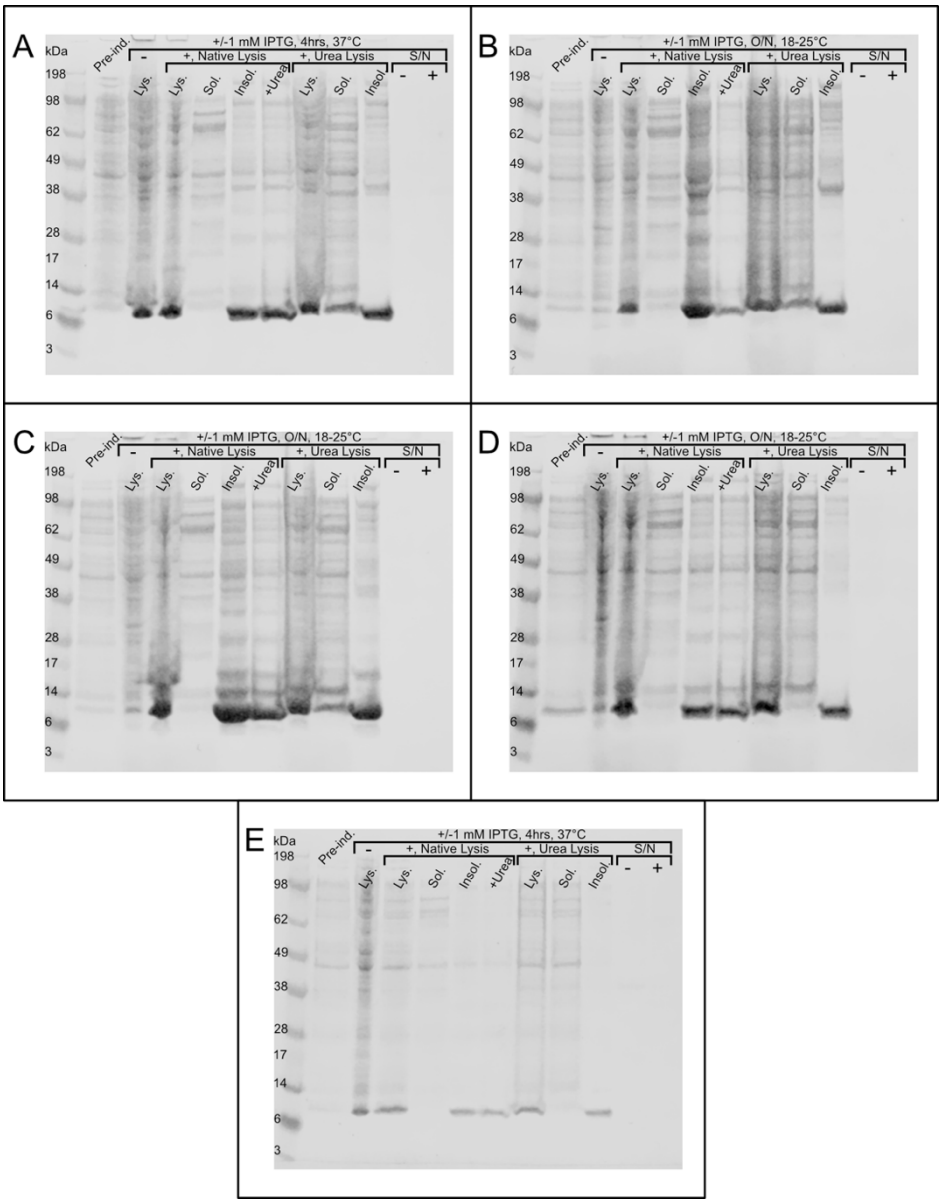

Figure 9.3: Results of small volume expression trials of a subset of 6His-tagged streptococcins adapted to assess methods for re-solubilisation of proteins in the insoluble fraction of cell lysates. Lys.: total cell lysate, Sol.: soluble fraction of lysate, Insol.: insoluble fraction of lysate. Panel A: streptococcin A allele 3, panel B: streptococcin B allele 1, panel C: streptococcin C allele 3, panel D: streptococcin D allele 1, panel E: streptococcin E allele 1.

