Genomic investigations of bacteriocins and bacteriophages in *Streptococcus pneumoniae* and other streptococci

Reza Rezaei Javan

(Nuffield Department of Medicine and St Catherine's College, University of Oxford)

Thesis submitted for examination for the degree of Doctor of Philosophy in Clinical Medicine

(Trinity Term 2019)

Primary supervisor: Prof Angela B. Brueggemann

Joint supervisor: Prof Paul Klenerman

Reza Rezaei Javan, St. Catherine's College
Thesis submitted for examination for the degree of DPhil in Clinical Medicine, Trinity Term 2019.

Genomic investigations of bacteriocins and bacteriophages in *Streptococcus pneumoniae* and other streptococci

Abstract

Streptococcus pneumoniae (the "pneumococcus") is a major public health problem, leading to significant morbidity and mortality worldwide. The global increase in antimicrobial-resistant pneumococci is of serious concern, which necessitates investigations into novel antimicrobial strategies. Two promising areas of research are the exploitation of phages (viruses that infect bacteria) and bacteriocins (antimicrobial peptides produced by bacterial species to kill other bacteria). This thesis uses genomic approaches to address questions related to the prevalence, diversity and molecular epidemiology of bacteriocins and prophages (phage genomes that are integrated into bacterial chromosomes).

The first part analysed >6,200 pneumococcal genomes and discovered 14 novel bacteriocin gene clusters. The molecular epidemiology of the bacteriocin clusters was investigated in the context of the pneumococcal population structure. The results revealed extraordinary bacteriocin diversity among pneumococci and the majority of bacteriocin clusters were also present in other streptococcal species. Genomic hotspots for the integration of different bacteriocin gene clusters were discovered. Bacteriocin genes were found to be transcriptionally active when the pneumococcus was under stress and when two different strains were co-cultured in broth.

The second part describes the molecular epidemiology of satellite prophages (prophages that rely on the host and an additional helper phage for survival) in a large global and historical pneumococcal dataset. Forty-four unique satellite prophages were newly identified, which had persistent associations with specific widely-circulating pneumococcal clonal complexes over many decades. Collaborative experimental work demonstrated that one of these satellite prophages was associated with virulence in a murine model of infection. RNA sequencing revealed that satellite prophage genes were overexpressed when pneumococci were grown planktonically versus in *in vitro* biofilm experimental conditions.

The final part analysed >1,300 genomes of 70 different *Streptococcus* species and identified nearly 800 prophages and satellite prophages. The data showed that prophages and satellite prophages were widely distributed among streptococci and constituted two distinct entities with little effective genetic exchange between them. Contrary to the current dogma that suggests prophages are bacterial species-specific, there was convincing evidence that transmission of prophages occurred between genetically different streptococcal species.

These results broaden our understanding of bacteriocins and phages among streptococci and uncover many areas for future studies.

Declaration

I hereby declare that this thesis is my own work and where collaborative work was undertaken this is indicated at the beginning of each Results chapter (Chapters 3, 4 and 5). None of the contents of this thesis have previously been submitted for any other degree. I completed all analyses under the supervision of Prof Brueggemann.

Most of the results described in Chapter 3 were presented as an oral presentation at the 13th European Meeting on the Molecular Biology of the Pneumococcus (EuroPneumo 2017), in Stockholm, Sweden in June 2017. Part of the results described in Chapter 3 were also presented at the 11th International Symposium on Pneumococci and Pneumococcal Diseases (ISPPD 2018), in Melbourne, Australia in April 2018.

The findings described in Chapter 4 were presented in-part at two conferences: i) in an oral presentation at ISPPD 2018; and ii) in an oral presentation at the 14th European Meeting on the Molecular Biology of the Pneumococcus (EuroPneumo 2019), in Greifswald, Germany in June 2019.

The results described in Chapter 5 were presented as an oral presentation at EuroPneumo 2019. A copy of all conference abstracts related to this thesis are provided in Appendices 1-6.

All Results chapters of this thesis have been accepted for publication in peer-reviewed journals; these manuscripts were written by me and Prof Brueggemann and were subsequently reviewed by each of the co-authors (as listed at the beginning of these chapters). A copy of all publications related to this thesis are provided at the end of the thesis. All other sections of this thesis were written by me.

Signed,

Reza Rezaei Javan

Acknowledgments

First and foremost, I thank Prof Angela Brueggemann, the primary supervisor of this thesis, for her excellent supervision, advice and overwhelming support. I have learned a tremendous amount during my DPhil course from her and had the opportunity to travel to several scientific conferences. Angela is always very generous with her time and I am privileged to have had the opportunity to learn from her vast knowledge and experience. I must extend my thanks to my other supervisor, Prof Paul Klenerman, who provided additional support during the time Angela was based away from Oxford.

I am grateful to Dr Andries van Tonder for introducing me to many useful bioinformatics and computational skills and for general discussion of ideas. I have also benefited from several productive discussions with Prof Sunetra Gupta, Dr Philippa Matthews and Dr Melissa Jansen van Rensburg. I thank the Peter Medawar Building for Pathogen Research and the Big Data Institute for providing me with a great environment to learn in and perform research.

Science is a highly collaborative endeavour and I am indebted to the numerous scientists who had previously developed the computational tools and algorithms that I used in this thesis, without whom this work would have been impossible. Special thanks must go to Prof Martin Maiden and Dr Keith Jolley for the set-up and maintenance of the BIGS database infrastructure, which was used extensively throughout my research. I acknowledge the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) project for providing additional computational resources. I am indebted to all the individuals who had contributed to the generation of the transcriptomic data, and the isolation, sequencing and assembly of the bacterial genomes that were used in this project. I thank Prof Jeremy Brown who arranged for the laboratory work described in Chapter 4 to take place in his laboratory at University College London, where Dr Elisa Ramos-Sevillano and Dr Asma Akter performed the experiments.

I thank St. Catherine's College for providing financial support, wonderful memories, and a warm home that will be dearly missed. I am also thankful to the Leatherseller’s Company for their generous scholarship.

Finally, I am eternally indebted to my parents who have unconditionally supported me in every imaginable way throughout my life. I am also grateful to Katherine Summers, my friends, sister, niece and nephew, for their support and encouragement.

List of contents

Title ***1***

Abstract ***2***

Declaration ***3***

Acknowledgments ***4***

List of contents ***5***

Abbreviations ***15***

1. General introduction ***18***

1.1 An overview of the Streptococcus genus ***18***

Table 1.1 The *Streptococcus* genus main subgroups based on 16S rRNA sequence phylogeny 20

Figure 1.1 Growth of genome sequences in GenBank from 2003 - 2019 22

1.2 The pneumococcus ***22***

1.2.1 A brief history 22

1.2.2 Identification and epidemiological characterisation 23

1.2.3 General biology 25

1.2.4 Interspecific competition between the pneumococcus and other bacterial species 28

1.2.5 Global disease burden 29

1.2.6 Vaccines 30

1.2.6.1 Early vaccines 30

1.2.6.2 Polysaccharide vaccines 31

1.2.6.3 Conjugate vaccines 32

1.2.6.4 Impact of vaccination on the pneumococcal population 33

1.2.7 Antimicrobials and resistance 34

1.2.7.1 A brief historical perspective 35

1.2.7.2 Beta-lactams 36

1.2.7.3 Macrolides 39

1.2.7.4 Fluoroquinolones 40

1.2.8 Need for novel antimicrobial strategies 42

1.3 Bacteriocins 43

1.3.1 General characteristics 43

1.3.2 Bacteriocin classification 45

1.3.3 Lantibiotics 46

1.3.4 Head-to-tail cyclised peptides 48

1.3.5 Sactipeptides 49

1.3.6 Lassopeptides 50

1.3.7 Lactococcin 972-like peptides 51

1.3.8 Bacteriocin discovery 51

1.4 Bacteriophages 54

1.4.1 General characteristics 54

1.4.2 Lytic *vs.* lysogenic phages 55

1.4.3 Impact of lysogeny on host cell 56

1.4.4 Constant arms race between bacteria and phages 56

1.4.5 Satellite prophages 57

1.4.6 Prophage discovery 59

1.5 Aim and outline of this thesis 61

2. General Methods 63

2.1 BIGSdb 63

2.2 Genome collection 63

2.3 RNA sequencing datasets 64

2.4 BLAST 65

2.5 Gene prediction and annotation 66

2.5.1 RAST 66

2.5.2 Prokka 67

2.5.3 CD-Search 67

2.5.4 STRING 68

2.6 Visualisation of phylogenetic trees 68

2.7 Estimation of prophage content among bacterial genomes 68

3. Genome sequencing reveals a large and diverse repertoire of antimicrobial peptides 70

3.1 Abstract 71

3.2 Introduction and aims 72

3.3 Methods 76

3.3.1 Genome mining for the discovery of novel bacteriocin clusters 76

Figure 3.1 Outline of the pipeline for discovery of novel bacteriocins. 77

3.3.2 Investigation of the predicted bacteriocin genes 78

3.3.3 Classification and nomenclature of the identified bacteriocin clusters 78

Table 3.1 Classification and nomenclature of bacteriocin clusters found among pneumococci. 79

3.3.4 Investigation of the molecular epidemiology of the bacteriocin clusters.....82

3.3.5 Generation of the core genome phylogenetic tree 82

3.3.6 Calculating the GC content of *Streptococcus* species 83

3.3.7 Analyses of bacteriocin cluster insertion sites.....83

3.3.8 RNA sequencing analyses 84

3.3.8.1 Heat experiment 84

3.3.8.2 Competition experiment 84

Figure 3.2 A visual summary of the methodology for analysing the RNA sequencing data from the co-colonisation experiment. 86

3.4 Results 87

3.4.1 Genome mining triples the number of known bacteriocins in *S. pneumoniae* .87

Table 3.2 List of the bacteriocins identified among pneumococci. 88

Table 3.3 Result of the BLAST searches of the bacteriocin clusters in the NCBI database. 88

Figure 3.3 GC content of pneumococcal bacteriocin clusters. 92

Figure 3.4 A schematic representation of each streptococcin cluster and their flanking regions 94

Figure 3.5 A distance matrix of the nucleotide sequence identity shared between genes of different streptococcin clusters. 95

Figure 3.6 A schematic representation of the finished genome of pneumococcal strain G54 with the locations of the five streptococcin clusters highlighted in red. 95

Figure 3.7 Genetic composition of streptolancidin, streptocyclicin, streptosactin, and streptolassin bacteriocins in the pneumococcal genomes. 96

Figure 3.8 Predicted bacteriocin genes identified in this study aligned against similar bacteriocin genes in other bacterial species for comparison.....97

3.4.2 Bacteriocin heterogeneity within a global pneumococcal population dataset ..98

Table 3.4 Molecular epidemiology of the bacteriocin clusters identified among a dataset of 571 pneumococci recovered since 1916 from patients of all ages residing in 39 different countries.....99

Figure 3.9 The number of bacteriocins per genome among the 571 pneumococcal genomes..... 100

Figure 3.10 Bacteriocin combinations found within a global pneumococcal dataset. ... 100

Figure 3.11 Diversity of bacteriocins within a global pneumococcal dataset. 101

Table 3.5 The overall prevalence of the streptococcin gene clusters..... 102

3.4.3 Genomic hotspots for the integration of different bacteriocin gene clusters...103

Figure 3.12 Whole genome-based population analysis reveals evidence for bacteriocin switching among pneumococci. 104

3.4.4 Transcriptome analyses demonstrate temporal dynamic changes in the expression of bacteriocin genes in response to heat 105

Figure 3.13 Dynamic changes in the expression of bacteriocin genes in response to bacterial stress. 106

3.4.5 Bacteriocins are induced in response to strain competition 107

Figure 3.14 Evidence for the upregulation of bacteriocin genes when two reference strains, PMEN3 and PMEN6, were co-cultured in broth media 107

3.5 Discussion 108

3.6 Supplementary information 112

4. Genomic investigation and molecular epidemiology of satellite prophages in S. pneumoniae 113

4.1 Abstract 115

4.2 Introduction and aims 116

Figure 4.1 – RNA sequencing analyses indicate evidence for phage gene expression. 118

4.3 Methods 120

4.3.1 Genome dataset 120

4.3.2 Identification of satellite prophages among 'partial prophage' sequences 120

Figure 4.2 Genomic organisation of a subset of satellite prophages previously reported among various bacterial species. 121

4.3.3 Sequence analyses of prophages 122

4.3.4 Investigation of prophage insertion sites 122

4.3.5 Construction of a pneumococcal core genome phylogenetic tree 123

4.3.6 RNA sequencing analyses 123

4.3.6.1 Dataset 1: prophage induction with mitomycin C 124

4.3.6.2 Dataset 2: biofilm versus planktonic mode of growth 124

4.3.7 Assessment of virulence in murine pneumococcal infection model 125

4.3.7.1 Bacterial strains, media and growth conditions 125

4.3.7.2 Construction of pneumococcus mutant strains lacking SpnSP38 and *vapE* 125

4.3.7.3 Experimental models of infection 126

4.3.7.4 C3b binding to pneumococci 127

4.3.7.5 Neutrophil killing assay 127

4.4 Results 128

4.4.1 Identification of 44 unique pneumococcal satellite prophages 128

Figure 4.3 Pneumococcal satellite prophages demonstrate well-conserved patterns in genome organisation and synteny. 129
Figure 4.4 Pneumococcal satellite prophages can be clustered into five major groups. 130
Figure 4.5 The average GC content of the representative pneumococcal satellite prophages. 131
Figure 4.6 Graphical representation of pneumococcal satellite and full-length prophages by genome size and number of genes. 132
Figure 4.7 The associations between representative satellite prophages, their integrase gene and insertion sites. 133
Figure 4.8 Insertion sites of satellite prophages within the pneumococcal genome. 134
4.4.2 Molecular epidemiology of satellite prophages within a global and historical pneumococcal dataset 135
Table 4.1 Epidemiological characteristics of 44 representative satellite prophages identified among a collection of pneumococcal isolates dating from 1939 onwards 136
Figure 4.9 A phylogenetic tree of the genomes in the study dataset labelled according to the presence of different prophages. 137
Figure 4.10 The average prophage content of the major pneumococcal clonal complexes. 138
4.4.3 Satellite prophages and *vapE* are associated with virulence 139
Table 4.2 The predicted function of genes among the 44 pneumococcal satellite prophage genomes identified in the study. 139
Figure 4.11 Construction of *ΔvapE* and *ΔSpnSP38* pneumococcal mutant strains. 141
Figure 4.12 Assessment of the virulence of *ΔSpnSP38* and Δ*vapE* mutant pneumococcal strains in murine infection. 142
4.4.4 The satellite prophage is required for optimum growth in sera but not for evasion of complement recognition or phagocytosis 143
4.4.5 Satellite prophage genes, including *vapE*, were overexpressed in planktonic versus biofilm samples 143

Figure 4.13 Satellite prophage genes are overexpressed when pneumococci is grown planktonically compared to as a biofilm. 145
Figure 4.14 RNA sequencing analyses indicate evidence for satellite prophage replication. 146

4.5 Discussion 147

4.6 Supplementary information 150

5. Prophages and satellite prophages are widespread among Streptococcus species 151

5.1 Abstract 152

5.2 Introduction and aims 153

5.3 Methods 154

5.3.1 Development of PhageMiner, a bioinformatics tool for prophage identification in bacterial genomes 154

5.3.2 Streptococcal genomes dataset 156

5.3.3 Sequence analyses of prophages 157

5.3.4 Investigation of prophage insertion sites 157

5.3.5 Estimate of phylogenetic relationships among *Streptococcus* species 158

5.4 Results 158

5.4.1 Prophages are a significant component of genomes of most clinically relevant streptococcal species 158
Table 5.1 Descriptive statistics for the identified full-length and satellite prophages. 159
Table 5.2 Average prophage content within each streptococcal species. 160

5.4.2 Full-length and satellite prophages are separate entities with little effective genetic exchange between them 161

Figure 5.1 Graphical representation of all prophages by average genome size and number of genes. 161

Figure 5.2 One satellite prophage that was found among pneumococci isolated between 1939 and 2006. 162

Figure 5.3 An unrooted phylogenetic tree of all streptococcal prophage genomes identified in the dataset. 163

Figure 5.4 Venn diagram showing genes shared between full-length prophages and satellite prophages at a threshold of >70% amino acid sequence similarity. 164

5.4.3 Streptococcal prophages demonstrate a structured population 164

Figure 5.5 Full-length and satellite prophages from different streptococcal species demonstrate similar patterns of genome organisation and synteny. 165

Figure 5.6 Schematic representation of examples of full-length and satellite prophage genomes from different non-streptococcal species. 166

Figure 5.7 A neighbour-joining phylogenetic tree of all streptococcal prophage genomes identified in the dataset. 167

Figure 5.8 Evidence for cross-species transmission of prophages 168

Figure 5.9 Distance matrix of pairwise similarity among one example of a cluster of full-length prophages that were found among multiple streptococcal species 169

5.4.4 Streptococcal prophages were more frequently inserted among genes involved in information storage and processing 170

Figure 5.10 The location of prophage insertion sites within the bacterial genomes. 171

Figure 5.11 Insertion sites of prophages within the streptococcal genomes. 172

Table 5.3 The proportions of genes in each COGs (clusters of orthologous groups) for streptococcal genomes on average compared to those flanking prophages 173

5.5 Discussion 174

5.6 Supplementary information 177

6. Summary and future work 178

6.1 Summary 178

6.1.1 The pneumococcus possesses an unexpectedly large bacteriocin repertoire 178

6.1.2 There is extraordinary bacteriocin diversity within the global pneumococcal population 179

6.1.3 Bacteriocin clusters missing one or more genes are not uncommon....180

6.1.4 There is evidence for the intra- and interspecies exchange of bacteriocin clusters....181

6.1.5 Satellite prophages are widespread among pneumococcal genomes and possess a structured population 182

6.1.6 The pneumococcal satellite prophages are maintained at specific sites on the bacterial chromosome....182

6.1.7 Satellite prophages may play a role in pneumococcal pathogenesis....183

6.1.8 Prophages are a major component of genomes of most clinically relevant streptococcal species and a significant driving force for bacterial strain diversification 184

6.1.9 Full-length and satellite prophages are separate entities with little effective genetic exchange between them 184

6.1.10 There is convincing evidence that cross-species transmission of prophages is not uncommon 185

6.1.11 Streptococcal prophages are frequently inserted among genes involved in information storage and processing....186

6.2 Future work....186

7. References....190

8. Appendices....241

Abbreviations

ABC	ATP-Binding Cassette
ACT	Artemis Comparison Tool
antiSMASH	Antibiotics and Secondary Metabolite Analysis Shell
ANOVA	Analysis of Variance
ANSORP	Asian Network for Surveillance of Resistant Pathogens
Avg	Average
BCH	Bacteriocin Cluster Hotspot
BHI	Brain-Heart Infusion
BIGSdb	Bacterial Isolate Genome Sequence Database
BLAST	Basic Local Alignment Search Tool
blp	Bacteriocin-Like Peptides
bp	Base Pair(s)
CC	Clonal Complex
CD-Search	Conserved Domain Search
CDD	Conserved Domain Database
CDS	Coding Sequence(s)
CI	Competitive Index
CO_2	Carbon Dioxide
COG	Cluster of Orthologous Genes
cps	Capsular Polysaccharide Synthesis
CSV	Comma-Separated Value
OD	Optical Density
cib	Competence-Induced Bacteriocin
CSP	Competence Stimulating Peptide

CSV	Comma-Separated Value
DLV	Double-Locus Variant
DNA	Deoxyribonucleic Acid
EARS-Net	European Antimicrobial Resistance Surveillance Network
eggNOG	Evolutionary Genealogy of Genes: Non-Supervised Orthologous Groups
GC	Guanine-Cytosine
GEO	Gene Expression Omnibus
GFF	General Feature Format
ICE	Integrative and Conjugative Element
IPD	Invasive Pneumococcal Disease
iTOL	Interactive Tree of Life
kb	Kilo Base
LPSN	List of Prokaryotic Names with Standing in Nomenclature
MGE	Mobile Genetic Element
MIC	Minimum Inhibitory Concentration
MLST	Multi-Locus Sequence Typing
MMR	Mismatch Repair
mRNA	Messenger RNA
n	Number
NCBI	National Center for Biotechnology Information
ONS	Office of National Statistics
ORF	Open Reading Frame
oriC	Origin of Replication of the Chromosome
OE-PCR	Overlap Extension Polymerase Chain Reaction
PBP	Penicillin Binding Protein

PBS	Phosphate-Buffered Saline
PCR	Polymerase Chain Reaction
PCV	Pneumococcal Conjugate Vaccine
PICI	Phage-Inducible Chromosomal Island
PMEN	Pneumococcal Molecular Epidemiology Network
PPV	Pneumococcal Polysaccharide Vaccine
PRCI	Phage-Related Chromosomal Island
Prokka	Rapid Prokaryotic Genome Annotation
RAST	Rapid Annotation using Subsystem Technology
RBC	Red Blood Cells
rMLST	Ribosomal Multi-Locus Sequence Type
rpm	Revolutions Per Minute
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
SaPI	*Staphylococcus aureus* Pathogenicity Island
SLV	Single-Locus Variant
SpyCI	*Streptococcus pyogenes* Phage-Like Chromosomal Island
SD	Standard Deviation
ST	Sequence Type
STRING	Search Tool for the Retrieval of Interacting Proteins
TLV	Triple-Locus Variant
THY	Todd-Hewitt Broth Supplemented with Yeast Extract
WCV	Whole-Cell Vaccine
WHO	World Health Organisation

1. General introduction

1.1 An overview of the *Streptococcus* genus

The term streptococcus (from Greek *streptos*, “chain”, and *kokkos*, “berry”) was first introduced by Albert Theodor Billroth in 1877 while working with skin and wound infections to describe bacteria with the morphological resemblance to a string of beads. The genus S*treptococcus* belongs to the family Streptococcaceae, within the order Lactobacillales (lactic acid bacteria), in the phylum Firmicutes [1]. At the time of writing, according to the List of Prokaryotic names with Standing in Nomenclature (LPSN) [2], 129 species and 23 subspecies are recognised in this genus (September 2019; http://www.bacterio.net). Most streptococcal species establish commensal relationships with their respective hosts and are often associated with the normal flora, predominantly of the skin, mouth, intestine and upper respiratory tract of warm-blooded animals [1]. However, several of these species are opportunistic human and domestic animal pathogens and have shown remarkable adaptability to new host species, resistance to antibiotics and immunological challenges. As a result, they have caused significant morbidity and mortality worldwide, leading to a substantial health and economic burden [1, 3-9].

Members of the *Streptococcus* genus are non-motile, non-sporulating, low guanine-cytosine (GC) content, Gram-positive bacteria. They can be classified into alpha, beta, or gamma groups based on their haemolytic properties on red blood agar. Alpha-haemolysis is characterised by a greenish halo around the colonies due to a partial lysis of red blood cells (RBCs). Beta-haemolysis represents a complete breakdown of RBCs, which results in a clearing of the blood around colonies and thus the area appears lightened (yellow) and transparent. The beta-

haemolytic streptococci are further divided into lettered groups by the type of carbohydrate present in the cell wall, a classification system known as the Lancefield grouping. Lastly, if a streptococcal isolate displays no haemolysis at all, it is considered to be gamma-haemolytic [1]. However, the use of biochemical and serological assays can be imprecise, as atypical haemolytic activity is sometimes observed and Lancefield groups do not always correlate with the species [1, 10].

From a medical perspective, the most important groups are the alpha and beta-haemolytic streptococci. Notable examples include *Streptococcus pneumoniae* (also known as the 'pneumococcus' and which is alpha-haemolytic), a leading cause of bacterial pneumonia, bacteraemia, and meningitis [3]; *Streptococcus pyogenes* (beta-hemolytic group A streptococci or 'GAS'), which ranks among the main causes of mortality from bacterial infections worldwide, contributing to sequelae such as rheumatic heart disease [5]; and *Streptococcus agalactiae* (beta-hemolytic group B streptococci, or 'GBS'), the most common cause of neonatal sepsis [6]. There are also other alpha and beta-haemolytic species such as *Streptococcus suis* and *Streptococcus equi* that rarely cause disease in humans but are important pathogens of domesticated animals [1, 4].

Recent advances in molecular biology means that the conventional microbiological and biochemical methods for classification of streptococcal isolates have been mostly superseded by more sophisticated DNA-based techniques. At the moment, streptococci are broadly divided into distinct groups based on their 16S ribosomal DNA sequences, namely the "Pyogenic", "Mitis", "Anginosus", "Bovis", "Mutans" and "Salivarius" groups (Table 1). Nonetheless, the majority of the *Streptococcus* species currently remain ungrouped [1, 10].

Table 1.1 - The *Streptococcus* genus main subgroups based on 16S rRNA sequence phylogeny [10]**.**

Subgroups	***Species***
Mitis	*Streptococcus pneumoniae* *Streptococcus mitis* *Streptococcus oralis* *Streptococcus pseudopneumoniae* *Streptococcus infantis* *Streptococcus cristatus* *Streptococcus gordonii* *Streptococcus sanguinis*
Pyogenic	*Streptococcus pyogenes* *Streptococcus agalactiae* *Streptococcus dysgalactiae* *Streptococcus iniae* *Streptococcus uberis* *Streptococcus equi*
Anginosus	*Streptococcus anginosus* *Streptococcus intermedius* *Streptococcus constellatus*
Salivarius	*Streptococcus salivarius* *Streptococcus thermophilus* *Streptococcus vestibularius*
Bovis	*Streptococcus bovis* *Streptococcus gallolyticus* *Streptococcus infantarius* *Streptococcus equinus*
Mutans	*Streptococcus mutans* *Streptococcus sobrinus* *Streptococcus downeii* *Streptococcus rattus* *Streptococcus cricetus*

Different streptococcal species can share an ecological niche and presumably compete with each other for resources. For instance, *S. pneumoniae*, *S. mitis* and *S. oralis* are frequently found to cohabitate in the respiratory tract of the human host [11]. Interestingly, the competition can also occur between different strains of the same streptococcal species [12, 13]. The outcome of such competition not only determines which bacterial isolates will establish within a particular niche, but also likely affects disease progression. For example, clinical studies have revealed that patients colonised by *S. oligofermentans* have a lower risk of dental caries resulting from *S. mutans* [14].

Furthermore, bacterial gene transfer is not uncommon among *Streptococcus* strains, whereby virulence factors and other important genes can be exchanged between species that share an environment [15]. For instance, streptococci of the mitis group, such as *S. mitis*, *S. oralis* and *S. infantis*, are frequent exchange partners of *S. pneumoniae*. Comparative sequence analyses of these streptococcal species suggest that nearly 30% of the sequence variability of *S. pneumoniae*, including pathogenicity genes, could be attributed to genetic exchange with *S. mitis* [16, 17].

Taken together, streptococci are intriguing subjects for studying bacterial competition, gene transfer, and the effects of viruses in bacterial fitness. The wide availability of affordable genome sequencing and policies around data sharing have led to a large number of freely-available genomes in public databases such as GenBank (Figure 1.1), ushering in an era of considerable excitement among biologists [18]. As will be shown in the subsequent chapters of this thesis, many

new types of experiments and analysis can now be conducted that were previously unfeasible.

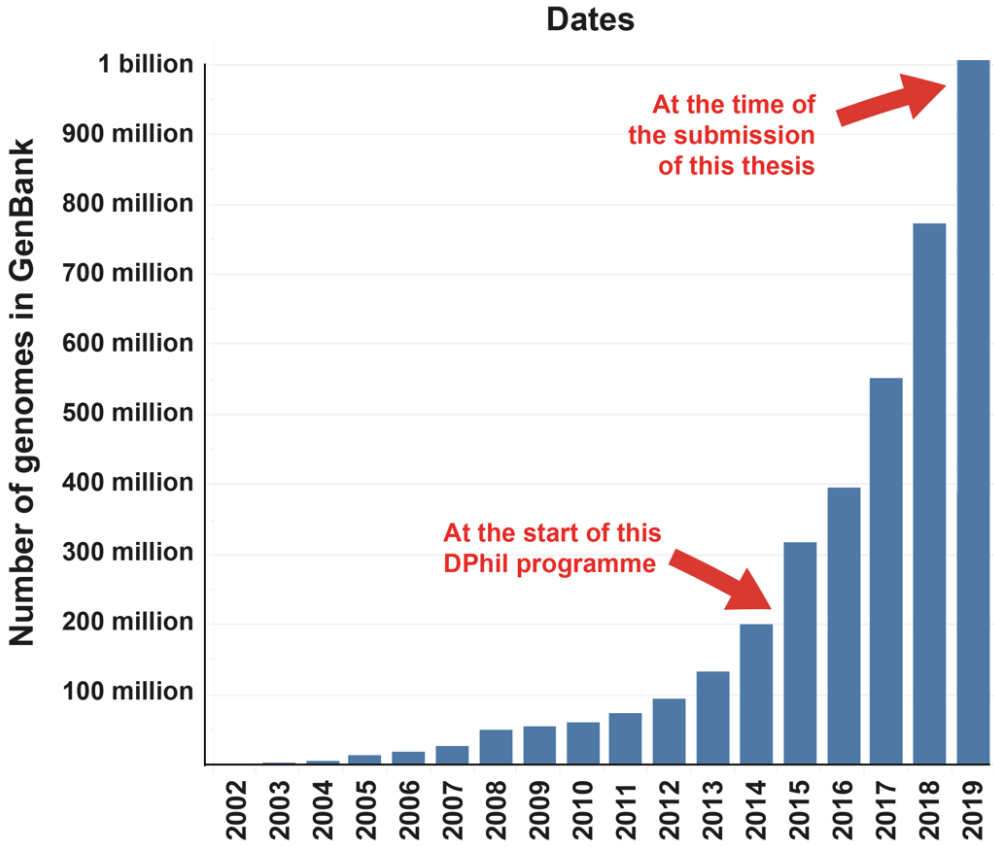


Figure 1.1 – Growth of genome sequences in GenBank from 2003 - 2019. Data collected in September 2019.

1.2 The pneumococcus

1.2.1 A brief history

S. pneumoniae was isolated separately by George Sternberg and Louis Pasteur for the first time in 1881. Pasteur named the organism *Microbe septicemique du salive* and Sternberg termed it *Micrococcus pasteuri*. In 1886, Fraenkel demonstrated that the organism Pasteur and Sternberg has identified caused pulmonary disease, which led to the term pneumococcus to be identified with this

organism. The term pneumococcus still perseveres to this day. The bacterium was later renamed *Diplococcus pneumoniae* in 1920 as isolates from pneumonia appeared as pairs of cells. It was not until 1974 that the organism was given its current scientific name, *Streptococcus pneumoniae*, based on its chain-like growth in liquid media [19]. Since then, the studies of this clinically-important bacterium have resulted in many fundamental contributions to biology and medicine, such as the identification of polysaccharides as antigens [20], bacterial gene transfer [21], the discovery of DNA as the genetic material [22], and the therapeutic use of penicillin [23].

1.2.2 Identification and epidemiological characterisation

The conventional methods used to identify pneumococci in clinical microbiology laboratories rely on biochemical and phenotypic characteristics. The pneumococcus is alpha-haemolytic under aerobic conditions but exhibits beta-haemolytic properties under anaerobic conditions [24]. Similar to other streptococci, the pneumococcus is facultatively anaerobic, oxidase- and catalase-negative, Gram-positive and appears as pairs or chains under the microscope. Unlike most other members of the genus, *S. pneumoniae* is both optochin-susceptible and soluble in bile [25, 26]; however, these assays are not strictly reliable as not all pneumococci are bile-soluble and optochin-resistant strains of pneumococci have been reported [26, 27].

Multilocus sequence typing (MLST) is a highly discriminatory typing strategy introduced by Maiden and colleagues in 1998 [28] and has since become the gold standard for genotyping various bacterial pathogens. MLST is used worldwide for microbial epidemiological surveillance, studies of bacterial population structure and

evolution. MLST allows the characterisation of pneumococcal isolates based on approximately 500 base-pair (bp) internal fragments of seven housekeeping genes (loci) throughout the genome. The genes used in the pneumococcal MLST scheme are: *aroE* (shikimate dehydrogenase), *gdh* (glucose-6-phosphate dehydrogenase), *gki* (glucose kinase), *recP* (transketolase), *spi* (signal peptidase I), *xpt* (xanthine phosphoribosyltransferase), and *ddl* (D-alanine-D-alanine ligase). Every unique sequence for each of the MLST loci is assigned an allele number, and each unique combination of seven allele numbers are assigned a sequence type (ST) [29].

Related STs can be assigned to clonal complexes (CCs) according to shared MLST alleles using the eBURST [30] and goeBURST [31] algorithms. CCs are defined as groups of closely-related STs that share six alleles with at least another ST in the group, which have likely emerged from a recent common ancestor (known as the group founder). Within each CC, the group founder is assigned as the ST that is linked to the highest number of single-locus variants (SLVs, defined as STs that differ by only a single allele). Subsequently, SLVs of each ST that is a SLV of the group founder (*i.e.* double-locus variants (DLVs) of the group founder) are also assigned, followed by triple-locus variants (TLV) of the group founder, and so forth. In this way, a network of related isolates is established, which can be represented diagrammatically using the goeBURST algorithm implemented in Phyloviz software [32], where the STs are displayed as nodes and the SLVs are linked by edges. The node size is proportional to the frequency of isolates assigned to that ST. STs that cannot be assigned to any group are referred to as singletons [33].

There are >100 MLST genotyping schemes and the alleles, ST designations, and

isolate metadata (*i.e.* provenance data) for each scheme are stored in the PubMLST online databases. Genome sequences for individual isolates may also be uploaded. At the time of writing, there were nearly 15,000 unique STs, over 45,000 isolate records and nearly 14,000 genome sequences in the pneumococcal PubMLST database, and new data are continuously submitted by many laboratories around the world (September 2019; http://pubmlst.org/spneumoniae).

1.2.3 General biology

The pneumococcus is commonly an asymptomatic coloniser of the nasopharynx in healthy individuals. Although this organism is primarily a commensal, it can transition to a pathogenic form, and young children (<5 years), the elderly (75+ years), and immunocompromised persons are most at risk of pneumococcal disease [34]. Despite its pathogenic potential, the pneumococcus is not related to other overtly pathogenic *Streptococcus* species, but falls in the Mitis group, members of which are generally deemed as prototypes of commensal bacteria, such as *S. mitis* and *S. oralis* (Table 1) [1]. Recent investigations suggest that *S. pneumoniae*, *S. mitis*, and *S. pseudopneumoniae* descended from a pneumococcus-like organism that was probably pathogenic to a humanlike ancestor [35, 36].

In general, pneumococci exist within the nasopharynx by adhering to the mucosa, and over longer periods this means of colonisation might develop into a biofilm (colonies embedded in a matrix composed of various polymeric substances). A biofilm structure can protect the pneumococci from host defenses and other environmental adversities [37, 38]. The duration of the nasopharyngeal colonisation period may be short or long (*i.e.* days or weeks) and the pneumococci

are presumed to reside asymptomatically during the period of colonisation [12, 39]. Factors (largely unknown) can also trigger the pneumococci to invade through the epithelial surface and enter normally sterile sites like the blood or cerebrospinal fluid; this enables rapid, widespread dissemination of the bacteria and could lead to life-threatening illnesses like bacteraemia or meningitis [37, 38].

Once pneumococci gain access to other sites of the body and if not cleared by the host immune response, they can cause a variety of diseases, the most serious of which are life-threatening infections like pneumonia, meningitis and septicaemia. Otitis media, sinusitis and conjunctivitis are also common infections caused by pneumococcus and although not life-threatening, they are significant causes of morbidity worldwide [3, 25, 38, 40]. Although the genetic basis for the pathogenicity and virulence of the pneumococcus is not yet fully understood, the polysaccharide capsule has long been recognised as a major virulence factor. As of today, almost 100 different capsule types (serotypes) have been identified, and certain serotypes are more frequently associated with disease and others with asymptomatic carriage [41].

Since colonisation is the initial stage of the disease process, asymptomatic carriage is considered an important risk factor for serious pneumococcal disease [42]. The pneumococcal carriage rate is particularly high in the paediatric population and children are the main reservoir for disease transmission. Pneumococcal colonisation arises very early in life and diminishes with increasing host age. Neonates can be colonised in the first days after birth, while all children harbour pneumococci at some point by the time they turn one [43-45].

Pneumococcal acquisition rates were found to be to two to six times higher in children less than nine years compared to adults 17–39 years [46].

Pneumococcal carriage starts to decline rapidly at the age of ten, and despite historical studies reporting carriage in nearly half of all adults at the beginning of the 20th century [47, 48], carriage in adults and elderly is now seldom detected [48-51]. Intriguingly, despite the low carriage rate among older adults, the incidence of pneumococcal disease is particularly high in this age group. The exact pathophysiology in this high-risk group is poorly understood; however, it is widely suggested that it may be related to immunosenescence, the gradual deterioration of the immune system caused by the aging process [52]. Notably, adults in close contact with infected children have higher pneumococcal colonisation rates, and thus an increased risk of developing pneumococcal infection [42].

The pneumococcal carriage rate is disproportionately higher in low income countries compared to industrialised countries, particularly in Africa and Asia, where the highest incidence of pneumococcal disease occurs [53] (see 1.2.6). In the Gambia, pneumococcal carriage rates were reported to be 80% in children under five years of age and 20% in adults [54]. In a recent meta-analysis from 57 studies on pneumococcal carriage in African countries, Usuf and colleagues [55] estimated the carriage rate to be 63% in children less than five years and 43% in children between the ages of five to 15. Similarly, high pneumococcal carriage rates have been reported in several developing countries in Asia such as Pakistan, the Philippines, and Bangladesh [43, 56, 57], as well as among certain indigenous populations [58, 59]. On the other hand, the pneumococcal carriage rate appears

to be lower in industrialised countries. For instance, in Sweden, only 12% of infants harboured pneumococci at three months, 30% at seven months of age, and 32% at 12–18 months [60]. Likewise, in the Netherlands, rates of pneumococcal carriage were reported to be 8% among infants, 31% at six months, and 44% at two years of age [61].

Pneumococcal carriage duration can vary from days to months and is strongly influenced by the serotype of the infecting strain [39, 62, 63]. Furthermore, co-colonisation with several genetically different strains as well as loss and acquisition of strains over time are common features of the pneumococcal carriage state [64, 65]. Nasopharyngeal competition can influence colonisation dynamics, strain prevalence, serotype distributions and, in turn, the potential for disease progression [12]. Intraspecific competition is believed to be (at least in part) mediated by bacteriocins, which are small antimicrobial peptides produced by many bacterial species to inhibit the growth of other bacteria in the same ecological niche. Bacteriocins are a major focus of this thesis and will be discussed in detail in Chapter 1.3.

1.2.4 Interspecific competition between the pneumococcus and other bacterial species

The human nasopharynx is a reservoir to a diverse range of bacterial species from many genera including *Streptococcus, Staphylococcus, Haemophilus, Moraxella* and *Neisseria.* Studies have shown that hydrogen peroxide secreted by the pneumococcus is able to inhibit the growth of a variety of organisms such as *Staphylococcus aureus, Haemophilus influenzae, Neisseria meningitidis and Moraxella catarrhalis* in *in vitro* settings [66, 67]. The prevalence of the

pneumococcus is reported to be negatively correlated with *S. aureus* carriage, which further suggests a competitive interaction between these two species [68]. Furthermore, to promote clearance of its competitors, the pneumococcus expresses a neuraminidase that can remove the sialic acid from the capsules of *N. meningitidis* and *H. influenzae*, which would otherwise protect these respiratory pathogens from the host immune system [69]. Conversely, *in vivo* studies suggest that *H. influenzae* can stimulate a neutrophil-mediated killing of the pneumococcus by the host [70]. Pneumococci and *H. influenzae* were found to be negatively correlated in children during infection; however, this correlation changes to positive when *M. catarrhalis* is also present, suggesting complex biological relationships within the nasopharynx that are yet to be fully elucidated [71].

1.2.5 Global disease burden

S. pneumoniae is a major public health problem globally, leading to approximately 14.5 million cases of invasive pneumococcal disease (IPD) such as meningitis, septicaemia and pneumonia [72]. *S. pneumoniae* was responsible for over one million deaths globally in 2016, the majority of which occurred in low income countries, and Africa and Asia have the highest burden of mortality [53, 72]. Pneumococcal meningitis leads to mortality rates between 16 and 37%, and up to half of the children surviving pneumococcal meningitis suffer permanent sequelae [73, 74]. Unsurprisingly, the economic burden of pneumococcal disease is substantial. For example, in the United States alone, the cost of pneumococcal disease was estimated at nearly $3.5 billion in 2004, of which 72% accounted for IPD and 16% for acute otitis media and sinusitis [75, 76]. Nearly 5,800 hospitalisations and 700,000 GP consultations in England and Wales are

estimated to occur annually due to pneumococcal disease [77]. According to data from the Office of National Statistics (ONS), approximately 50% of hospitalisations for pneumococcal disease in these two countries occur among individuals aged 65 years and older [77]. Because the elderly are a major high-risk group for pneumococcal disease, the ageing human population, especially in developed countries, constitute a serious challenge that could lead to an increase in the health and economic burden for pneumococcal diseases in the future [78].

1.2.6 Vaccines

1.2.6.1 Early vaccines

Attempts to develop a vaccine against pneumococcus began as early as the 1880s. George Sternberg, one of the first scientists studying the pneumococcus (see 1.2.1), was first to show that rabbits injected with dead pneumococci were protected against infection in following injections [79]. This observed protection in animals formed the basis for the development of vaccine in humans. In 1911, British physician Sir Almroth Wright and his colleagues began a series of clinical trials in order to test the first pneumococcal vaccine administered to humans [80]. The vaccine consisted of dead pneumococci (a whole-cell vaccine; WCV) and was administered to several thousand gold miners in South Africa, a specific population of individuals who suffered from a high burden of morbidity and mortality caused by pneumococcal pneumonia. The vaccine led to a reduction in the cases of pneumococcal diseases during the four months following vaccine administration, but protection was lost over time [80]. Development and evaluation of WCVs continued sporadically without major success; the serotype-specificity of pneumococcus was realised in 1940s and shortly thereafter WCVs were replaced

by serotype-specific polysaccharide vaccines [81].

1.2.6.2 Polysaccharide vaccines

Pneumococcal polysaccharide vaccines (PPVs) are composed of purified preparations of pneumococcal capsular polysaccharide. The first successful PPV was tested on American airmen during the Second World War, where it was found to reduce the incidence of vaccine-type pneumococcal pneumonia by 85% [82]. The success of this trial led to the development of two hexavalent vaccines that were made commercially available in the United States in 1946: one for adults (serotypes 1, 2, 3, 5, 7 and 8), and one for children (serotypes 1, 4, 6, 14, 18 and 19). After the introduction of penicillin in the early 20th century, the mortality caused by the pneumococcus fell significantly, which led to a temporary loss of interest in developing pneumococcal vaccines and consequently, these vaccines were withdrawn in 1954 [81, 83].

Interest in vaccines re-emerged following the rise of antibiotic-resistant pneumococci in 1960s (see 1.2.8.1). A 14-valent PPV was licensed for use in the United States in 1977. This vaccine was expanded to a 23-valent PPV (PPV-23) in 1983 that contained serotypes 1, 2, 3, 4, 5, 6B, 7F, 8, 9N, 9V, 10A, 11A, 12F, 14, 15B, 17F, 18C, 19F, 19A, 20, 22F, 23F and 33F, which is still in use [84, 85]. However, despite the introduction of PPV-23, the incidence of disease amongst children remained mostly unchanged [86-88]. This is because children under five are not able to produce an anamestic immune response to the plain polysaccharides contained in the vaccine [89, 90]. As a result, PPV-23 is currently only recommended in the elderly and those who are at higher risk of complications from pneumococcal disease, such as immunocompromised

adults [91]. Since PPV-23 was shown to be ineffective in protecting the paediatric population, the main reservoir for pneumococcal disease transmission, this led to the development of conjugate vaccines which are capable of eliciting a protective immune response in children [81].

1.2.6.3 Conjugate vaccines

Pneumococcal conjugate vaccines (PCVs) are composed of purified pneumococcal capsular polysaccharide covalently linked to a carrier protein, which enhances the recipient's immune response to the vaccine. The idea of a conjugate vaccine initially emerged due to experiments in rabbits in 1927, when the protective immune response to the pneumococcal serotype 3 polysaccharide was found to be enhanced by attaching the polysaccharide antigen to a horse serum globulin [92, 93]. The first use of this technology for vaccines in humans was for the capsulated *H. influenzae* type b (Hib) in the 1990s, which proved to be highly effective in significantly reducing Hib carriage and disease in infants [94]. The outstanding success of the Hib vaccine provoked interests in developing an equivalent conjugate vaccine for the pneumococcus.

The first PCV, named Prevnar (US) or Prevenar (Europe), was a 7-valent vaccine licenced for use in infants in the United States in 2000 [95]. Since one limitation of conjugate vaccine technology is the number of serotypes that can be incorporated in the vaccine formulation, PCVs have tended to include serotypes that are associated with IPD and/or are highly resistant to antimicrobials. PCV7 contained seven capsular polysaccharides (serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F), which were attached to the inactivated diphtheria toxin CRM197, a highly immunogenic but non-toxic carrier protein. This combination proved to be capable

of eliciting a protective immune response in children <5 years of age [95]. Since then, two additional PCVs have been licenced, which have now replaced PCV7: PCV10 (PCV7 serotypes, plus 1, 5 and 7F) and PCV13 (PCV7 serotypes, plus 1, 3, 5, 6A, 7F and 19A) [81]. These two vaccines were developed simultaneously by two pharmaceutical companies: PCV10 was developed by GlaxoSmithKline and introduced in November 2009, while PCV13 was developed by Wyeth Pharmaceuticals (now acquired by Pfizer) and introduced in May 2010. As of April 2019, PCV10 or PCV13 are in use in national immunisation programmes in 142 countries and have successfully reduced the burden of morbidity and mortality caused by the pneumococcus [96, 97].

1.2.6.4 Impact of vaccination on the pneumococcal population

Although pneumococcal vaccines are safe and highly effective in reducing the incidence of pneumococcal disease, vaccine use has led to a change in the serotype distribution among pneumococcal populations [98-103]. As a result, there has been an increased prevalence of colonisation by non-vaccine pneumococcal serotypes, many of which are now responsible for causing severe illness. The emergence and spread of non-vaccine serotypes can be attributed to two phenomena: serotype replacement and capsular switching.

Serotype replacement reflects an expansion of non-vaccine serotypes due to the removal from the population of vaccine serotypes that compete with them. By targeting a small subset of serotypes, PCVs may create a vacant ecological niche that will be filled by pneumococcal serotypes not covered by the vaccine. Serotype replacement could arise by either an increase in the prevalence of previously rare non-vaccine serotypes present in the population or the emergence and spread of

non-vaccine serotypes previously absent from the population that were unable to compete with the vaccine serotypes [98,101,102].

The other phenomenon, capsular switching, occurs via the horizontal transfer of capsular genes among colonising strains by the processes of transformation and recombination. Through this process, a strain of *S. pneumoniae* may acquire capsular genes that lead to the expression of a serotypically different capsule [98-103]. The switching of capsular genes from a vaccine serotype to a non-vaccine serotype allows for the evasion of vaccine-induced immunity and provides an opportunity for the emergence of vaccine escape genetic lineages [103].

Capsular switching events are a regular occurrence among pneumococcal populations. Wyres *et al.* [103] investigated 426 pneumococcal isolates recovered from 1937 through 2007 and identified 36 independent capsular switching events, which in some cases extended to the nearby penicillin-binding protein (PBP) genes (which when altered confer penicillin resistance, see 1.2.8.2). Such recombinational events not only enable vaccine escape but also facilitate the spread of antimicrobial resistance determinants among pneumococci, which is a major concern worldwide.

1.2.7 Antimicrobials and resistance

Several families of antimicrobials can be used against pneumococcal infections. The most commonly used are beta-lactams and macrolides, followed by fluoroquinolones [104]. However, with the widespread availability and overuse of antimicrobials, antimicrobial resistant pneumococci have become prevalent and rates of resistance are increasing at an alarming pace, which will be discussed in the subsequent sections.

1.2.7.1 A brief historical perspective

During the pre-antibiotic era, mortality associated with bacteraemic pneumococcal infections was nearly 80% [105]. After the introduction of sulphonamides and penicillin in the early 20th century, the mortality caused by this pathogen fell significantly [106]. For the next few decades, the pneumococcus was considered to be entirely susceptible to antimicrobials, which led to temporary loss of interest in this organism as a pathogen during the middle of the 20th century [107].

The first reports of penicillin non-susceptible pneumococci emerged in 1967 [108] and by the turn of the century, antimicrobial resistance among pneumococci had become a global health concern. Penicillin non-susceptible pneumococci spread rapidly to many countries around the world, including to Australia, Papua New Guinea, Spain, South Africa, Taiwan and the United States [19, 109-113]. Of particular concern was the emergence of multidrug-resistant pneumococci, which are strains resistant to penicillin plus at least two additional classes of antimicrobials. The first multidrug-resistant pneumococcal strain appeared in South Africa in 1977, which was found to be resistant to penicillin, tetracycline, erythromycin, clindamycin, trimethoprim-sulfamethoxazole, and chloramphenicol [114]. Over the next few years, multiresistant pneumococci were observed in the United Kingdom, United States, and continental Europe [19, 110, 115, 116].

The Pneumococcal Molecular Epidemiology Network (PMEN) was established in 1997 when it became clear that highly similar genetic lineages of resistant pneumococci were circulating in more than one country, thus the aim of PMEN was to identify, characterise and name major antimicrobial-resistant lineages of pneumococci. Later it was realised that antimicrobial-susceptible lineages also

circulated widely and these were also included in the PMEN collection [117]. PMEN currently recognises 43 antimicrobial-susceptible and -resistant globally-distributed lineages (July 2019; http://pneumogen.net/pmen).

In the most recent report on global antimicrobial resistance, released by the World Health Organization (WHO) in 2017, the pneumococcus was considered as one of the twelve bacteria for which new antibiotics are urgently needed [118]. The rates of resistance vary widely between different type of antimicrobials and geographic area, which will be discussed in the following sections. What follows is a brief introduction to their mode of action and mechanisms and the extent of resistance in different geographical regions.

1.2.7.2 Beta-lactams

Beta-lactam antimicrobials include the penicillins and cephalosporins and have a beta-lactam ring as their core chemical structure. The mechanism of action is to inhibit cell wall synthesis by binding to penicillin-binding proteins (PBPs) [119]. There are six PBPs in pneumococci, three of which (PBP2X, PBP1A and PBP2B) are the most important with respect to the development of resistance. Beta-lactam resistance is the result of genetic modifications in the binding site of these enzymes, which reduces their affinity for beta-lactam antibiotics [120, 121]. The genes that encode altered PBPs may be acquired by recombination events with other pneumococci or with related streptococcal species, such as *S. mitis* and *S. oralis* [122, 123]. Alterations in other genes, such as those in the *murMN* operon [124], *ciaRH* operon [125] and the *clpL* locus [126], have also been shown to be important for the acquisition of high-level beta-lactam non-susceptibility in the pneumococcus.

Beta-lactam non-susceptibility rates vary widely between different geographical regions. The European Antimicrobial Resistance Surveillance Network (EARS-Net) is a European wide network of national surveillance systems, monitoring antimicrobial resistance in important pathogens. The EARS-Net collects antimicrobial resistance data from invasive pneumococcal isolates obtained from both children and adults. A 2017 surveillance study by the EARS-Net reported that the rates of penicillin non-susceptibility ranged from 0.2% in Belgium to 45% in Greece [127].

The SENTRY program in the United States is a network of hospitals tracking antimicrobial susceptibility of pathogens. Nearly 19,000 pneumococcal isolates were recovered from individuals of all ages in medical centres across the country from 1998 to 2011. Nearly three-fourths of all samples were isolated from carriage and the remaining isolates were from invasive disease. The results of this surveillance study indicated an increase in the rate of penicillin non-susceptibility from 3% in 1998 to 15% in 2011 [128].

The Canadian Bacterial Surveillance Network performed a surveillance study whereby a total of 26,081 pneumococcal isolates were recovered from people of all age groups in 119 medical laboratories across Canada from 1998 to 2009 [146]. Of the recovered samples, 10,127 (39%) were isolated from blood or other sterile sites, 10,500 (41%) were recovered from nonsterile respiratory sites, and the remaining 5,454 (20%) were either from an unknown source or recovered from other sites such as eye or ear swabs. During this time period, the rate of penicillin non-susceptibility increased from 6% in 1998 to 8.3% in 2009.

Penicillin non-susceptibility is also common in Asia [129, 130]. The Asian Network for Surveillance of Resistant Pathogens (ANSORP) performed a surveillance study whereby 685 clinical pneumococcal isolates were recovered from patients of all ages with pneumococcal disease from 14 centres in 11 Asian countries from 2000 to 2001. Nearly 50% of all tested isolates were non-susceptible to penicillin, with non-susceptibility rates ranging from 8% in Vellore, India to 92% in Ho Chi Minh City, Vietnam [130].

Multi-centre studies of pneumococcal antimicrobial susceptibilities in Africa are scarce, albeit penicillin non-susceptible pneumococci have been reported to be prevalent and on the rise in several African countries. Emgard and colleagues investigated pneumococci isolated at six different health facilities in northern Tanzania between 2013 to 2015 from carriage in children less than 2 years of age. During this period, penicillin non-susceptibility in the recovered isolates increased significantly from 31% in 2013, to 47% in 2014 and 53% in 2015 [131]. Annual surveys were conducted in Kenya during 2009 and 2010 among children aged <5 years, whereby 657 pneumococci were isolated from carriage and tested for antibiotic susceptibility. The authors reported nonsusceptibility to penicillin in nearly 82% of the isolates [132]. Ginsburg *et al* [133] conducted a meta-analysis of 74 smaller studies from 1978 to 2011 that provided data on nearly 42,000 invasive and carriage pneumococcal isolates recovered from individuals of all ages in 29 African countries. The results of the meta-analyses indicated that ~19% of pneumococcal isolates were not susceptible to penicillin [133].

Penicillin non-susceptibility is also a concern in Latin America. Pneumococci isolated between 1997 to 2001 from patients of all ages with disease residing in

seven different Latin American countries were investigated. The authors reported high rates of penicillin non-susceptibility that ranged from 22% in Brazil to 67% in Mexico [134].

1.2.7.3 Macrolides

Macrolides are broad-spectrum antibiotics that exert their antibacterial action by binding to the bacterial 50S ribosomal subunit, inhibiting the initiation of messenger RNA (mRNA) binding and thereby halting protein synthesis in bacteria [135]. Macrolides are a good alternative to beta-lactams due to their relatively low toxicity and good tolerability; however, the widespread emergence of macrolide-resistant pneumococci is problematic in some parts of the world. Macrolide resistance in pneumococcus is mediated by two major mechanisms. One involves acquisition of *ermB* (erythromycin-resistance methylase), which blocks the binding of the macrolide to the ribosome by post-transcriptional modification of the ribosomal RNA. The other mechanism involves acquisition of a macrolide efflux system, encoded by *mefE* or *mefA* (macrolide efflux), which pumps out the antibiotic from inside the bacterial cell to maintain a concentration below that required for binding to ribosomes [136].

Macrolide resistance among pneumococci is geographically variable but surveillance studies have revealed an increase in resistance rates worldwide. The results of the 14-year longitudinal surveillance study by the SENTRY program in the United States (described in 1.2.8.2) indicated a striking increase in the rates of macrolide non-susceptibility, from 18% in 1998 to 45% in 2011 [128]. A 2017 surveillance study by the EARS-Net reported that the rates of macrolide non-susceptibility in Europe ranged from 4% in Denmark to

37% in Malta [127]. For most European countries, macrolide non-susceptibility was more frequent than penicillin non-susceptibility; however, non-susceptibility to both penicillin and erythromycin was also fairly common, with several countries such as Cyprus, Malta, and Romania reporting this phenotype for over 20% of the tested isolates [127]. According to ANSORP, in Asian countries, the rates of macrolide resistance have increased from 46% in 1996-1997 to 73% in 2008-2009, with a particularly high prevalence in China (96%) and Taiwan (85%) [129].

1.2.7.4 Fluoroquinolones

Fluoroquinolones are a family of broad-spectrum antimicrobials that have been in clinical use for treatment of various infections since the late 1980s [137]. Ciprofloxacin, introduced in 1987, was the first member of the fluoroquinolone class of antimicrobial to be widely available. The 1990s saw the release of many other subsequent fluoroquinolones, such as ofloxacin (1991), enoxacin (1992), lomefloxacin (1992), levofloxacin (1997), sparfloxacin (1997) and moxifloxacin (1999). Despite their potency, broad spectrum of activity and oral bioavailability, many newer fluoroquinolones were swiftly withdrawn from worldwide markets due to severe adverse effects in patients. Notwithstanding rare side effects, due to their effectiveness, several fluoroquinolones are still in use today, with ciprofloxacin, levofloxacin and moxifloxacin being the most commonly prescribed [137, 138].

Fluoroquinolones exert their antimicrobial action by targeting and inhibiting GyrA and ParC proteins of the DNA gyrase and topoisomerase IV, respectively, both of which are essential enzymes required for bacterial DNA replication. Resistance to fluoroquinolones in pneumococcus is predominantly achieved by mutations in

the *gyrA* and *parC* genes [137, 139]. Resistance can occur rapidly during the course of therapy, and there are numerous documented reports of fluoroquinolone treatment failures due to pneumococcal isolates with the so-called “first step” and “second step” mutations [140, 141]. Resistance typically arises in two stages: a first-step mutation in a primary target, which decreases the binding affinity to that target and thus reduces fluoroquinolone activity, followed by a second-step mutation in a secondary target, which causes a further reduction in the fluoroquinolone activity. To avoid therapeutic failure, the detection of strains with first-step mutations is important, as these strains have a higher likelihood of subsequent mutations that could lead to the strains becoming highly resistant [140-142].

As would be expected from the selection pressure imposed by the use of broad-spectrum antimicrobials (see 1.2.9), the level of fluoroquinolone consumption is linked to the rates of resistance [129, 143, 144]. In an attempt to maintain their effectiveness, the use of fluoroquinolones is currently generally reserved for severe or resistant infections [129]. As a consequence, the prevalence of fluoroquinolone-resistance pneumococci has remained relatively low, albeit with notable geographical variation.

Data from the SENTRY surveillance program in the United States (described in 1.2.8.2) reported the rate of nonsusceptibility to levofloxacin to be 0.2% in 1998 and 1% in 2011 [128]. A 2012 surveillance study was conducted in Spain using 3,621 pneumococcal isolates recovered from both children and adults from hospitals nationwide; nearly two-thirds of all samples were isolated from invasive disease, and the remaining isolates were from carriage. The findings from the study

indicated the rates of non-susceptibility to ciprofloxacin to be 2%, identical to rates reported previously in 2002 and 2006 [145]. Similarly, the longitudinal surveillance study by the Canadian Bacterial Surveillance Network (see 1.2.8.2) indicated that the rates of ciprofloxacin non-susceptibility across Canada have remained unchanged between 1998 and 2009 at less than 2% [146]. In contrast, higher levels have been documented in Asian countries. For instance, a surveillance study from Hong Kong that investigated antimicrobial susceptibility for pneumococci isolated from invasive disease between 2001 and 2007, indicated that 11% were not susceptible to levofloxacin [147]. The ANSORP investigated nearly 2,000 pneumococcal isolates recovered from patients of all ages with pneumococcal disease in 11 Asian countries from 2008 to 2009. Nearly 13% of all tested isolated were non-susceptible to ciprofloxacin, with isolates from China showing the highest non-susceptibility rates at 26% [148].

1.2.8 Need for novel antimicrobial strategies

Antimicrobial resistance in pneumococcus is of serious concern, which necessitates investigations into novel antimicrobial strategies. Many of the currently available antimicrobials for treating pneumococcal infections are broad spectrum, which affect a wide range of bacteria and thus exert selection pressure for the development of antimicrobial-resistant bacteria. Using broad-spectrum antibiotics can also have a detrimental impact on the commensal human microbiota, since they act indiscriminately on both the target and non-target organisms [149-151].

The human microbiota harbours a complex and diverse community of numerous different bacterial species, some of which provide a variety of beneficial effects to

the host [152]. Disturbances in this population are increasingly correlated with various conditions and disorders, such as immune dysregulation and inflammation. Studies have shown that some populations of bacteria do not recover after treatment with broad-spectrum antibiotics. The suspected link between the use of broad-spectrum antibiotics and an increased risk of developing atopic and autoimmune disorders is alarming [149-151]. Consequently, there is an urgent need for the development of narrow-spectrum substitutes. Two promising areas of research are the exploitation of bacteriophages (phages: viruses that infect bacteria) and bacteriocins [153, 154]. Bacteriocins and bacteriophages are the main focus of this thesis and will be discussed in detail in subsequent chapters.

1.3 Bacteriocins

1.3.1 General characteristics

Bacteriocins are ribosomally-synthesised antimicrobial peptides produced by many bacterial species to inhibit the growth of closely-related bacterial strains in the same ecological niche. Bacteriocin production is a common attribute among many bacteria that reside in polymicrobial communities [155]. These peptides can have a very broad or narrow spectrum of antimicrobial activity; some bacteriocins exhibit interspecies activity, while others are only active against other strains of the same species. Most bacteriocins are extremely potent and exert their killing activity at nanomolar concentrations [156].

Bacteriocin toxins are initially synthesised as inactive pre-peptides, with an N-terminal leader sequence and a C-terminal pro-peptide part. During the transport of the bacteriocin to the outside of the producing cells, the leader sequence is proteolytically removed, which in turn converts the inactive pro-peptide into its active

form [157]. Bacteriocin systems generally utilise dedicated ABC transporters that can export the bacteriocin without altering or destroying the peptide. The bacteriocin producer strain also harbours a dedicated immunity system that confers protection from its own toxin. The mode of action of immunity proteins is not understood, but they may protect the producing cell via scavenging bacteriocins or by serving as an antagonistic receptor to prevent the bacteriocins from binding to the membrane [157]. Bacteriocin systems are commonly encoded within “bacteriocin clusters”, whereby genes involved in toxin production, immunity, modification and transport are situated adjacent to each other in the bacterial genome [157]. The expression of genes in the bacteriocin cluster is often part of bacterial quorum-sensing mechanisms and may also be induced by other stress factors such as heat, nutrient limitations and exposure to other antibiotics [158, 159].

Bacteriocin production has been associated with more efficient colonisation of a host by the producer strain, owing to the ability of these toxins to remove competitors, and therefore, one promising approach for treatment and prevention of pathogenic bacterial species is the application of commensal bacteriocin-producing species as probiotics in order to outcompete the pathogenic ones [160]. Furthermore, bacteriocins such as nisin have been commercialised and are used globally as food preservatives [153]. Understanding the relevance of bacteriocins on the structure of the bacterial population is crucial in the context of understanding and predicting the emergence and spread of genetic lineages, as well as in assessing vaccine impact, specifically in the context of vaccine-induced changes in the overall population structure after vaccine implementation.

1.3.2 Bacteriocin classification

The nomenclature of bacteriocins is complex and has caused considerable confusion over the years. Several classification approaches are currently in place and the debates about these various schemes are ongoing [161].

The first classification scheme started shortly after the discovery of colicins as the first bacteriocins and several key properties, including the route of entry into the target cells and structural features, were used to define colicins. However, identification of subsequent new colicin-like bacteriocins revealed significant dissimilarities among these compounds and consequently, the colicin-based classification system became obsolete [162]. Later attempts involved dividing bacteriocins into eight types (I-VIII) according to various characteristics such as heat tolerance, host range, trypsin sensitivity and cross reactivity [163]. This scheme was also later abandoned and replaced by that proposed by Klaenhammer [164], who classified bacteriocins into four classes on the basis of their structure, molecular weight and thermal stability. This scheme is still in use, although another classification scheme by Cotter and co-workers was later defined and widely adopted, resulting in continued confusion within the field.

The classification scheme proposed by Cotter and colleagues broadly divides bacteriocins into two different categories; the class I “lantibiotics” (containing lanthionine) and the class II “non-lanthionine-containing bacteriocins”, both of which are further divided into different subclasses [161, 165]. However, the increasing number of newly-discovered bacteriocins has revealed the heterogeneity of these compounds and various bacteriocin families have recently emerged that are too dissimilar for classification under any of the bacteriocin

classification schema mentioned above [157].

At the moment, the most recent and widely adopted classification is arguably the one proposed by Arnison and colleagues [157], which expands on previous classification schema to encompass the majority of the bacteriocins identified to date. Accordingly, bacteriocins are classified into over a dozen distinct families based on the enzymes involved in their production and structural features [157]. Several common bacteriocin families as defined by Arnison *et al*, particularly those relevant to this thesis, are described below and what follows is a brief introduction to their general characteristics, structural features and modes of action.

1.3.3 Lantibiotics

Lantibiotics are the most extensively studied family of bacteriocins and are produced by bacteria from diverse environments. To date, nearly 100 lantibiotics have been discovered, the majority of which are secreted by Gram-positive bacteria [166]. This group of bacteriocins has gained considerable interest during the last decades due to their potent antimicrobial activity and low probability of developing resistance [166]. One of the best-known examples is nisin, which has been in use as a food preservative for over half a century in more than 80 countries [157, 167].

The name 'lantibiotic', short for 'lanthionine-containing antibiotic', is derived from their content of unusual amino acids, such as lanthionine (Lan) and methyllanthionine (MeLan). The Lan and MeLan residues result from extensive post-translational modifications by lanthionine synthetases, involving dehydration and cyclisation in the peptide backbone [157, 168].

Based on the differences between the biosynthetic enzymes that introduce the Lan

and MeLan residues, lantibiotics are categorised into four different classes (I to IV). For class I lantibiotics (which include nisin) the dehydration and cyclisation are carried out by two separate biosynthetic enzymes, a dehydratase (LanB) and a cyclase (Lan*C*). For the remaining classes of lantibiotics, dehydration and cyclisation are carried out by a single bifunctional lanthionine synthetase. These modification enzymes, named LanM (class II), LanKC (class III) or LanL (class IV), differ in their conserved domains and sequence homology among each lantibiotic cluster. Other genes required for the production of a functioning lantibiotic peptide are generally designated as follow: pre-peptide (*lanA*), protease (*lanP*), transporter (*lanT*), immunity (*lanE*, *lanF*, *lanG*, *lanH* and *lanI*), receptor histidine kinase (*lanK*) and transcriptional response regulator (*lanR*) [157, 168].

Although various modes of action have been described for lantibiotics, the majority seem to exert their antimicrobial affect by attacking the bacterial membrane. For instance, lantibiotics, such as nisin, mersacidin, and epidermin effectively bind to lipid II found in the membrane of the target bacteria. Lipid II is an essential precursor of bacterial cell wall biosynthesis and the binding of the lantibiotic to this molecule results in the inhibition of peptidoglycan synthesis. This binding may also lead to the formation of pores in the target membrane, resulting in the release of small cytoplasmic compounds and eventually cell death [157, 168, 169].

Another known mode of action of lantibiotics is inhibition of the outgrowth of germinated bacterial endospores. Members of several bacterial genera, including but not limited to the genus *Bacillus* and *Clostridium* form metabolically dormant endospores, typically as a response to cellular nutrient starvation. The endospore structure provides protection during adverse environmental conditions such as

starvation, heat, UV radiation and chemical or oxidative stress. The endospore will remain dormant until a variety of environmental stimuli (generally indicative of the return of more favourable conditions), trigger germination, allowing outgrowth into a replicative, metabolically active vegetative form [170]. Nisin has been shown to prevent the outgrowth of many *Clostridium* and *Bacillus* species. Although a clear molecular basis of lantibiotic inhibition of spore outgrowth remains to be elucidated, it is thought to involve either covalent binding to a sensitive cellular target or disruption of membrane integrity [170, 171].

1.3.4 Head-to-tail cyclised peptides

Head-to-tail cyclised peptides are a relatively large family of bacteriocins that contain a peptide bond between N- and C-termini that results in a circular molecule [172]. The precise mechanism of cyclisation and the enzymes involved in the post-translational modification of these bacteriocins remain largely unknown.

The first circular bacteriocin, Enterocin AS-48, was first reported in 1986 [173], but its circular structure only became known in 1994 [174]. Since then a number of additional circular bacteriocins have been discovered from various bacterial species, particularly Gram-positive bacteria of the phylum Firmicutes [172, 175]. Circular bacteriocins are generally broad-spectrum antimicrobials, mainly active against other Firmicutes. Antimicrobial activity against other Gram-positive phyla has also been reported, including against common food-borne pathogens such as *Clostridium* and *Listeria* species [172]. The antimicrobial action of circular bacteriocins is thought to be through pore formation in the lipid membrane of target cells, which results in leakage of ions and water and ultimately cell death [172].

The bacteriocin cluster is generally composed of five to seven genes, which are

often designated alphabetically by order of occurrence in the cluster. These include genes encoding a pre-peptide, an immunity protein, which is often hydrophobic, and ABC transporters and membrane proteins, which are involved in exporting the bacteriocin outside the producer cell [157, 172].

1.3.5 Sactipeptides

Sactipeptides are a newly discovered family of bacteriocins that possess a characteristic thioether bridge (sactionine linkage) that is added post-translationally and is absolutely required for their antimicrobial activity [157, 176, 177]. The first sactipeptide, subtilosin A, was isolated from *Bacillus subtilis* in 1985 [178]. Subtilosin A shows a narrow spectrum of activity against several human pathogens, such as *Gardnerella vaginalis*, *Listeria monocytogenes*, and *S. agalactiae* [179]. Since the discovery of subtilosin A, a number of additional members of this bacteriocin family have been identified, which show antimicrobial activities against closely related Gram-positive bacteria relative to the producer organisms [157, 180]. The narrow-spectrum activity of these bacteriocins has attracted interest for their use as a therapeutic tool to treat and control gastrointestinal infections without damaging the normal gut microbiota [180].

One of the most well-characterised sactipeptides is thuricin CD, which was first isolated from *Bacillus thuringiensis* in 2010 [181]. Thuricin CD demonstrates high potency against *Clostridium difficile*, with an activity at least equivalent to commonly used antimicrobials. Unlike commonly used broad-spectrum antimicrobials in *C. difficile* treatment, thuricin CD has been shown to specifically inhibit the growth of this pathogen without having any substantial impact on other members of the microbiota [181].

The exact mechanisms responsible for the antimicrobial activity of sactipeptides are not yet entirely clear; however, due to their high potency and narrow spectrum activity, it is widely speculated that sactipeptides require binding to a specific receptor in the outer membrane of the target cell [176]. Subtilosin A has been found to be partially buried in lipid bilayers, possibly exerting its antimicrobial activity by penetrating the outer membrane of the cell [176].

1.3.6 Lassopeptides

Lassopeptides are a family of bacteriocins that are characterised by short peptides of 16 to 21 amino acids containing a slipknot-like structural motif that resembles a lasso rope, hence the name. The knotted structure of these biomolecules confers resistance against thermal, chemical and proteolytic degradation and is key to their intrinsic antibacterial property. Due to their stability in harsh conditions, lasso peptides are considered attractive candidates for drug discovery [182].

The first lasso peptide, anantin, was isolated from a strain of *Streptomyces coerulescens* in 1991 [183]. Since then, over 40 new lasso peptides have been discovered, most of which have shown narrow spectrum antimicrobial activity against strains closely related to or in competition with the producer organism in its ecological niche [182, 184]. However, some lasso peptides, such as propeptin and aborycin, are known to display inhibitory activity against both Gram-positive and Gram-negative bacteria [185, 186].

Antimicrobial activity of lassopeptides appears to be mediated by inhibiting transcription via binding to essential molecular machines of the sensitive cells, such as bacterial RNA polymerase and chaperones [187]. It remains an open question whether other modes of actions, such as stimulating the production of

superoxide and interfering with the glucose uptake system, are also involved in the antimicrobial actions of these compounds [187, 188].

1.3.7 Lactococcin 972-like peptides

Lactococcin 972-like peptides are a family of bacteriocins originally discovered in *Lactococcus lactis* IPLA 972 in 1996 [189]. Since then, additional lactococcin 972-like bacteriocins have been discovered from a variety of different bacterial species belonging to Firmicutes and Actinobacteria phyla [190].

Lactococcin 972-like bacteriocins do not appear to exhibit pore formation activities, rather, they are the only family of non-lantibiotic bacteriocins described thus far that specifically inhibit septum formation by binding to lipid II, although the exact mechanism has not been elucidated. Since their mode of action is through blocking the formation of septum, they are only effective against growing bacteria and have little to no effect on non-dividing cells [189, 191, 192]. Given the recent discovery of this family of bacteriocins, only a few studies have assessed their spectrum of antibacterial activity, which so far suggests a narrow range of bacterial targets. Lactococcin 972 and garvicin, both isolated from *Lactococcus* species, were found to be active only against other closely related lactococci [192, 193].

Compared to other bacteriocin gene clusters, the genetic structure of lactococcin 972-like bacteriocins is relatively simple as they often contain only three genes, which are designated as A, B, and C, encoding a bacteriocin pre-peptide, an immunity protein and an ABC transporter, respectively [190].

1.3.8 Bacteriocin discovery

The traditional approach to identifying new bacteriocins involves culture-based

isolation of potential bacteriocin-producing organisms from diverse environments, followed by screening for antimicrobial activity. Generally, these methods rely on detecting the ability of one bacterial strain (producer) to inhibit the growth of another (indicator). Antimicrobial activity can be assessed using various techniques, among which the well-diffusion technique is the most common. This method involves the inoculation of molten agar with susceptible indicator bacteria that is poured into a Petri plate and allowed to set. Wells are cut into the agar and are filled with either cell-free supernatant from the producer (potentially harbouring a bacteriocin peptide) or purified peptide. After incubation under optimum growth conditions, the wells exhibiting clear zones of inhibition will suggest that the strain tested produces bacteriocins [194]. Additional characterisation methods, often involving traditional purification methods combined with mass spectrum analysis, are required in order to assess the exact nature of the detected compound and to confirm its novelty [195, 196].

With the advances in genome sequencing, the identification of new bacteriocins is changing from traditional culture-based screening towards *in silico* mining of bacterial genome sequences. The rapidly increasing number of available whole genome sequences provides the ability to perform searches for bacteriocin genes. To facilitate these efforts, a number of dedicated bacteriocin data repositories have been developed such as BACTIBASE and BAGEL [197, 198]. These repositories contain manually curated bacteriocin sequences to facilitate the detection of novel bacteriocin peptides using standard sequence homology algorithms such as the basic local alignment search tool (BLAST) [199]. Such an approach has previously been successfully employed to identify many novel bacteriocins such as lichenicidin and sakacins [200, 201]. At the time of writing, there are >200 unique

bacteriocin peptide sequences deposited in Bactibase (September 2019), and this number is continually increasing.

However, finding novel bacteriocins by focusing on precursor peptide discovery is challenging and not always fruitful. This is because many bacteriocin peptide genes are very small and exhibit significant sequence diversity, and thus are often missed by homology searches. To circumvent this issue, conserved bacteriocin-associated genes such as those encoding for bacteriocin self-immunity and transporters can also be used as guides to identify new bacteriocin gene clusters. antiSMASH (antibiotics and Secondary Metabolite Analysis Shell) is a fully automated pipeline that combines direct screening for the precursor peptides along with indirect screening for other bacteriocin-related genes in order to extensively scan the genomes [202].

During the past decade, various genome mining projects have resulted in the discovery of novel bacteriocins [157, 190, 200, 201, 203]. As a result, bacteriocins are now known to be significantly more prevalent in various phyla than previously realised. For instance, using a combination of antiSMASH, BAGEL and BACTIBASE databases, Letzel and coworkers scanned >200 anaerobe genomes from 18 distinct phyla and found that 25% of the genomes contain a bacteriocin cluster. In their study, 81 bacteriocin clusters were detected, 43 of which were newly-discovered [190]. Overall, genome mining has proven to be an incredibly useful tool for the identification of novel bacteriocin gene clusters.

1.4 Bacteriophages

1.4.1 General characteristics

Bacteriophages (phages) are viruses that infect and replicate within bacteria. They are proposed to be the most abundant biological entities on the planet, with the estimated global number of about 10^{31} phage particles (196). Upon entering a target bacterial cell, phages act as intracellular parasites by hijacking the host machinery to manufacture new phages, which typically results in the death of the infected cell. In this way, these natural predators of bacteria play an immense role in the turnover of the bacterial ecosystem [204, 205].

Initially discovered independently by Frederick Twort in 1915 and Felix d'Herelle in 1917, phages were immediately recognised as a great tool for eliminating bacterial pathogens [206]. Prior to the antibiotic revolution, they were conceived as a “magic bullet”; a drug that can specifically attack its target without damaging human cells. As a result, their initial discovery was followed by research on phages as a cure for human illnesses; however, the majority of these experiments were carried out without a clear understanding of the viral nature of phages and their bactericidal mode of action, and thus, success was unpredictable and infrequent [206]. The viral nature of phages, disputed for a time, with sceptics arguing that phages were enzymes rather than viruses, was definitely recognised after visualisation of phages by electron microscopy in 1940 [206].

After the introduction of antibiotics in the 1940s, the initial enthusiasm towards the phage approach quickly faded in most Western countries, but its use continued in several countries such as Russia and Poland, where they are still used to this day [206]. In the West, phages were continued to be studied as model organisms in

molecular biology, which resulted in many fundamental contributions to the field, such as the elucidation of the triplet nature of the genetic code [207] and the discovery of messenger RNA [208]. Moreover, phages have been a valuable source of numerous enzymes used routinely in bacterial genetics experiments and biotechnology [209]. In recent years, interest in phage-based therapeutics for bacterial infections has re-emerged due to the increase in prevalence of antibiotic-resistant bacteria [154, 206] (see 1.2.8).

1.4.2 Lytic *vs.* lysogenic phages

Phage infection starts with the adsorption of the phage to the host bacterium. This is mediated by the phage tail proteins that recognise and bind to specific receptors located on the host cell surface. Subsequently, the phage genome is injected into the host cell and replication of the phage particle can begin [206].

Phages are typically divided into two main types according to their replication strategies: lytic and lysogenic. Lytic phages enter a bacterial cell and immediately start producing new virus particles that burst out of the bacterial cell. The escape from the infected cell is facilitated by the expression of the phage-encoded hydrolysing enzymes such as lysin, which permeabilise ("make holes in") the bacterial cell membrane. The newly-released phage particles then go on to infect other susceptible host cells [206].

Alternatively, lysogenic phages do not necessarily start multiplying immediately after host entry and may instead integrate their genomes into the host genome to be activated at a later time. The integration of the phage genome is mediated by the activity of a phage-encoded integrase and generally occurs at a unique attachment site in the bacterial genome [210]. Once a phage is integrated into the

host bacterial genome, it is termed a prophage and the genes encoding the prophage can be carried over to the bacterial daughter cells after cell division. The prophage genes can remain dormant within the host cell for many generations until the cell experiences some type of stress, which may result in activation of prophage and initiation of the lytic cycle [211, 212].

1.4.3 Impact of lysogeny on host cell

Unlike lytic phages, lysogenic phages often maintain a long-term relationship with their host. In this context, since prophages, like other obligate intercellular parasites, rely solely on their host for survival, they often express genes that enhance the fitness of the host bacteria. Prophages can have a variety of profound phenotypic effects on the bacterial host, for example by encoding toxins that increase bacterial virulence [213], promoting the binding of bacteria to human platelets [214] or cells [215], helping the host bacteria to overcome immune defences [216], or enhancing the survival of intracellular bacteria by protecting them from oxidative stress [217]. Prophage integration can also modulate bacterial gene expression through gene disruption or genome rearrangements [218].

1.4.4 Constant arms race between bacteria and phages

Phages outnumber bacteria by an estimated factor of ten to one, exerting a constant challenge to bacterial populations [205]. As with all predator-prey relationships, phages and bacteria are inseparably tangled in a constant arms race, whereby the bacterial hosts evolve a broad range of defence mechanisms to protect themselves and phages co-evolve counterstrategies to bypass these defences [219]. To name a few, bacteria may lose or alter the target receptor of phages in order prevent virus cell entry, and phages bypass this defence by

mutations in their receptor-recognition sites or by using different receptors in order to attach to the cell. Bacteria can also use extracellular polysaccharides as a means to hinder phage attachment to the cell surface. In response, many phages have evolved degradative enzymes which allow them to overcome this physical barrier. Most bacteria also harbour restriction endonuclease enzymes that can recognise and degrade phage DNA in the cytoplasm; however, phages are shown to be able to evade these enzymes by corrupting recognition sequences in their genomes [219]. These coevolution dynamics are further complicated by satellite prophages, which are prophages that are parasites of other phages [220].

1.4.5 Satellite prophages

The term “satellite” has been in use since the 1960s to describe a virus that is not capable of multiplying in cells without the assistance of another virus (referred to as ‘helper’) and is not necessary for the infection cycle of the helper virus [221]. Satellite prophages are a type of prophage that lacks all the necessary genetic information for replication on its own, and thus depends on hijacking the machinery of a helper phage for its survival and spread [222-224]. Essentially they might be considered as ‘parasites of parasites’ [225]. Satellite prophages have been discovered through different circumstances in various organisms and thus there are different terms used to describe this particular type of mobile genetic element (MGE) in the literature: staphylococcal pathogenicity islands (SaPIs) [226]; phage-related chromosomal islands (PRCIs) [223]; phage-inducible chromosomal islands (PICIs) [224] and phage-like *Streptococcus pyogenes* chromosomal islands (SpyCIs) [227]. For simplicity, the generic term satellite

prophages will continue be used throughout this thesis to refer to this class of MGE, of which SaPIs, PRCIs, PICIs or SpyCIs are thus a subset.

Similar to other prophages, satellite prophages reside at specific locations in the host chromosome. They remain quiescent until activated by a certain stimulus, which might be superinfection by a new phage or the induction of another resident prophage in the host genome [228]. Upon activation, they interfere with helper phage replication and divert both the bacterial host and helper virus machinery to the production of satellite prophage particles, and in this way, they promote the survival of the bacterial population [222-224]. For instance, many encode genes that remodel the helper virus capsid to better accommodate the smaller satellite prophage genome and exclude the larger helper virus genome [229, 230]. Although the specific cell confronting superinfection might eventually be killed, this death is advantageous at the bacterial population level, as the neighbouring cells are less likely to undergo lysis due to fewer infectious phage particles in the surrounding environment.

Satellite prophages are known to be vectors for the spreading of toxin genes and other virulence factors. One of the best-known examples is the *Staphylococcus aureus* pathogenicity island 1 (SaPI1), which was discovered because it carries the gene encoding toxic shock syndrome toxin-1 (*tst*), the causative agent of toxic shock syndrome [222]. A small number of satellite prophages have been discovered in streptococcal species [231], however, whether they contribute to the virulence of the host organism was unknown.

1.4.6 Prophage discovery

Traditionally, the most commonly used approach to identifying prophages was to induce the phage via exposure of the host cell to DNA damaging agents or other stressful conditions in order to release phage particles, followed by assessing the presence of phage particles using various experimental methods [232]. The most traditional and widely used method to assess the presence of phage is the plaque assay, originally developed by one of the discoverers of phages, Felix d'Herelle, in 1917. The plaque assay is performed by growing a thin layer of host cells onto a culture dish and observing plaques (clear zones formed on the bacterial lawn due to the lysis of bacterial cells). Serial dilutions of the samples are often prepared to create plaques derived from individual phage clones. Subsequently, the phage from each plaque may be isolated and further analysed using DNA isolation, sequencing and electron microscopy [232].

Prophages can also be detected using DNA probe-based detection of conserved prophage-associated genes with molecular biological techniques such as polymerase chain reaction (PCR). Primers specific to target genes known to be carried by prophages can be designed and used to confirm the presence of the prophage within the bacterial genome [232, 233]. However, the major disadvantage of these techniques is that they rely on the use of primers complementary to the prophage sequences, and thus previously unidentified prophages may be overlooked [232].

While experimental work still plays an important role in phage research, the explosion of whole genome sequence data has drastically expedited the identification and genetic analysis of prophages. Owing to the decrease in the cost

of sequencing whole bacterial genomes (226) the identification of novel prophages is shifting from traditional culture-based screening towards *in silico* mining of bacterial genome sequences. A handful of *in silico* prophage detection tools are currently available, among which Prophinder [234] and PHAST [235] are the most well-known. These tools scan bacterial host genomes using a database of previously known phage genomes in order to identify prophages. As a result, their performance is directly dependent on the extensiveness of their database and hampered by the fact that different phages have little sequence similarity and very diverse genome sizes. Such tools are particularly inadequate when applied to less well characterised bacterial species or when aiming to identify a recently discovered type of viruses, such as streptococcal satellite prophages (see Chapter 4). Consequently, in order to ensure a thorough discovery of previously unidentified prophages, prophage predictions need to undergo lengthy manual (human) curation and inspection [212, 236-238].

One manual computational approach that has resulted in the discovery of many previously unidentified prophages [212] involves generating a list containing the location and the annotation of each open reading frame (ORF) in the host genome, which is then manually reviewed for clusters of bacteriophage-related genes. The content, order and orientation of the ORFs within the putative prophage genome (and its flanking ORFs) are then carefully evaluated using a variety of different software. However, this process is slow and not scalable. Taken together, there is currently a need for the development of new *in silico* screening tools that can rapidly and effectively identify prophages in large genome studies.

1.5 Aim and outline of this thesis

This thesis will investigate a number of questions related to the prevalence, diversity and molecular epidemiology of bacteriocins and prophages. In the present chapter, the literature review and studies relevant to this thesis have been discussed. The chapter that follows moves on to describes the methods more generally used throughout this thesis.

In Chapter 3, a genome mining approach was employed to discover potential novel bacteriocin systems in a large and diverse set of historical and modern pneumococcal genomes. Fourteen novel bacteriocin clusters were discovered among pneumococci, tripling the number of known bacteriocins in this species. The molecular epidemiology of the identified bacteriocin clusters was conducted in the context of the pneumococcal population structure. Furthermore, RNA sequencing transcriptome analyses were used to investigate whether the identified bacteriocin clusters were transcriptionally active. This chapter also explores whether pneumococcal bacteriocin clusters are also present in other streptococcal species.

Chapter 4 presents the identification of 44 unique and novel satellite prophages among the pneumococcal genomes. The prevalence, diversity, genetic stability and molecular epidemiology of the pneumococcal satellite prophages were investigated using a large and diverse set of historical and modern pneumococcal genomes isolated between 1937 and 2008. This chapter also uses existing RNA sequencing datasets to investigate the expression of prophage genes following mitomycin C induction, as well as in planktonic growth and biofilm conditions. Satellite prophage genomes were screened for the presence of virulence genes, which led to the discovery of a gene that is associated with virulence in a murine infection model.

Chapter 5 attempts to perform genus-wide analyses of the genomic diversity and population structure of streptococcal prophages using >1,300 genomes from 70 different *Streptococcus* species. With this aim, Chapter 5 outlines the development of a new bioinformatics software called PhageMiner, which can streamline the manual curation process for prophage discovery. Using the tool, nearly 800 full-length and satellite prophages were identified, the majority of which were newly discovered. This chapter also aimed to estimate the average prophage gene content within various streptococcal species and provides convincing evidence that cross-species transmission of prophages is not uncommon. The location of prophage insertion sites and the flanking genes upstream and downstream of each prophage were also investigated within streptococcal genomes. The final chapter, Chapter 6, summarises the main findings of this thesis and identifies areas for further research.

2. General Methods

This chapter describes the methods more generally used within this thesis. Methods specific to a particular Results chapter are given within the chapter itself.

2.1 BIGSdb

BIGSdb (Bacterial Isolate Genome Sequence Database) is a software platform used to store and analyse sequence data for bacterial isolates [239]. The BIGSdb software can store sequence data ranging from single sequence reads of a few base pairs to whole genomes, and each isolate entry can be linked to its associated metadata, such as serotype, country of isolation, date of isolation, among many other fields. Furthermore, various individual modules exist within the BIGSdb infrastructure that facilitate downstream analyses of sequence data. For example, the Gene Presence module allows a set of sequences to be defined as references, which can then be used to scan bacterial isolates for the absence or presence of certain genes, such as those found within a given bacteriocin gene cluster. Other modules within BIGSdb that were used throughout this thesis include third party tools such as BLAST (see 2.4) and iTOL (see 2.6), which are described below. All genomes used in this thesis were stored in a private BIGSdb, referred to as the Brueggemann BIGSdb.

2.2 Genome collection

All genomes used in this thesis were previously sequenced and assembled by others and not by me. Most genomes were already submitted to the Brueggemann BIGSdb by various former members of the Brueggemann group over many years

after being newly sequenced or retrieved from other published databases. Those not already present in the Brueggemann BIGSdb were retrieved from the ribosomal MLST database (https://pubmlst.org/rmlst) and uploaded to the Brueggemann BIGSdb database. The specific genome datasets used for particular analyses are detailed in the relevant Results chapters.

2.3 RNA sequencing datasets

Four RNA sequencing datasets were used in this thesis: three were generated by our research group and were not yet analysed, the other was generated by Blanchette *et al* [240]. Raw RNA sequencing data from Blanchette *et al* was downloaded from the publicly available database Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo) using the accession number GSE85196.

The datasets from our group were generated by culturing pneumococci in 10 ml tubes of brain-heart infusion (BHI) broth for 6 hours. An aliquot of 0.5 ml broth was removed and the absorbance at OD_{600} was measured at each time point from 0 to 6 hours to measure increased bacterial growth. In one experiment, BHI broth were inoculated with pneumococcal reference strain PMEN3 and incubated at 37°C + 5% CO_2. Mitomycin C (Sigma-Aldrich) was added to the broth culture tubes to a final concentration of 2 μ g/ml after 3 hours of incubation ($OD_{600} \geq 0.5$) to facilitate phage induction. In the second experiment, BHI broth were inoculated with pneumococcal strain 2/2 and incubated at 40°C + 5% CO_2 to stress the bacterial cells by growth at high temperature. Controls for both experiments were the pneumococci incubated at standard conditions (37°C + 5% CO_2). In the third experiment, pneumococcal reference strains PMEN3 and PMEN6 were cultured in

the same BHI broth aliquot and incubated at 37°C + 5% CO_2, and the experimental controls were generated by growing each reference strain individually in the BHI broth. RNA extractions were performed on samples from all the three experiments from five time points (2, 3, 4, 5 and 6 hours of incubation). A 0.5 ml aliquot was used to measure the absorbance and the RNA was stabilised in the remaining 9.5 ml of broth culture by the addition of 19 ml of RNAprotect Bacteria Reagent (Qiagen). RNA was immediately extracted from the samples using the Promega Maxwell® 16 Instrument and LEV simplyRNA Cells purification kit, following the manufacturer's protocol. Extracted RNA samples were sent to the Oxford Genomics Centre where library preps were made using RNA-Seq Ribozero kits (Illumina, Inc) and sequencing was performed on the MiSeq (Illumina, Inc).

Methods used for calculating differential gene expression levels between the control and experimental conditions varied for each specific analysis and are detailed in the relevant Results chapters.

2.4 BLAST

BLAST (Basic Local Alignment Search Tool) [199] is one of the most widely used bioinformatics tools. It performs similarity searches for nucleotide or amino acid sequences against a library of sequences and quantifies the sequence similarity shared by the results (referred to as a "subject") and the query sequence. Different versions of BLAST exist for comparison of different kinds of sequences, such as BLASTN for nucleotide queries, BLASTP for protein queries, and TBLASTX, which translates nucleotide queries in all six frames for comparison to other protein queries or databases.

The BLAST output includes E-value, query coverage, identity value and total score. E-value denotes the expected number of chance alignments; the smaller the E-value, the less likely the alignment was by chance. Query coverage refers to the percent of the query length that is included in the aligned segments. Identity value is the percentage of perfect matches between subject and query. The total score is the sum of scores of all aligned sequences.

2.5 Gene prediction and annotation

All whole genome bacterial sequences used in this thesis were annotated using the automated pipelines Prokka (Rapid Prokaryotic Genome Annotation) and RAST (Rapid Annotation using Subsystem Technology). Any specific genes of interest were further investigated manually using CD-Search (Conserved Domain Search) and STRING (Search Tool for the Retrieval of Interacting proteins). Further details regarding these tools are provided below.

2.5.1 RAST

The RAST server is a free, publicly-available, fully-automated software pipeline for annotating bacterial and archaeal genomes. Inputs are uploaded to the RAST server (http://rast.nmpdr.org) in the form of a set of contigs in FASTA format, where protein-encoding genes are predicted and annotated via a web interface. The RAST pipeline works by projecting gene annotations onto newly submitted genomes using evidence from different sources such as ORF prediction via Glimmer [241], homology searches for conserved protein families (FIGfams) [242], and a BLAST search versus a large and rapidly growing protein database [243]. Annotation data in the RAST server database is manually curated by skilled human

curators and organised into subsystems (sets of logically related functional roles), which provides a consistent and accurate pipeline for initial annotation of bacterial genomes [243].

2.5.2 Prokka

Similar to the RAST server described above, Prokka is a tool designed for rapid annotation of prokaryotic genomes and it works by assigning gene annotations derived from a protein sequence database [244]. Annotation was performed using Prokka version v1.10 with default parameters ('Prokka Kingdom Bacteria' as primary reference database). The primary reason for the use of Prokka in addition to RAST was because Prokka will generate a number of output files in a format specifically required as inputs by other bioinformatics software used in downstream analyses, such as GFF (General Feature Format), a plain text file used to describe gene structure.

2.5.3 CD-Search

CD-Search [245] is a tool used in protein annotation. CD-Search identifies the conserved structural and functional domains present in a protein sequence by searching a comprehensive collection of domain models stored in Conserved Domain Database (CDD) [246] curated by the National Center for Biotechnology Information (NCBI). A protein or nucleotide query sequence can be submitted in the form of a protein or nucleotide using the NCBI's online interface (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The results of CD-Search are displayed as an annotation of protein domains on the input sequence. Strong

associations between a query sequence and conserved domains are reported as hits.

2.5.4 STRING

STRING [247] is a large, publicly-available, protein-protein interaction database containing both known and predicted interactions based on experimental data and mining of databases and literature. Protein sequences for any genes of interest can be submitted to the STRING database to perform searches for any previously reported relationship to other genes. STRING produce a score to estimate the accuracy of any identified association, which ranges from 0 to 1 (1 indicates the highest degree of confidence that an interaction is non-random). STRING can be accessed at http://string-db.org.

2.6 Visualisation of phylogenetic trees

All phylogenetic trees in this thesis were annotated using the online tool iTOL (Interactive Tree of Life) [248], which was used to overlay relevant metadata on or near the tree branches, such as the presence or absence of a particular bacteriocin cluster or prophage in any given tree cluster or sets of isolates. The iTOL tool is available as a plugin within BIGSdb (see 2.1) or can be reached at http://itol.embl.de.

2.7 Estimation of prophage content among bacterial genomes

I developed a Python script named PhageContentCalculator, which calculates the prophage content as the percentage of prophage genes within a given bacterial genome. This script first uses Prodigal software in the Prokka annotation [244] suite (version 1.10) to predict ORFs in three separate groups of sequences:

(i) full-length prophage genomes, (ii) satellite prophage genomes and (iii) a single bacterial genome of interest for which the phage content is to be calculated. Next, the script uses Roary [249] set at a 70% similarity threshold to extract, combine and cluster individual ORF nucleotide sequences from all three groups. Any ORFs in the bacterial genome that were also present in at least one prophage genome were deemed to be phage-related and this information was used by the script to output the total percentage of phage-related ORFs in the given bacterial genome. The PhageContentCalculator script is available in GitHub (https://github.com/RezaRezaeiJavan/PhageContentCalculator).

3. Genome sequencing reveals a large and diverse repertoire of antimicrobial peptides

Most of the results described in this chapter were presented as an oral presentation at the 13th European Meeting on the Molecular Biology of the Pneumococcus (EuroPneumo 2017), in Stockholm, Sweden in June 2017. Part of the results were also presented at the 11th International Symposium on Pneumococci and Pneumococcal Diseases (ISPPD 2018), in Melbourne, Australia in April 2018.

This chapter was published as:

Reza Rezaei Javan, Andries J. van Tonder, James P. King, Caroline L. Harrold, and Angela B. Brueggemann. 2018. Genome Sequencing Reveals a Large and Diverse Repertoire of Antimicrobial Peptides. *Frontiers in Microbiology.* doi: 10.3389/fmicb.2018.02012

The bacterial genomes used in this chapter were all previously sequenced and assembled by others and not by me. These genomes were submitted to the Brueggemann BIGSdb database by various former members of the Brueggemann group over many years after being newly sequenced or retrieved from other published databases. I downloaded these genomes and their associated metadata from the Brueggemann BIGSdb database. Dr Keith Jolley and Prof Martin Maiden in the Department of Zoology, University of Oxford had developed the BIGSdb infrastructure, which allowed these data to be easily accessible.

Before I joined the Brueggemann group, the molecular epidemiology of one of the previously known lantibiotic clusters (herein, called streptolancidin G) had been

partially characterised by Dr James King, a medical student undertaking his final year honours project in the Brueggemann group, and Dr Andries van Tonder, a DPhil student under the supervision of Prof Brueggemann at that time.

I developed the genome mining pipeline for the identification of bacteriocins and discovered all the 14 novel bacteriocin clusters reported in this study. Subsequently, I developed a Python script for calculating the prevalence, molecular epidemiology and co-occurrence patterns of all the bacteriocin clusters among pneumococci. I classified all pneumococcal bacteriocins according to a logical nomenclature format and performed further detailed genomic characterisation on these bacteriocins. My supervisor, Prof Angela Brueggemann, provided advice throughout this process.

The RNA sequencing experiments were conducted in the Brueggemann laboratory by Dr Andries van Tonder, Caroline Harrold and Prof Angela Brueggemann. Extracted RNA samples were sent to the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics for sequencing. I used the raw reads generated from these experiments and assembled them to the reference genomes. Subsequently, I performed the differential expression analyses, which involved the development of a novel computational approach for the competition experiment analyses.

3.1 Abstract

Competition among bacterial members of the same ecological niche is mediated by bacteriocins: antimicrobial peptides produced by bacterial species to kill other bacteria. Bacteriocins are also promising candidates for novel antimicrobials. Here, 14 novel bacteriocin gene clusters were discovered by screening >6,200

pneumococcal genomes. The molecular epidemiology of the bacteriocin clusters was investigated using a large global and historical pneumococcal dataset dating from 1916. This work revealed extraordinary bacteriocin diversity among pneumococci and the majority of bacteriocin clusters were also present in other streptococcal species. Genomic hotspots for the integration of different bacteriocin gene clusters were discovered. Bacteriocin genes were found to be transcriptionally active when the pneumococcus was under stress and when two different strains were co-cultured in broth. The findings from this chapter reveal much more diversity among bacterial defense mechanisms than previously appreciated, which fundamentally change our view of bacteriocins and nasopharyngeal competition among pneumococci.

3.2 Introduction and aims

The preferential niche of the pneumococcus is the nasopharynx, a polymicrobial environment where competitive interactions shape the microbial community [12]. While the genetic basis for the pathogenicity and virulence of the pneumococcus is not yet fully understood, it is known that certain pneumococcal lineages and serotypes are predominately associated with disease whilst others with asymptomatic nasopharyngeal carriage [41]. Therefore, understanding the factors that affect the makeup of the bacterial population is crucial, as it is indirectly related to pathogenicity. The pneumococcus is known to possess bacteriocins that mediate intra- and interspecies competition within the human nasopharynx. To date, seven bacteriocin clusters have been identified in this organism, which are briefly reviewed below.

The most well-characterised pneumococcal bacteriocin system is the *blp* (bacteriocin-like peptide) cluster. Early work by two research groups carefully

described the genetics and experimental activity of the *blp* cluster in pneumococcus [250, 251], and the Brueggemann group later showed that the *blp* bacteriocin system is ubiquitous among pneumococci and genetically highly diverse [175]. The number of bacteriocin peptides present within the bacteriocin gene cluster varies among different strains of the pneumococcus [13, 175, 251-253]. For instance, some pneumococci were found to possess up to six bacteriocin genes and eight putative immunity genes in their *blp* cluster, whereas other pneumococci contained only one of each [175]. The production of *blp* bacteriocins is thought to result in more efficient colonisation of the host due to the ability of these toxins to eliminate competitors; a strain harbouring the *blp* cluster has a competitive advantage in a mouse colonisation model [13] and *in vitro* experiments have demonstrated that *blp* bacteriocins are able to inhibit not only other pneumococci lacking the associated immunity genes but also closely related streptococci, such as *S. pyogenes*, *S. mitis*, and *S. oralis* [13, 251, 254].

Another well characterised bacteriocin system in the pneumococcus is the *cib* (competence-induced bacteriocin) cluster, which encodes two peptide bacteriocins, CibAB, and the corresponding bacteriocin self-immunity protein CibC. CibAB bacteriocin peptides are implicated in competence-mediated fratricide, which results in lysis of noncompetent cells [255]. The prevalence of the cib bacteriocin cluster among the pneumococcal population is currently unknown.

A novel circular bacteriocin named pneumocyclicin was recently discovered *in silico* by the Brueggemann group [175]. This bacteriocin is similar in genetic organisation to other well-known members of the head-to-tail cyclised bacteriocin family (see 1.3.4), such as uberolysin from *Streptococcus uberis* and circularin A

from *Clostridium beijerinckii*. Bogaardt *et al.* [175] investigated 336 diverse pneumococcal isolates belonging to more than 50 distinct clonal complexes and demonstrated that the pneumocyclicin cluster was present in nearly one-third of the pneumococcal genomes and was particularly prevalent among major pneumococcal clonal complexes [175].

Recently, the presence of four lantibiotic bacteriocin clusters (see 1.3.3 for a review on lantibiotics) has been reported in this bacterium, however, their prevalence and diversity remain poorly understood. These include pneumolancidin and three other lantibiotic gene clusters that have not been named, highlighting the need for a unified nomenclature for the pneumococcal bacteriocin clusters. Pneumolancidin has been shown to inhibit a wide range of Gram-positive bacteria, including pneumococci that do not harbour the pneumolancidin bacteriocin cluster and thus lack the genes required for immunity to this bacteriocin, as well as isolates of *L. monocytogenes*, *S. pyogenes, S. agalactiae*, and *L. lactis* [256].

As for the remaining three unnamed lantibiotic clusters, one was found to be under the control of a quorum sensing system known as TprA/PhrA, which induces their expression in the presence of galactose and represses them when under high glucose growth conditions [257]. Intriguingly, galactose is found in abundance in the nasopharynx, whereas glucose is scarce (although abundant in the blood), suggesting that the TprA/PhrA system may mediate the expression of this lantibiotic cluster to aid in competition for limited nutrient resources during nasopharynx colonisation [257]. Details regarding the antimicrobial activity of the lantibiotic cluster controlled by TprA/PhrA is currently scare, however, experiments suggest that the bacteriocins produced by this system are active against *Micrococcus flavus*, a

bacterium that colonises humans [258].

The second unnamed lantibiotic gene cluster was identified computationally within an integrative and conjugative element (ICE) present in the pandemic pneumococcal strain ATCC700669 [259]. It has been suggested that the presence of this lantibiotic cluster may have been a contributing factor to the success of this strain, although no antimicrobial activity has so far been reported from this lantibiotic cluster [259].

A recent genome mining effort (see 1.3.8 for more details on bacteriocin discovery) has identified a third unnamed lantibiotic cluster within the genome of pneumococcal strain SP23-BS72 [200]; however, the toxin genes encoded within this lantibiotic cluster in this particular strain were found to be frameshifted, which is predicted to result in the loss of antimicrobial activity [200]. Whether a functional version of this lantibiotic cluster is present among other strains of the pneumococcus is yet to be investigated.

At the moment, excluding the *blp* and pneumocyclicin bacteriocins, details regarding the prevalence, genetic composition and molecular epidemiology of the remaining pneumococcal bacteriocin clusters in the context of the pneumococcal population are mostly unknown. Importantly, it remains an open question whether other bacteriocin clusters are also present among pneumococci and are still awaiting discovery. Therefore, the main aims of this chapter were to: a) employ a genome mining approach to discover potential novel bacteriocin clusters in a large and diverse set of historical and modern pneumococcal genomes isolated between 1937 and 2008; b) determine the prevalence, genetic composition and molecular epidemiology of all known bacteriocins with respect to global population structure of pneumococci; c) study the genetic stability of each bacteriocin cluster over time;

d) explore co-occurrence patterns between different bacteriocins; e) characterise bacteriocin cluster insertion sites; f) explore whether pneumococcal bacteriocin clusters are also present in other streptococcal species; and g) investigate the expression of pneumococcal bacteriocin genes under various stress conditions.

3.3 Methods

3.3.1 Genome mining for the discovery of novel bacteriocin clusters

A large dataset of assembled pneumococcal genomes (n=6,244) was compiled from previously published studies [212, 260-266] (Supplementary File 3.1). Genomes were scanned for the presence of bacteriocin gene clusters using a number of different bioinformatics tools and databases, such as antiSMASH [202] (to identify putative gene clusters that encode microbial secondary metabolites), bactibase [197] and Bagel [198] (to screen our genome sequences for homology to known bacteriocin genes from a diverse range of bacterial species) and InterProScan (to assess the putative function of encoded proteins and identify protein domains and key sites) [267] (data collected in March 2016) (Figure 3.1). Predicted bacteriocin-associated genes from the outputs were analysed manually and potential bacteriocin gene clusters were further scrutinised using extensive BLAST searches. An in-house pipeline was generated to automate part of the analysis workflow. Notably, no single tool or sequence alignment was capable of identifying all the bacteriocins, but rather a combination of tools was necessary to confidently identify each bacteriocin gene cluster.

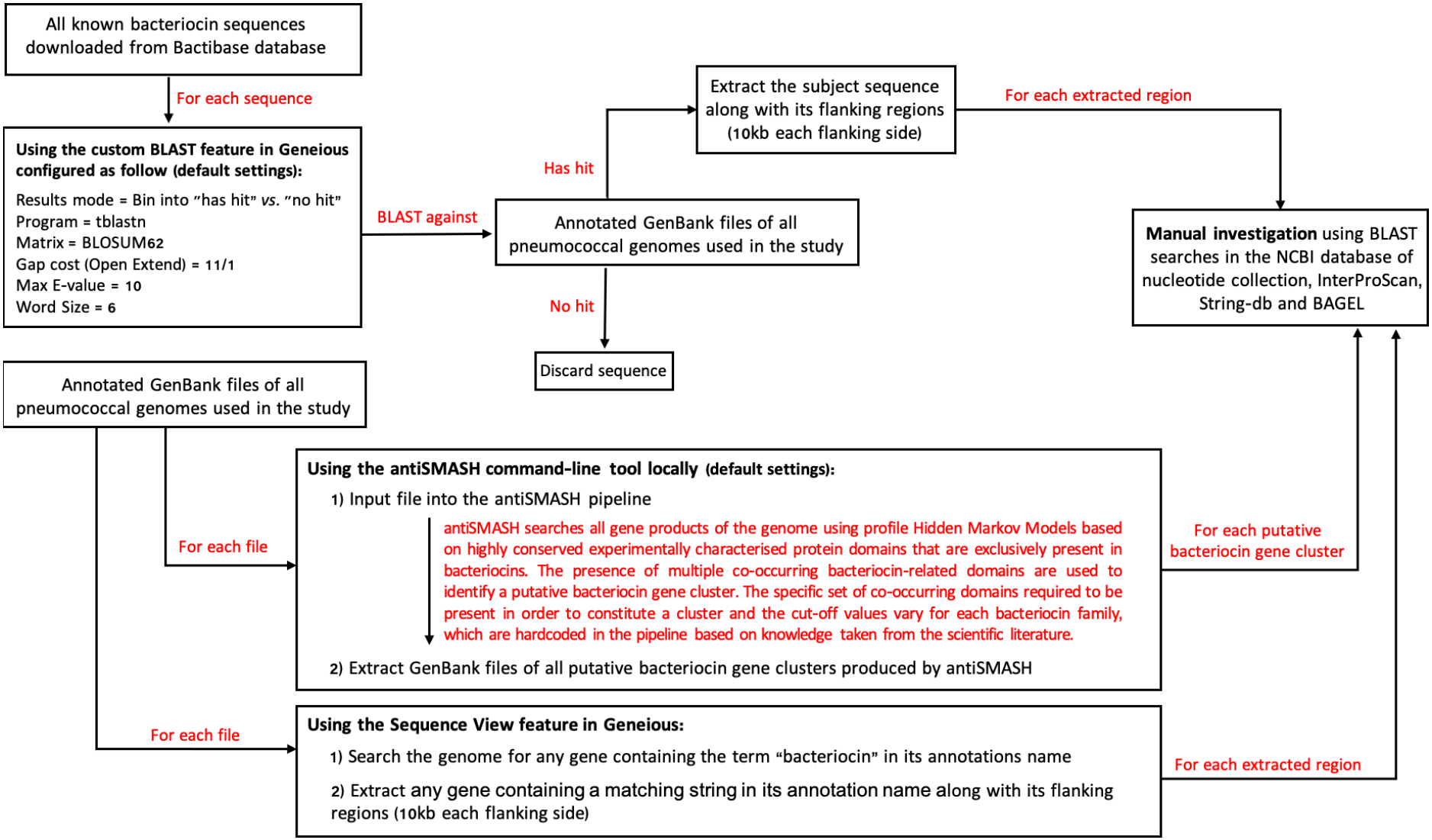


Figure 3.1 - Outline of the pipeline for discovery of novel bacteriocins. No single tool was capable of identifying all the bacteriocins, but rather a combination of tools and substantial manual effort and inspection of sequences were necessary to confidently identify each bacteriocin gene cluster.

3.3.2 Investigation of the predicted bacteriocin genes

Putative bacteriocin genes were annotated based on homology searches to other known bacteriocin genes, combined with structure-based searches. The structural and functional domains in protein sequences were predicted by CD-Search [245] using the NCBI Conserved Domain Database [246]. The protein sequences of genes of interest were submitted to the STRING database [247] to perform searches for any previously reported relationship to other genes (see general methods). Multiple sequence alignments were generated in Geneious version 9.1 (Biomatters Ltd) using the ClustalW algorithm [268] with default parameters (Gap open cost=15, Gap extend cost=6.66), and the output was used to calculate percentage identity matrices. Gene organisation diagrams were produced using Geneious and further edited in Inkscape (http://inkscape.org).

3.3.3 Classification and nomenclature of the identified bacteriocin clusters

Based on their predicted biosynthetic machinery and structural features, bacteriocin clusters were classified according to Arnison *et al.* [157] and associated genes were named following the standard nomenclature for bacteriocins (Table 3.1). Since most bacteriocin clusters were not exclusive to *S. pneumoniae*, but were also present in other closely-related streptococci (see below), the prefix “strepto-” was used to name the bacteriocin clusters followed by an abbreviation of the bacteriocin family: “streptococcins” for those belonging to the lactococcin 972-like family, “streptolancidins” for lanthipeptides, “streptocyclicins” for the head-to-tail cyclised peptides, “streptosactins” for the sactipeptides, and “streptolassins” for the lassopeptide family of bacteriocins. In cases where more than one bacteriocin cluster from the same family were present, they were named alphabetically in the chronological order that they were discovered in the analyses.

Table 3.1 (part 1). Classification and nomenclature of bacteriocin clusters found among pneumococci.

Bacteriocin	Family	Gene	Synonym(s)	Predicted function
Streptococcin A	Lactococcin 972	*scaA*	-	Precursor peptide
		scaB	-	Immunity
		scaC	-	Transport
Streptococcin B	Lactococcin 972	*scbA*	-	Precursor peptide
		scbB	-	Immunity
		scbC	-	Transport
Streptococcin C	Lactococcin 972	*sccA*	-	Precursor peptide
		sccB	-	Immunity
		sccC	-	Transport
Streptococcin D	Lactococcin 972	*scdA*	-	Precursor peptide
		scdB	-	Immunity
		scdC	-	Transport
Streptococcin E	Lactococcin 972	*sceA*	-	Precursor peptide
		sceB	-	Immunity
		sceC	-	Transport
Streptolancidin A	Lanthipeptide class: II	*slaA1*	*pldA1*, SrnA	Precursor peptide
		slaA2	*pldA2*, SrnA	Precursor peptide
		slaA3	*pldA3*, SrnA	Precursor peptide
		slaA4	*pldA4*	Precursor peptide
		slaA5	-	Precursor peptide
		slaF	*pldF, srnX*	Transporter involved in immunity
		slaE	*pldE, srnY*	Transporter involved in immunity
		slaK	*pldK, srnK*	Histidine kinase
		slaR	*pldR, srnR*	Response regulator
		slaM	*pldM, srnM*	Bifunctional modification enzyme
		slaT	*pldT, srnT*	Transport
Streptolancidin B	Lanthipeptideclass: II	*slbF*	-	Transporter involved in immunity
		slbG	-	Transporter involved in immunity
		slbE	-	Transporter involved in immunity
		slbA	*lcpA*	Precursor peptide
		slbM	*lcpM*	Bifunctional modification enzyme
		slbT	*lcpT*	Transport
Streptolancidin C	Lanthipeptide class: IV	*slcA*	-	Precursor peptide
		slcX	-	Unknown
		slcL	-	Bifunctional modification enzyme
		slcT	-	Transport
Streptolancidin D	Lanthipeptide class: I	*sldA*	-	Precursor peptide
		sldB	-	Dehydratase
		sldC	-	Cyclase
		sldT	-	Transport

Table 3.1 (part 2). Classification and nomenclature of bacteriocin clusters found among pneumococci.

Bacteriocin	Family	Gene	Synonym(s)	Predicted function
Streptolancidin E	Lanthipeptide class: II	*sleM1*	LanM	Bifunctional modification enzyme
		sleA1	-	Precursor peptide
		sleA2	-	Precursor peptide
		sleM2	LanM	Bifunctional modification enzyme
		sleM3	LanM	Bifunctional modification enzyme
		sleT	LanT	Transport
		sleX1	-	Unknown
		sleF	MrsF	Transporter involved in immunity
		sleG	MutG	Transporter involved in immunity
		sleX2	-	Unknown
Streptolancidin F	Lanthipeptide class: IV	*slfA*	-	Precursor peptide
		slfL	-	Bifunctional modification enzyme
Streptolancidin G	Lanthipeptide class: II	*slgA1*	LanA1, *pneA1*	Precursor peptide
		slgA2	LanA2, *pneA2*	Precursor peptide
		slgM	LanM	Bifunctional modification enzyme
		slgD	LanD	FAD-dependent flavoprotein
		slgP1	-	Peptidase
		slgT	LanT	Transport
		slgP2	LanP	Peptidase
Streptolancidin H	Lanthipeptide class: I	*slhP*	-	Peptidase
		slhR	-	Response regulator
		slhK	-	Histidine kinase
		slhF	-	Transporter involved in immunity
		slhE	-	Transporter involved in immunity
		slhG	-	Transporter involved in immunity
		slhX1	-	Unknown
		slhX2	-	Unknown
		slhA	-	Precursor peptide
		slhB	-	Dehydratase
		slhT	-	Transport
		slhC	-	Cyclase
		slhI	-	Immunity

Table 3.1 (part 3). Classification and nomenclature of bacteriocin clusters found among pneumococci.

Bacteriocin	Family	Gene	Synonym	Predicted function
Streptolancidin I	Lanthipeptide class: I	*sliP*	-	Peptidase
		sliR	-	Response regulator
		sliK	-	Histidine kinase
		sliF	-	Transporter involved in immunity
		sliE	-	Transporter involved in immunity
		sliG	-	Transporter involved in immunity
		sliA	-	Precursor peptide
		sliB	-	Dehydratase
		sliT	-	Transport
		sliC	-	Cyclase
		sliI	-	Immunity
Streptolancidin J	Lanthipeptide class: IV	*sljA1*	-	Precursor peptide
		sljL	-	Bifunctional modification enzyme
		sljP	-	Peptidase
		sljT1	-	Transport
		sljT2	-	Transport
		sljT3	-	Transport
		sljA2	-	Precursor peptide
Streptolancidin K	Lanthipeptide class: IV	*slkA*	-	Precursor peptide
		slkL	-	Bifunctional modification enzyme
		slkT	-	Transport
Streptocyclicin	Head-to-tail cyclized peptides	*scyA*	*pcyA*	Precursor peptide
		scyB	*pcyB*	Maturation/Immunity
		scyC	*pcyC*	Maturation/Immunity
		scyD	*pcyD*	Transport
		scyE	*pcyE*	Immunity/Transport
Streptolassin	Lasso peptides	*slsA*	-	Precursor peptide
		slsC	-	Cyclase
		slsB1	-	Peptide chaperone
		slsB2	-	Dehydratase
		slsF	-	Transporter involved in immunity
		slsE	-	Transporter involved in immunity
		slsG	-	Transporter involved in immunity
		slsR	-	Response regulator
		slsK	-	Histidine kinase
Streptosactin	Sactipeptides	*ssaA*	-	Precursor peptide
		ssaCD	-	Radical-SAM protein
		ssaX1	-	Unknown
		ssaX2	-	Unknown
		ssaP	-	Peptidase
		ssaX3	-	Unknown

3.3.4 Investigation of the molecular epidemiology of the bacteriocin clusters

A global representative dataset (n=571) was compiled by selecting a large and diverse collection of pneumococcal genomes recovered between 1916 and 2009 from individuals of all ages living in 39 different countries (Supplementary File 3.2). A private BIGSdb database [239] was used to store the genome sequences for the pneumococcal isolates along with their metadata (serotype, country of isolation, date of isolation *etc.*) (see general methods). The BIGSdb database platform was then used to construct a presence/absence matrix considering all the different known bacteriocin genes in all genomes in the study dataset. Using this matrix (79,940 genes) as an input, a Python script was developed to determine the prevalence, molecular epidemiology and co-occurrence patterns of the bacteriocins clusters.

3.3.5 Generation of the core genome phylogenetic tree

Genomes were annotated using the Prokka v1.10 program [244]. The annotation files were then input into Roary [249] and clustered using a threshold of 90% sequence similarity at the amino-acid level (parameter: -i 90). Genes present in all genomes were identified using a core genome threshold of 100% (parameter: -cd 100) and were aligned using Roary. FastTreeMP [269] was used to generate the phylogenetic tree, followed by ClonalFrameML [270] to reconstruct the tree adjusted for recombination. The tree was visualised using iTOL [248] and Inkscape.

3.3.6 Calculating the GC content of *Streptococcus* species

In total, 1,395 assembled genomes from 70 different species of the genus *Streptococcus* were selected for this analysis (Supplementary File 3.3). 571 genomes belonged to the pneumococcal global representative dataset described above. The remaining 824 genomes belonged to 69 different *Streptococcus* species, and up to 50 genomes per species were selected for analyses from the ribosomal MLST database (https://pubmlst.org/rmlst) [271]. When more than 50 genomes were available, the population structure of the species was depicted using PHYLOViZ [272] and genomes were selected to maximise the population-level diversity of the species from the available genomes. The average GC content values were calculated within the Geneious environment.

3.3.7 Analyses of bacteriocin cluster insertion sites

The DNA sequences of each bacteriocin gene cluster were used as queries to BLAST against all other genomes the dataset using the custom BLAST feature in Geneious. The matching region plus additional flanking regions were consecutively inspected manually using the query-centric view function of the Geneious program. Using extensive manual curation, regions of DNA with different bacteriocin clusters but similar flanking genes among different genomes were identified and further inspected using the Artemis Comparison Tool (ACT) [273]. Figures illustrating multiple sequence alignments of bacteriocin cluster insertion sites were made using Geneious, ACT and Inkscape.

3.3.8 RNA sequencing analyses

3.3.8.1 Heat experiment

An RNA sequencing dataset was previously generated from a pneumococcal strain 2/2 isolate growing at a higher temperature than normal (40 *vs.* 37°C) in order to elicit a stress response (see Chapter 2). Herein, the sequenced forward and reverse reads were paired and mapped against the annotated *S. pneumoniae* strain 2/2 genome using Bowtie2 [274] with the highest sensitivity option. Differences in gene expression were calculated in Geneious using the DESeq [275] method. Genes were considered differentially expressed if the adjusted P value was <0.05. The raw transcriptomic sequence data was deposited to the GEO database and is accessible through accession number GSE103778.

3.3.8.2 Competition experiment

An existing whole-genome RNA sequencing dataset produced from two genetically distinct reference strains, PMEN3 and PMEN6, was analysed in order to explore whether bacteriocin genes are induced in response to competition for space and nutrients. The dataset was previously generated by growing the reference strains together in the same BHI broth for 6 hours. The controls were generated by growing each reference strain individually in BHI broth for 6 hours. RNA from samples were extracted at five time points (2, 3, 4, 5 and 6 hours of incubation) during growth and sequenced using the procedures described in general methods.

Due to the complexity of the generated data, i.e. sequencing reads obtained from two strains of the same species, I developed a novel computational approach to perform this analysis. Firstly, a pseudo-reference genome was created using Bowtie2, Velvet [276] and MeDuSa [277] to separate the genes only found in PMEN3 and PMEN6 plus the genes shared between the two strains (Figure 3.2A). Next, sequencing reads from all time points were combined computationally to minimise variability caused by different growth rates of strains, and were subsequently mapped to the pseudo-reference genome using Bowtie2 with the highest sensitivity option (Figure 3.2B). Finally, differential expression analyses were performed using the DESeq method by comparing reads obtained when different strains were grown together versus those obtained from when two strains where grown individually, but the reads were combined *in silico* (Figure 3.2B).

In theory, the control contained twice the amount of reads relative to the *in vivo* competition experiment as it was pooled *in silico* from two sets of samples (Figure 3.2B). Consequently, the downregulation of genes could not reliably assessed, but I could be more confident of the assessment of differential upregulation, since the levels of expression must surpass that of the *in silico* combined controls to be significant. Nevertheless, one drawback of this approach is that it only yields relative and not absolute values, and the fold-change ratios for the upregulated genes between experimental and control samples were likely underestimated. The raw transcriptomic sequence data for this analysis is accessible through accession number GSE110750.

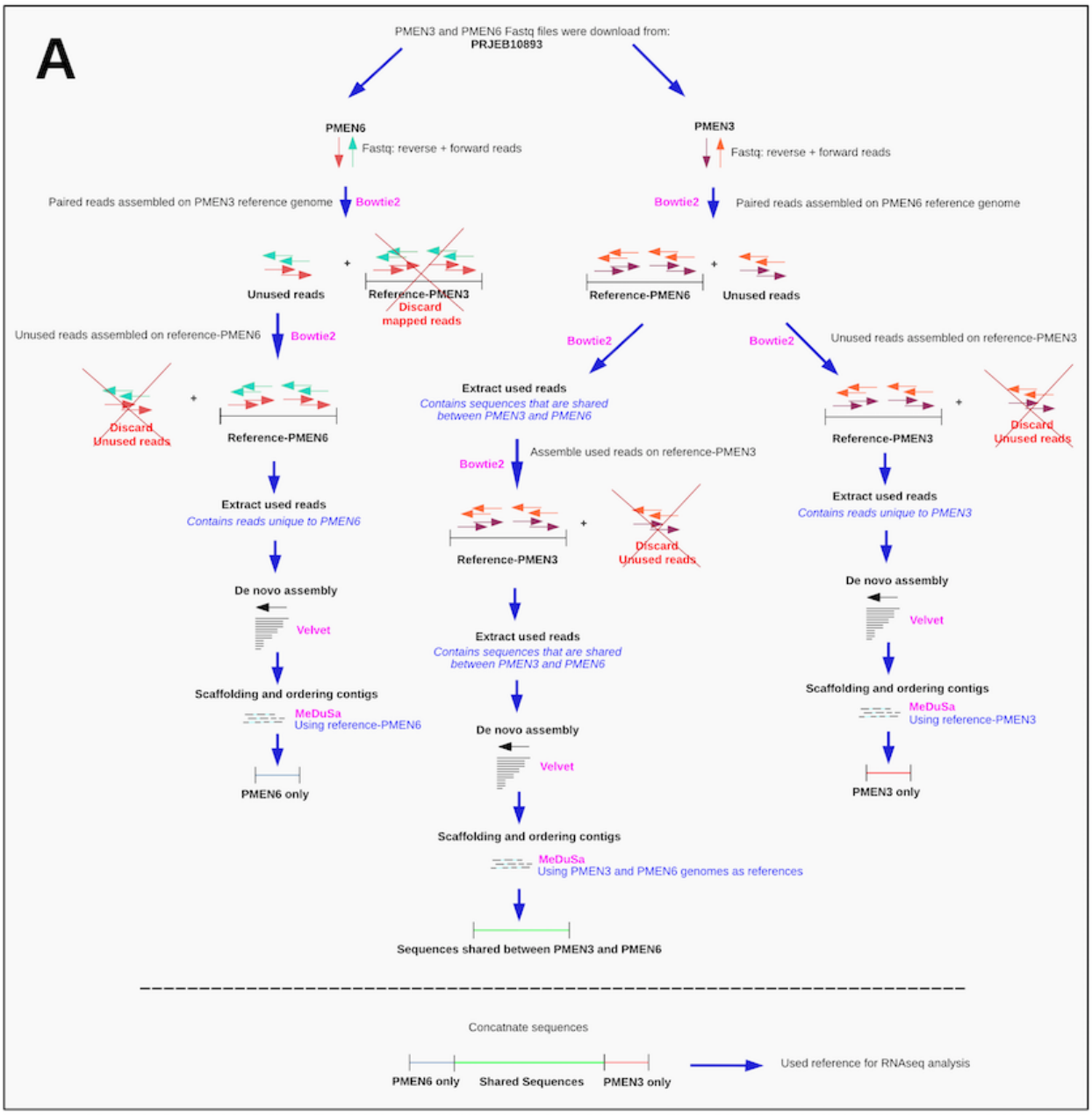


Figure 3.2 - **A visual summary of the methodology for analysing the RNA sequencing data from the co-colonisation experiment. A**, Steps involved in generating the pseudo-reference genome. The names of the tool used in each step are shown in pink. **B**, Schematic describing the combination of RNA sequence reads in order to calculate differential expression levels.

3.4 Results

3.4.1 Genome mining triples the number of known bacteriocins in *S. pneumoniae*

A large and diverse set of 571 historical and modern pneumococcal genomes (Supplementary File 3.2) were scanned for the presence of bacteriocin gene clusters, which resulted in the identification of 14 newly-discovered bacteriocin clusters, increasing the number of known bacteriocins in this species to 21 (Table 3.2). Among the identified bacteriocin clusters, several belonged to three distinct bacteriocin families of lactococcin 972, lassopeptide and sactipeptide [157, 189, 190], which were previously not known to be harboured by pneumococci. I subsequently expanded the search for bacteriocins to a much larger dataset of 5,673 published pneumococcal genomes (Supplementary File 3.1), but no additional bacteriocins were found; thus, the detailed description of the bacteriocins in this chapter is restricted to those identified in the dataset of 571 genomes. BLAST searches in the NCBI database of nucleotide collection revealed that the majority of these clusters were not exclusive to *S. pneumoniae* but were also present in other closely related streptococci (Table 3.3).

Table 3.2 - List of the bacteriocins identified among pneumococci.

Bacteriocin	Bacteriocin family[a]	Synonym(s)[b]	Newly discovered
streptococcin A	Lactococcin 972	-	Yes
streptococcin B	Lactococcin 972	-	Yes
streptococcin C	Lactococcin 972	-	Yes
streptococcin D	Lactococcin 972	-	Yes
streptococcin E	Lactococcin 972	-	Yes
streptolancidin A	Lanthipeptide (class: II)	pneumolancidin [256] and salivaricin E [278]	-
streptolancidin B	Lanthipeptide (class: II)	lcpAMT [279] and ICESp23FST81 lantibiotic [259]	-
streptolancidin C	Lanthipeptide (class: IV)	-	Yes
streptolancidin D	Lanthipeptide (class: I)	-	Yes
streptolancidin E	Lanthipeptide (class: II)	SP23-BS72 lantibiotic [200]	-
streptolancidin F	Lanthipeptide (class: IV)	-	Yes
streptolancidin G	Lanthipeptide (class: II)	phr lantibiotic [257]	-
streptolancidin H	Lanthipeptide (class: I)	-	Yes
streptolancidin I	Lanthipeptide (class: I)	-	Yes
streptolancidin J	Lanthipeptide (class: IV)	-	Yes
streptolancidin K	Lanthipeptide (class: IV)	-	Yes
streptocyclicin	Head-to-tail cyclised	pneumocyclicin [175]	-
streptolassin	Lasso peptides	-	Yes
streptosactin	Sactipeptides	-	Yes
cib	Unclassified	*cibAB* [255]	-
blp	Unclassified	*spi* and *pnc* [175]	-

The family of each bacteriocin according to Arnison *et al.* [157] (**a**) and synonym(s) for the previously identified bacteriocins (**b**) are provided.

Table 3.3 (part 1). Result of the BLAST searches of the bacteriocin clusters in the NCBI database.

streptococcin A (IS6)						
Organism	**Strain**	**Total score[a]**	**Query cover[b]**	**E value[c]**	**Identity[d]**	**Accession**
Streptococcus oralis	Uo5	3760	99.00%	0	89.00%	FR720602.1
Streptococcus mitis B6	B6	3563	89.00%	0	90.00%	FN568063.1
Streptococcus spp.	VT 162	3411	88.00%	0	89.00%	CP007628.2
Streptococcus pseudopneumoniae	IS7493	2966	99.00%	0	84.00%	CP002925.1

streptococcin B (IS6)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus mitis B6	KCOM 1350	4963	100%	0	96.00%	CP012646.1
Streptococcus pseudopneumoniae	IS7493	4290	89.00%	0	95.00%	CP002925.1

Table 3.3 (part 2). Result of the BLAST searches of the bacteriocin clusters in the NCBI database.

streptococcin C (IS6)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus pseudopneumoniae	IS7493	5234	100%	0	97.00%	CP002925.1
Streptococcus mitis	B6	4313	86.00%	0	96.00%	FN568063.1
Streptococcus mitis	KCOM 1350	4113	86.00%	0	94.00%	CP012646.1
Streptococcus spp.	oral taxon 431	3241	100%	0	82.00%	CP014264.1
Streptococcus spp.	VT 162	2802	86.00%	0	82.00%	CP007628.2

streptococcin D (IS6)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus spp.	A12	2495	88.00%	0	84.00%	CP013651.1
Streptococcus oligofermentans	AS 1.3089	2401	87.00%	0	83.00%	CP004409.1
Streptococcus gordonii	Challis	2303	88.00%	0	82.00%	CP000725.1
Streptococcus parasanguinis	ATCC 15912	2274	87.00%	0	82.00%	CP002843.1
Streptococcus parasanguinis	FW213	2239	87.00%	0	82.00%	CP003122.1

streptococcin E (IS6)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus suis	6407	4713	100%	0	95.00%	CP008921.1
Streptococcus equi	MGCS10565	3229	97.00%	0	86.00%	CP001129.1
Streptococcus equi	H70	3182	97.00%	0	86.00%	FM204884.1
Streptococcus equi	CY	3173	98.00%	0	86.00%	CP006770.1
Streptococcus equi	4047	1893	49.00%	0	87.00%	FM204883.1

streptolancidin A (PN1)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus salivarius	JH	22580	92.00%	0	97.00%	KT032116.1
Streptococcus mitis	B6	10223	46.00%	0	96.00%	FN568063.1
Streptococcus oralis	osk_001	3396	18.00%	0	97.00%	AP018338.1

streptolancidin B (USA35)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
No significant similarity found.	N/A	N/A	N/A	N/A	N/A	N/A

streptolancidin C (19F/5)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
No significant similarity found.	N/A	N/A	N/A	N/A	N/A	N/A

Table 3.3 (part 3). Result of the BLAST searches of the bacteriocin clusters in the NCBI database.

streptolancidin D (CGSP14)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus dysgalactiae	ATCC 12394	4370	99.00%	0	81.00%	CP002215.1
Streptococcus dysgalactiae	167	4298	99.00%	0	81.00%	AP012976.1
Streptococcus thermophilus	S9	4279	98.00%	0	81.00%	CP013939.1
Streptococcus thermophilus	CS8	4274	98.00%	0	81.00%	CP016439.1
Streptococcus thermophilus	CNRZ1066	4274	98.00%	0	81.00%	CP000024.1

streptolancidin E (Tennessee 23F/4)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
No significant similarity found.	N/A	N/A	N/A	N/A	N/A	N/A

streptolancidin F (45/4)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
No significant similarity found.	N/A	N/A	N/A	N/A	N/A	N/A

streptolancidin G (D39)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
No significant similarity found.	N/A	N/A	N/A	N/A	N/A	N/A

streptolancidin H (UoS3138)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus mitis	KCOM 1350	9850	37.00%	0	98.00%	CP012646.1

streptolancidin I (ERR063921)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus agalactiae	SG-M25	24421	100%	0	99.00%	CP021867.1
Streptococcus agalactiae	C001	24408	100%	0	99.00%	CP008813.1
Streptococcus suis	JS14	24393	100%	0	99.00%	CP002465.1
Streptococcus suis	SC070731	24387	100%	0	99.00%	CP003922.1
Streptococcus pasteurianus	ATCC 43144	23878	100%	0	99.00%	AP012054.1
Streptococcus uberis	42	16610	96.00%	0	93.00%	DQ146939.1

streptolancidin J (1/5)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
No significant similarity found.	N/A	N/A	N/A	N/A	N/A	N/A

streptolancidin K (ERR054267)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus pseudopneumoniae	IS7493	2250	29.00%	0	98.00%	CP002925.1

Table 3.3 (part 4). Result of the BLAST searches of the bacteriocin clusters in the NCBI database.

streptolancidin K (ERR054267)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus pseudopneumoniae	IS7493	2250	29.00%	0	98.00%	CP002925.1

streptosactin (VICE0913)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus intermedius	B196	11329	100%	0	99.00%	CP003857.1
Parvimonas micra	KCOM 1535	11318	100%	0	99.00%	CP009761.1

Streptolassin (37/3)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus mitis	SVGS_061	15187	100%	0	99.00%	CP014326.1
Streptococcus suis	SC19	12331	99.00%	0	93.00%	CP020863.1
Streptococcus suis	SS2-1	12331	99.00%	0	93.00%	CP018908.1
Streptococcus suis	ZY05719	12331	99.00%	0	93.00%	CP007497.1
Streptococcus equinus	FDARGOS	12191	47.00%	0	95.00%	CP020439.1

Streptocyclicin (2/2)						
Organism	**Strain**	**Total score**	**Query cover**	**E value**	**Identity**	**Accession**
Streptococcus pseudopneumoniae	IS7493	5339	100%	0	99.00%	CP002925.1

a, The total score is the sum of scores of all aligned sequences. **b,** Query coverage refers to the percent of the query length that is included in the aligned segments. **c,** E-value denotes the expected number of chance alignments: the smaller the E-value, the less likely the resulting alignment was assembled by chance. **d**, Identity value is the percentage of perfect matches between subject and query.

Further support for the interspecies exchange of bacteriocin clusters was provided by investigating the DNA base composition of the clusters. The average guanine-cytosine (GC) content of pneumococcal genomes in the dataset was 39.6%, however, the average GC content for different bacteriocin groups ranged from: streptococcins, 36.3–42.4%; streptolancidins subset 1, 31.2–33.4%; streptolassin, 31.2%; streptolancidins subset 2, 28.9–29.6%; streptocyclicin, 27.0%; and streptosactin, 25.5%. For comparison, the average GC content of other non-pneumococcal *Streptococcus* species was calculated using 824 genomes of 69 different *Streptococcus* species, and varied from 33.2 to 44.6% (Figure 3.3).

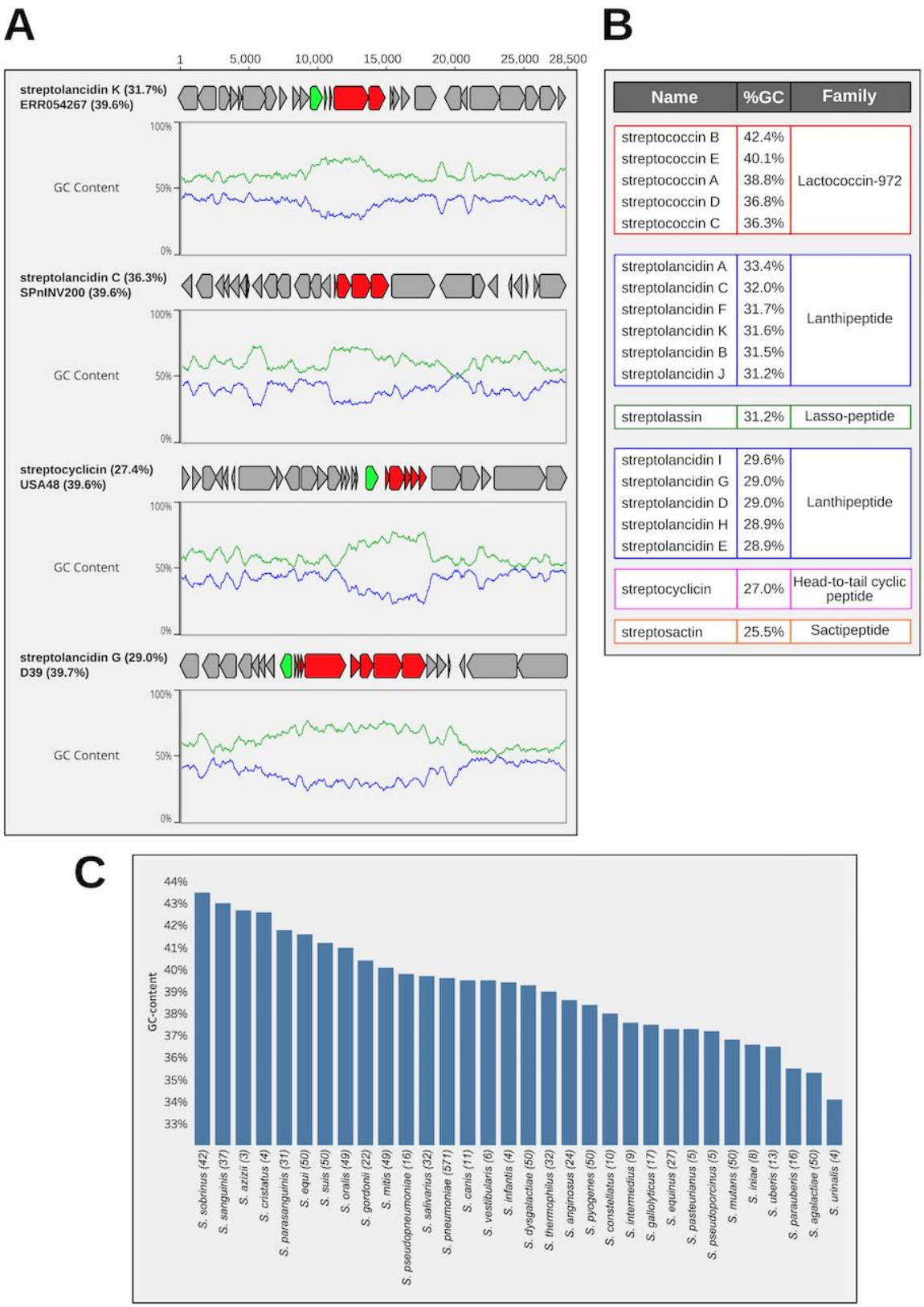


Figure 3.3 – GC content of pneumococcal bacteriocin clusters. A, Four examples of GC plots showing the percentage GC content of bacteriocin cluster genes (red), transcriptional regulator genes (green), and other adjacent pneumococcal genes (grey). The names of the bacteriocin and the genome harbouring it are provided, with the percentage GC content of each in brackets. The green and blue lines in each graph depicts adenine (A) thymine (T) content, respectively. **B,** Average GC content for each bacteriocin cluster type, grouped by bacteriocin type or class. The lanthipeptides formed two subsets based on GC content of <30 or >31%. **C,** Average GC content for genomes of non-pneumococcal streptococcal species. The number of genomes for each species used in the analyses are given in brackets.

Five streptococcins were identified among pneumococci. Despite gene synteny among the streptococcins (Figure 3.4), nucleotide sequence similarity was low: bacteriocin genes, 37–59%; immunity genes, 27–48%; and transporter genes, 49–63% (Figure 3.5). Different streptococcins were found in different but consistent locations within the bacterial chromosome and one example of this is illustrated in Figure 3.6. Furthermore, eleven different streptolancidins, one streptosactin, one streptocyclicin and one streptolassin were also present among pneumococcal genomes (Figure 3.7). The amino acid sequences of bacteriocins from these clusters shared very little sequence similarity to other known bacteriocins (Figure 3.8).

There was a pattern in the genes flanking the pneumococcal bacteriocin clusters: genes commonly present in the flanking regions were predicted to encode CAAX proteins (thought to be involved in self-immunity from bacteriocin toxins [201]), Rgg and PlcR transcriptional regulators (thought to be involved in bacterial quorum sensing [280]), transporters and mobile element proteins (Figure 3.4 and 3.7).

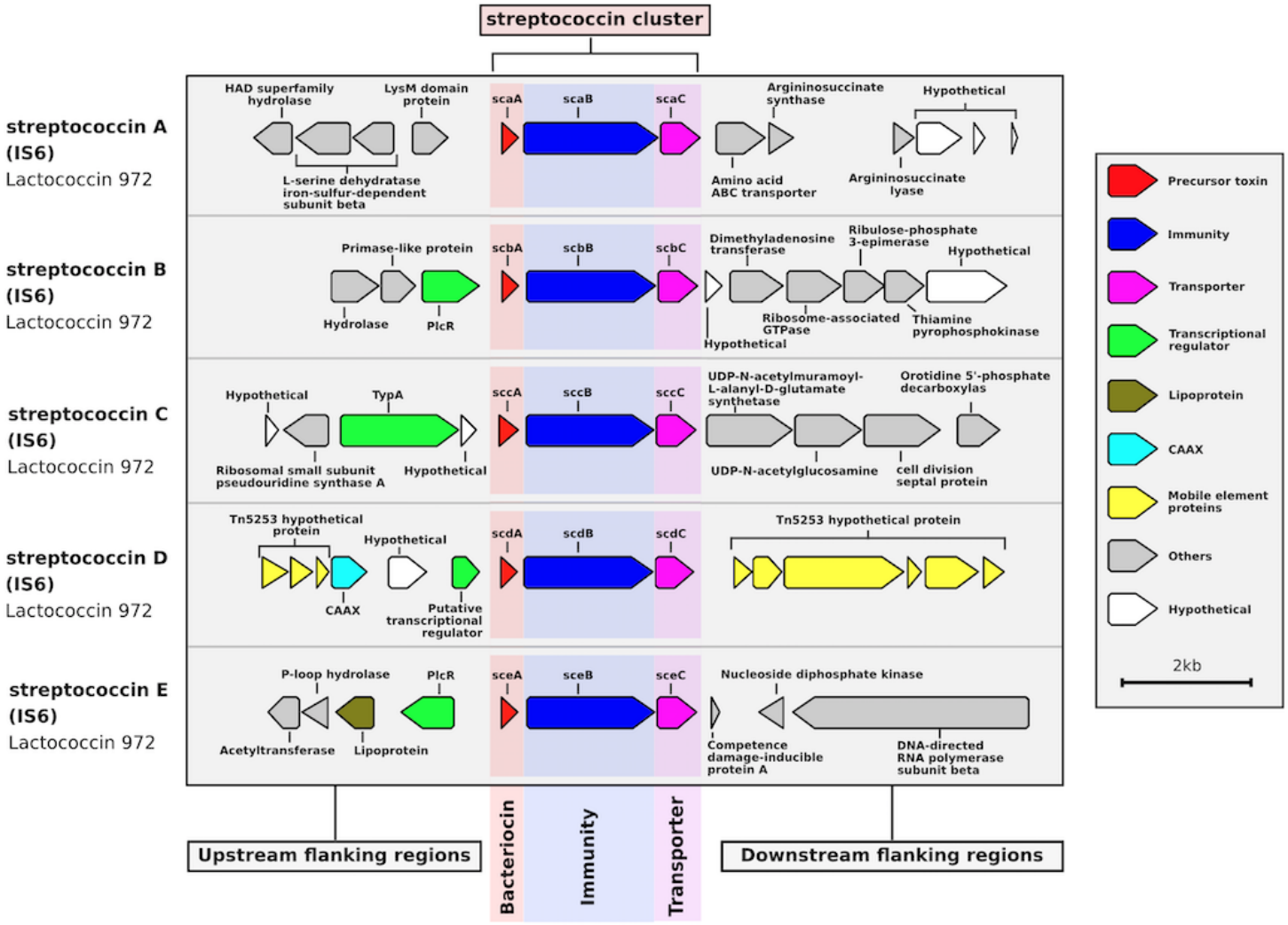


Figure 3.4 - A schematic representation of each streptococcin cluster and their flanking regions. The coding regions were derived from the genome of pneumococcal strain IS6, which contains five different streptococcin clusters.

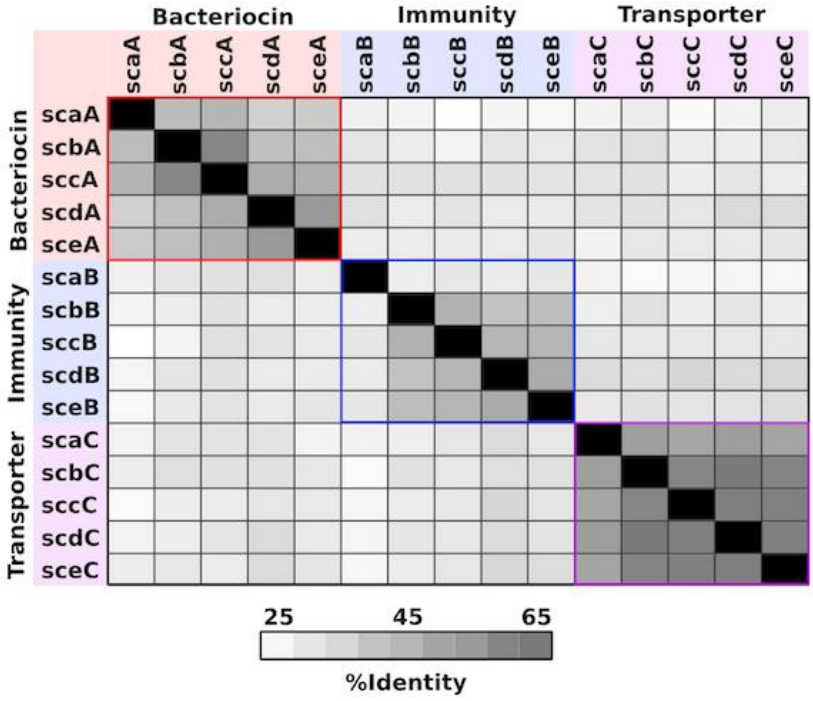


Figure 3.5 - A distance matrix of the nucleotide sequence identity shared between genes of different streptococcin clusters. The coding sequences were derived from the genome of pneumococcal strain IS6, which possess five different streptococcin clusters.

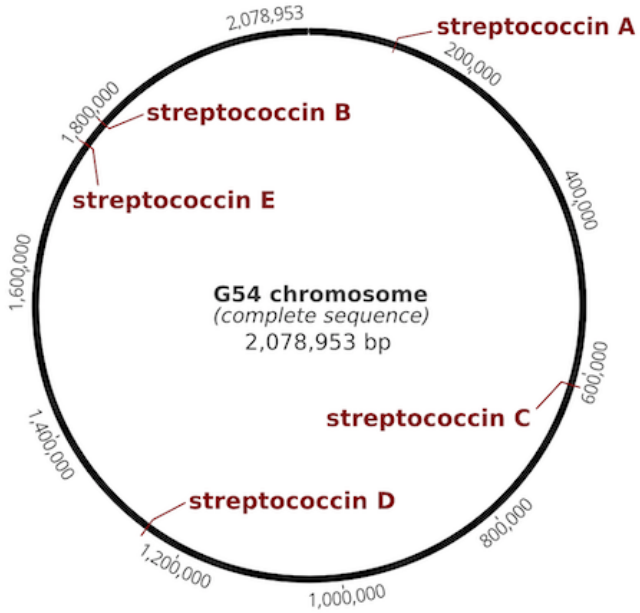


Figure 3.6 - A schematic representation of the finished genome of pneumococcal strain G54 with the locations of the five streptococcin clusters highlighted in red.

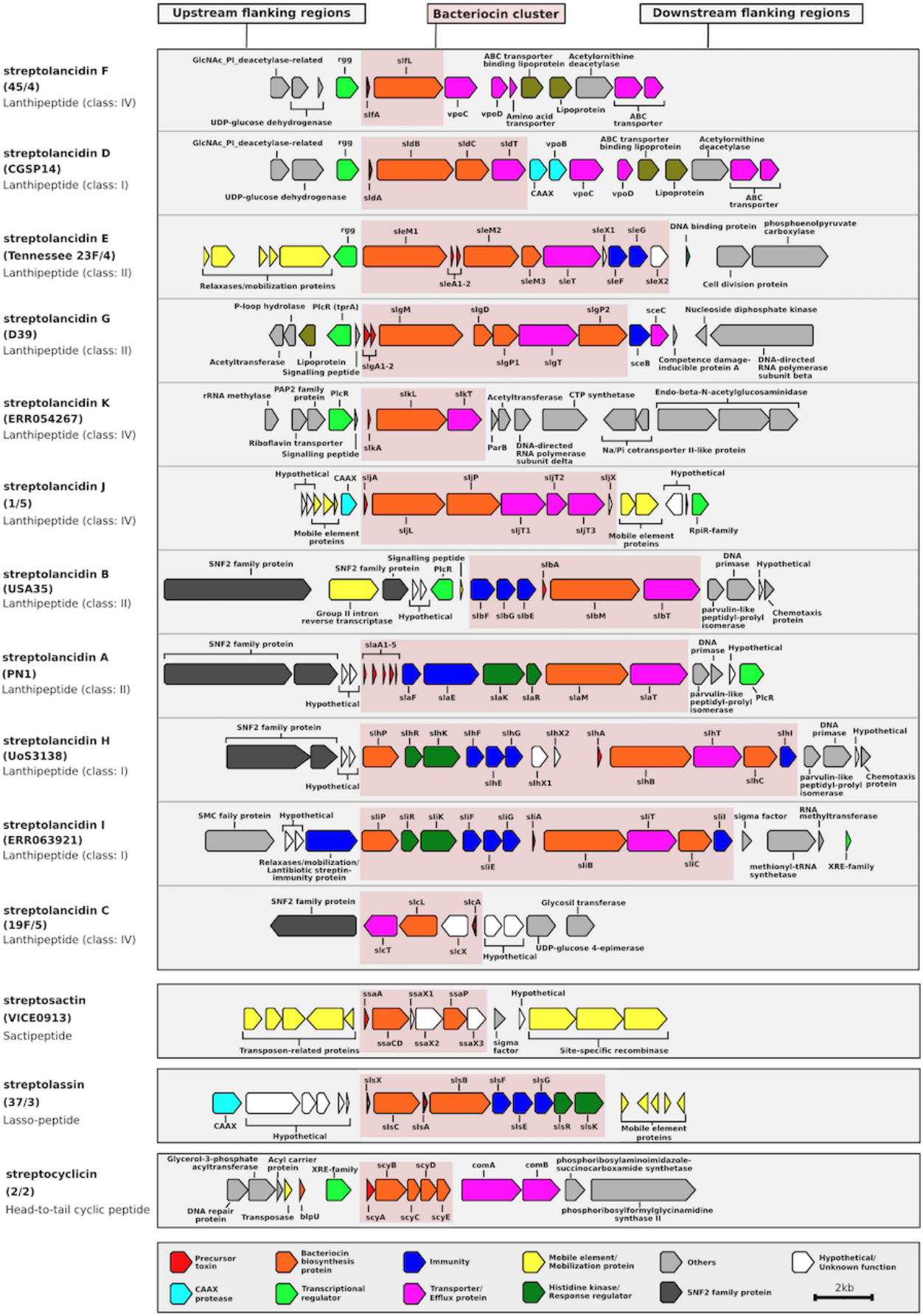


Figure 3.7 - Genetic composition of streptolancidin, streptocyclicin, streptosactin, and streptolassin bacteriocins in the pneumococcal genomes. The names of the isolates containing each bacteriocin cluster are given in brackets and the class of each bacteriocin cluster is given underneath the names of isolates.

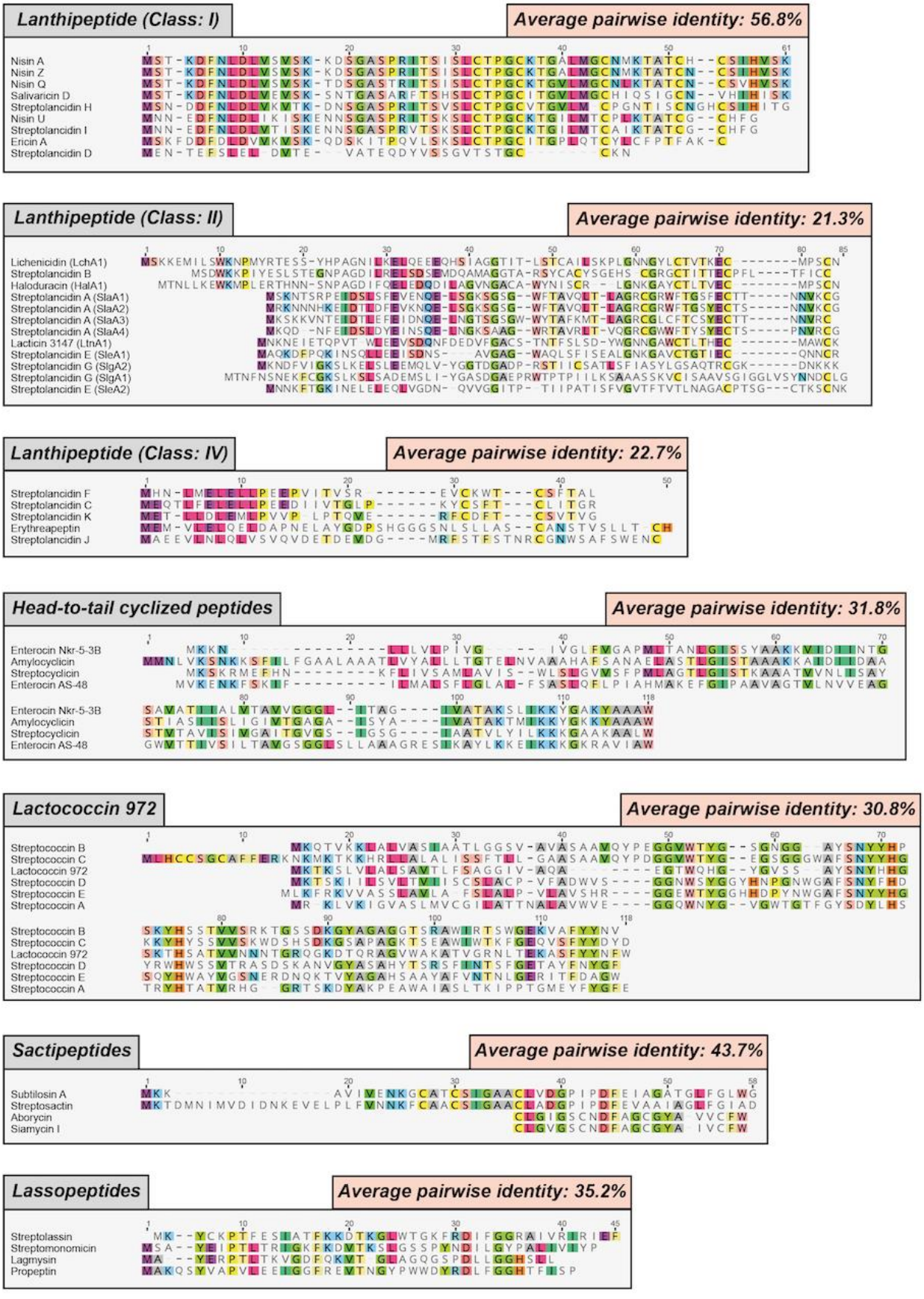


Figure 3.8 - Predicted bacteriocin genes identified in this study aligned against similar bacteriocin genes in other bacterial species for comparison.

3.4.2 Bacteriocin heterogeneity within a global pneumococcal population dataset

I further assessed these bacteriocins in the context of the pneumococcal population structure. The study dataset consisted of a diverse collection of 571 pneumococci isolated between 1916 and 2009 from patients and healthy individuals of all ages residing in 39 different countries across six continents. Eighty-eight pneumococcal serotypes and 99 different clonal complexes were represented (Table 3.4 and Supplementary File 3.2). All bacteriocins detected more than once in the dataset were identified among pneumococci isolated over several decades and from a variety of different countries (Table 3.4). Some bacteriocins were found to be ubiquitous among all genomes in the dataset, while others were only detected in certain clonal complexes; however, bacteriocins found in more than one genome were present in pneumococci isolated over several decades from various countries (Table 3.4). The number of bacteriocin clusters present in each genome varied widely among pneumococci, ranging from 6 to 11 (Figure 3.9). Nevertheless, certain combinations of bacteriocins were more commonly represented than others (Figure 3.10).

Bacteriocin clusters lacking one or more genes compared to the largest clearly described cluster were deemed as partial. The percentages of partial and complete clusters varied between different bacteriocins (Figure 3.11). For instance, streptococcin E was found in nearly all pneumococci (99.8%) but was partial in the majority (83.5%) of cases, while streptococcin C was present in all genomes as a complete cluster (Table 3.4).

Table 3.4 – Molecular epidemiology of the bacteriocin clusters identified among a dataset of 571 pneumococci recovered since 1916 from patients of all ages residing in 39 different countries.

Bacteriocin	Genome(s)			Year(s) of Isolation	Countries (n)	Clonal complexes* (n)	Serotypes (n)
	Complete	Partial	Total				
streptococcin A	374 (65.5%)	31 (5.4%)	405 (70.9%)	1916 - 2009	36	80	76
streptococcin B	415 (72.7%)	156 (27.3%)	571 (100%)	1916 - 2009	39	99	88
streptococcin C	571 (100%)	0 (0.0%)	571 (100%)	1916 - 2009	39	99	88
streptococcin D	6 (1.1%)	0 (0.0%)	6 (1.1%)	1968 - 2005	5	3	4
streptococcin E	74 (13.0%)	496 (86.9%)	570 (99.8%)	1916 - 2009	39	98	88
streptolancidin A	1 (0.2%)	1 (0.2%)	2 (0.4%)	1972 - 2006	2	2	2
streptolancidin B	45 (7.9%)	48 (8.4%)	93 (16.3%)	1939 - 2006	16	11	10
streptolancidin C	73 (12.8%)	213 (37.3%)	286 (50.1%)	1937 - 2009	27	52	52
streptolancidin D	49 (8.6%)	0 (0.0%)	49 (8.6%)	1938 - 2006	13	13	11
streptolancidin E	3 (0.5%)	161 (28.2%)	164 (28.7%)	1937 - 2009	25	21	38
streptolancidin F	23 (4.0%)	0 (0.0%)	23 (4.0%)	1937 - 2006	7	4	11
streptolancidin G	186 (32.6%)	13 (2.3%)	199 (34.9%)	1916 - 2009	25	38	46
streptolancidin H	1 (0.2%)	0 (0.0%)	1 (0.2%)	2006-2008	1	0	1
streptolancidin I	1 (0.2%)	0 (0.0%)	1 (0.2%)	2009	1	1	1
streptolancidin J	140 (24.5%)	237 (41.5%)	377 (66.0%)	1916 - 2009	33	64	74
streptolancidin K	1 (0.2%)	0 (0.0%)	1 (0.2%)	2009	1	0	1
streptocyclicin	205 (35.9%)	4 (0.7%)	209 (36.6%)	1937 - 2009	20	46	59
streptolassin	20 (3.5%)	0 (0.0%)	20 (3.5%)	1939 - 1996	8	7	10
streptosactin	1 (0.2%)	0 (0.0%)	1 (0.2)	2009	1	1	1
cib	551 (96.5%)	14 (2.5%)	557 (97.5%)	1916 - 2009	39	97	88
blp	N/A	N/A	571 (100%)	1916 - 2009	39	99	88

*****Singletons (single genotypes with no closely related variant) were excluded from the count.

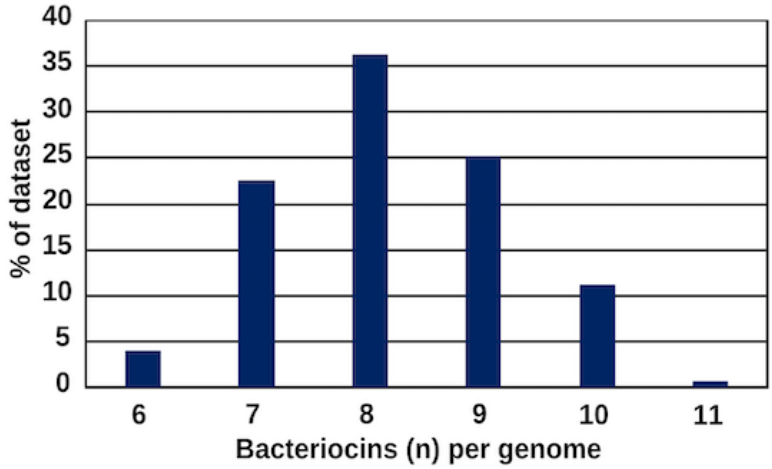


Figure 3.9 - The number of bacteriocins per genome among the 571 pneumococcal genomes.

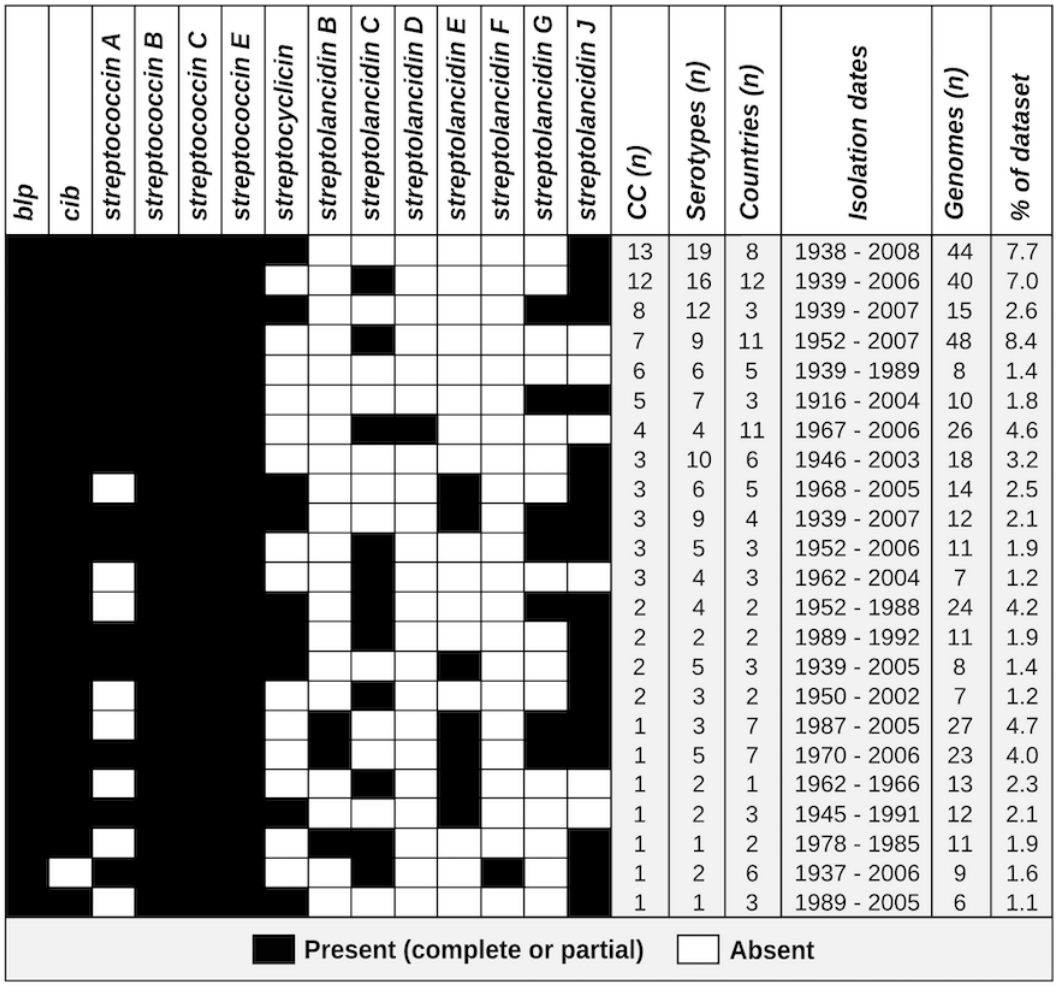


Figure 3.10 - Bacteriocin combinations found within a global pneumococcal dataset. Black and white boxes indicate the presence or absence of each bacteriocin, respectively. The epidemiological characteristics of the pneumococci that possessed each combination of bacteriocins are provided in the columns in the right side of the figure.

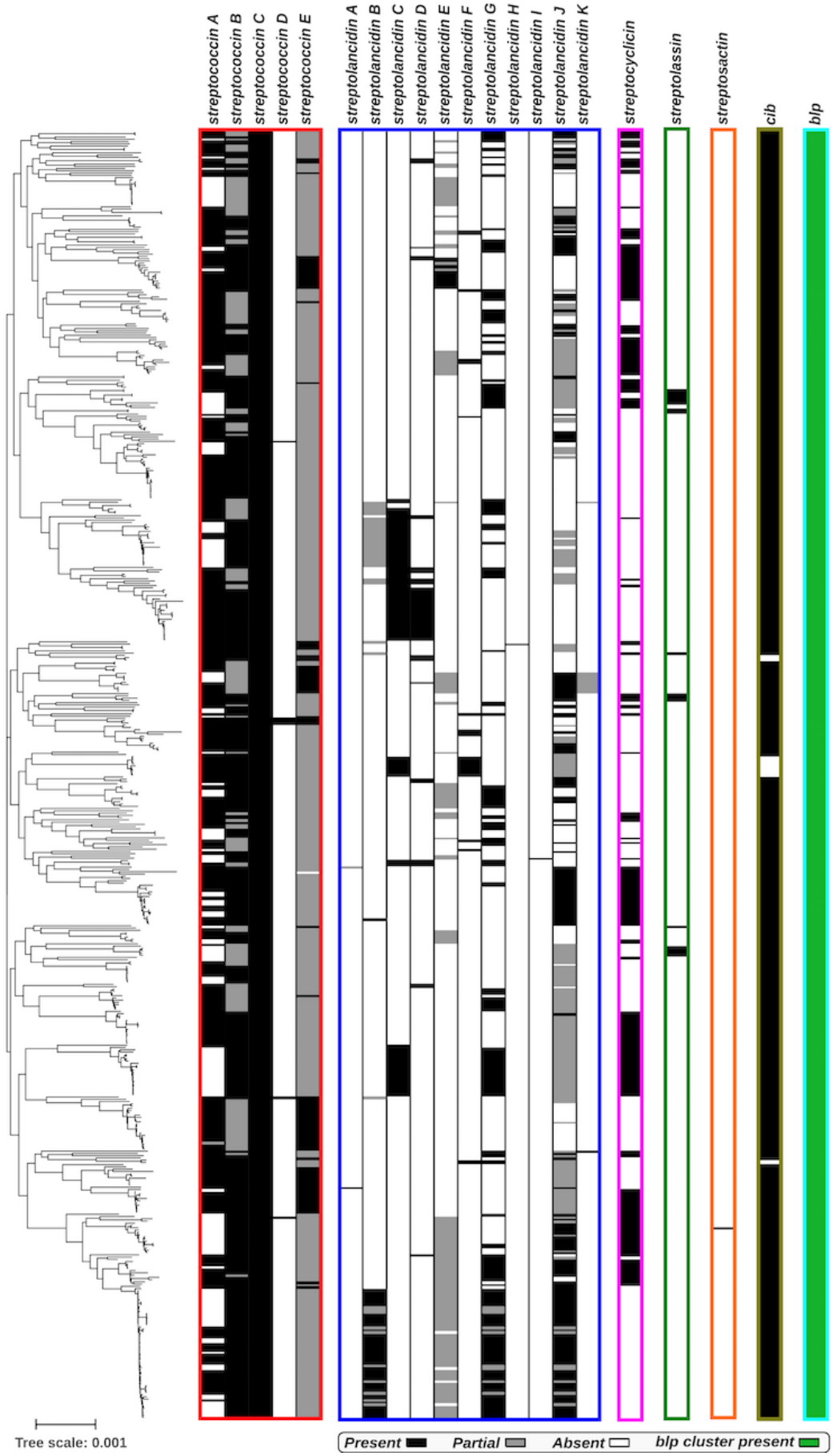


Figure 3.11 - Diversity of bacteriocins within a global pneumococcal dataset. A phylogenetic tree of the genomes in the study dataset labelled according to the presence of different bacteriocins. Due to complicated genetic composition of the *blp* clusters [175], a similar classification between partial (those with missing genes) and complete clusters could not be applied. Instead, their presence (irrespective of being partial or complete) is depicted by green.

Owing to their relatively simple genomic structure (being composed of only three genes), streptococcins were selected as models for further scrutinising the pattern of missing genes in partial bacteriocin clusters. A detailed investigation of streptococcins showed that the majority of the partial clusters do not harbour the gene encoding the bacteriocin pre-peptide, while the immunity and/or the transporter genes are still retained (Table 3.5). This observation is in accordance with the “cheater”- hypothesis, which suggests that some strains retain the immunity and transporter genes in order to protect themselves from neighbouring bacteria that express the bacteriocin, without bearing the cost of toxin production [252, 281].

Table 3.5 - The overall prevalence of the streptococcin gene clusters.

	streptococcin A	*streptococcin B*	*streptococcin C*	*streptococcin D*	*streptococcin E*
ABC	374	415	571	6	74
AB	0	0	0	0	0
AC	0	0	0	0	0
BC	31	154	0	0	300
A	0	0	0	0	0
B	0	0	0	0	0
C	0	2	0	0	196
Total	**405**	**571**	**571**	**6**	**570**

The patterns of gene presence (column 1: A, bacteriocin; B, immunity; C, transporter) and frequency of those patterns (indicated by numbers) among streptococcins A–E within the study dataset are provided.

3.4.3 Genomic hotspots for the integration of different bacteriocin gene clusters

I investigated the flanking regions of each bacteriocin gene cluster among genomes in the study dataset, which led to the identification of three specific locations in the pneumococcal chromosome that are putative hotspots for the integration of bacteriocin clusters. These loci, which I termed bacteriocin cluster hotspots (BCHs), consisted of regions of DNA in the genome where different bacteriocin clusters with identical flanking genes were detected in different isolates (Figure 3.12). The presence of multiple BCHs among pneumococcal genomes suggested a switching mechanism whereby different bacteriocin clusters may replace one another through genetic recombination.

Up to three different bacteriocin clusters were found within a single BCH (Figure 3.12). The acquisition of streptolancidin G appeared to have resulted in a partial streptococcin E cluster by replacement of the bacteriocin pre-peptide gene and part of the gene encoding the bacteriocin immunity protein (Figure 3.12A). Streptolancidin G was not detected in genomes that possessed a complete streptococcin E cluster (Figure 3.11), whereas the remnant genes of streptococcin E partial clusters were conserved in pneumococcal genomes isolated over nearly a century (Figure 3.11 and Table 3.5).