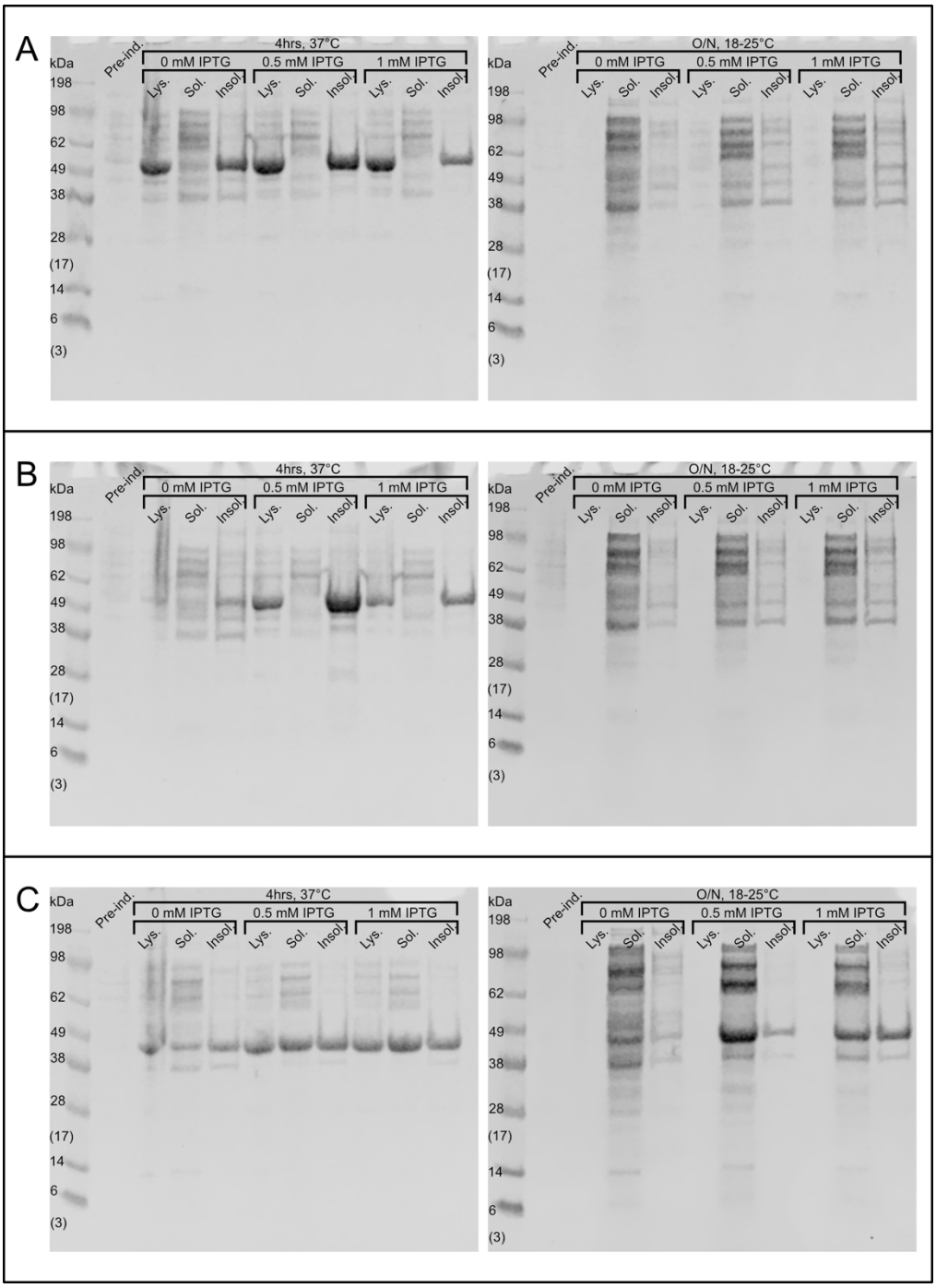

Figure 9.4: Small volume expression trials using MBP-tagged streptococcin A allele 3 (panel A), MBP-tagged streptococcin B allele 1 (panel B) and the empty MBP expression vector (panel C). For each construct, both induction conditions are shown. The pre-induction sample represents the total cell lysate. Lys.: total cell lysate, Sol.: soluble fraction of lysate, Insol.: insoluble fraction of lysate. Total cell lysate samples were viscous and, in some cases, could not be loaded onto SDS-PAGE, resulting in some empty lanes.

9.4.2 Summary of cloning, expression, and purification of tagged streptococcins

Table 9.17: Summary of the cloning, expression, and purification of tagged streptococcins.