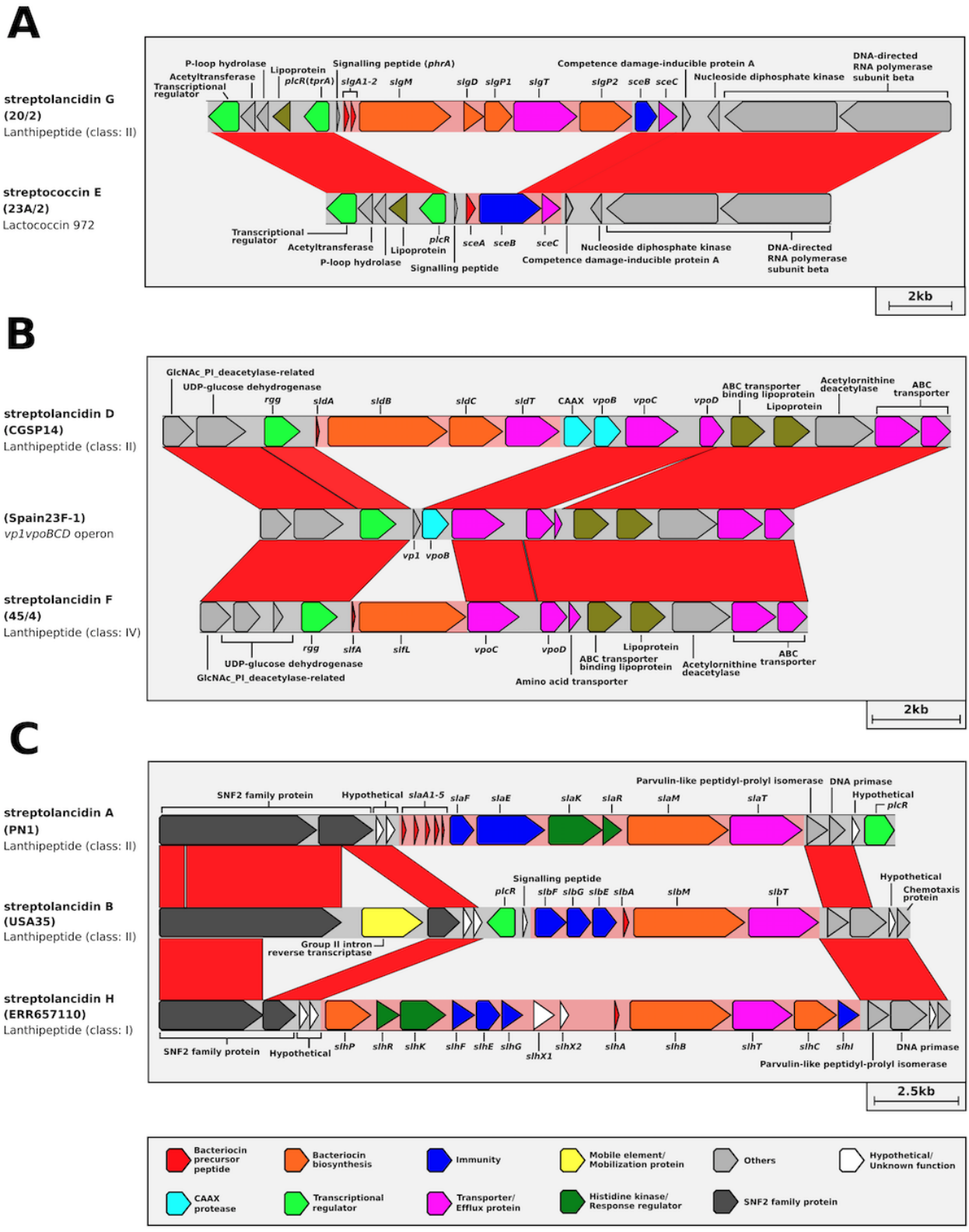


Figure 3.12 - Whole genome-based population analysis reveals evidence for bacteriocin switching among pneumococci. Bacteriocin cluster hotspots (BCHs) were defined as regions of DNA where different bacteriocin clusters with identical flanking genes were found among pneumococcal genomes. Linear comparisons of (**A**) BCH-1, (**B**) BCH-2, and (**C**) BCH-3 are displayed. The isolate names are shown in brackets. The family of each bacteriocin cluster is provided underneath the isolate name.

3.4.4 Transcriptome analyses demonstrate temporal dynamic changes in the expression of bacteriocin genes in response to heat

An existing whole-genome RNA sequencing dataset generated from a pneumococcal strain 2/2 isolate growing in broth culture at a higher incubation temperature than normal (40 *vs.* 37°C) (NCBI GEO accession number GSE103778) was investigated in order to assess whether the bacteriocin clusters are transcriptionally active. A number of genes belonging to multiple bacteriocin clusters were found to be differentially expressed compared to the control over several time points (Figure 3.13), revealing that many of these bacteriocin genes were likely to be transcribed in response to external stress.

The strain 2/2 harbours two partial bacteriocin clusters (streptolancidin J and streptococcin E). Despite lacking several genes, the remaining genes in these clusters were found to be upregulated in response to heat and furthermore, the timing of gene expression differed among bacteriocin clusters, with some being induced at an earlier time point than others, while several were up-regulated concurrently (Figure 3.13). The regulatory mechanisms involved in modulating the expression of bacteriocin genes remain to be investigated. A full list of the differentially expressed genes and their sequence is provided in the Supplementary File 3.4.

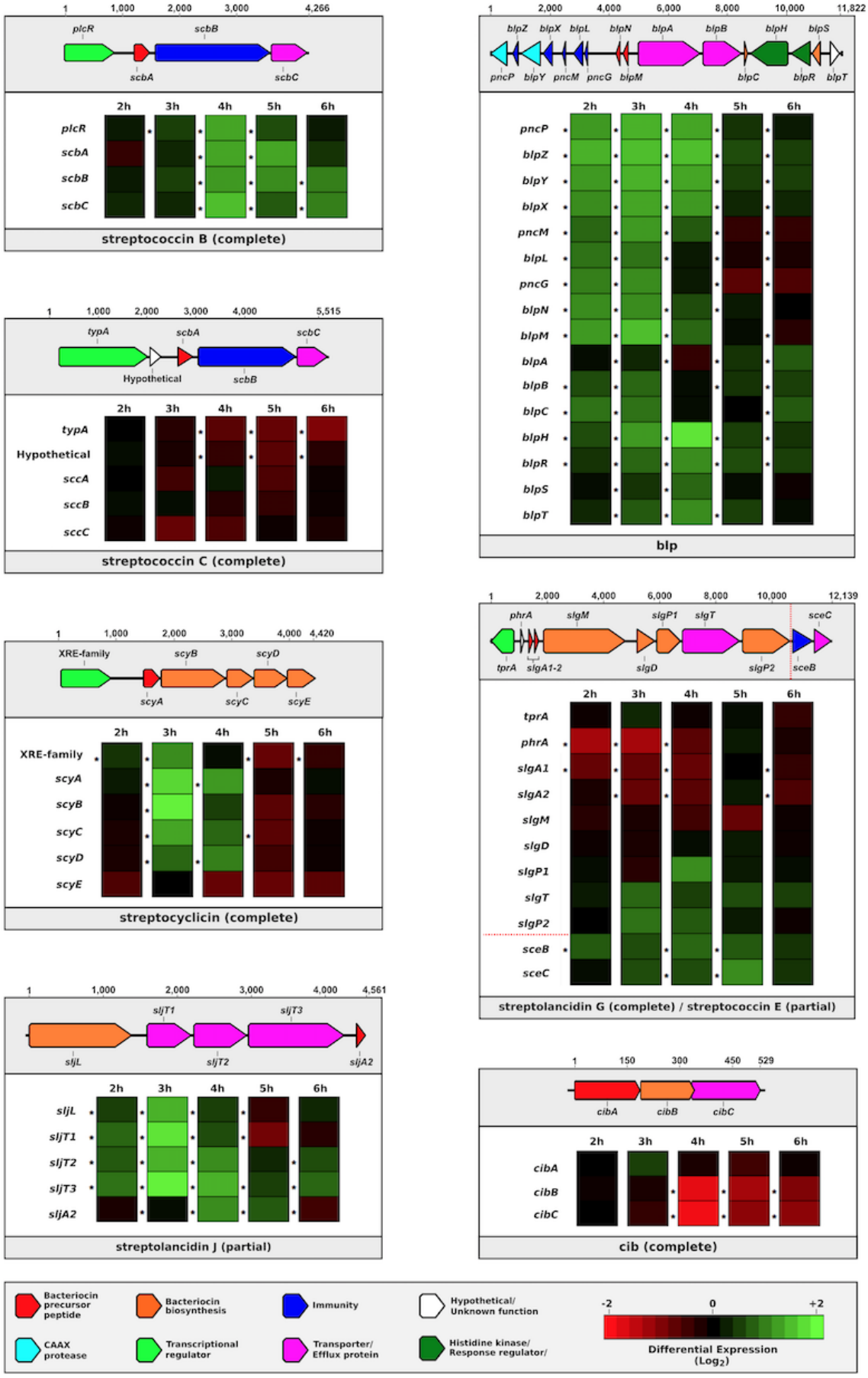


Figure 3.13 - Dynamic changes in the expression of bacteriocin genes in response to bacterial stress. The differential expression levels for each bacteriocin clusters found in pneumococcal strain 2/2 when incubated at 40°C *vs.* 37°C are shown. Genes are represented by rows and differential expression levels at different time points are indicated in columns. An asterisk to the left of a cell indicates a statistically significant differential expression level ($p < 0.05$).

3.4.5 Bacteriocins are induced in response to strain competition

I analysed an existing whole-genome RNA sequencing dataset (NCBI GEO accession number GSE110750) generated from two genetically distinct reference strains, PMEN3 and PMEN6 in order to explore whether bacteriocin genes are induced in response to competition for space and nutrients. Many pneumococcal genes were found to be significantly upregulated when the two strains were co-cultured compared to when cultured individually including those that would be expected during growth (*e.g.* metabolic genes), as well as 29 different bacteriocin genes (Figure 3.14 and Supplementary File 3.5).

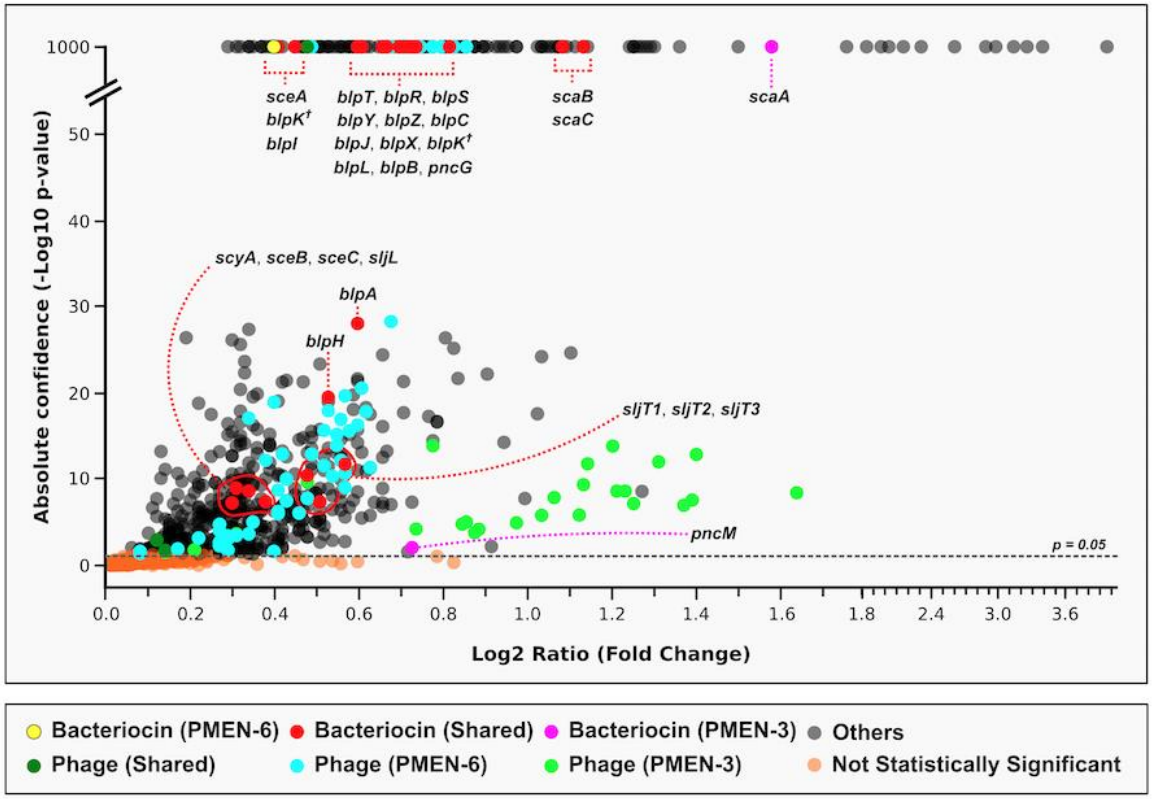


Figure 3.14 - Evidence for the upregulation of bacteriocin genes when two reference strains, PMEN3 and PMEN6, were co-cultured in broth media. Only genes that were upregulated compared to the controls (strains cultured individually) are displayed in the figure. Genes of interest that were unique to each reference strain and those shared by both strains are marked in different colours. Two copies of *blpK* were found in different locations (one in the *blp* cluster as expected and one elsewhere in the genome) in both genomes and are marked here with a cross. A full list of the upregulated genes can be found in the Supplementary File 3.5.

Since the genomes of both PMEN3 and PMEN6 were already sequenced, the original intention was to map the RNA sequencing reads generated for each strain back to their corresponding genome sequence; however, many of the bacteriocin genes were present in similar allelic versions in both strains and therefore it was not possible to establish with confidence whether the gene was overexpressed in only one strain or whether both strains upregulated similar genes. Nevertheless, a number of bacteriocin genes were present in only one of the two test strains: the gene encoding the bacteriocin toxin of streptococcin A (*scaA*) and a putative immunity gene (*pncM*) from the *blp* bacteriocin cluster were found to be significantly upregulated in PMEN3; whereas PMEN6 significantly upregulated the gene encoding the bacteriocin toxin of streptococcin E (*sceA*). The list of significantly upregulated genes in the reference strains also included genes associated with prophages (Figure 3.14 and Supplementary File 3.5).

3.5 Discussion

A clear understanding of the role bacteriocins play in pneumococcal biology is central to understanding microbial interactions within the ecological niche (the nasopharynx). The importance of intraspecies competition to pneumococcal ecology is reflected in the changes in prevalence of different pneumococcal serotypes and genotypes in the nasopharynx over time, and understanding competition dynamics is important in the context of understanding vaccine impact [282, 283]. As discussed in the general introduction (see 1.2.7.4), pneumococcal vaccines are disruptive to the pneumococcal population structure and alter the composition of microbes competing for space and nutrients in the nasopharynx.

The effects of this disruption are not yet fully understood but can lead to increased disease in human populations [284-286].

The findings presented in this chapter underscore the extraordinarily complexity of the bacteriocins possessed by *S. pneumoniae*. This study revealed that not only do pneumococci possess a substantially greater and more varied array of bacteriocins than previously recognised, the bacteriocins (often in a particular combination) are associated with specific clonal complexes. This is fundamental, as it provides the framework on which to investigate the mechanisms underpinning specific bacteriocin-pneumococcus combinations, particularly among epidemiologically successful clonal complexes, and the activity of specific bacteriocin and immunity genes. Additionally, the 14 newly-discovered bacteriocins as part of this study are potential candidates for further investigation as novel antimicrobials. In fact, several of the bacteriocins identified as part of this chapter are currently in the process of being synthesised and tested in collaboration with an industrial partner. This work also provides a unified and easily accessible nomenclature for the pneumococcal bacteriocin clusters and their genes.

The number of genomes used in the population-based analyses made it possible to more confidently deduce the set of genes expected to be present in a complete bacteriocin gene cluster. This is valuable for designing future studies that assess their antimicrobial potential, as it reduces the risk of inadvertently choosing an isolate that harbours a partial cluster (missing genes necessary for bacteriocin biosynthesis). Furthermore, the information regarding the prevalence of bacteriocins is insightful in informing decisions as to which bacteriocin cluster might be the best candidate for drug development; for instance, if a bacteriocin is rarely found in the

natural population, fewer bacteria are likely to be resistant to it; additionally, a bacteriocin that is extremely conserved and ubiquitously present among all pneumococci, such as streptococcin C, is likely to play an important role in the biology of the pneumococcus, and therefore may serve as a potential drug target.

Several locations in the pneumococcal chromosome were identified where different bacteriocin clusters were found among different isolates, suggesting a switching mechanism, whereby pneumococci can exchange bacteriocin clusters via recombination events (Figure 3.12). Recombination is well documented to occur at several other locations in the pneumococcal genomes. For instance, genetic exchange between genes at the capsular polysaccharide locus occurs frequently and can lead to changes in serotype, a phenomenon known as serotype switching [99, 100, 103]. Likewise, the DpnI, DpnII and DpnIII clusters, each possessing a unique restriction modification system, can replace one another at the *dpn* locus, which is implicated in providing protection against phage attack at the population level [287, 288].

The current study identified multiple examples of the proposed bacteriocin cluster switching events that had taken place adjacent to different quorum-sensing transcriptional regulators TprA and Rgg. Interestingly, although the genes known to be under the regulation of TprA and Rgg were replaced, the quorum-sensing transcriptional regulators remained conserved (Figure 3.12A-B). As described above (see 3.2), TprA is shown to regulate the expression of its downstream bacteriocin genes according to the levels of galactose and glucose present in the environment, which is thought to facilitate nasal colonisation by helping pneumococci to compete for resources [257, 289]. Similarly, Rgg is known to

control the expression of its adjacent genes, and this is thought to be mediated by sensing amino acid levels in the cellular community [290]. An explanatory hypothesis might be that bacteriocin cluster switching provides a mechanism by which the existing intricate quorum-sensing signalling network required for coordinating population-level behaviours is accessible by the newly acquired bacteriocin cluster. This remains to be experimentally verified.

RNA sequencing revealed that bacteriocin genes are transcriptionally active when pneumococci are under stress or competing with different strains during bacterial co-culture, which must be considered when designing laboratory experiments aimed at assessing the role of an individual bacteriocin. Moreover, the data from the co-culture experiment provides proof-of-concept that such a method can be utilised to investigate the differential expression of important bacterial genes when different strains of the same species are in competition for limited resources. It remains to be shown whether functional bacteriocin proteins are produced *in vivo*, and further experiments are needed to evaluate their antibacterial activities.

Interestingly, other genes that were significantly upregulated when pneumococci were competing during bacterial co-culture were genes associated with unique prophages present in each of the PMEN strains (Figure 3.14). The extent to which prophages are influencing pneumococcal biology and perhaps competition between strains is not yet understood and needs further investigation. Having characterised and revealed a large and diverse repertoire of bacteriocins among streptococci in this chapter, the two result chapters that follow move on to characterise and investigate prophages of pneumococci and other members of the *Streptococcus* genus (see Chapter 4 and 5).

3.6 Supplementary information

Supplementary File 3.1. List of pneumococcal genomes used for the genome mining of bacteriocins.

Supplementary File 3.2. Descriptive data for the genomes included in the global representative dataset.

Supplementary File 3.3. List of genomes used for calculating the GC content of *Streptococcus* species.

Supplementary File 3.4. RNA sequencing data from the heat experiment.

Supplementary File 3.5. RNA sequencing data from the competition experiment.

4. Genomic investigation and molecular epidemiology of satellite prophages in *S. pneumoniae*

The findings described in this chapter were presented in part at two conferences: i) in an oral presentation at the 11th International Symposium on Pneumococci and Pneumococcal Diseases, in Melbourne, Australia in April 2018; and ii) in an oral presentation at the 14th European Meeting on the Molecular Biology of the Pneumococcus, in Greifswald, Germany in June 2019.

This chapter and chapter 5 were combined into one manuscript, which has been accepted for publication in *Nature Communications* (in press). A pre-print version was published in bioRxiv in December 2018 as:

Reza Rezaei Javan, Elisa Ramos-Sevillano, Asma Akter, Jeremy Brown and Angela B Brueggemann. 2018. Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis. *bioRxiv* [Preprint]. doi: 10.1101/203398

All of the bacterial genomes used in this chapter were retrieved from the Brueggemann BIGSdb database as described in Chapter 3. A large-scale prophage study was published by the Brueggemann group at the start of my DPhil work and one of my contributions to this paper was the analysis of the RNA sequencing data. The RNA sequencing dataset was generated in the Brueggemann laboratory by Dr Andries van Tonder, Caroline Harrold and Prof Angela Brueggemann. Extracted RNA samples were sent for sequencing to the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics. I used the raw RNA sequencing reads generated from these experiments, assembled them to the reference genomes, and performed

the differential expression analyses for the transcriptomic analyses that were presented in the published paper (data presented in Figure 4.1):

Angela B. Brueggemann, Caroline L. Harrold, Reza Rezaei Javan, Andries J. van Tonder, Angus J. McDonnell and Ben A. Edwards. 2017. Pneumococcal prophages are diverse, but not without structure or history. *Scientific Reports* 7:42976. doi: 10.1038/srep42976

The *Scientific Reports* paper focused on full-length prophages and only performed limited analyses of the 'partial prophages', which were of interest but the large volume of data meant we focused on primarily characterising and reporting the full-length prophages first. The 403 partial phage sequences identified in the study were extracted via manual curation by Prof Angela Brueggemann and stored in a BIGSdb database, and I reassessed these in the work described here in this chapter.

A second RNA sequencing dataset was investigated as part of my work on satellite prophages, and these transcriptomic data were generated and published by Blanchette *et al* [240]. I downloaded the raw RNA sequencing data from the GEO repository and reanalysed the Blanchette data as described in this chapter.

The laboratory work that was performed to assess the role of the putative *vapE* virulence gene and the satellite prophage genome harbouring it (data presented in Figure 4.11B and 4.12) were part of a collaborative study with Prof Jeremy Brown and Dr Elisa Ramos-Sevillano at University College London. The experimental work was performed by Dr Elisa Ramos-Sevillano and Dr Asma Akter (a post-doc in Prof Brueggemann's research group). Dr Ramos-Sevillano and Dr Akter created the genetic mutants and Dr Elisa Ramos-Sevillano performed the animal experiments.

4.1 Abstract

Prophages are phage genomes that are integrated into bacterial chromosomes and are of interest because they can be reservoirs of genes that are involved in bacterial pathogenesis. Satellite prophages are a recently discovered type of prophage that do not have the ability to replicate on their own and have a life cycle that is dependent on the bacterial host and an additional helper phage. Satellite prophages have a demonstrated role in virulence in several bacterial species, but the prevalence, diversity and genetic stability of satellite prophages among pneumococci are currently unknown, and whether or not pneumococcal satellite prophages encode any virulence factors has not yet been investigated. Here, I investigated 403 'partial prophage' sequences identified as part of a recent study by the Brueggemann group with the aim of identifying whether any satellite prophages exist among these sequences. This work led to the identification of 44 unique and novel satellite prophages, and the molecular epidemiology of these elements was conducted in the context of the pneumococcal population structure. A putative virulence gene, *vapE*, was revealed in some of the satellite prophages. Collaborative experimental work demonstrated that one of these satellite prophages was associated with virulence in a murine model of infection. Furthermore, RNA sequencing data were used to investigate the expression of satellite prophage genes, which revealed that they were overexpressed when pneumococci were grown planktonically in broth versus in *in vitro* biofilm experimental conditions. Overall, the findings from this study demonstrated that satellite prophages are widespread among the pneumococci and suggested that they play a role in pneumococcal pathogenesis.

4.2 Introduction and aims

In a recent study by the Brueggemann group, whole genome sequences of nearly 500 pneumococcal isolates were used to assess prophage prevalence, diversity and distribution among pneumococci, which revealed that every pneumococcal genome contained prophage DNA [212]. Furthermore, 66 representative full-length phage genomes and 403 partial phage sequences were identified. While these findings reveal that prophages are widespread among pneumococci, to what extent pneumococcal prophages play a role in virulence and/or pathogenesis is not well defined.

Previous investigators demonstrated that the presence of the MM-1 prophage in a pneumococcal genome enhanced adherence to pharyngeal cells in an *in vitro* study, and this was thought to confer an advantage during colonisation [291]. Another study identified two prophage-encoded surface proteins (PblA and PblB) in some strains of *S. mitis*, which mediated binding to human platelets and was associated with an increased risk of endocarditis [214]. More recently, prophages expressing PblB were detected within pneumococcal genomes and expression of PblB enhanced adherence to lung epithelial cells [292]. The presence of PblB was associated with promoting persistence in the nasopharynx and lungs in a murine model [292]. Our recent prophage study revealed that the majority (72%) of the prophages we characterised possessed *pblA* and/or *pblB* [212], thus to what extent these specific genes are associated with increased adherence or whether there are other additional contributory prophage genes is yet unknown. More generally, given the diversity of prophages and the fact that the function of the vast majority of prophage genes remains unknown, it is conceivable that there are additional prophage-encoded virulence genes among streptococci that are still awaiting discovery.

Previously in the Brueggemann *et al* paper, we reported on the gene expression levels after mitomycin C induction of a partial prophage sequence (IPPX439) and two full-length prophage sequences (SP195_1 and SP195_2) in pneumococcal reference strain PMEN3 [212]. We revealed that the genes of the two full-length phage genomes were significantly upregulated an hour after the addition of mitomycin C and furthermore, that several genes in the partial prophage were also differentially expressed, which occurred nearly an hour after the induction of the two full-length prophages (Figure 4.1). The finding that the genes associated with the partial phage sequence were transcriptionally active warranted further investigation. The function and role of the partial phage sequences in the pneumococci genomes are currently unknown, although one possibility is that some of these sequences are satellite prophages.

As discussed in the general introduction (see 1.4.5), satellite prophages are a recently discovered type of prophages, which do not have the ability to replicate on their own and have a life cycle that is dependent on a helper phage. Satellite prophages are able to exploit phages integrated within bacterial chromosome as helpers by manipulating the phage life cycle to enable their own replication and promiscuous spread throughout genomes. Satellite prophages have been shown to be vectors for the spreading of toxin genes and other virulence factors, *e.g.* SaPI1, which possesses the gene responsible for causing toxic shock syndrome [224].

To date, a small number of satellite prophages have been discovered in streptococcal species [231]; however, despite their demonstrated role in virulence in other bacterial species, the prevalence, diversity and genetic stability of satellite prophages among pneumococci had not yet been evaluated. This is in part because, due to their smaller

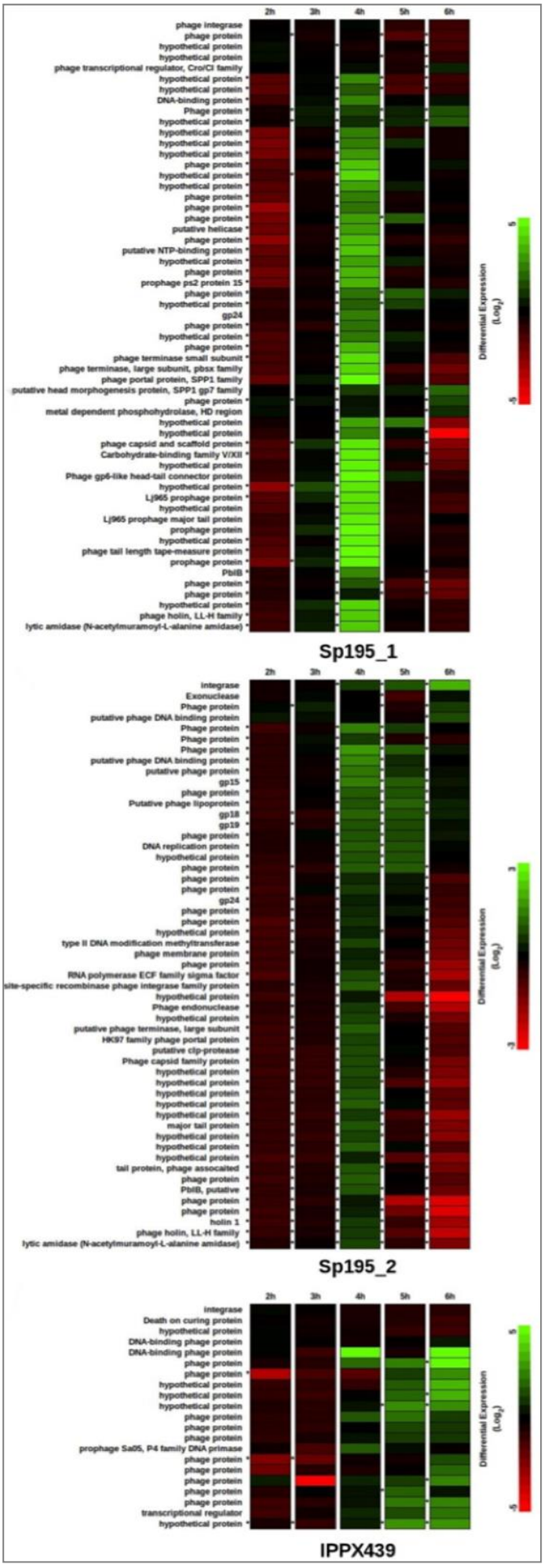


Figure 4.1 – RNA sequencing analyses indicate evidence for phage gene expression. The heat maps display differential gene expression following mitomycin C induction of two full-length phages (Sp195_1 and Sp195_2) and a partial phage (IPPX439) integrated within the PMEN3 genome. Mitomycin C was added to bacterial cultures after three hours of incubation. Phage genes are represented by rows and differential expression levels at each time point are displayed in columns. Statistically-significant ($p<0.05$) expression levels are indicated by an asterisk to the left of a cell (see 4.3.6.1 for further details). Figure is reproduced from Brueggemann *et al* [212].

genome and apparent lack of essential genes, streptococcal satellite prophage sequences have historically been disregarded as “remnant” or “defective” prophages in a state of mutational decay [218, 227, 293, 294]. They are also often very difficult to distinguish from other integrative elements within genomes, and therefore, frequently overlooked in many studies.

The best characterised streptococcal satellite prophage is referred to as *S. pyogenes* phage-like chromosomal island M1 (SpyCIM1), which is maintained as an episome in the bacterial cytosol during exponential growth. However, during stationary phase, SpyCIM renders the mismatch repair (MMR) system non-functional via integration in-between *mutS* and *mutL* genes. The MRR system is responsible for the correction of randomly occurring mutations in the bacterial genome, and thus, its disruption leads to higher mutation rates, which can be beneficial in certain situations, as it allows the bacteria to evolve and thus adapt more quickly to external stress. Upon returning to exponential growth, SpyCIM1 momentarily leave the genome through precise site-specific excision, leading to reactivation of the MMR system via genomic rearrangement. In this way, SpyCIM1 can serve as a genetic switch for modulating the rate of host spontaneous mutation according to cell growth stage [218, 227]. This finding emphasises the importance of investigating prophage insertion sites, as it is possible that other similar dynamic process of regulating a bacterial system via prophage excision and reintegration are in place and are yet to be discovered.

The main aims of this chapter were to: a) investigate the 403 ‘partial prophage’ sequences identified as part of the recent study with the aim of identifying whether any satellite prophages exist among these sequences; b) analyse the prevalence, diversity and genetic stability of any identified satellite prophages in the context of the

pneumococcal population structure; c) characterise satellite prophage insertion sites among pneumococcal genomes; d) screen satellite prophage genomes for the presence of virulence genes; and e) analyse existing RNA sequencing datasets to investigate prophage gene expression under biofilm versus planktonic conditions.

4.3 Methods

4.3.1 Genome dataset

The genome dataset was comprised of 482 pneumococcal genomes [212] (Supplementary File 4.1). All genomes were stored in BIGSdb database [239] and annotated using the RAST server [243] (http://rast.nmpdr.org). The prophage content for all genomes in the study dataset were calculated using the PhageContentCalculator script (see Chapter 2)

4.3.2 Identification of satellite prophages among ‘partial prophage’ sequences

Using the genome dataset described above (see 4.3.1), we had previously determined the prevalence, diversity and molecular epidemiology of full-length prophages [212]. Full-length prophage DNA sequences were defined as those that started with an integrase gene, ended with an amidase or lysin gene and were >28 kb in length. Many shorter prophage sequences (n= 403) were also identified in that study, which were simply classified as ‘partial prophage’ sequences and not characterised further at the time. These consisted of any sequence regions of between 2–25 Kb in length within pneumococcal genomes that contained one or more coding region annotated as 'phage'.

To facilitate computational analysis for identifying novel satellite prophages among ‘partial prophage’ sequences, I used CD-HIT [295] to make non-redundant sequence sets with a threshold of ≥90% nucleotide sequence identity. Remaining sequences

were manually investigated for evidence of satellite prophage using Geneious version 11.1 (Biomatters Ltd.) and BLAST [199]. Satellite prophage sequences were defined as those lacking extensive nucleotide sequence homology with full-length prophages, as well as having a genomic organisation that is similar to previously reported satellite prophages among other bacterial species (Figure 4.2) [224]; sequences that started with an integrase gene, adjacent to two divergently oriented genes encoding lysogeny regulator proteins, followed by a replication module, containing genes that encode a primase and/or a replication initiator.

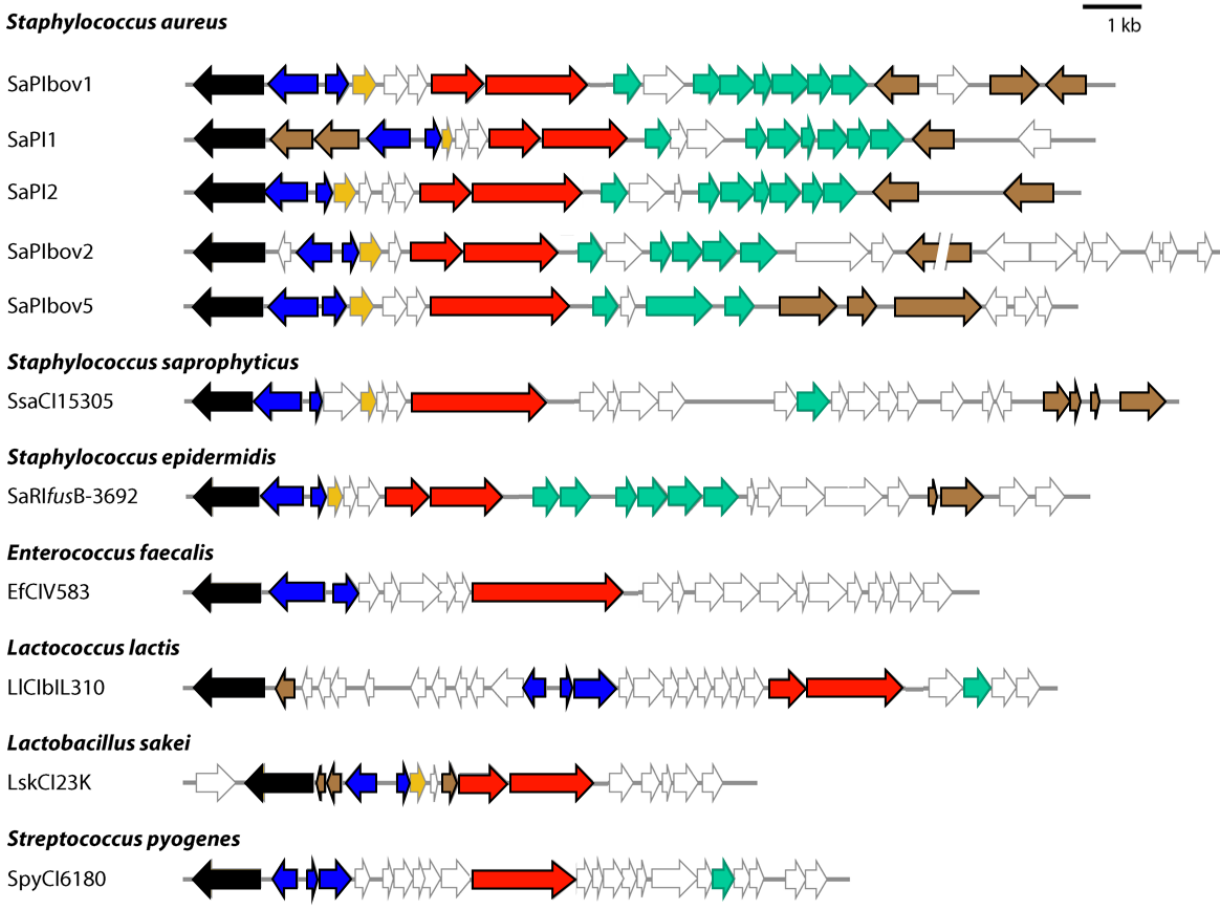


Figure 4.2 – Genomic organisation of a subset of satellite prophages previously reported among various bacterial species. Genomes are aligned according to prophage convention, with the integrase gene at the left end. Genes are coloured according to predicted function: integrase genes are black, genes encoding lysogeny regulators are blue; genes involved in prophage replication are red, genes involved in phage interference are green; virulence genes are brown; excisionase genes are orange; genes encoding hypothetical proteins are white. Figure was modified from Penades *et al* [224].

For further validation, the DNA sequences of the identified pneumococcal satellite prophage genomes were used as queries to BLAST against all host genomes in the study dataset. The matching region plus additional flanking regions were consecutively inspected manually using the Geneious program in order to confirm the presence or absence of each unique pneumococcal satellite prophage in each pneumococcal genome. This process necessitated substantial manual effort and inspection of sequences, but the return was a lengthy list of previously unidentified complete satellite prophage genomes for investigation.

4.3.3 Sequence analyses of prophages

The total number of ORFs and the average GC content of each prophage were calculated within the Geneious environment. All multiple sequence alignments were performed using ClustalW [268] with default parameters (Gap open cost = 15, Gap extend cost = 6.66). Phylogenetic trees were constructed based upon sequence alignments using FastTreeMP [269]. Putative prophage genes were scrutinised using homology searches to other known prophage genes, combined with structure-based searches. The structural and functional domains in protein sequences were predicted using the NCBI Conserved Domain Database [246]. Genes of interest were submitted to the STRING database [247] to perform searches for any previously reported relationship to other genes. Schematic diagrams of the coding regions of the prophages were produced in Geneious and edited using Adobe Illustrator (Adobe Inc.).

4.3.4 Investigation of prophage insertion sites

The nucleotide sequences of all identified pneumococcal satellite prophage genomes were used as queries to BLAST against all host genomes in the study dataset using

the custom BLAST feature in Geneious. The matching region plus flanking regions were inspected manually using the query-centric view function of the Geneious software in order to identify insertion sites among pneumococcal genomes. The satellite prophage integrases were divided into different categories using the CD-HIT program at a threshold of ≥95% sequence identity. The genome diagram representing the satellite prophage insertion sites among pneumococcal genomes was created using Geneious and edited using Adobe Illustrator.

4.3.5 Construction of a pneumococcal core genome phylogenetic tree

The 482 genomes in the dataset were annotated using the Prokka v1.10 program [244] in order to generate GFF files compatible with tools used in the downstream analyses. The GFF files were then input into Roary [249] and clustered using a threshold of 90% amino acid sequence similarity (parameter: -i 90). Genes present in all genomes were selected using a core genome threshold of 100% (parameter: -cd 100) and were aligned using Roary. FastTreeMP [269] was used to create the phylogenetic tree, followed by ClonalFrameML [270] to reconstruct the tree adjusted for recombination. The tree was visualised using iTOL [248] and further edited using Adobe Illustrator.

4.3.6 RNA sequencing analyses

I analysed two RNA sequencing datasets as part of my work on satellite prophages: one was generated by our group (prophage induction by mitomycin C) and were not yet analysed, the other one (biofilm versus planktonic mode of growth) was generated by Blanchette *et al* [240].

4.3.6.1 Dataset 1: prophage induction with mitomycin C

Total bacterial RNA sequencing was performed on RNA extracted from a pneumococcus that was subjected to mitomycin C treatment to facilitate phage induction (see Chapter 2). The sequenced forward and reverse reads were paired and mapped onto the annotated PMEN3 reference genome using Bowtie2 [33] with the highest sensitivity option. Differential gene expression between the control and experimental treatment were assessed in Geneious using the DESeq [275] method. Genes with an adjusted P value less than 0.05 were deemed to be differentially expressed. The full list of the genes and their sequences and expression levels can be found in the Supplementary File 4.2. Raw RNA sequencing data from this study have been deposited in the publicly-available GEO repository (http://www.ncbi.nlm.nih.gov/geo) and can be accessed using the accession number GSE89200.

4.3.6.2 Dataset 2: biofilm versus planktonic mode of growth

A whole-genome RNA sequencing dataset was published by Blanchette *et al* in 2016 [240]. In brief, samples were collected in three biological replicates from a pneumococcal strain Sp6A-10 isolate (serotype 6A; ST460) growing in Todd-Hewitt broth either planktonically or in polystyrene six-well plates as two-day-old biofilms. Total RNA from each sample was extracted and sequenced using the Illumina HiSeq4000 sequencing platform. For use in the current study, raw RNA sequencing data was retrieved from the GEO repository (accession number GSE85196). Reads from the control planktonic (THB_PK1, THB_PK2, THB_PK3) and biofilm (THB_BF1, THB_BF2, THB_BF3) samples were paired and mapped onto the pneumococcal Sp6A-10 genome using Bowtie2 with the highest sensitivity option. Differential gene expression levels were computed in Geneious using the DESeq2 [296] method.

Genes with an adjusted p value <0.001 were deemed to be differentially expressed. A volcano plot was generated within the Geneious environment and further edited using Adobe Illustrator. The full list of the genes and their sequences and expression levels can be found in the Supplementary File 4.3.

4.3.7 Assessment of virulence in murine pneumococcal infection model

4.3.7.1 Bacterial strains, media and growth conditions

Strains, plasmids and primers used for this study are listed in Supplementary File 4.4. *S. pneumoniae* strains were cultured in Columbia agar supplemented with 5% horse blood, or in Todd-Hewitt broth supplemented with 0.5% yeast-extract (THY) at 37°C in 5% CO_2. Mutant strains were selected by using antibiotics (150 µg/ml spectinomycin). Broth culture growth was monitored by measuring optical density at 580 nm and stocks of pneumococcus were stored as single use aliquots at −70°C. Data for growth curve measurements were collected using 96-well plates in a Tecan Spark microtiter plate reader essentially as described before [297], measuring the optical density at 595 nm (OD_{595}) in 30 minutes intervals. For growth in THY and serum, 10^6 CFU (colony-forming unit) of each strain was added to 200 µl of medium or serum and incubated at 37°C plus 5% CO_2.

4.3.7.2 Construction of pneumococcus mutant strains lacking SpnSP38 *and vapE*

Pneumococcus mutant strains lacking SpnSP38 *(*Δ*SpnSP38*) *and vapE* (Δ*vapE*) were created using overlap extension polymerase chain reaction (OE-PCR) [298] in a serotype 6B pneumococcal strain (BHN418) using a transformation fragment in which *vapE* or the entire satellite prophage were replaced by the spectinomycin resistance cassette *aadA9*. To construct the Δ*SpnSP38* mutant, two products

corresponding to 762 base pairs upstream (primers SpnSP_UpF and SpnSP_UpspecR) and 872 base pairs downstream (primers SpnSP_Downspec_F and SpnSP_DownR) of the satellite prophage were amplified from BHN418 genomic DNA by PCR carrying 3' and 5' linkers complementary to the 5' and 3' portion of the *aacA9* gene respectively. Using primers SpnSP_Upspec_F and SpnSP_Downspec_R, *aadA9* was amplified by PCR from pR412 plasmid [299]. Likewise, to create the Δ*vapE* mutant, a construct was generated wherein 820 base pairs of flanking DNA upstream (primers VapE_UpF and VapE_UpspcR) and 526 base pairs of flanking DNA downstream (primers VapE_DownspecF and VapE_DownR) from the *vapE* gene were amplified by PCR and fused with the *aadA9* cassette by OE-PCR [298]. The resulting constructs were then transformed into the BHN418 strain using standard protocols [299, 300].

4.3.7.3 Experimental models of infection

Studies investigating pneumococcal sepsis or pneumonia were conducted using six-week-old mice and infected as previously described elsewhere [301]. In brief, in the sepsis model mice were challenged with 5×10^6 CFU/ml of the wild-type BHN418 strain or the constructed mutants (Δ*SpnSP38 and* Δ*vapE*) in a volume of 150 µl by the intraperitoneal route, whereas for pneumonia, mice were inoculated intranasally with 50 µl containing 10^7 CFU/mouse of the wild-type strain or the mutants. Animals were killed at 24 or 28 hours after challenge and bacterial counts were made from samples recovered from lung and blood. Lungs and spleens were homogenised through a 0.2 µm filter. Results were presented as $\log_{10}$ CFU/ml of bacteria recovered from the different sites.

For mixed infection experiments, mice were inoculated with a 50:50 mixture of wild-type and mutant strains. The competitive index (CI) was defined as the ratio of the

test strain (mutant strain) compared to the control strain (wild-type strain) recovered from mice divided by the ratio of the test strain to the control strain in the inoculum [302, 303]. A CI of <1 denotes that the test strain is attenuated in virulence compared to the control strain and the lower the CI the more attenuated is the mutant strain. Statistical analyses were conducted using analysis of variance (ANOVA) for multiple comparisons in GraphPad Prism 7.0 (GraphPad Software, San Diego, CA).

4.3.7.4 C3b binding to pneumococci

Serum samples from five healthy male volunteer controls (median age 40 years) were obtained according to institutional guidelines and stored as single-use aliquots at -70°C to use as a source of complement. C3b deposition was analysed using a flow cytometry assay [304]. Briefly, C3b deposition was investigated by incubating 10^7 CFU of pneumococci with 10 µl of pooled human serum (diluted to 20% in phosphate-buffered saline; PBS) for 30 minutes at 37°C. C3b bound to the different strains was labelled with 50 µl of a 1/500 dilution of fluorescein isothiocyanate-conjugated polyclonal goat anti-human C3b antibody after two washes in PBS-Tween 20 (0.01%). The detection of C3b binding was performed using flow cytometry with gating based on the analysis of at least 10,000 bacteria. Experiments were repeated three times and the results were expressed as the proportion of C3b deposition on the surface of the different mutants compared to the C3b deposition on the 6B wild-type strain.

4.3.7.5 Neutrophil killing assay

Frozen aliquots of pneumococci were thawed and washed twice with PBS-Tween 20 (0.01%) by centrifugation for 5 minutes at 13,000 revolutions per minute (rpm). 100 µl of the bacterial suspension, diluted to 10^3 CFU, was added to each well in the presence of 25% baby rabbit complement. After 30 minutes of incubation at 37˚C, 100 µl of

neutrophils (10^5 cells) previously isolated from human blood using MACSxpress® was added to each well and incubated at 37˚C with shaking. Sample aliquots were taken at 15 and 30 minutes, spotted onto Columbia blood agar plates and incubated at 37°C plus 5% CO_2. Bacterial colony counts were performed after overnight incubation.

4.4 Results

4.4.1 Identification of 44 unique pneumococcal satellite prophages

Extensive *in silico* analyses of a large and diverse dataset of 482 historical and modern pneumococcal genomes (described in 4.3.2) led to the identification of 44 unique and newly-discovered putative satellite prophage genomes, which clustered into five major groups (Figure 4.3A and 4.4). The average GC content of the satellite prophages was lower than that of the pneumococcal host but varied among each group (Figure 4.5).

Similar to previously reported satellite prophages among other bacterial species [226], the genomic organisation of the pneumococcal satellite prophages indicated conserved modular structures, with genes clustered according to function (Figure 4.3). Putative satellite prophage genomes started with an integrase gene, adjacent to two divergently oriented lysogeny-related genes (often with several intervening accessory genes), followed by a replication module consisting of a primase and/or a replication initiator gene(s). Accessory genes (*i.e.* genes that do not appear to be involved in the prophage lifecycle), such as those encoding toxin-antitoxin systems, were situated between the integrase and the lysogeny-related genes. In addition to this well-conserved gene organisation, the pneumococcal satellite prophages were distinguished from full-length sequences by their smaller genome size (Figure 4.6), the lack of phage structural and lysis genes (Supplementary File 4.5) and the fact that their insertion sites were never occupied by full-length prophages.

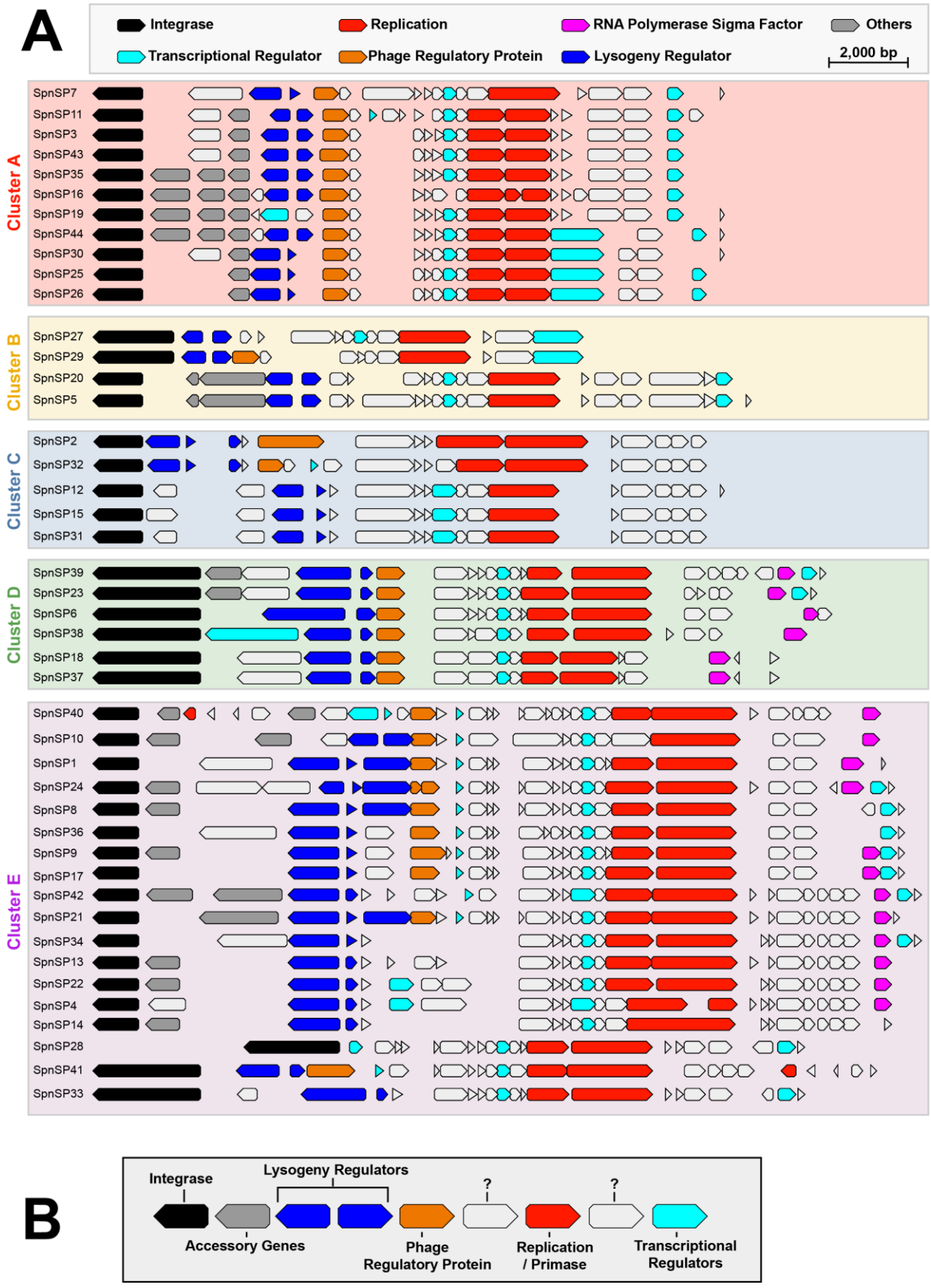


Figure 4.3 – Pneumococcal satellite prophages demonstrate well-conserved patterns in genome organisation and synteny. A, A schematic representation of the novel pneumococcal satellite prophage genomes. **B,** A schematic representation of the conserved modular structures observed among different satellite prophage genomes. Genes are coloured based on putative function.

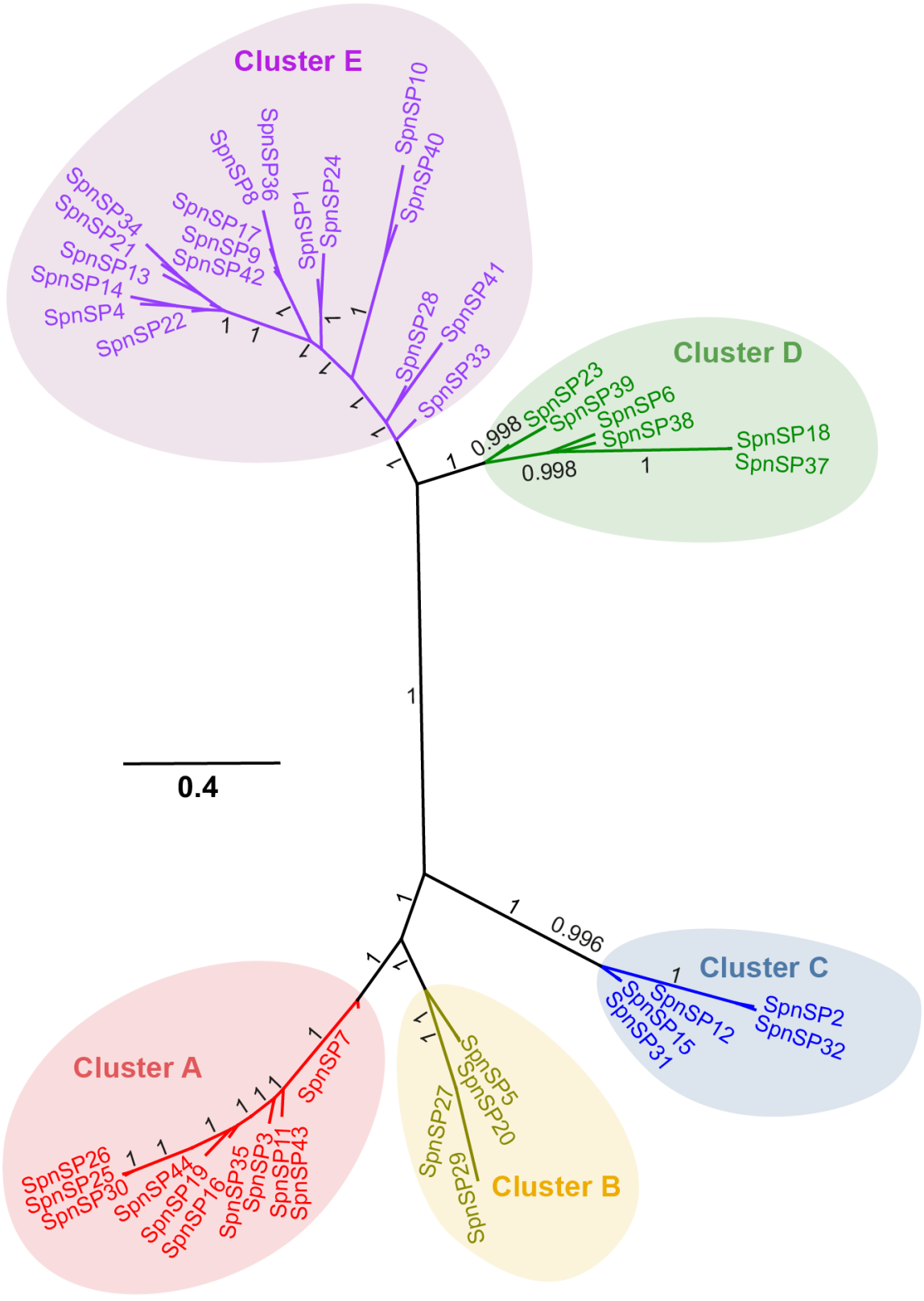


Figure 4.4 – Pneumococcal satellite prophages can be clustered into five major groups. An unrooted phylogenetic tree demonstrating that the 44 representative satellite prophages can be clustered into five major groups based upon nucleotide similarity.

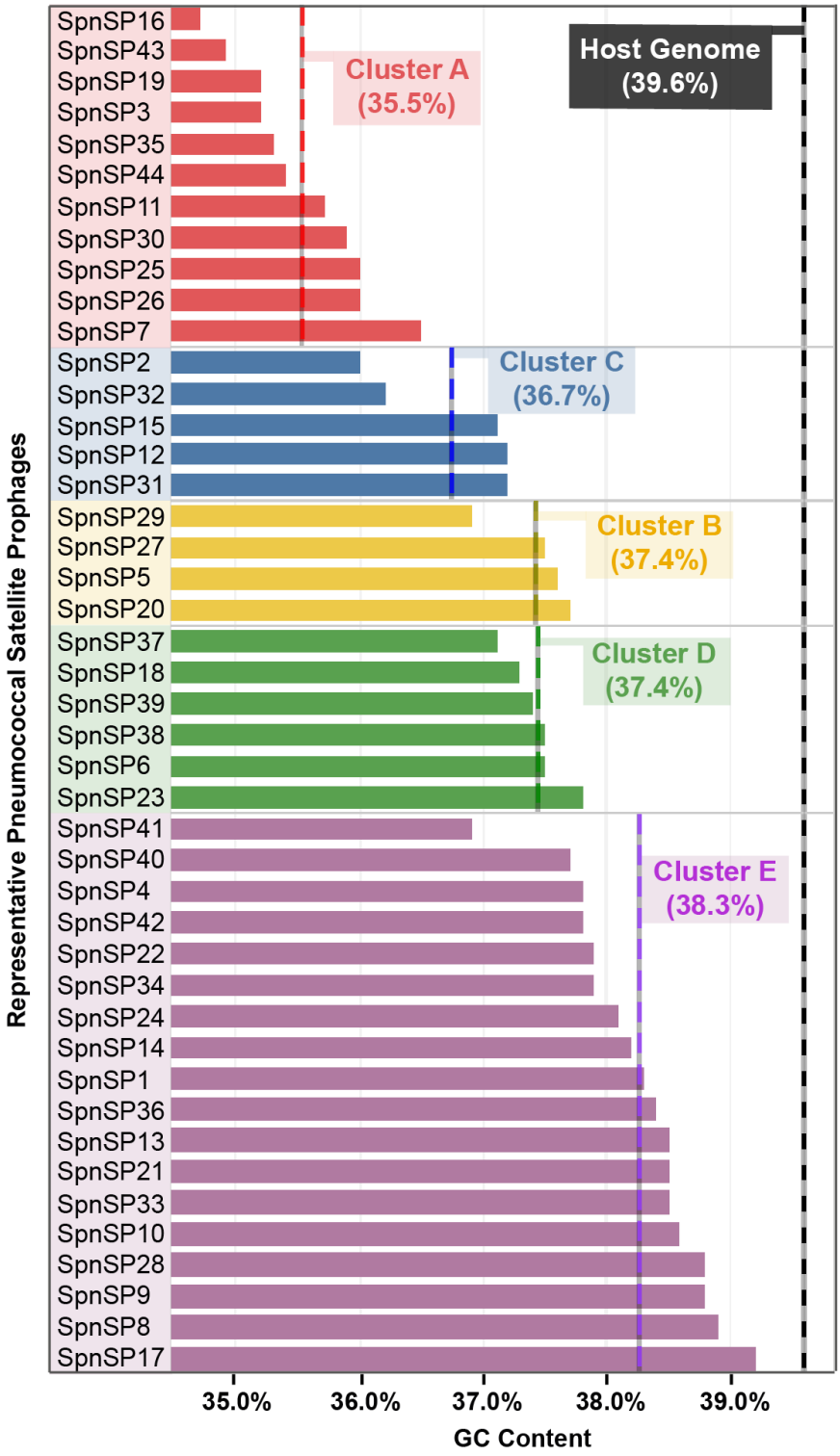


Figure 4.5 – The average GC content of the representative pneumococcal satellite prophages.

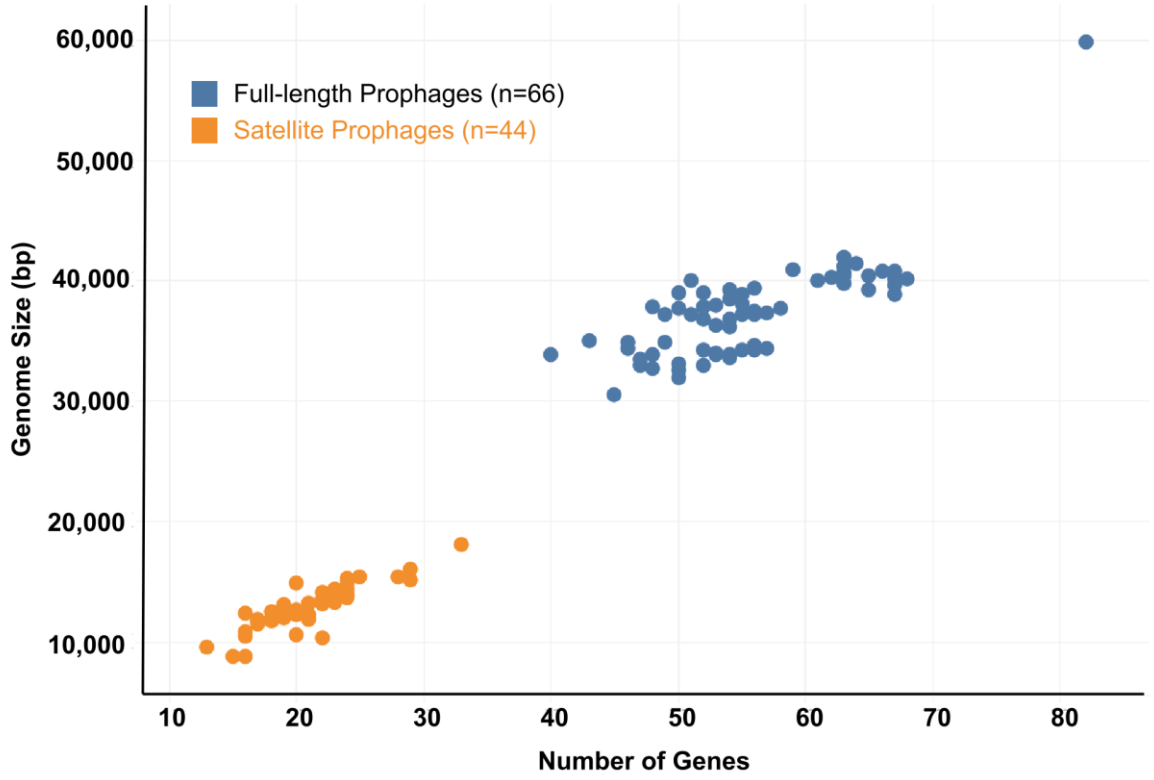


Figure 4.6 – Graphical representation of pneumococcal satellite and full-length prophages by genome size and number of genes.

Satellite prophages were consistently inserted in seven specific locations (a-f) of the host genome, each of which was directly associated with the nucleotide sequence of the integrase gene they harboured (Figure 4.7 and 4.8). The 44 representative satellite prophage integrases were divided into seven different categories with ≥95% nucleotide sequence similarity within each category. Each integrase category was associated with insertion at a single location on the pneumococcal genome, aside from integrase category I, which was associated with five different locations (Figure 4.7). 28.3% of pneumococcal satellite prophages were integrated at site a, which was situated very close to the origin of replication (*oriC*) (Figure 4.8), warranting further studies to explore the intriguing possibility that factors other than the integrase sequence may determine the prophage insertion

site (will be explored in more details in Chapter 5). The majority (88%) of the identified satellite prophages were found to be situated in three particular insertion sites (a-c), suggesting that the presence of suitable region for integration in the host may act as a major determining factor in phage sensitivity.

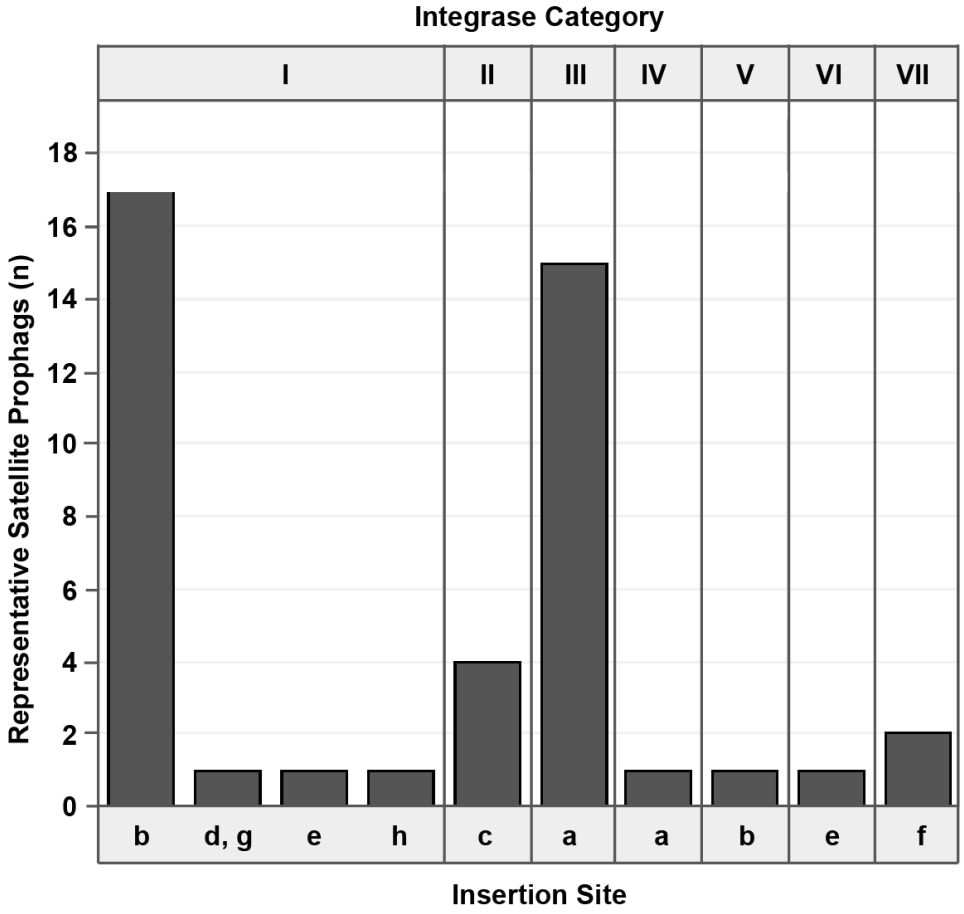


Figure 4.7 – The associations between representative satellite prophages, their integrase gene and insertion sites. Satellite prophages were found inserted in seven specific locations (a-f) of the pneumococcal chromosome. The integrase sequences of the 44 representative satellite prophages were divided into seven different categories based upon ≥95% nucleotide similarity.

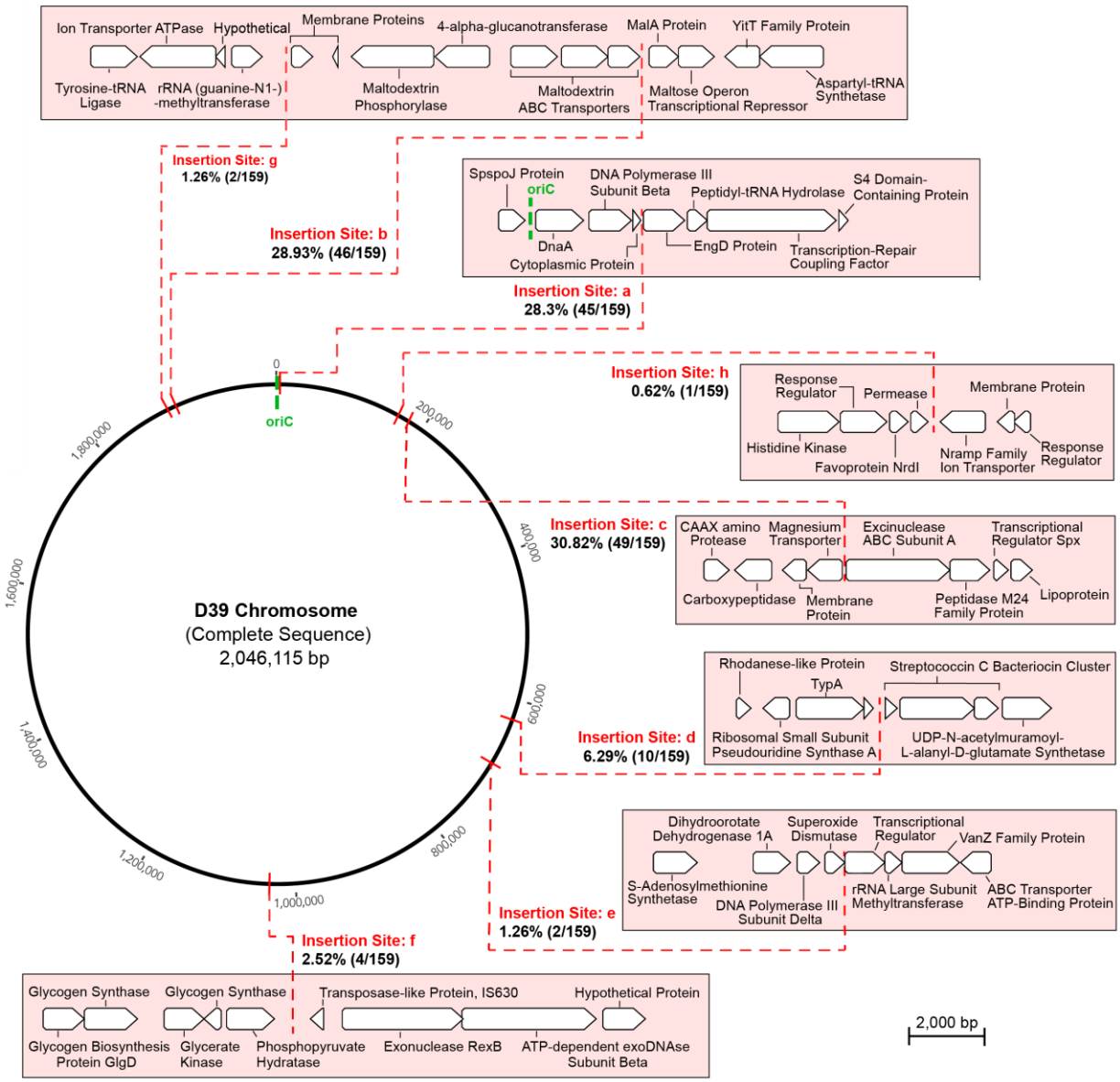


Figure 4.8 - Insertion sites of satellite prophages within the pneumococcal genome. Pneumococcal satellite prophages were integrated in seven locations (a-f) within the host genome. Percentages and numbers in brackets refer to the proportion and number out of all 159 satellite prophages that were inserted in that particular location. The origin of replication of the chromosome (*oriC*) is shown in green.

4.4.2 Molecular epidemiology of satellite prophages within a global and historical pneumococcal dataset

I further assessed the 44 representative satellite prophages in the context of the pneumococcal population structure. The study dataset was comprised of 482 pneumococci isolated between 1937 and 2008, recovered from both healthy and diseased individuals of all ages residing in 36 different countries. Ninety-one serotypes and 94 different clonal complexes were represented in the dataset (Supplementary File 4.1).

Satellite prophages were found to be widespread among the pneumococci in our dataset: 35% of the genomes contained at least one satellite prophage and 5% of the genomes contained two. Some satellite prophages were found to be present in up to six different clonal complexes, whereas others were only present in Singletons (genotypes with no closely related variants; Table 4.1 and Figure 4.9). The majority (68.2%) of satellite prophages were found within only one clonal complex or singletons (Table 4.1). Those satellite prophages identified in more than one genome were often found among pneumococci recovered over a decade or more (Table 4.1). The average prophage content for each of the major clonal complexes ranged from 2.2-6.5%, and with only one exception (CC7232), all of these are widely-circulating pneumococcal clonal complexes (Figure 4.10; https://pubmlst.org/spneumoniae).

Table 4.1 - Epidemiological characteristics of 44 representative satellite prophages identified among a collection of pneumococcal isolates dating from 1939 onwards.

Satellite Prophage		Pneumococci						
Name	Cluster	Clonal Complex (n)	Genomes (n)	Isolation dates	Countries (n)	Serotypes (n)	Insertion Site	Integrase Category
SpnSP16	A	3	4	1939 - 1982	2	4	b	I
SpnSP3	A	2	3	1981 - 2004	2	2	b	I
SpnSP26	A	2	3	1985 - 2000	1	2	b	I
SpnSP35	A	2	2	1952 - 1952	1	2	b	I
SpnSP43	A	2	2	1939 - 2004	2	2	b	I
SpnSP30	A	1	5	1978 - 1978	1	1	b	I
SpnSP44	A	1	2	1939 - 1962	1	1	b	I
SpnSP7	A	1	1	1968	1	1	b	I
SpnSP25	A	1	1	1999	1	1	b	I
SpnSP19	A	Singleton[a]	2	1939 - 1952	2	2	b	V
SpnSP11	A	Singleton	1	1952	1	1	b	I
SpnSP5	B	5	15	1939 - 2007	3	7	d, g	I
SpnSP29	B	1	15	1978 - 1988	1	2	b	I
SpnSP27	B	1	1	2006	1	1	b	I
SpnSP20	B	Singleton	1	1954	1	1	b	I
SpnSP2	C	2	4	1984 - 2005	3	2	f	VII
SpnSP31	C	2	2	1983 - 2005	1	2	b	I
SpnSP12	C	1	1	1968	1	1	b	I
SpnSP15	C	1	1	1943	1	1	b	I
SpnSP32	C	1	1	1986	1	1	f	VII
SpnSP37	D	5	9	1939 - 1988	4	7	c	II
SpnSP38	D	4	30	1972 - 2006	6	5	c	II
SpnSP6	D	3	8	1939 - 1991	3	3	c	II
SpnSP23	D	2	11	1962 - 2008	3	4	a	III
SpnSP39	D	1	2	2005 - 2007	1	1	a	III
SpnSP18	D	Singleton	2	1939 - 1952	2	2	c	II
SpnSP24	E	6	23	1939 - 2006	6	4	a	III
SpnSP33	E	2	3	1952 - 1998	1	2	a	III
SpnSP1	E	1	5	1978 - 1988	1	1	b	I
SpnSP40	E	1	3	2001	2	2	a	III
SpnSP8	E	1	1	1988	1	1	a	III
SpnSP9	E	1	1	1957	1	1	a	III
SpnSP13	E	1	1	1943	1	1	a	III
SpnSP14	E	1	1	1995	1	1	a	III
SpnSP17	E	1	1	1972	1	1	a	IV
SpnSP22	E	1	1	1971	1	1	a	III
SpnSP28	E	1	1	2003	1	1	a	III
SpnSP34	E	1	1	1990	1	1	a	III
SpnSP36	E	1	1	1963	1	1	a	III
SpnSP42	E	1	1	1994	1	1	a	III
SpnSP4	E	Singleton	1	1982	1	1	e	I
SpnSP10	E	Singleton	1	N/A	1	1	h	I
SpnSP21	E	Singleton	1	1954	1	1	e	VI
SpnSP41	E	Singleton	1	1983	1	1	a	III

a, Singletons are genotypes with no closely related variants.

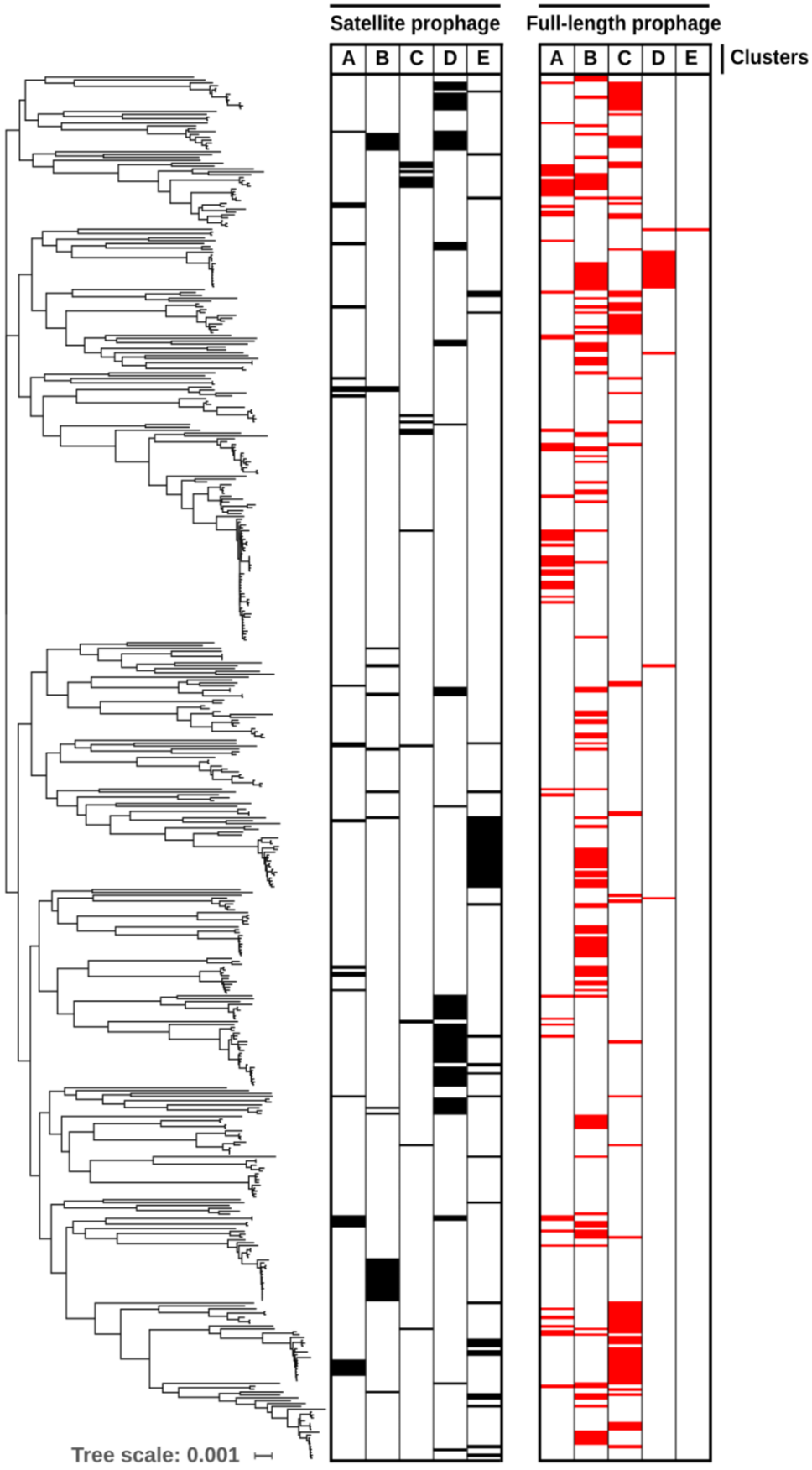


Figure 4.9 - A phylogenetic tree of the genomes in the study dataset labelled according to the presence of different prophages.

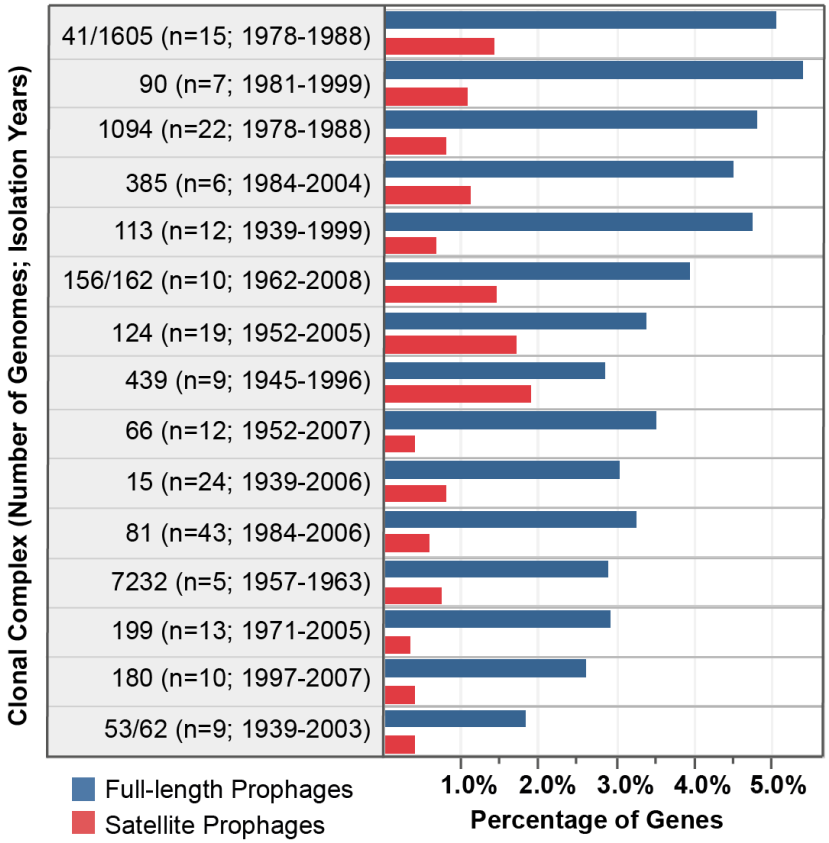


Figure 4.10 – The average prophage content of the major pneumococcal clonal complexes. The average prophage content for each of the major pneumococcal clonal complexes is depicted as a percentage of the total number of genes in the host pneumococcal genome (~2 Mb).

4.4.3 Satellite prophages and *vapE* are associated with virulence

The majority (57.5%) of satellite prophages genes in pneumococcal genomes in the study dataset were annotated as hypothetical, while the remaining were mostly predicted to be involved in replication and/or interfering with helper phage machinery (Table 4.2). A full list of the genes and their sequences is provided in the Supplementary File 4.5.

Table 4.2 – The predicted function of genes among the 44 pneumococcal satellite prophage genomes identified in the study.

Predicted function	Number of genes	Percentage
Unknown	532	57.5%
Transcriptional regulator	89	9.6%
Lysogeny regulator	88	9.5%
DNA replication	81	8.8%
Integrase	44	4.8%
Helper phage interference	38	4.1%
RNA polymerase sigma-70 domain containing protein	18	1.9%
Toxin-antitoxin system	17	1.8%
Lipoprotein	10	1.1%
DNA-damage inducible protein	8	0.9%
Grand total	925	100.0%

Investigation of the poorly characterised genes using homology searches and detection of conserved domains led to the identification of a gene that is a homologue of the 'virulence-associated gene E' (*vapE*) in *S. suis*. The gene encoding VapE in *S. suis* genomes has previously been described to have a role in *S. suis* virulence through an unknown mechanism [305]. 30/44 (68.2%) of the representative pneumococcal satellite prophages harboured *vapE*.

To examine whether the *vapE* homologue in the pneumococcal satellite prophage is also associated with virulence, a series of *in vivo* experiments were performed using a murine pneumococcal infection model and one example of a satellite prophage containing *vapE* identified in this study (Figure 4.11A). Deletion mutant strains were generated in a serotype 6B pneumococcal strain (BHN418) in which either *vapE (*Δ*vapE*) or the entire satellite prophage genome sequence (Δ*SpnSP38*) were replaced by a spectinomycin resistance cassette 205 (aadA9) in the BHN418 strain (Figure 4.11B). For each of the mutant strains, a competitive index (CI) was calculated using a competitive infection experiment in a mouse model of pneumonia. The CI was significantly <1 in the lungs after mixed infection with Δ*SpnSP38* and the wild-type serotype 6B or Δ*vapE* and the wild-type serotype 6B, indicating a role for the satellite prophage and *vapE* in the establishment of pneumococcal pneumonia (Figure 4.12A).

To further assess the degree of attenuation in virulence of the Δ*SpnSP38* and Δ*vapE* strains, infection experiments were repeated with pure inocula of each strain in both the pneumonia and sepsis models. There were no significant differences in bacterial CFU recovered from the lungs of infected mice at 24 hours between either mutant and the parental wild-type strain (data not shown) and the majority of the mice developed fatal infection by this point. However, in the sepsis model the mice infected with the wild-type serotype 6B strain had significantly greater blood and spleen CFU than the Δ*SpnSP38* mutant (Figure 4.12B and 4.12C), indicating that the satellite prophage is directly involved in pneumococcal virulence during bacterial dissemination in the systemic circulation. Although the Δ*vapE* strain had lower spleen CFU compared to the wild-type, this difference was not statistically significant, suggesting that loss of the whole satellite prophage has a more marked effect on the attenuation of virulence during sepsis than loss of VapE alone.

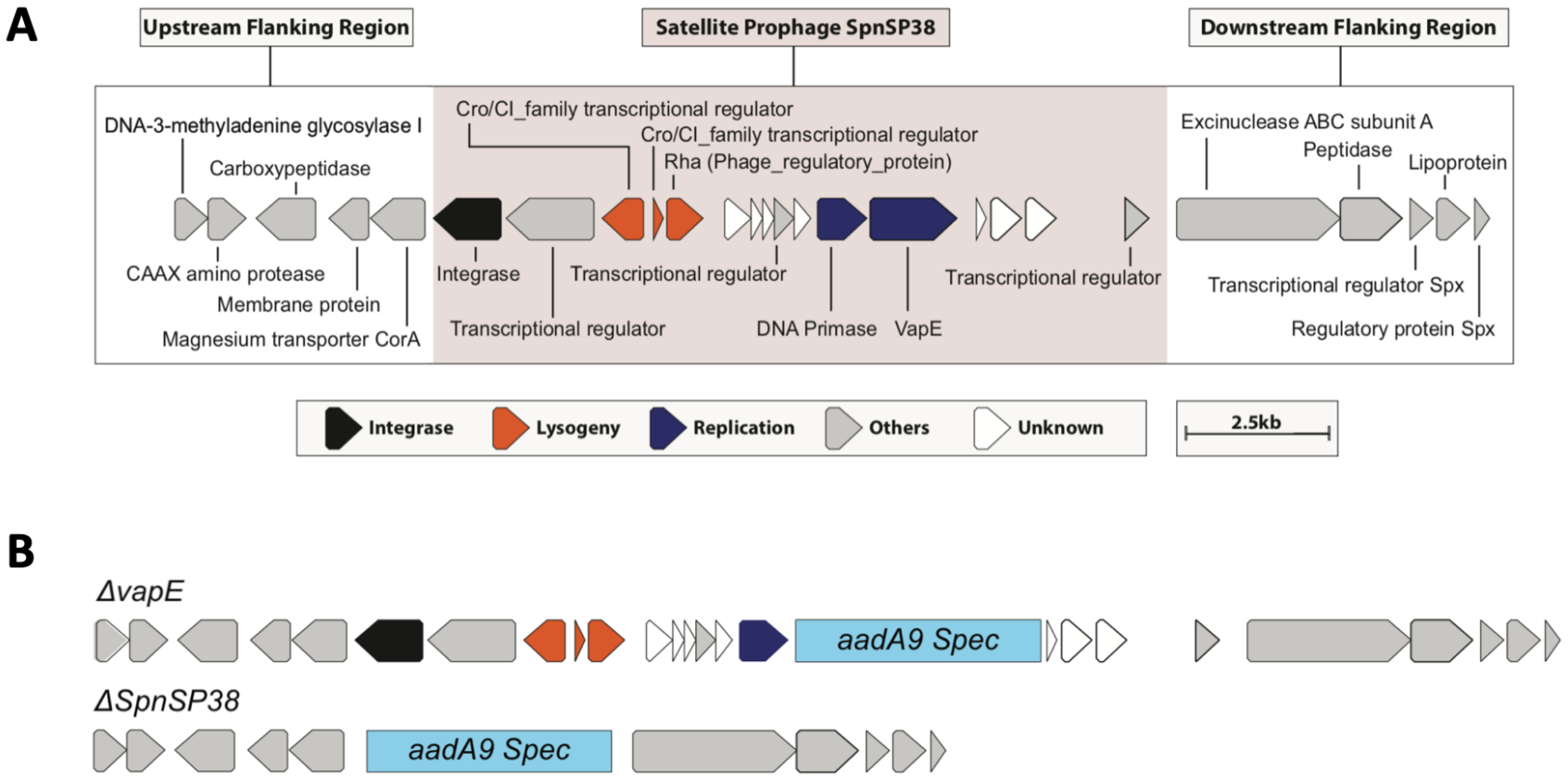


Figure 4.11 - Construction of *ΔvapE* and *ΔSpnSP38* pneumococcal mutant strains. A, A diagram depicting the satellite prophage genes integrated within the BHN418 genome and flanking pneumococcal genes. **B**, A diagram depicting Δ*SpnSP38* and Δ*vapE* mutants with the addition of the spectinomycin resistance cassette aadA9.