Construct	Vector cloned	Small volume expression trials	Re-solubilisation trial	Scaled up expression	Purification with IMAC	Refolding				Concentration and storage
						On-column	One step dialysis	One step dialysis (low pH)	Incremental dialysis	
6His-scaA2										
6His-scaA3										
6His-scbA1										
6His-sccA2										
6His-sccA3										
6His-scdA1										
6His-sceA1										
6His-sceA2										
MBP-scaA3										
MBP-scbA1										

Note: Green shading indicates the procedure was performed successfully and a result was obtained, orange shading indicates a purification optimisation that was attempted unsuccessfully (and resulted in precipitation of the product), and grey indicates that the procedure was not attempted. IMAC: immobilised metal affinity chromatography.

9.4.3 Optimisation of 6His-tagged streptococcin A refolding

On-column refolding of the 6His-tagged streptococcin A was attempted using a modified IMAC procedure to switch to a native buffer during the wash steps. The column was first washed with denaturing buffer (Table 6.6, buffer W1), then with three additional 10 mL washes of a native buffer (0 M urea, Table 6.6, buffer W2), before elution with a native buffer (Table 6.6, buffer E2). Additionally, the purification and dialysis procedures were adjusted to trial refolding by dialysis at pH 6.0, rather than pH 8.0. Cell lysis proceeded as above (at pH 8.0), the IMAC washes and elutions used pH 7.0 buffers (as an acidic pH may prevent binding of 6His-tagged proteins to the IMAC resin), and the eluted product was dialysed overnight into a native pH 6.0 buffer (Table 6.6, buffers L3, L4, W3, E3, D11).

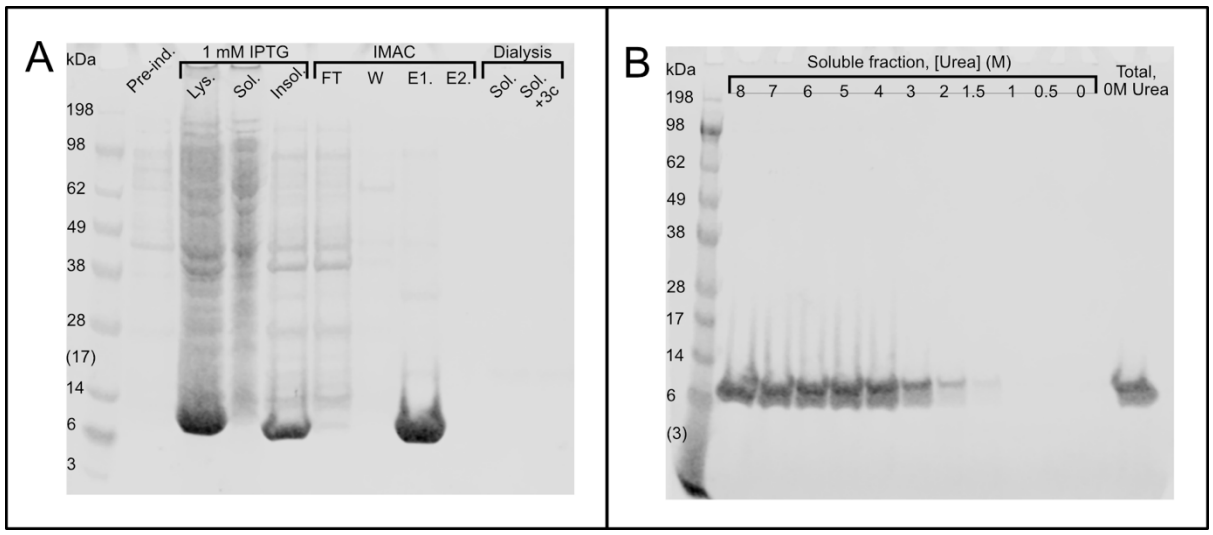

Figure 9.5: Purification and attempted refolding of 6His-tagged streptococcin A. Panel A: expression and purification of 6His-tagged streptococcin A by immobilised metal affinity chromatography (IMAC) and attempted refolding with a single step dialysis. Pre-ind.: pre-induction, Lys: total cell lysate, Sol.: soluble cell fraction, Insol.: insoluble cell fraction, FT: flow through, W: wash, E1: elution 1, E2: elution 2, Dialysis Sol.: soluble fraction of dialysed product, 3c: 3c protease. Panel B: Attempted refolding of 6His-tagged streptococcin A by incremental dialysis. Soluble fraction of the dialysed product shown at each concentration of urea, final lane shows the total product at 0M urea including precipitated protein.