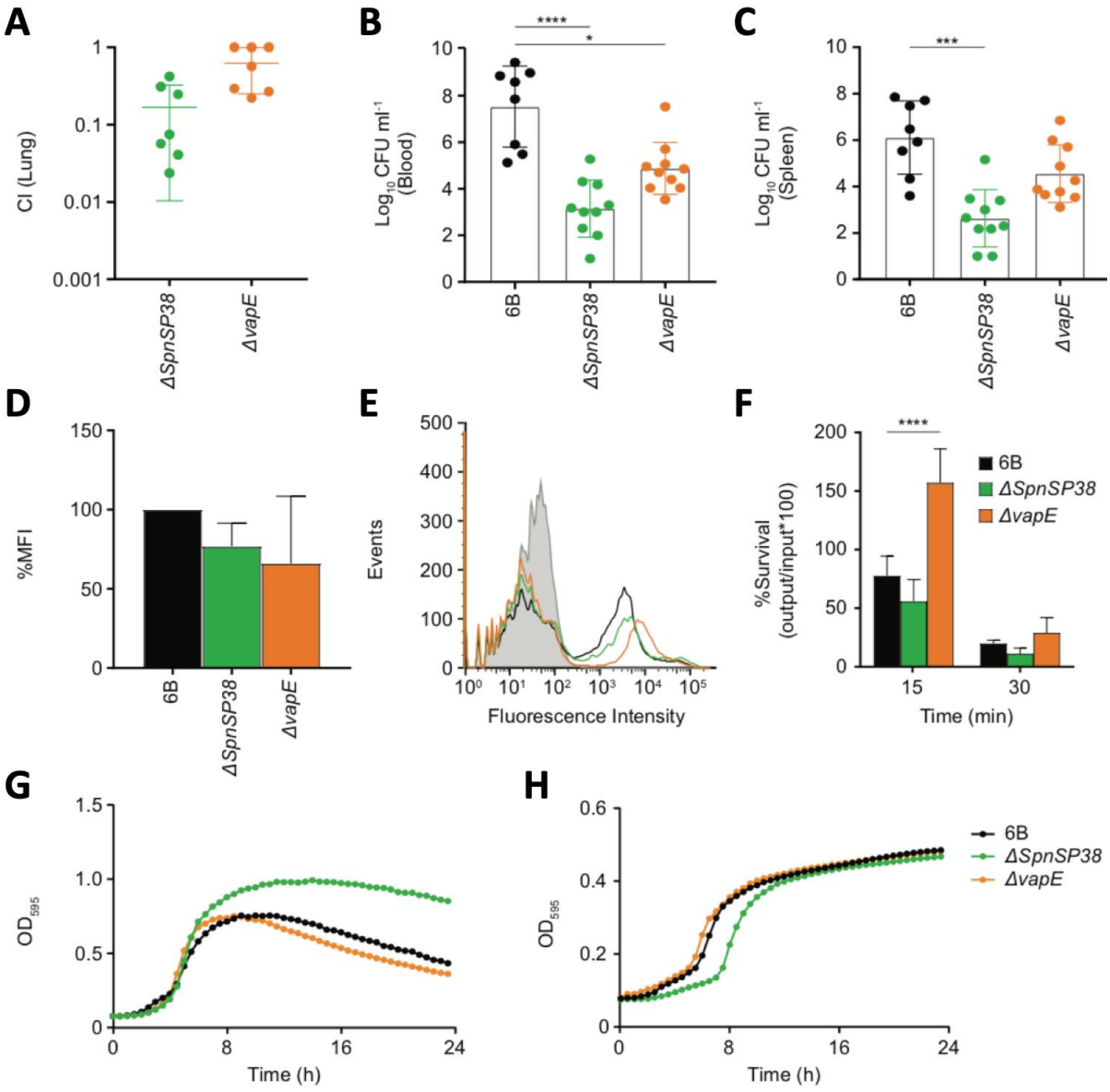


Figure 4.12 - Assessment of the virulence of *ΔSpnSP38* and *ΔvapE* mutant pneumococcal strains in murine infection. A, Plots of the competitive index (CI) for the *ΔSpnSP38* and *ΔvapE* mutant strains versus the wild-type strain in a mouse model of pneumonia. Each symbol represents the CI for a single animal and bars represent the median and range. **B** and **C**, Mean bacterial colony-forming units (CFU) recovered at 24 hour (h) from blood (**B**) or spleen (**C**) homogenates after intraperitoneal inoculation of 5×10^6 CFU/strain. Each symbol represents data for a single animal and error bars represent standard deviation (two-sided, Kruskall-Wallis with Dunn's post hoc test to identify significant differences between groups, *, $p<0.05$; ***, $p<0.001$; ****, $p<0.0001$). **D**, Median standard deviation (SD) mean fluorescence intensity of C3b deposition on the surface of the wild-type and mutant strains as measured by a flow cytometry assay. **E**, Example of a flow cytometry histogram for the C3b deposition data. **F**, Bacterial survival in a neutrophil killing assay (multiplicity of infection: 1 bacterium/100 neutrophils) represented as % CFU/ml recovered after 15 to 30 minutes incubation compared to the input bacteria. Error bars represent standard deviation and asterisks represent statistical significance compared to the wild- type strain (Kruskal-Wallis test with Dunn's correction for multiple comparisons, ****, $p<0.0001$). **G** and **H**, Growth curves as measured by the optical density (OD) of wild-type and mutant strains cultured in Todd- Hewitt broth supplemented with 0.5% yeast-extract (THY) (**G**) or 100% human serum (**H**).

4.4.4 The satellite prophage is required for optimum growth in sera but not for evasion of complement recognition or phagocytosis

Reduced systemic virulence of Δ*SpnSP38* or Δ*vapE* mutants could reflect poor growth under physiological conditions, or evasion of host innate immune killing, which is largely dependent on complement-mediated neutrophil killing. Using a flow cytometry assay, the binding of complement component C3b was not demonstrably different between the mutant strains and wild-type strain (Figure 4.12D and 4.12E). Furthermore, survival of the Δ*SpnSP38* and Δ*vapE* mutants in the presence of neutrophils after 30 minutes was similar to the wild-type BHN418 strain (Figure 4.12F). These data indicate that the satellite prophage and *vapE* are not required for evasion of complement or neutrophil killing, and that the reduced virulence of the Δ*SpnSP38* strain could reflect delayed growth in serum. Growth rates of both mutant strains in THY were not significantly different to the parental wild-type strain (Figure 4.12G); however, culture in serum demonstrated a small but significant delay in growth of the Δ*SpnSP38* strain compared to the wild-type and Δ*vapE* strains (Figure 4.12H).

4.4.5 Satellite prophage genes, including *vapE*, were overexpressed in planktonic versus biofilm samples

Given the association of the satellite prophage and *vapE* with virulence in the murine pneumococcal infection model (see 4.4.5), it was hypothesised that satellite prophage genes would be overexpressed when pneumococci were grown planktonically in broth versus in a biofilm. To evaluate this hypothesis, I performed a comparative transcriptome analyses of planktonic and biofilm pneumococci using an existing RNA sequencing dataset generated by Blanchette and colleagues [240]. In their study, pneumococcal reference strain Sp6A-10, which contained two full-length prophages

(Spn_6A-10_FP1 and Spn_6A-10_FP2) and one satellite prophage (SpnSP33, 58.7% identical to SpnSP38), was grown planktonically and as a two-day old biofilm. Three biological replicates were collected from each of the growth conditions and the corresponding RNA samples were extracted and sequenced.

The Blanchette transcriptomic data was analysed to assess prophage gene expression under these two experimental conditions, and the data demonstrated significantly higher satellite prophage and full-length prophage gene expression when the host pneumococcus was grown in broth as compared to growth in a biofilm pneumococcus (Figure 4.13, 4.14 and Supplementary File 4.3). The full complement of satellite prophage genes were significantly expressed, and many of the genes of the two full-length prophages, mainly structural and lysis genes, were also significantly upregulated. Notably, among the 20 most significantly upregulated genes, 60% (n=12) were satellite prophage genes and *vapE* was the third most upregulated gene in the entire genome. Among the 50 most highly expressed genes, just over half were prophage-related genes: 15 (30%) were satellite prophage genes; 7 (14%) were genes of one full-length prophage; and 4 (8%) were genes of the second full-length prophage (Figure 4.13 and Supplementary File 4.3).

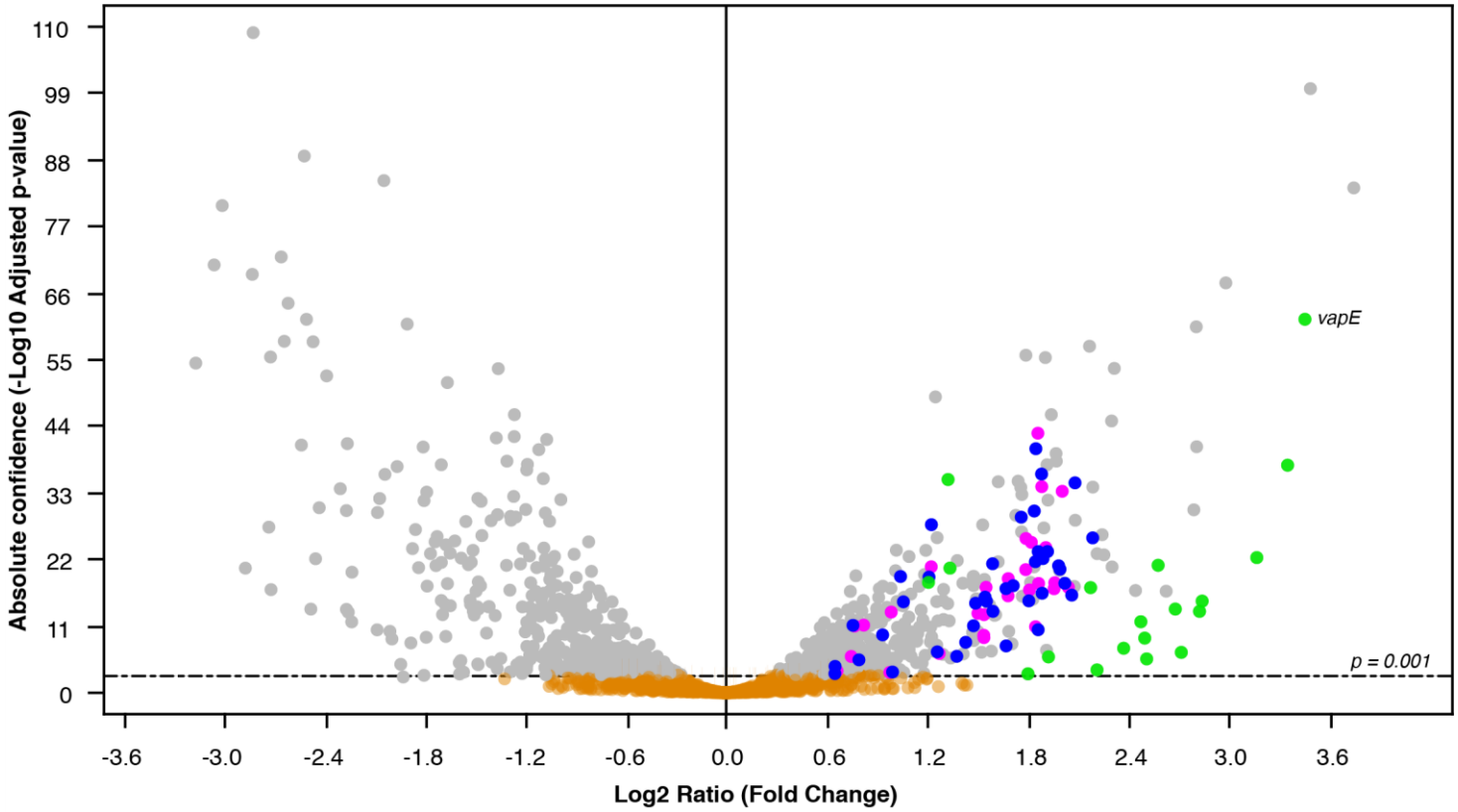


Figure 4.13 – Satellite prophage genes are overexpressed when pneumococci is grown planktonically compared to as a biofilm. The dataset was generated from a pneumococcal strain 6A-10 isolate, containing two full-length prophages (Spn_6A-10_FP1 and Spn_6A-10_FP2) and one satellite prophage (SpnSP38). Genes belonging to SpnSP38 are shown in green, while those belonging to Spn_6A-10_FP1 and Spn_6A-10_FP2 are shown in Blue and Magenta, respectively. Higher fold change ratio denotes increased expression levels in planktonic form compared to the biofilm. A full list of the genes depicted here and their sequences can be found in the Supplementary File 4.3.

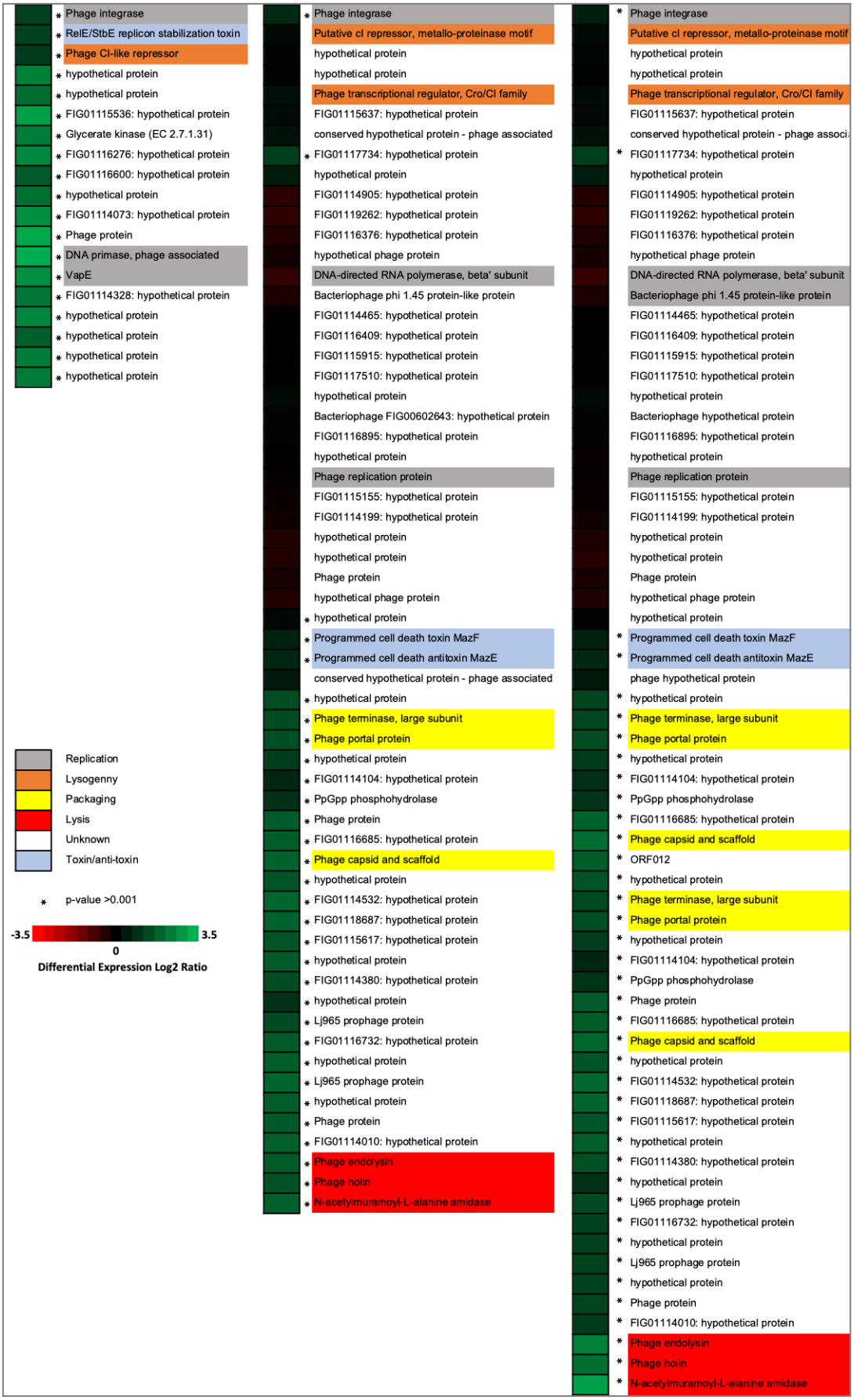


Figure 4.14 – RNA sequencing analyses indicate evidence for satellite prophage replication. Dataset was generated for pneumococcus growing planktonically versus as a biofilm. The heat maps display differential gene expression of a satellite prophage SpnSP38 (left) and two full-length prophages Spn_6A-10_FP1 (middle) and Spn_6A-10_FP2 (right) integrated within the 6A-10 genome.

4.5 Discussion

The findings from this chapter reveal that satellite prophages are widespread among the pneumococci. Forty-four novel and unique pneumococcal satellite prophages were discovered, and they demonstrated a structured population. Satellite prophages were found to have persistent associations with specific clonal complexes of the host bacteria over many decades. This is important, since these findings provide a framework by which to explore why particular combinations of prophages and bacteria exist and whether the presence of certain prophages among genomes might contribute to the virulence or the epidemiological success of a bacterial genetic lineage.

RNA sequencing analyses demonstrated that several of the newly discovered pneumococcal satellite prophages were, at the very least, transcriptionally active. Notably, the observed patterns of gene expression in these studies were consistent with the life cycle of satellite prophages as phage parasites. Firstly, in the mitomycin C experiment (Figure 4.1), satellite prophage genes were only upregulated an hour after the induction of full-length prophages, suggesting that induction of the satellite prophage may have occurred as a response to helper prophage replication. Furthermore, during the period of planktonic growth in broth, the non-structural genes of the two full-length prophage genomes were not differentially expressed (Figure 4.14), suggesting that the full-length prophage genome replication is unlikely. However, all of the satellite prophage genes and the structural genes of full-length prophages were upregulated (Figure 4.14), which is a gene expression pattern that would be expected during active satellite prophage replication. Previous works in different bacterial species have shown

that, due to the absence of structural genes, satellite prophages modulate the expression and assembly of the helper phage's structural proteins (*e.g.* capsid proteins) for their own replication and propagation [224].

These analyses led to the identification of a putative virulence-associated gene (*vapE*) in some of the satellite prophages among pneumococci. *In vivo* studies using a murine pneumococcal infection model and one example of a satellite prophage containing *vapE* showed that deletion of the whole prophage or *vapE* alone had a significant effect on pneumococcal virulence, and deletion of the whole prophage had a particularly strong effect and reduced recovered CFU for the sepsis model to less than 10^4 $\log_{10}$ (Figure 4.12). While this is a promising result, it does not span the range of the 44 satellite prophages described here, which necessitate further study in order to assess the generalisability of the findings. Nevertheless, a different satellite prophage, containing a homologue of *vapE* and carried by a genetically distinct pneumococcal host, was significantly overexpressed among pneumococci growing in planktonic form, which is akin to pneumococcal bacteraemia, rather than in a biofilm (a state in which pneumococci are less likely to be virulent [37]) (Figure 4.13), which provides further support for the hypothesis that satellite prophages play a role in pneumococcal virulence.

In vitro characterisation of the mutant strains indicated that the reduced virulence of the prophage mutant was related to impaired growth in serum rather than avoidance of opsonophagocytic killing. How the prophage influences pneumococcal growth in serum will require more detailed investigation, but the stronger phenotypic effect of loss of the whole prophage compared to *vapE* alone suggests that additional prophage genes are involved in virulence. For example,

the prophage is predicted to contain regulatory genes, which could potentially improve growth in serum by altering the expression of metabolic and transporter genes. Overall, while the specific mechanism driving virulence is not yet clear, the findings presented in this chapter are a proof-of-principle that experimental investigations of pneumococcal satellite prophages should be pursued, as such findings may hold the key to understanding important aspects of pneumococcal biology.

A genomic comparison between pneumococcal prophages and those belonging to other species of *Streptococcus* has not been thoroughly studied to date. For instance, it would be interesting to investigate how closely related the prophages identified in this chapter might be to those found in the other streptococcal species. Therefore, the next chapter will extend the investigation of prophages to the entire *Streptococcus* genus.

4.6 Supplementary information

Supplementary File 4.1. List of pneumococcal genomes used in this study.

Supplementary File 4.2. RNA sequencing data from the mitomycin C experiment.

Supplementary File 4.3. RNA sequencing data from the biofilm versus planktonic mode of growth experiment.

Supplementary File 4.4. Strains, plasmids and primers used in the study.

Supplementary File 4.5. A list of all the genes present among the 44 representative pneumococcal satellite prophage genomes.

5. Prophages and satellite prophages are widespread among *Streptococcus* species

The results described in this chapter were presented as an oral presentation at the 14th European Meeting on the Molecular Biology of the Pneumococcus, in Greifswald, Germany in June 2019. This chapter and chapter 4 were combined into one manuscript, which has been accepted for publication in *Nature Communications* (in press). A pre-print version of the manuscript was published in bioRxiv in December 2018 as:

Reza Rezaei Javan, Elisa Ramos-Sevillano, Asma Akter, Jeremy Brown and Angela B Brueggemann. 2018. Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis. *bioRxiv* [Preprint]. doi: 10.1101/203398

1,306 assembled genomes from 70 different *Streptococcus* species were used for this study. 482 pneumococcal genomes were retrieved from the Brueggemann BIGSdb database as described in Chapters 3 and 4. I downloaded the remaining 824 genomes belonging to non-pneumococcal *Streptococcus* species from the rMLST database and uploaded them to the Brueggemann BIGSdb database. The non-pneumococcal genome dataset was also analysed within a different study that is not included in this thesis and published as:

Ayako Kurioka†, Bonnie van Wilgenburg†, Reza Rezaei Javan†, Ryan Hoyle†, Andries J van Tonder, Caroline L Harrold, Tianqi Leng, Lauren J Howson, Dawn Shepherd, Vincenzo Cerundolo, Angela B Brueggemann and Paul Klenerman. 2018. Diverse *Streptococcus pneumoniae* Strains Drive a Mucosal-Associated Invariant T-Cell Response Through Major Histocompatibility Complex class I–Related Molecule–Dependent and Cytokine-Driven Pathways. *The Journal of Infectious Diseases.* doi: 10.1093/infdis/jix647

† = joint first-authorship

5.1 Abstract

Prophages are abundant within the bacterial world and are major contributors to the diversity of bacterial gene repertoires. Large numbers of bacterial genomes of a wide array of species are now in the public domain and this presents opportunities for genome mining of prophages at a scale that was previously impossible. To facilitate this process, I developed PhageMiner, a bioinformatics software used to identify prophage sequences in bacterial genomes. Using PhageMiner, I screened >1,300 genomes of 70 different *Streptococcus* species for prophage sequences. A diverse collection of nearly 800 prophages and satellite prophages were identified, the majority of which were discovered for the first time. Prophages and satellite prophages were widely distributed among streptococci and, contrary to the current dogma that suggests prophages are bacterial species-specific, there was convincing evidence that transmission of prophages occurred between genetically different streptococcal species. This study also revealed a remarkable intra- and interspecies variability in the amount of prophage sequences found within an individual bacterial genome. Furthermore, genes flanking prophage insertion sites were found to be more frequently involved in information storage and processing compared to the average for streptococcal genomes. Overall, the findings presented in this chapter suggested that prophages are likely to be influencing streptococcal biology and epidemiology to a much greater extent than previously appreciated.

5.2 Introduction and aims

The extent to which prophages make up the bacterial host genome and contribute to interstrain genetic variability remains an unanswered empirical question for the majority of the *Streptococcus* species, many of which have not previously been analysed for evidence of prophages. Genomic studies have thus far revealed that prophage-related sequences are highly prevalent within *S. pneumoniae* [212, 233, 306, 307], *S. pyogenes* [308] and *S. agalactiae* genomes [309], however, genus-wide analyses of the genomic diversity and population structure of streptococcal prophages are still lacking.

As discussed in the general introduction (see 1.4), virulence determinants associated with the shift of a bacterium from a commensal to a pathogen may include genes associated with prophages. Nevertheless, despite their potential role in the emergence and spread of new pathogenic clones, the host specificities of streptococcal prophages have not been systematically investigated.

The main aims of this chapter were to: a) use genome mining to identify prophages in a large and diverse set of >1,300 genomes of 70 different *Streptococcus* species; b) analyse the genomic diversity and population structure of streptococcal prophages; c) estimate the average prophage gene content within various streptococcal species; d) investigate the extent to which prophages serve as the source of interstrain variation in gene content; e) assess the proportion of shared genes between prophages and satellite prophages; and f) investigate the prophage insertion sites and the flanking genes upstream and downstream of each prophage.

5.3 Methods

5.3.1 Development of PhageMiner, a bioinformatics tool for prophage identification in bacterial genomes

Some *in silico* prophage detection tools are available that identify prophages using a reference database of known prophage genomes, thus their performance is strongly influenced by the size and composition of the reference dataset [234, 235]. In order to ensure a thorough discovery of previously unidentified prophages, manual curation of annotated genomes is required, however, this is not feasible for large genome studies [212, 236, 310]. To address these issues, a new user-supervised semi-automated computational tool, named PhageMiner, was developed in order to streamline the manual curation process for prophage sequence discovery. PhageMiner uses a mean shift algorithm combined with annotation-based genome mining in order to rapidly identify prophage sequences within complete or draft bacterial genomes. Notably, while PhageMiner significantly facilitates the manual curation process, it is not a fully automated pipeline and requires manual input by the user during key decisions, thus ensuring careful inspection of putative prophage clusters.

The PhageMiner pipeline consists of the following steps:

1) The bacterial genome of interest is annotated using the RAST annotation server [243] (http://rast.nmpdr.org) in order to create an annotated GenBank file, which is then input into the PhageMiner Python script.

2) The location and the annotated name of each ORF in the host genome is retrieved from the annotated GenBank file and saved to a comma-separated value (CSV) file using the Biopython package (http://biopython.org) [311].

3) A number of predefined user-adjustable phage-associated keywords are used to scan the CSV file generated in the previous step. Any ORF containing a matching string (*e.g.*, “phage”, “lytic amidase”, "tail fiber protein", *etc.*) in its annotation name is deemed a ‘hit’.

4) An additional set of predefined user-adjustable keywords are used to discard any matching hits with annotation names that resemble phages but are not prophage genes (*e.g.* ‘macrophage’).

5) Using the mean shift clustering method in Scikit-Learn machine learning library (https://scikit-learn.org) [312], the location of the remaining phage hits respective to each other and to the size of the host genome are used to identify clusters of bacteriophage-related genes. During this step, minimal manual inputs by the user are requested in order to ensure correct identification of prophage regions. If necessary, clustering can be repeated with a different sensitivity as redefined by the user, or alternatively, the coordinates corresponding to each suspected prophage region can be entered manually. The pipeline is aborted at this stage if no clusters of bacteriophage-related genes are detected or manually defined by the user.

6) Once clusters of bacteriophage-related genes are identified, PhageMiner creates various figures and tables related to each of the suspected prophage regions, the most important of which are: a schematic diagram of the coding regions; the location of the prophage region in the chromosome, including the flanking genes adjacent to the prophage region; the presence of any assembly

gaps; and the nucleotide sequences of the ORFs in the cluster. If necessary, the number of flanking genes displayed in each figure can be manually adjusted.

7) Based on the decisions made by the user, the putative prophage genomes are either rejected or extracted as a separate GenBank file and categorised into three groups: full-length prophages, satellite prophages and unknown phage-related regions.

The source code of PhageMiner is deposited in GitHub at https://github.com/RezaRezaeiJavan/PhageMiner.

5.3.2 Streptococcal genomes dataset

In total, 1,306 assembled genomes from 70 different species of the genus *Streptococcus* were selected for this analysis (Supplementary File 5.1). 482 genomes belonged to the pneumococcal global representative dataset described in Chapter 4 (Supplementary File 4.1). The remaining 824 genomes belonged to 69 different *Streptococcus* species, and up to 50 genomes per species were selected for analyses from the ribosomal MLST database (https://pubmlst.org/rmlst) [271]. When more than 50 genomes were available, the population structure of the species was depicted using PHYLOViZ [272] and genomes were selected to maximise the population-level diversity of the species from the available genomes. All genomes were stored in a BIGSdb database [239] and annotated using the RAST server [243]. The prophage content for all genomes in the study dataset were calculated using the PhageContentCalculator script (see Chapter 2).

5.3.3 Sequence analyses of prophages

All putative prophage sequences were inspected manually using Artemis Genome Browser version 8 [313] and those containing ambiguous bases (Ns) and/or assembly gaps (n=411) were excluded from further analyses. The total number of ORFs, overall length of genome and GC content of each prophage were calculated using Geneious version 11.1 (Biomatters Ltd.). Schematic diagrams of the coding regions of the prophages were produced in Geneious and edited using Adobe Illustrator (Adobe Inc.).

5.3.4 Investigation of prophage insertion sites

The prophage insertion sites within the bacterial genomes were investigated for all streptococcal species for which at least one complete (closed) bacterial genome was available (n=29). Prophage insertion sites containing ambiguous bases or assembly gaps were excluded from further analyses. The genomes were divided into eight equally-sized segments and the prevalence of prophages per segment was calculated, in order to assess the relative location of prophages within streptococcal bacterial genomes.

To investigate the location of prophages relative to the putative function of the flanking bacterial genes, the sequences of the five bacterial genes both upstream and downstream of each prophage were retrieved. Bacterial gene sequences were categorised into Clusters of Orthologous Groups (COGs) using eggNOG-mapper [314] based on eggNOG 4.5 orthology data [315]. For comparison, a reference set of 70 streptococcal genomes, each representing a different streptococcal species, was compiled. All bacterial genes were assigned a COGs category using eggNOG

and the average prevalence of each COG category across the combined set of 70 reference streptococcal genomes was calculated.

5.3.5 Estimate of phylogenetic relationships among *Streptococcus* species

A phylogenetic tree was generated using concatenated sequence data from 53 ribosomal loci among all streptococcal genomes used in this study using the BIGSdb PhyloTree plugin [239]. The tree was graphically simplified to the species level by collapsing clades containing genomes from the same species into a single leaf using iTOL [248]. The tree was further modified and coloured using Adobe Illustrator.

5.4 Results

5.4.1 Prophages are a significant component of genomes of most clinically relevant streptococcal species

Applying PhageMiner to a collection of 1,306 genomes from 70 different streptococcal species resulted in the identification of 415 full-length and 348 satellite prophage genomes (Table 5.1). The prophage content within each streptococcal genome was estimated and revealed wide variability between streptococcal species, ranging from 0.4% of the *Streptococcus thermophilus* genome to 9.5% of the *S. pyogenes* genome (Table 5.2 and Supplementary File 5.1). These data also revealed variability in prophage content among genomes of the same streptococcal species, for example, between 1-19% of the genes within individual *S. pyogenes* genomes were of full-length prophages (Table 5.2). The overall prevalence of satellite prophages also varied, ranging from 0.1% among *S. mutans* and *Streptococcus sanguinis* genomes to 4.5% of *the Streptococcus dysgalactiae* genomes (Table 5.2).

Table 5.1 - Descriptive statistics for the identified full-length and satellite prophages.

Host Genomes		Full-length Prophages				Satellite Prophages			
Streptococcus species (n[a] of genomes)	Avg[b] GC[c] content	n	Avg size (bp[d])	Avg GC content	Avg n of genes	n	Avg size (bp)	Avg GC content	Avg n of genes
S. pneumoniae (482)	39.6%	66	37,346	39.6%	56	44	12,936	37.2%	21
S. pyogenes (50)	38.4%	67	39,113	38.5%	56	21	12,825	36.3%	23
S. agalactiae (50)	35.3%	37	37,393	39.4%	51	19	14,096	35.7%	23
S. dysgalactiae (50)	39.3%	34	39,344	38.6%	57	22	13,455	35.5%	24
S. suis (50)	41.2%	28	35,696	41.1%	52	24	10,216	38.1%	17
S. equi (50)	41.6%	18	38,096	39.0%	51	4	12,813	36.3%	20
S. parauberis (16)	35.5%	14	38,455	35.8%	55	17	13,564	32.7%	20
S. mitis (49)	40.1%	12	37,559	40.0%	55	35	12,074	36.2%	18
S. oralis (49)	41.0%	12	35,654	39.7%	50	14	11,649	36.0%	17
S. anginosus (24)	38.6%	9	37,077	39.6%	51	16	10,921	36.8%	17
S. equinus (27)	37.3%	10	40,614	39.1%	56	13	10,075	34.2%	15
S. constellatus (10)	38.0%	10	43,497	39.9%	51	4	10,347	35.3%	21
S. uberis (13)	36.5%	9	39,936	37.7%	54	11	11,652	33.2%	18
S. gallolyticus (17)	37.5%	8	37,207	38.1%	52	9	9,759	35.1%	14
S. canis (11)	39.5%	8	37,776	40.4%	51	3	13,895	36.3%	25
S. urinalis (4)	34.1%	7	38,072	37.2%	53	7	9,251	34.1%	14
S. parasanguinis (31)	41.8%	7	38,470	41.4%	54	6	9,435	34.4%	11
S. gordonii (22)	40.4%	7	35,522	40.1%	47	3	10,069	35.4%	15
*S. pseudopneumoniae (*16)	39.8%	5	36,007	39.7%	62	9	11,870	38.2%	19
S. iniae (8)	36.6%	5	35,285	36.8%	52	1	12,203	30.0%	18
S. salivarius (32)	39.7%	4	40,116	42.2%	42	4	9,577	37.7%	15
S. infantis (4)	39.4%	4	35,734	39.3%	49	1	11,158	38.0%	14
S. porcinus (2)	36.7%	4	37,741	39.6%	51	0	-	-	-
S. pseudoporcinus (5)	37.2%	3	38,992	38.3%	62	2	8,343	33.8%	19
S. entericus (1)	44.6%	2	41,400	43.0%	59	2	10,349	42.4%	18
S. himalayensis (1)	41.3%	2	39,219	40.8%	49	2	11,811	38.8%	16
S. marmotae (1)	40.9%	2	37,995	43.0%	66	2	14,088	40.6%	21
S. infantarius (2)	37.6%	2	33,058	38.4%	48	1	9,627	35.7%	14
S. *azizii* (3)	42.7%	2	43,572	41.2%	58	0	-	-	-
S. henryi (2)	38.6%	2	40,013	39.8%	58	0	-	-	-
S. ictaluri (2)	38.1%	2	24,012	38.4%	32	0	-	-	-
S. intermedius (9)	37.6%	1	33,366	38.0%	49	6	13,935	36.4%	23
S. pasteurianus (5)	37.3%	1	35,546	38.0%	45	4	10,524	35.0%	15
S. lutetiensis (2)	37.6%	1	37,997	39.0%	51	3	10,881	35.1%	15
S. halotolerans (1)	39.2%	1	41,477	38.1%	52	2	13,216	36.9%	20
S. hyovaginalis (1)	39.9%	1	39,028	37.6%	60	2	12,180	37.0%	22
S. acidominimus (1)	42.6%	1	38,191	40.1%	52	1	9,826	39.5%	15
S. cristatus (4)	42.6%	1	38,959	39.8%	52	1	13,630	40.4%	17
S. macedonicus (2)	37.5%	1	38,767	38.5%	52	1	10,632	33.8%	14
S. phocae (2)	39.5%	1	46,987	37.4%	76	1	12,626	36.2%	20
S. cuniculi (1)	43.4%	1	28,958	40.4%	44	0	-	-	-
S. orisratti (1)	38.5%	1	37,447	41.2%	42	0	-	-	-
S. porci (1)	40.8%	1	40,394	35.3%	45	0	-	-	-
S. thoraltensis (1)	38.4%	1	40,339	37.5%	57	0	-	-	-
S. thermophilus (32)	39.0%	0	-	-	-	16	8,429	37.3%	14
S. sanguinis (37)	43.0%	0	-	-	-	6	11,336	38.5%	18
S. vestibularis (6)	39.5%	0	-	-	-	3	10,768	36.5%	14
S. castoreus (1)	37.8%	0	-	-	-	2	10,268	36.1%	18
S. merionis (1)	41.7%	0	-	-	-	1	10,429	39.5%	17
S. oligofermentans (1)	42.1%	0	-	-	-	1	10,319	36.2%	16
S. caballi (1)	40.4%	0	-	-	-	1	9,970	38.7%	17
S. plurextorum (1)	41.1%	0	-	-	-	1	13,633	35.0%	27
Other *strep. s*pecies (111)	-	0	-	-	-	0	-	-	-
Total	**-**	**415**	**-**	**-**	**-**	**348**	**-**	**-**	**-**
Avg.	**39.4%**	**-**	**38,003**	**39.2%**	**54**	**-**	**11,767**	**36.2%**	**19**

a, n = number. **b,** Avg = Average. **c,** GC = guanine and cytosine. **d,** bp = base pairs.

Table 5.2 – Average prophage content within each streptococcal species.

Species (number of genomes)	Full-length Prophages				Satellite Prophages			
	Min[a]	Max[b]	Avg[c]	SD[d]	Min	Max	Avg	SD
S. pyogenes (50)	1.0%	19.0%	9.5%	4.1	0.4%	3.3%	1.3%	0.9
S. urinalis (4)	6.8%	9.1%	8.0%	1.0	1.5%	1.6%	1.5%	0.1
S. pseudopneumoniae (16)	3.9%	7.0%	5.4%	1.2	0.7%	2.8%	1.7%	0.7
S. parauberis (16)	0.4%	10.5%	5.4%	2.9	0.4%	3.2%	1.6%	1.0
S. canis (11)	2.0%	9.5%	4.5%	2.9	2.0%	2.6%	2.3%	0.2
S. dysgalactiae (50)	0.7%	11.9%	3.8%	2.3	0.3%	4.5%	1.3%	0.8
S. suis (50)	0.5%	9.1%	3.7%	2.3	0.4%	3.7%	1.3%	0.7
S. equi (50)	0.5%	11.1%	4.4%	3.4	0.2%	2.4%	0.6%	0.5
S. agalactiae (50)	0.3%	8.7%	4.0%	2.1	0.2%	2.4%	1.0%	0.6
S. infantis (4)	1.5%	7.6%	4.2%	2.2	0.4%	1.3%	0.8%	0.3
S. uberis (13)	0.5%	6.7%	3.2%	2.0	0.9%	2.9%	1.7%	0.7
S. constellatus (10)	1.4%	6.7%	3.9%	2.1	0.4%	1.3%	0.9%	0.4
S. azizii (3)	3.7%	5.5%	4.3%	0.9	0.5%	0.5%	0.5%	0.0
S. pseudoporcinus (5)	1.9%	6.7%	3.8%	2.3	0.5%	1.0%	0.7%	0.2
S. mitis (49)	0.4%	9.3%	2.5%	2.1	0.3%	3.5%	1.6%	0.8
S. pasteurianus (5)	1.3%	5.2%	2.5%	1.6	0.8%	2.3%	1.4%	0.5
S. gallolyticus (17)	0.5%	6.2%	2.9%	1.4	0.4%	1.8%	0.9%	0.4
S. pneumoniae (482)	0.4%	8.4%	2.9%	2.1	0.2%	3.3%	0.9%	0.6
S. anginosus (24)	0.4%	7.6%	2.4%	2.1	0.4%	2.3%	1.2%	0.6
S. oralis (49)	0.5%	6.4%	2.1%	1.8	0.2%	2.8%	1.0%	0.6
S. iniae (8)	0.5%	8.6%	2.3%	3.0	0.2%	1.5%	0.6%	0.4
S. parasanguinis (31)	0.4%	5.4%	2.2%	1.6	0.3%	2.0%	0.7%	0.5
S. equinus (27)	0.4%	7.2%	1.9%	2.0	0.2%	2.7%	0.9%	0.7
S. cristatus (4)	0.8%	3.1%	1.7%	0.9	0.5%	1.5%	0.8%	0.4
S. intermedius (9)	0.5%	3.4%	1.1%	0.8	0.5%	1.8%	1.3%	0.5
S. gordonii (22)	0.4%	5.6%	1.9%	1.6	0.2%	1.3%	0.5%	0.3
S. vestibularis (6)	0.4%	0.8%	0.5%	0.1	0.5%	2.3%	1.6%	0.6
S. salivarius (32)	0.4%	3.0%	1.2%	1.0	0.2%	1.7%	0.5%	0.4
S. thermophilus (32)	0.3%	0.9%	0.4%	0.1	0.2%	1.4%	0.9%	0.4
S. sanguinis (37)	0.3%	3.4%	0.8%	0.7	0.1%	1.3%	0.4%	0.4
S. sobrinus (42)	0.4%	0.7%	0.5%	0.1	0.3%	0.6%	0.4%	0.1
S. mutans (50)	0.2%	1.5%	0.5%	0.2	0.1%	0.3%	0.2%	0.1

a, Min = Minimum. **b,** Max = Maximum. **c,** Avg = Average. **d,** SD = Standard Deviation. Values for streptococcal species for which at least three genomes were available in the study dataset are shown. A more comprehensive table detailing the number of prophage related genes in each streptococcal genome in the study dataset is provided in Supplementary File 5.1.

5.4.2 Full-length and satellite prophages are separate entities with little effective genetic exchange between them

Genomic comparisons between full-length and satellite prophage genomes revealed that satellite prophages had a lower GC-content than full-length prophages and were about a third of the size both in terms of length of sequence and the number of genes they contained (Figure 5.1).

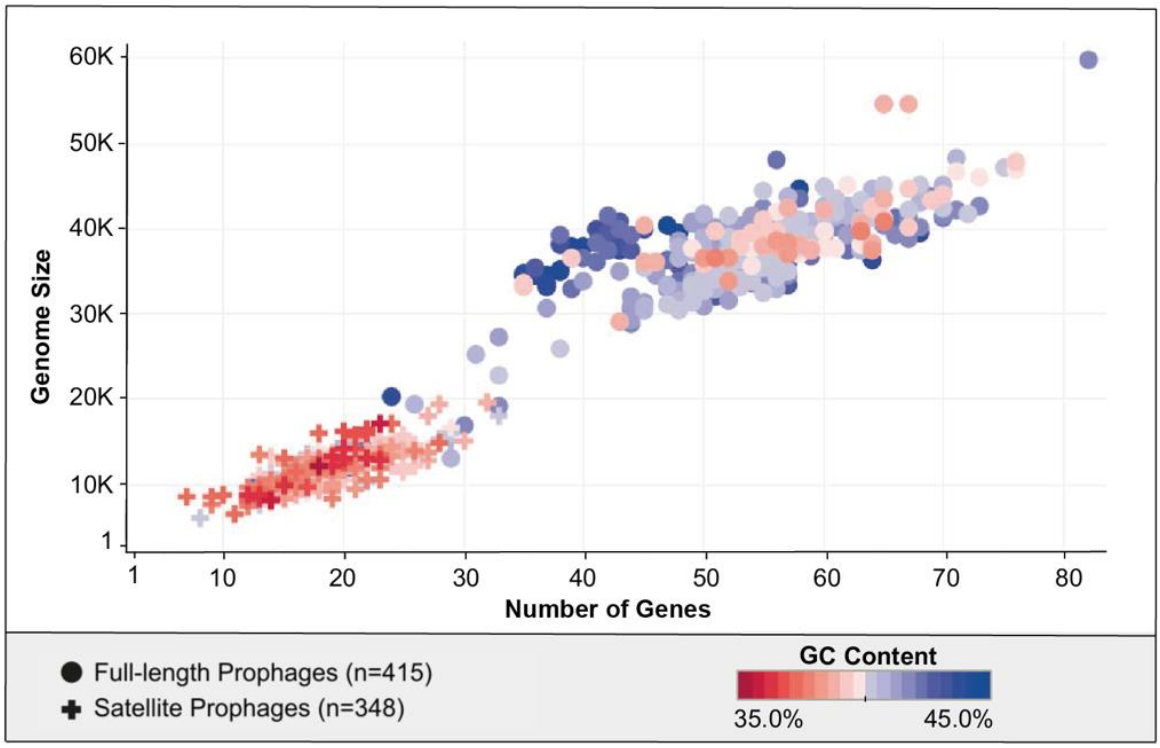


Figure 5.1 - Graphical representation of all prophages by average genome size and number of genes. Each prophage is coloured to depict its average GC content**.**

Streptococcal satellite prophage sequences have historically been regarded as "remnant" or "defective" prophages in a state of mutational decay [218, 227, 293, 294]. However, this study found that satellite prophage sequences were highly conserved over many decades, *e.g.* one satellite prophage sequence was present

among pneumococci with isolation dates ranging from 1939 to 2006, and it had retained >99.98% nucleotide sequence similarity across the entire sequence (Figure 5.2), suggesting that it is under strong positive evolutionary pressure and likely provides an important biological function. The highly conserved nature of this satellite prophage was particularly striking given that the pneumococcus has long been known to be a highly recombinant organism [316, 317].

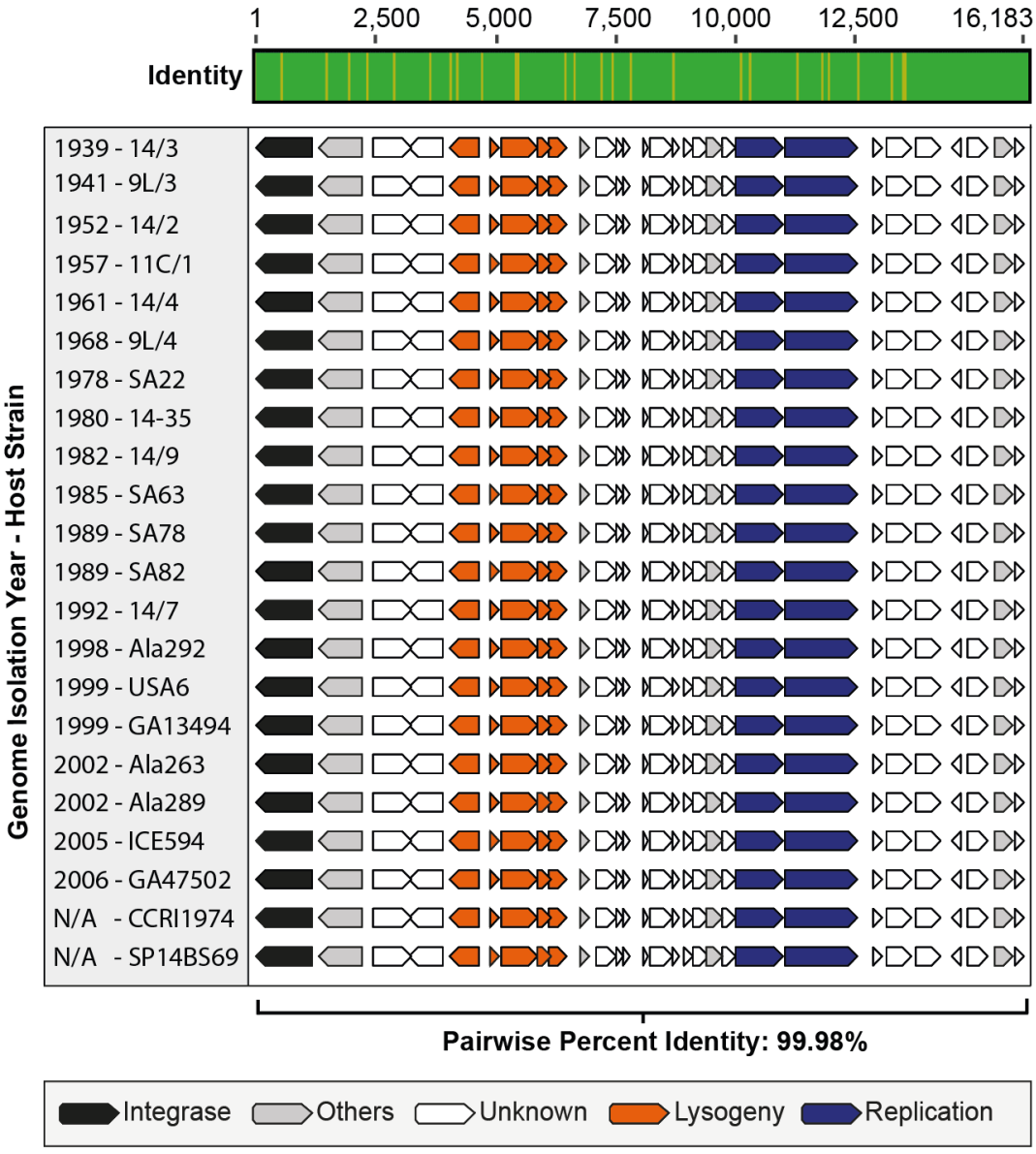


Figure 5.2 - One satellite prophage that was found among pneumococci isolated between 1939 and 2006. All of these satellite prophage sequences were nearly identical at the nucleotide level.

An unrooted phylogenetic tree of all streptococcal prophage genomes in the dataset depicted full-length and satellite prophages as two clearly distinct groups (Figure 5.3). Furthermore, the genes of satellite prophages were unique to those of full-length prophages; 93% of all satellite prophage genes were not found in any full-length prophage (Figure 5.4). Five flanking genes upstream and downstream of each prophage in the dataset were investigated, which revealed that prophages and satellite prophages never shared an insertion site (Supplementary File 5.2). Taken together, these findings suggest that these satellite prophage sequences were not recent remnants of previous lysogenisation by full-length prophages, but rather that they belong to a unique family of mobile genetic elements (MGEs).

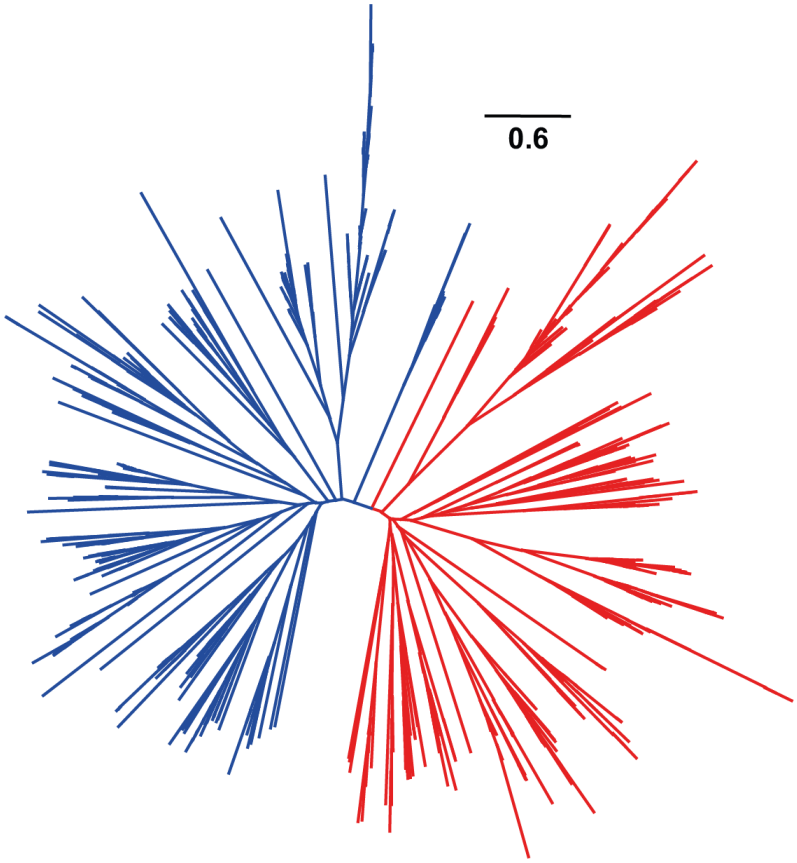


Figure 5.3 - An unrooted phylogenetic tree of all streptococcal prophage genomes identified in the dataset. Blue branches mark full-length prophages and red branches mark satellite prophages.

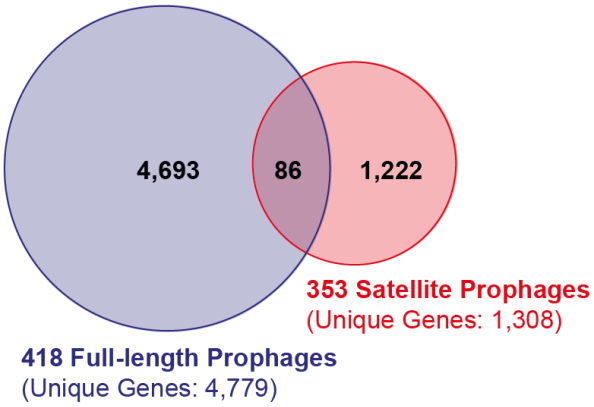


Figure 5.4 – Venn diagram showing genes shared between full-length prophages and satellite prophages at a threshold of >70% amino acid sequence similarity.

5.4.3 Streptococcal prophages demonstrate a structured population

Both full-length and satellite prophages demonstrated well-conserved patterns of genome organisation and synteny, regardless of the species that they were isolated from (Figure 5.5). Similar to other non-streptococcal prophages (Figure 5.6), genes encoding specific functions were often found clustered together in the prophage genome, although note that the function of many genes is still unknown and therefore the delineation of discrete gene clusters remains problematic (Figure 5.5).

Comparisons of all the identified prophage sequences depicted major and minor clusters for both full-length and satellite prophages (Figure 5.7 and Supplementary File 5.3). Phages are generally believed to be bacterial species-specific and even specific to one or a few strains of a single bacterial species [206, 318-320]; however, these data showed that prophages isolated from different streptococcal species were often found in the same phylogenetic cluster, suggesting that cross-species transmissions are more common than hitherto realised. Intriguingly, despite the close relatedness of their prophages, the bacterial hosts were not necessarily closely related (phylogenetically) (Figure 5.7, 5.8 and 5.9).

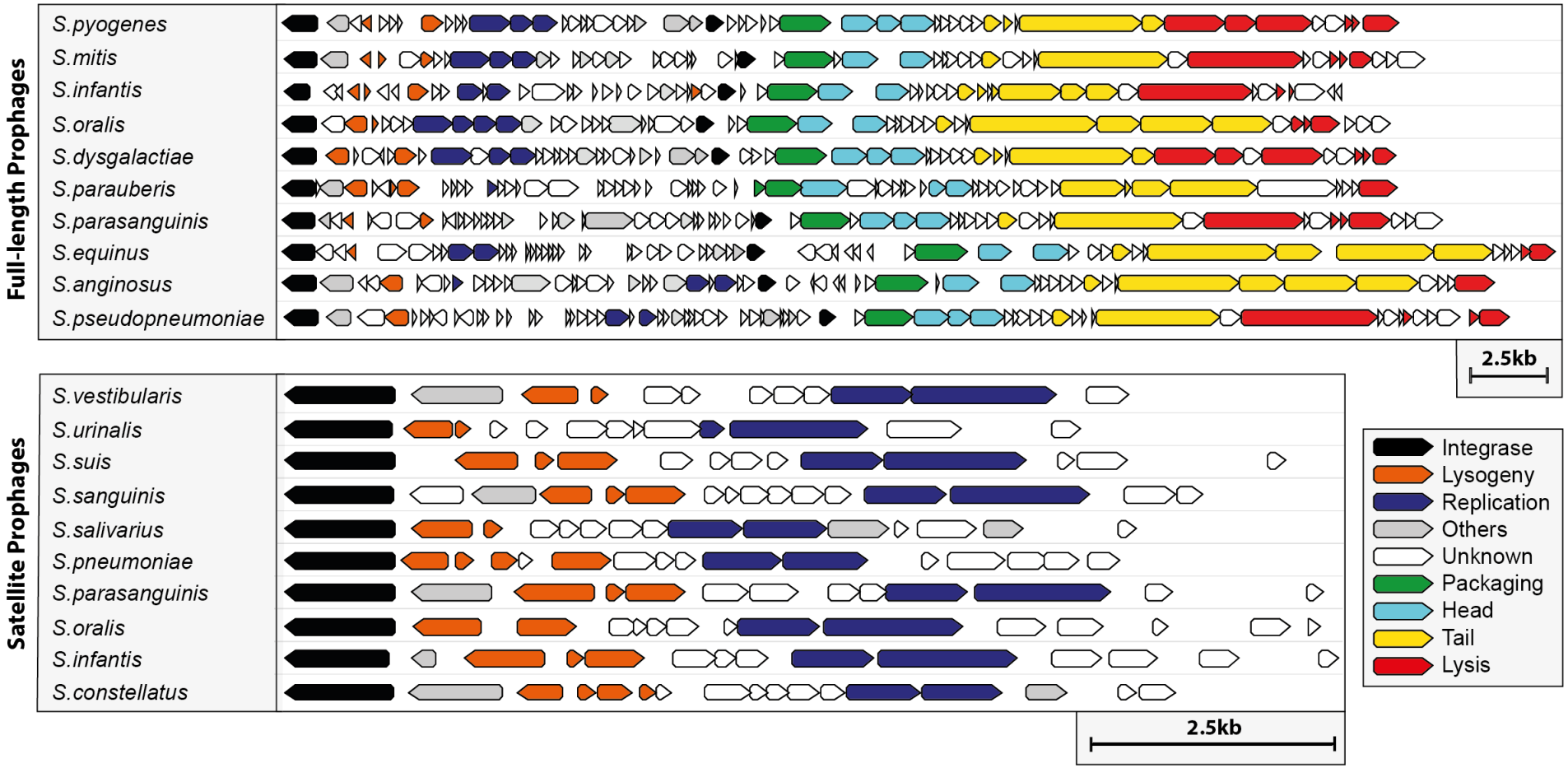


Figure 5.5 - Full-length and satellite prophages from different streptococcal species demonstrate similar patterns of genome organisation and synteny.

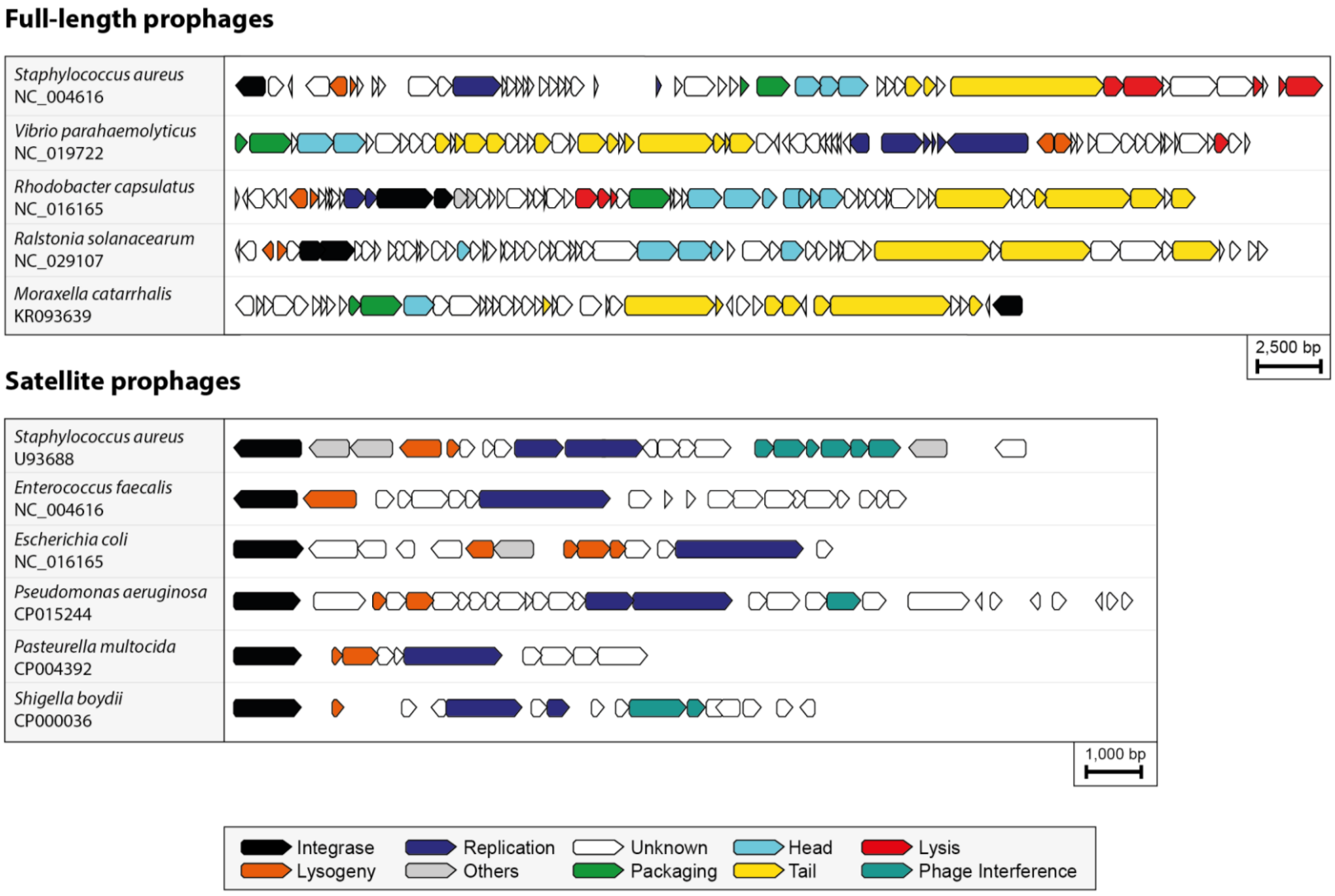


Figure 5.6 – Schematic representation of examples of full-length and satellite prophage genomes from different non-streptococcal species.

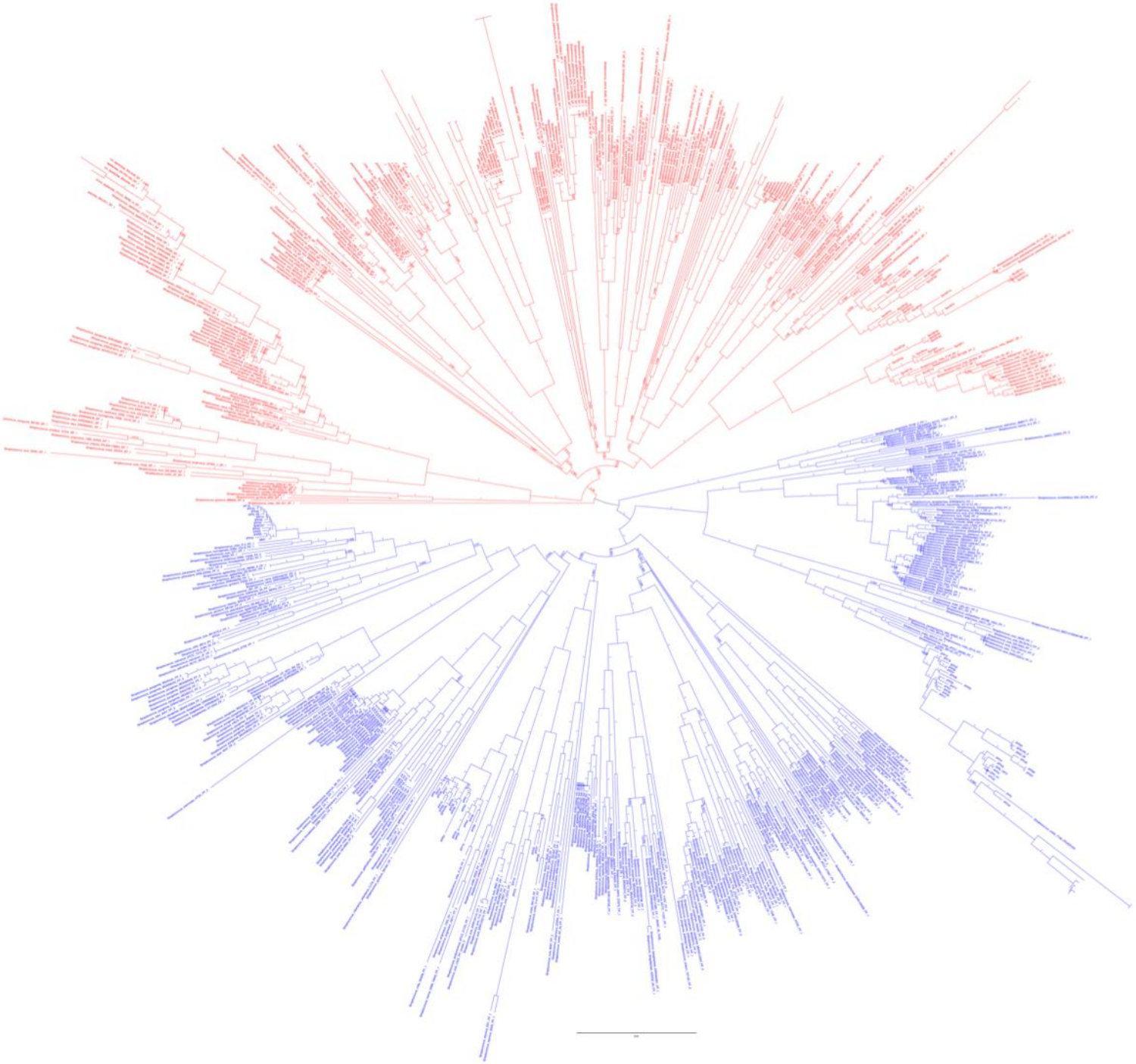


Figure 5.7 – A neighbour-joining phylogenetic tree of all streptococcal prophage genomes identified in the dataset. Blue branches mark full-length prophages and red branches mark satellite prophages (see Supplementary File 5.3 for a larger version of the tree).

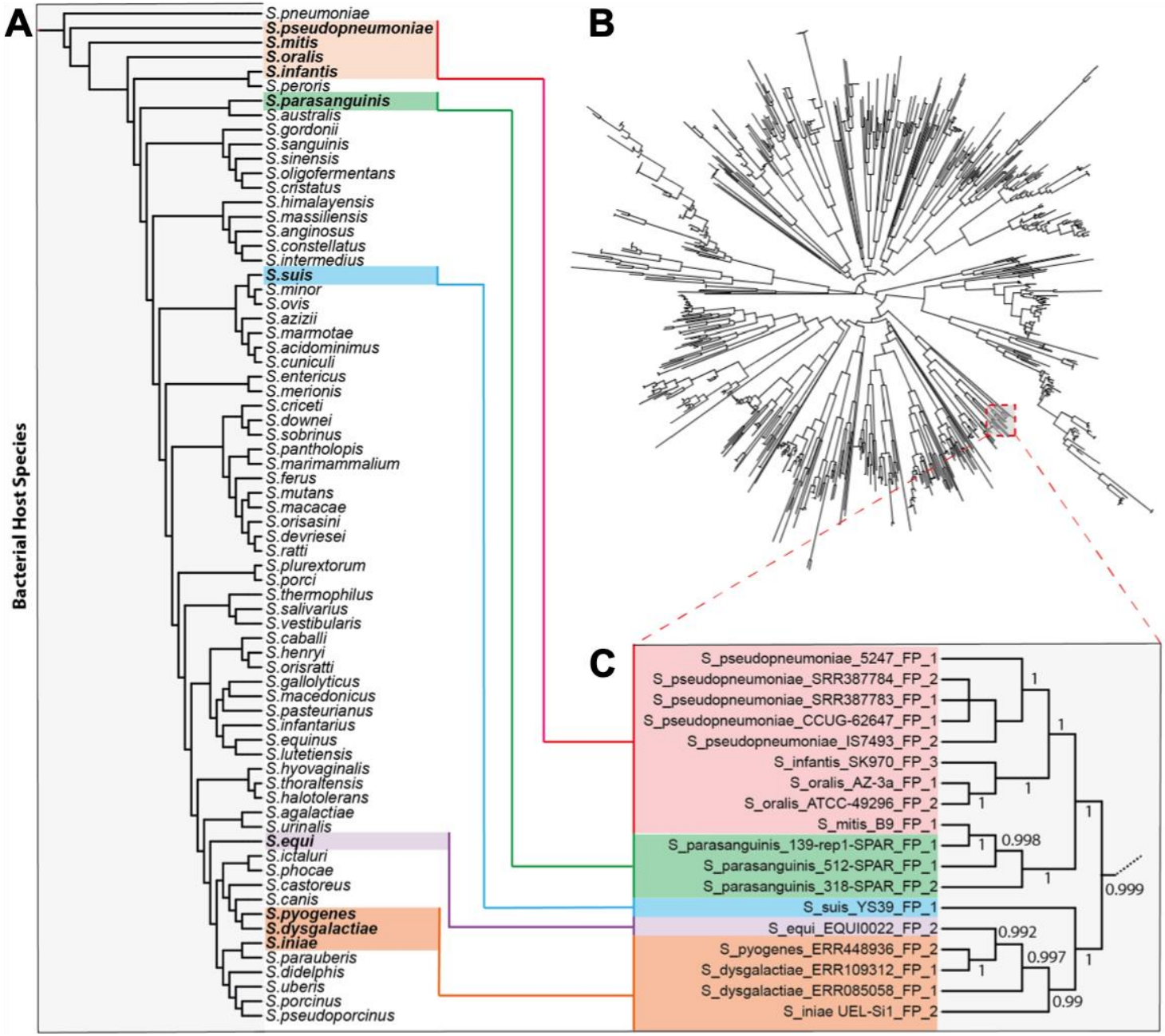


Figure 5.8 - Evidence for cross-species transmission of prophages. **A**, Phylogenetic tree of bacterial host genomes constructed on the basis of the concatenated sequences of 53 ribosomal multilocus sequence type loci among all streptococcal genomes used in this study. The tree was graphically simplified to the species level by collapsing clades containing genomes from the same species. **B** and **C**, Phylogenetic tree depicting all prophages identified in this study (**B**) (see Supplementary File 5.3 for a large version of the tree) and a zoomed-in branch from the same tree (**C**) (with branch lengths ignored for illustrative purposes) depicting one example of a cluster of full-length prophages that were found among multiple streptococcal species (see also Figure 5.9 for a distance matrix of pairwise similarity among these 18 prophages).

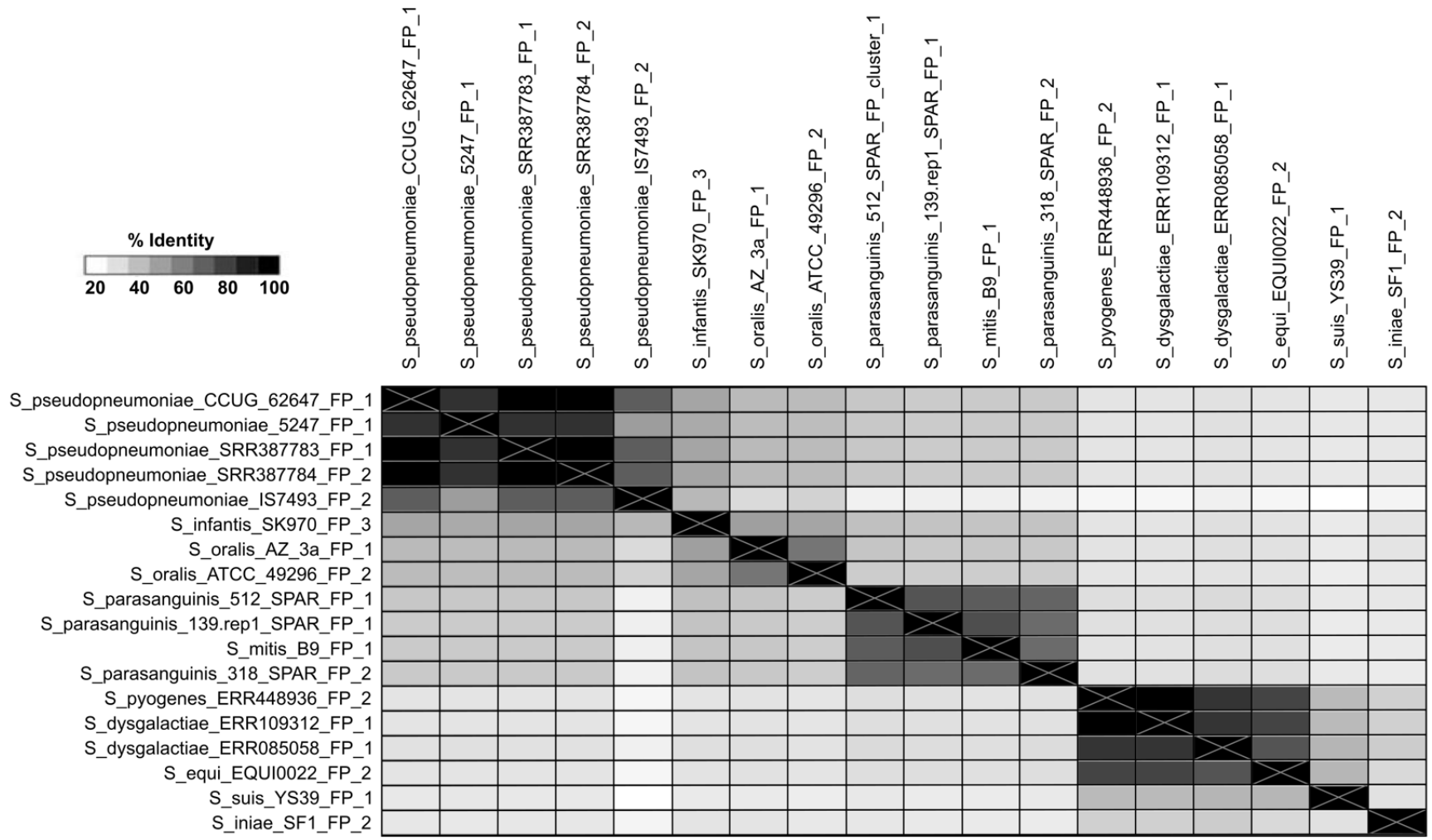


Figure 5.9 - Distance matrix of pairwise similarity among one example of a cluster of full-length prophages that were found among multiple streptococcal species. See Figure 5.8 for a phylogenetic tree depicting these 18 prophages. Phylogenetic tree depicting all prophages identified in this study is shown in Figure 5.7.

5.4.4 Streptococcal prophages were more frequently inserted among genes involved in information storage and processing

In Chapter 4, 28.3% of satellite prophage sequences were inserted in a specific site that was very close to the origin of replication (*oriC*) (see 4.4.1). This prompted further investigation into whether other factors beyond the integrase sequence determined the prophage insertion site. Complete (finished) genomes of 29 different streptococcal species were analysed and the location of prophage sequences within the genome varied widely between different species. There was no obvious pattern or preference towards the specific location of prophage insertion across the streptococcal genomes (Figure 5.10).

Five flanking genes upstream and downstream of each prophage in the dataset were retrieved for functional classification using gene ontology analyses. This revealed that nearly one-third of all the bacterial flanking genes were involved in replication, recombination, DNA repair, transcription, translation and ribosomal structure and biogenesis (Figure 5.11 and Table 5.3). One-quarter of flanking genes were involved in metabolic processes, but equally, one-quarter of all flanking genes were of unknown function. The remaining flanking genes were involved in other cellular processes and signalling (Table 5.3).

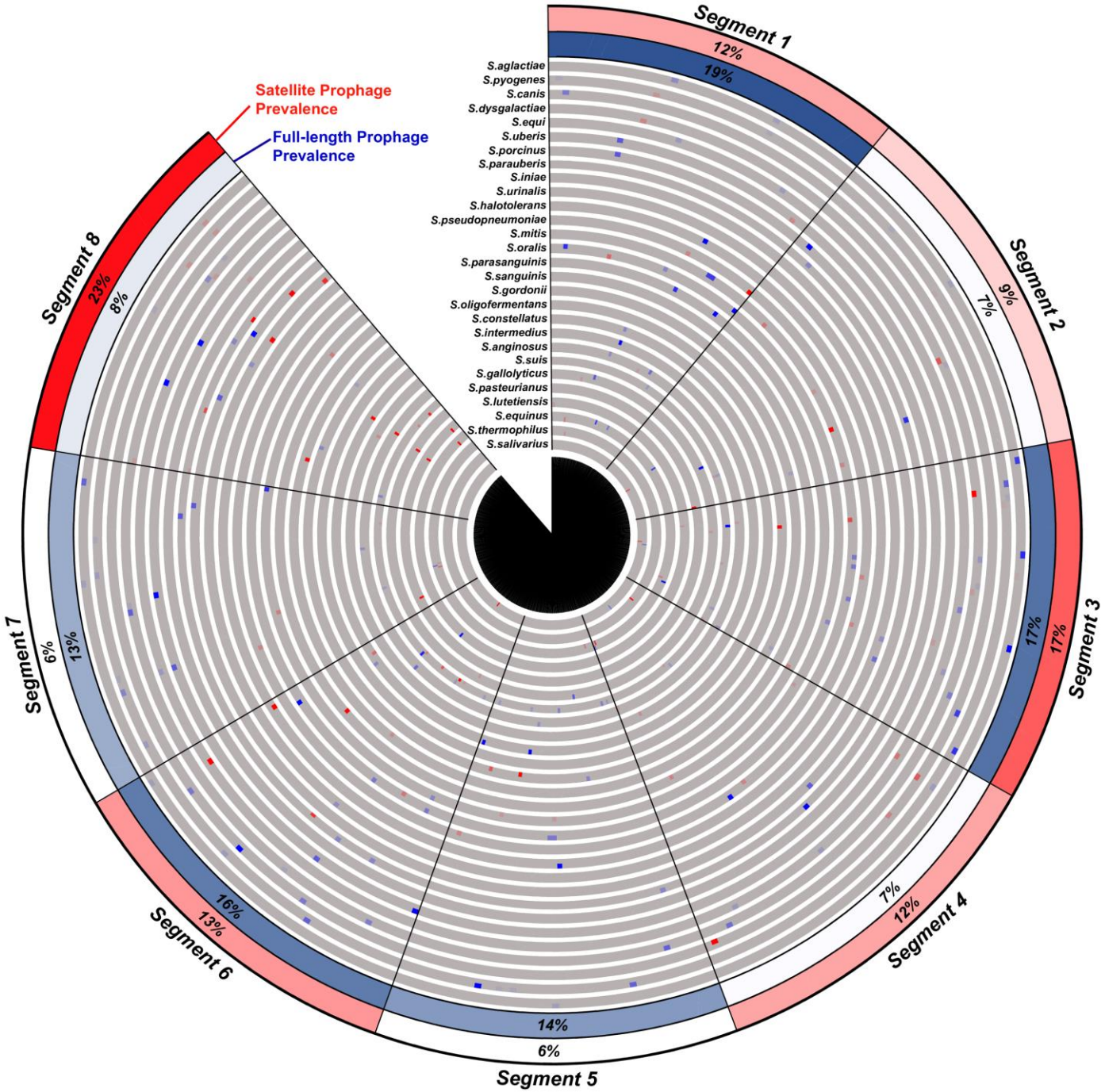


Figure 5.10 – The location of prophage insertion sites within the bacterial genomes. One finished genome of each of 29 streptococcal species was divided into eight non-overlapping segments of equal length according to the number of base pairs, and the percentages of prophages situated in each segment were quantified.

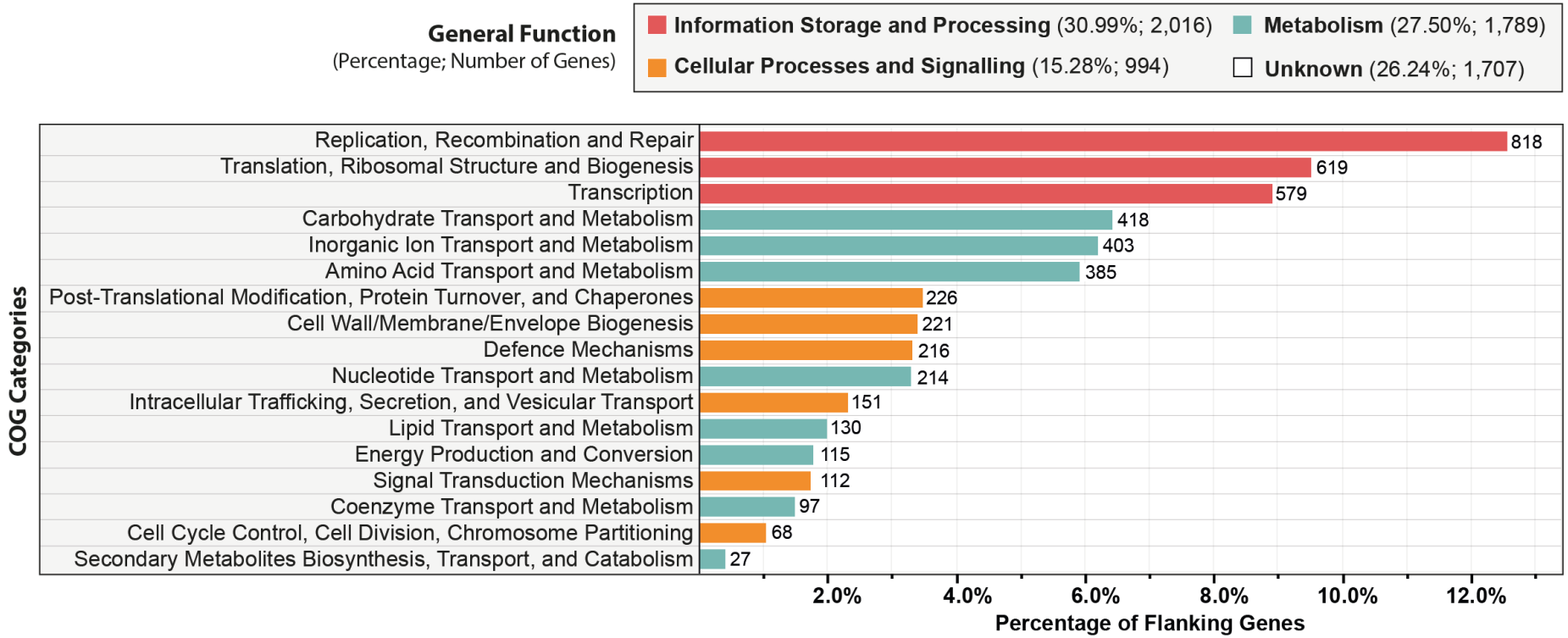


Figure 5.11 - Insertion sites of prophages within the streptococcal genomes. The streptococcal flanking genes upstream and downstream of all integrated full-length and satellite prophages were retrieved for functional classification and are depicted here based upon their COG (clusters of orthologous groups) classifications.

Table 5.3 – The proportions of genes in each COGs (clusters of orthologous groups) for streptococcal genomes on average compared to those flanking prophages.

Type	Functional categories	Flanking prophage genes		*Streptococcus* spp. genes[a]		Difference
		Total (n)	Percentage	Total (n)	Percentage	
Cellular processes and signalling	Posttranslational modification, protein turnover, chaperones	226	3.5%	3,930	3.2%	
	Cell wall/membrane/envelope biogenesis	221	3.4%	6,492	5.2%	
	Defense mechanisms	216	3.3%	4,119	3.3%	
	Intracellular trafficking, secretion, and vesicular transport	151	2.3%	1,657	1.3%	
	Signal transduction mechanisms	112	1.7%	3,187	2.6%	
	Cell cycle control, cell division, chromosome partitioning	68	1.0%	1,367	1.1%	
Subtotal:		**994**	**15.3%**	**20,752**	**16.7%**	-1.4%
Information storage and processing	Replication, recombination and repair	818	12.6%	10,255	8.2%	
	Translation, ribosomal structure and biogenesis	619	9.5%	10,082	8.1%	
	Transcription	579	8.9%	9,066	7.3%	
Subtotal:		**2,016**	**31.0%**	**29,403**	**23.6%**	7.4%
Metabolism	Carbohydrate transport and metabolism	418	6.4%	9,779	7.9%	
	Inorganic ion transport and metabolism	403	6.2%	5,868	4.7%	
	Amino acid transport and metabolism	385	5.9%	9,704	7.8%	
	Nucleotide transport and metabolism	214	3.3%	4,308	3.5%	
	Lipid transport and metabolism	130	2.0%	2,472	2.0%	
	Energy production and conversion	115	1.8%	4,295	3.5%	
	Coenzyme transport and metabolism	97	1.5%	2,804	2.3%	
	Secondary metabolites biosynthesis, transport and catabolism	27	0.4%	744	0.6%	
Subtotal:		**1,789**	**27.5%**	**39,974**	**32.1%**	-4.6%
Poorly characterised	Function unknown	1,707	26.2%	34,377	27.6%	-1.4%
Total[b]		**6,506**	**100%**	**124,526**	**100%**	

a, Sum total of genes in each category, as represented by a dataset consisting of one reference genome for each of 70 different *Streptococcus*. **b,** Chi square test of 2x4 contingency table (flanking prophage genes *vs.* streptococcal genes among four COG groups): p value <0.00001.

For comparison, one genome of each of the 70 different streptococcal species were selected and used to determine the clusters of orthologous groups (COGs) for all streptococcal genes, and subsequently, the genome-wide streptococcal data were compared to the COGs represented by the prophage flanking genes in the overall dataset (Table 5.4). This demonstrated that the distributions of COGs categories were significantly different, and while prophage flanking genes were more likely to be in the information storage and processing COGs category, the most prevalent COGs category among all streptococcal genes was metabolism (32.1% of all genes). A list of all prophage insertion sites and their flanking genes is available in Supplementary File 5.2.

5.5 Discussion

This study was the first large-scale analysis of the genomic diversity and population structure of prophages across a wide range of streptococcal species and the data revealed that prophages and satellite prophages were two clearly different entities and were widely distributed among streptococci. The main overall finding was that prophages are likely influencing streptococcal biology and epidemiology to a much larger extent than previously realised, given the high proportion of prophage DNA found in many streptococcal species. Many of the streptococci investigated in this study are important human and animal pathogens, and prophages were found to be a significant component of the genomes of most clinically-relevant streptococci.

Another major finding was the evidence for cross-species transmission of prophages between unrelated streptococcal species. Historically, the prevailing dogma is that phages have a narrow host range [206, 318-320], but the data

presented in this chapter challenge this view and suggest that prophage transmission across bacterial species is more common than previously recognised. Other investigators have also recently suggested that some phages may have a broader host range than previously appreciated [321].

The dataset was designed to be comprised of streptococci that were genetically different and geographically widely distributed, rather than from a very defined population. In the context of a highly diverse dataset, there are two plausible explanations for finding the same or highly similar prophages in different species, the most likely of which is cross-species transmissions of prophages or at least prophage sequences. The alternative explanation is a shared common ancestor, but this is far less likely given the overall variation among prophage sequences, at least on any reasonable time frame. The implications of these findings are that host specificity should be considered when trying to understand the precise role of prophages in streptococcal biology and when considering whether phages might be used in any therapeutic interventions. One of the commonly cited advantages of phage therapy over traditional antibiotics is the narrow host range of the phages, which allows a particular phage to target pathogenic bacteria without harming the normal microflora [322, 323]. Conversely, the same narrow spectrum of activity is a major limitation of phage therapy, as it restricts the use of a particular phage to a limited number of potential pathogens, and thus, requires more through diagnosis [322, 323].

A limitation of this study is that the number of complete prophage genomes reported within streptococcal genomes is likely to be an underestimate. This is because the vast majority of the genomes in the study dataset were unfinished

genomes, meaning they were comprised of several assembly contigs. Consequently, any prophage genome split over different contigs would not be identified using the current methodology, as the presence of multiple prophages in the same genome meant that accurate prophage assemblies could not be ensured. However, genome sequencing and assembly technologies are consistently improving, and more recently generated genome assemblies are comprised of only a few contigs [324], which will circumvent this limitation in the future.

Overall, the results from this study enhance our understanding of how prophages drive evolution of streptococci and reveal numerous areas for future studies. A summary of the main findings of this thesis, together with suggestions for future research, is provided in the next chapter.

5.6 Supplementary information

Supplementary File 5.1. List of genomes used in this study.

Supplementary File 5.2. A list of all the identified prophage insertion sites and their flanking genes.

Supplementary File 5.3. A phylogenetic tree of all prophage genomes identified in the dataset.

6. Summary and future work

6.1 Summary

The pneumococcus is a major public health problem, leading to significant morbidity and mortality worldwide with a substantial health and economic burden [1, 3-9]. The increase in antibiotic-resistant pneumococci is of serious concern, which necessitates investigations into novel antimicrobial strategies. Despite their potential as therapeutic weapons for bacterial diseases, our knowledge of phages and bacteriocins has been heavily bottlenecked by technical difficulties in their detection and characterisation. Before the advent of affordable genome sequencing, most bacteriocins and phages were identified using experimental culture-based methods; however, the explosion of whole genome sequence data has drastically expedited the identification and genetic analysis of phages and bacteriocins. The data presented in this thesis demonstrated how genomic approaches can be employed to enhance our knowledge of the prevalence, diversity and molecular epidemiology of bacteriocins and phages.

The principal findings of the Results chapters of this thesis are summarised below:

Chapter 3. Genome sequencing reveals a large and diverse repertoire of antimicrobial peptides

6.1.1 The pneumococcus possesses an unexpectedly large bacteriocin repertoire

A bioinformatics investigation of >6,200 pneumococcal genomes resulted in the discovery of 14 novel and different bacteriocin clusters. This was a substantial increase on the previously identified pneumococcal bacteriocin clusters, which tripled the number of known bacteriocins in this species. Among the newly-

discovered bacteriocin clusters, several belong to three distinct bacteriocin families of lactococcin 972-like, lassopeptide and sactipeptide, which were previously not known to be harboured by pneumococci. RNA sequencing clearly demonstrated that bacteriocin genes were, at the very least, transcriptionally active when the pneumococcus was under stress or in competition with another strain during bacterial co-culture. These findings fundamentally expand our knowledge of the bacteriocin repertoire of the pneumococcus, which is central to understanding how bacteriocins drive changes within the pneumococcal population and the wider microbial community. There is obvious potential for the development of bacteriocins as novel antimicrobials, and at a time when the challenges of antimicrobial-resistant microbes have never been more acute these data provide many new areas of investigation.

6.1.2 There is extraordinary bacteriocin diversity within the global pneumococcal population

Investigation of the identified bacteriocins within a globally-distributed pneumococcal dataset revealed that some bacteriocins were ubiquitous among all genomes, while others were only found in specific clonal complexes. Overall, each pneumococcal genome possessed between 6 to 11 different bacteriocins. The fact that any single pneumococcal genome possesses multiple bacteriocin clusters should be carefully considered when designing laboratory experiments aimed at assessing the activity of an individual bacteriocin.

Bacteriocins, often in particular combinations, were associated with specific clonal complexes. This is important, since these findings provide a framework on which to examine the specific contribution bacteriocins might make to the success of

genetic lineages, and thus in shaping the bacterial community. It is plausible that harbouring particular bacteriocin combinations promote the dominance of certain strains. Given that certain pneumococcal lineages and serotypes are predominately associated with disease whilst others with asymptomatic nasopharyngeal carriage [41], it would be clinically relevant to investigate any relationships between the production of bacteriocins and the epidemiological success of bacterial genetic lineages. Moreover, an improved understanding of the relevance of the bacteriocin production in pneumococcal biology may help guide the development of appropriate interventions, such as predicting vaccine impact. If bacteriocins are essential to the microbial competitive strategy, by directly influencing changes in microbial populations, they might indirectly be affecting the effectiveness of vaccines in the longer term: after vaccine use in human populations, the target bacterial population is significantly changed and those bacteria must now compete within the altered microbiome.

6.1.3 Bacteriocin clusters missing one or more genes are not uncommon

The percentages of partial and complete clusters varied between different bacteriocins, *i.e.* some were partial in the majority of genomes, while others were present in most genomes as a complete cluster, which adds further layers of complexity when trying to elucidate the roles bacteriocins play in streptococcal biology. Furthermore, the remnant genes in the partial bacteriocin clusters were highly conserved for at least several decades. RNA sequencing also revealed that partial clusters were transcriptionally active. Taken together, these results suggest that the remnant genes in the partial bacteriocin clusters likely play an important biological function.

Streptococcins were used as models for further investigating the pattern of missing genes in partial clusters, which revealed that the majority of the partial clusters lacked the bacteriocin gene, while still retaining the immunity and/or the transporter genes. These findings provide further support for the “cheater” hypothesis, which suggests that some strains retain the immunity and transporter genes in order to protect themselves from neighbouring bacteria that express the bacteriocin, without bearing the cost of toxin production [252, 281].

6.1.4 There is evidence for the intra- and interspecies exchange of bacteriocin clusters

The majority of the bacteriocin clusters identified among pneumococcal genomes were not exclusive to *S. pneumoniae* but were also present in other unrelated streptococcal species. The interspecies exchange of bacteriocin clusters was further evidenced by the differences in the GC content of the bacteriocin clusters relative to the pneumococcal genome. Moreover, this study identified genomic hotspots for the integration of bacteriocin gene clusters in the pneumococcal chromosome. Up to three different bacteriocin clusters were found to be associated with a single hotspot, suggesting a switching mechanism whereby different bacteriocin clusters may replace one another through genetic recombination. Taken together, these findings suggest that the intra- and interspecies exchange of bacteriocin clusters can occur, which fundamentally expands our understanding of bacteriocins relative to intraspecies and interspecies nasopharyngeal competition.

These results are also significant to the broader community as one promising approach for treatment and prevention of pathogenic bacterial species is the use of commensal bacteriocin-producing species as probiotics to outcompete the

pathogenic ones [160]. The findings that the bacteriocin gene clusters can be exchanged among streptococcal strains may have implication for the future of this approach. While such exchange events have not yet been experimentally confirmed, nor is it yet clear how prevalent or efficient they would be, it is possible to envision a scenario wherein the bacteriocin gene cluster of the probiotic strain could be integrated into the genomes of other bacterial species, including the target pathogen, which would cause the bacteriocin to become ineffective.

Chapter 4. Genomic investigation and molecular epidemiology of satellite prophages in S. pneumoniae

6.1.5 Satellite prophages are widespread among pneumococcal genomes and possess a structured population

A bioinformatic investigation of 482 historical and modern pneumococcal genomes resulted in the discovery of 44 novel and unique satellite prophages, which clustered into five major groups. Satellite prophages were found to have persistent associations with specific widely-circulating pneumococcal clonal complexes over many decades. These findings are important as they provide a framework on which to explore whether the presence of certain prophages among genomes might contribute to the virulence or the epidemiological success of a bacterial genetic lineage.

6.1.6 The pneumococcal satellite prophages are maintained at specific sites on the bacterial chromosome

The pneumococcal satellite prophages were consistently inserted in seven specific locations of the host chromosome, which were directly associated with

the nucleotide sequence of the integrase gene they harboured. The majority (88%) of the identified satellite prophages were found in three particular locations among the pneumococcal genomes, suggesting that the presence of a suitable region(s) for integration in the host genome may act as a major determining factor in phage sensitivity.

6.1.7 Satellite prophages may play a role in pneumococcal pathogenesis

An extensive investigation of pneumococcal satellite prophage genes led to the identification of a virulence-associated gene, which was found to be involved in pneumococcal pneumonia and sepsis in a murine infection model, demonstrating experimentally for the first time that satellite prophages were directly involved in pneumococcal virulence. Moreover, a comparative transcriptome analysis of planktonic and biofilm-grown pneumococci clearly showed significantly higher satellite prophage gene expression in planktonic growth compared to the biofilm. *vapE* was the third most upregulated gene in the entire genome, further indicating a role for the satellite prophage and *vapE* in pneumococcal virulence. While the specific mechanism driving virulence is not yet clear, these data provide clear evidence that experimental investigations of pneumococcal satellite prophages should be pursued, as such findings may hold the key to understanding important aspects of pneumococcal biology and pathogenesis.

Chapter 5. Prophages and satellite prophages are widespread among Streptococcus species

6.1.8 Prophages are a major component of genomes of most clinically relevant streptococcal species and a significant driving force for bacterial strain diversification

A bioinformatic investigation of >1,300 genomes of 70 different *Streptococcus* species resulted in the identification of nearly 800 prophages and satellite prophages, the majority of which were newly discovered. The prophage content within each streptococcal genome was calculated, which revealed significant variability among streptococcal species, ranging from 0.4% of genes in *S. thermophilus* to 9.5% in *S. pyogenes*. Many of the streptococci investigated are important human and animal pathogens and prophages constituted a significant portion of their genomes. Overall, these findings suggest that prophages are likely to be influencing streptococcal biology and epidemiology to a much greater extent than previously appreciated, given the high proportion of prophage DNA present in many streptococcal species – many of which had not previously been analysed for evidence of prophages. Furthermore, significant differences in prophage content were detected among different genomes of the same streptococcal species, highlighting the important role of prophages in contributing to interstrain genetic variability.

6.1.9 Full-length and satellite prophages are separate entities with little effective genetic exchange between them

An unrooted phylogenetic tree of all streptococcal prophage genomes depicted full-length and satellite prophages as two clearly distinct groups. Prophages and satellite prophages never shared an insertion site. Furthermore, 93% of all satellite prophage genes were not found in any full-length prophage. Satellite prophage genomes were found to be highly conserved among pneumococcal

genomes over many decades, signifying that they are under strong positive evolutionary pressure, possibly due to an important biological function. Taken together, these findings indicate that satellite prophage sequences are not recent remnants of previous lysogenisation by full-length prophages, but rather that they belong to a unique family of MGE.

6.1.10 There is convincing evidence that cross-species transmission of prophages is not uncommon

Whole genome comparisons of all the identified prophage sequences revealed several major and minor clusters for both full-length and satellite prophages. Surprisingly, streptococcal prophages isolated from different species were often in the same cluster, suggesting that cross-species transmission occurs more commonly than previously realised. Intriguingly, despite the close relatedness of their prophages, the hosts were not necessarily the closest phylogenetically related species. A speculative hypothesis might be that streptococcal prophages are evolving separately from their bacterial hosts, and thus, different factors such as ecological relatedness may dominate over evolutionary relatedness of the host bacteria. The overall implications of these findings are that host specificity should be taken into account when aiming to elucidate the exact role of prophages in streptococcal biology and when evaluating the potential of phages for therapeutic purposes.

6.1.11 Streptococcal prophages are frequently inserted among genes involved in information storage and processing

The location of prophages residing along the chromosome of the *Streptococcus* species varies greatly between different species. Genes flanking prophage insertion sites are more frequently involved in information storage and processing compared to the average for streptococcal genomes. One possible explanation might be that there is an evolutionary pressure on selecting prophage integrase genes that insert in-between essential host genes in order to enhance prophage stability, as prophage genomes integrated within essential genes are less likely to be lost during host replication.

6.2 Future work

The genome datasets used in this thesis were specifically compiled to address questions related to prevalence and diversity. Further research, using a more focused dataset, should be undertaken to investigate how bacteriocins and phages drive changes within the pneumococcal population and the wider microbial community. Genome datasets comprised of sequential samples isolated from a particular geographical environment over years could be used to track bacteriocin and prophage stability in lineages/species and contextualise their movement across species. Furthermore, the use of a more focused genome dataset would enable addressing questions such as how frequent processes such as bacteriocin cluster switching is taking place in nature. It would also be interesting to investigate whether the acquisition of a new bacteriocin cluster within a specific bacterial lineage, would lead to another

existing bacteriocin cluster becoming redundant, and thus accumulate detrimental mutations, particularly in their toxin gene.

As discussed in the general introduction (see 1.4.4), the long abiding race between bacteria and phage has led to the evolution of resistance systems that protect the host bacteria from infection by phages. Recently, several new phage defense systems have been described, and some of these are reported to be present among streptococcal genomes [325]. A further study could assess whether the presence of any of these newly-discovered phage defense systems are mutually exclusive with particular phages. These investigations have the potential to greatly improve our understanding of the complex interplay between the phages and the host.

Classically, individual bacteriocins are named after the species of organism that produces them; for example, colicin and pesticin bacteriocins are produced by *Escherichia coli* [326] and *Yersinia pestis* [327], respectively. Likewise, phages are generally named after the bacterial species and strain for which they are supposedly specific; for instance, Dp1 (*diplococcus pneumoniae*; the pneumococcus), PA DP4 (*Pseudomonas aeruginosa*), KP DP1 (*Klebsiella pneumoniae*), SA DP1 (S*. aureus*) and EC DP3 (*E. coli*) [328]. However, the findings presented in this thesis revealed that identical or highly similar versions of the bacteriocin gene clusters and phages could be found in other unrelated species, highlighting a need for new naming schemes. The lack of an adequate systematic scheme for naming bacteriocins and phages can result in confusion and potentially results in duplication of some of the *in vitro* work. For instance, as shown in Chapter 3, a similar bacteriocin cluster has been named differently in two

different bacteria (pneumolancidin in pneumococcus [256], and Salivaricin E in *Streptococcus salivarius* [278]). Based on the findings of this thesis, it is recommended that, in future investigations, newly-identified bacteriocins and phages are named based on their nucleotide sequence rather than the species that they target or were isolated from.

There is potential for the use of bacteriocins and phages as an alternative to existing antibiotics and the lengthy list of phages and bacteriocins identified as part of this thesis provides a valuable resource for future studies. A natural progression of this work is to investigate the *in vivo* production and functional activities of bacteriocin and phage gene products in the laboratory settings. For instance, susceptibility assays [329] need to be carried out in order to evaluate the spectrum of activity and potency of the identified bacteriocins and phages. In fact, several of the bacteriocins identified as part of this thesis are currently in the process of being synthesised and tested in collaboration with an industrial partner.

The lengthy list of prophages and bacteriocins identified as part of this thesis also provide a useful resource for the development of the novel bioinformatics tools aimed at *in silico* identification of bacteriocins and prophages. The coding sequences from the newly identified bacteriocin and phages ORFs can be used as query to easily perform genome mining in additional genome datasets. Furthermore, results from this thesis revealed that the prophages and bacteriocins are commonly located at particular location in the genome, and this information can be used to develop pipelines that can identify the majority of these elements by simply scanning known insertion sites.

Overall, the findings from this thesis clearly highlight the importance of performing population genomics studies for understanding bacterial infection. The heterogeneity observed amongst pneumococcal populations in terms of bacteriocins and prophages was astonishing, which cautions against making broader conclusions about pneumococcal biology based on results obtained from the use of a single bacterial strain.

The rapidly increasing number of available whole genome sequences presents unprecedented opportunities to enhance our understanding of microbial competition and infection at a depth and scale not previously feasible. As evidenced throughout this thesis, large population genomics approaches can be employed to develop hypotheses, design experiments and choose the most suitable strains for further *in vitro* experiments. The results of this thesis uncover many areas for future studies, the outcomes of which will enhance our understanding of the potential roles bacteriocins and phages play in streptococcal biology, pathogenesis, ecology, epidemiology and evolution.

7. References

1. Hardie JM and Whiley RA: **The genus *Streptococcus*.** Williams & Wilkins Co, Baltimore 1995:55-124.

2. Euzéby JP. **List of bacterial names with standing in nomenclature: a folder available on the internet.** *Int. J. Syst. Evol. Microbiol.* 1997, **47**(2):590-592.

3. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T *et al*: **Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates.** *Lancet* 2009, **9693**(374):893-902.

4. Krzyściak W, Pluskwa K, Jurczak A and Kościelniak D: **The pathogenicity of the *Streptococcus* genus.** *Eur. J. Clin. Microbiol. Infect. Dis.* 2013, **32**(11):1361-1376.

5. Carapetis JR, Steer AC, Mulholland EK and Weber M: **The global burden of group A streptococcal diseases.** *Lancet Infect Dis* 2005, **5**(11):685-694.

6. Vornhagen J, Adams Waldorf, Kristina M and Rajagopal L: **Perinatal group B streptococcal infections: virulence factors, immunity, and prevention strategies.** *Trends Microbiol* 2017, **25**(11):919-931.

7. Carroll RK, Beres SB, Sitkiewicz I, Peterson L, Matsunami RK, Engler DA, Flores AR, Sumby P and Musser JM: **Evolution of diversity in epidemics revealed by analysis of the human bacterial pathogen group A *Streptococcus*.** *Epidemics* 2011, **3**(3-4):159-170.

8. Evans JJ, Bohnsack JF, Klesius PH, Whiting AA, Garcia JC, Shoemaker CA and Takahashi S: **Phylogenetic relationships among *Streptococcus agalactiae* isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan.** *J Med Microbiol* 2008, **57**(11):1369-1376.

9. Low DE, Nakashima K, Vuopio-Varkila J, Salmenlinna S, Dou S, Musser JM, Naidich S, Grigsby D, Pan X, Liu M *et al*: **Rapid selection of complement-inhibiting protein variants in group A *Streptococcus* epidemic waves.** *Nat Med* 1999, **5**(8):924-929.

10. Facklam R. **What happened to the streptococci: overview of taxonomic and nomenclature changes.** *Clin Microbiol Rev* 2002, **15**(4):613-630.

11. Dickson RP, Erb-Downward JR, Martinez FJ and Huffnagle GB: **The microbiome and the respiratory tract.** *Annu Rev Physiol* 2016, **78**(1):481-504.

12. Shak JR, Vidal JE and Klugman KP: **Influence of bacterial interactions on pneumococcal colonization of the nasopharynx.** *Trends Microbiol* 2012, **21**(3):129-135.

13. Dawid S, Roche AM and Weiser JN: **The *blp* bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both *in vitro* and *in vivo*.** *Infect Immun* 2007, **75**(1):443-451.

14. Tong H, Chen W, Merritt J, Qi F, Shi W and Dong X: ***Streptococcus oligofermentans*** **inhibits** ***Streptococcus mutans*** **through conversion of lactic acid into inhibitory H_2O_2: a possible counteroffensive strategy for interspecies competition.** *Mol Microbiol* 2007, **63**(3):872-880.

15. Marri PR, Hao W and Golding BG: **Gene gain and gene loss in *Streptococcus*: is it driven by habitat?** *Mol Biol Evol* 2006, **23**(12):2379-2391.

16. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR *et al*: **Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species.** *Genome Biol* 2010, **11**(10):R107.

17. Sanguinetti L, Toti S, Reguzzi V, Bagnoli F and Donati C: **A novel computational method identifies intra- and inter-species recombination events in *Staphylococcus aureus* and *Streptococcus pneumoniae*.** *PLoS Comput Biol* 2012, **8**(9):e1002668.

18. Land M, Hauser L, Jun S, Nookaew I, Leuze MR, Ahn T, Karpinets T, Lund O, Kora G, Wassenaar T *et al*: **Insights from 20 years of bacterial genome sequencing.** *Funct Integr Genomics* 2015, **15**(2):141-161.

19. Watson DA, Musher DM, Jacobson JW and Verhoef J: **A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery.** *Clin Infect Dis* 1993, **17**(5):913-924.

20. Avery OT, Dochez AR, Cole RI and Chickering HT: **Acute lobar pneumonia; prevention and serum treatment.** Monographs of The Rockefeller Institute for Medical Research, New York 1917.

21. Griffith F. **The significance of pneumococcal types.** *J Hyg (Lond)* 1928, **27**(2):113-159.

22. Avery OT, Macleod CM and 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.

23. Tillett WS, Cambier MJ and McCormack JE: **The treatment of lobar pneumonia and pneumococcal empyema with penicillin.** *Bull N Y Acad Med* 1944, **20**(3):142.

24. Howden R. **Use of anaerobic culture for the improved isolation of *Streptococcus pneumoniae*.** *J Clin Pathol* 1976, **29**(1):50-53.

25. Tuomanen EI, Mitchell TJ, Morrison DA and Spratt BG: **The pneumococcus.** ASM press, Washington, D.C 2004.

26. James A. Kellogg, David A. Bankert, Carol J. Elder, Joanne L. Gibbs and Marie C. Smith: **Identification of *Streptococcus pneumoniae* revisited.** *J Clin Microbiol* 2001, **39**(9):3373-3375.

27. Kontiainen S and Sivonen A: **Optochin resistance in *Streptococcus pneumoniae* strains isolated from blood and middle ear fluid.** *Eur. J. Clin. Microbiol. Infect. Dis.* 1987, **6**(4):422.

28. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA *et al*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci U S A* 1998, **95**(6):3140-3145.

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

30. Spratt BG, Hanage WP, Li B, Aanensen DM and Feil EJ: **Displaying the relatedness among isolates of bacterial species – the eBURST approach.** *FEMS Microbiol Lett* 2004, **241**(2):129-134.

31. Francisco AP, Bugalho M, Ramirez M and Carriço JA: **Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach.** *BMC Bioinformatics* 2009, **10**(1):152.

32. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA and Vaz C: **PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods.** *Bioinformatics* 2017, **33**(1):128-129.

33. Feil EJ, Li BC, Aanensen DM, Hanage WP and Spratt BG: **eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data.** *J Bacteriol* 2004, **186**(5):1518-1530.

34. Infante AJ, McCullers JA and Orihuela CJ: **Chapter 19 - mechanisms of predisposition to pneumonia: infants, the elderly, and viral infections.** In: Brown J, Hammerschmidt S, Orihuela C, **Mechanisms of Predisposition to Pneumonia**, Academic Press, Amsterdam 2016:363-382.

35. Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H and Sørensen UBS: **Evolution of *Streptococcus pneumoniae* and its close commensal relatives.** *PLoS One* 2008, **3**(7):e2683.

36. Kilian M, Riley DR, Jensen A, Brüggemann H and Tettelin H: **Parallel evolution of *Streptococcus pneumoniae* and *Streptococcus mitis* to pathogenic and mutualistic lifestyles.** *mBio* 2014, **5**(4):e01490.

37. Gilley RP and Orihuela CJ: **Pneumococci in biofilms are non-invasive: implications on nasopharyngeal colonization.** *Front Cell Infect Microbiol* 2014, **4**:163.

38. Weiser JN, Ferreira DM and Paton JC: ***Streptococcus pneumoniae*: transmission, colonization and invasion.** *Nat Rev Microbiol* 2018, **16**(6):355-367.

39. Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, Moxon ER, Crook DW and Peto TEA: **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.

40. Drijkoningen JJC and Rohde GGU: **Pneumococcal infection in adults: burden of disease.** *Clin Microbiol Infect* 2014, **20**(s5):45-51.

41. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW and 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.

42. Darboe MK, Fulford AJ, Secka O and Prentice AM: **The dynamics of nasopharyngeal *Streptococcus pneumoniae* carriage among rural Gambian mother-infant pairs.** *BMC Infect Dis* 2010, **10**(1):195.

43. Granat S, Mia Z, Ollgren J, Herva E, Das M, Piirainen L, Auranen K and Mäkelä P: **Longitudinal study on pneumococcal carriage during the first year of life in Bangladesh.** *Pediatr Infect Dis J* 2007, **26**(4):319-324.

44. Hill PC, Akisanya A, Sankareh K, Cheung YB, Saaka M, Lahai G, Greenwood BM and Adegbola RA: **Nasopharyngeal carriage of *Streptococcus pneumoniae* in Gambian villagers.** *Clin Infect Dis* 2006, **43**(6):673-679.

45. Gray BM, Converse GM and Dillon HC: **Epidemiologics studies of *Streptococcus pneumoniae* in infants: acquisition, carriage, and infection during the first 24 months of life.** *J Infect Dis* 1980, **142**(6):923-933.

46. Mosser JF, Grant LR, Millar EV, Weatherholtz RC, Jackson DM, Beall B, Craig MJ, Reid R, Santosham M and O'Brien KL: **Nasopharyngeal carriage and transmission of *Streptococcus pneumoniae* in American Indian households after a decade of pneumococcal conjugate vaccine use.** *PLoS One* 2014, **9**(1):e79578.

47. Rosenaum MJ, Felton LD and Atwater RM: **An epidemiologic study of pneumonia and its mode of spread.** *Am J Hyg* 1926, **6**(3):463.

48. Krone CL, van de Groep K, Trzciński K, Sanders E and Bogaert D: **Immunosenescence and pneumococcal disease: an imbalance in host–pathogen interactions.** *Lancet Respir Med* 2014, **2**(2):141-153.

49. Esposito S, Mari D, Bergamaschini L, Orenti A, Terranova L, Ruggiero L, Ierardi V, Gambino M, Croce FD and Principi N: **Pneumococcal colonization in older adults.** *Immun Ageing* 2016, **13**(1):2.

50. Almeida ST, Nunes S, Santos Paulo AC, Valadares I, Martins S, Breia F, Brito-Avô A, Morais A, de Lencastre H and Sá-Leão R: **Low prevalence of pneumococcal carriage and high serotype and genotype diversity among adults over 60 years of age living in Portugal.** *PLoS One* 2014, **9**(3):e90974.

51. Jomrich N, Kellner S, Djukic M, Eiffert H and Nau R: **Absence of *Streptococcus pneumoniae* in pharyngeal swabs of geriatric inpatients.** *J Infect Dis* 2015, **47**(7):504-509.

52. Krone CL, van de Groep K, Trzciński K, Sanders EA and Bogaert D: **Immunosenescence and pneumococcal disease: an imbalance in host–pathogen interactions.** *Lancet Respir Med* 2014, **2**(2):141-153.

53. Troeger C, Blacker B, Khalil IA, Rao PC, Cao J, Zimsen SRM, Albertson SB, Deshpande A, Farag T, Abebe Z *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.

54. Lloyd-Evans N, O'Dempsey TJ, Baldeh I, Secka O, Demba E, Todd JE, Mcardle TF, Banya WS and Greenwood B: **Nasopharyngeal carriage of pneumococci in Gambian children and in their families.** *Pediatr Infect Dis J* 1996, **15**(10):866-871.

55. Usuf E, Bottomley C, Adegbola RA and Hall A: **Pneumococcal carriage in sub-saharan Africa - a systematic review.** *PLoS One* 2014, **9**(1):e85001.

56. Nisar MI, Nayani K, Akhund T, Riaz A, Irfan O, Shakoor S, Muneer S, Muslim S, Hotwani A, Kabir F *et al*: **Nasopharyngeal carriage of *Streptococcus pneumoniae* in children under 5 years of age before introduction of pneumococcal vaccine (PCV10) in urban and rural districts in Pakistan.** *BMC Infect Dis* 2019, **19**(1):114.

57. Lankinen KS, Leinonen M, Tupasi TE, Haikala R and Ruutu P: **Pneumococci in nasopharyngeal samples from Filipino children with acute respiratory infections.** *J Clin Microbiol* 1994, **32**(12):2948-2952.

58. Millar E, O'Brien K, Zell E, Bronsdon M, Reid R and Santosham M: **Nasopharyngeal carriage of *Streptococcus pneumoniae* in Navajo and White Mountain Apache children before the introduction of pneumococcal conjugate vaccine.** *Pediatr Infect Dis J* 2009, **28**(8):711-716.

59. O'Brien KL, Shaw J, Weatherholtz R, Reid R, Watt J, Croll J, Dagan R, Parkinson AJ and Santosham M: **Epidemiology of invasive *Streptococcus pneumoniae* among Navajo children in the era before use of conjugate pneumococcal vaccines, 1989-1996.** *Am J Epidemiol* 2004, **160**(3):270-278.

60. Aniansson G, Alm B, Andersson B, Larsson P, Nylén O, Peterson H, Rignér P, Svanborg M and Svanborg C: **Nasopharyngeal colonization during the first year of life.** *J Infect Dis* 1992, **165**(Supplement 1):S38-S42.

61. Labout JAM, Duijts L, Arends LR, Jaddoe VWV, Hofman A and De Groot R: **Factors associated with pneumococcal carriage in healthy Dutch infants: the generation R study.** *J Pediatr* 2008, **153**(6):77-776.e1.

62. Lipsitch M, Abdullahi O, D'Amour A, Xie W, Weinberger DM, Tchetgen ET and Scott JAG: **Estimating rates of carriage acquisition and clearance and competitive ability for pneumococcal serotypes in Kenya with a Markov transition model.** *Epidemiology* 2012, **23**(4):510-519.

63. Turner P, Turner C, Jankhot A, Helen N, Lee SJ, Day NP, White NJ, Nosten F and Goldblatt D: **A longitudinal study of *Streptococcus pneumoniae* carriage in a cohort of infants and their mothers on the Thailand-Myanmar border.** *PLoS One* 2012, **7**(5):e38271.

64. Turner P, Hinds J, Turner C, Jankhot A, Gould K, Bentley SD, Nosten F and Goldblatt D: **Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes by use of latex agglutination or molecular serotyping by microarray.** *J Clin Microbiol* 2011, **49**(5):1784-1789.

65. Gratten M, Montgomery J, Gerega G, Gratten H, Siwi H, Poli A and Koki G: **Multiple colonization of the upper respiratory tract of Papua New Guinea children with *Haemophilus influenzae* and *Streptococcus pneumoniae*.** *Southeast Asian J Trop Med Public Health* 1989, **20**(4):501.

66. Pericone CD, Overweg K, Hermans PWM and Weiser JN: **Inhibitory and bactericidal effects of hydrogen peroxide production by *Streptococcus pneumoniae* on other inhabitants of the upper respiratory tract.** *Infect Immun* 2000, **68**(7):3990-3997.

67. Regev-Yochay G, Trzciński K, Thompson CM, Malley R and Lipsitch M: **Interference between *Streptococcus pneumoniae* and *Staphylococcus aureus*: *in vitro* hydrogen peroxide-mediated killing by *Streptococcus pneumoniae*.** *J Bacteriol* 2006, **188**(13):4996-5001.

68. Dunne EM, Smith-Vaughan HC, Robins-Browne RM, Mulholland EK and Satzke C: **Nasopharyngeal microbial interactions in the era of pneumococcal conjugate vaccination.** *Vaccine* 2013, **31**(19):2333-2342.

69. Shakhnovich EA, King SJ and Weiser JN: **Neuraminidase expressed by *Streptococcus pneumoniae* desialylates the lipopolysaccharide of *Neisseria meningitidis* and *Haemophilus influenzae*: a paradigm for interbacterial competition among pathogens of the human respiratory tract.** *Infect Immun* 2002, **70**(12):7161-7164.

70. Lysenko ES, Ratner AJ, Nelson AL and Weiser JN: **The role of innate immune responses in the outcome of interspecies competition for colonization of mucosal surfaces.** *PLoS Pathog* 2005, **1**(1):e1.

71. Pettigrew MM, Gent JF, Revai K, Patel JA and Chonmaitree T: **Microbial interactions during upper respiratory tract infections.** *Emerg Infect Dis* 2008, **14**(10):1584-1591.

72. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T *et al*: **Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates.** *Lancet* 2009, **374**:893-902.

73. van de Beek D, de Gans J, Tunkel AR and Wijdicks EFM: **Community-acquired bacterial meningitis in adults.** *N Engl J Med* 2006, **354**(1):44-53.

74. Weisfelt M, van de Beek D, Spanjaard L, Reitsma JB and de Gans J: **Clinical features, complications, and outcome in adults with pneumococcal meningitis: a prospective case series.** *Lancet Neurol* 2006, **5**(2):123-129.

75. Huang S, Johnson K, Ray G, Wroe P, Lieu T, Zell E, Linder J, Grijalva C, Metlay J and Finkelstein J: **Healthcare utilization and cost of pneumococcal disease in the United States.** *Vaccine* 2011,:3398-412.

76. Weycker D, Strutton D, Edelsberg J, Sato R and Jackson LA: **Clinical and economic burden of pneumococcal disease in older US adults.** *Vaccine* 2010, **28**(31):4955-4960.

77. Melegaro A, Edmunds WJ, Pebody R, Miller E and George R: **The current burden of pneumococcal disease in England and Wales.** *J Infect* 2006, **52**(1):37-48.

78. Wroe PC, Finkelstein JA, Ray GT, Linder JA, Johnson KM, Rifas-Shiman S, Moore MR and Huang SS: **Aging population and future burden of pneumococcal pneumonia in the United States.** *J Infect Dis* 2012, **205**(10):1589-1592.

79. Sternberg GM. **Induced septicæmia in the rabbit.** *Am J Med Sci* 1882, **84**(167):69-76.

80. Wright AE, Parry Morgan W, Colebrook L and Dodgson RW: **Observations on prophylactic inoculation against pneumococcus infections and on the results which have been achieved by it.** *Lancet* 1914, **183**(4715):87-95.

81. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen HB and Nahm MH: **Pneumococcal capsules and their types: past, present, and future.** *Clin Microbiol Rev* 2015, **28**(3):871-899.

82. MacLeod CM and Hodges RG: **Prevention of pneumococcal pneumonia by immunization with specific capsular polysaccharides.** *J Exp Med* 1945, **82**:445.

83. Gibbons JH and Lashof JC: **A review of selected Federal vaccine and immunization policies: based on case studies of pneumococcal vaccine.** Government Printing Office, Washington, DC 1979.

84. Shapiro ED and Clemens JD: **A controlled evaluation of the protective efficacy of pneumococcal vaccine for patients at high risk of serious pneumococcal infections.** *Ann Intern Med* 1984, **101**(3):325.

85. Bolan G, Broome CV, Facklam RR, Plikaytis BD, Fraser DW and Schlech WF: **Pneumococcal vaccine efficacy in selected populations in the United States.** *Ann Intern Med* 1986, **104**(1):1.

86. Mäkelä PH, Leinonen M, Pukander J and Karma P: **A study of the pneumococcal vaccine in prevention of clinically acute atttacks of recurrent otitis media.** *Rev Infect Dis* 1981, **3 Suppl**:S124-S132.

87. Sloyer JLJ, Ploussard JH and Howie VM: **Efficacy of pneumococcal polysaccharide vaccine in preventing acute otitis media in infants in Huntsville, Alabama.** *Rev Infect Dis* 1981, **3**:S119-S123.

88. Teele DW, Klein JO, Bratton L, Fisch GR, Mathieu OR, Porter PJ, Starobin SG, Tarlin LD and Younes RP: **Use of pneumococcal vaccine for prevention of recurrent acute otitis media in infants in Boston.** *Rev Infect Dis* 1981, **3**:S113-S118.

89. Douglas RM, Paton JC, Duncan SJ and Hansman DJ: **Antibody response to pneumococcal vaccination in children younger than five years of age.** *J Infect Dis* 1983, **148**(1):131-137.

90. Sell SH, Wright PF, Vaughn WK, Thompson J and Schiffman G: **Clinical studies of pneumococcal vaccines in infants. I. reactogenicity and immunogenicity of two polyvalent polysaccharide vaccines.** *Rev Infect Dis* 1981, **3**(Supplement_1):S97-S107.

91. Nuorti PJ, Butler JC and Breiman RF: **Prevention of pneumococcal disease: recommendations of the advisory committee on immunization practices (ACIP)**. *MMWR Recomm Rep* 1997:24.

92. Avery OT and Goebel WF: **Chemo-immunological studies on conjugated carbohydrate-proteins: II. immunological specificity of synthetic sugar-protein antigens.** *J Exp Med* 1929, **50**(4):533-550.

93. Goebel WF and Avery OT: **Chemo-immunological studies on conjugated carbohydrate-proteins. I. the synthesis of p-aminophenol b-glucoside, p-aminophenol b-galactoside, and their coupling with serum globulin.** *J Exp Med* 1938, **68**(4):640.

94. Adams WG, Deaver KA, Cochi SL, Plikaytis BD, Zell ER, Broome CV, Wenger JD, Stephens DS, Farley MM, Harvey C *et al*: **Decline of childhood *Haemophilus influenzae type b* (Hib) disease in the Hib vaccine era.** *JAMA* 1993, **269**(2):221-226.

95. Black S, Shinefield H, Fireman B, Lewis E, Ray P, Hansen JR, Elvin L, Ensor KM, Hackell J, Siber G *et al*: **Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. Northern California Kaiser Permanente Vaccine Study Center Group.** *Pediatr Infect Dis J* 2000, **19**(3):187-195.

96. World Health Organization. **Pneumococcal conjugate vaccines in infants and children under 5 years of age: WHO position paper – February 2019.** *Wkly Epidemiol Rec* 2019, **94**(8):85-103.

97. Temple B, Toan NT, Dai VTT, Bright K, Licciardi PV, Marimla RA, Nguyen CD, Uyen DY, Balloch A, Huu TN *et al*: **Immunogenicity and reactogenicity of ten-valent versus 13-valent pneumococcal conjugate vaccines among infants in Ho Chi Minh City, Vietnam: a randomised controlled trial.** *Lancet Infect Dis* 2019, **19**(5):497-509.

98. Feikin DR, Kagucia EW, Loo JD, Link-Gelles R, Puhan MA, Cherian T, Levine OS, Whitney CG, O'Brien KL, Moore MR *et al*: **Serotype-specific changes in invasive pneumococcal disease after pneumococcal conjugate vaccine introduction: a pooled analysis of multiple surveillance sites.** *PLoS Med* 2013, **10**(9):e1001517.

99. Brueggemann AB, Pai R, Crook DW and Beall B: **Vaccine escape recombinants emerge after pneumococcal vaccination in the United States.** *PLoS Pathog* 2007, **3**(11):e168.

100. Coffey TJ, Enright MC, Daniels M, Morona JK, Morona R, Hryniewicz W, Paton JC and Spratt BG: **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.

101. Moore MR, Gertz RE, Woodbury RL, Barkocy-Gallagher GA, Schaffner W, Lexau C, Gershman K, Reingold A, Farley M, Harrison LH *et al*: **Population snapshot of emergent *Streptococcus pneumoniae* serotype 19A in the United States, 2005.** *J Infect Dis* 2008, **197**(7):1016-1027.

102. Ladhani SN, Collins S, Djennad A, Sheppard CL, Borrow R, Fry NK, Andrews NJ, Miller E and Ramsay ME: **Rapid increase in non-vaccine serotypes causing invasive pneumococcal disease in England and Wales, 2000–17: a prospective national observational cohort study.** *Lancet Infect Dis* 2018, **18**(4):441-451.

103. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, Gottberg Av, Liñares J, R. Jacobs M, Kristinsson KG, Beall BW, Klugman KP *et al*: **Pneumococcal capsular switching: a historical perspective.** *J Infect Dis* 2013, **207**(3):439-449.

104. van Boeckel TP, Gandra S, Ashok A, Caudron Q, Grenfell BT, Levin SA and Laxminarayan R: **Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data.** *Lancet Infect Dis* 2014, **14**(8):742-750.

105. Tilghman RC and Finland M: **Clinical significance of bacteremia in pneumococcic pneumonia.** *Arch Intern Med* 1937, **59**(4):602-619.

106. Austrian R and Gold J: **Pneumococcal bacteremia with especial reference to bacteremic pneumococcal pneumonia.** *Ann Intern Med* 1964, **60**(5):759.

107. Kazanjian P. **Changing interest among physicians toward pneumococcal vaccination throughout the twentieth century.** *J Hist Med Allied Sci* 2004, **59**(4):555-587.

108. Hansman D and Bullen MM: **A resistant pneumococcus.** *Lancet* 1967, **290**(7509):264-265.

109. Chiou CC, Liu YC, Huang TS, Hwang WK, Wang JH, Lin HH, Yen MY and Hsieh KS: **Extremely high prevalence of nasopharyngeal carriage of penicillin-resistant *Streptococcus pneumoniae* among children in Kaohsiung, Taiwan.** *J Clin Microbiol* 1998, **36**(7):1933-1937.

110. Kim L, McGee L, Tomczyk S and Beall B: **Biological and epidemiological features of antibiotic-resistant *Streptococcus pneumoniae* in pre- and post-conjugate vaccine eras: a United States perspective.** *Clin Microbiol Rev* 2016, **29**(3):525-552.

111. Pérez JL, Linares J, Bosch J, López de Goicoechea, M J and Martín R: **Antibiotic resistance of *Streptococcus pneumoniae* in childhood carriers.** *J Antimicrob Chemother* 1987, **19**(2):278-280.

112. Hansman D and Morris S: **Pneumococcal carriage amongst children in Adelaide, South Australia.** *Epidemiol Infect* 1988, **101**(2):411-417.

113. Henderson FW, Gilligan PH, Wait K and Goff DA: **Nasopharyngeal carriage of antibiotic-resistant pneumococci by children in group day care.** *J Infect Dis* 1988, **157**(2):256-263.

114. Jacobs MR, Koornhof HJ, Robins-Browne RM, Stevenson CM, Vermaak ZA, Freiman I, Miller GB, Witcomb MA, Isaäcson M, Ward JI *et al*: **Emergence of multiply resistant pneumococci.** *N Engl J Med* 1978, **299**(14):735-740.

115. Sá-Leão R, Tomasz A, Sanches IS, Brito-Avô A, Vilhelmsson SE, Kristinsson KG and de Lencastre H: **Carriage of internationally spread clones of *Streptococcus pneumoniae* with unusual drug resistance patterns in**

children attending day care centers in Lisbon, Portugal. *J Infect Dis* 2000, **182**(4):1153-1160.

116. Samore MH, Magill MK, Alder SC, Severina E, Morrison-De Boer L, Lyon JL, Carroll K, Leary J, Stone MB, Bradford D *et al*: **High rates of multiple antibiotic resistance in *Streptococcus pneumoniae* from healthy children living in isolated rural communities: association with cephalosporin use and intrafamilial transmission.** *Pediatrics* 2001, **108**(4):856-865.

117. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, Hakenbeck R, Hryniewicz W, Lefévre JC, Tomasz A *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.

118. World Health Organization. **Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics.** *WHO. Globa* 2017.

119. Waxman DJ and Strominger JL: **Penicillin-binding proteins and the mechanism of action of beta-lactam antibiotics.** *Annu Rev Biochem* 1983, **52**:825-869.

120. Hakenbeck R, Grebe T, Zähner D and Stock JB: **β-Lactam resistance in *Streptococcus pneumoniae*: penicillin-binding proteins and non-penicillin-binding proteins.** *Mol Microbiol* 1999, **33**(4):673-678.

121. Coffey TJ, Dowson CG, Daniels M and Spratt BG: **Genetics and molecular biology of beta-lactam-resistant pneumococci.** *Microb Drug Resist* 1995, **1**(1):29-34.

122. Jensen A, Valdórsson O, Frimodt-Møller N, Hollingshead S and Kilian M: **Commensal streptococci serve as a reservoir for beta-lactam resistance genes in *Streptococcus pneumoniae*.** *Antimicrob Agents Chemother* 2015, **59**(6):3529-3540.

123. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Liñares J, Jacobs MR, Kristinsson KG, Beall BW, Klugman KP *et al*: **The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success.** *Genome Biol* 2012, **13**(11):R103.

124. Smith AM and 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.

125. Mascher T, Heintz M, Zähner D, Merai M and 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 β-Lactam resistance.** *J Bacteriol* 2006, **188**(5):1959-1968.

126. Tran TD, Kwon H, Kim E, Kim K, Briles DE, Pyo S and Rhee D: **Decrease in penicillin susceptibility due to heat shock protein ClpL in *Streptococcus pneumoniae*.** *Antimicrob Agents Chemother* 2011, **55**(6):2714-2728.

127. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). **Surveillance of antimicrobial resistance in Europe 2017.** *European Centre for Disease Prevention and Control; Stockholm, Sweden* 2018,.

128. Jones RN, Sader HS, Mendes RE and Flamm RK: **Update on antimicrobial susceptibility trends among *Streptococcus pneumoniae* in the United States: report of ceftaroline activity from the SENTRY Antimicrobial Surveillance Program (1998–2011).** *Diagn Microbiol Infect Dis* 2013, **75**(1):107-109.

129. Song J. **Advances in pneumococcal antibiotic resistance.** *Expert Rev Respir Med* 2013, **7**(5):491-498.

130. Song J, Jung S, Soo Ko K, Kim NY, Son JS, Chang H, Ki HK, Oh WS, Suh JY, Peck KR *et al*: **High prevalence of antimicrobial resistance among clinical *Streptococcus pneumoniae* isolates in Asia (an ANSORP study).** *Antimicrob Agents Chemother* 2004, **48**(6):2101-2107.

131. Emgård M, Msuya SE, Nyombi BM, Mosha D, Mosha V, Gonzales-Siles L, Nordén R, Geravandi S, Blomqvist J, Franzén S *et al*: **Carriage of penicillin-non-susceptible pneumococci among children in northern Tanzania in the 13-valent pneumococcal vaccine era.** *Int J Infect Dis* 2019, **81**:156-166.

132. Kobayashi M, Conklin LM, Bigogo G, Jagero G, Hampton L, Fleming-Dutra KE, Junghae M, Carvalho MdG, Pimenta F, Beall B *et al*: **Pneumococcal carriage and antibiotic susceptibility patterns from two cross-sectional colonization surveys among children aged** *BMC Infect Dis* 2017, **17**(1):25.

133. Ginsburg AS, Tinkham L, Riley K, Kay NA, Klugman KP and Gill CJ: **Antibiotic non-susceptibility among *Streptococcus pneumoniae* and *Haemophilus influenzae* isolates identified in African cohorts: a meta-analysis of three decades of published studies.** *Int J Antimicrob Agents* 2013, **42**(6):482-491.

134. Castanheira M, Gales AC, Mendes RE, Jones RN and Sader HS: **Antimicrobial susceptibility of *Streptococcus pneumoniae* in Latin America: results from five years of the SENTRY Antimicrobial Surveillance Program.** *Clin Microbiol Infect* 2004, **10**(7):645-651.

135. Schroeder MR, Stephens DS, Stephens DS and Stephens DS: **Macrolide resistance in *Streptococcus pneumoniae*.** *Front Cell Infect Microbiol* 2016,.

136. Cheng AC and Jenney AWJ: **Macrolide resistance in pneumococci-is it relevant?** *Pneumonia (Nathan)* 2016, **8**(1):10.

137. Redgrave LS, Sutton SB, Webber MA and Piddock LJV: **Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success.** *Trends Microbiol* 2014, **22**(8):438-445.

138. Marchant J. **When antibiotics turn toxic.** *Nature* 2018, **555**(7697):431-433.

139. Stanhope MJ, Walsh SL, Becker JA, Italia MJ, Ingraham KA, Gwynn MN, Mathie T, Poupard JA, Miller LA, Brown JR *et al*: **Molecular evolution perspectives on intraspecific lateral DNA transfer of topoisomerase and gyrase loci in *Streptococcus pneumoniae*, with implications for**

fluoroquinolone resistance development and spread. *Antimicrob Agents Chemother* 2005, **49**(10):4315-4326.

140. Fuller JD and Low DE: **A review of *Streptococcus pneumoniae* infection treatment failures associated with fluoroquinolone resistance.** *Clin Infect Dis* 2005, **41**(1):118-121.

141. Davidson R, Cavalcanti R, Brunton JL, Bast DJ, de Azavedo, Joyce C. S, Kibsey P, Fleming C and Low DE: **Resistance to levofloxacin and failure of treatment of pneumococcal pneumonia.** *N Engl J Med* 2002, **346**(10):747-750.

142. Brueggemann AB, Coffman SL, Rhomberg P, Huynh H, Almer L, Nilius A, Flamm R and Doern GV: **Fluoroquinolone resistance in *Streptococcus pneumoniae* in United States since 1994-1995.** *Antimicrob Agents Chemother* 2002, **46**(3):680-688.

143. Simoens S, Verhaegen J, Bleyenbergh Pv, Peetermans WE and Decramer M: **Consumption patterns and *in vitro* resistance *of Streptococcus pneumoniae* to fluoroquinolones.** *Antimicrob Agents Chemother* 2011, **55**(6):3051-3053.

144. Adam HJ, Hoban DJ, Gin AS and Zhanel GG: **Association between fluoroquinolone usage and a dramatic rise in ciprofloxacin-resistant *Streptococcus pneumoniae* in Canada, 1997–2006.** *Int J Antimicrob Agents* 2009, **34**(1):82-85.

145. Domenech A, Tirado-Vélez JM, Fenoll A, Ardanuy C, Yuste J, Liñares J and de la Campa, Adela G: **Fluoroquinolone-resistant pneumococci: dynamics of serotypes and clones in Spain in 2012 compared with those from 2002 and 2006.** *Antimicrob Agents Chemother* 2014, **58**(4):2393-2399.

146. Patel SN, McGeer A, Melano R, Tyrrell GJ, Green K, Pillai DR, Low DE and Canadian Bacterial Surveillance Network: **Susceptibility of *Streptococcus pneumoniae* to fluoroquinolones in Canada.** *Antimicrob Agents Chemother* 2011, **55**(8):3703-3708.

147. Ho PL, Cheng VC and Chow KH: **Decreasing prevalence of levofloxacin-resistant *Streptococcus pneumoniae* in Hong Kong, 2001 to 2007.** *J Antimicrob Chemother* 2009, **63**(4):836-838.

148. Kim SH, Song J, Chung DR, Thamlikitkul V, Yang Y, Wang H, Lu M, So TM, Hsueh P, Yasin RM *et al*: **Changing trends in antimicrobial resistance and serotypes of *Streptococcus pneumoniae* isolates in Asian countries: an Asian Network for Surveillance of resistant pathogens (ANSORP) study.** *Antimicrob Agents Chemother* 2012, **56**(3):1418-1426.

149. Finlay BB, Russell SL and Willing BP: **Shifting the balance: antibiotic effects on host–microbiota mutualism.** *Nat Rev Microbiol* 2011, **9**(4):233-243.

150. Blaser M. **Stop the killing of beneficial bacteria.** *Nature* 2011, **476**(7361):393.

151. Cotter PD, Stanton C, Ross RP and Hill C: **The impact of antibiotics on the gut microbiota as revealed by high throughput DNA sequencing.** *Discov Med* 2012, **13**(70):193.

152. Lloyd-Price J, Abu-Ali G and Huttenhower C: **The healthy human microbiome.** *Genome Med* 2016, **8**(1).

153. Cotter PD, Ross RP and Hill C: **Bacteriocins — a viable alternative to antibiotics?** *Nat Rev Microbiol* 2013, **11**(2):95-105.

154. Czaplewski L, Bax R, Clokie M, Dawson M, Fairhead H, Fischetti VA, Foster S, Gilmore B, Hancock REW and Jones B: **Alternatives to antibiotics - a pipeline portfolio review.** *Lancet Infect Dis* 2016, **2**(16):239-51.

155. Riley MA and Wertz JE: **Bacteriocins: evolution, ecology, and application.** *Annu Rev Microbiol* 2002, **56**(1):117-137.

156. Jenssen H, Hamill P and Hancock REW: **Peptide antimicrobial agents.** *Clin Microbiol Rev* 2006, **19**(3):491-511.

157. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J *et al*: **Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature.** *Nat Prod Rep* 2012, **3**(1):18-16.

158. Gillor O, Vriezen JAC and Riley MA: **The role of SOS boxes in enteric bacteriocin regulation.** *Microbiology* 2008, **154**(6):1783-1792.

159. Kjos M, Miller E, Slager J, Lake FB, Gericke O, Roberts IS, Rozen DE and Veening J: **Expression of *Streptococcus pneumoniae* bacteriocins is induced by antibiotics via regulatory interplay with the competence system.** *PLoS Pathog* 2016, **12**(2):e1005422.

160. Dobson A, Cotter PD, Ross RP and Hill C: **Bacteriocin production: a probiotic trait?** *Appl Environ Microbiol* 2012, **78**(1):1-6.

161. Rea MC, Ross RP, Cotter PD and Hill C: **Classification of bacteriocins from Gram-positive bacteria.** Springer New York, New York 2011:29-53.

162. Jack RW, Tagg JR and Ray B: **Bacteriocins of Gram-positive bacteria.** *Microbiol Rev* 1995, **59**(2):171-200.

163. Geis A, Singh J and Teuber M: **Potential of lactic streptococci to produce bacteriocin.** *Appl Environ Microbiol* 1983, **45**(1):205-211.

164. Klaenhammer TR. **Genetics of bacteriocins produced by lactic acid bacteria.** *FEMS Microbiol Rev* 1993, **12**(1):39-85.

165. Cotter PD, Hill C and Ross RP: **Bacteriocins: developing innate immunity for food.** *Nat Rev Microbiol* 2005, **3**(10):777-788.

166. Dischinger J, Basi Chipalu S and Bierbaum G: **Lantibiotics: promising candidates for future applications in health care.** *Int J Med Microbiol* 2013, **304**(1):51-62.

167. Shin JM, Gwak JW, Kamarajan P, Fenno JC, Rickard AH and Kapila YL: **Biomedical applications of nisin.** *J Appl Microbiol* 2016, **120**(6):1449-1465.

168. Asaduzzaman SM, Al-Mahin A, Bashar T and Noor R: **Lantibiotics: A candidate for future generation of antibiotics.** *Stamford J Microbiology* 1970, **1**(1):1-12.

169. Draper LA, Cotter PD, Hill C and Ross RP: **Lantibiotic resistance.** *Microbiol Mol Biol Rev* 2015, **79**(2):171-191.

170. Egan K, Field D, Rea MC, Ross RP, Hill C and Cotter PD: **Bacteriocins: novel solutions to age old spore-related problems?** *Front Microbiol* 2016, **7**:461.

171. Gut IM, Blanke SR and van der Donk, Wilfred A: **Mechanism of inhibition of *Bacillus anthracis* spore outgrowth by the lantibiotic nisin.** *ACS Chem Biol* 2011, **6**(7):744-752.

172. Gabrielsen C, Brede DA, Nes IF and Diep DB: **Circular bacteriocins: biosynthesis and mode of action.** *Appl Environ Microbiol* 2014, **80**(22):6854-6862.

173. Montoya E, Gálvez A, Maqueda M, Valdivia E and Quesada A: **Characterization and partial purification of a broad spectrum antibiotic AS-48 produced by *Streptococcus faecalis*.** *Can J Microbiol* 1986, **32**(10):765-771.

174. Samyn B, Martinez-Bueno M, Devreese B, Maqueda M, Gálvez A, Valdivia E, Coyette J and Van Beeumen J: **The cyclic structure of the enterococcal peptide antibiotic AS-48.** *FEBS Lett* 1994, **352**(1):87-90.

175. Bogaardt C, van Tonder AJ and Brueggemann AB: **Genomic analyses of pneumococci reveal a wide diversity of bacteriocins – including pneumocyclicin, a novel circular bacteriocin.** *BMC Genomics* 2015, **16**(1):554.

176. Mahanta N, Hudson GA and Mitchell DA: **Radical S-adenosylmethionine enzymes involved in RiPP biosynthesis.** *Biochemistry* 2017, **56**(40):5229-5244.

177. Grove TL, Himes PM, Hwang S, Yumerefendi H, Bonanno JB, Kuhlman B, Almo SC and Bowers AA: **Structural insights into thioether bond formation in the biosynthesis of sactipeptides.** *J Am Chem Soc* 2017, **139**(34):11734-11744.

178. Babasaki K, Takao T, Shimonishi Y and Kurahashi K: **Subtilosin A, a new antibiotic peptide produced by *Bacillus subtilis* 168: isolation, structural analysis, and biogenesis.** *J Biochem* 1985, **98**(3):585-603.

179. Sutyak KE, Wirawan RE, Aroutcheva AA and Chikindas ML: **Isolation of the *Bacillus subtilis* antimicrobial peptide subtilosin from the dairy product-derived *Bacillus amyloliquefaciens*.** *J Appl Microbiol* 2008, **104**(4):1067-1074.

180. Rajakovich LJ and Balskus EP: **Metabolic functions of the human gut microbiota: the role of metalloenzymes.** *Nat Prod Rep* 2019, **36**(4):593-625.

181. Rea MC, Dobson A, O'Sullivan O, Crispie F, Fouhy F, Cotter PD, Shanahan F, Kiely B, Hill C, Ross RP *et al*: **Effect of broad- and narrow-spectrum**

antimicrobials on *Clostridium difficile* and microbial diversity in a model of the distal colon. *Proc Natl Acad Sci U S A* 2011, **108**(Supplement 1):4639-4644.

182. Zhao N, Pan Y, Cheng Z and Liu H: **Lasso peptide, a highly stable structure and designable multifunctional backbone.** *Amino Acids* 2016, **48**(6):1347-1356.

183. Weber W, Fischli W, Hochuli E, Kupfer E and Weibel EK: **Anantin-A peptide antagonist of the atrial natriuretic factor(ANF). I. producing organism, fermentation, isolation and biological activity.** *J Antibiot* 1991, **44**(2):164-171.

184. Destoumieux-Garzón D, Peduzzi J and Rebuffat S: **Focus on modified microcins: structural features and mechanisms of action.** *Biochimie* 2002, **84**(5):511-519.

185. Esumi Y, Suzuki Y, Itoh Y, Uramoto M, Kimura K, Goto M, Yoshihama M and Ichikawa T: **Propeptin, a new inhibitor of prolyl endopeptidase produced by microbispora II. determination of chemical structure.** *J Antibiot* 2002, **55**(3):296-300.

186. Potterat O, Stephan H, Metzger JW, Gnau V, Zähner H and Jung G: **Aborycin—a tricyclic 21-peptide antibiotic isolated from *Streptomyces griseoflavus.*** *Liebigs Ann Chem* 1994, **25**(47):no.

187. Severinov K, Semenova E, Kazakov A, Kazakov T and Gelfand MS: **Low-molecular-weight post-translationally modified microcins.** *Mol Microbiol* 2007, **66**(1):277.

188. Wu C, Biswas S, Garcia De Gonzalo, Chantal V and van der Donk, Wilfred A: **Investigations into the mechanism of action of sublancin.** *ACS Infect Dis* 2019,.

189. Martinez B, Suarez JE and Rodriguez A: **Lactococcin 972: a homodimeric lactococcal bacteriocin whose primary target is not the plasma membrane.** *Microbiology* 1996, **142**(9):2393-2398.

190. Letzel A, Pidot SJ and Hertweck C: **Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria.** *BMC Genomics* 2014, **15**(1):983.

191. Martínez B, Böttiger T, Schneider T, Rodríguez A, Sahl H and 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.

192. Martinez B, Rodriguez A and Suarez JE: **Lactococcin 972, a bacteriocin that inhibits septum formation in lactococci.** *Microbiology* 2000, **146**(4):949-955.

193. Maldonado-Barragán A, Cárdenas N, Martínez B, Ruiz-Barba JL, Fernández-Garayzábal JF, Rodríguez JM and Gibello A: **Garvicin A, a novel class IId**

bacteriocin from *Lactococcus garvieae* that inhibits septum formation in *L. garvieae* strains. *Appl Environ Microbiol* 2013, **79**(14):4336-4346.

194. Cintas LM, Casaus P, Håvarstein LS, Hernández PE and Nes IF: **Biochemical and genetic characterization of enterocin P, a novel sec-dependent bacteriocin from *Enterococcus faecium* P13 with a broad antimicrobial spectrum.** *Appl Environ Microbiol* 1997, **63**(11):4321-4330.

195. Kjos M, Borrero J, Opsata M, Birri DJ, Holo H, Cintas LM, Snipen L, Hernández PE, Nes IF and Diep DB: **Target recognition, resistance, immunity and genome mining of class II bacteriocins from Gram-positive bacteria.** *Microbiology* 2011, **157**(Pt 12):3256-3267.

196. Montalbán-López M, Zhou L, Buivydas A, van Heel AJ and Kuipers OP: **Increasing the success rate of lantibiotic drug discovery by Synthetic Biology.** *Expert Opin Drug Discov* 2012, **7**(8):695-709.

197. Hammami R, Zouhir A, Ben Hamida J and Fliss I: **BACTIBASE: a new web-accessible database for bacteriocin characterization.** *BMC Microbiol* 2007, **7**(1):89.

198. de Jong A, van Hijum S, Bijlsma JJE, Kok J and Kuipers OP: **BAGEL: a web-based bacteriocin genome mining tool.** *Nucleic Acids Res* 2006, **34**(suppl_2):W273-W279.

199. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.

200. Begley M, Cotter PD, Hill C and 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.

201. Kjos M, Snipen L, Salehian Z, Nes IF and Diep DB: **The Abi proteins and their involvement in bacteriocin self-immunity.** *J Bacteriol* 2010, **192**(8):2068-2076.

202. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, Los Santos, De, Emmanuel L.C, Kim HU, Nave M *et al*: **AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification.** *Nucleic Acids Res* 2017, **45**(W1):W36-W41.

203. Wang H, Fewer DP and Sivonen K: **Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria.** *PLoS One* 2011, **6**(7):e22384.

204. Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH and Prangishvili D: **Exploring the prokaryotic virosphere.** *Res Microbiol* 2008, **159**(5):306-313.

205. Hendrix RW. **Bacteriophages: evolution of the majority.** *Theor Popul Biol* 2002, **61**(4):471-480.

206. Wittebole X, De Roock S and Opal SM: **A historical overview of bacteriophage therapy as an alternative to antibiotics for the treatment of bacterial pathogens.** *Virulence* 2014, **5**(1):226-235.

207. Yanofsky C. **Establishing the triplet nature of the genetic code.** *Cell* 2007, **128**(5):815-818.

208. Cobb M. **Who discovered messenger RNA?** *Curr Biol* 2015, **25**(13):526-532.

209. Drulis-Kawa Z, Majkowska-Skrobek G and Maciejewska B: **Bacteriophages and phage-derived proteins—application approaches.** *Curr Med Chem* 2015, **22**(14):1757-1773.

210. Landy A and Ross W: **Viral integration and excision: structure of the lambda att sites.** *Science* 1977, **197**(4309):1147-1160.

211. Ramisetty BCM and Sudhakari PA: **Bacterial 'grounded' prophages: hotspots for genetic renovation and innovation.** *Front Genet* 2019, **10**.

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

213. Boyd EF and Brüssow H: **Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved.** *Trends Microbiol* 2002, **10**(11):521-529.

214. Barbara A. Bensing, Ian R. Siboo and Paul M. Sullam: **Proteins PblA and PblB of *Streptococcus mitis*, which promote binding to human platelets, are encoded within a lysogenic bacteriophage.** *Infect Immun* 2001, **69**(10):6186-6192.

215. Vaca Pacheco S, García González O and Paniagua Contreras GL: **The *lom* gene of bacteriophage λ is involved in *Escherichia coli* K12 adhesion to human buccal epithelial cells.** *FEMS Microbiol Lett* 1997, **156**(1):129-132.

216. Mirold S, Rabsch W, Rohde M, Stender S, Tschäpe H, Rüssmann H, Igwe E and Hardt W: **Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain.** *Proc Natl Acad Sci U S A* 1999, **96**(17):9845-9850.

217. Figueroa-Bossi N, Uzzau S, Maloriol D and Bossi L: **Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*.** *Mol Microbiol* 2001, **39**(2):260-272.

218. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I and Herskovits AA: **A new perspective on lysogeny: prophages as active regulatory switches of bacteria.** *Nat Rev Microbiol* 2015, **13**(10):641-650.

219. Koskella B and Brockhurst MA: **Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities.** *FEMS Microbiol Rev* 2014, **38**(5):916-931.

220. Frígols B, Quiles-Puchalt N, Mir-Sanchis I, Donderis J, Elena SF, Buckling A, Novick RP, Marina A and Penadés JR: **Virus satellites drive viral evolution and ecology.** *PLoS Genet* 2015, **11**(10):e1005609.

221. Palukaitis P. **Chapter 50 - Satellite viruses and satellite nucleic acids.** In: Hadidi A, Flores, R, Randles JW, Palukaitis P, **Viroids and Satellites**, Academic Press, London 2017:545-552.

222. Novick RP. **Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of *Staphylococcus aureus*.** *Plasmid* 2003, **49**(2):93-105.

223. Christie GE, Novick RP and Penadés JR: **The phage-related chromosomal islands of Gram-positive bacteria.** *Nat Rev Microbiol* 2010, **8**(8):541-551.

224. Penadés JR and Christie GE: **The phage-inducible chromosomal islands: a family of highly evolved molecular parasites.** *Annu Rev Virol* 2015, **2**(1):181-201.

225. Varon M and Levisohn R: **Three-membered parasitic system: a bacteriophage, *Bdellovibrio bacteriovorus*, and *Escherichia coli*.** *J Virol* 1972, **9**(3):519-525.

226. Novick RP and Ram G: **Staphylococcal pathogenicity islands—movers and shakers in the genomic firmament.** *Curr Opin Microbiol* 2017, **38**:197-204.

227. Scott J, Nguyen SV, King CJ, Hendrickson C and McShan WM: **Phage-like *Streptococcus pyogenes* chromosomal islands (SpyCI) and mutator phenotypes: control by growth state and rescue by a SpyCI-encoded promoter.** *Front Microbiol* 2012, **3**:317.

228. Campoy S, Mir I, Novick RP, Lasa Í, Barbé J, Christie GE, Tormo-Más MÁ, Penadés JR, Tallent SM and Shrestha A: **Moonlighting bacteriophage proteins derepress staphylococcal pathogenicity islands.** *Nature* 2010, **465**(7299):779-782.

229. Ubeda C, Olivarez NP, Barry P, Wang H, Kong X, Matthews A, Tallent SM, Christie GE and Novick RP: **Specificity of staphylococcal phage and SaPI**

DNA packaging as revealed by integrase and terminase mutations. *Mol Microbiol* 2009, **72**(1):98.

230. Poliakov A, Chang JR, Spilman MS, Damle PK, Christie GE, Mobley JA and Dokland T: **Capsid size determination by *Staphylococcus aureus* pathogenicity island SaPI1 involves specific incorporation of SaPI1 proteins into procapsids.** *J Mol Biol* 2008, **380**(3):465-475.

231. Martínez-Rubio R, Quiles-Puchalt N, Martí M, Humphrey S, Ram G, Smyth D, Chen J, Novick RP and Penadés JR: **Phage-inducible islands in the Gram-positive cocci.** *ISME J* 2017, **11**(4):1029-1042.

232. Clokie MRJ, Kropinski AM and Lavigne R: **Bacteriophages: Methods and Protocols.** Humana Press, New York 2018.

233. Ramirez M, Severina E and Tomasz A: **A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*.** *J Bacteriol* 1999, **181**(12):3618-3625.

234. Lima-Mendez G, Van Helden J, Toussaint A and Leplae R: **Prophinder: a computational tool for prophage prediction in prokaryotic genomes.** *Bioinformatics* 2008, **24**(6):863-865.

235. Zhou Y, Liang Y, Lynch KH, Dennis JJ and Wishart DS: **PHAST: a fast phage search tool.** *Nucleic Acids Res* 2011, **39**(suppl_2):W347-W352.

236. Crispim JS, Dias RS, Vidigal PMP, de Sousa MP, da Silva CC, Santana MF and de Paula SO: **Screening and characterization of prophages in *Desulfovibrio* genomes.** *Sci Rep* 2018, **8**(1):9273-10.

237. Castillo D, Kauffman K, Hussain F, Kalatzis P, Rørbo N, Polz MF and Middelboe M: **Widespread distribution of prophage-encoded virulence factors in marine *Vibrio* communities.** *Sci Rep* 2018, **8**(1):9973-9.

238. Fu Y, Wu Y, Yuan Y and Gao M: **Prevalence and diversity analysis of candidate prophages to provide an understanding on their roles in *Bacillus thuringiensis*.** *Viruses* 2019, **11**(4):388.

239. Jolley KA and Maiden MCJ: **BIGSdb: scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics* 2010, **11**(1):595.

240. Blanchette KA, Shenoy AT, Milner J, Gilley RP, McClure E, Hinojosa CA, Kumar N, Daugherty SC, Tallon LJ, Ott S *et al*: **Neuraminidase A exposed galactose promotes *Streptococcus pneumoniae* biofilm formation during colonization.** *Infect Immun* 2016, **84**(10):2922-2932.

241. Delcher AL, Bratke KA, Powers EC and Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer.** *Bioinformatics* 2007, **23**(6):673-679.

242. Meyer F, Overbeek R and Rodriguez A: **FIGfams: yet another set of protein families.** *Nucleic Acids Res* 2009, **37**(20):6643-6654.

243. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M *et al*: **The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST).** *Nucleic Acids Res* 2014, **42**:D206-D214.

244. Seemann T. **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics* 2014, **30**(14):2068-2069.

245. Marchler-Bauer A and Bryant SH: **CD-Search: protein domain annotations on the fly.** *Nucleic Acids Res* 2004, **32**:W327-W331.

246. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR *et al*: **CDD: a conserved domain database for the functional annotation of proteins.** *Nucleic Acids Res* 2011, **39**:D225-D229.

247. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al*: **STRING v10: protein–protein interaction networks, integrated over the tree of life.** *Nucleic Acids Res* 2015, **43**:D447-D452.

248. Letunic I and Bork P: **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation.** *Bioinformatics* 2007, **23**(1):127-128.

249. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA and Parkhill J: **Roary: rapid large-scale prokaryote pan genome analysis.** *Bioinformatics* 2015, **31**(22):3691-3693.

250. Reichmann P and Hakenbeck R: **Allelic variation in a peptide-inducible two-component system of *Streptococcus pneumoniae*.** *FEMS Microbiol Lett* 2000, **190**(2):231-236.

251. Lux T, Nuhn M, Hakenbeck R and Reichmann P: **Diversity of bacteriocins and activity spectrum in *Streptococcus pneumoniae*.** *J Bacteriol* 2007, **189**(21):7741-7751.

252. Son MR, Shchepetov M, Adrian PV, Madhi SA, de Gouveia L, von Gottberg A, Klugman KP, Weiser JN and Dawid S: **Conserved mutations in the pneumococcal bacteriocin transporter gene, *blpA*, result in a complex population consisting of producers and cheaters.** *mBio* 2011, **2**(5).

253. de Saizieu A, Gardes C, Flint N and Wagner C: **Microarray-based identification of a novel *Streptococcus pneumoniae* regulon controlled by an autoinduced peptide.** *J Bacteriol* 2000, **182**(17):4696-4703.

254. Wholey W, Abu-Khdeir M, Yu EA, Siddiqui S, Esimai O and Dawid S: **Characterization of the competitive pneumocin peptides of *Streptococcus pneumoniae*.** *Front Cell Infect Microbiol* 2019, **9**:55.

255. Guiral S, Mitchell TJ, Martin B, Claverys J and Losick RM: **Competence-programmed predation of noncompetent cells in the human pathogen *Streptococcus pneumoniae*: genetic requirements.** *Proc Natl Acad Sci U S A* 2005, **102**(24):8710-8715.

256. Maricic N, Anderson ES, Opipari AE, Yu EA and Dawid S: **Characterization of a multipeptide lantibiotic locus in *Streptococcus pneumoniae*.** *mBio* 2016, **7**(1):e01656.

257. Hoover SE, Perez AJ, Tsui HT, Sinha D, Smiley DL, DiMarchi RD, Winkler ME and Lazazzera BA: **A new quorum-sensing system (TprA/PhrA) for**

***Streptococcus pneumoniae* D39 that regulates a lantibiotic biosynthesis gene cluster.** *Mol Microbiol* 2015, **97**(2):229-243.

258. Majchrzykiewicz JA, Lubelski J, Moll GN, Kuipers A, Bijlsma JJE, Kuipers OP and Rink R: **Production of a class II two-component lantibiotic of *Streptococcus pneumoniae* using the class I nisin synthetic machinery and leader sequence.** *Antimicrob Agents Chemother* 2010, **54**(4):1498-1505.

259. Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, Mitchell AM, Quail MA, Andrew PW, Parkhill J *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.

260. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, Linden Mvd, McGee L, Gottberg Av, Song JH, Ko KS *et al*: **Rapid pneumococcal evolution in response to clinical interventions.** *Science* 2011, **331**(6016):430-434.

261. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP and Lipsitch M: **Population genomics of post-vaccine changes in pneumococcal epidemiology.** *Nat Genet* 2013, **45**(6):656-663.

262. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, Feltwell T, Fookes MC, Harris SR, Dordel J, Salter SJ *et al*: **Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection.** *PLoS Genet* 2013, **9**(10):e1003868.

263. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ *et al*: **Dense genomic sampling identifies highways of pneumococcal recombination.** *Nat Genet* 2014, **46**(3):305-309.

264. van Tonder AJ, Mistry S, Bray JE, Hill DMC, Cody AJ, Farmer CL, Klugman KP, von Gottberg A, Bentley SD, Parkhill J *et al*: **Defining the estimated core genome of bacterial populations using a Bayesian decision model.** *PLoS Comput Biol* 2014, **10**(8):e1003788.

265. van Tonder AJ, Bray JE, Roalfe L, White R, Zancolli M, Quirk SJ, Haraldsson G, Jolley KA, Maiden MCJ, Bentley SD *et al*: **Genomics reveals the worldwide distribution of multidrug-resistant serotype 6E pneumococci.** *J Clin Microbiol* 2015, **53**(7):2271-2285.

266. Gladstone RA, Jefferies JM, Tocheva AS, Beard KR, Garley D, Chong WW, Bentley SD, Faust SN and Clarke SC: **Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction.** *Vaccine* 2015, **33**(17):2015-2021.

267. Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al*: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30**(9):1236-1240.

268. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**(21):2947-2948.

269. Price MN, Dehal PS and Arkin AP: **FastTree 2 – approximately maximum-likelihood trees for large alignments.** *PLoS One* 2010, **5**(3):e9490.

270. Didelot X and Wilson DJ: **ClonalFrameML: efficient inference of recombination in whole bacterial genomes.** *PLoS Comput Biol* 2015, **11**(2):e1004041.

271. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ *et al*: **Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain.** *Microbiology* 2012, **158**(Pt 4):1005-1015.

272. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M and Carriço JA: **PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods.** *BMC Bioinformatics* 2012, **13**(1):87.

273. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG and Parkhill J: **ACT: the Artemis comparison tool.** *Bioinformatics* 2005, **21**(16):3422-3423.

274. Langmead B and Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**(4):357-359.

275. Anders S and Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11**(10):R106.

276. Zerbino DR and Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**(5):821-829.

277. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M, Lió P, Crescenzi P, Fani R and Fondi M: **MeDuSa: a multi-draft based scaffolder.** *Bioinformatics* 2015, **31**(15):2443-2451.

278. Walker GV, Heng NCK, Carne A, Tagg JR and 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.

279. Kadam A, Eutsey RA, Rosch J, Miao X, Longwell M, Xu W, Woolford CA, Hillman T, Motib AS, Yesilkaya H *et al*: **Promiscuous signaling by a regulatory system unique to the pandemic PMEN1 pneumococcal lineage.** *PLoS Pathog* 2017, **13**(5):e1006339.

280. Perez-Pascual D, Monnet V and Gardan R: **Bacterial cell–cell communication in the host via RRNPP peptide-binding regulators.** *Front Microbiol* 2016, **7**:706.

281. Brown SP, West SA, Diggle SP and Griffin AS: **Social evolution in micro-organisms and a Trojan horse approach to medical intervention strategies.** *Philos Trans R Soc Lond B Biol Sci* 2009, **364**(1533):3157-3168.

282. García-Rodríguez JA and Fresnadillo Martínez MJ: **Dynamics of nasopharyngeal colonization by potential respiratory pathogens.** *J Antimicrob Chemother* 2002,(90003):59-74.

283. Auranen K, Mehtälä J, Tanskanen A and 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.

284. Huang SS, Platt R, Rifas-Shiman SL, Pelton SI, Goldmann D and Finkelstein JA: **Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004.** *Pediatrics* 2005, **116**(3):e408-e413.

285. Aguiar SI, Brito MJ, Gonçalo-Marques J, Melo-Cristino J and Ramirez M: **Serotypes 1, 7F and 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7 years of heptavalent conjugate vaccine use.** *Vaccine* 2010, **28**(32):5167-5173.

286. Kaplan SL, Barson WJ, Lin PL, Stovall SH, Bradley JS, Tan TQ, Hoffman JA, Givner LB and Mason J, Edward O: **Serotype 19A is the most common serotype causing invasive pneumococcal infections in children.** *Pediatrics* 2010, **125**(3):429-436.

287. Eutsey RA, Powell E, Dordel J, Salter SJ, Clark TA, Korlach J, Ehrlich GD and Hiller NL: **Genetic stabilization of the drug-resistant PMEN1 Pneumococcus lineage by its distinctive DpnIII restriction-modification system.** *mBio* 2015, **6**(3):e00173.

288. Johnston C, Polard P and Claverys J: **The DpnI/DpnII pneumococcal system, defense against foreign attack without compromising genetic exchange.** *Mob Genet Elements* 2013, **3**(4):e25582.

289. Bidossi A, Mulas L, Decorosi F, Colomba L, Ricci S, Pozzi G, Deutscher J, Viti C and Oggioni MR: **A functional genomics approach to establish the complement of carbohydrate transporters in *Streptococcus pneumoniae*.** *PLoS One* 2012, **7**(3):e33320.

290. Cuevas RA, Eutsey R, Kadam A, West-Roberts JA, Woolford CA, Mitchell AP, Mason KM and Hiller NL: **A novel streptococcal cell–cell communication peptide promotes pneumococcal virulence and biofilm formation.** *Mol Microbiol* 2017, **105**(4):554-571.

291. Jutta M. Loeffler and Vincent A. Fischetti: **Lysogeny of *Streptococcus pneumoniae* with MM1 phage: improved adherence and other phenotypic changes.** *Infect Immun* 2006, **74**(8):4486-4495.

292. Hsieh Y, Lin T, Lin C and Wang J: **Identification of PblB mediating galactose-specific adhesion in a successful *Streptococcus pneumoniae* clone.** *Sci Rep* 2015, **5**(1):12265.

293. Canchaya C, Desiere F, McShan WM, Ferretti JJ, Parkhill J and Brüssow H: **Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain SF370.** *Virology* 2002, **302**(2):245-258.

294. Bobay L, Touchon M and Rocha EP: **Pervasive domestication of defective prophages by bacteria.** *Proc Natl Acad Sci U S A* 2014, **111**(33):12127-12132.

295. Li W and Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics* 2006, **22**(13):1658-1659.

296. Love MI, Huber W and Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15**(12):550.

297. Kjos M, Aprianto R, Fernandes VE, Anew PW, van Strijp, Jos A G, Nijland R and Veening J: **Bright fluorescent *Streptococcus pneumoniae* for live cell imaging of host-pathogen interactions.** *J Bacteriol* 2015, **197**(5):807-818.

298. Pease LR and Heckman KL: **Gene splicing and mutagenesis by PCR-driven overlap extension.** *Nat Protoc* 2007, **2**(4):924-932.

299. Khandavilli S, Homer KA, Yuste J, Basavanna S, Mitchell T and Brown JS: **Maturation of *Streptococcus pneumoniae* lipoproteins by a type II signal peptidase is required for ABC transporter function and full virulence.** *Mol Microbiol* 2008, **67**(3):541-557.

300. Basavanna S, Chimalapati S, Maqbool A, Rubbo B, Yuste J, Wilson RJ, Hosie A, Ogunniyi AD, Paton JC, Thomas G *et al*: **The effects of methionine acquisition and synthesis on *Streptococcus pneumoniae* growth and virulence.** *PLoS One* 2013, **8**(1):e49638.

301. Ramos-Sevillano E, Urzainqui A, Campuzano S, Moscoso M, Gonzalez-Camacho F, Domenech M, de Cordoba SR, Sanchez-Madrid F, Brown JS and Garcia E: **Pleiotropic effects of cell wall amidase LytA on *Streptococcus***

pneumoniae **sensitivity to the host immune response.** *Infect Immun* 2015, **83**(2):591-603.

302. Beuzón CR and Holden DW: **Use of mixed infections with *Salmonella* strains to study virulence genes and their interactions *in vivo*.** *Microbes Infect* 2001, **3**(14):1345-1352.

303. Yuste J, Botto M, Paton JC, Holden DW and Brown JS: **Additive inhibition of complement deposition by pneumolysin and PspA facilitates *Streptococcus pneumoniae* septicemia.** *J Immunol* 2005, **175**(3):1813-1819.

304. Ramos-Sevillano E, Moscoso M, García P, García E and Yuste J: **Nasopharyngeal colonization and invasive disease are enhanced by the cell wall hydrolases LytB and LytC of *Streptococcus pneumoniae*.** *PloS One* 2011, **6**(8):e23626.

305. Ji X, Sun Y, Liu J, Zhu L, Guo X, Lang X and Feng S: **A novel virulence-associated protein, *vapE*, in *Streptococcus suis* serotype 2.** *Mol Med Rep* 2016, **13**(3):2871-2877.

306. Bernheimer HP. **Lysogenic pneumococci and their bacteriophages.** *J Bacteriol* 1979, **138**(2):618-624.

307. López E, Domenech A, Ferrándiz M, Frias MJ, Ardanuy C, Ramirez M, García E, Liñares J and Campa, Adela G de la: **Induction of prophages by fluoroquinolones in *Streptococcus pneumoniae*: implications for**

emergence of resistance in genetically-related clones. *PLoS One* 2014, **9**(4):e94358.

308. Beres SB, Sylva GL, Barbian KD, Lei B, Hoff JS, Mammarella ND, Liu M, Smoot JC, Porcella SF, Parkins LD *et al*: **Genome sequence of a serotype M3 strain of group A *Streptococcus*: Phage-encoded toxins, the high-virulence phenotype, and clone emergence.** *Proc Natl Acad Sci U S A* 2002, **99**(15):10078-10083.

309. van der Mee-Marquet N, Diene SM, Barbera L, Courtier-Martinez L, Lafont L, Ouachée A, Valentin AS, Santos SD, Quentin R and François P: **Analysis of the prophages carried by human infecting isolates provides new insight into the evolution of Group B *Streptococcus* species.** *Clin Microbiol Infect* 2018, **24**(5):514-521.

310. Langille MGI, Brinkman FSL and Hsiao WWL: **Detecting genomic islands using bioinformatics approaches.** *Nat Rev Microbiol* 2010, **8**(5):373-382.

311. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics* 2009, **25**(11):1422-1423.

312. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V *et al*: **Scikit-Learn: machine learning in Python.** *J Mach Learn Res* 2011,.

313. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream M and Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**(10):944-945.

314. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C and Bork P: **Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper.** *Mol Biol Evol* 2017, **34**(8):2115-2122.

315. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M *et al*: **eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences.** *Nucleic Acids Res* 2016, **44**(D1):D286-D293.

316. Brian G. Spratt and Martin C. J. Maiden: **Bacterial population genetics, evolution and epidemiology.** *Philos Trans R Soc Lond B Biol Sci* 1999, **354**(1384):701-710.

317. Feil EJ, Smith JM, Enright MC and Spratt BG: **Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data.** *Genetics* 2000, **154**(4):1439-1450.

318. Weinbauer MG. **Ecology of prokaryotic viruses.** *FEMS Microbiol Rev* 2004, **28**(2):127-181.

319. Ackermann HW and DuBow MS: **Phage multiplication.** CRC Press, Inc., Florida 1987:49-85.

320. Ackermann H, Audurier A, Berthiaume L, Jones LA, Mayo JA and Vidaver AK: **Guidelines for bacteriophage characterization.** *Adv. Virus Res* 1978, **23**:1-24.

321. Ross A, Ward S and Hyman P: **More is better: selecting for broad host range bacteriophages.** *Front Microbiol* 2016, **7**:1352.