9.4.4 Optimised protocol for the expression and purification of streptococcin B allele 1

Whole procedure summarised in Figure 6.2, all named buffer compositions can be found in Table 6.6.

9.4.4.1 Large volume expression

1) Inoculate 10 mL of autoclaved LB supplemented with 50 μg/mL kanamycin with a scrape from a glycerol stock of NiCo21-De3 *E. coli* transformed with the expression vector pET-17b/streptococcin B allele 1.
2) Grow overnight at room temperature with shaking (220 rpm).
3) Use the overnight culture to inoculate 500 mL of autoclaved LB supplemented with kanamycin. If multiple 500 mL batches are being used, split the overnight culture evenly between flasks.
4) Grow at 37 ˚C with shaking (220 rpm) until the OD600 reaches 0.6.
5) Take a 1 mL pre-induction culture sample, spin down at 3,000 rpm for 2 minutes and store the cell pellet at -20 ˚C.
6) Induce the OD600 expression cultures with 1 mM IPTG from a sterile 1000x stock.
7) Grow induced cultures at room temperature (18 - 25 ˚C) overnight with shaking (220 rpm).
8) Take a 1 mL post-induction culture sample, spin down at 3,000 rpm for 2 minutes and store the cell pellet at -20 ˚C.
9) Harvest cells by centrifugation at 3,000 rpm for 15 minutes.
10) Freeze cell pellets at -80 ˚C.
11) Assess induction by lysing the frozen cell pellets from each 1 mL culture sample and run the total, soluble and insoluble fractions of the lysate on SDS-PAGE.

9.4.4.2 Cell lysis

1) Defrost cell pellets from a large volume expression.
2) Resuspend the pellets in BugBuster master mix supplemented with protease inhibitors (buffer L5). Use 25 mL of lysis buffer per 500 mL of original expression culture.
3) Incubate at room temperature with agitation for 20 minutes.
4) Separate the soluble and insoluble cell fractions by ultracentrifugation at 10,000 rpm for 45 minutes at 4 ˚C.
5) Resuspend the cell pellet in the same volume of denaturing buffer (buffer L4) as used for initial resuspension of the cell pellets. Incubate at room temperature for 60 minutes.
6) Centrifuge the solution at 2,500 rpm until any remaining solid debris are pelleted and remove the supernatant. This is the IMAC load.

9.4.4.3 Purification

1) Take a 20 µL sample of the IMAC load for SDS-PAGE.
2) Prepare a His GraviTrap column by pouring off the storage buffer, cutting open the spout and washing with 10 mL of the load/wash buffer (same buffer as the insoluble fraction of the cell lysate is in, buffer W1).
3) Load the insoluble cell fraction onto the column and collect the flow-through.
4) Take a 20 µL sample of the flow-through for SDS-PAGE.
5) Wash the column with 10 mL of the load buffer (buffer W1) and collect the flow-through.
6) Take a 20 µL sample of the wash flow-through for SDS-PAGE.
7) Elute the column with 3 mL elution buffer containing 500 mM imidazole (buffer E1).
8) Take a 20 µL sample of the elution for SDS-PAGE.
9) Run all samples on SDS-PAGE to assess the purification.
10) If the elution is high in streptococcin B and low in contaminants, proceed to re-folding.

9.4.4.4 Refolding by incremental dialysis

1) Dilute the elution from the IMAC 10-fold in the elution buffer (buffer E1).
2) Take a 20 µL pre-dialysis/8 M urea sample for SDS-PAGE and store at - 20 ˚C.
3) Load the diluted protein solution into sufficient dialysis tubing and dialyse against the first dialysis buffer (buffer D2) for 8 -24 hours at 4 ˚C.
4) Following dialysis, open the tubing a carefully take a 20 µL sample for SDS-PAGE and store at - 20 ˚C.
5) Repeat steps 3-4 for dialysis buffers D3 - D11. Once urea concentration is below 2 M, dialyse for at least 18 hours. Record the point at which visible precipitation occurs. After this point, spin down SDS-PAGE samples at 2,500 rpm for 3 minutes and store only the supernatant (containing the soluble, refolded protein).
6) Finish by dialysing as in step 3 into buffer D1 (0 M urea).
7) Recover the sample from the dialysis tubing and take two 20 µL samples for SDS-PAGE. Store one sample immediately and spin the other down as before.
8) Run the samples from each intermediate stage of dialysis to assess the extent of re-folding.
9) If a band is visible in the soluble fraction of the final dialysed product, assess the yield of re-folded product by A_{280}.