322. Loc-Carrillo C and Abedon ST: **Pros and cons of phage therapy.** *Bacteriophage* 2011, **1**(2):111-114.

323. Nilsson AS. **Phage therapy-constraints and possibilities.** *Ups J Med Sci* 2014, **119**(2):192-198.

324. Phillippy AM. **New advances in sequence assembly.** *Genome Res* 2017, **27**(5):xi-xiii.

325. Shany D, Sarah M, Gal O, Azita L, Anna L, Mai K, Gil A and Rotem S: **Systematic discovery of antiphage defense systems in the microbial pangenome.** *Science* 2018, **359**(6379):4120.

326. Cascales E, Buchanan SK, Duché D, Kleanthous C, Lloubès R, Postle K, Riley M, Slatin S and Cavard D: **Colicin Biology.** *Microbiol Mol Biol Rev* 2007, **71**(1):158-229.

327. Kol'tsova EG, Suchkov YG and Lebedeva SA: **Transmission of a bacteriocinogenic factor in *Pasteurella pestis*.** *Sov Genet* 1971, **7**(4):507.

328. Pallavali RR, Degati VL, Lomada D, Reddy MC and Durbaka VRP: **Isolation and *in vitro* evaluation of bacteriophages against MDR-bacterial isolates from septic wound infections.** *PLoS One* 2017, **12**(7):e0179245.

329. Lambert RJW and Pearson J: **Susceptibility testing: accurate and reproducible minimum inhibitory concentration (MIC) and non-inhibitory concentration (NIC) values.** *J Appl Microbiol* 2000, **88**(5):784-790.

8. Appendices

Appendix 1 – Oral presentation

Conference:

The 13th European Meeting on the Molecular Biology of the Pneumococcus, in Stockholm, Sweden in June 2017

Title:

Genome mining and transcriptome analyses of bacteriocins in *Streptococcus pneumoniae*

Authors:

Reza Rezaei Javan[1], Angela Brueggemann[1, 2]

Authors' affiliations:

[1]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[2]Department of Medicine, Imperial College London, London, United Kingdom

Abstract:

Competition among bacterial members of the nasopharyngeal microbiome is believed to be mediated by bacteriocins: antimicrobial toxins produced by many bacterial species to inhibit other bacteria in the same ecological niche. The producer strain also encodes an immunity protein to protect itself from its own bacteriocin. Their role as weapons of defence makes the bacteriocins attractive candidates for the development of new antibiotics. A genome mining approach was employed to discover bacteriocin systems in a large, diverse set of historical and modern pneumococcal genomes isolated from 1937-2016. Nearly 7,000 pneumococcal genomes isolated from 40 countries were screened in silico and the bacteriocins were investigated in the context of the pneumococcal population

structure. Our study revealed that the pneumococcus possesses more bacteriocins than previously thought. Some of the bacteriocin cassettes were found in all pneumococci, while others were limited to a small percentage of isolates. Furthermore, RNA sequencing transcriptome analyses were used to investigate the triggers of bacteriocin expression, which currently remain poorly understood. Our study found that external factors like heat stress can induce the expression of bacteriocin genes. These findings fundamentally change our view of bacteriocins and nasopharyngeal competition among pneumococci.

Appendix 2 – Oral presentation

Conference:

The 11th International Symposium on Pneumococci and Pneumococcal Diseases, in Melbourne, Australia in April 2018

Title:

Genomic investigation of phage-inducible chromosomal islands in *Streptococcus pneumoniae*

Authors:

Reza Rezaei Javan[1], Angela Brueggemann[1, 2]

Authors' affiliations:

[1]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[2]Department of Medicine, Imperial College London, London, United Kingdom

Abstract:

Background and Aims: Phage-inducible chromosomal islands (PICIs) are a recently discovered type of satellite virus that do not have the ability to replicate on their own and have a life cycle dependent on a helper virus. These elements can exploit bacteriophages integrated within a bacterial chromosome as helpers, by manipulating the bacteriophage life cycle to enable their own replication and promiscuous spread. Originally identified in *Staphylococcus aureus*, they were shown to be vectors for the spreading of toxin genes and other virulence factors. The prevalence, diversity and genetic stability of PICIs in pneumococcus are virtually unknown. We aimed to identify and categorise pneumococcal PICIs and investigate their molecular epidemiology in the context of the pneumococcal

population structure.

Methods: We analysed the genomes of 482 diverse pneumococci recovered since 1916 from ill and healthy persons in 36 different countries for evidence of PICIs. PICI sequences were defined as those having a genomic organisation similar to the previously reported staphylococcal PICIs.

Results: We identified 45 unique pneumococcal PICIs. The genomic organisation of the PICIs indicated conserved modular structures, with genes clustered according to function. Multiple sequence alignment analyses suggested that pneumococcal PICIs can be categorised into four major groups. 34.4% (166/482) of the pneumococcal genomes harboured at least one PICI and 5.6% (n = 27) contained two. Some PICIs were detected in multiple clonal complexes (CC), while others were found exclusively in a single CC. Several PICIs persisted for decades.

Conclusion: PICIs are widespread among the pneumococci and demonstrate a structured population.

Appendix 3 – Poster presentation

Conference:

The 11th International Symposium on Pneumococci and Pneumococcal Diseases, in Melbourne, Australia in April 2018

Title:

Molecular epidemiology of 14 newly-discovered putative bacteriocins in *Streptococcus pneumoniae*

Authors:

Reza Rezaei Javan[1], Andries J. van Tonder[1], James P. King[1], Angela Brueggemann[1, 2]

Authors' affiliations:

[1]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[2]Department of Medicine, Imperial College London, London, United Kingdom

Abstract:

Background and Aims: Competition among bacterial members of the nasopharynx is believed to be mediated by bacteriocins: antimicrobial toxins produced by bacterial species to inhibit the growth of other closely-related bacteria. Their ability to kill bacteria makes bacteriocins attractive candidates for the development of novel antimicrobials. We recently identified 14 newly-discovered putative bacteriocins, tripling the number of known bacteriocins among pneumococci. We aimed to further assess these bacteriocins in the context of the pneumococcal population structure.

Methods: Using a large global and historical dataset of 571 pneumococci isolated from 1916-2009, we determined the prevalence, molecular epidemiology and co-

occurrence patterns of the pneumococcal bacteriocins.

Results: The number of different bacteriocins within each genome varied from 6 to 11 among pneumococci and certain combinations of bacteriocins were more frequently represented in the dataset. All bacteriocins detected more than once in the dataset were identified among pneumococci isolated over several decades and from a variety of different countries. We observed that many pneumococcal genomes harbour partial bacteriocin clusters that lack the toxin gene, while still retaining the immunity and/or the transporter genes. Whole-genome-based population analyses identified three genomic regions that are putative hotspots for the integration of bacteriocin clusters in the pneumococcal chromosome. Our data suggest a switching mechanism, whereby different bacteriocin clusters can replace one another via genetic recombination.

Conclusion: We revealed that not only do pneumococci possess a substantially greater and more varied array of bacteriocins than previously recognised, the bacteriocins, often in a particular combination, are associated with specific genetic lineages.

Appendix 4 – Oral presentation

Conference:

The 14th European Meeting on the Molecular Biology of the Pneumococcus, in Greifswald, Germany in June 2019.

Title:

Genus-wide analyses of the genomic diversity and population structure of streptococcal prophages

Authors:

Reza Rezaei Javan[1], Angela Brueggemann[1, 2]

Authors' affiliations:

[1]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[2]Department of Medicine, Imperial College London, London, United Kingdom

Abstract:

Prophages (viral genomes integrated within a host bacterial genome) are abundant within the bacterial world and are of interest because they can be reservoirs of genes that are involved in pathogenesis and antibiotic resistance. Satellite prophages are parasites of parasites that rely on the bacterial host and another helper prophage for survival. We developed PhageMiner, a powerful new user-friendly bioinformatics software for finding prophage in bacterial genomes. Using this tool, we analysed >1,300 genomes of 70 different *Streptococcus* species and identified an extraordinarily diverse collection of nearly 800 prophages and satellite prophages, the majority of which were discovered for the first time. We show that prophages and satellite prophages are a significant component of the genomes of clinically-relevant *Streptococcus* species. Contrary to the current dogma, we found

convincing evidence that cross-species transmission of prophages is not uncommon, which may have implications for the future of phage therapy. Our study also revealed a remarkable variability in prophage content among various streptococcal species as well as different genomes of the same species. We calculated the average prophage content for several of the major and widely-circulating pneumococcal genetic lineages. Furthermore, we performed a systematic investigation of prophage insertion sites and found that prophages are more frequently inserted adjacent to genes involved in information storage and processing. Overall, our findings suggest that prophages are likely to be influencing streptococcal biology and epidemiology to a much greater extent than previously appreciated.

Appendix 5 – Oral presentation

(Co-presented with Elisa Ramos-Sevillano)

Conference:

The 14th European Meeting on the Molecular Biology of the Pneumococcus, in Greifswald, Germany in June 2019.

Title:

A satellite prophage is involved in pneumococcal pneumonia and sepsis in a murine infection model

Authors:

Reza Rezaei Javan[1], Elisa Ramos-Sevillano[2], Asma Akter[3], Jeremy Brown[2], Angela Brueggemann[1,3]

Authors' affiliations:

[1]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[2]UCL Respiratory, Division of Medicine, University College London
[3]Department of Medicine, Imperial College London, London, United Kingdom.

Abstract:

Satellite prophages are a recently discovered type of prophages (viral genomes integrated within a host bacterial genome), which lack all the necessary genetic information to replicate on their own and are reliant on hijacking the machinery of another inducing 'helper' virus to replicate. These 'parasites of parasites' have been shown to be vectors for the spreading of toxin genes and other virulence factors, *e.g. Staphylococcus aureus* pathogenicity islands 1, which possesses the gene responsible for causing toxic shock syndrome. An *in-silico* investigation of pneumococcal satellite prophage genes led to the identification of a gene that is a homologue of the 'virulence-associated gene E' (*vapE*) in *Streptococcus suis*. To

investigate whether the *vapE* homologue in the pneumococcal satellite prophage is also associated with virulence, we performed *in vivo* studies using a murine pneumococcal infection model. Deletion mutant stains were constructed in a serotype 6B pneumococcal strain (BHN418) in which either *vapE* (Δ*vapE*) or the entire satellite prophage sequence (Δ*SpnSP38)* were knocked out. For each of the mutant strains, a competitive index (CI) was determined using a competitive infection experiment in a mouse model of pneumonia, which indicated a role for the satellite prophage and *vapE* in the establishment of pneumococcal pneumonia. Furthermore, in a sepsis model, the mice infected with the wildtype serotype 6B strain had significantly greater blood and spleen colony-forming units (CFU) than the Δ*SpnSP38* mutant, indicating that the satellite prophage is directly involved in pneumococcal virulence during bacterial dissemination in the systemic circulation. Overall, our findings suggest that satellite prophages may play a role in pneumococcal pathogenes.

Appendix 6 – Poster presentation

Conference:

The 11th International Symposium on Pneumococci and Pneumococcal Diseases, in Melbourne, Australia in April 2018.

Title:

Diverse pneumococcal strains drive a MAIT cell response through MR1-dependent and cytokine-driven pathways

Authors:

Ayako Kurioka[1], Bonnie van Wilgenburg[1], Reza Rezaei Javan[1], Ryan Hoyle[1], Andries J van Tonder[1], Caroline L Harrold[1], Tianqi Leng[1], Lauren J Howson[2], Dawn Shepherd[2], Vincenzo Cerundolo[2], Angela B Brueggemann, [1, 3], Paul Klenerman[1]

Authors' affiliations:

[1]Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[2]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, United Kingdom
[3]Department of Medicine, Imperial College London, London, United Kingdom.

Abstract:

Background and Aims: Mucosal Associated Invariant T (MAIT) cells represent an innate T cell population of emerging significance. These abundant cells can recognize ligands generated by microbes utilizing the riboflavin synthesis pathway, presented via the MHC-like molecule MR1 and binding of specific T cell receptors (TCR). They also possess a functional programme (shared by innate T cell populations expressing CD161) allowing microbial sensing in a cytokine-dependent, TCR-independent manner. We aimed to determine whether MAIT cells recognise pneumococci and assess the genomics and transcriptomics of the

pneumococcal riboflavin operon (Rib genes).

Methods: We examined the expression of Rib genes in pneumococci at rest and in response to metabolic stress using RNA sequencing, and linked this to MAIT cell activation *in vitro*. We analysed 571 diverse pneumococcal genomes from 39 countries dating back to 1916, and 824 genomes of 69 different non-pneumococcal *Streptococcus* species genomes for evidence of Rib gene sequences.

Results: We observed robust recognition of pneumococcal strains at rest and following stress, using both TCR-dependent and TCR-independent pathways. The pathway used was highly dependent on the antigen-presenting cell, but was maintained across a wide range of clinically-relevant strains. The riboflavin operon was highly conserved among pneumococci, and different versions of the riboflavin operon were also identified in other streptococcal species.

Conclusion: These data indicated an important functional relationship between MAIT cells and pneumococci, which may be tuned by local factors, including the metabolic state of the organism and the antigen-presenting cell that it encounters.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 27 August 2018
doi: 10.3389/fmicb.2018.02012

Genome Sequencing Reveals a Large and Diverse Repertoire of Antimicrobial Peptides

*Reza Rezaei Javan[1], Andries J. van Tonder[1], James P. King[1], Caroline L. Harrold[1] and Angela B. Brueggemann[1,2]**

[1] Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom, [2] Department of Medicine, Imperial College London, London, United Kingdom

Competition among bacterial members of the same ecological niche is mediated by bacteriocins: antimicrobial peptides produced by bacterial species to kill other bacteria. Bacteriocins are also promising candidates for novel antimicrobials. *Streptococcus pneumoniae* (the "pneumococcus") is a leading cause of morbidity and mortality worldwide and a frequent colonizer of the human nasopharynx. Here, 14 newly discovered bacteriocin gene clusters were identified among >6,200 pneumococcal genomes. The molecular epidemiology of the bacteriocin clusters was investigated using a large global and historical pneumococcal dataset dating from 1916. These analyses revealed extraordinary bacteriocin diversity among pneumococci and the majority of bacteriocin clusters were also found in other streptococcal species. Genomic hotspots for the integration of different bacteriocin gene clusters were discovered. Experimentally, bacteriocin genes were transcriptionally active when the pneumococcus was under stress and when two strains were co-cultured in broth. These findings reveal much more diversity among bacterial defense mechanisms than previously appreciated, which fundamentally broaden our understanding of bacteriocins relative to intraspecies and interspecies nasopharyngeal competition and bacterial population structure.

Keywords: bacteriocins, pneumococcus, genomics, antimicrobials, population biology

OPEN ACCESS

Edited by:
Santi M. Mandal,
Indian Institute of Technology Kharagpur, India

Reviewed by:
Piyush Baindara,
National Centre for Biological Sciences, India
Anirban Chakraborty,
The University of Texas Medical Branch at Galveston, United States

***Correspondence:**
Angela B. Brueggemann
angela.brueggemann@ndm.ox.ac.uk

Specialty section:
This article was submitted to Antimicrobials, Resistance and Chemotherapy, a section of the journal Frontiers in Microbiology

Received: *12 June 2018*
Accepted: *09 August 2018*
Published: *27 August 2018*

Citation:
Rezaei Javan R, van Tonder AJ, King JP, Harrold CL and Brueggemann AB (2018) Genome Sequencing Reveals a Large and Diverse Repertoire of Antimicrobial Peptides. Front. Microbiol. 9:2012. doi: 10.3389/fmicb.2018.02012

INTRODUCTION

Competition among bacterial members of the nasopharyngeal microbiome is mediated at least in part by bacteriocins, which are ribosomally synthesized antimicrobial peptides produced by bacterial species to inhibit the growth of other closely related bacteria. The producer strain also encodes an immunity protein to protect itself from its own bacteriocin (Dawid et al., 2006; Czaplewski et al., 2016). Bacteriocin production has been associated with more efficient colonization of a host by the producer strain, owing to the ability of these peptides to remove competitors (Dawid et al., 2006; Shak et al., 2013). Their ability to kill bacteria makes bacteriocins attractive potential candidates for the development of new antimicrobials. Several bacteriocins (e.g., nisin and pediocin PA-1) have already been commercialized and are widely used as food preservatives (Czaplewski et al., 2016).

From a genetic perspective, bacteriocins are found in gene clusters, whereby the genes involved in bacteriocin production, immunity and transport (exporting and processing the bacteriocin

peptide) are situated adjacent to each other in the bacterial genome. The *blp* (bacteriocin-like peptides) cluster is the best-characterized bacteriocin among pneumococci (Reichmann and Hakenbeck, 2000; Dawid et al., 2006; Shak et al., 2013; Bogaardt et al., 2015; Czaplewski et al., 2016). Previous work by our group showed that the *blp* cluster is ubiquitous among pneumococci recovered from the early 1900s onward and is highly diverse. We also identified a novel bacteriocin cluster that we named pneumocyclicin (Bogaardt et al., 2015). Five additional bacteriocins have been reported among pneumococci (Guiral et al., 2005; Begley et al., 2009; Hoover et al., 2015; Maricic et al., 2016; Kadam et al., 2017), however, their prevalence, genetic composition and molecular epidemiology in the context of the pneumococcal population are unknown.

Pneumococci are a leading cause of severe infections such as pneumonia, bacteraemia, and meningitis, and are among the most common causes of otitis media, sinusitis, and conjunctivitis. All age groups are susceptible to pneumococcal infection, but young children, the elderly, and immunocompromised persons are most at risk (Bogaert et al., 2004). Despite the use of antimicrobials and pneumococcal conjugate vaccines, the pneumococcus remains a major global health problem, causing approximately 14.5 million cases of serious disease and 826,000 deaths annually in children <5 years of age (O'Brien et al., 2009). Antimicrobial-resistant pneumococci have been a serious and increasing concern for several decades (Klugman, 1990; Doern et al., 2001; Tadesse et al., 2017). The pneumococcus is now considered to be a "priority pathogen" – defined as antimicrobial-resistant bacteria that pose the greatest threat to global health–by the World Health Organization, 2017.

The pneumococcus is normally an asymptomatic colonizer of the nasopharynx in healthy young children; however, colonization is also the initial stage of the disease process. The polysaccharide capsule is the main virulence factor of the pneumococcus: nearly 100 distinct capsular antigenic types (serotypes) have been described and certain serotypes are predominantly associated with disease whilst others are largely associated with nasopharyngeal colonization (Brueggemann et al., 2003). Pneumococci frequently co-colonize with other pneumococci and non-pneumococcal bacterial species, and intraspecies and interspecies competition can influence colonization dynamics, strain prevalence, serotype distributions and consequently, the potential for disease progression (Shak et al., 2013).

The abundance of whole genome sequence data expedites the detection and genetic analysis of bacteriocin clusters. Here, we report 14 newly discovered pneumococcal bacteriocin clusters and describe the molecular epidemiology of all the currently known bacteriocins within a collection of pneumococci isolated over the past 90 years. We also found that identical or highly similar versions of the pneumococcal bacteriocin gene clusters could be found in other unrelated streptococcal species. We provide transcriptomic evidence that multiple bacteriocin clusters were induced in response to external stress and in response to competition for space and nutrients in broth co-culture.

MATERIALS AND METHODS

Genome Mining for Bacteriocin Clusters

In total, 6,244 assembled pneumococcal genomes from studies previously published by us and others (Croucher et al., 2011, 2013a,b; Chewapreecha et al., 2014; van Tonder et al., 2014, 2015; Gladstone et al., 2015; Brueggemann et al., 2017) as listed in the **Supplementary Table S1** were screened for the presence of bacteriocin clusters using a variety of bioinformatic tools and databases, including antiSMASH (Weber et al., 2015) (to identify putative gene clusters that encode microbial secondary metabolites), BACTIBASE (Hammami et al., 2010) and BAGEL (van Heel et al., 2013) (to screen our genome sequences for homology to known bacteriocin genes from a diverse range of bacterial species) and InterProScan (Jones et al., 2014) (to assess the putative function of encoded proteins and identify protein domains and key sites). An in-house pipeline was developed to automate part of this process. Predicted gene clusters from each of the database outputs were examined manually and further scrutinized using extensive BLAST searches. No single program or sequence alignment was sufficient to definitively identify every bacteriocin cluster in its entirety, but rather a combination of tools was used to be confident of the identification of each bacteriocin gene cluster.

Analyses of the Putative Bacteriocin Genes

Putative bacteriocin genes were annotated using homology to other known bacteriocin genes (**Supplementary Figure S1**) in all available databases mentioned above, as well as structure-based searches. Protein domains were examined using the Conserved Domain search feature at NCBI (Marchler-Bauer et al., 2010). Genes of interest were screened against the STRING database (Szklarczyk et al., 2015) to search for any previously reported relationship to other genes. Multiple sequence alignments of the genes in streptococcin clusters were performed in Geneious version 9.1 (Biomatters Ltd.) using the ClustalW algorithm (Larkin et al., 2007) with default parameters (Gap open cost = 15, Gap extend cost = 6.66). The multiple sequence alignment output was used within the Geneious environment to calculate percentage identity matrices. The figures of the coding regions of the bacteriocin clusters and their flanking genes were generated in Geneious and edited using Inkscape[1].

Classification and Nomenclature of Bacteriocin Clusters

Putative bacteriocin clusters were classified based on their predicted biosynthetic machinery and structural features (Arnison et al., 2013) and their corresponding genes were designated per the standards of nomenclature for bacteriocins (**Supplementary Table S2**). BLAST searches in the NCBI database revealed that the majority of these clusters are also present in other closely related streptococci (**Supplementary Table S3**); therefore, the clusters were named

[1]http://inkscape.org

using the prefix "strepto-" followed by an abbreviation of their bacteriocin class: "streptococcins" for those that were lactococcin 972-like, "streptolancidins" for lanthipeptides, "streptocyclicins" for the head-to-tail cyclized peptides, "streptosactins" for the sactipeptides, and "streptolassins" for the lassopeptide group of bacteriocins. When more than one bacteriocin cluster from the same class was present, they were lettered alphabetically by the order of their discovery in our analyses.

Molecular Epidemiology of the Bacteriocin Clusters

We compiled a global and historical dataset ($n = 571$) by selecting a diverse set of pneumococcal genomes isolated between 1916 and 2009 from people of all ages residing in 39 different countries. Pneumococci from both carriage and disease, 88 different serotypes and 99 different clonal complexes were represented in this dataset (**Supplementary Table S4**). Genomes and their associated metadata were stored in a BIGSdb database (Jolley and Maiden, 2010). The BIGSdb database platform was used to generate a presence/absence matrix of all the known bacteriocin genes in all 571 genomes. (Note that there were two small frameshifted gene remnants of both streptolancidins C and E present in some genomes, but these were not analyzed further in this study.) Using this matrix (79,940 genes) as an input, an in-house python script was developed (available upon request) to calculate the prevalence, molecular epidemiology, and co-occurrence patterns of all the bacteriocin clusters in the study dataset.

Construction of the Core Genome Phylogenetic Tree

All genomes in the study dataset were annotated using the Prokka prokaryotic annotation pipeline (Seemann, 2014). The annotation files were input into Roary (Page et al., 2015) and clustered using a sequence identity threshold of 90%. Genes present in every genome were selected using a core genome threshold of 100% and were aligned using Roary. FastTreeMP (Price et al., 2010) was used to construct the phylogenetic tree using generalized time-reversible model (parameters: FastTreeMP -nt -gtr). ClonalFrameML (Didelot and Wilson, 2015) was then applied to reconstruct the phylogenetic tree adjusted for recombination. The tree was annotated using iTOL (Letunic and Bork, 2016) and Inkscape.

Investigation of Bacteriocin Cluster Insertion Sites

Bacteriocin cluster sequences were used as queries to BLAST against genomes in the study dataset using the custom BLAST implemented in Geneious. The matching region plus additional flanking regions were visualized using the query-centric alignment feature within the Geneious environment. Regions of DNA with different bacteriocin clusters but identical flanking genes among different isolates were identified and further investigated using the Artemis Comparison Tool (ACT) (Carver et al., 2005). Linear comparison figures were generated using Geneious, ACT, and Inkscape.

RNA Sequencing Analyses

In the first experiment, total bacterial RNA sequencing was performed on RNA extracted from pneumococcal strain 2/2 grown at a higher incubation temperature than normal (40°C vs. 37°C) to induce a bacterial stress response (Kurioka et al., 2017). Pneumococci were grown in brain–heart infusion broth for 6 h and RNA extractions were performed on samples from five time points (2, 3, 4, 5, and 6 h of incubation) using the Promega Maxwell® 16 Instrument and LEV simplyRNA Cells purification kit, following the manufacturer's protocol. Extracted RNA samples were sent to the Oxford Genomics Centre for sequencing on the Illumina platform (NCBI GEO accession number GSE103778). The sequenced forward and reverse reads were paired and mapped onto the annotated pneumococcal strain 2/2 genome using Bowtie2 (Langmead and Salzberg, 2012) with the highest sensitivity option. Differential gene expression was assessed in Geneious using the DESeq (Anders and Huber, 2010) method. Genes with an adjusted P-value <0.05 were deemed to be differentially expressed.

In a second experiment, pneumococcal reference strains PMEN-3 (Spain^{9V}-3) and PMEN-6 (Hungary^{19A}-6) were grown together in brain–heart infusion broth for 6 h. The controls were prepared by growing each strain individually in brain–heart infusion broth for 6 h. Total bacterial RNA sequencing was performed on RNA extracted from broth cultures at 2, 3, 4, 5, and 6 h after incubation using the procedures described above (NCBI GEO accession number GSE110750).

A pseudo-reference genome was constructed using Bowtie2, Velvet (Zerbino and Birney, 2008), and MeDuSa (Bosi et al., 2015) to sort genes into three categories: those unique to PMEN-3, those unique to PMEN-6, and those shared between the two (**Supplementary Figure S2**). The RNA sequencing reads for each individual control strain at all time points were pooled *in silico*. Data from all time points were combined to minimize variability caused by different growth rates of strains. Data from all time points sampled in the broth culture of PMEN-3 + PMEN-6 were also combined *in silico* and mapped to the pseudo-reference genome using Bowtie2 with the highest sensitivity option. Differential expression analyses were performed using the DESeq method by comparing sequence reads generated when strains were co-cultured to those from when strains were grown individually (**Supplementary Figure S2**).

Theoretically, the control contained double the amount of reads in comparison to the *in vivo* competition experiment due to being compiled *in silico* from two sets of samples. While this meant that the downregulation of genes could not reliably be assessed, one could have confidence that upregulated genes were differentially expressed (since expression levels must exceed that of the combined controls). However, this approach can provide only relative and not absolute values, and true fold-change ratios for the upregulated genes are most likely underestimated using this method.

RESULTS

Genome Mining Triples the Number of Known Bacteriocins Among Pneumococci

Our investigation of a large and diverse dataset of 571 historical and modern pneumococcal genomes resulted in the identification of 14 newly discovered bacteriocin clusters, increasing the number of known bacteriocins in this species to 21 (**Table 1**). We identified several clusters similar to lactococcin 972, sactipeptide and lassopeptide bacteriocins (Martińez et al., 1999; Arnison et al., 2013; Letzel et al., 2014), which were hitherto not known to be harbored by the pneumococcus (**Table 1** and **Supplementary Table S2**). We subsequently expanded our search for bacteriocins to a much larger dataset of 5,673 published pneumococcal genomes, but no additional bacteriocins were found; thus, the detailed description of the bacteriocins here is restricted to those identified in the dataset of 571 genomes.

BLAST searches of the NCBI database of non-redundant nucleotide sequences revealed that the majority of these clusters were not exclusive to pneumococcus, but that identical or similar versions were present in other streptococci (**Supplementary Table S3**). Therefore, we used the prefix "strepto" and named each cluster according to its type of bacteriocin. Among the putative bacteriocins, five were similar to lactococcin 972 (Martińez et al., 1999) and we called these streptococcins (**Figure 1** and **Supplementary Figure S1**). Despite gene synteny among the streptococcins (**Figure 1A**), nucleotide sequence similarity was low: bacteriocin genes, 37–59%; immunity genes, 27–48%; and transporter genes, 49–63% (**Figure 1B**). Different streptococcins were found in different, but consistent, locations within the bacterial chromosome (**Figures 1A,C**). Eleven different streptolancidins, one streptosactin, one streptolassin, and one streptocyclicin [what we previously called pneumocyclicin (Bogaardt et al., 2015)] were also identified among pneumococcal genomes, as shown in **Figure 2** (Arnison et al., 2013 and **Supplementary Figure S1**). The bacteriocin clusters were commonly flanked by genes predicted to encode CAAX proteins (possibly involved in self-immunity from bacteriocins; Kjos et al., 2010), Rgg and PlcR transcriptional regulators (implicated in bacterial quorum sensing; Perez-Pascual et al., 2016), transporters, lipoproteins and mobilization proteins.

Further support for interspecies exchange of bacteriocin gene clusters was provided by the guanine (G) and cytosine (C) content of the bacteriocin clusters. The average GC-content of all pneumococcal genomes in this dataset was 39.6%, whereas the range of values for different bacteriocin groups was as follows: streptococcins, 36.3–42.4%; streptolancidins subset 1, 31.2–33.4%; streptolassin, 31.2%; streptolancidins subset 2, 28.9–29.6%; streptocyclicin, 27.0%; and streptosactin, 25.5% (**Supplementary Figure S3**). For comparison, the GC-content of other non-pneumococcal *Streptococcus* species in another study ranged between 33.2 and 44.6% (Kurioka et al., 2017 and **Supplementary Figure S3**).

Extraordinary Bacteriocin Diversity Within a Globally Distributed Pneumococcal Dataset Dating From 1916

We further assessed these bacteriocins in the context of the pneumococcal population structure. The study dataset consisted of a diverse collection of 571 pneumococci isolated between 1916 and 2009 from patients and healthy individuals of all ages residing in 39 different countries across six continents. Eighty-eight pneumococcal serotypes and 99 different clonal complexes were represented (**Table 1** and **Supplementary Table S4**). All bacteriocins detected more than once in the dataset were identified among pneumococci isolated over several decades and from a variety of different countries (**Table 1**). Some bacteriocins were present in all pneumococci, whilst others were limited to specific clonal complexes (genetic lineages). Bacteriocin clusters missing one or more genes relative to the largest clearly defined cluster in the group were defined as partial clusters. The percentages of partial and complete clusters varied between different bacteriocins, e.g., streptococcin E was present in all apart from one pneumococcal genome but as a partial cluster in the majority of genomes, while streptococcin C was found in all genomes as a complete cluster (**Table 1** and **Figure 3A**). We constructed a core genome phylogenetic tree of all isolates and labeled each genome according to the presence or absence of each bacteriocin cluster (**Supplementary Figure S4**). Overall, we found that the number of bacteriocins within each genome varied from 5 to 11 per genome and that certain combinations of bacteriocins were well represented in the dataset (**Figures 3B,C** and **Supplementary Figure S4**).

Due to their relative simplicity (containing only three genes), we chose streptococcins as models for further investigating the pattern of missing genes in partial clusters (**Figure 3A**). Scrutinizing these clusters revealed that the majority of the partial clusters lack the bacteriocin gene, while still retaining the immunity and/or the transporter genes. This could support the general idea of a "cheater" phenotype, whereby the immunity and transporter genes are conserved to protect the pneumococcus from neighboring bacteria that express the bacteriocin, but the cheater strain does not bear the cost of producing the bacteriocin (Brown et al., 2009; Son et al., 2011).

Multiple Hotspots for the Integration of Bacteriocin Clusters in the Pneumococcal Genome

By conducting a population genomics-based assessment of the bacteriocin cluster insertion sites, we identified three genomic regions that are putative hotspots for the integration of bacteriocin clusters in the pneumococcal chromosome. These bacteriocin cluster hotspots (BCHs) are specific locations in the genome where different bacteriocin clusters were found in different pneumococci (**Figures 4A–C**). This suggested a switching mechanism whereby different clusters can replace one another via homologous recombination. Up to three different bacteriocin clusters were found to be associated with a single

TABLE 1 | Bacteriocin clusters identified among a dataset of 571 pneumococci recovered since 1916 from patients of all ages residing in 39 different countries.

Bacteriocin	Genome(s)			Year(s) of isolation	Countries (*n*)	CCs[a] (*n*)	Serotypes (*n*)
	Complete	Partial	Total				
Streptococcin A	404 (70.8%)	1 (0.2%)	405 (70.9%)	1916–2009	36	80	76
Streptococcin B	414 (72.5%)	157 (27.5%)	571 (100%)	1916–2009	39	99	88
Streptococcin C	571 (100%)	0 (0.0%)	571 (100%)	1916–2009	39	99	88
Streptococcin D	6 (1.1%)	0 (0.0%)	6 (1.1%)	1968–2005	5	3	4
Streptococcin E	93 (16.3%)	477 (83.5%)	570 (99.8%)	1916–2009	39	98	88
Streptolancidin A[b]	1 (0.2%)	1 (0.2%)	2 (0.4%)	1972–2006	2	2	2
Streptolancidin B[c]	49 (8.6%)	44 (7.7%)	93 (16.3%)	1939–2006	16	11	10
Streptolancidin C	96 (16.8%)	0 (0.0%)	96 (16.8%)	1937–2006	13	15	12
Streptolancidin D	49 (8.6%)	0 (0.0%)	49 (8.6%)	1938–2006	13	13	11
Streptolancidin E[d]	12 (2.1%)	156 (27.3%)	168 (29.4%)	1937–2009	26	21	38
Streptolancidin F	23 (4.0%)	0 (0.0%)	23 (4.0%)	1937–2006	7	4	11
Streptolancidin G[e]	190 (33.3%)	9 (1.6%)	199 (34.9%)	1916–2009	25	38	46
Streptolancidin H	1 (0.2%)	0 (0.0%)	1 (0.2%)	2006–2008	1	0	1
Streptolancidin I	1 (0.2%)	0 (0.0%)	1 (0.2%)	2009	1	1	1
Streptolancidin J	185 (32.4%)	195 (34.2%)	380 (66.5%)	1916–2009	33	64	74
Streptolancidin K	1 (0.2%)	10 (1.8%)	11 (1.9%)	1943–2009	4	2	3
Streptocyclicin[f]	209 (36.6%)	0 (0.0%)	209 (36.6%)	1937–2009	20	46	59
Streptolassin	20 (3.5%)	0 (0.0%)	20 (3.5%)	1939–1996	8	7	10
Streptosactin	1 (0.2%)	0 (0.0%)	1 (0.2)	2009	1	1	1
cib[g]	557 (97.5%)	0 (0.0%)	557 (97.5%)	1916–2009	39	97	88
blp[h]	N/A	N/A	571 (100%)	1916–2009	39	99	88

[a]*CC, clonal complex (genetic lineage). Singletons (single genotypes with no closely related variant) were excluded from the count. Synonym(s) for the previously identified bacteriocins are as follows:* [b]*Pneumolancidin (Maricic et al., 2016) and salivaricin E (Walker et al., 2016);* [c]*lcpAMT (Kadam et al., 2017) and ICESp23FST81 lantibiotic (Croucher et al., 2008);* [d]*SP23-BS72 lantibiotic (Begley et al., 2009);* [e]*Phr lantibiotic (Hoover et al., 2015);* [f]*Pneumocyclicin (Bogaardt et al., 2015);* [g]*cibAB (Guiral et al., 2005);* [h]*spi and pnc (Bogaardt et al., 2015).*

BCH (**Figure 4C**). The acquisition of streptolancidin G appears to have rendered the streptococcin E partial by replacing the bacteriocin and part of its immunity gene (**Figure 4A**), which is in accordance with the fact that streptolancidin G could not be found in genomes that harbored a complete streptococcin E cluster (**Supplementary Figure S4**). Nonetheless, the remnant genes of the streptococcin E partial clusters were conserved in samples collected over many decades (**Figure 3A**).

Bacteriocin Gene Expression in Response to Heat Stress and Strain Competition

To explore whether the bacteriocin clusters were transcriptionally active, we analyzed a whole-genome RNA sequencing dataset from a broth culture of pneumococcal strain 2/2, which was incubated at a higher than normal temperature (40 vs. 37°C) to induce a bacterial stress response (NCBI GEO accession number GSE103778; Kurioka et al., 2017). Multiple genes in the bacteriocin clusters were differentially expressed compared to the control over several time points, indicating that many of these bacteriocin genes were transcribed in response to external stress (**Figure 5**).

Notably, there were two bacteriocin clusters (streptolancidin J and streptococcin E) that were missing genes and yet the remaining genes were clearly being upregulated. Moreover, the timing of gene expression varied across bacteriocin clusters, e.g., genes within a cluster were induced in a specific pattern, whilst at any particular time point during the sampling period several genes in different clusters were upregulated simultaneously. The location of and variation within the promoter regions, and the regulatory governance over which bacteriocin genes are expressed at any given point during bacterial growth remain to be determined.

A second whole-genome RNA sequencing experiment was designed to test whether bacteriocin genes were transcribed when two genetically different reference strains, PMEN-3 and PMEN-6, were cultured together in the same broth culture using standard incubation conditions. These strains were competing for space and nutrients as the incubation time progressed and cell density increased. Samples were taken for RNA sequencing at multiple time points and sequenced. Sequencing reads from all time points were combined *in silico*. The sequencing reads were mapped to a pseudo-reference genome that was constructed to include the genes unique to PMEN3 and PMEN6 plus the genes shared between both strains.

Pneumococcal genes that were significantly upregulated included many genes that would be expected to be expressed during growth (e.g., metabolic genes), in addition to the significant upregulation of 29 bacteriocin genes (**Figure 6** and **Supplementary Table S5**). Many of the bacteriocin genes were present in highly similar allelic versions in both PMEN-3 and

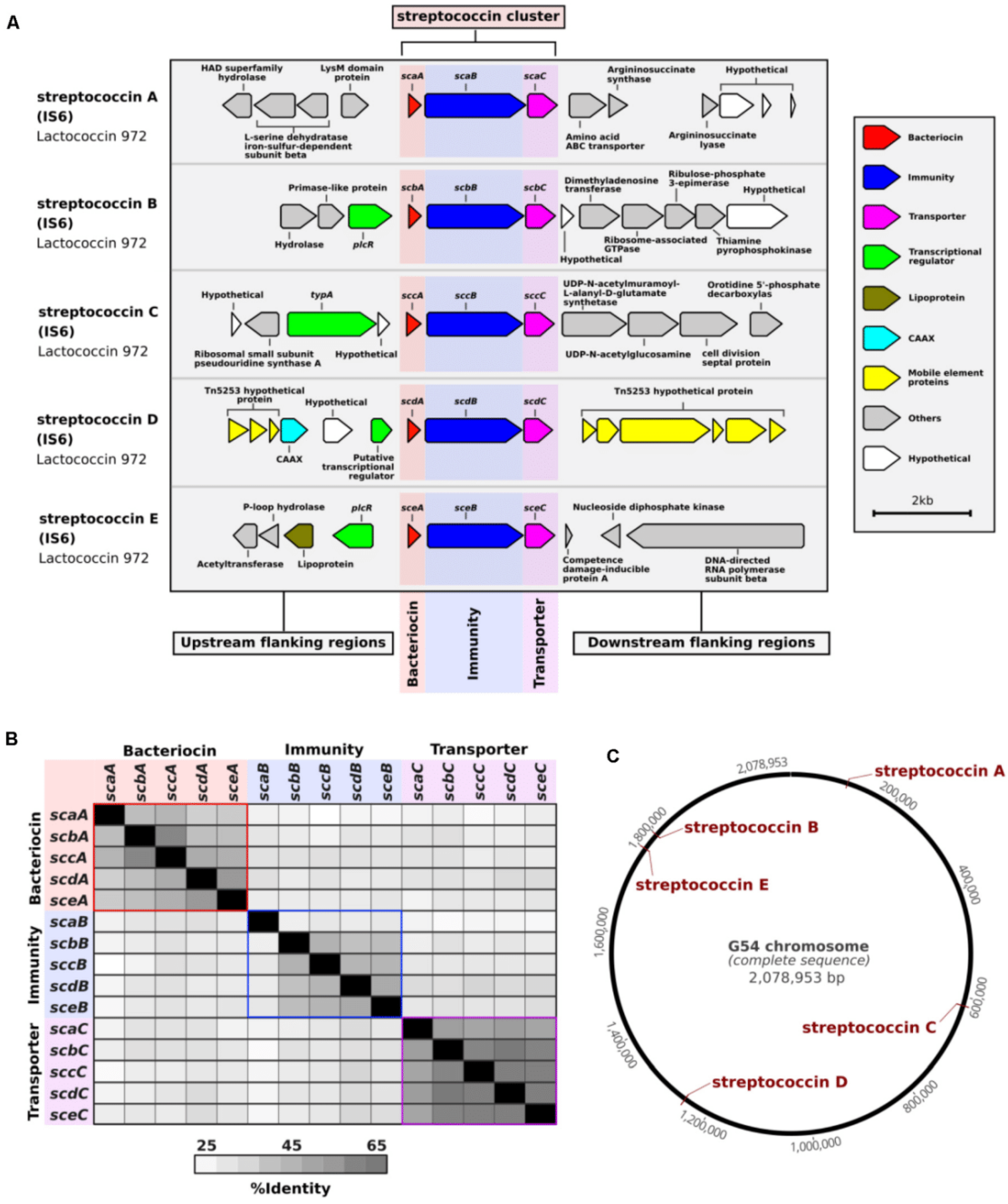


FIGURE 1 | Streptococcin bacteriocins discovered among a large collection of pneumococcal genomes. **(A)** A schematic representation of each streptococcin cluster and their flanking regions is depicted. The coding regions were derived from the genome of pneumococcal strain IS6, which harbors five different complete streptococcin clusters. **(B)** A distance matrix of the nucleotide sequence identity shared between genes of different streptococcin clusters found in pneumococcal strain IS6 is shown. **(C)** Schematic of the finished genome of pneumococcal strain G54 with the locations of the five streptococcin clusters highlighted in red.

PMEN-6, thus it was not possible to determine with confidence whether both versions were upregulated or whether one strain overexpressed a similar gene. However, there were three genes within the bacteriocin clusters that were unique: PMEN-3 significantly upregulated the bacteriocin gene of streptococcin A (*scaA*) and a putative immunity gene (*pncM*) from the *blp* bacteriocin cluster, and PMEN-6 significantly upregulated the bacteriocin gene of streptococcin E (*sceA*).

Interestingly, other genes that were significantly upregulated were genes associated with unique prophages present in each

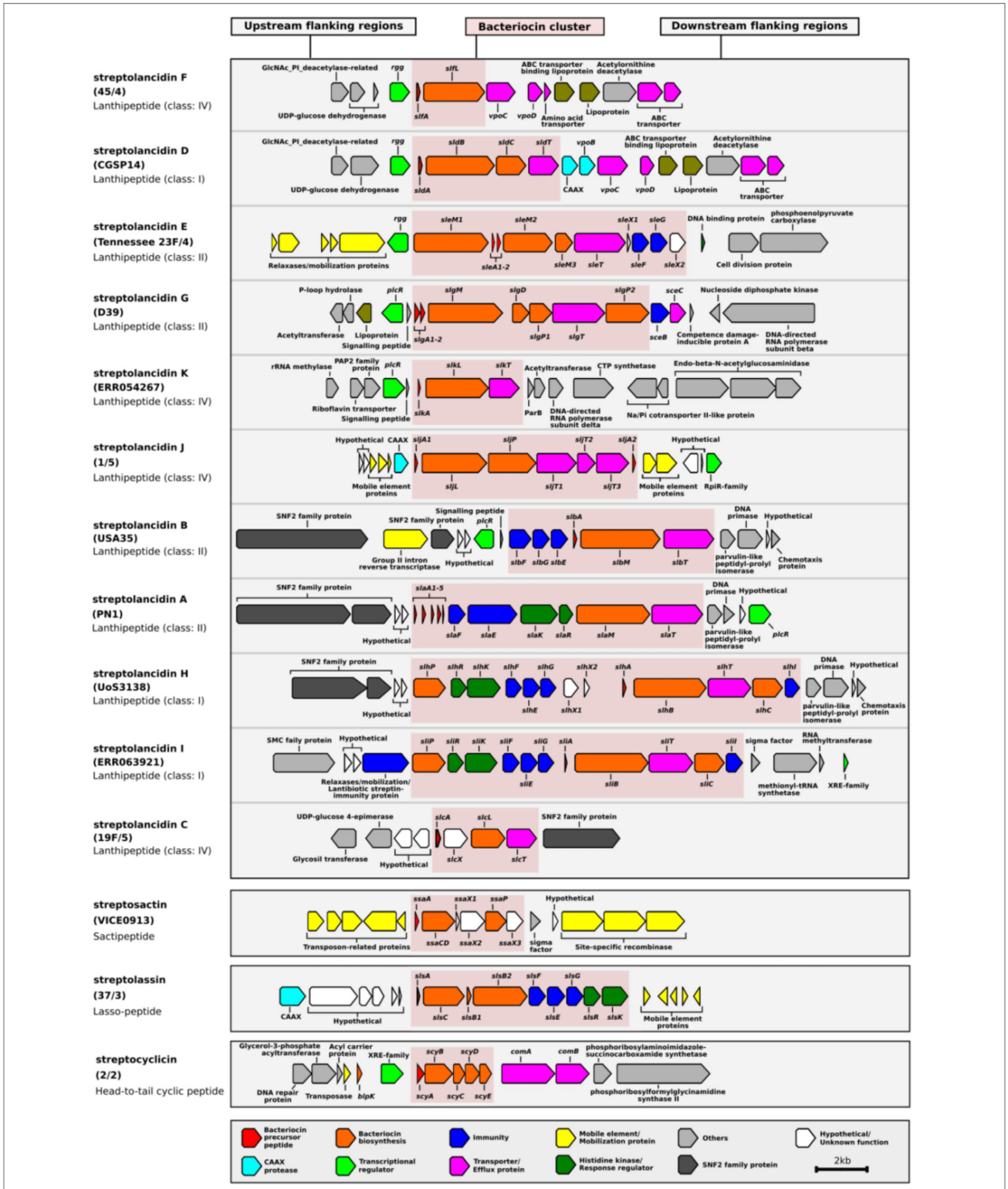
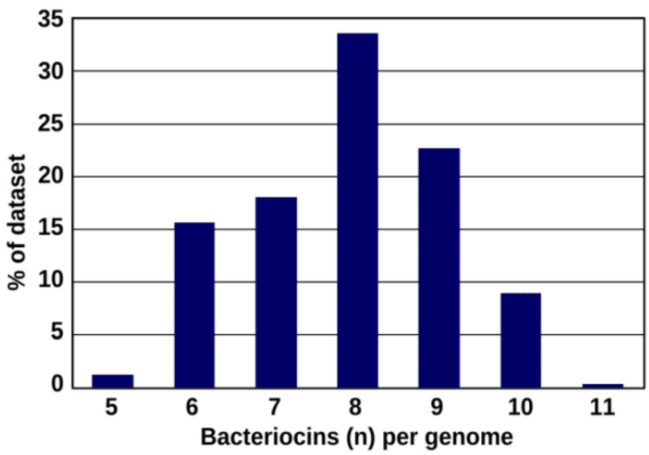


FIGURE 2 | Genetic composition of streptolancidin, streptocyclicin, streptosactin, and streptolassin bacteriocins in the pneumococcal genomes. A schematic representation of 11 different streptolancidins, one streptocyclicin, one streptosactin, and one streptolassin identified among pneumococcal genomes is depicted. The names of the isolates containing each bacteriocin cluster are given in brackets and the class of each bacteriocin cluster is given underneath the names of isolates.

A

	streptococcin A	*streptococcin B*	*streptococcin C*	*streptococcin D*	*streptococcin E*
ABC	404	414	571	6	93
AB	0	1	0	0	0
AC	0	0	0	0	0
BC	1	156	0	0	476
A	0	0	0	0	0
B	0	0	0	0	0
C	0	0	0	0	1
Total	**405**	**571**	**571**	**6**	**570**

B



C

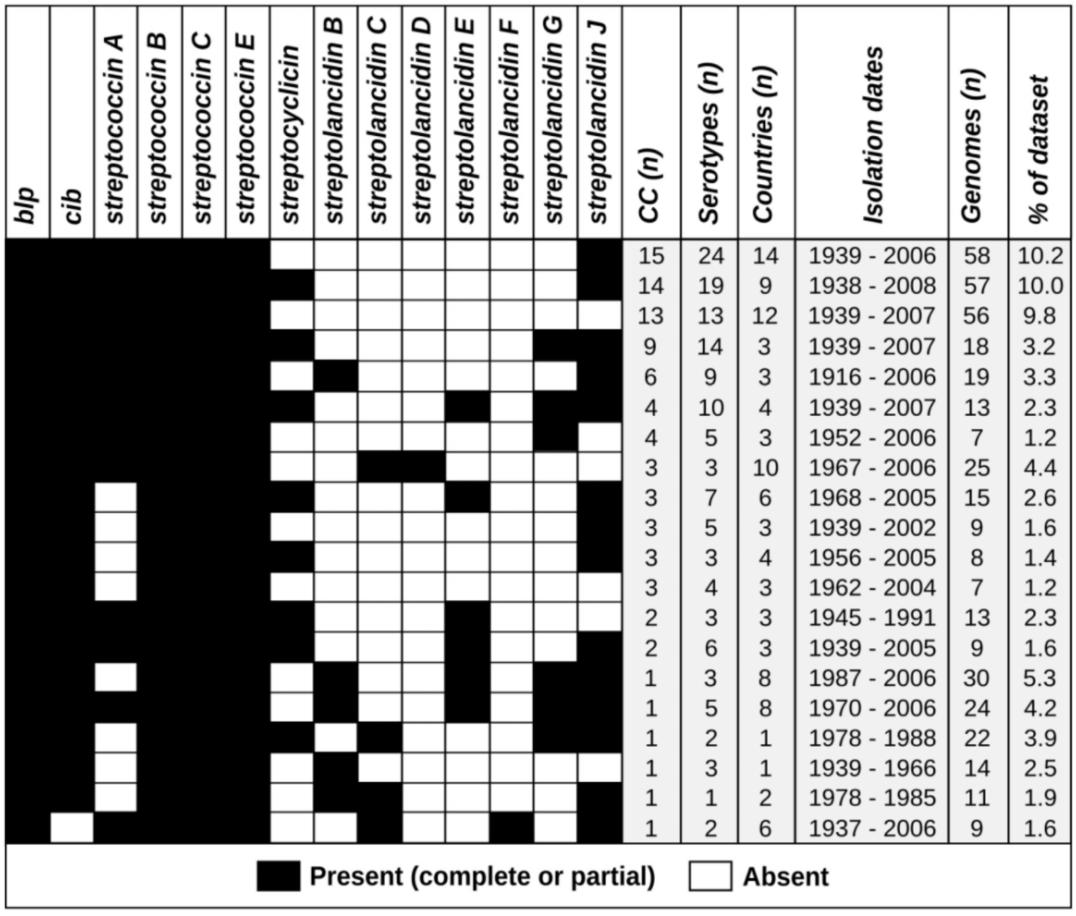


FIGURE 3 | The prevalence, molecular epidemiology and co-occurrence patterns of bacteriocins within the global pneumococcal population. **(A)** The patterns of gene presence (column 1: A, bacteriocin; B, immunity; C, transporter) and frequency of those patterns (indicated by numbers) among streptococcins A–E within the study dataset are shown. **(B)** The number of bacteriocins (of any type) per genome among the 571 pneumococcal genomes is summarized. **(C)** Bacteriocin combinations found in >1 percent of the genomes in the dataset are presented as black and white boxes indicating the presence or absence, respectively, of each bacteriocin. The epidemiological characteristics of the pneumococci that possessed each combination of bacteriocins are summarized in the columns to the right of the black and white boxes.

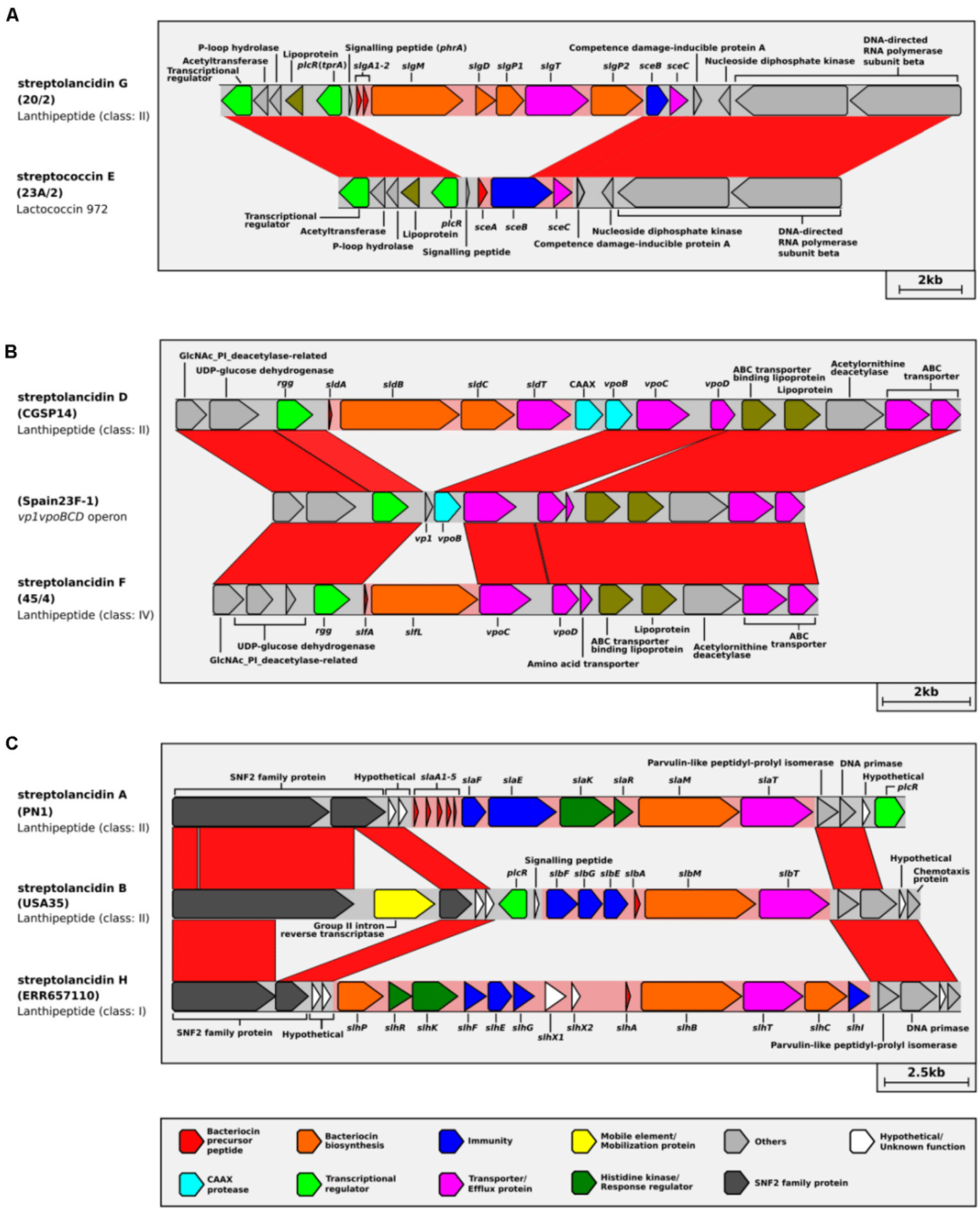


FIGURE 4 | Whole-genome-based population analysis reveals evidence for bacteriocin switching. Bacteriocin cluster hotspots (BCHs) were defined as regions of DNA where different bacteriocin clusters with identical flanking genes were found among pneumococcal genomes. Linear comparisons of **(A)** BCH-1, **(B)** BCH-2, and **(C)** BCH-3 are shown. The isolate names are given in brackets. The class of each bacteriocin cluster is given underneath the isolate name.

of the PMEN strains. We have recently demonstrated that prophage genes are ubiquitous among pneumococcal genomes, but to what extent prophages are influencing pneumococcal biology and perhaps competition between strains is not yet understood (Brueggemann et al., 2017). Overall, the data from this experiment demonstrate proof-of-principle that such methodology can be used to test for the differential expression of key bacterial genes within a competitive environment.

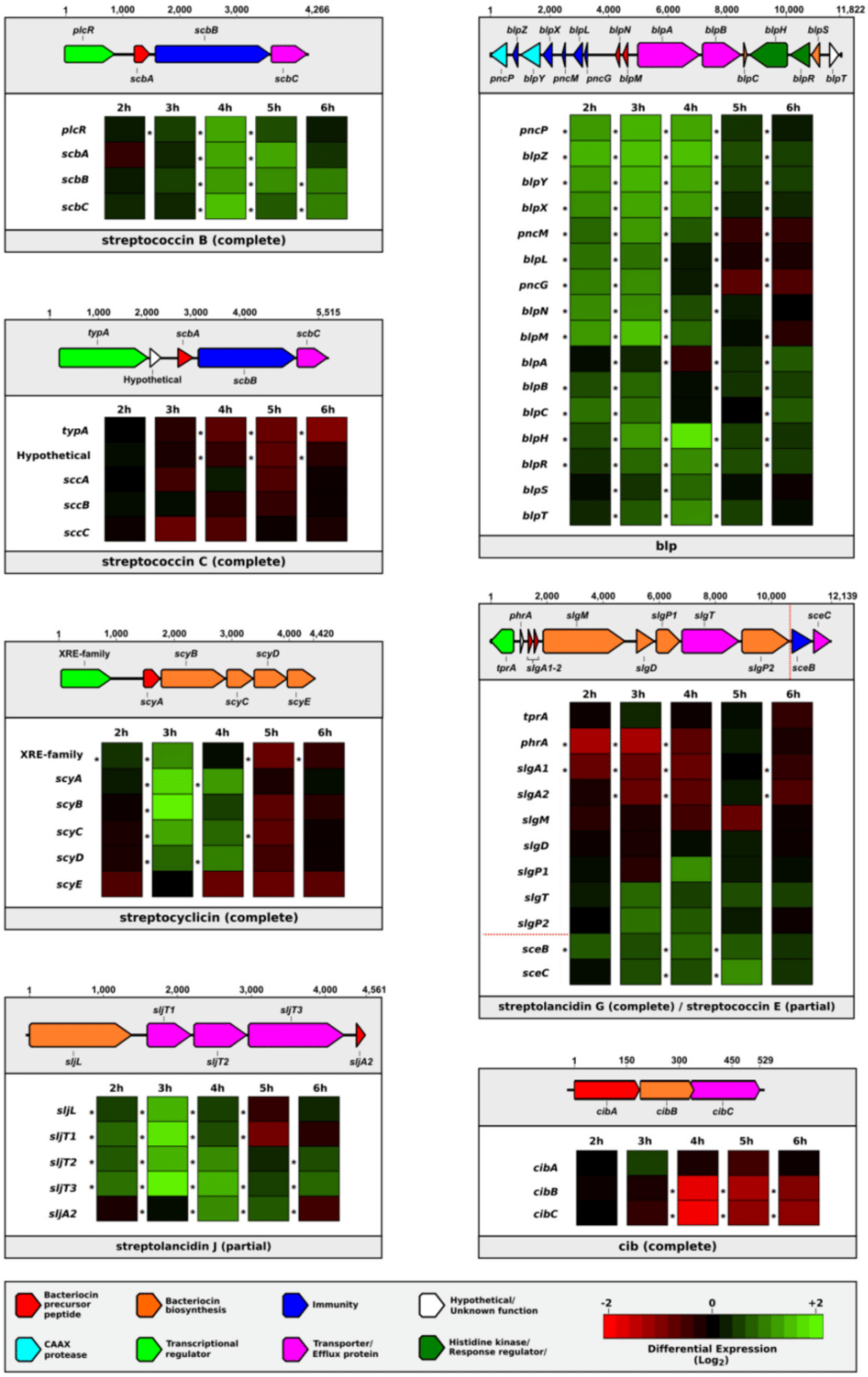


FIGURE 5 | Dynamic changes in the expression of bacteriocin genes in response to bacterial stress. The differential expression levels of bacteriocin gene clusters found in pneumococcal strain 2/2 when incubated at 40°C vs. 37°C are displayed in individual boxes. A schematic representation of each bacteriocin cluster is provided above each box. Genes are represented by rows and differential expression levels at different time points are indicated in columns. An asterisk to the left of a cell indicates a statistically significant differential expression level ($p < 0.05$).

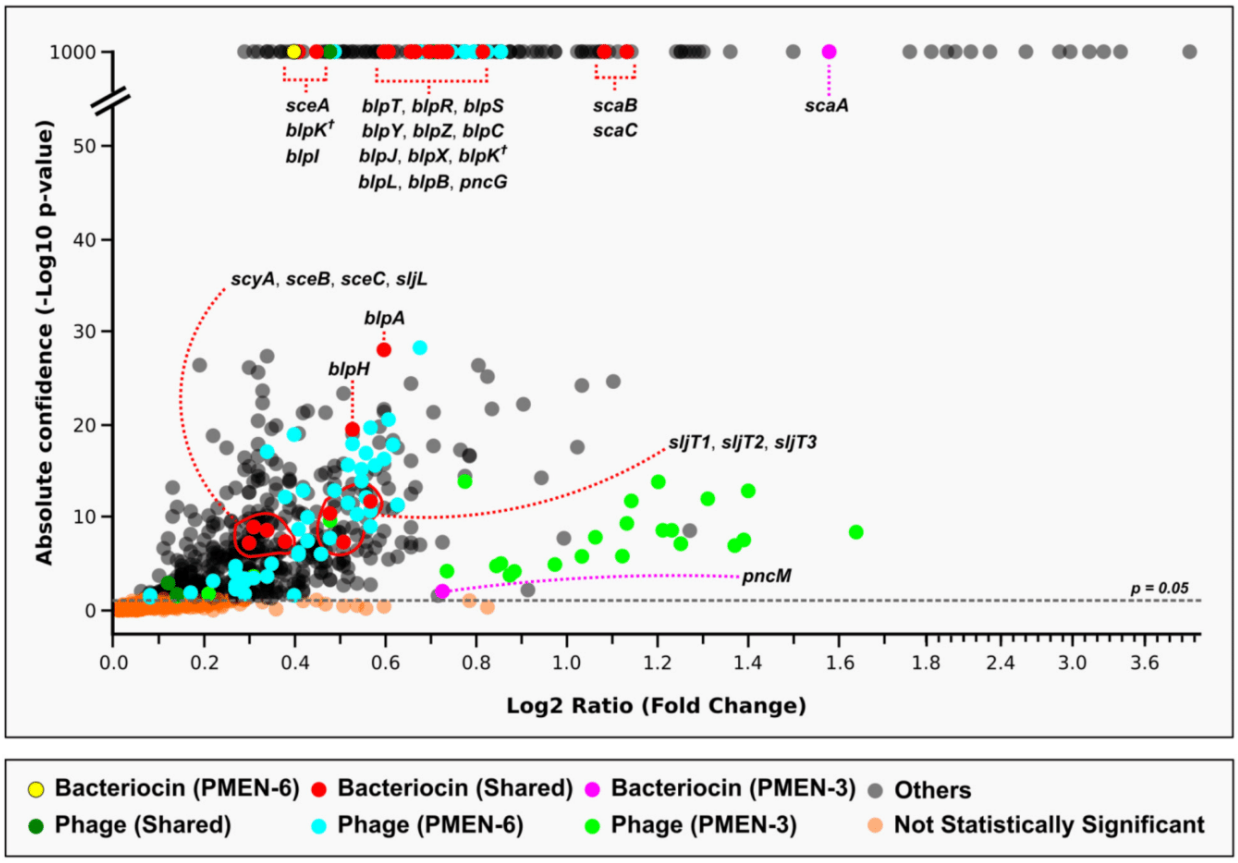


FIGURE 6 | Evidence for the upregulation of bacteriocin genes when two reference strains, PMEN-3 and PMEN-6, were co-cultured in broth media. Only genes that were significantly upregulated compared to the controls (strains cultured individually) are shown in the figure. Genes of interest that were unique to each reference strain and those shared by both strains are marked in different colors. Two copies of *blpK* were found in different locations (one in the *blp* cluster as expected and one elsewhere in the genome) in both genomes and are marked here with a cross. A full list of the genes depicted here and their sequences is found in the **Supplementary Table S5**.

DISCUSSION

A clear understanding of the role bacteriocins play in pneumococcal biology is central to understanding microbial interactions within the ecological niche (the nasopharynx). The importance of intraspecies competition to pneumococcal ecology is reflected in the changes in prevalence of different pneumococcal serotypes and genotypes in the nasopharynx over time, and understanding competition dynamics is important in the context of understanding vaccine impact (Garcia-Rodriguez and Fresnadillo Martínez, 2002; Crook et al., 2004; Auranen et al., 2009). Pneumococcal conjugate vaccines are disruptive to the pneumococcal population structure and alter the composition of microbes competing for space and nutrients in the nasopharynx. The effects of this disruption are not yet fully understood but can lead to increased disease in human populations (Huang et al., 2005; Aguiar et al., 2010; Kaplan et al., 2010).

We revealed here that not only do pneumococci possess a substantially greater and more varied array of bacteriocins than previously recognized, the bacteriocins (often in a particular combination) are associated with specific genetic lineages. This is fundamental, as it provides the framework on which to investigate the mechanisms underpinning specific bacteriocin-pneumococcus combinations, particularly among epidemiologically successful genetic lineages, and the activity of specific bacteriocin and immunity genes. RNA sequencing clearly demonstrated that bacteriocin genes were transcriptionally active when the pneumococcus was under stress or in competition with another strain during bacterial co-culture. The *in vivo* production of bacteriocin gene products and their functional activities is under further investigation.

Interestingly, we identified several specific locations in the pneumococcal genome that harbored different bacteriocin clusters, which suggested that recombination events had occurred at these locations and resulted in a switching of bacteriocin clusters. The profound impact of recombination on the pneumococcal genome has been described for >25 years and recombination events are well documented in other locations within the pneumococcal genome, most commonly at the capsular polysaccharide locus (conferring a change of serotype)

and at penicillin binding protein genes (conferring penicillin resistance if the proteins are altered) (Coffey et al., 1991; Laible et al., 1991; Maynard Smith et al., 1991; Golubchik et al., 2012; Wyres et al., 2012). Similarly, the DpnI, DpnII, and DpnIII clusters, each containing a distinct restriction modification system, can replace one another at the *dpn* locus. This is believed to be of protective value in mixed pneumococcal populations against bacteriophages, which are constant invaders of pneumococcal genomes (Johnston et al., 2013; Eutsey et al., 2015; Brueggemann et al., 2017).

We found several examples of presumed bacteriocin cluster switching events that had occurred adjacent to distinct quorum-sensing transcriptional regulators TprA and Rgg. Intriguingly, while the genes that are known to be under the control of TprA and Rgg were replaced, the quorum-sensing transcriptional regulators remained conserved. It is known that TprA controls the expression of its downstream bacteriocin genes: it induces them when pneumococcal cells are at high density in the presence of galactose and represses them when under high glucose growth conditions (Hoover et al., 2015). Galactose is plentiful in the nasopharynx, whereas glucose is scarce (although abundant in the blood), suggesting that the TprA may mediate the expression of its adjacent bacteriocin genes to aid pneumococci to compete for resources during nasopharynx colonization (Bidossi et al., 2012; Hoover et al., 2015). Likewise, the Rgg quorum-sensing transcriptional regulator has been shown to mediate the expression of its adjacent genes, and this is thought to be directed by sensing amino acid levels in the cellular community (Cuevas et al., 2017). An explanatory hypothesis might be that bacteriocin cluster switching provides a mechanism by which the existing intricate quorum-sensing signaling network required for coordinating population-level behaviors is accessible by the newly acquired bacteriocin cluster. This remains to be experimentally verified.

Our current work is significant for the broader community in that among the 21 different pneumococcal bacteriocin clusters now identified, many homologs are found in other unrelated streptococcal species. We also provide here a unified nomenclature for the pneumococcal bacteriocin clusters and their genes. Finally, the fact that any single pneumococcal genome possesses multiple bacteriocin clusters should be carefully considered when designing laboratory experiments aimed at assessing the activity of an individual bacteriocin.

Overall, these population genomic and transcriptomic analyses reveal an extraordinary complexity of bacteriocins among pneumococci and underscore the need to determine precisely how these bacteriocins drive changes within the pneumococcal population and the wider microbial community. Such findings are interesting not only for their population biology and ecology insights, but also because bacteriocins potentially have a huge impact on public health. By directly influencing changes in microbial populations, bacteriocins might indirectly be affecting the effectiveness of vaccines in the longer term: after vaccine use in human populations, the target bacterial population is significantly changed and those bacteria must now compete within the altered microbiome. If bacteriocins are essential to the microbial competitive strategy, then the composition of bacteriocins possessed by the post-vaccination bacterial community is important to understand. Moreover, there is obvious potential for the development of bacteriocins as novel antimicrobials, and at a time when the challenges of antimicrobial-resistant microbes have never been more acute these data provide many new areas of investigation.

DATA AVAILABILITY

Representative examples of the newly discovered bacteriocin clusters have been deposited at GenBank under the accession numbers of MF990778–MF990796. Accession numbers for all genomes used in this study are listed in the **Supplementary Table S1**. Raw transcriptomic sequence data used in this study is available under GEO accession numbers GSE103778 and GSE110750.

AUTHOR CONTRIBUTIONS

RR designed the genome mining pipelines. AT, CH, and AB performed the RNA sequencing experiments. RR, AT, CH, JK, and AB analyzed the data. RR and AB wrote the manuscript. All authors reviewed and agreed on the final version of the manuscript.

FUNDING

This work was supported by the Wellcome Trust (Research Fellowship 083511/Z/07/Z and Investigator Award 206394/Z/17/Z to AB), the University of Oxford John Fell Fund (Grant 123/734 to AB). BIGSdb genome database support was provided by a Wellcome Trust Biomedical Research Fund award (Grant 04992/Z/14/Z) awarded to Martin J. C. Maiden, Keith A. Jolley, and AB at the University of Oxford. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We wish to acknowledge computational assistance and helpful comments on the manuscript from Dr. Melissa Jansen van Rensburg. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics for the generation of the RNA sequencing data. A pre-print version of this article was published in bioRxiv (Rezaei Javan et al., 2017).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02012/full#supplementary-material

FIGURE S1 | Amino acid sequence alignments of bacteriocin genes identified in this study. Putative bacteriocin genes that were identified among

the pneumococcal genomes in this study were aligned against similar bacteriocin genes in other bacterial species for comparison.

FIGURE S2 | Methodology for analyzing the RNA sequencing data from the co-colonization experiment. **(A)** Steps involved in creating the pseudo-reference genome sequence. The name of the tool used in each step is shown in pink. **(B)** Schematic describing the combination of RNA sequence reads to assess differential expression levels.

FIGURE S3 | Guanine (G) and cytosine (C) content of pneumococcal bacteriocin clusters. **(A)** Four examples of GC plots depicting the percentage GC-content of bacteriocin cluster genes (red), transcriptional regulator genes (green), and other adjacent pneumococcal genes (grey). The names of the bacteriocin and the genome in which it was found are given, with the percentage GC-content of each in brackets. Each graph depicts GC-content and adenine (A) thymine (T) content by the green and blue lines, respectively. **(B)** Average GC-content for each bacteriocin cluster type, organized by bacteriocin type or class. The lanthipeptides formed two subsets based on GC-content of <30 or >31%. **(C)** Average GC-content for 3 genomes of non-pneumococcal streptococcal species analyzed in a previous study (Kurioka et al., 2017).

FIGURE S4 | Diversity of bacteriocins within a global pneumococcal dataset. A phylogenetic tree of all genomes in the study dataset is depicted and labeled according to the presence of different bacteriocins. Clusters with missing genes were defined as partial. The exceptions were the *blp* clusters: their highly diverse and complicated genetic compositions among pneumococci genomes meant that a similar classification between partial and complete clusters could not be applied (Bogaardt et al., 2015). Instead, their presence (irrespective of being partial or complete) is depicted by the green color.

TABLE S1 | List of pneumococcal genomes used in this study.

TABLE S2 | Classification and nomenclature of bacteriocin clusters found among pneumococci.

TABLE S3 | Result of the BLAST searches of the bacteriocin clusters in the NCBI database.

TABLE S4 | Descriptive data for the genomes included in the global representative dataset.

TABLE S5 | RNA-seq data from the competition experiment.

REFERENCES

Aguiar, S., Brito, M., Gonçalo-Marques, J., Melo-Cristino, J., and Ramirez, M. (2010). Serotypes 1, 7F and 19A became the leading causes of pediatric invasive pneumococcal infections in Portugal after 7 years of heptavalent conjugate vaccine use. *Vaccine* 28, 5167–5173. doi: 10.1016/j.vaccine.2010.06.008

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Arnison, P., Bibb, M., Bierbaum, G., Bowers, A., Bugni, T., Bulaj, G., et al. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat. Prod. Rep.* 30, 108–160. doi: 10.1039/c2np20085f

Auranen, K., Mehtälä, J., Tanskanen, A. S., and Kaltoft, M. (2009). Between-strain competition in acquisition and clearance of pneumococcal carriage—epidemiologic evidence from a longitudinal study of day-care children. *Am. J. Epidemiol.* 171, 169–176. doi: 10.1093/aje/kwp351

Begley, M., Cotter, P., Hill, C., and Ross, R. (2009). Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. *Appl. Environ. Microbiol.* 75, 5451–5460. doi: 10.1128/AEM.00730-09

Bidossi, A., Mulas, L., Decorosi, F., Colomba, L., Ricci, S., Pozzi, G., et al. (2012). A functional genomics approach to establish the complement of carbohydrate transporters in *Streptococcus pneumoniae*. *PLoS One* 7:e33320. doi: 10.1371/journal.pone.0033320

Bogaardt, C., van Tonder, A., and Brueggemann, A. B. (2015). Genomic analyses of pneumococci reveal a wide diversity of bacteriocins – Including pneumocyclicin, a novel circular bacteriocin. *BMC Genomics* 16:554. doi: 10.1186/s12864-015-1729-4

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

Bosi, E., Donati, B., Galardini, M., Brunetti, S., Sagot, M., Lió, P., et al. (2015). MeDuSa: a multi-draft based scaffolder. *Bioinformatics* 31, 2443–2451. doi: 10.1093/bioinformatics/btv171

Brown, S., West, S., Diggle, S., and Griffin, A. (2009). Social evolution in micro-organisms and a Trojan horse approach to medical intervention strategies. *Philos. Trans. R. Soc. B* 364, 3157–3168. doi: 10.1098/rstb.2009.0055

Brueggemann, A. B., Griffiths, D., Meats, E., Peto, T., Crook, D., and Spratt, B. (2003). Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. *J. Infect. Dis.* 187, 1424–1432. doi: 10.1086/374624

Brueggemann, A. B., Harrold, C., Rezaei Javan, R., van Tonder, A., McDonnell, A., and Edwards, B. (2017). Pneumococcal prophages are diverse, but not without structure or history. *Sci. Rep.* 7:42976. doi: 10.1038/srep42976

Carver, T., Rutherford, K., Berriman, M., Rajandream, M., Barrell, B., and Parkhill, J. (2005). ACT: the artemis comparison tool. *Bioinformatics* 21, 3422–3423. doi: 10.1093/bioinformatics/bti553

Chewapreecha, C., Harris, S., Croucher, N., Turner, C., Marttinen, P., Cheng, L., et al. (2014). Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* 46, 305–309. doi: 10.1038/ng.2895

Coffey, T., Dowson, C., Daniels, M., Zhou, J., Martin, C., Spratt, B., et al. (1991). Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae*. *Mol. Microbiol.* 5, 2255–2260. doi: 10.1111/j.1365-2958.1991.tb02155.x

Crook, D. W., Brueggemann, A. B., Sleeman, K., and Peto, T. E. A. (2004). "Pneumococcal carriage," in *The Pneumococcus,* eds E. I. Tuomanen, T. J. Mitchell, D. A. Morrison, and B. Spratt (Washington, D.C.: ASM Press), 136–147.

Croucher, N., Finkelstein, J., Pelton, S., Mitchell, P., Lee, G., Parkhill, J., et al. (2013a). Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* 45, 656–663. doi: 10.1038/ng.2625

Croucher, N., Mitchell, A., Gould, K., Inverarity, D., Barquist, L., Feltwell, T., et al. (2013b). Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. *PLoS Genet.* 9:e1003868. doi: 10.1371/journal.pgen.1003868

Croucher, N., Harris, S., Fraser, C., Quail, M., Burton, J., van der Linden, M., et al. (2011). Rapid pneumococcal evolution in response to clinical interventions. *Science* 331, 430–434. doi: 10.1126/science.1198545

Croucher, N., Walker, D., Romero, P., Lennard, N., Paterson, G., Bason, N. C., et al. (2008). Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F-ST81. *J. Bacteriol.* 191, 1480–1489. doi: 10.1128/JB.01343-08

Cuevas, R., Eutsey, R., Kadam, A., West-Roberts, J., Woolford, C., Mitchell, A. P., et al. (2017). A novel streptococcal cell-cell communication peptide promotes pneumococcal virulence and biofilm formation. *Mol. Microbiol.* 105, 554–571. doi: 10.1111/mmi.13721

Czaplewski, L., Bax, R., Clokie, M., Dawson, M., Fairhead, H., Fischetti, V. A., et al. (2016). Alternatives to antibiotics-a pipeline portfolio review. *Lancet Infect. Dis.* 16, 239–251. doi: 10.1016/S1473-3099(15)00466-1

Dawid, S., Roche, A., and Weiser, J. (2006). The blp bacteriocins of *Streptococcus pneumoniae* mediate intraspecies competition both in vitro and in vivo. *Infect. Immun.* 75, 443–451. doi: 10.1128/IAI.01775-05

Didelot, X., and Wilson, D. (2015). ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* 11:e1004041. doi: 10.1371/journal.pcbi.1004041

Doern, G., Heilmann, K., Huynh, H., Rhomberg, P., Coffman, S., and Brueggemann, A. B. (2001). Antimicrobial resistance among clinical isolates of *Streptococcus pneumoniae* in the United States during 1999-2000, including a comparison of resistance rates since 1994-1995. *Antimicrob. Agents Chemother.* 45, 1721–1729. doi: 10.1128/AAC.45.6.1721-1729.2001

Eutsey, R., Powell, E., Dordel, J., Salter, S., Clark, T., Clark, T. A., et al. (2015). Genetic stabilization of the drug-resistant PMEN1 pneumococcus lineage by its distinctive DpnIII restriction- modification system. *mBio* 6:e173-15. doi: 10.1128/mBio.00173-15

Garcia-Rodriguez, J., and Fresnadillo Martínez, M. (2002). Dynamics of nasopharyngeal colonization by potential respiratory pathogens. *J. Antimicrob. Chemother.* 50, 59–74. doi: 10.1093/jac/dkf506

Gladstone, R., Jefferies, J., Tocheva, A., Beard, K., Garley, D., Chong, W. W., et al. (2015). Five winters of pneumococcal serotype replacement in UK carriage following PCV introduction. *Vaccine* 33, 2015–2021. doi: 10.1016/j.vaccine.2015.03.012

Golubchik, T., Brueggemann, A. B., Street, T., Gertz, R., Spencer, C., Ho, T., et al. (2012). Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat. Genet.* 44, 352–355. doi: 10.1038/ng.1072

Guiral, S., Mitchell, T., Martin, B., and Claverys, J. (2005). Competence-programmed predation of noncompetent cells in the human pathogen *Streptococcus pneumoniae*: genetic requirements. *Proc. Natl. Acad. Sci. U.S.A.* 102, 8710–8715. doi: 10.1073/pnas.0500879102

Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J., and Fliss, I. (2010). BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiol.* 10:22. doi: 10.1186/1471-2180-10-22

Hoover, S., Perez, A., Tsui, H., Sinha, D., Smiley, D., DiMarchi, R. D., et al. (2015). A new quorum-sensing system (TprA/PhrA) for *Streptococcus pneumoniae* D39 that regulates a lantibiotic biosynthesis gene cluster. *Mol. Microbiol.* 97, 229–243. doi: 10.1111/mmi.13029

Huang, S., Platt, R., Rifas-Shiman, S., Pelton, S., Goldmann, D., and Finkelstein, J. (2005). Post- PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004. *Pediatrics* 116:e408-13. doi: 10.1542/peds.2004-2338

Johnston, C., Polard, P., and Claverys, J. (2013). The DpnI/DpnII pneumococcal system, defense against foreign attack without compromising genetic exchange. *Mob. Genet. Elements* 3:e25582. doi: 10.4161/mge.25582

Jolley, K., and Maiden, M. (2010). BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* 11:595. doi: 10.1186/1471-2105-11-595

Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome- scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kadam, A., Eutsey, R., Rosch, J., Miao, X., Longwell, M., Xu, W., et al. (2017). Promiscuous signaling by a regulatory system unique to the pandemic PMEN1 pneumococcal lineage. *PLoS Pathog.* 13:e1006339. doi: 10.1371/journal.ppat.1006339

Kaplan, S., Barson, W., Lin, P., Stovall, S., Bradley, J., Tan, T. Q., et al. (2010). Serotype 19A is the most common serotype causing invasive pneumococcal infections in children. *Pediatrics* 125, 429–436. doi: 10.1542/peds.2008-1702

Kjos, M., Snipen, L., Salehian, Z., Nes, I., and Diep, D. (2010). The Abi proteins and their involvement in bacteriocin self-immunity. *J. Bacteriol.* 192, 2068–2076. doi: 10.1128/JB.01553-09

Klugman, K. (1990). Pneumococcal resistance to antibiotics. *Clin. Microbiol. Rev.* 3, 171–196. doi: 10.1128/CMR.3.2.171

Kurioka, A., van Wilgenburg, B., Rezaei Javan, R., Hoyle, R., van Tonder, A. J., Harrold, C. L., et al. (2017). Diverse *Streptococcus pneumoniae* strains drive a MAIT cell response through MR1- dependent and cytokine-driven pathways. *J. Infect. Dis.* 217, 988–999. doi: 10.1093/infdis/jix647

Laible, G., Spratt, B. G., and Hakenbeck, R. (1991). Interspecies recombinational events during the evolution of altered PBP 2x genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. *Mol. Microbiol.* 5, 1993–2002. doi: 10.1111/j.1365-2958.1991.tb00821.x

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44, W242–W245. doi: 10.1093/nar/gkw290

Letzel, A., Pidot, S., and Hertweck, C. (2014). Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC Genomics* 15:983. doi: 10.1186/1471-2164-15-983

Marchler-Bauer, A., Lu, S., Anderson, J., Chitsaz, F., Derbyshire, M., DeWeese-Scott, C., et al. (2010). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189

Maricic, N., Anderson, E., Opipari, A., Yu, E., and Dawid, S. (2016). Characterization of a multipeptide lantibiotic locus in *Streptococcus pneumoniae*. *mBio* 7:e1656-15. doi: 10.1128/mBio.01656-15

Martińez, B., Rodrıiguez, A., Suráez, J., and Fernández, M. (1999). Synthesis of lactococcin 972, a bacteriocin produced by *Lactococcus lactis* IPLA 972, depends on the expression of a plasmid-encoded bicistronic operon. *Microbiology* 145, 3155–3161. doi: 10.1099/00221287-145-11-3155

Maynard Smith, J., Dowson, C. G., and Spratt, B. G. (1991). Localized sex in bacteria. *Nature* 349, 29–31. doi: 10.1038/349029a0

O'Brien, K., Wolfson, L., Watt, J., Henkle, E., Deloria-Knoll, M., McCall, N., et al. (2009). Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet* 374, 893–902. doi: 10.1016/S0140-6736(09)61204-6

Page, A., Cummins, C., Hunt, M., Wong, V., Reuter, S., Holden, M. T., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Perez-Pascual, D., Monnet, V., and Gardan, R. (2016). Bacterial cell–cell communication in the host via RRNPP peptide-binding regulators. *Front. Microbiol.* 7:706. doi: 10.3389/fmicb.2016.00706

Price, M., Dehal, P., and Arkin, A. (2010). FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Reichmann, P., and Hakenbeck, R. (2000). Allelic variation in a peptide-inducible two- component system of *Streptococcus pneumoniae*. *FEMS Microbiol. Lett.* 190, 231–236. doi: 10.1111/j.1574-6968.2000.tb09291.x

Rezaei Javan, R., van Tonder, A. J., King, J. P., Harrold, C. L., and Brueggemann, A. B. (2017). *Streptococcus pneumoniae* possesses an unexpectedly large bacteriocin repertoire. *bioRxiv* [Preprint]. doi: 10.1101/203398

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shak, J., Vidal, J., and Klugman, K. (2013). Influence of bacterial interactions on pneumococcal colonization of the nasopharynx. *Trends Microbiol.* 21, 129–135. doi: 10.1016/j.tim.2012.11.005

Son, M., Shchepetov, M., Adrian, P., Madhi, S., de Gouveia, L., von Gottberg, A., et al. (2011). Conserved mutations in the pneumococcal bacteriocin transporter gene, blpA, result in a complex population consisting of producers and cheaters. *mBio* 2:e179-11. doi: 10.1128/mBio.00179-11

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

Tadesse, B., Ashley, E., Ongarello, S., Havumaki, J., Wijegoonewardena, M., Gonzalez, I. J., et al. (2017). Antimicrobial resistance in Africa: a systematic review. *BMC Infect. Dis.* 17:616. doi: 10.1186/s12879-017-2713-1

van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J., and Kuipers, O. P. (2013). BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res.* 41, W448–W453. doi: 10.1093/nar/gkt391

van Tonder, A., Bray, J., Roalfe, L., White, R., Zancolli, M., Quirk, S. J., et al. (2015). Genomics reveals the worldwide distribution of multidrug-resistant serotype 6E pneumococci. *J. Clin. Microbiol.* 53, 2271–2285. doi: 10.1128/JCM.00744-15

van Tonder, A., Mistry, S., Bray, J., Hill, D., Cody, A., Farmer, C. L., et al. (2014). Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Comput. Biol.* 10:e1003788. doi: 10.1371/journal.pcbi.1003788

Walker, G., Heng, N., Carne, A., Tagg, J., and Wescombe, P. (2016). Salivaricin E and abundant dextranase activity may contribute to the anti-cariogenic potential of the probiotic candidate *Streptococcus salivarius* JH. *Microbiology* 162, 476–486. doi: 10.1099/mic.0.000237

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H., Bruccoleri, R., et al. (2015). antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 43, W237–W243. doi: 10.1093/nar/gkv437

World Health Organization (2017). *Global Priority List of Antibiotic-resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics*. Geneva: World Health Organization. Available at: http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria

Wyres, K., Lambertsen, L., Croucher, N., McGee, L., von Gottberg, A., Liñares, J., et al. (2012). Pneumococcal capsular switching: a historical perspective. *J. Infect. Dis.* 207, 439–449. doi: 10.1093/infdis/jis703

Zerbino, D., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

SCIENTIFIC REPORTS

OPEN

Pneumococcal prophages are diverse, but not without structure or history

Received: 01 November 2016
Accepted: 17 January 2017
Published: 20 February 2017

Angela B. Brueggemann, Caroline L. Harrold, Reza Rezaei Javan, Andries J. van Tonder, Angus J. McDonnell & Ben A. Edwards

Bacteriophages (phages) infect many bacterial species, but little is known about the diversity of phages among the pneumococcus, a leading global pathogen. The objectives of this study were to determine the prevalence, diversity and molecular epidemiology of prophages (phage DNA integrated within the bacterial genome) among pneumococci isolated over the past 90 years. Nearly 500 pneumococcal genomes were investigated and RNA sequencing was used to explore prophage gene expression. We revealed that every pneumococcal genome contained prophage DNA. 286 full-length/putatively full-length pneumococcal prophages were identified, of which 163 have not previously been reported. Full-length prophages clustered into four major groups and every group dated from the 1930–40 s onward. There was limited evidence for genes shared between prophage clusters. Prophages typically integrated in one of five different sites within the pneumococcal genome. 72% of prophages possessed the virulence genes *pblA* and/or *pblB*. Individual prophages and the host pneumococcal genetic lineage were strongly associated and some prophages persisted for many decades. RNA sequencing provided clear evidence of prophage gene expression. Overall, pneumococcal prophages were highly prevalent, demonstrated a structured population, possessed genes associated with virulence, and were expressed under experimental conditions. Pneumococcal prophages are likely to play a more important role in pneumococcal biology and evolution than previously recognised.

Infectious diseases are a leading cause of early childhood deaths, among which pneumonia is the most common: an estimated 1.3 million children died of pneumonia in 2013. The leading cause of pneumonia was *Streptococcus pneumoniae* (the pneumococcus), which is also a major cause of paediatric meningitis and bacteraemia. Children who survive life-threatening pneumococcal diseases often have profound disabilities[1,2].

Safe and effective pneumococcal vaccines are available to prevent disease, but vaccines are not yet universally administered among all countries, particularly those in the geographical regions in need of the greatest protection. Furthermore, although they are safe and effective, current vaccines are limited in the protection they provide. Antimicrobials can be used to treat pneumococcal infection, but antimicrobial-resistant pneumococci are found worldwide and resistance can result in treatment failures. Therefore, despite the availability of vaccines and antibiotics, pneumococcal disease remains a major global challenge[3–5]. Crucially, a long-standing unresolved question is what determines pathogenicity, because pneumococci predominantly reside in the healthy paediatric nasopharynx and it is still unclear why some pneumococci cause devastating disease and others do not.

Phages infect bacteria by attaching to surface-exposed host receptors and injecting their DNA into the cell. There are virulent phages that are only lytic, and temperate phages that can be both lytic and lysogenic. In the lytic cycle, phages synthesise new progeny, cause the host to lyse and release new viruses, while in the lysogenic stage, the viral genome is stably integrated into the bacterial DNA and replicates along with the bacterial chromosome. Prophages (temperate phage DNA that is integrated within the bacterial genome) can become lytic in a process called induction, which is triggered spontaneously or by external stimuli such as exposure to mitomycin C. Prophages are one type of mobile genetic element, transferring DNA to and from the bacterial genome[6,7].

Expressed prophage genes often have a phenotypic effect on the host bacterium, for example by producing a toxin that increases bacterial virulence (*Vibrio cholerae* and *Staphylococcus aureus*), enhancing bacterial adherence to platelets (*Streptococcus mitis*), or evading immune defences (*Pseudomonas aeruginosa*)[6–8]. Importantly, such genes may be functional even if the prophage is defective or the prophage genome is incomplete. Host

Nuffield Department of Medicine, University of Oxford, United Kingdom. Correspondence and requests for materials should be addressed to A.B.B. (email: angela.brueggemann@ndm.ox.ac.uk)

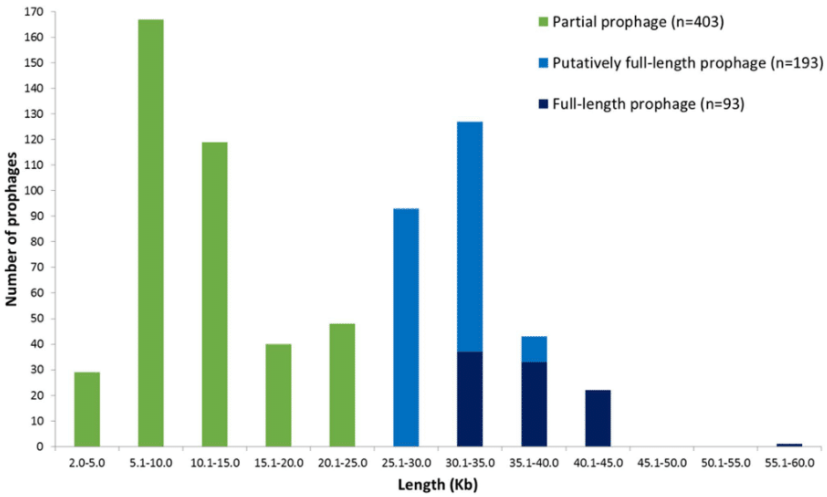


Figure 1. Graph depicting the number of prophages in each category: partial, putatively full-length or full-length. Bars represent the number of prophages relative to the length of the prophage sequence that was identified.

bacteria can also be infected with multiple prophages: for example, one strain of *Streptococcus pyogenes* contains six different prophages or prophage-like elements that together comprise approximately 12% of the bacterial genome and have been shown to directly contribute to the high virulence of that lineage[9].

Pneumococcal phages were first reported in 1975 and subsequent work using simple DNA probe-based detection of one prophage-associated gene suggested that the majority of pneumococcal genomes contain prophages; however, prevalence data has been limited and little is known about how a prophage affects the pneumococcal host[10–13]. We showed that a globally-distributed multidrug-resistant pneumococcal lineage called PMEN1 likely gained some of its epidemiological success from the acquisition of three mobile genetic elements including the MM1 prophage, which has been shown *in vitro* to enhance adherence to human cells[14–16]. We also discovered a surprising link between prophages and antibiotic resistance: a pneumococcal prophage harboured the genes that confer tetracycline resistance[17].

Prophages are widespread and demonstrably linked to increased infection, pathogenesis and virulence in many bacterial species. Prophages are likely to be highly prevalent within the pneumococcal population, but the genetics, diversity and molecular epidemiology of pneumococcal prophages are not well understood. If pneumococcal prophages are widely distributed, genetically highly diverse and frequently recombining then it will be a challenge to associate specific prophages with pneumococcal genotypes and phenotypes. Alternatively, if pneumococcal prophages are highly prevalent but the prophage population is structured, then there will be a framework on which to investigate specific prophage/pneumococcus interactions and prophage-derived mechanisms that may contribute to pneumococcal phenotype, disease potential and epidemiological success. Uniquely, we have included a set of historical genomes dating from 1916 onwards, allowing us to investigate the potential for the persistence of prophages over many decades.

Therefore, the aims of this study were to: i) determine the prevalence of prophage DNA within a large global and historical pneumococcal genome dataset; ii) assess the genetic diversity and molecular epidemiology of prophages among pneumococci isolated over a 90 year period; and iii) investigate associations between these prophage sequences and pneumococcal genetic lineages.

Results

High prevalence of prophage sequences among a diverse collection of pneumococcal genomes. The genomes analysed in this study comprised a diverse set of 482 pneumococci recovered between 1916 and 2008 from people of all ages residing in 36 different countries. Pneumococci were recovered from both healthy individuals and those with disease. 91 pneumococcal serotypes, 258 multilocus sequence types (STs) and 94 different clonal complexes (CCs; clusters of closely-related pneumococci) were represented in the dataset (Table S1).

Overall, 100% of the pneumococcal genomes contained coding sequences annotated as being phage-associated. In particular, 689 full-length, putatively full-length and partial prophage sequences at least 2 Kb in length (see Methods for definitions) were identified among 326 (68%) of the 482 pneumococcal genomes (Fig. 1). Phage sequences <2 Kb in length were not investigated further in this study. The total amount of prophage sequence (≥ 2 Kb in length) within a single pneumococcal genome ranged from 4–135 Kb, which was up to 6% of the pneumococcal genome.

In total, we revealed 163 previously unreported full-length or putatively full-length pneumococcal prophages in this dataset. 93 (66 representative examples) full-length prophage sequences were identified and these were 29–60 Kb (median 37.5 Kb) in length (Fig. 1; Table S2). 76 (56 representative examples) of these full-length

prophage sequences were new. 193 putatively full-length prophages ranged in size from 25–38 Kb; 129 of these sequences were unique and 107 of the unique prophage sequences were new (12 matched to prophage sequences found in GenBank and 10 matched to new full-length prophages in this study). In addition, 403 partial prophage sequences were identified among the 482 pneumococcal genomes and these ranged in length from 2–25 Kb.

45% (219/482) of the pneumococcal genomes harboured at least one full-length and/or putatively full-length prophage sequence. Among these 219 genomes, 28% (n = 63) were poly-lysogenic, i.e. contained >1 full-length or putatively full-length prophage sequence within the pneumococcal genome.

Prophage sequence alignments demonstrate clusters of prophages that persist for decades. The 66 representative full-length prophage sequences were aligned and the average pairwise identity was only 40.2% (range 20.8–97.4%); however, pairwise comparisons of prophage sequences and an unrooted phylogenetic tree depicted four major prophage clusters and one single prophage (Fig. 2a,b and Figure S1). Remarkably, all four major clusters included prophages integrated in host pneumococci that were recovered in the 1930–40 s as well as among modern pneumococci, demonstrating the persistence of the prophage clusters and some prophages over several decades.

The evidence for genes shared between prophage clusters was limited. Figure 2c depicts all of the different genes (<90% amino acid sequence similarity) identified among all representative prophages of each cluster, and whether any of those genes were found among any prophage(s) of any other cluster. Small numbers of individual genes were shared between some prophages in different clusters, e.g. one holin gene sequence was found in 19 prophages distributed across all five clusters (Fig. 2c and Table S3); however, the vast majority of genes within a prophage cluster were only found among that cluster of prophages (Fig. 2c). There were three larger groups of shared genes (cluster C and either cluster A (n = 16), B (n = 27) or D (n = 29), respectively) and generally these were sets of genes, sometimes contiguous genes, which were shared by a subset of prophages from two clusters (Table S3). The majority of the shared genes were predicted to encode proteins with unknown functions.

Prophages within every cluster possessed genes involved in DNA packaging, the production of the phage capsid, tail and other structural proteins, and genes encoding hypothetical proteins that were identical or highly similar (>95% identical at a nucleotide level) to the other prophages in that cluster (Figs 3 and 4). Prophage sequences also possessed a cluster-specific set of other identical and highly similar genes including those that encoded the integrase, other enzymes, DNA binding and replication proteins, holin proteins and/or lytic amidase (Figs 3 and 4; Table S4).

Thirteen representative cluster A prophages had an average pairwise nucleotide sequence identity of 83.6% (range 75.9–95.1%; Fig. 3). These were identified among 15 pneumococci recovered from 1939–2005 (Table S2). Cluster B prophages (n = 28) subdivided into 4 smaller clusters (Figs 2a,b and 3). Clusters B1 and B2 each depicted three representative prophages: cluster B1 prophages shared 91.9% average pairwise identity (range 90.6–93.8%) and were found in three different pneumococci recovered in 1957, 1993 and 2005; cluster B2 prophages were 78.9–91.8% identical (average 83.3%) and were identified in five host pneumococci from 1988–2000 (Fig. 3 and Table S2). Fourteen representative cluster B3 prophages (average pairwise identity of 77.9%, range 70.2–96.8%) were identified among 21 pneumococci recovered from 1941–2007. Eight representative cluster B4 prophages averaged 78.7% pairwise identity (range 66.1–95.3%) and were found in 13 pneumococci recovered from 1938–2007 (Table S2).

Nineteen representative cluster C prophages were identified among 21 pneumococci recovered from 1940–2008 and shared only 69.6% (range 55.1–97.4%) pairwise identity overall (Fig. 4 and Table S2). Cluster C included the IPP61 prophage sequence that harboured the tetracycline resistance mobile genetic element[17,18], which was not present in any of the other prophages. Five representative cluster D prophages and one cluster E prophage were also more diverse: the overall pairwise identity among cluster D prophages was 65.6% (range 57.8–81.9%) and the five host pneumococci were identified from 1939–1985. The cluster E prophage, IPP24, was identified in a host pneumococcus from 1985 (Fig. 4).

Relationships between prophage integration sites and prophage integrase sequences. The 93 full-length prophages were consistently integrated in specific locations of the pneumococcal genome. There were five major categories of integration sites, within which the patterns of pneumococcal genes flanking the prophage integrase and amidase genes were consistent (Table 1). The flanking pneumococcal genes were involved in transcriptional regulation, metabolic enzyme production and/or activity, purine nucleotide biosynthesis, regulation of DNA repair, and competence. The 193 putatively full-length prophages were inserted in the same five major insertion site categories (Table S5).

The integration site was directly associated with the nucleotide sequence of the prophage integrase and the individual prophage clusters (Table 1). Based upon a threshold of ≥90% nucleotide sequence identity, the 93 full-length prophage integrase sequences clustered into two major groups (I and II) that together described 75.5% of the prophage integrases. There were also two minor integrase groups (III and IV), identical integrase sequences in a pair of prophages, and four single prophages that each had a unique integrase sequence. There was little to no sequence similarity between integrase groups.

Cluster A prophages were split in terms of integrase groups: seven prophages (IPP48, IPP53, SPN_18, SPN_H2, IPP66 and IPP52) had a group I integrase and six (MM1, IPP55, IPP65, BHN167, IPP14, IPP54 and IPP39) possessed a group III integrase. All cluster B prophages possessed an integrase from group II, apart from IPP5, which had a unique integrase. Cluster C prophages mainly had an integrase from group I, apart from four prophages (IPP64, IPP46, IPP15 and IPP69) that had a group IV integrase. One prophage in cluster D (IPP26) had a group I integrase sequence whereas the four others possessed unique integrase sequences. The single prophage IPP24 contained an integrase sequence from group I.

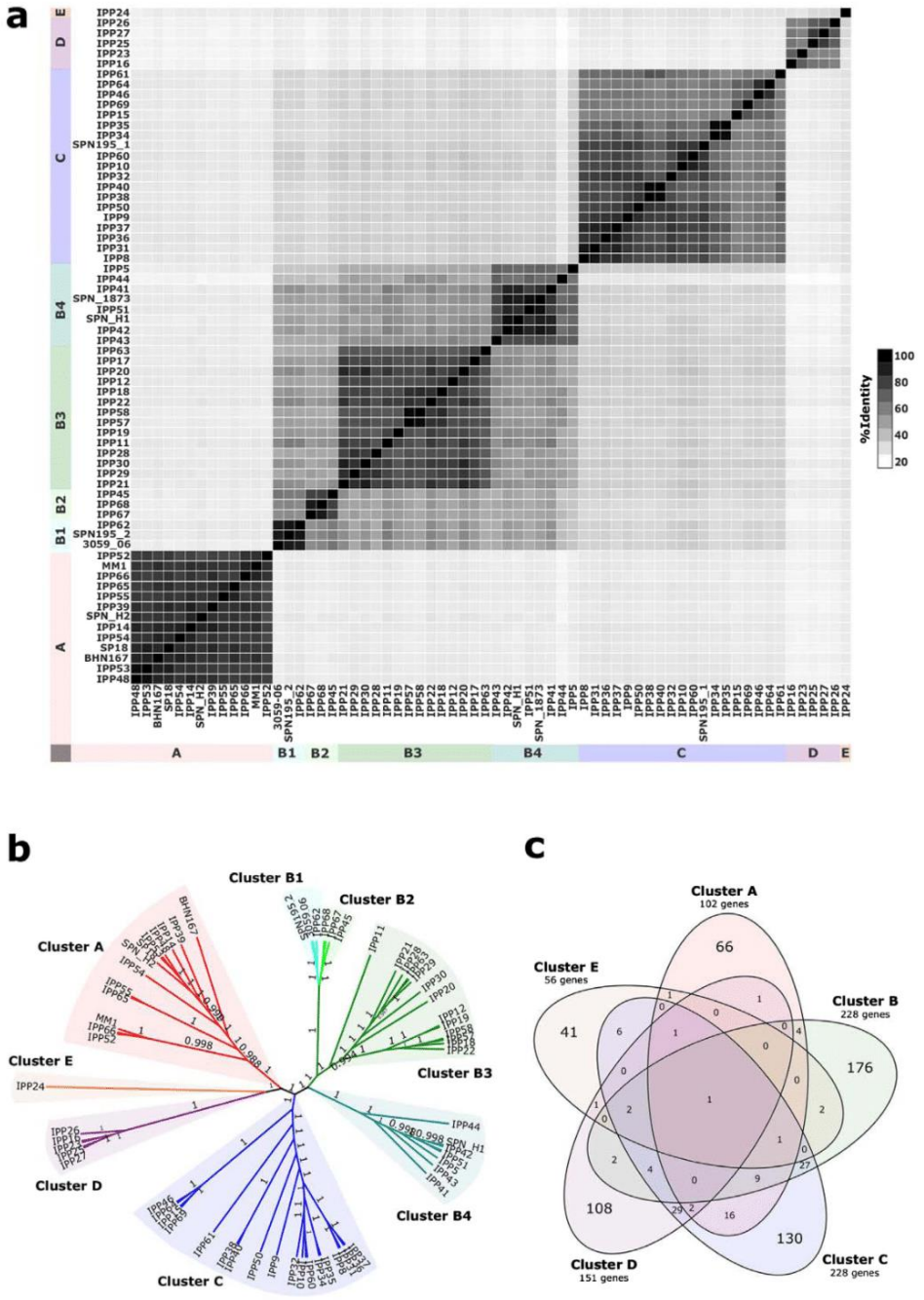


Figure 2. Description of nucleotide sequence similarity, phylogenetic clustering and shared genes among the 66 representative full-length prophages identified in the pneumococcal genome dataset. (**a**) Heat map depicting the percentage nucleotide sequence identity shared between pairs of full-length prophage sequences. Groups of similar prophages are marked A-E and correspond to the clusters seen in part b. (**b**) Clusters of prophage sequences based upon nucleotide sequence identity. Bootstrap values are marked on branches of each cluster. (**c**) Venn diagram depicting genes unique to each prophage cluster and genes shared between clusters.

***pblA* and *pblB*, putative virulence factors.** The phage tail is used for host recognition, attachment and delivery of the phage DNA to the host cell. *pblA* and *pblB* are phage tail protein genes that were first described in *S. mitis* and were shown to be associated with platelet binding and an increased risk of endocarditis[8]. Pneumococcal prophage-associated *pblB* was recently shown to be associated with pneumococcal adherence to human lung epithelial cells and increased persistence in the nasopharynx and lung in a murine model of pneumococcal infection[19].

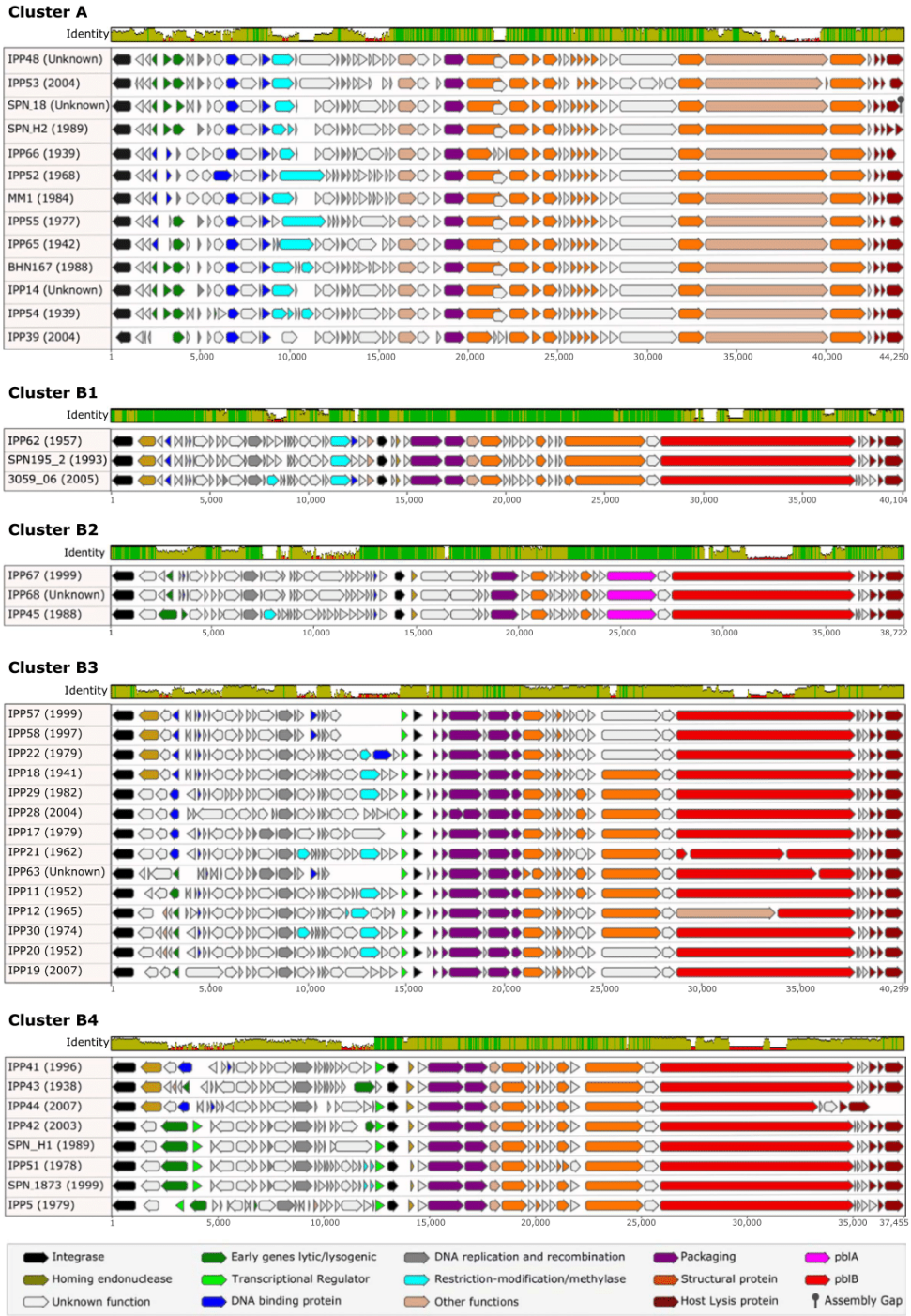


Figure 3. Nucleotide sequence alignments of representative full-length prophages from clusters A and B. The coloured bar at the top of each cluster indicates the mean pairwise nucleotide sequence identity over all pairs in the column: bright green = 100% identity; green-brown = <100% but >30% identity; and red = < 30% identity. Prophage genes are coloured based on putative or known function. The names of each prophage are given, followed by the year of isolation (in brackets) of the oldest known pneumococcus harbouring that prophage.

In the current study, only prophages in clusters B and C (71.6% of the 66 unique full-length prophage sequences) possessed *pblA* and/or *pblB* (Figs 3 and 4). The three prophages in cluster B2 were the only prophages that possessed *pblA* and those *pblA* sequences were virtually identical (99.9% pairwise identity), although the pairwise sequence identity among the three versions of *pblB* possessed by these three prophages was only 67.1%.

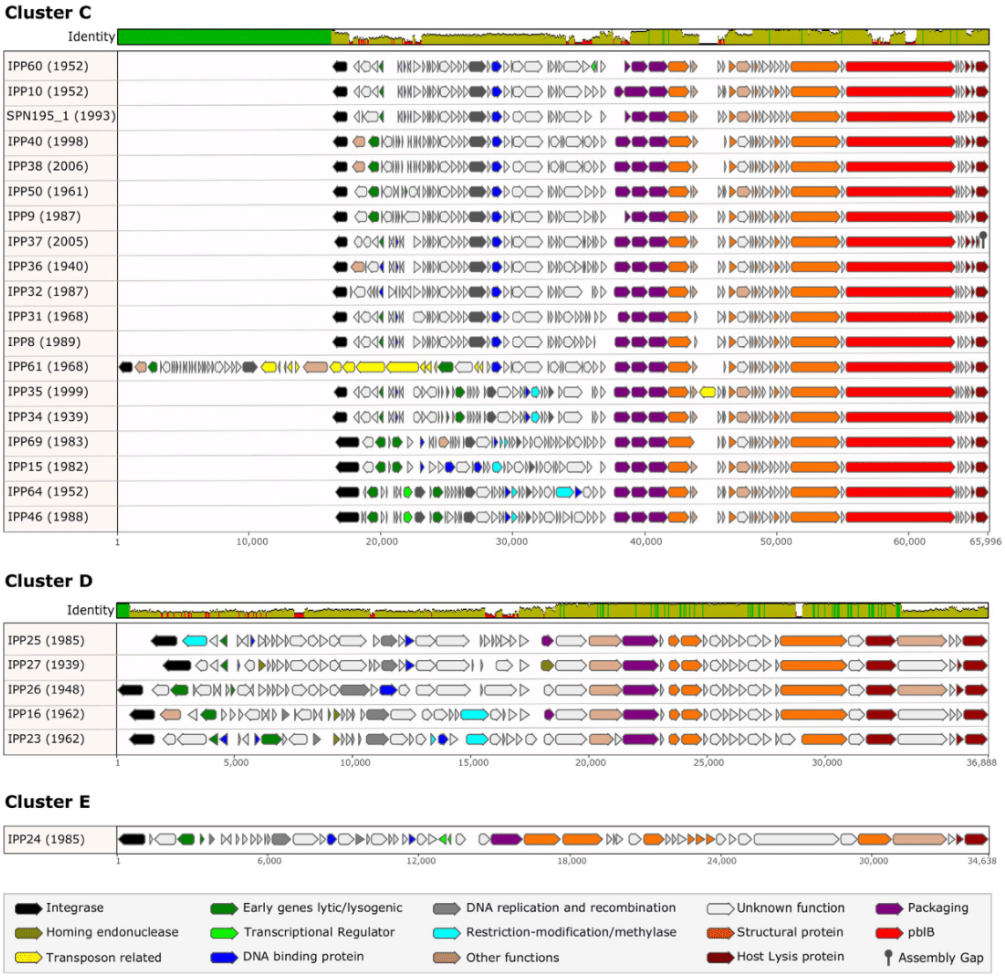


Figure 4. Nucleotide sequence alignments of representative full-length prophages from clusters C, D and E. The coloured bar at the top of each cluster indicates the mean pairwise nucleotide sequence identity over all pairs in the column: bright green = 100% identity; green-brown = <100% but >30% identity; and red = < 30% identity. Prophage genes are coloured based on putative or known function. The names of each prophage are given, followed by the year of isolation (in brackets) of the oldest known pneumococcus harbouring that prophage.

Overall, *pblB* sequence diversity was high, which was perhaps not surprising since the phage tail interacts with the host. The average pairwise sequence identity was 63.5% among all 47 *pblB* sequences, although within-cluster sequence identity was higher: clusters B1 (83.6%); B2 (67.3%); B3 (68.0%); B4 (69.6%); and C (79.7%). The sequence variation within *pblB* led to sequencing challenges associated with this particular gene, i.e. the majority of the putatively full-length prophage sequences in this study had a fragmented *pblB* and as a consequence, an incomplete sequence assembly. There were no complete matches to the *pblB* sequence from the murine model study (NTUH-P15 *pblB*)[19]: the best matches to any prophage *pblB* in this study were 97–98% similar matches to only ~1.2–4.4 Kb of the NTUH-P15 *pblB* sequence (data not shown).

Molecular epidemiology and persistence of full-length and putatively full-length prophage sequences. Collectively, the 286 full-length and putatively full-length prophages were identified among 56 different CCs and 36 Singletons (unclustered pneumococci) overall. Five or more prophages were identified among the pneumococci representing 17 different CCs (Fig. 5). The two CCs with the greatest number of prophages, CC1094^{6A} (CC^{serotype}; n = 32) and CC41/1605^{19A} (n = 18), were lineages of pneumococci recovered in South Africa during the 1970–80 s, apart from one pneumococcus in CC41/1605^{19A} that was recovered in the USA in 1985. Twelve different prophages were detected among the 17 pneumococci in CC1094^{6A} and 15 of the 17 pneumococci each contained two different prophages. Three different prophages were detected among 14 pneumococci in CC41/1605^{19A} and four of the pneumococci each contained the same pair of prophages. The pneumococci of CC1094^{6A} and CC41/1605^{19A} were isolated from both healthy people and those with invasive

Cat[a]	N[b]	Genes upstream of phage integrase			PI[c]	Genes downstream of phage amidase		
a	35[d]	P-loop-containing kinase	transporter	putative sporulation transcription regulator whiA	I	pyridine nucleotide-disulfide oxidoreductase	ABC transporter permease	lipoprotein
b	7	transposase	recombination regulator RecX	adenylosuccinate synthetase	II	hypothetical protein	cytidine/deoxycytidylate deaminase	deoxyuridine 5'-triphosphate nucleotidohydrolase
b	6	transposase	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	hypothetical protein	cytidine/deoxycytidylate deaminase
b	5	IS1167 transposase	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	cytidine/deoxycytidylate deaminase	deoxyuridine 5'-triphosphate nucleotidohydrolase
b	3	IS1167 transposase	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	hypothetical protein	cytidine/deoxycytidylate deaminase
b	2	tRNA-Asn	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	cytidine/deoxycytidylate deaminase	deoxyuridine 5'-triphosphate nucleotidohydrolase
b	2	tRNA-Asn	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	hypothetical protein	cytidine/deoxycytidylate deaminase
b	2	transposase	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	cytidine/deoxycytidylate deaminase	phage integrase
b	1	assembly gap	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	hypothetical protein	cytidine/deoxycytidylate deaminase
b	1	transposase	recombination regulator RecX	adenylosuccinate synthetase	II	hypothetical protein	arginine deiminase	arginine deiminase
b	1	transposase-like protein, IS630	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	cytidine/deoxycytidylate deaminase	deoxyuridine 5'-triphosphate nucleotidohydrolase
b	1	transposase-like protein, IS630	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	hypothetical protein	cytidine/deoxycytidylate deaminase
b	1	tRNA-Asn	recombination regulator RecX	adenylosuccinate synthetase	II	DNA-binding protein	integrase	assembly gap
b	1	—	—	assembly gap	II	DNA-binding protein	cytidine/deoxycytidylate deaminase	deoxyuridine 5'-triphosphate nucleotidohydrolase
c	3	recombination regulator RecX	adenylosuccinate synthetase	cytidine/deoxycytidylate deaminase	III	deoxyuridine 5'-triphosphate nucleotidohydrolase	phosphoglycerate mutase	DNA repair protein RadA
c	1	phage lytic amidase	DNA-binding protein	cytidine/deoxycytidylate deaminase	III	deoxyuridine 5'-triphosphate nucleotidohydrolase	phosphoglycerate mutase	DNA repair protein RadA
c	1	DNA-binding protein	hypothetical protein	cytidine/deoxycytidylate deaminase	III	deoxyuridine 5'-triphosphate nucleotidohydrolase	phosphoglycerate mutase	DNA repair protein RadA
c	1	recombination regulator RecX	adenylosuccinate synthetase	cytidine/deoxycytidylate deaminase	V	deoxyuridine 5'-triphosphate nucleotidohydrolase	phosphoglycerate mutase	DNA repair protein RadA
d	3	superoxide dismutase	competence protein CglA	competence protein CglB	IV	hypothetical protein	competence protein CglC	competence protein CglD
e	3	leucyl-tRNA synthetase	acetyltransferase	GNAT family acetyltransferase	II	DNA-binding protein	Holliday junction DNA helicase RuvB	hypothetical protein ProS

Table 1. Pneumococcal genes flanking either side of the 93 integrated full-length prophage sequences. [a]Cat = pneumococcal genome insertion site category. 13 examples of genomes with miscellaneous prophage integration sites, including assembly gaps on either side of the prophage sequence, are not shown here. A more detailed list of the integration sites and flanking genes may be found in Table S5, which includes the flanking genes associated with the 193 putatively full-length prophage sequences. [b]N = number of prophages. [c]PI = prophage integrase group, determined by the nucleotide sequence of the integrase. [d]Two pneumococcal genomes had assembly gaps after the prophage amidase.

disease. CC1094[6A] remains a major pneumococcal lineage causing invasive disease in South Africa[19], but whether contemporary pneumococci of CC1094[6A] continue to maintain these prophages is currently unknown.

CCs 113[18C], 15[14], 180[3] and 199[19A] are all globally-distributed lineages of pneumococci (reference strains PMEN36, PMEN9, PMEN31, and PMEN37, respectively) that are prevalent in carriage and disease worldwide[20,21]. Each of these lineages possessed two predominant prophages plus other single prophages (Fig. 5). In

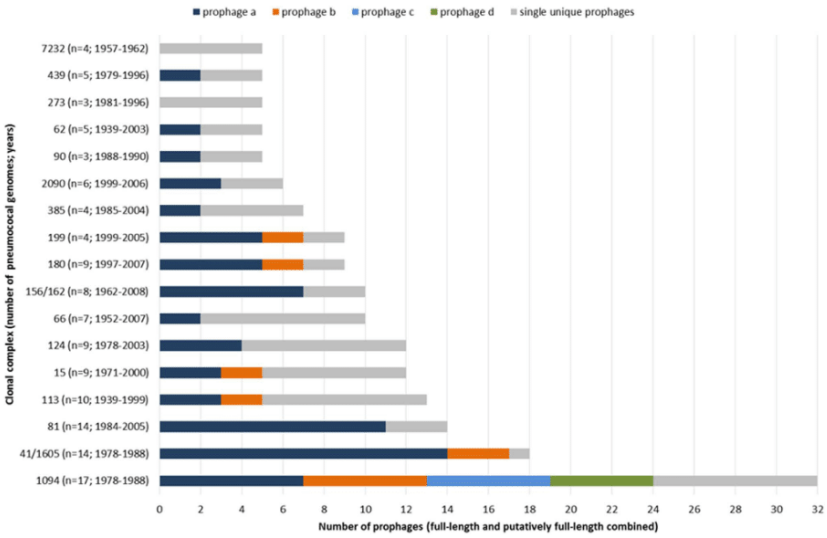


Figure 5. Illustration of the number of different full-length and putatively full-length prophages identified in each of the major pneumococcal clonal complexes. Clonal complexes are labelled on the y-axis followed by brackets containing the number of pneumococcal genomes within that complex and the years of isolation of those pneumococci. See Table 2 and S2 for details of the specific prophages identified within each clonal complex.

contrast, CCs 81[23F] and 156/162[9V] are also globally-distributed lineages (PMEN1 and PMEN3, respectively) in disease and carriage, but the majority of pneumococci in these CCs harboured one main prophage.

Furthermore, there was generally a strong association between individual prophages and the host pneumococcal genetic lineage, that is, unique prophages were usually found in one (or one predominant) CC. Nineteen different prophages were found at least three times within the entire dataset and of these, 13 prophages were found exclusively in a single CC (Table 2). Most notably, some of these prophages persisted for decades: IPP34 (CC113[18C]; ≥60 y); IPP12 (CC615[1] and CC217[1], ≥54 y); IPP29 (CC124[14]; ≥24 y); and MM1 (CC81[23F]; ≥18 y). Serotypes 18C, 1 and 14 have a high invasive disease potential[22], but whether the specific prophages associated with these serotypes and genetic lineages contribute to pathogenesis remains to be determined. There was no obvious association between specific prophages and pneumococcal serotypes, independent of the well-recognised serotype association with CC (Table S2)[20].

RNA sequencing demonstrated clear evidence for prophage gene expression. RNA sequencing was utilised to explore whether prophage gene expression could be detected. The pneumococcal reference strain PMEN3 (Spain[9V]−3), which contained two full-length prophages and one partial prophage, was grown in broth culture and mitomycin C was added to prompt prophage induction. PMEN3 culture samples were taken at sequential time points and RNA was extracted and sequenced. The RNA-seq data clearly showed prophage gene expression among all three prophage sequences (Fig. 6). Both full-length prophages significantly expressed nearly their full complement of genes after mitomycin C induction, while downregulating the genes involved in integration and lysogeny. Remarkably, most of the genes associated with the partial prophage sequence also were expressed after the addition of mitomycin C, and some of the gene expression levels were also statistically significant. The functions of nearly all of the partial prophage genes are currently unknown.

Discussion

The high prevalence of prophages among pneumococci was predicted some years ago, but the advent of affordable genome sequencing has made it possible to more easily identify prophages. That said, a rapid and straightforward identification of the many prophages in this study by screening tools and pipelines was hindered since so many of the prophages were newly discovered. This study required significant manual effort and inspection of sequences, but the return was a lengthy list of prophages to investigate here and in future studies. The key overall finding was that the breadth and depth of prophage sequences within pneumococcal genomes was astonishing.

The second most important finding was that the pneumococcal prophage population does appear to be structured. Many papers describe the vast diversity of phages among bacterial species – and high phage diversity is unequivocal[23] – but our data and analyses also clearly demonstrated clusters of closely-related pneumococcal prophages and in some cases, clear associations of specific prophages with pneumococcal genetic lineages. Population structure (genetic clustering) in bacterial populations reveals clusters of closely-related individuals and suggests that the population is diversifying from a common ancestor, and this diversification might lead to a beneficial outcome like increased virulence or antibiotic resistance. The pneumococcal population structure is reasonably well understood in terms of major circulating lineages and their association with disease, carriage, antibiotic resistance, geographical location, and so on. The fact that prophages can be associated with specific genetic lineages within the pneumococcal population provides a framework on which to assess the specific

Clonal complex	Prophage frequency (years of isolation)[a]																			
	IPP32 (1977–1988)	MM1 (1984–2002)	IPPX302 (1978–1985)	SPN195_1 (1993–2008)	IPPX3 (1978–1984)	IPPX4 (1978–1987)	3059_06_phage (1978–1985)	IPPX300 (1999–2003)	IPP57 (1992–2005)	IPP12 (1948–2002)	IPP29 (1978–2002)	IPPX328 (1978)	IPP34 (1939–1999)	IPP62 (1999–2006)	IPP67 (1957–1962)	SPN_1873 (1999–2000)	IPPX215 (1999–2006)	IPP8 (1989–2000)	IPPX100 (1965–1966)	Total
1094			7		6	6		5												24
41/1605	14											3								17
81		11																		11
156/162				7																7
199							5										1			6
124							1				4									5
180									5											5
2090																3	1			4
3122/5977																			3	3
113													3							3
15															3					3
439																		2		2
Singleton	1																1			2
66														2						2
217										2										2
615										2										2
236/271/320																		1		1
7232														1						1
68/1656	1																			1
Total	16	11	7	7	6	6	6	5	5	4	4	3	3	3	3	3	3	3	3	101

Table 2. Associations between full-length and putatively full-length prophages and pneumococcal genetic lineages. [a]Only the full-length and putatively full-length prophages present ≥3 times in the study dataset were included here. The years of isolation refer to when the host pneumococci were isolated.

contribution prophages might make to a genetic lineage, particularly in the context of increased cell adherence, pathogenesis, virulence, etc. Additionally, the function of many of the prophage genes is unknown and thus these prophages represent a vast reservoir of new information to be revealed.

Furthermore, although the prophage genes were organised in 'modular' groups, e.g. genes encoding phage structural components were located adjacent to each other in the prophage sequence, we could find relatively little evidence (given the large numbers of prophages and prophage genes investigated in this study) for genes that were similar at a sequence level and shared between prophage clusters. This is relevant because one theory of phage evolution is that modular groups of genes are exchanged between different phages; however, that does not appear to be an accurate reflection of evolution among pneumococcal prophages[24]. More broadly, this study highlighted the importance of analysing a large collection of diverse bacterial genomes to reveal the similarities and differences among prophages within the same bacterial species.

The third major finding was that some prophages persisted over long time periods – sometimes relatively unchanged at a nucleotide sequence level. Given the high level of recombination and frequent exchange of genetic material between unrelated pneumococci, this suggests that some form of selective pressure may be acting to maintain the integration of certain prophages within specific pneumococcal lineages. Moreover, the prophages represent large regions of DNA and therefore the maintenance of so much foreign DNA could suggest that the prophages provide a benefit to the host pneumococcus.

Furthermore, RNA sequencing of the prophages found in the reference strain of one globally-distributed, multidrug-resistant pneumococcal lineage clearly demonstrated prophage gene expression in a simple *in vitro* experiment. At a minimum, this provided evidence for a functional set of prophage genes and proof-in-principle that RNA sequencing could be a useful experimental tool to understand prophage gene expression. It remains to be shown whether functional phage proteins are produced *in vivo*, but the potential for RNA sequencing to contribute to our understanding of pneumococcal prophages is evident.

It is also important to note that the full-length prophage prevalence in our study was almost certainly underestimated. Some of the older genomes sequenced by us and others were comprised of many assembly contigs; this made full-length prophage sequence identification very challenging and for this reason we resequenced some of our older genomes to improve sequence quality. We could not do this for all pneumococci, thus will have missed some full-length prophage sequences if they were split over multiple contigs and the presence of other prophage sequences in the same genome meant that correct prophage assemblies could not be guaranteed. DNA sequencing technology and genome assembly tools are continually improving and newer genome assemblies are comprised of only a few contigs, meaning that incomplete prophage sequences will be less frequent and problematic going forward.

The pneumococcal genome collection was compiled to capture population-level diversity and uniquely investigate historical pneumococci, but much of the pneumococcal metadata were unknown and thus it was difficult

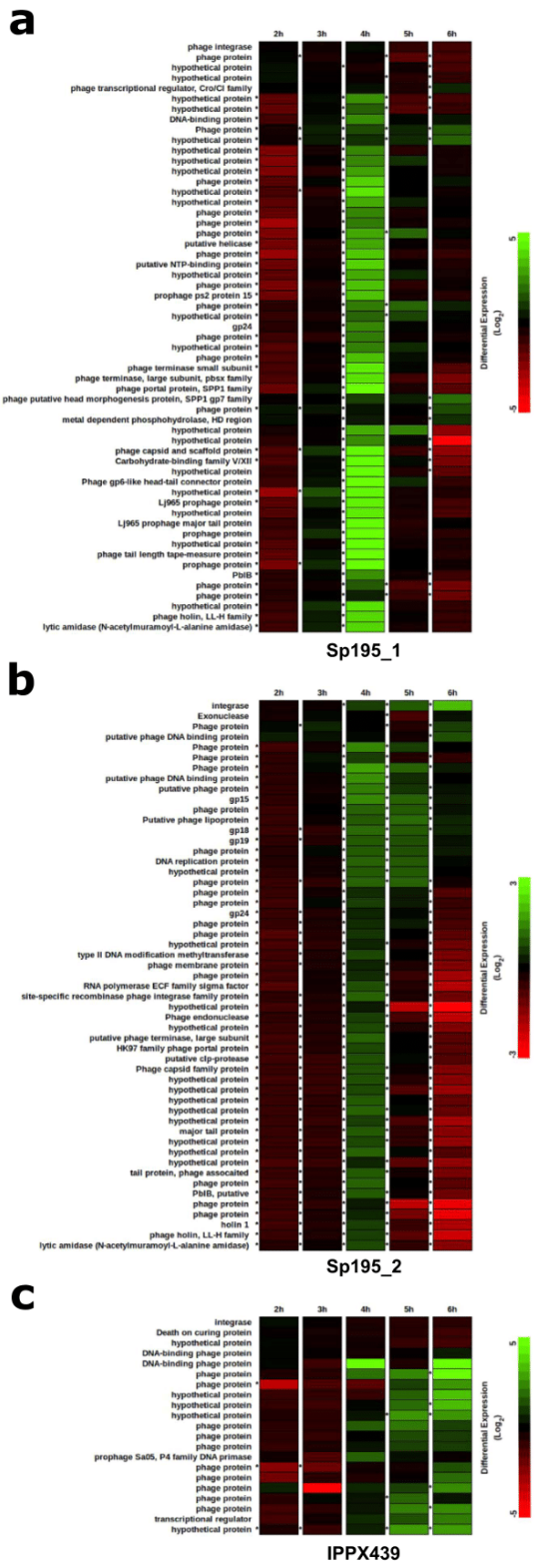


Figure 6. Heat maps describing the results of the RNA-seq experiment. Prophage genes are depicted by rows and differential expression levels at each of five time points are presented in columns. An asterisk to the left of a cell indicates a statistically significant differential level of expression ($p < 0.05$). Prophage gene expression levels are given for two full-length prophages, Sp195_1 (**a**) and Sp195_2 (**b**), and one partial prophage sequence, IPPX439 (**c**). Mitomycin C was added to the broth culture after 3 h of incubation.

to associate specific prophage(s) and pneumococcal phenotype from this dataset. However, our work provides the foundation for future studies to explore associations between unique prophages and pneumococcal phenotypes among well-sampled genome datasets.

Importantly, it will be essential to understand whether the prophage genes with unknown functions are directly contributing to (or strongly influencing) pneumococcal pathogenesis and virulence. Pneumococcal pathogenesis is complicated and whether or not there is a significant prophage contribution is not yet well understood, but the vast array of prophage DNA within these pneumococcal genomes warrants a detailed exploration.

Methods

Compilation of the pneumococcal genome dataset. The study dataset consisted of 482 historical and modern pneumococcal genomes (Table S1) described in our previously published studies[16–17,25–28], but 42 of these were re-sequenced to obtain genome assemblies with fewer contigs than the original assemblies generated some years ago. Full-length prophage sequence detection was difficult if the genomes contained many assembly gaps. In addition, 46 historical pneumococci were added from an isolate collection donated by the Statens Serum Institut, and these were also sequenced as described below. Quality control statistics for all 482 pneumococcal genome sequences are detailed in Table S6.

Pneumococci were cultured using standard protocols and DNA was extracted using the Promega Maxwell® 16 Instrument and Buccal Swab LEV DNA purification kits. Extracted DNA samples were sent to the Oxford Genomics Centre, where libraries were made, genome sequencing was performed on the Illumina® platform, and de novo sequence assemblies were generated using Velvet[29]. Genome assembly quality was further improved using SSPACE and GapFiller[30,31]. Genes were annotated using Prokka and sequences were visualised using Artemis[32,33]. Pneumococcal sequences and metadata were stored in a BIGSdb database[34]. STs were automatically tagged and defined in BIGSdb via links to the PubMLST website[20] and CCs were defined using Phyloviz[35]. The 66 representative full-length phage sequences were deposited in GenBank (accession numbers in Table S2). All assembled pneumococcal genome sequences with corresponding metadata can be accessed from the PubMLST website[20].

Interrogation of pneumococcal genomes for phage sequences. An initial screen of the first 336 genomes in the dataset using PHAST resulted in a prophage DNA hit rate of 86% among the pneumococcal genomes[36]. Subsets of the data were explored further, which led to the identification of ten newly-discovered prophage sequences. It also became clear that while PHAST was generally accurate in detecting prophage sequence, the identification of specific prophages was suboptimal. Therefore, the entire dataset of 482 pneumococcal genomes was screened using a reference prophage database we compiled and an in-house pipeline that facilitated BLAT searches of the pneumococcal genome dataset for evidence of prophage sequence[37]. The prophage reference sequence database consisted of 81 previously-characterised streptococcal species prophage sequences downloaded from GenBank, published prophage sequences obtained from Timothy Mitchell[13], plus the ten new prophage sequences we identified in the initial screen.

The screening pipeline returned the three best BLAT hits per pneumococcal genome and each of these was manually examined. This resulted in the identification of seven more new prophage sequences, which were added to our reference prophage database and the entire genome dataset was rescreened. The screens also returned hits to >1 full-length prophage sequences in different locations within a single pneumococcal genome, all of which were manually inspected to confirm poly-lysogeny (the presence of different prophage sequences in the same pneumococcal genome). The screens also frequently revealed >10 Kb highly similar hits to reference prophages, and investigation of some of these long sequence matches revealed that PHAST and our bespoke screens were missing many full-length prophages, so the gene annotation files for all 482 genomes were manually investigated for evidence of prophage sequences. Ultimately, new prophages were defined as those with a nucleotide sequence <98% identical to any other known prophage in the prophage reference sequence database or our study dataset. Extensive cross-referencing and BLAST searching of previously-identified prophages to the many new prophages identified by manual inspection confirmed the presence and novelty of the detected prophages.

Altogether, these analyses revealed three categories of prophage sequences. 'Full-length' prophage DNA sequences were defined as those that started with an integrase gene, ended with an amidase or lysin gene and were >28 kb in length. 'Putatively full-length' prophages were defined as those that started with an integrase and nearly met the full-length criteria; however, there was an assembly gap at the end (usually at or around *pblB*) such that the last few gene sequences were missing, but these could be found elsewhere in the genome assembly. However, since many genomes contained multiple prophage sequences it was often impossible to reconnect with confidence the two or more prophage sequence fragments bioinformatically, so only the longest contiguous part of the prophage sequence was extracted and labelled as IPPX(n). 'Partial' prophage sequences contained a series of contiguous phage-associated genes, may or may not have included an integrase or amidase gene, were between 2–25 Kb in length, and were also labelled as IPPX(n).

Genomic analyses of phage sequences. 93 full-length prophages were identified and the individual gene nucleotide sequences were clustered using Roary set at a 90% similarity threshold[38]. The 66 unique full-length prophage genomes were visually confirmed and reverse-complemented as required in order to be in the same orientation using Artemis. Multiple sequence alignment (MSA) was performed in Geneious version 9.1 (Biomatters Ltd) using the ClustalW algorithm with default parameters (Gap open cost = 15, Gap extend cost = 6.66)[39]. The MSA output was used within Geneious to calculate percentage identity matrices and create gene alignment figures. Tree building was performed using the Jukes–Cantor model with default parameters in FastTree[40]. Roary was used to identify all of the genes present within each prophage cluster and then an in-house Python script was developed to identify those genes found in more than one prophage cluster. Figures were edited using Inkscape[41].

RNA-sequencing experiments and analyses. Seven 10 ml tubes of brain-heart infusion broth were inoculated with pneumococcal reference strain PMEN3 and incubated at 37 °C + 5% CO_2. An aliquot of 0.5 ml broth was removed and the absorbance at OD_{600} was measured at each time point from 0–6 h to measure increased bacterial growth. Mitomycin C (Sigma-Aldrich) was added to the broth culture tubes to a final concentration of 2 μg/ml after 3 h of incubation ($OD_{600} \geq 0.5$). Just prior to the addition of mitomycin C at 3 h and after 4, 5 and 6 h of incubation, respectively, broth cultures were removed from the incubator for processing. A 0.5 ml aliquot was used to measure the absorbance and the RNA was stabilised in the remaining 9.5 ml of broth culture by the addition of 19 ml of RNAprotect Bacteria Reagent (Qiagen). RNA was immediately extracted from the samples using the Promega Maxwell® 16 Instrument and LEV simplyRNA Cells purification kit, following the manufacturer's protocol.

RNA extracts were sent to the Oxford Genomics Centre where library preps were made using RNA-Seq Ribozero kits (Illumina, Inc) and sequencing was performed on the MiSeq (Illumina, Inc). The sequenced forward and reverse reads were paired and mapped to reference genomes using Bowtie2 with the highest sensitivity option[42]. Differential gene expression was assessed in Geneious using the DESeq method[43]. Genes with an adjusted p-value < 0.05 were deemed to be differentially expressed. RNAseq data are provided in Table S7. Sequence data have been deposited in the Gene Expression Omnibus repository with accession numbers GSM2360598-GSM2360607[44].

References

1. Anthony, D. & Mullerbeck, E. Committing to child survival: a promise renewed. *United Nations Children's Fund (UNICEF)* (2013).
2. Liu, L. *et al.* Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. *Lancet* **385,** 430–440 (2015).
3. World Health Organization. Pneumococcal vaccines WHO position paper – 2012. *Wkly. Epidemiol. Rec.* **87,** 129–144 (2012).
4. Weinberger, D. M., Malley, R. & Lipsitch, M. Serotype replacement in disease after pneumococcal vaccination. *Lancet* **378,** 1962–73 (2011).
5. Kang, C. I. & Song, J. H. Antimicrobial resistance in Asia: current epidemiology and clinical implications. *Infect. Chemother.* **45,** 22–31 (2013).
6. Calendar, R. *The Bacteriophages*, 2nd edition. New York: Oxford University Press (2006).
7. Casjens, S. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbial.* **49,** 277–300 (2003).
8. Bensing, B. A., Siboo, I. R. & Sullam, P. M. Proteins PblA and PblB of *Streptococcus mitis*, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect. Immun.* **69,** 6186–92 (2001).
9. Beres, S. B., *et al.* Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc. Natl. Acad. Sci. USA* **99,** 10078–83 (2002).
10. McDonnell, M., Ronda, C. & Tomasz, A. "Diplophage": a bacteriophage of *Diplococcus pneumoniae*. *Virology* **63,** 577–582 (1975).
11. Bernheimer, H. P. Lysogeny in pneumococci freshly isolated from man. *Science* **195,** 66–68 (1977).
12. Ramirez, M., Severina, E. & Tomasz, A. A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*. *J. Bacteriol.* **181,** 3618–25 (1999).
13. Romero, P., García, E. & Mitchell, T. J. Development of a prophage typing system and analysis of prophage carriage in *Streptococcus pneumoniae*. *Appl. Environ. Microbiol.* **75,** 1642–9 (2009).
14. Obregon, V., García, J. L., García, E., Lopez, R. & García, P. Genome organization and molecular analysis of the temperate bacteriophage MM1 of *Streptococcus pneumoniae*. *J. Bact.* **185,** 2362–2368 (2003).
15. Loeffler, J. M. & Fischetti, V. A. Lysogeny of *Streptococcus pneumoniae* with MM1 phage: improved adherence and other phenotypic changes. *Infect. Immun.* **74,** 4486–4495 (2006).
16. Wyres, K. L. *et al.* The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. *Genome Biol.* **13,** R103 (2012).
17. Wyres, K. L. *et al.* Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated prior to 1974. *BMC Genomics* **14,** 500 (2013).
18. Hsieh, Y.-C., Lin, C.-M. & Wang, J.-T. Identification of PblB mediating galactose-specific adhesion in a successful *Streptococcus pneumoniae* clone. *Sci. Rep.* **5,** 1226 (2015).
19. Ndlangisa, K. M., du Plessis, M., Wolter, N., de Gouveia, L., Klugman, K. P., von Gottberg, A. & GERMS-SA. Population snapshot of *Streptococcus pneumoniae* causing invasive disease in South Africa prior to introduction of pneumococcal conjugate vaccines. *PLoS One* **18,** 9, e107666 (2014).
20. Jolley, K. A. & Maiden, M. C. PubMLST: Public databases for molecular typing and microbial genome diversity. http://pubmlst.org/spneumoniae/ (Date of access: 01/10/2016).
21. McGee, L. *et al.* Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. *J. Clin. Microbiol.* **39,** 2565–71 (2001).
22. Brueggemann, A. B., Griffiths, D. T., Meats, E., Peto, T., Crook, D. W. & Spratt, B. G. Clonal relationships between invasive and carriage *Streptococcus pneumoniae*, and serotype- and clone-specific differences in invasive disease potential. *J. Infect. Dis.* **187,** 1424–32 (2003).
23. Hatfull G. F. Dark matter of the biosphere: the amazing world of bacteriophage diversity. *J. Virol.* **89,** 8107–10 (2015).
24. Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E. & Hatfull, G. F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Nat. Acad. Sci.* **96,** 2192–2197 (1999).
25. Bogaardt, C., van Tonder, A. J. & Brueggemann, A. B. Genomic analyses of pneumococci reveal a wide diversity of bacteriocins – including pneumocyclicin, a novel circular bacteriocin. *BMC Genomics* **16,** 554 (2015).
26. van Tonder, A. J. *et al.* Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Comput. Biol.* **10,** e1003788 (2014).
27. van Tonder, A. J. *et al.* Genomics reveals the worldwide distribution of multidrug-resistant serotype 6E pneumococci. *J. Clin. Microbiol.* **53,** 2271–85 (2015).
28. van Tonder, A. J. *et al.* Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5,405 pneumococcal genomes. *Microbial Genomics* **2** (2016).
29. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
30. Boetze, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15,** 211 (2014).
31. Boetze, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol.* **13,** R56 (2012).
32. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).
33. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).

34. Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinform* **11,** 595 (2010).
35. Francisco, A. P. *et al.* PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinform.* **13,** 87 (2012).
36. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: A Fast Phage Search Tool. *Nucl. Acids Res.* **1,** 6 (2011).
37. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
38. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3 (2015).
39. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* **2,** Unit 2.3 (2002).
40. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One.* **5,** e9490 (2010).
41. Inkscape: an open-source vector graphics editor. www.inkscape.org/ (Date of access: 15/10/2016).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* **9,** 357–9 (2012).
43. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
44. Gene Expression Omnibus: a public functional genomics data repository. http://www.ncbi.nlm.nih.gov/geo/ (Date of access: 25/10/2016).

Acknowledgements

Historical isolates of pneumococci were kindly donated by Steen Hoffman and Lotte Lambertsen at the Statens Serum Institut, Copenhagen. This work was supported by a Wellcome Trust fellowship (083511/Z/07/Z) to ABB and a University of Oxford John Fell Fund award (123/734) to ABB. BIGSdb genome database support is provided by a Wellcome Trust Biomedical Research Fund award (04992/Z/14/Z) awarded to Martin JC Maiden, Keith A Jolley and ABB at the University of Oxford. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant 090532/Z/09/Z) for the generation of the RNA sequencing data and selected DNA sequencing data.

Author Contributions

C.L.H. and A.J.v.T. prepared D.N.A. extracts for genome sequencing. A.J.v.T. developed the prophage screening pipeline. C.L.H., A.J.v.T. and A.B.B. performed R.N.A. sequencing experiments. R.R.J., AJ.v.T. and A.B.B. analysed RNA sequencing data. R.R.J. prepared manuscript figures. All authors analysed data. A.B.B. wrote the manuscript and all authors reviewed the manuscript before submission.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Brueggemann, A. B. *et al.* Pneumococcal prophages are diverse, but not without structure or history. *Sci. Rep.* **7**, 42976; doi: 10.1038/srep42976 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis

Reza Rezaei Javan[1], Elisa Ramos-Sevillano[2], Asma Akter[3], Jeremy Brown[2] and Angela B Brueggemann[1,3,4]

[1]Nuffield Department of Medicine, University of Oxford
[2]UCL Respiratory, Division of Medicine, University College London
[3]Department of Medicine, Imperial College London
[4] Nuffield Department of Population Health, University of Oxford

Corresponding author: angela.brueggemann@ndph.ox.ac.uk

Abstract

Prophages (viral genomes integrated within a host bacterial genome) are abundant within the bacterial world and are of interest because they often confer various phenotypic traits to their hosts, such as by encoding genes that increase pathogenicity. Satellite prophages are 'parasites of parasites' that rely on the bacterial host and another helper prophage for survival. We analysed >1,300 genomes of 70 different *Streptococcus* species for evidence of prophages and identified nearly 800 prophages and satellite prophages, the majority of which are reported here for the first time. We show that prophages and satellite prophages were widely distributed among streptococci, were two clearly different entities and each possessed a structured population. There was convincing evidence that cross-species transmission of prophages is not uncommon. Furthermore, *Streptococcus pneumoniae* (pneumococcus) is a leading human pathogen worldwide, but the genetic basis for its pathogenicity and virulence is not yet fully understood. Here we report that over one-third of pneumococcal genomes possessed satellite prophages and demonstrate for the first time that a satellite prophage was associated with virulence in a murine model of infection. Overall, our findings demonstrate that prophages are widespread components of *Streptococcus* species and suggest that they play a role in pneumococcal pathogenesis.

Main

The genus *Streptococcus* comprises a wide variety of pathogens responsible for causing significant morbidity and mortality worldwide[1]. Some of the most important species causing disease in humans include: *Streptococcus pneumoniae* (pneumococcus), a leading cause of pneumonia, bacteraemia, and meningitis[2]; *Streptococcus pyogenes* (group A streptococci), a major cause of pharyngitis, scarlet fever and necrotising fasciitis[3]; and *Streptococcus agalactiae* (group B streptococci), the most common cause of neonatal sepsis[4]. Additionally, *Streptococcus suis* and *Streptococcus equi* rarely cause disease in humans but are important animal pathogens[1].

Bacteriophages (phages) are intracellular parasites of bacteria. Lytic phages hijack the host bacterial machinery, produce new phages and destroy the infected bacterial cell, whereas lysogenic phages do not necessarily initiate replication immediately upon host entry and may integrate their genome within the bacterial genome to be activated at a later stage. An integrated phage is termed a prophage and those genes can be passed down to the bacterial daughter cells. Since survival depends on their bacterial hosts, prophages often express genes that increase host cell fitness[5,6]. Prophages can exert a range of phenotypic effects on the host bacteria: encode toxins that increase virulence[5], promote binding to human platelets[7] or cells[8], evade immune defences[9,10], or protect from oxidative stress[11]. Prophage integration can also regulate bacterial populations by altering bacterial gene expression[12,13].

Prophages and their hosts, like other predator and prey relationships, are embroiled in a complex evolutionary arms race whereby bacteria evolve various strategies to defend themselves and prophages co-evolve to overcome these barriers[14]. These coevolutionary dynamics are complicated by satellite prophages, which lack all the necessary genetic information to replicate on their own and are reliant on hijacking the machinery of another inducing 'helper' prophage to replicate. Satellite prophages might be thought of as 'parasites of parasites'[15,16].

Satellite prophages adversely interfere with helper prophage replication and thus promote bacterial survival[17-19]. Satellite prophages have been discovered through different circumstances and thus there are different terms used to describe this particular type of mobile genetic element (MGE) in the literature, including *Staphylococcus aureus* pathogenicity islands (SaPIs), phage-related chromosomal islands (PRCIs) and phage-inducible chromosomal islands (PICIs), among others[17-23].

Satellite prophages have been shown to be vectors for the spreading of toxin genes and other virulence factors, e.g. SaPI1, which possesses the gene responsible for causing toxic shock syndrome[24]. The prevalence, diversity, genetic stability and molecular epidemiology of satellite prophages in streptococcal species are largely unknown. A small number of satellite prophages have been identified in streptococcal species, although whether they are associated with virulence remains to be investigated[25]. Previous work has shown that prophage-related sequences are highly prevalent within pneumococcal[26-28], *S. pyogenes*[29,30] and *S. agalactiae* genomes[31]; however, genus-wide analyses of the genomic diversity and population structure of streptococcal prophages have not yet been reported.

Here we report the discovery of nearly 800 prophages among >1,300 streptococcal genomes and provide detailed insights into prophage genomics and population structure. Using the pneumococcus as the model organism, the molecular epidemiology of satellite prophages was investigated within a large globally-distributed collection of pneumococci isolated over a 90-year period. Finally, we demonstrated that a satellite prophage was associated with virulence in a murine infection model.

Results

Prophage sequences are a significant component of the genomes of clinically-relevant *Streptococcus* species

We analysed 1,306 genomes from 70 different streptococcal species and identified 415 full-length prophages and 348 satellite prophage genomes (Supplementary Table 1). We estimated the prophage gene content within each streptococcal genome and this revealed a substantial difference in the average prophage content among various streptococcal species, ranging from 0.4% of the *Streptococcus thermophilus* genome to 9.5% of the *S. pyogenes* genome (Figure 1a; Supplementary Table 2). Furthermore, we observed significant variability in prophage content among different genomes of the same bacterial species, e.g. full-length prophages comprised up to 19% of the genes in some *S. pyogenes* genomes, while in others they made up <1% of the genome (Figure 1a). The prevalence of satellite prophages ranged from 0.1% among *Streptococcus mutans* and *Streptococcus sanguinis* genomes to 4.5% of the *Streptococcus dysgalactiae* genomes (Figure 1a).

Full-length and satellite prophages are separate entities with little effective genetic exchange between them

Satellite prophages had a lower guanine (G) and cytosine (C) content than full-length prophages and were about a third of the size in terms of both length of sequence and the number of genes they harboured (Figure 1b). Due to their relatively small genome and apparent lack of essential genes, streptococcal satellite prophage sequences have historically often been regarded as "remnant" or "defective" prophages in a state of mutational decay[13,22,32-34]. Our data reveal that satellite prophage sequences can be highly conserved over many decades, e.g. one satellite prophage was present among pneumococcal genomes with isolation dates ranging from 1939 to 2006 and had maintained >99.98% nucleotide similarity across its entire genome (Figure 1c), suggesting that it is under very strong evolutionary pressure and likely provides an important biological function. The highly conserved nature of this satellite prophage is particularly striking given that the pneumococcus has long been known to be a highly recombinant organism[35-36].

An unrooted phylogenetic tree of all streptococcal prophage genomes in our dataset depicted full-length and satellite prophages as two clearly distinct groups (Figure 1d). We observed that the genes of satellite prophages are unique and differ to those of full-length prophages, as 93% of all satellite prophage genes (>70% amino acid sequence similarity) are not found in any full-length prophages (Figure 1e). Taken together, these findings confirm that satellite prophage sequences are not recent remnants of previous lysogenisation by full-length prophages, but rather that they belong to a unique family of mobile genetic elements.

Streptococcal prophages have a structured population

We found that both full-length and satellite streptococcal prophages demonstrated well-conserved patterns in genome organisation and synteny, regardless of the species that they were isolated from (Figure 2a). Similar to other non-streptococcal prophages (Supplementary Figure 1), genes encoding specific functions were often found clustered together in the prophage genome, although note that the function of many genes is still unknown and therefore the delineation of discrete gene clusters remains problematic (Figure 2a). Whole genome comparisons of all prophage sequences in our dataset depicted major and minor clusters for both full-length and satellite prophages (Figure 2b; Supplementary Figure 2).

Phages are generally believed to be bacterial species-specific and even specific to genetic lineages within a single bacterial species[37]. Surprisingly, we often found prophages from different bacterial species within the same phylogenetic cluster, suggesting that cross-species transmissions are more common among streptococcal prophages than previously realised. Remarkably, despite the relatedness of their prophages, the bacterial hosts were not necessarily the closest phylogenetically-related species (Figure 2b; Supplementary Figure 3). One possible explanation could be that streptococcal prophages are evolving separately from their microbial hosts, and therefore, other factors such as ecological relatedness may dominate over evolutionary relatedness of the host bacteria.

Molecular epidemiology of satellite prophages within a global pneumococcal dataset dating from 1916

We had previously determined the prevalence, diversity and molecular epidemiology of full-length prophages in a global and historical pneumococcal genome dataset[26]. Many shorter prophage sequences were also identified in that study, which were simply classified as partial prophage sequences and not characterised further at the time. Here, we used this genome dataset to further investigate satellite prophages in the context of the pneumococcal population structure. The genome collection was comprised of 482 pneumococci recovered from both healthy and diseased individuals between 1916 and 2009. Pneumococci were isolated from people of all ages residing in 36 different countries. Ninety-one serotypes and 94 different clonal complexes (genetic lineages) were represented in the dataset.

A reinvestigation of the 'partial prophage' sequences resulted in the identification of 44 representative pneumococcal satellite prophages, which clustered into five major groups (Figure 3a). The average GC content of the satellite prophages was lower than their pneumococcal host but varied among each group (Figure 3b).

We found that 35% of the pneumococci in our dataset contained at least one satellite prophage and 5% of the genomes contained two. Some satellite prophages were present in up to six different clonal complexes, whereas others were only found in Singletons (genotypes with no closely related variants; Table 1 and Supplementary Figure 4). Those satellite prophages identified in more than one genome were often found among pneumococci recovered over a decade or more (Table 1). The average prophage content for each of the major clonal complexes ranged from 2.2-6.5%, and with only one exception (CC7232), all of these are widely-circulating pneumococcal genetic lineages (Figure 1c; https://pubmlst.org/spneumoniae).

Prophages are more frequently inserted adjacent to genes involved in information storage and processing

We previously reported that pneumococcal full-length prophages were consistently integrated in specific locations within the genome[26]. Likewise, pneumococcal satellite prophages were consistently integrated in seven precise locations (a-f) within the host genome, each of which was directly associated with the integrase gene they harboured (Figure 3d; Figure 4a). The 44 representative satellite prophage integrases were divided into seven different categories with ≥95% nucleotide sequence similarity within each category. Each integrase category was associated with insertion at a single location on the pneumococcal genome, apart from integrase category I, which was associated with five different locations (Figure 3d). 28.3% of pneumococcal satellite prophages were inserted at site a, which was very close to the origin of replication (oriC) (Figure 4a) and prompted us to investigate whether factors other than the integrase sequence determined the prophage insertion site.

We investigated the location of prophage insertion sites within the genome sequences of non-pneumococcal streptococci for which at least one complete (finished) genome was available (n=29). We divided the genome of each species into eight non-overlapping segments of equal length according to the number of base pairs, and the percentages of prophages situated in each segment were quantified. Overall, we observed no strong preference for prophage insertion in any of the eight segments and the location of prophages residing within the genome varied greatly between different species (Supplementary Figure 5).

Among pneumococcal and non-pneumococcal streptococcal genomes, five flanking genes upstream and downstream of each prophage were retrieved for functional classification using gene ontology analyses. This revealed that nearly one-third of all the bacterial flanking genes were involved in replication, recombination, DNA repair, transcription, translation and ribosomal structure and biogenesis (Figure 4b). One-quarter of flanking genes were involved in metabolic processes, but equally, one-quarter of all flanking genes did not have a defined functional classification. The remaining flanking genes were involved in other cellular processes

and signalling. A list of all prophage insertion sites and their flanking genes is available in Supplementary Table 3.

For comparison, we selected one genome of each of the 70 different streptococcal species, determined the clusters of orthologous groups (COGs) for all streptococcal genes, and then compared those genome-wide streptococcal data to the COGs represented by the prophage flanking genes in the overall dataset. This demonstrated that the distributions of COGs categories were significantly different, and while prophage flanking genes were more likely to be in the information storage and processing COGs category, the most prevalent COGs category among all streptococcal genes was metabolism (32.1% of all genes; Supplementary Table 4).

Satellite prophages and *vapE* are involved in pneumococcal pneumonia and sepsis in a murine infection model

Our investigation of pneumococcal satellite prophage genes led to the identification of a gene that is a homologue of the 'virulence-associated gene E' (*vapE*) in *S. suis*[38]. We investigated *vapE* in *S. suis* genomes and confirmed that it is carried by a satellite prophage. We searched for *vapE* in the representative pneumococcal satellite prophages and found that 30/44 (68.2%) contained *vapE*. To investigate whether the *vapE* homologue in the pneumococcal satellite prophage is also associated with virulence, we performed *in vivo* studies using a murine pneumococcal infection model and one example of a satellite prophage containing *vapE* identified in this study (Figure 5a).

Deletion mutant stains were constructed in a serotype 6B pneumococcal strain, BHN418, which contains a satellite prophage sequence (SpnSP38; GenBank accession number MK448645) and no full-length prophage sequences (see Supplementary Table 5 for details of the gene content of BHN418). Either *vapE* only (*ΔvapE*) or the entire satellite prophage sequence (*ΔSpnSP38)* were replaced by a spectinomycin resistance cassette (*aadA9*) in the BHN418 strain (Figure 5a). For each of the mutant strains a competitive index (CI) was determined using a highly sensitive competitive infection experiment in a mouse model of pneumonia.

The CI was significantly <1 in the lungs after mixed infection with *ΔSpnSP38* and the wild-type serotype 6B or *ΔvapE* and the wild-type serotype 6B, indicating a role for the satellite prophage and *vapE* in the establishment of pneumococcal pneumonia (Figure 5b). To further assess the degree of attenuation in virulence of the *ΔSpnSP38* and *ΔvapE* strains, infection experiments were repeated with pure inocula of each strain in both the pneumonia and sepsis models. There were no significant differences in bacterial CFU recovered from the lungs

of infected mice at 24 h between either mutant and the parental wild-type strain (data not shown) and the majority of the mice developed fatal infection by this point. However, in the sepsis model the mice infected with the wild-type serotype 6B strain had significantly greater blood and spleen CFU than the *ΔSpnSP38* mutant (Figure 5c and 5d), indicating that the satellite prophage is directly involved in pneumococcal virulence during bacterial dissemination in the systemic circulation. Although the *ΔvapE* strain had lower spleen CFU compared to the wild-type, this difference was not statistically significant, suggesting that loss of the whole satellite prophage has a more marked effect on the attenuation of virulence during sepsis than loss of VapE alone.

The satellite prophage is required for optimum growth in sera but not for evasion of complement recognition or phagocytosis

Reduced systemic virulence of *ΔSpnSP38* or *ΔvapE* mutants could reflect poor growth under physiological conditions, or evasion of host innate immune killing, which is largely dependent on complement-mediated neutrophil killing. Using a flow cytometry assay, the binding of complement component C3b was not demonstrably different between the mutant strains and wild-type strain (Figures 5e and 5f). Furthermore, survival of the *ΔSpnSP38* and *ΔvapE* mutants in the presence of neutrophils after 30 min was similar to the wild-type BHN418 strain (Figure 5g). These data indicate that the satellite prophage and *vapE* are not required for evasion of complement or neutrophil killing, and that the reduced virulence of the *ΔSpnSP38* strain could reflect delayed growth in serum. Growth rates of both mutant strains in THY were not significantly different to the parental wild-type strain (Figure 5h); however, culture in serum demonstrated a small but significant delay in growth of the *ΔSpnSP38* strain compared to the wild-type and Δ*vapE* strains (Figure 5i).

Satellite prophage genes, including *vapE*, were overexpressed in planktonic versus biofilm samples

Given the association of the satellite prophage and *vapE* with virulence in our murine pneumococcal infection model, we hypothesised that satellite prophage genes would be overexpressed when pneumococci were grown planktonically in broth versus in a biofilm. To evaluate this hypothesis, we performed comparative transcriptome analyses of planktonic and biofilm pneumococci using an existing RNA sequencing dataset generated by Blanchette and colleagues[39]. In their study, pneumococcal reference strain Sp6A-10, which contained two full-length prophages and one satellite prophage (SpnSP33, 58.7% identical to SpnSP38; GenBank accession number MK448640), was grown planktonically and as a two-day old biofilm. Three biological replicates were collected from each of the growth conditions and the corresponding RNA samples were extracted and sequenced.

We analysed the Blanchette transcriptomic data to assess prophage gene expression under these two experimental conditions, and the data demonstrated significantly higher satellite prophage and full-length prophage gene expression when the host pneumococcus was grown in broth as compared to growth in a biofilm (Figure 6; Supplementary Table 6). The full complement of satellite prophage genes were significantly expressed, and many of the genes of the two full-length prophages, mainly structural and lysis genes, were also significantly upregulated. These gene expression patterns were consistent with the hypothesis that the satellite prophage was exploiting the other full-length prophages in the pneumococcal genome as helper prophages, since the satellite prophage does not possess phage structural genes.

Notably, among the 20 most significantly upregulated genes, 60% (n=12) were satellite prophage genes and *vapE* was the third most upregulated gene in the entire genome. Among the 50 most highly expressed genes, just over half were prophage-related genes: 15 (30%) were satellite prophage genes; 7 (14%) were genes of one full-length prophage; and 4 (8%) were genes of the second full-length prophage (Figure 6; Supplementary Figure 6; Supplementary Table 6). These experimental data further support a significant role for satellite prophages and *vapE* (and full-length prophages) in pneumococcal biology.

Discussion

In this study we sampled a large collection of streptococcal genomes and revealed a diverse collection of full-length prophages and satellite prophages among streptococcal species. What was striking about these findings was that prophages and satellite prophages were two clearly different entities and both had a structured population. Specifically, among pneumococci there were full-length prophages and satellite prophages with persistent associations to major, epidemiologically successful genetic lineages of pneumococci over long periods of time. This is crucial, since these data allow for the exploration of *why* certain combinations of prophages and bacteria exist and whether the prophages might be contributing to the epidemiological success of bacterial genetic lineages.

Our findings suggest that prophages are likely to be influencing bacterial biology and epidemiology to a much greater extent than previously appreciated, given the high proportion of prophage DNA present in many streptococcal species – many of which have not previously been analysed for evidence of prophages. Prophages are mobile genetic elements and genetically similar prophages were frequently detected between different streptococcal species. Historically, the prevailing dogma is that phages have a narrow host range, but our data challenge this view and suggest that prophage transmission across bacterial species is more

common than previously recognised. Other investigators have also recently suggested that some phages may have a broader host range than previously appreciated[40].

Our dataset was designed to be comprised of streptococci that were genetically different and geographically widely distributed, rather than from a very defined population. These data demonstrated high prophage diversity overall, given the breadth and depth of the dataset, and what was remarkable was the similarity among prophages in different bacterial species. In the context of a highly diverse dataset, there are two plausible explanations for finding the same or highly similar prophages in different species, the most likely of which is cross-species transmissions of prophages or at least prophage sequences. The alternative explanation is a shared common ancestor, but this is far less likely given the overall variation among prophage sequences, at least on any reasonable time frame. The implications of these findings are that host specificity should be taken into account when trying to understand the precise role of prophages in streptococcal biology and when considering whether phages might be used in any therapeutic interventions.

Many of the streptococci we investigated are important human and animal pathogens, raising the question whether prophages influence host virulence potential. To investigate this, we assessed the effects of deleting a previously unrecognised pneumococcal satellite prophage sequence on virulence in a murine model of infection. This prophage contains *vapE*, a gene that has previously been described to have a role in *S. suis* virulence through an unknown mechanism. The results showed that deletion of the whole prophage or *vapE* alone had a significant effect on pneumococcal virulence, and deletion of the whole prophage had a particularly strong effect and reduced recovered CFU for the sepsis model by approaching 10^4 $\log_{10}$. *In vitro* characterisation of the mutant strains indicated that the reduced virulence of the prophage mutant was related to impaired growth in serum rather than avoidance of opsonophagocytic killing. How the prophage influences pneumococcal growth in serum will require more detailed investigation, but the stronger phenotypic effect of loss of the whole prophage compared to *vapE* alone suggests that additional prophage genes are involved in virulence. For example, the prophage is predicted to contain regulatory genes, which could potentially improve growth in serum by altering the expression of metabolic and transporter genes.

Furthermore, when we analysed the transcriptomic data from Blanchette *et al*, these data demonstrated that all satellite prophage genes (including *vapE*) and many genes of the two full-length prophages were among the most significantly upregulated among pneumococci growing in planktonic form, which is akin to pneumococcal bacteraemia, rather than in a biofilm (a state in which pneumococci are less likely to be virulent[41]). While the specific mechanism driving virulence is not yet clear, this work provides clear evidence

that experimental investigations of pneumococcal prophages and satellite prophages can reveal central aspects of the bacteria/prophage relationship among pneumococci and other streptococci.

The increasingly large volume of genome sequence data in the public domain presents many new opportunities for understanding bacterial infection and pathogenesis at a depth and breadth never before experienced. Large population-level analyses such as this alter our perspective on how bacterial and prophage populations interact and drive evolution of both parasite and host. As demonstrated here, population genomics studies can and should be used to generate hypotheses, design experiments, and select the most appropriate strains for testing. The findings of this study reveal numerous areas for further investigation, the results of which will increase our knowledge of prophage and bacterial biology, epidemiology and evolution.

Methods

Development of PhageMiner, a bioinformatics tool for prophage identification in bacterial genomes

Some *in silico* prophage detection tools are available that identify prophages by comparison to a reference database of known prophage genomes, thus their performance is strongly influenced by the size and composition of the reference dataset[42-43]. In order to ensure a thorough discovery of previously unidentified prophages, manual curation of annotated genomes is required, however, this is not feasible for large genome studies[26,44-45]. To address these issues, we developed a user-supervised semi-automated computational tool called PhageMiner in order to streamline the manual curation process for prophage sequence discovery. PhageMiner uses a mean shift algorithm combined with annotation-based genome mining in order to rapidly identify prophage sequences within complete or draft bacterial genomes.

In brief, the PhageMiner pipeline proceeded as follows: 1) bacterial genomes were scanned for the presence of phage-related genes based on genome annotation; 2) the presence of multiple phage-related genes in close proximately within the host genome was used as a possible indication that a prophage genome might be present in that location; 3) suspected prophage regions were displayed to the user via diagrams and tables for further manual inspection; and 4) based on the inputs provided by the user, the suspected prophage sequences were either rejected, or extracted and named systematically. Notably, while PhageMiner significantly facilitates the manual curation process, it is not a fully automated pipeline and requires manual input by the user to make key decisions, thus ensuring careful inspection of putative prophage clusters. A detailed description of the PhageMiner pipeline is provided in Supplementary Methods. The source code of PhageMiner is deposited in GitHub (https://github.com/RezaRezaeiJavan/PhageMiner).

Genomes used in this study

In total, 1,316 assembled genomes from 70 different species of the genus *Streptococcus* were selected for this study. 482 genomes belonged to a pneumococcal dataset previously characterised by us[26]. This collection was designed to be highly diverse and consisted of pneumococci recovered from both ill and healthy individuals of all ages residing in 36 different countries between 1916 and 2009. These pneumococci represented 91 serotypes and 94 different clonal complexes (Supplementary Table 7).

The remaining 834 streptococcal genomes were selected from a non-pneumococcal *Streptococcus* species genome dataset previously compiled by us[46]. In brief, 69 different *Streptococcus* species were included in this dataset and up to 50 genomes per species were selected for analyses from the ribosomal MLST database (https://pubmlst.org/rmlst)[47]. When more than 50 genomes were available, the population structure of the species was depicted using PHYLOViZ[48] and genomes were selected to maximise the population-level diversity of the species from the available genomes. All streptococcal genome sequences were stored in a BIGSdb database[49] and annotated using the RAST server (http://rast.nmpdr.org).

Sequence analyses of prophages

All putative prophage sequences were inspected manually using Geneious version 11.1 (Biomatters Ltd.) and those containing ambiguous bases (N's) and/or assembly gaps (n = 411) were excluded from further analyses. The total number of open reading frames (ORFs), overall sequence length and GC content of each prophage were calculated within the Geneious environment. All multiple sequence alignments were performed using ClustalW (version 2.1)[50] with default parameters (Gap open cost = 15, Gap extend cost = 6.66). Phylogenetic trees were constructed based upon sequence alignments using FastTreeMP (version 2.1.5)[51]. Unique integrase sequences were identified using the CD-HIT program (version 4.6.6)[52] and a threshold of ≥95% sequence identity. Schematic diagrams of the coding regions of the prophages were produced in Geneious and edited using Adobe Illustrator.

Estimation of prophage content within bacterial genomes

The phage content was estimated based on the percentage of prophage genes within a given bacterial genome. To do this, we developed a Python script that first used Prodigal software in the Prokka annotation suite (version 1.10)[53] to predict ORFs in three separate groups of sequences: (i) all identified full-length prophage genomes, (ii) all identified satellite prophage genomes and (iii) a single bacterial genome of interest

for which the phage content is to be estimated. Next, the individual ORF nucleotide sequences from all three groups were extracted, combined and clustered using Roary[54] set at a 70% similarity threshold. Any ORFs in the bacterial genome that were also present in at least one prophage genome were deemed to be phage-related, and this information was used to output the total percentage of phage-related ORFs in the given bacterial genome. The PhageContentCalculator script is available in GitHub (https://github.com/RezaRezaeiJavan/PhageContentCalculator).

Investigation of prophage insertion sites and functional annotation of the flanking bacterial genes

The prophage insertion sites within the bacterial genomes were investigated among the representative pneumococcal prophages and any streptococcal species for which at least one complete bacterial genome was available. Prophage insertion sites containing ambiguous bases or assembly gaps were excluded from the analyses. In order to assess the relative location of prophages within streptococcal bacterial genomes, the genomes were divided into 8 equally-sized segments and the prevalence of prophages per segment was calculated.

To investigate the location of prophages relative to the putative function of the flanking bacterial genes, the sequences of the five bacterial genes both upstream and downstream of each prophage were retrieved. Bacterial gene sequences were categorised into Clusters of Orthologous Groups (COGs) using eggNOG-mapper, which is based on eggNOG 4.5 orthology data[55-56]. For comparison, a reference set of 70 streptococcal genomes, each representing a different streptococcal species, was compiled. All bacterial genes were assigned a COGs category using eggNOG and the average prevalence of each COG category across the combined set of 70 reference streptococcal genomes was calculated.

Construction of a pneumococcal core genome phylogenetic tree

The 482 pneumococcal genomes in the study dataset were annotated using Prokka in order to create GFF3 files compatible with downstream analysis scripts. Genes present in all strains were clustered at 90% sequence identity threshold and aligned using Roary. The phylogenetic tree was generated using FastTreeMP[51] using a generalized time-reversible model and then was reconstructed using ClonalFrameML (version 1.11)[57] to account for recombination. The tree was annotated using iTOL (version 4.3.3)[58] and Adobe Illustrator (Adobe Inc.).

Estimate of phylogenetic relationships among *Streptococcus* species

A phylogenetic tree was constructed using concatenated sequence data from 53 ribosomal loci among all streptococcal genomes in the study dataset using the BIGSdb PhyloTree plugin. The tree was graphically simplified to the species level by collapsing clades containing genomes from the same species into a single leaf using iTOL.

Bacterial strains, media and growth conditions

Pneumococcal strains were cultured in the presence of 5% CO_2 at 37°C on Columbia agar (Oxoid) supplemented with 5% horse blood, or in Todd-Hewitt broth supplemented with 0.5% yeast-extract (THY; Oxoid). Mutant strains were selected by using 150 µg/ml spectinomycin. Growth of pneumococcal strains in broth was monitored by measuring optical density at 580 nm (OD_{580}) and stocks of pneumococci were stored as single use 0.5 ml aliquots of THY broth culture (OD_{580} 0.4–0.5) at −70°C in 10% glycerol. Data for growth curve measurements were collected using 96-well plates in a Tecan Spark microtiter plate reader essentially as described before[59], measuring the optical density at 595 nm (OD_{595}) in 30 min intervals. For growth in THY and serum, 10^6 CFU of each strain was added to 200 µl of medium or serum and incubated at 37°C plus 5% CO_2.