9.4.4.5 Long term storage

1) Dialyse the re-folded protein into buffer D12, an intermediate salt buffer.
2) Dialyse the re-folded protein into buffer D13, the long-term storage buffer.
3) Optional: verify the presence of protein by SDS-PAGE.

4) Assess the yield of protein by A_{280}.
5) Concentrate the protein using spin concentrators as much as possible without precipitation. Monitor the concentration by A_{280}. His6-streptococcin should be soluble at 100 - 130 μg/mL.
6) Divide the final product into 250 - 500 μL aliquots and store at -80 ˚C.

9.5 Conference Abstracts

9.5.1 Abstract for EuroPneumo 2019

Genomic studies in an Icelandic dataset reveal complexity in bacteriocin prevalence and distribution

Butler, MEB (1); Jansen van Rensburg, MJ (1); van Tonder, AJ (2); Quirk, S (3); Haraldsson, G (3); Haraldsson, A (4); Erlendsdóttir, H (3); Kristinsson, KG (3); Brueggemann, AB (1, 5)

1. Department of Medicine, Imperial College London, London, United Kingdom
2. Infection Genomics, Wellcome Sanger Institute, Hinxton, United Kingdom
3. University of Iceland and Landspitali University Hospital, Reykjavík, Iceland
4. University of Iceland and Children´s Hospital Iceland, Reykjavík, Iceland
5. Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

Bacteriocins are antimicrobial peptides produced by many bacteria and are believed to inhibit competing strains. Genomic studies identified a diverse range of pneumococcal bacteriocins, but their competitive influence on bacterial populations is poorly defined. 'Cheater' bacteriocin gene clusters have been reported, which lack the toxin gene (avoiding cost of production) but retain immunity genes to protect the cheater strain. We investigated the prevalence, diversity and distribution of 20 different bacteriocins in 1,916 genomes of carriage and invasive pneumococci recovered from 2009-2014, pre/post-PCV introduction. 18 bacteriocin clusters were represented. The overall prevalence of each ranged from 0.1% to 100%, and 4 - 9 different bacteriocins were found in every genome. Four of the most prevalent were streptococcins A (80%), B (100%), C (100%) and E (96%). Between 19% (streptococcin A) and 62% (streptococcin E) of streptococcins were potential cheaters and were distributed between 18 (streptococcin A) and 35 (streptococcin E) distinct lineages, although cheater prevalence varied among pneumococci within each lineage. There were differences in the prevalence of complete and cheater clusters in pre- and post-vaccination periods, which can be explained by PCV-

induced population restructuring and changes in the prevalence of specific lineages. Genomes from particular lineages contained consistent sets of streptococcins, such as those from the multi-drug resistant CC236/271/320 lineage, which typically had a complete streptococcin A and C, and a putative cheater streptococcin B and E. Work is ongoing to understand how lineage-specific complexity in the distribution of both complete and cheater bacteriocins influences the overall pneumococcal population structure.

9.5.2 Abstract for ECCMID 2021

Vaccine-induced population restructuring alters the prevalence of pneumococcal bacteriocins in Icelandic and Kenyan genomic datasets

Madeleine EB Butler (1), Melissa J Jansen van Rensburg (2), Angela Karani (3), Benedict Mvera (3), Donald Akech (3), Asma Akter (1), Calum Forrest (1), Andries J van Tonder (4), Sigríður J Quirk (5), Gunnsteinn Haraldsson (5), Stephen D Bentley (6), Helga Erlendsdóttir (5), Ásgeir Haraldsson (7), Karl G Kristinsson (5), J Anthony G Scott (3,8), Angela B Brueggemann (1,2)

1. Imperial College London, London, United Kingdom
2. University of Oxford, Oxford, United Kingdom
3. KEMRI Wellcome Trust Programme, Kilifi, Kenya
4. University of Cambridge, Cambridge, United Kingdom
5. University of Iceland and Landspitali University Hospital, Reykjavík, Iceland
6. Wellcome Sanger Institute, Hinxton, United Kingdom
7. University of Iceland and Children´s Hospital Iceland, Reykjavík, Iceland
8. London School of Hygiene and Tropical Medicine, London, United Kingdom

Background

Streptococcus pneumoniae (pneumococcus) is an opportunistic pathogen that colonises the paediatric nasopharynx, from where it can invade to cause pneumonia or invasive disease. The most important pneumococcal virulence factor is the polysaccharide capsule

and 100 serotypes have been identified to date. Current pneumococcal conjugate vaccines (PCVs) target 10 or 13 serotypes and PCV use within a human population perturbs the bacterial population structure. The consequences of such perturbations include changes in nasopharyngeal colonisation and the competition dynamics among pneumococci within this ecological niche. Bacteriocins are antimicrobial peptides produced by bacteria to inhibit competing strains and 20 different putative pneumococcal bacteriocins have been identified. This study aimed to compare the bacteriocin distributions among pneumococci recovered pre- and post-PCV10 introduction in two countries.

Methods

Carriage and disease pneumococci collected from children and adults in Iceland (n=1,912; 2009-2014) and Kenya (n=3,159; 2003-2017) spanning PCV10 introduction (2011 in both countries) were sequenced. Genome assemblies were screened for the 20 different bacteriocin sequences. The distribution and combinations of bacteriocins within pneumococci recovered pre- and post-PCV10 introduction were assessed and stratified by country, genetic lineage and serotype.

Results

PCV10 use led to a significant reduction in the prevalence of vaccine serotypes and associated genetic lineages in both countries. Overall, 18 of 20 bacteriocins were detected in 0.1% to 100% of pneumococci, and between 4 and 11 different bacteriocins were detected per genome. Two bacteriocins were not detected in either dataset. Post-PCV10, three bacteriocins were significantly altered in prevalence in the Icelandic dataset, and these three plus three additional bacteriocins differed significantly in the Kenyan dataset. Bacteriocins were associated with genetic lineages that also significantly changed in prevalence. The specific combinations of bacteriocins within a genome were inconsistent across all representatives of some genetic lineages.

Conclusions

Bacteriocin distributions changed significantly post-PCV10 introduction in both Iceland and Kenya as a result of pneumococcal population restructuring, which suggests altered competition dynamics in post-vaccine populations. This may affect which pneumococci

are more likely to colonise the nasopharynx and therefore have the opportunity to cause disease. The consequences for pneumococcal disease in the long term remain to be determined.

9.5.3 Abstract for ISPPD-12, June 2022

Investigation of 7000 genomes reveals that streptococcin bacteriocins are shared between pneumococci and non-pneumococcal *Streptococcus* species, contributing to the diversification of bacteriocins

Madeleine EB Butler (1), Melissa J Jansen van Rensburg (2), Femke M Ahlers (2), James E Bray (3), Keith A Jolley (3), Angela B Brueggemann (1,2)

1. Imperial College London, London, United Kingdom
2. Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
3. Department of Zoology, University of Oxford, Oxford, United Kingdom

Background

Pneumococcal bacteriocins are antimicrobial peptides believed to be used during nasopharyngeal competition. Streptococcins are one type of bacteriocin, encoded by a toxin gene and two immunity genes, and five different streptococcins (A-E) have been identified among pneumococci. The aims of this study were to investigate the diversity and distribution of pneumococcal streptococcins and determine whether pneumococcal streptococcins were present in non-pneumococcal *Streptococcus* species.

Methods

Three curated genomic datasets were used: carriage and disease pneumococci collected in Iceland (n=1,916; 2009-2014) and Kenya (n=3,257; 2003-2017); and 1,825 genomes of 55 non-pneumococcal *Streptococcus* species. All 6,998 genomes were screened to

identify streptococcin gene clusters, using a semi-automated methodology in BIGSdb (pubmlst.org/software/bigsdb/) followed by manual curation.

Results

Among 5,173 pneumococcal genomes, streptococcins B, C and E were ubiquitous (100%), streptococcin A was prevalent (80%), and streptococcin D was rare (<3%). Streptococcin cluster composition varied: the toxin gene was frequently absent among streptococcin B (27%) and E (55%) gene clusters; and disruptions to the toxin gene, presumably leading to loss of function, were observed among all streptococcins apart from D (2-45% per streptococcin). Identical toxin genes were associated with divergent immunity genes within streptococcins A, B, C and E. Identical streptococcin gene clusters were observed within the same pneumococcal lineage and across different lineages. Finally, pneumococcal streptococcin gene clusters were identified among 10 different non-pneumococcal *Streptococcus* species, most frequently among viridans streptococci (21-100% of each species).

Conclusion

Horizontal genetic exchange likely mediates the wide distribution and heterogeneity of streptococcins observed among pneumococci and the nasopharyngeal microbiome.