Construction of Δ*vapE* and Δ*SpnSP38* pneumococcal mutant strains

Strains, plasmids and primers used for this study are described in Supplementary Table 8. Both mutants, Δ*vapE* and *ΔSpnSP38,* were generated by overlap extension PCR[60-61] in the pneumococcal serotype 6B BHN418 strain (multilocus sequence type (ST)138) using a transformation fragment in which the *Spn_00749* gene (*vapE*) or the entire satellite prophage, *Spn_00738-Spn_00753*, were replaced by the spectinomycin resistance cassette *aadA9*. For the satellite prophage, two products corresponding to 762 bp upstream (primers SpnSP_UpF and SpnSP_UpspecR) and 872 bp downstream (primers SpnSP_Downspec_F and SpnSP_DownR) of the satellite prophage were amplified from pneumococcal genomic DNA by PCR carrying 3' and 5' linkers complementary to the 5' and 3' portion of the *aacA9* gene respectively. *aadA9* was amplified from the pR412 plasmid (a gift from M. Domenech) using PCR and primers SpnSP_Upspec_F and SpnSP_Downspec_R[60].

Similarly, for the in-frame deletion of *vapE*, a construct was created in which 820 bp of flanking DNA upstream of the *vapE* ATG (primers VapE_UpF and VapE_UpspcR) and 526 bp of flanking DNA downstream

from the *vapE* ORF (starting from the ATG of the overlapping Spn_00750 ORF, primers VapE_DownspecF and VapE_DownR) were amplified by PCR and fused with the *aadA9* cassette by overlap extension PCR[62]. The resulting constructs were then transformed into the BHN418 strain by homologous recombination and allelic replacement using a mix of CSP-1 and CSP-2 and standard protocols[63-64]. The mutations were confirmed by PCR analysis and sequencing.

Experimental models of infection

6-week-old female CD-1 mice were obtained from Charles River Laboratory and bred in a conventional animal facility at University College of London (UCL). Animal procedures were performed according to United Kingdom (UK) national guidelines for animal use and care and approved by the UCL Biological Services Ethical Committee and the UK Home Office (Project Licence PPL70/6510). Studies investigating pneumococcal sepsis or pneumonia were performed using 6-week-old mice and infected as previously described[65].

Briefly, in the sepsis model, mice were challenged with 5×10^6 CFU/ml of the serotype 6B strain or the correspondent mutants in a volume of 150 µl by the intraperitoneal route, whereas for pneumonia, mice under anaesthesia with isofluorane were inoculated intranasally with 50 µl containing 10^7 CFU/mouse of the serotype 6B strain or the mutants. A lethal dose of pentobarbital was administered at 24 or 28 h after challenge and bacterial counts were determined from samples recovered from lung and blood. Lungs and spleens were homogenised through a 0.2 µm filter. Results were expressed as $\log_{10}$ CFU/ml of bacteria recovered from the different sites.

For mixed infection experiments, mice were inoculated with a 50/50 mixture of wild-type and mutant pneumococci. The competitive index (CI) was defined as the ratio of the test strain (mutant strain) compared to the control strain (wild-type strain) recovered from mice, divided by the ratio of the test strain to the control strain in the inoculum[66-67]. A CI of <1 indicates that the test strain is attenuated in virulence compared to the control strain, and the lower the CI the more attenuated the strain. Statistical analyses were performed using analysis of variance (ANOVA) for multiple comparisons. GraphPad Prism 7.0 (GraphPad Software, San Diego, CA) was used for statistical analyses.

C3b binding to pneumococci

Serum samples from five healthy male volunteer controls (median age 40 y) were obtained according to institutional guidelines and stored as single-use aliquots at -70°C to use as a source of complement. C3b deposition was analysed using a flow cytometry assay[68]. Briefly, C3b deposition was investigated by incubating 10^7 CFU of pneumococci with 10 µl of pooled human serum (diluted to 20% in PBS) for 30 min at 37°C. C3b bound to the different strains was labelled with 50 µl of a 1/500 dilution of fluorescein isothiocyanate-conjugated polyclonal goat anti-human C3b antibody (ICN) after two washes in PBS-Tween 20 (0.01%). The detection of C3b binding was performed using flow cytometry with gating based on the analysis of at least 10,000 bacteria. Experiments were repeated three times and the results were expressed as the proportion of C3b deposition on the surface of the different mutants compared to the C3b deposition on the 6B wild-type strain.

Neutrophil killing assay

Frozen aliquots of pneumococci were thawed and washed twice with PBS-Tween 20 (0.01%) by centrifugation for 5 min at 13,000 rpm. 100 µl of the bacterial suspension, diluted to 10^3 CFU, was added to each well in the presence of 25% baby rabbit complement. After 30 min of incubation at 37˚C, 100 µl of neutrophils (10^5 cells) previously isolated from human blood using MACSxpress® was added to each well and incubated at 37˚C with shaking. Sample aliquots were taken at 15 and 30 min, spotted onto Columbia blood agar plates and incubated at 37°C plus 5% CO_2. Bacterial colony counts were performed after overnight incubation.

Transcriptomic analyses of prophage gene expression when the host pneumococci were grown in broth culture and as a biofilm

The RNA sequencing data used in this study were originally generated by Blanchette *et al*[39]. In brief, samples were collected in three biological replicates from a pneumococcal strain Sp6A-10 isolate (serotype 6A; ST460) growing in Todd-Hewitt broth either planktonically or in polystyrene six-well plates as two-day-old biofilms. Total RNA from each sample was extracted and sequenced using the Illumina HiSeq4000 sequencing platform. For use in the current study, raw RNA sequencing data was retrieved from the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/; accession number GSE85196). Reads from the control planktonic (THB_PK1, THB_PK2, THB_PK3) and biofilm (THB_BF1, THB_BF2, THB_BF3) samples were paired and mapped onto the pneumococcal Sp6A-10 genome using Bowtie2[69] with the highest sensitivity option. Differential gene expressions were computed in Geneious using the DESeq2

method[70]. Genes with an adjusted p value <0.001 were deemed to be differentially expressed. A volcano plot was generated within the Geneious environment and further edited using Adobe Illustrator.

Data availability

The newly discovered full-length and satellite prophage genomes have been deposited at GenBank under accession numbers MK448665-MK449012 (Supplementary Table 3). The sequence of the *vapE* gene is available via GenBank accession number QBX13222.1. Accession numbers for all genomes used in this study are listed in Supplementary Table 2.

Code availability

The PhageMiner script is available at GitHub (https://github.com/RezaRezaeiJavan/PhageMiner). The PhageContentCalculator script is available in GitHub (https://github.com/RezaRezaeiJavan/PhageContentCalculator).

Author contributions

R.R.J. and A.B.B. conceived and designed the overall study. R.R.J. wrote the computer code. E.R.S. and J.B. designed the pneumococcal mutants and animal experiments. E.R.S. and A.A. created the genetic mutants and E.R.S. performed the animal experiments. R.R.J., A.B.B., E.R.S. and J.B. analysed the data. R.R.J. and A.B.B. wrote the manuscript. All authors read and reviewed the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

We thank Prof Carlos Orihuela and Dr Herve Tettelin for providing the genome sequence of pneumococcal strain Sp6A-10 that was used in the transcriptomic analyses.

References

1. Krzyściak W, Pluskwa K, Jurczak A, Kościelniak D. The pathogenicity of the *Streptococcus* genus. *Eur J Clin Microbiol Infect Dis* **32**,1361-1376 (2013).

2. O'Brien K, Wolfson L, Watt J, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T; Hib and Pneumococcal Global Burden of Disease Study Team. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. *Lancet*. **374**, 893-902 (2009).

3. Carapetis J, Steer A, Mulholland E, Weber M. The global burden of group A streptococcal diseases. *Lancet Infect Dis*. **5**, 685-694 (2005).

4. Vornhagen J, Adams Waldorf K, Rajagopal L. Perinatal group B Streptococcal infections: virulence factors, immunity, and prevention strategies. *Trends Microbiol*. **25**, 919-931 (2017).

5. Boyd E, Brüssow H. Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol*. **10**, 521-529 (2002).

6. Casjens S. Prophages and bacterial genomics: what have we learned so far? *Molec Microbiol.* **49**, 277-300 (2003).

7. Bensing B, Siboo I, Sullam P. Proteins PblA and PblB of *Streptococcus mitis*, which promote binding to human platelets, are encoded within a lysogenic bacteriophage. *Infect Immun*. **69**, 6186-6192 (2001).

8. Vaca Pacheco S, Garcıća González O, Paniagua Contreras G. The lom gene of bacteriophage λ is involved in *Escherichia coli* K12 adhesion to human buccal epithelial cells. *FEMS Microbiol Lett*. **156**, 129-132 (2006).

9. Mirold S, Rabsch W, Rohde M, Stender S, Tschape H, Russmann H, Igwe E, Hardt WD. Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic *Salmonella typhimurium* strain. *Proc Nat Acad Sci USA*. **96**, 9845-9850 (1999).

10. Bulgin R, Raymond B, Garnett J, Frankel G, Crepin V, Berger C, Arbeloa A. Bacterial guanine nucleotide exchange factors SopE-like and WxxxE effectors. *Infect Immun*. **78**, 1417-1425 (2010).

11. Figueroa-Bossi N, Uzzau S, Maloriol D, Bossi L. Variable assortment of prophages provides a transferable repertoire of pathogenic determinants in *Salmonella*. *Molec Microbiol*. **39**, 260-272 (2001).

12. Menouni R, Hutinet G, Petit M, Ansaldi M. Bacterial genome remodeling through bacteriophage recombination. *FEMS Microbiol Lett*. **362**, 1-10 (2015).

13. Feiner R, Argov T, Rabinovich L, Sigal N, Borovok I, Herskovits A. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat Rev Microbiol.* **13**, 641-650 (2015).

14. Koskella B, Brockhurst M. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol Rev*. **38**, 916-931 (2014).

15. Varon M, Levisohn R. Three-membered parasitic system: a bacteriophage, *Bdellovibrio bacteriovorus*, and *Escherichia coli*. *J Virol*. **9**, 519-525 (1972).

16. Belfort M. Bacteriophage introns: parasites within parasites? *Trends Genet*. **5**, 209-213 (1989).

17. Novick R. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of *Staphylococcus aureus*. *Plasmid*. **49**, 93-105 (2003).

18. Novick R, Christie G, Penadés J. The phage-related chromosomal islands of Gram-positive bacteria. *Nat Rev Microbiol*. **8**, 541-551 (2010).

19. Penadés J, Christie G. The phage-inducible chromosomal islands: a family of highly evolved molecular parasites. *Ann Rev Virol*. **2**, 181-201 (2015).

20. Frígols B, Quiles-Puchalt N, Mir-Sanchis I, Donderis J, Elena S, Buckling A, Novick RP, Marina A, Penadés JR. Virus satellites drive viral evolution and ecology. *PLOS Genetics*. **11**, e1005609 (2015).

21. O'Neill A, Larsen A, Skov R, Henriksen A, Chopra I. Characterization of the epidemic European fusidic acid-resistant impetigo clone of *Staphylococcus aureus*. *J Clin Microbiol*. **45**, 1505-1510 (2007).

22. Scott J, Nguyen S, King C, Hendrickson C, McShan W. Phage-Like *Streptococcus pyogenes* Chromosomal Islands (SpyCI) and Mutator Phenotypes: Control by Growth State and Rescue by a SpyCI-Encoded Promoter. *Front Microbiol*. **3** (2012).

23. Seed K, Lazinski D, Calderwood S, Camilli A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*. **494**, 489-491 (2013).

24. Lindsay J, Ruzin A, Ross H, Kurepina N, Novick R. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in *Staphylococcus aureus*. *Molec Microbiol.* **29**, 527-543 (1998).

25. Martínez-Rubio R, Quiles-Puchalt N, Martí M, Humphrey S, Ram G, Smyth D, Chen J, Novick RP, Penadés JR. Phage-inducible islands in the Gram-positive cocci. *ISME J*. **11**, 1029-1042 (2016).

26. Brueggemann A, Harrold C, Rezaei Javan R, van Tonder A, McDonnell A, Edwards B. Pneumococcal prophages are diverse, but not without structure or history. *Sci Rep*. **7** (2017).

27. Romero P, García E, Mitchell TJ. Development of a prophage typing system and analysis of prophage carriage in *Streptococcus pneumoniae*. *Appl. Environ. Microbiol*. **75**, 1642–9 (2009).

28. Ramirez M, Severina E, Tomasz A. A high incidence of prophage carriage among natural isolates of *Streptococcus pneumoniae*. *J. Bacteriol*. **181**, 3618–25 (1999).

29. Beres S, Sylva G, Barbian K, Lei B, Hoff J, Mammarella N, Liu MY, Smoot JC, Porcella SF, Parkins LD, Campbell DS, Smith TM, McCormick JK, Leung DY, Schlievert PM, Musser JM. Genome sequence of a serotype M3 strain of group A Streptococcus: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc Nat Acad Sci USA*. **99**, 10078-10083 (2002).

30. McShan WM, Nguyen SV. The bacteriophages of *Streptococcus pyogenes*. In: Ferretti JJ, Stevens DL, Fischetti VA, editors. *Streptococcus pyogenes*: Basic Biology to Clinical Manifestations. University of Oklahoma Health Sciences Center (2016). Available from: https://www.ncbi.nlm.nih.gov/books/NBK333409/

31. van der Mee-Marquet N, Diene S, Barbera L, Courtier-Martinez L, Lafont L, Ouachée A et al. Analysis of the prophages carried by human infecting isolates provides new insight into the evolution of group B *Streptococcus* species. *Clin Microbiol Infect*. **24**, 514-521 (2018).

32. Canchaya C, Desiere F, McShan W, Ferretti J, Parkhill J, Brüssow H. Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain SF370. *Virol*. **302**, 245-258 (2002).

33. Davies E, Winstanley C, Fothergill J, James C. The role of temperate bacteriophages in bacterial infection. *FEMS Microbiol Lett*. **363**, fnw015 (2016).

34. Bobay L, Touchon M, Rocha E. Pervasive domestication of defective prophages by bacteria. *Proc Nat Acad Sci USA* **111**, 12127-12132 (2014).

35. Spratt BG, Maiden MC. Bacterial population genetics, evolution and epidemiology. *Philos Trans R Soc Lond B Biol Sci* **354**, 701-10 (1999).

36. Feil EJ, Smith JM, Enright MC, Spratt BG. Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154**, 1439-50 (2000).

37. Ackermann H, Audurier A, Berthiaume L, Jones L, Mayo J, Vidaver A. Guidelines for Bacteriophage Characterization. *Adv Virus Res* **23**, 1-24 (1978).

38. Ji X, Sun Y, Liu J, Zhu L, Guo X, Lang X, Feng S. A novel virulence-associated protein, VapE, in *Streptococcus suis* serotype 2. *Mol Med Rep*. **13**, 2871-7. (2016)

39. Blanchette KA, Shenoy AT, Milner J, II, Gilley RP, McClure E, Hinojosa CA, Kumar N, Daugherty SC, Tallon LJ, Ott S, King SJ, Ferreira DM, Gordon SB, Tettelin H, Orihuela CJ. Neuraminidase A-exposed galactose promotes *Streptococcus pneumoniae* biofilm formation during colonization. *Infect Immun* **84**, 2922–2932 (2016).

40. Ross A, Ward S, Hyman P. More Is Better: Selecting for Broad Host Range Bacteriophages. *Front Microbiol* **7**, 352 (2016).

41. Gilley RP, Orihuela CJ. Pneumococci in biofilms are non-invasive: implications on nasopharyngeal colonization. *Front Cell Infect Microbiol* **4**, 163 (2014).

42. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*. **24**, 863-865 (2008).

43. Zhou Y, Liang Y, Lynch K, Dennis J, Wishart D. PHAST: A Fast Phage Search Tool. *Nucl Acids Res*. **39**, W347-W352 (2011).

44. Crispim J, Dias R, Vidigal P, de Sousa M, da Silva C, Santana M, de Paula SO. Screening and characterization of prophages in Desulfovibrio genomes. *Sci Rep*. **8** (2018).

45. Langille M, Hsiao W, Brinkman F. Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* **8**, 373-382 (2010).

46. Kurioka A, Wilgenburg B, Rezaei Javan R, Hoyle R, van Tonder AJ, Harrold CL, Leng T, Phalora P, Howson LJ, Shepherd D, Cerundolo V, Brueggemann AB, Klenerman P. Diverse *Streptococcus pneumoniae* strains drive a MAIT cell response through MR1-dependent and cytokine-driven pathways. *J Infect Dis* **217**, 988–999. (2018)

47. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MC. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiol* **158**,1005-15. (2012)

48. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*. **13**:87 (2012).

49. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*. **11**:595 (2010).

50. Thompson JD, Higgins DG, GibsonTJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680 (1994).

51. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLOS ONE.* **5**:e9490 (2010).

52. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. **22**:1658-1659 (2006).

53. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30**:2068-2069 (2014).

54. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. **31**:3691-3693 (2015).

55. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol* **34**, 2115-2122 (2017).

56. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**:D286-293 (2016).

57. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLOS Comput Biol*. **11**:e1004041 (2015).

58. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**:W475-478 (2011).

59. Kjos M, Aprianto R, Fernandes VE, Andrew PW, van Strijp JA, Nijland R, Veening JW. Bright fluorescent *Streptococcus pneumoniae* for live-cell imaging of host-pathogen interactions. *J Bacteriol* **197**, 807-18 (2015).

60. Khandavilli S, Homer KA, Yuste J, Basavanna S, Mitchell T, Brown JS. Maturation of *Streptococcus pneumoniae* lipoproteins by a type II signal peptidase is required for ABC transporter function and full virulence. *Mol Microbiol*. **67**, 541-57 (2008).

61. Basavanna S, Chimalapati S, Maqbool A, Rubbo B, Yuste J, Wilson RJ, Hosie A, Ogunniyi AD, Paton JC, Thomas G, Brown JS. The effects of methionine acquisition and synthesis on *Streptococcus pneumoniae* growth and virulence. *PLOS One*. **8**, e49638 (2013).

62. Heckman KL, Pease LR. Gene splicing and mutagenesis by PCR-driven overlap extension. *Nat Protoc* **2**, 924-32 (2007).

63. Håvarstein LS, Coomaraswamy G, Morrison DA. An unmodified heptadecapeptide pheromone induces competence for genetic transformation in *Streptococcus pneumoniae*. *Proc Natl Acad Sci U S A*. **92**, 11140-4 (1995).

64. Lau GW, Haataja S, Lonetto M, Kensit SE, Marra A, Bryant AP, McDevitt D, Morrison DA, Holden DW. A functional genomic analysis of type 3 *Streptococcus pneumoniae* virulence. *Mol Microbiol* **40**, 555-71 (2001).

65. Ramos-Sevillano E, Urzainqui A, Campuzano S, Moscoso M, González-Camacho F, Domenech M, Rodríguez de Córdoba S, Sánchez-Madrid F, Brown JS, García E, Yuste J. Pleiotropic effects of cell wall amidase LytA on *Streptococcus pneumoniae* sensitivity to the host immune response. *Infect Immun*. **83**, 591-603 (2015).

66. 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.* **175**, 1813-9 (2005).

67. Beuzón CR, Holden DW. Use of mixed infections with Salmonella strains to study virulence genes and their interactions in vivo. *Microbes Infect*. **3**, 1345-52 (2001).

68. Ramos-Sevillano E, Moscoso M, García P, García E, Yuste J. Nasopharyngeal colonization and invasive disease are enhanced by the cell wall hydrolases LytB and LytC of *Streptococcus pneumoniae*. *PLoS One* **6**, e23626 (2011).

69. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).

70. Geneious 11.1.5. https://www.geneious.com

Table 1. Epidemiological characteristics of 44 representative satellite prophages identified among a collection of pneumococcal isolates dating from 1939 onwards.

		Pneumococci						
Satellite Prophage Name	**Satellite Prophage Cluster**	**Clonal Complex (n)**	**Genomes (n)**	**Isolation dates**	**Countries (n)**	**Serotypes (n)**	**Insertion Site**	**Integrase Category**
SpnSP16	A	3	4	1939 - 1982	2	4	b	I
SpnSP3	A	2	3	1981 - 2004	2	2	b	I
SpnSP26	A	2	3	1985 - 2000	1	2	b	I
SpnSP35	A	2	2	1952 - 1952	1	2	b	I
SpnSP43	A	2	2	1939 - 2004	2	2	b	I
SpnSP30	A	1	5	1978 - 1978	1	1	b	I
SpnSP44	A	1	2	1939 - 1962	1	1	b	I
SpnSP7	A	1	1	1968	1	1	b	I
SpnSP25	A	1	1	1999	1	1	b	I
SpnSP19	A	Singleton[a]	2	1939 - 1952	2	2	b	V
SpnSP11	A	Singleton	1	1952	1	1	b	I
SpnSP5	B	5	15	1939 - 2007	3	7	d, g	I
SpnSP29	B	1	15	1978 - 1988	1	2	b	I
SpnSP27	B	1	1	2006	1	1	b	I
SpnSP20	B	Singleton	1	1954	1	1	b	I
SpnSP2	C	2	4	1984 - 2005	3	2	f	VII
SpnSP31	C	2	2	1983 - 2005	1	2	b	I
SpnSP12	C	1	1	1968	1	1	b	I
SpnSP15	C	1	1	1943	1	1	b	I
SpnSP32	C	1	1	1986	1	1	f	VII
SpnSP37	D	5	9	1939 - 1988	4	7	c	II
SpnSP38	D	4	30	1972 - 2006	6	5	c	II
SpnSP6	D	3	8	1939 - 1991	3	3	c	II
SpnSP23	D	2	11	1962 - 2008	3	4	a	III
SpnSP39	D	1	2	2005 - 2007	1	1	a	III
SpnSP18	D	Singleton	2	1939 - 1952	2	2	c	II
SpnSP24	E	6	23	1939 - 2006	6	4	a	III
SpnSP33	E	2	3	1952 - 1998	1	2	a	III
SpnSP1	E	1	5	1978 - 1988	1	1	b	I
SpnSP40	E	1	3	2001	2	2	a	III
SpnSP8	E	1	1	1988	1	1	a	III
SpnSP9	E	1	1	1957	1	1	a	III
SpnSP13	E	1	1	1943	1	1	a	III
SpnSP14	E	1	1	1995	1	1	a	III
SpnSP17	E	1	1	1972	1	1	a	IV
SpnSP22	E	1	1	1971	1	1	a	III
SpnSP28	E	1	1	2003	1	1	a	III
SpnSP34	E	1	1	1990	1	1	a	III
SpnSP36	E	1	1	1963	1	1	a	III
SpnSP42	E	1	1	1994	1	1	a	III
SpnSP4	E	Singleton	1	1982	1	1	e	I
SpnSP10	E	Singleton	1	N/A	1	1	h	I
SpnSP21	E	Singleton	1	1954	1	1	e	VI
SpnSP41	E	Singleton	1	1983	1	1	a	III

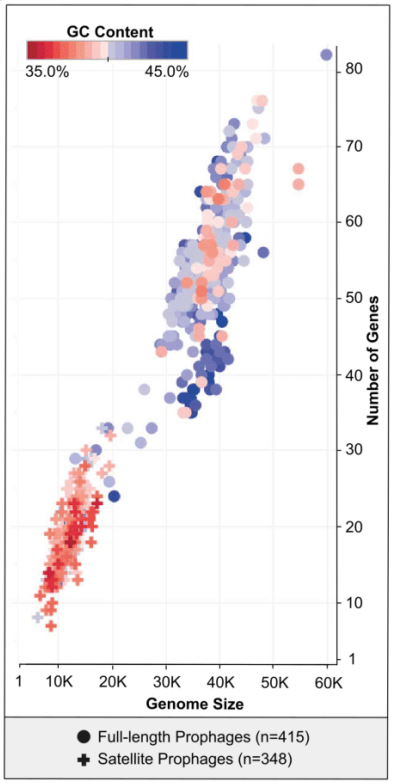
a. Singletons are genotypes with no closely related variants.

785

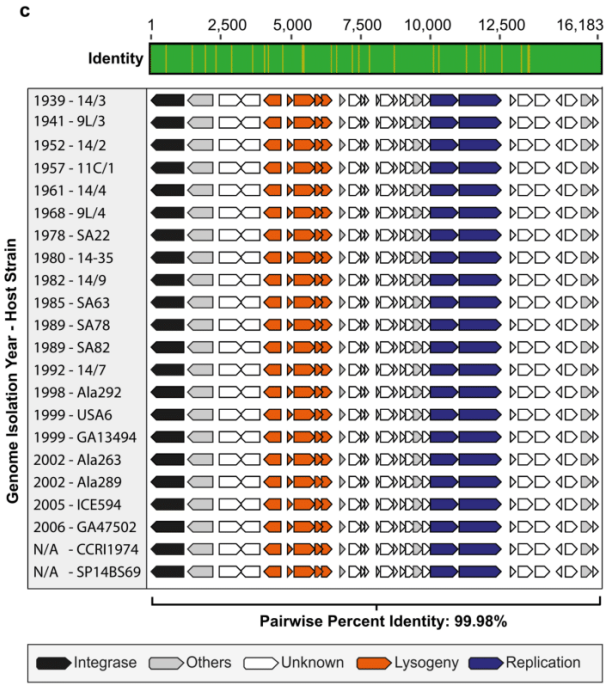
a

Species	Full-length Prophages				Satellite Prophages			
	Min	Max	Mean	SD	Min	Max	Mean	SD
S. pyogenes(50)	1.0%	19.0%	9.5%	4.1	0.4%	3.3%	1.3%	0.9
S. urinalis (4)	6.8%	9.1%	8.0%	1.0	1.5%	1.6%	1.5%	0.1
S. pseudopneumoniae (16)	3.9%	7.0%	5.4%	1.2	0.7%	2.8%	1.7%	0.7
S. parauberis (16)	0.4%	10.5%	5.4%	2.9	0.4%	3.2%	1.6%	1.0
S. canis (11)	2.0%	9.5%	4.5%	2.9	2.0%	2.6%	2.3%	0.2
S. dysgalactiae (50)	0.7%	11.9%	3.8%	2.3	0.3%	4.5%	1.3%	0.8
S. suis (50)	0.5%	9.1%	3.7%	2.3	0.4%	3.7%	1.3%	0.7
S. equi (50)	0.5%	11.1%	4.4%	3.4	0.2%	2.4%	0.6%	0.5
S. agalactiae (50)	0.3%	8.7%	4.0%	2.1	0.2%	2.4%	1.0%	0.6
S. infantis (4)	1.5%	7.6%	4.2%	2.2	0.4%	1.3%	0.8%	0.3
S. uberis (13)	0.5%	6.7%	3.2%	2.0	0.9%	2.9%	1.7%	0.7
S. constellatus (10)	1.4%	6.7%	3.9%	2.1	0.4%	1.3%	0.9%	0.4
S. azizii (3)	3.7%	5.5%	4.3%	0.9	0.5%	0.5%	0.5%	0
S. pseudoporcinus (5)	1.9%	6.7%	3.8%	2.3	0.5%	1.0%	0.7%	0.2
S. mitis (49)	0.4%	9.3%	2.5%	2.1	0.3%	3.5%	1.6%	0.8
S. pasteurianus (5)	1.3%	5.2%	2.5%	1.6	0.8%	2.3%	1.4%	0.5
S. gallolyticus (17)	0.5%	6.2%	2.9%	1.4	0.4%	1.8%	0.9%	0.4
S. pneumoniae (482)	0.4%	8.4%	2.9%	2.1	0.2%	3.3%	0.9%	0.6
S. anginosus (24)	0.4%	7.6%	2.4%	2.1	0.4%	2.3%	1.2%	0.6
S. oralis (49)	0.5%	6.4%	2.1%	1.8	0.2%	2.8%	1.0%	0.6
S. iniae (8)	0.5%	8.6%	2.3%	3.0	0.2%	1.5%	0.6%	0.4
S. parasanguinis (31)	0.4%	5.4%	2.2%	1.6	0.3%	2.0%	0.7%	0.5
S. equinus (27)	0.4%	7.2%	1.9%	2.0	0.2%	2.7%	0.9%	0.7
S. cristatus (4)	0.8%	3.1%	1.7%	0.9	0.5%	1.5%	0.8%	0.4
S. intermedius (9)	0.5%	3.4%	1.1%	0.8	0.5%	1.8%	1.3%	0.5
S. gordonii (22)	0.4%	5.6%	1.9%	1.6	0.2%	1.3%	0.5%	0.3
S. vestibularis (6)	0.4%	0.8%	0.5%	0.1	0.5%	2.3%	1.6%	0.6
S. salivarius (32)	0.4%	3.0%	1.2%	1.0	0.2%	1.7%	0.5%	0.4
S. thermophilus (32)	0.3%	0.9%	0.4%	0.1	0.2%	1.4%	0.9%	0.4
S. sanguinis (37)	0.3%	3.4%	0.8%	0.7	0.1%	1.3%	0.4%	0.4
S. sobrinus (42)	0.4%	0.7%	0.5%	0.1	0.3%	0.6%	0.4%	0.1
S. mutans (50)	0.2%	1.5%	0.5%	0.2	0.1%	0.3%	0.2%	0.1

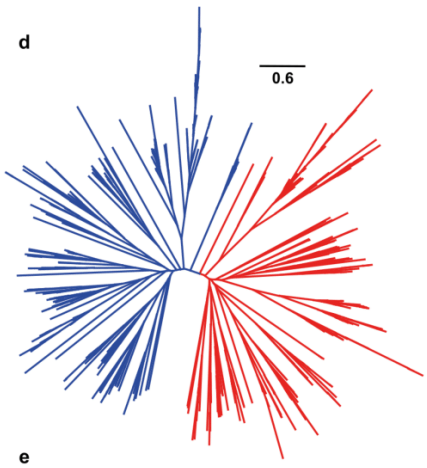
b



c



d



e

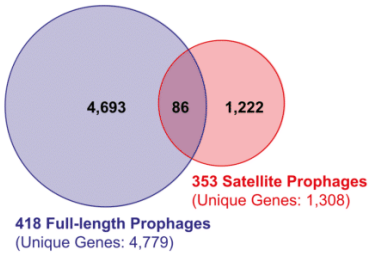


Fig. 1. Full-length and satellite prophages identified among streptococcal genomes. a, Average prophage content within each streptococcal species. SD, standard deviation. **b**, Graphical representation of all prophages by average genome size and number of genes. Each prophage is coloured to represent its average guanine (G) and cytosine (C) content. **c**, Satellite prophage SpnSP24 was represented among pneumococci isolated between 1939 and 2006 and all of these satellite prophages were nearly identical at the nucleotide level. **d**, An unrooted phylogenetic tree of all streptococcal prophage genomes identified in the dataset. Blue branches mark full-length prophages and red branches mark satellite prophages. **e**, Venn diagram depicting the number of genes found exclusively in full-length prophages or in satellite prophages (at a threshold of >70% amino acid sequence similarity) and those genes that are shared.

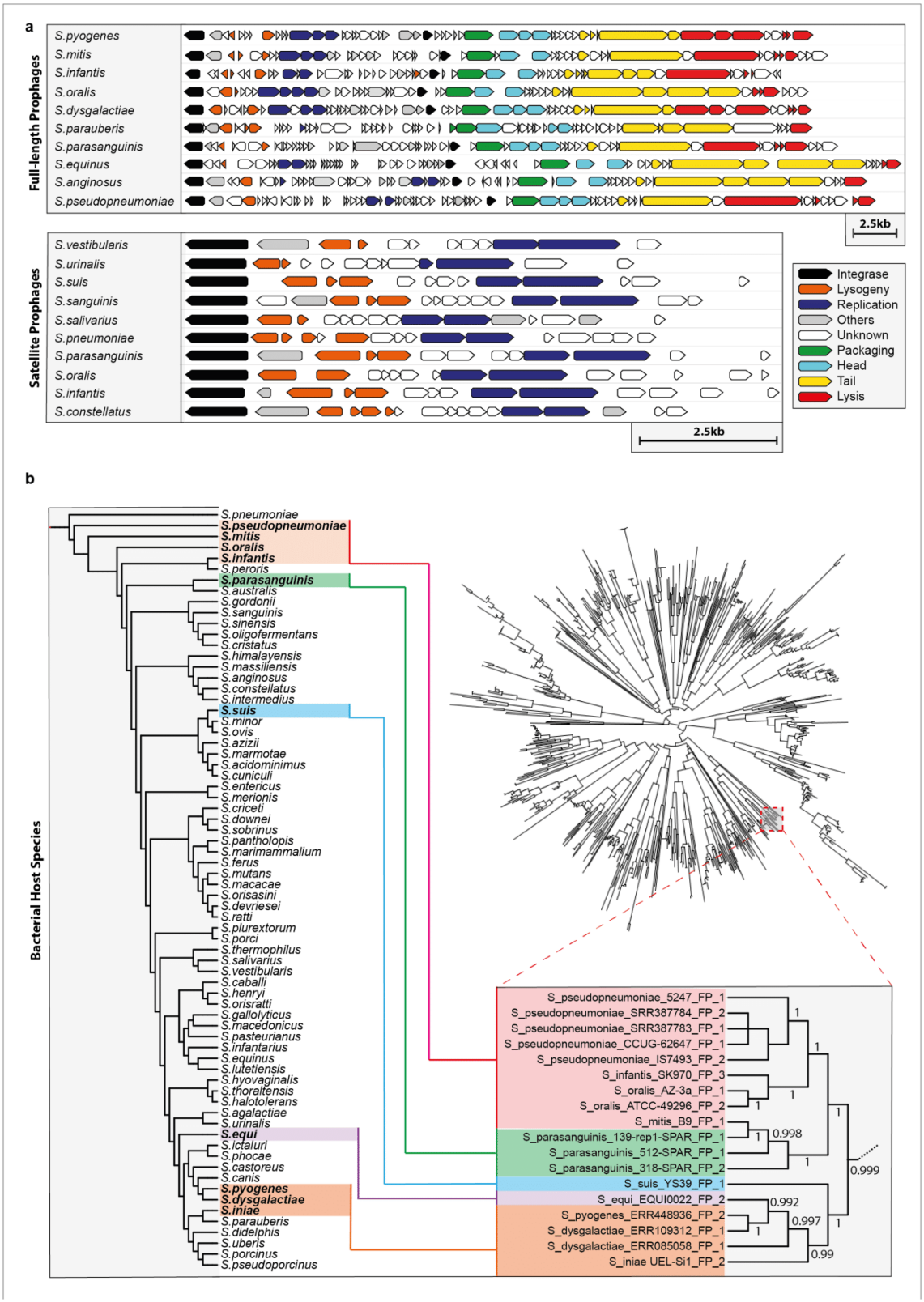


Fig. 2. Similarities among streptococcal prophages and evidence for cross-species transmission of prophages. a, Full-length and satellite prophages identified among different streptococcal species shared a

similar pattern in gene orientation and synteny. **b**, Phylogenetic tree depicting all prophages detected in this study (see Supplementary Figure 2 for a larger version of the tree) and a zoomed-in branch (with branch lengths ignored for illustrative purposes) depicting one example of a cluster of full-length prophages that were found among multiple streptococcal species (see also Supplementary Figure 3 for a distance matrix of pairwise
805 similarity among these 18 prophages).

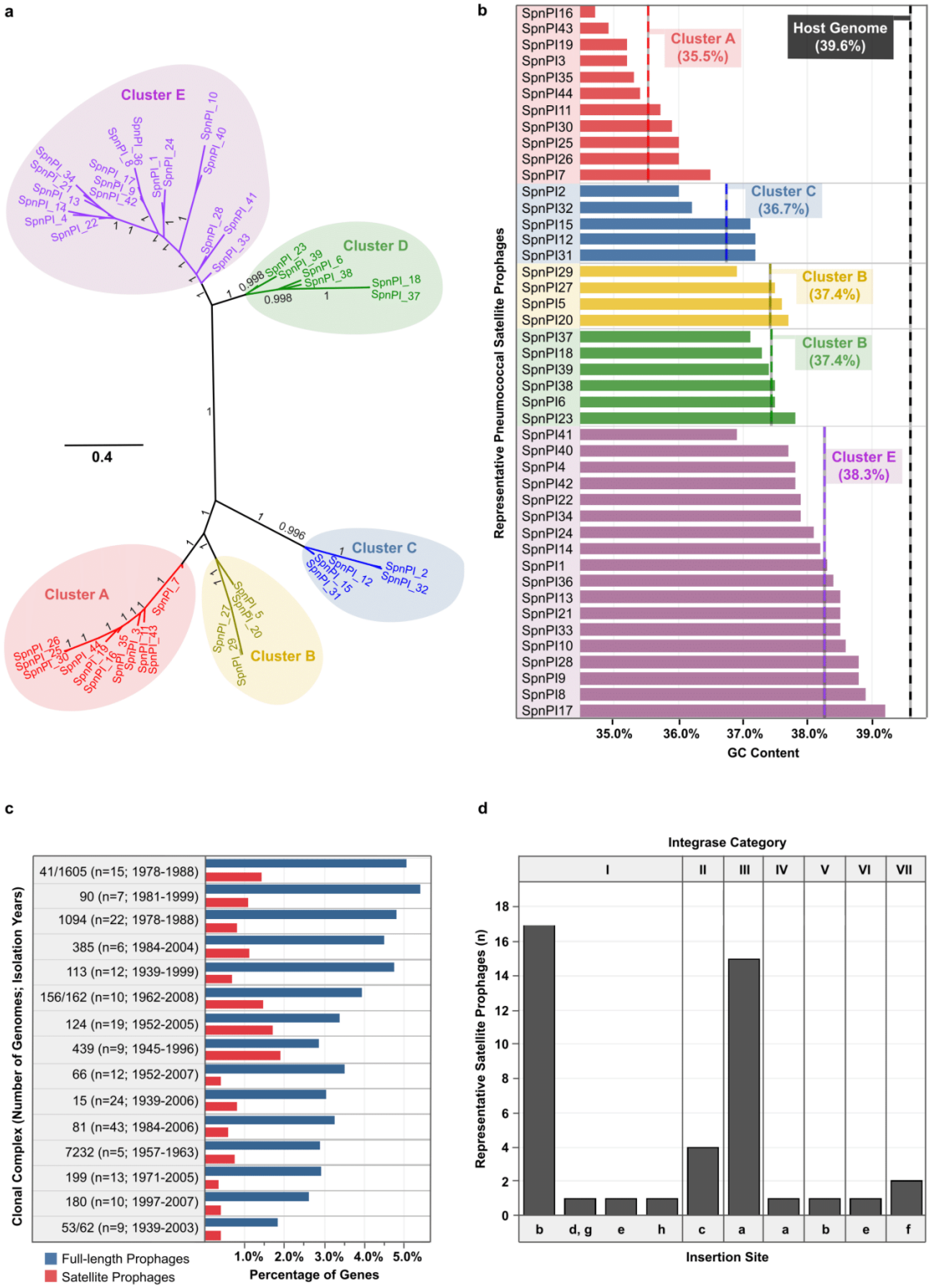
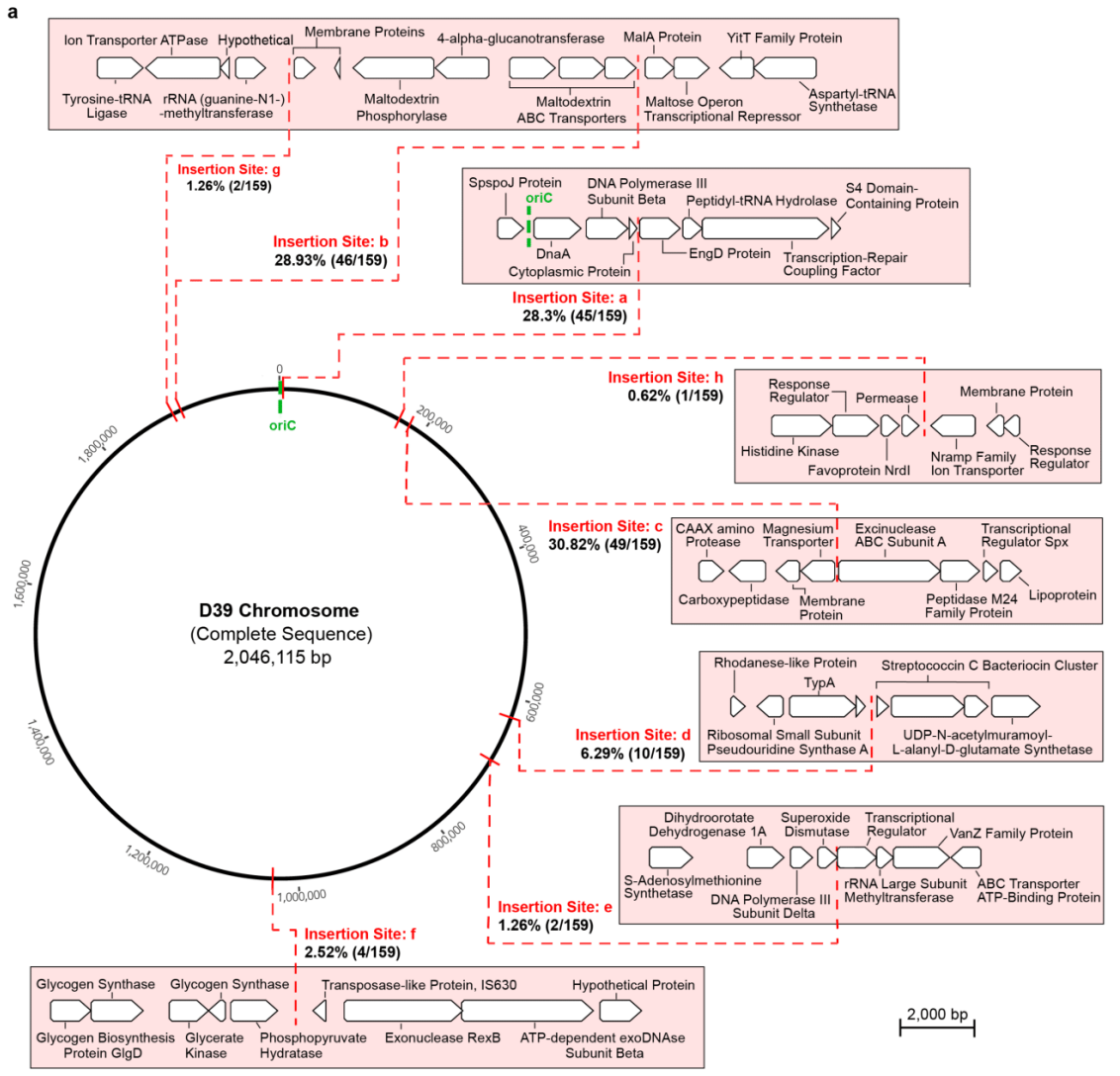

a
Cluster E
Cluster D
Cluster C
Cluster A
Cluster B
0.4
b
Host Genome
(39.6%)
Cluster A
(35.5%)
Cluster C
(36.7%)
Cluster B
(37.4%)
Cluster B
(37.4%)
Cluster E
(38.3%)
Representative Pneumococcal Satellite Prophages
35.0%
36.0%
37.0%
38.0%
39.0%
GC Content
c
Clonal Complex (Number of Genomes; Isolation Years)
41/1605 (n=15; 1978-1988)
90 (n=7; 1981-1999)
1094 (n=22; 1978-1988)
385 (n=6; 1984-2004)
113 (n=12; 1939-1999)
156/162 (n=10; 1962-2008)
124 (n=19; 1952-2005)
439 (n=9; 1945-1996)
66 (n=12; 1952-2007)
15 (n=24; 1939-2006)
81 (n=43; 1984-2006)
7232 (n=5; 1957-1963)
199 (n=13; 1971-2005)
180 (n=10; 1997-2007)
53/62 (n=9; 1939-2003)
Full-length Prophages
Satellite Prophages
1.0% 2.0% 3.0% 4.0% 5.0%
Percentage of Genes
d
Integrase Category
I
II
III
IV
V
VI
VII
Representative Satellite Prophages (n)
0
2
4
6
8
10
12
14
16
18
b
d, g
e
h
c
a
a
b
e
f
Insertion Site

Fig. 3. Satellite prophages found among a large collection of nearly 500 diverse pneumococcal genomes. a, An unrooted phylogenetic tree demonstrated that the 44 representative satellite prophages could be clustered into five major groups based upon nucleotide similarity. **b**, The average guanine/cytosine (GC) content (stated in brackets) of the satellite prophages varied by genetic cluster and was lower than the GC content of the pneumococcal host. **c**, The average prophage content for each of the major clonal complexes (genetic lineages) is depicted as a percentage of the total number of genes in the host pneumococcal genome (~2 Mb). **d**, The integrase sequences of the 44 representative satellite prophages were divided into seven different categories based upon ≥95% nucleotide similarity.


a
Ion Transporter ATPase
Hypothetical
Membrane Proteins
4-alpha-glucanotransferase
MalA Protein
YitT Family Protein
Tyrosine-tRNA Ligase
rRNA (guanine-N1-) -methyltransferase
Maltodextrin Phosphorylase
Maltodextrin ABC Transporters
Maltose Operon Transcriptional Repressor
Aspartyl-tRNA Synthetase
Insertion Site: g
1.26% (2/159)
Insertion Site: b
28.93% (46/159)
SpspoJ Protein
oriC
DNA Polymerase III Subunit Beta
Peptidyl-tRNA Hydrolase
S4 Domain-Containing Protein
DnaA
Cytoplasmic Protein
EngD Protein
Transcription-Repair Coupling Factor
Insertion Site: a
28.3% (45/159)
Insertion Site: h
0.62% (1/159)
Response Regulator
Permease
Membrane Protein
Histidine Kinase
Favoprotein NrdI
Nramp Family Ion Transporter
Response Regulator
Insertion Site: c
30.82% (49/159)
CAAX amino Protease
Magnesium Transporter
Excinuclease ABC Subunit A
Transcriptional Regulator Spx
Carboxypeptidase
Membrane Protein
Peptidase M24 Family Protein
Lipoprotein
D39 Chromosome
(Complete Sequence)
2,046,115 bp
oriC
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000
Insertion Site: d
6.29% (10/159)
Rhodanese-like Protein
TypA
Streptococcin C Bacteriocin Cluster
Ribosomal Small Subunit Pseudouridine Synthase A
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate Synthetase
Insertion Site: e
1.26% (2/159)
Dihydroorotate Dehydrogenase 1A
Superoxide Dismutase
Transcriptional Regulator
VanZ Family Protein
S-Adenosylmethionine Synthetase
DNA Polymerase III Subunit Delta
rRNA Large Subunit Methyltransferase
ABC Transporter ATP-Binding Protein
Insertion Site: f
2.52% (4/159)
Glycogen Synthase
Glycogen Synthase
Transposase-like Protein, IS630
Hypothetical Protein
Glycogen Biosynthesis Protein GlgD
Glycerate Kinase
Phosphopyruvate Hydratase
Exonuclease RexB
ATP-dependent exoDNAse Subunit Beta
2,000 bp
b
General Function
(Percentage; Number of Genes)
Information Storage and Processing (30.99%; 2,016)
Metabolism (27.50%; 1,789)
Cellular Processes and Signalling (15.28%; 994)
Unknown (26.24%; 1,707)
COG Categories
Replication, Recombination and Repair 818
Translation, Ribosomal Structure and Biogenesis 619
Transcription 579
Carbohydrate Transport and Metabolism 418
Inorganic Ion Transport and Metabolism 403
Amino Acid Transport and Metabolism 385
Post-Translational Modification, Protein Turnover, and Chaperones 226
Cell Wall/Membrane/Envelope Biogenesis 221
Defence Mechanisms 216
Nucleotide Transport and Metabolism 214
Intracellular Trafficking, Secretion, and Vesicular Transport 151
Lipid Transport and Metabolism 130
Energy Production and Conversion 115
Signal Transduction Mechanisms 112
Coenzyme Transport and Metabolism 97
Cell Cycle Control, Cell Division, Chromosome Partitioning 68
Secondary Metabolites Biosynthesis, Transport, and Catabolism 27
2.0%
4.0%
6.0%
8.0%
10.0%
12.0%
Percentage of Flanking Genes

820 **Fig. 4. Insertion sites of prophages within the streptococcal genomes. a**, Pneumococcal satellite prophages were integrated in seven locations (a-f) within the host genome. Percentages and numbers in brackets refer to the proportion and number out of all 159 satellite prophages that were inserted in that particular location. **b**, The flanking genes upstream and downstream of all integrated full-length and satellite prophages within the streptococcal genomes were retrieved for functional classification and are depicted here based upon their
825 COG (clusters of orthologous groups) classifications.

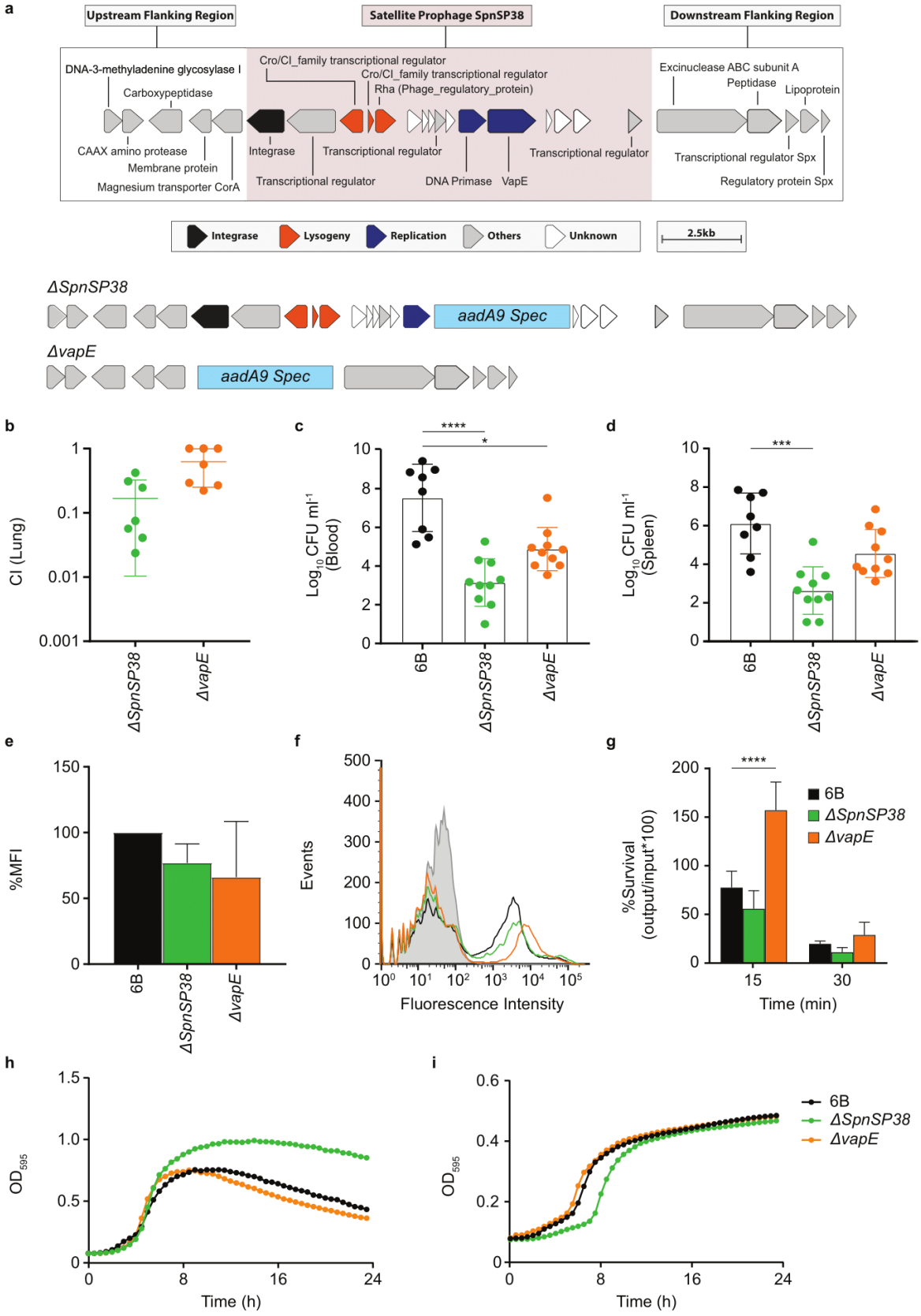
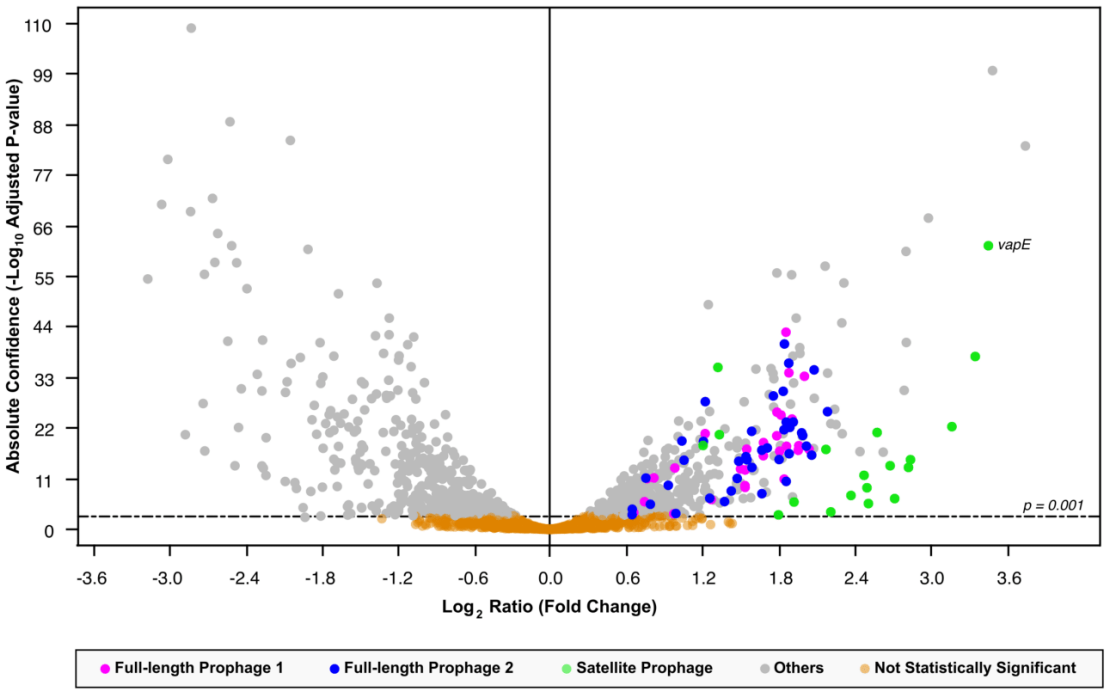

a
Upstream Flanking Region
Satellite Prophage SpnSP38
Downstream Flanking Region
DNA-3-methyladenine glycosylase I
Carboxypeptidase
CAAX amino protease
Membrane protein
Magnesium transporter CorA
Cro/CI_family transcriptional regulator
Cro/CI_family transcriptional regulator
Rha (Phage_regulatory_protein)
Integrase
Transcriptional regulator
Transcriptional regulator
DNA Primase
VapE
Transcriptional regulator
Excinuclease ABC subunit A
Peptidase
Lipoprotein
Transcriptional regulator Spx
Regulatory protein Spx
Integrase
Lysogeny
Replication
Others
Unknown
2.5kb
ΔSpnSP38
aadA9 Spec
ΔvapE
aadA9 Spec
b
CI (Lung)
1
0.1
0.01
0.001
ΔSpnSP38
ΔvapE
c
Log10 CFU ml-1 (Blood)
****
*
10
8
6
4
2
0
6B
ΔSpnSP38
ΔvapE
d
Log10 CFU ml-1 (Spleen)
***
10
8
6
4
2
0
6B
ΔSpnSP38
ΔvapE
e
%MFI
150
100
50
0
6B
ΔSpnSP38
ΔvapE
f
Events
500
400
300
200
100
0
10^0 10^1 10^2 10^3 10^4 10^5
Fluorescence Intensity
g
%Survival (output/input*100)
****
200
150
100
50
0
15
30
Time (min)
6B
ΔSpnSP38
ΔvapE
h
OD595
1.5
1.0
0.5
0
0
8
16
24
Time (h)
i
OD595
0.6
0.4
0.2
0
0
8
16
24
Time (h)
6B
ΔSpnSP38
ΔvapE

Fig. 5. Assessment of the virulence of *ΔSpnSP38* (deletion of entire satellite prophage) and *ΔvapE* (deletion of *vapE* only) mutant pneumococcal strains in murine infection. **a**, Upper part depicts the satellite prophage genes integrated within the BHN418 genome and flanking pneumococcal genes, and the lower part depicts the *ΔSpnSP38* and *ΔvapE* mutants with the addition of the spectinomycin resistance cassette aadA9. **b**, Plots of the competitive index (CI) for the *ΔSpnSP38* and *ΔvapE* mutant strains versus the wild-type strain in a mouse model of pneumonia. Each symbol represents the CI for a single animal and bars represent the median and range. **c** and **d**, Mean bacterial colony-forming units (CFU) recovered at 24h from blood (**c**) or spleen (**d**) homogenates after intraperitoneal inoculation of 5×10^6 CFU/strain. Each symbol represents data for a single animal and error bars represent standard deviation (two-sided, Kruskall-Wallis with Dunn's post hoc test to identify significant differences between groups, *, p<0.05; ***, p<0.001; ****, p<0.0001). **e**, Median standard deviation (SD) mean fluorescence intensity (MFI) of C3b deposition on the surface of the wild-type and mutant strains as measured by a flow cytometry assay. **f**, Example of a flow cytometry histogram for the C3b deposition data. **g**, Bacterial survival in a neutrophil killing assay (multiplicity of infection: 1 bacterium/100 neutrophils) represented as % CFU/ml recovered after 15 to 30 min incubation compared to the input bacteria. Error bars represent standard deviation and asterisks represent statistical significance compared to the wild-type strain (Kruskal-Wallis test with Dunn's correction for multiple comparisons, ****, p<0.0001). **h** and **i**, Growth curves as measured by the optical density (OD) of wild-type and mutant strains cultured in Todd-Hewitt broth supplemented with 0.5% yeast-extract (**h**) or 100% human serum (**i**).



850

Fig 6. Satellite prophage genes were overexpressed when pneumococci were grown planktonically versus in a biofilm. The data were generated from pneumococcal reference strain Sp6A-10, which contains two full-length prophages (Spn_6A-10_FP1 and Spn_6A-10_FP2) and one satellite prophage (SpnSP33). Genes belonging to SpnSP33 are shown in green, while those belonging to Spn_6A-10_FP1 and Spn_6A-10_FP2 are
855 shown in blue and magenta, respectively. A higher $\log_2$ ratio denotes increased expression levels in planktonic growth as compared to growth in a biofilm. A full list of the genes depicted here, their expression levels and sequences may be found in Supplementary Table 6. The annotated genes and relative expression levels of all three prophages are found in Supplementary Figure 6.

860

Supplementary Figure Legends

Supplementary Figure 1. Genetic diagrams of full-length and satellite prophages in different bacterial species. The diagrams illustrate the similarity in gene composition and synteny among different prophages identified in a variety of unrelated bacterial species.

Supplementary Figure 2. Full-length and satellite prophages identified among streptococcal genomes. An unrooted phylogenetic tree of all streptococcal prophage genomes identified in the dataset. Blue branches mark full-length prophages and red branches mark satellite prophages. Note that this is the annotated version of Figure 1d in the main paper.

Supplementary Figure 3. A distance matrix of pairwise similarity among the 18 prophages highlighted in Figure 2b. Eighteen full-length prophage sequences were compared in a pairwise fashion and the percentage similarity is given for each pair of prophages. Cells are shaded in grey to indicate the level of similarity, increasing to black for pairs of sequences that are 100% identical.

Supplementary Figure 4. A pneumococcal core genome phylogenetic tree. The tree is annotated with the corresponding satellite and full-length prophage clusters.

Supplementary Figure 5. The location of prophage insertion sites within the bacterial genomes. One finished genome of each of 29 streptococcal species was divided into eight non-overlapping segments of equal length according to the number of base pairs, and the percentages of prophages situated in each segment were quantified.

Supplementary Figure 6. Heat maps describing the differential expression levels of three prophages during planktonic growth vs growth in a biofilm. Prophage genes are annotated and the differential expression level of each gene during planktonic bacterial growth relative to growth in a biofilm is marked by shades of red (down-regulated genes) or green (up-regulated genes). An asterisk to the right of a cell indicates a statistically-significant differential level of expression ($p < 0.05$).

The Journal of Infectious Diseases

MAJOR ARTICLE

Diverse *Streptococcus pneumoniae* Strains Drive a Mucosal-Associated Invariant T-Cell Response Through Major Histocompatibility Complex class I–Related Molecule–Dependent and Cytokine-Driven Pathways

Ayako Kurioka,[1,a] Bonnie van Wilgenburg,[1,a] Reza Rezaei Javan,[1,a] Ryan Hoyle,[1,a] Andries J. van Tonder,[1] Caroline L. Harrold,[1] Tianqi Leng,[1] Lauren J. Howson,[2] Dawn Shepherd,[2] Vincenzo Cerundolo,[2] Angela B. Brueggemann,[1,3] and Paul Klenerman[1]

[1]Nuffield Department of Medicine and [2]MRC Human Immunology Unit, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford, and [3]Department of Medicine, Imperial College London, London, United Kingdom

Mucosal-associated invariant T (MAIT) cells represent an innate T-cell population that can recognize ligands generated by the microbial riboflavin synthesis pathway, presented via the major histocompatibility complex class I–related molecule (MR1). *Streptococcus pneumoniae* is a major human pathogen that is also associated with commensal carriage; thus, host control at the mucosal interface is critical. The recognition of pneumococci by MAIT cells has not been defined nor have the genomics and transcriptomics of the riboflavin operon. We observed robust recognition of pneumococci by MAIT cells, using both MR1-dependent and MR1-independent pathways. The pathway used was dependent on the antigen-presenting cell. The riboflavin operon was highly conserved across a range of 571 pneumococci from 39 countries, dating back to 1916, and different versions of the riboflavin operon were also identified in related *Streptococcus* species. These data indicate an important functional relationship between MAIT cells and pneumococci.

Keywords. MAIT cells; pneumococcus; riboflavin; innate; macrophages; cytokines; MR1; T cells.

The pneumococcus is the most common cause of community-acquired pneumonia and is associated with significant morbidity and mortality, especially among young children and older adults [1, 2]. Pneumococci also cause invasive diseases, such as meningitis and bacteremia, and upper respiratory tract infections, such as otitis media and sinusitis [3]. Antimicrobial-resistant strains are widespread and pose problems in the treatment of infections, leading the World Health Organization to include pneumococci on their list of priority pathogens [4]. The available pneumococcal conjugate vaccines prompt immune responses to polysaccharide capsules (differentiated as serotypes) and are highly effective at preventing invasive pneumococcal disease due to vaccine-serotype strains; however, current vaccines only protect against a small number of the possible serotypes, leading to increases in rates of disease from nonvaccine-serotype pneumococci [5, 6]. Therefore, pneumococcal disease remains a serious problem, and better understanding of the host defense against pneumococci may facilitate design of novel interventions.

There is increasing appreciation of the role of unconventional T cells in orchestrating early responses to pathogens [7]. Mucosal-associated invariant T (MAIT) cells are a recently described innate T-cell population, abundant in the lung, blood, and liver [8–10]. They express a semi-invariant T-cell receptor (TCR) chain, Vα7.2-Jα33/Jα12/Jα20, paired with a limited repertoire of Vβ chains [11]. This TCR can recognize ligands presented by the conserved major histocompatibility complex (MHC)–related protein 1 (MR1) [8]. MR1 binds vitamin B–based precursors from the riboflavin-biosynthesis pathway, conserved across various bacteria and fungi [10, 12, 13]. Human MAIT cells can also respond to innate cytokines even without TCR signaling [14, 15]. Upon activation, they produce immunomodulatory cytokines, including interferon γ (IFN-γ), tumor necrosis factor α, and interleukin 17.

MAIT cells are critical for the control of bacterial infections in mice, particularly in the lungs [16–18]. For instance, aerosol-based infection models with *Mycobacterium bovis* bacillus Calmette-Guérin and the live vaccine strain of *Francisella tularensis* demonstrated that MAIT cells were essential for the early control of the bacterial burdens [18, 19]. Indeed, early lung MAIT cell activation by *F. tularensis* was required for the differentiation of dendritic cells and subsequent recruitment of activated $CD4^+$ T cells [20]. Thus, rapid activation of MAIT cells in response to pulmonary bacteria is critical for bridging innate and adaptive systems.

Despite these data, it remains unclear whether MAIT cells play a role in the defense against pneumococcal infection. Here, we show that MAIT cells responded to pneumococci in

Received 12 September 2017; editorial decision 7 December 2017; accepted 14 December 2017; published online December 15, 2017.

[a]A. K., B. v. W., R. R. J., and R. H. contributed equally to this work.

Correspondence: P. Klenerman, F Med Sci, The Peter Medawar Building for Pathogen Research, South Parks Road, University of Oxford, OX1 3SY (paul.klenerman@ndm.ox.ac.uk).

The Journal of Infectious Diseases® 2018;217:988–99

DOI: 10.1093/infdis/jix647

an MR1-dependent manner in the presence of macrophages but not monocytes and that this was dependent on costimulation provided by innate cytokines. Furthermore, using a population-level genomics approach, we found that the riboflavin synthesis pathway is ubiquitous and highly conserved amongst pneumococci. Riboflavin operon genes were also found among other nonpneumococcal *Streptococcus* species, including *Streptococcus agalactiae* (group B streptococci), which suggests that the observations made here are relevant to other human-associated *Streptococcus* species infections.

METHODS

Cells

Whole-blood specimens were obtained from leukocyte cones (NHS Blood and Transplant), and peripheral blood mononuclear cells (PBMCs) were isolated by density gradient centrifugation (Lymphoprep Axis-Shield). All samples were collected with written consent and local research ethics committee approval (COREC 04.OXA.010). Monocyte-derived macrophages were generated by enriching for monocytes using CD14 microbeads (Miltenyi Biotech) before culturing with 50 ng/mL granulocyte-macrophage colony-stimulating factor (Miltenyi Biotech) in Roswell Park Memorial Institute 1640 medium, penicillin/streptomycin, L-glutamine, and 10% human serum (all from Sigma Aldrich) for 6–8 days. For details of the Jurkat-MAIT cell line, see the Supplementary Methods.

Bacteria

Pneumococcal Molecular Epidemiological Network (PMEN) strains (Supplementary Methods) were cultured from freezer stocks to Columbia blood agar plates (Oxoid), incubated overnight, and then transferred to Todd Hewitt broth (THB; Sigma Aldrich) with 0.5% yeast extract (THB-Y; Sigma Aldrich) and incubated overnight, unless indicated otherwise. Where indicated, bacteria were grown in riboflavin-free medium (ie, riboflavin assay medium [BD Difco] or THB alone) [21]. *Escherichia coli* (DH5a; Invitrogen) was cultured in LB medium overnight in a shaking incubator.

Pneumococci or *E. coli* were fixed in 2% paraformaldehyde for 15 minutes and washed extensively (except in a single set of experiments in which live bacteria were used for comparison). A negative control was prepared identically.

In Vitro Stimulation of MAIT Cells

THP1 cells (ATCC, Middlesex, United Kingdom) were incubated overnight with paraformaldehyde-fixed pneumococci or *E. coli* at a ratio of 30 bacteria/cell or with sterile control. For stimulation experiments, in which activation of MAIT cells was examined (eg, IFN-γ production), THP1 cells were washed, and PBMCs or enriched $CD8^+$ T cells were added to THP1 cells overnight. Brefeldin A (eBioscience) was added for the final 4 hours of the stimulation before intracellular cytokine staining. For internal staining, cells were fixed with 1% formaldehyde (Sigma Aldrich) and permeabilized with permeabilization buffer (eBioscience). Alternatively, for the assessment of degranulation, anti-CD107a PE-Cy7 (BioLegend) was added from the start of the stimulation. For blocking experiments, anti-MR1, anti–interleukin 12p40/70 (IL-12p40/70), and anti–interleukin 18 (IL-18) antibodies (all BioLegend) or the appropriate isotype controls were added for the duration of the experiment. Cells were acquired on the MACSQuant Analyser (Miltenyi Biotech) and analyzed using FlowJo v9.8 (TreeStar). Graphs and statistical analyses were completed using GraphPad Prism 6. All data are presented as mean values with standard errors of the mean (SEMs). For further details and antibodies used, see the Supplementary Methods.

RNA Sequencing

Pneumococcal strain 2/2 was cultured in brain-heart infusion broth and incubated at 40°C for 6 hours to mimic heat shock. Identical experimental controls were incubated at 37°C. Broth cultures at 2, 3, 4, 5, and 6 hours were removed from the incubator, and RNAprotect Bacteria Reagent (Qiagen) was added to stabilize the RNA. RNA was extracted from the samples, using the Promega Maxwell 16 Instrument and LEV simplyRNA Cells purification kit, following the manufacturer's protocol. Extracted RNA samples were sent to the Oxford Genomics Centre for processing. Library preps were made using RNA-Seq Ribozero kits (Illumina), and sequencing was performed on the MiSeq (Illumina). The Gene Expression Omnibus accession number is pending.

The sequenced forward and reverse reads were paired and mapped to pneumococcal strain 2/2 genome, using Bowtie2 with the highest-sensitivity option [22]. Differential gene expression was analyzed in Geneious, version 9.1 (Biomatters), using the DESeq method [23]. Genes with an adjusted *P* value of <.05 were considered to be differentially expressed.

Compilation of the Genome Data Sets

Two large genome data sets were compiled for this study, and data were stored in a BIGSdb database [24]. The pneumococcal data set consisted of 571 historical and modern genomes isolated during 1916–2009 from people of all ages residing in 39 different countries. The pneumococci were recovered from individuals with carriage and those with disease, and 89 serotypes and 296 multilocus sequence types were represented in this data set (Supplementary Table 1). A total of 486 pneumococcal genome sequences were compiled from previously published studies or were downloaded from GenBank [25]. The remaining 85 pneumococcal genomes were recently sequenced. Pneumococcal cultures were prepared as described above, before DNA was extracted using the Promega Maxwell 16 Instrument and Buccal Swab LEV DNA purification kits according to the manufacturers' protocols. DNA extracts were

sent to the Oxford Genomics Centre, where libraries were made and DNA was sequenced on the Ilumina platform. Velvet was used to make de novo genome assemblies, which were further improved using SSPACE and GapFiller [26–28].

The nonpneumococcal *Streptococcus* species data set contained 834 genomes of 69 different streptococcal species (Supplementary Table 2). Thirty-four genomes were newly sequenced as described above, and the rest were downloaded from the ribosomal multilocus sequence typing database [29]. Further details are provided in the Supplementary Methods.

RESULTS

Pneumococci Possess a Highly Conserved Riboflavin Synthesis Operon That Is Upregulated With Heat Stress

The genes encoding the riboflavin biosynthetic enzymes of pneumococci (*ribD*, *ribE*, *ribA*, and *ribH*) were found to be clustered together in the same orientation in a predicted 3.4-Kb operon structure (Figure 1A). The prevalence and sequence diversity of the coding regions of the riboflavin genes were investigated in a large, global, and historical genome data set of pneumococci isolated between 1916 and 2008 from people of all ages residing in 36 different countries. A total of 561 pneumococcal genomes (98.2%) contained the riboflavin operon. Nine of 10 genomes that lacked the operon were of a single multilocus sequence type ($ST^{serotype}$), $ST13^{14/nontypable}$, and the other belonged to $ST695^{4}$ (Supplementary Table 1). All genes in the riboflavin operon were found to be highly conserved: nucleotide and amino acid sequence identity were >99% (Table 1). The dN/dS analysis revealed a higher prevalence of synonymous versus nonsynonymous mutations, supporting the importance of maintaining the riboflavin operon (Table 1).

Total bacterial RNA sequencing was performed on RNA extracted from a pneumococcus that was subjected to metabolic stress by incubation at a higher temperature than normal (40°C vs 37°C). Differential expression analysis revealed that all of the riboflavin operon genes were significantly upregulated after 2–4 hours of incubation under heat stress as compared to the control (Figure 1B). Subsequently, the riboflavin operon was found to be significantly downregulated after 5–6 hours of incubation. The concurrent increase and decrease in the expression of the 4 riboflavin genes suggested that these genes are transcriptionally coupled.

MAIT Cells Are Activated by Pneumococci

Human MAIT cell responses to bacteria can be readily analyzed in vitro, using fluorescence-activated cell-sorting analysis of PBMCs. Following incubation with bacterially loaded antigen-presenting cells, activation of MAIT cells and control cells can be tracked in parallel by analysis of surface markers of activation (eg, CD69) and functional responses (ie, IFN-γ release and degranulation). To determine whether MAIT cells were able to respond to pneumococci, 10 PMEN reference strains were used to probe the activation of MAIT cells in the presence of the cell line THP1. PBMCs were cultured with paraformaldehyde-fixed pneumococci and THP1 cells overnight, and interferon production by MAIT cells was examined using intracellular cytokine staining and flow cytometry (Figure 2A, B). There was clear production of IFN-γ by MAIT cells across all strains, although there was variability in the responses: production by 7 of 10 strains reached statistical significance. Similarly, CD69 expression was induced by all 10 strains, as measured by geometric mean fluorescence intensity, and reached significance

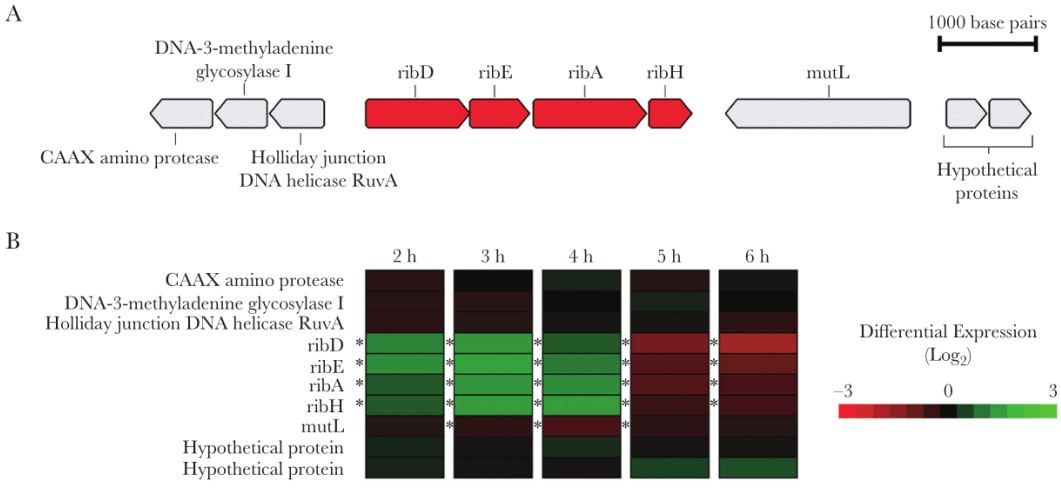


Figure 1. Genetic and transcriptomic data related to the riboflavin operon in pneumococci. *A*, The riboflavin operon is depicted with riboflavin genes *ribD, ribE, ribA,* and *ribH* (red) and flanking genes (gray). *B*, RNA expression data at 5 time points (2–6 hours after initial incubation) are illustrated for each riboflavin gene and the flanking genes. Genes marked with differential expression levels in green were upregulated, and those in red were downregulated during incubation at 40°C as compared to normal incubation at 37°C. *$P < .05$.

Table 1. Description of the 4 Riboflavin Operon Gene Sequences Within 571 Pneumococcal Genomes

Gene	Present, No. (%)[a]	Nucleotide Pair-Wise Identity, %	Amino Acid Pair-Wise Identity, %	Mean dN/dS	Gene Annotation
ribD	559 (98.2)	99.6	99.8	0.36	Diaminohydroxyphosphoribosylamino-pyrimidine deaminase
ribE	561 (98.2)	99.6	99.9	0.28	Riboflavin synthase
ribA	561 (98.2)	99.7	99.9	0.42	3,4-dihydroxy-2-butanone-4-phosphate synthase
ribH	559 (98.2)	99.5	99.1	0.12	6,7-dimethyl-8-ribityllumazine synthase

[a]Two genomes possessed sequence assembly gaps in *ribD* and *ribH*, and 10 genomes were missing all 4 riboflavin genes (see Results).

for 7 strains. In comparison, there was negligible activation, as measured by IFN-γ or CD69 expression, of non-MAIT cells (ie, $CD161^{-}CD8^{+}$ T cells, which act as a negative control because they do not respond to the bacterial ligand and/or accompanying cytokine signals; Figure 2B), suggesting that pneumococci specifically activated MAIT cells.

MAIT Cell Activation by Pneumococci in the Presence of Monocytes Is Not MR1 Dependent

We previously found that the response of MAIT cells to *E. coli* is codependent on both MR1 and the innate cytokines IL-12 and IL-18 [14]. In these experiments, blockade of either MR1 or cytokines alone only yielded partial inhibition, while combined blockade abrogated responsiveness. To investigate the mechanism of the MAIT cell response to pneumococci, we cultured paraformaldehyde-fixed pneumococci with PBMCs and THP1 cells in the presence of anti-MR1, anti–IL-12, and anti–IL-18 blocking antibodies (Figure 3A). As expected, IFN-γ expression in response to *E. coli* was blocked significantly by anti-IL-12 and anti-IL-18 blocking antibodies, with full blocking only seen with the addition of an anti-MR1 blocking antibody (in these experiments, MR1 blockade alone had a limited effect). We found in parallel that blockade of MR1 alone had no effect on pneumococcal MAIT cell activation. Instead, in contrast to *E. coli*, blocking IL-12 and IL-18 completely abrogated MAIT cell activation across all strains tested. This suggested that, although pneumococci possess the riboflavin synthesis pathway, activation of MAIT cells by pneumococci in the presence of THP1 cells was cytokine dependent.

To confirm these findings, Jurkat cells engineered to express the MAIT cell TCR Vα7.2-Jα33 (Jurkat-MAIT cells) were cultured with fixed pneumococcal strains overnight in the presence of THP1 cells (Figure 3B). There was no significant change in the expression of CD69 by Jurkat-MAIT cells in the presence of any of the pneumococcal strains (CD69 was chosen as a robust marker for activation since the cells do not produce IFN-γ). Thus, in the presence of primary monocytes and THP1 cells, there was very little activation of MAIT cells by pneumococci through the MR1 pathway.

Given that we observed the upregulation of the riboflavin synthesis pathway in pneumococci upon heat stress (Figure 1B), we tested whether changing environmental factors such as temperature and modulating the availability of riboflavin would increase riboflavin synthesis, increase the availability of the MAIT cell ligand, and trigger activation of MAIT cells through the MR1 pathway. Pneumococcal strain PMEN34 was grown for 16 hours in THB-Y at 36°C and then transferred either to a riboflavin-containing medium and incubated at 40°C or to riboflavin-free medium and incubated at 36°C or 40°C for 4 hours, before the bacteria were fixed (Figure 3C). Although there was a slight increase in the fraction of MAIT cells expressing IFN-γ when bacteria were cultured in riboflavin-free assay medium regardless of temperature, this increase was not dependent on MR1.

We also tested whether using the live strain PMEN34 or the supernatant of pneumococcal growth culture, instead of fixed bacteria, would stimulate MAIT cells through the MR1 pathway (Supplementary Figure 1). These responses were small and could not be significantly blocked by an anti-MR1 blocking antibody; responses were similarly small when using enriched $CD8^{+}$ T cells. Thus, in the presence of monocytes or THP1 cells, MAIT cells are activated mainly through innate cytokines rather than through MR1, regardless of temperature or riboflavin availability.

MR1-Dependent Activation of MAIT Cells by Pneumococci in the Presence of Macrophages

We next tested whether monocyte-derived macrophages can present the MR1 ligand to activate MAIT cells more effectively through MR1, because alveolar macrophages play an important role in the immune response to pneumococci [30]. Furthermore, we investigated whether temperature or the abundance of riboflavin in the medium influenced the availability of the ligand (through riboswitch-mediated modulation of the operon [31]) and therefore affected MR1-dependent activation. For this evaluation, strain PMEN34 was grown for 16 hours in THB-Y or THB at 36°C or 40°C (Figure 4A–B).

We found that when using monocyte-derived macrophages, pneumococci induced IFN-γ expression from MAIT cells that was significantly reduced by MR1 blockade. This MR1-dependent activation was seen regardless of the temperature and medium in which the pneumococci were grown. Interestingly, there was a clear increase in activation induced by bacteria grown in the basic medium, THB, as compared to bacteria grown in THB-Y (which contains additional riboflavin). This is consistent with an increase in riboflavin production in the absence of riboflavin or with induction of the operon through heat stress (or both) and, thus, is consistent with increased ligand availability.

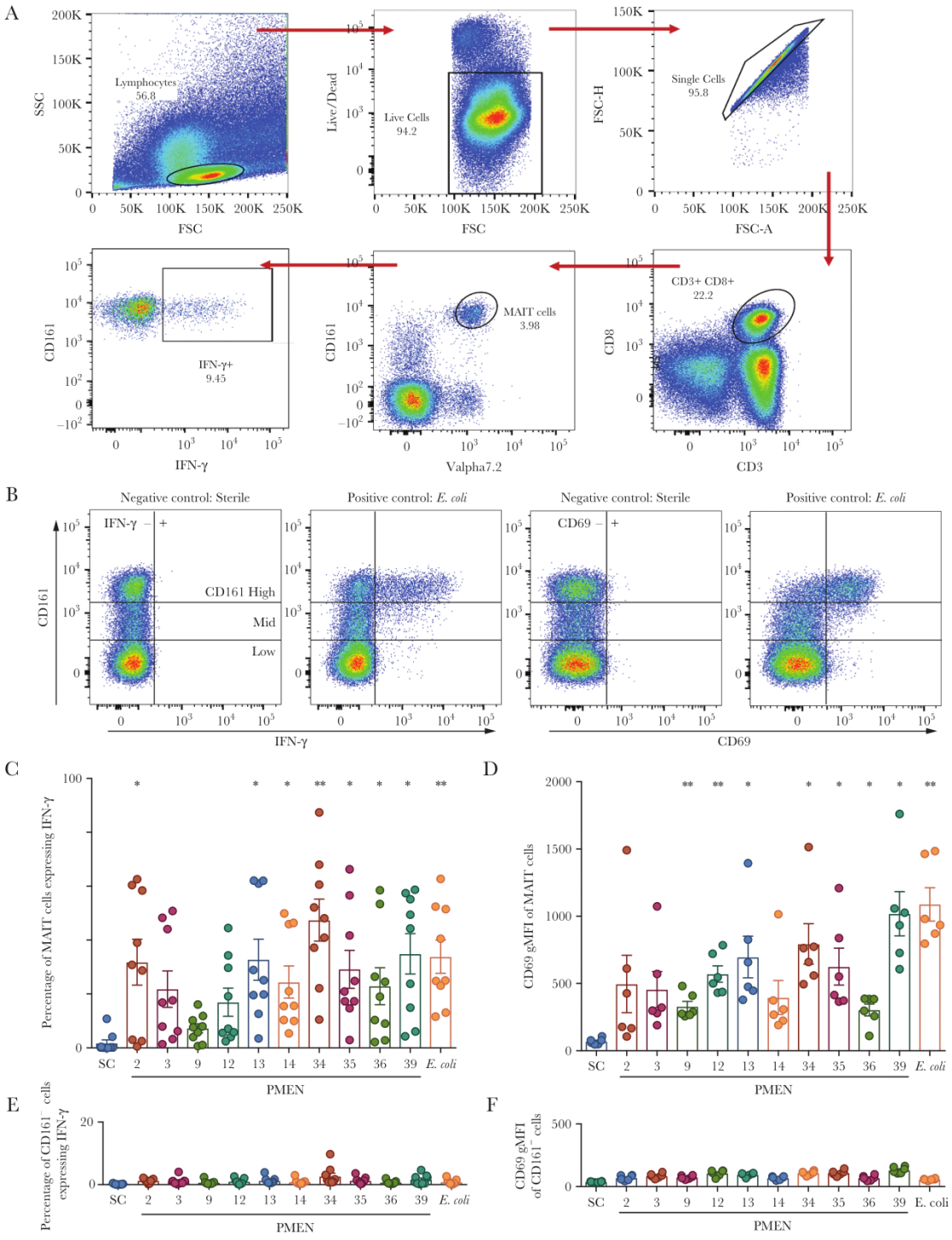


Figure 2. Mucosal-associated invariant T (MAIT) cells are activated by pneumococci. *A*, Gating strategy for analysis of MAIT cell–derived interferon γ (IFN-γ) following activation of peripheral blood mononuclear cells. *B*, Example fluorescence-activated cell-sorting plots showing upregulation of CD69 and IFN-γ in unstimulated and stimulated MAIT cells (gating on $CD3^+CD8^+$ live T cells). *C* and *D*, Ten Pneumococcal Molecular Epidemiological Network (PMEN) reference strains were used to probe the activation of MAIT cells following coculture of peripheral blood mononuclear cells in the presence of the monocytic cell line, THP1. *Escherichia coli* was added as a positive control. Frequency of cells expressing IFN-γ among MAIT cells (*C*) or $CD161^-CD8^+$ T cells (*D*) are shown (n = 9). *E* and *F*, CD69 expression measured by geometric mean fluorescence intensity (gMFI) in MAIT cells (*E*) or $CD161^-CD8^+$ T cells (*F*) are shown (n = 6). $^{**}P < .01$ and $^{*}P < .05$ by repeated measures 1-way analysis of variance with the Dunnett multiple comparisons test, compared with the sterile control (SC). Numbers indicate the PMEN reference strains. FSC, forward scatter; SSC, side scatter.

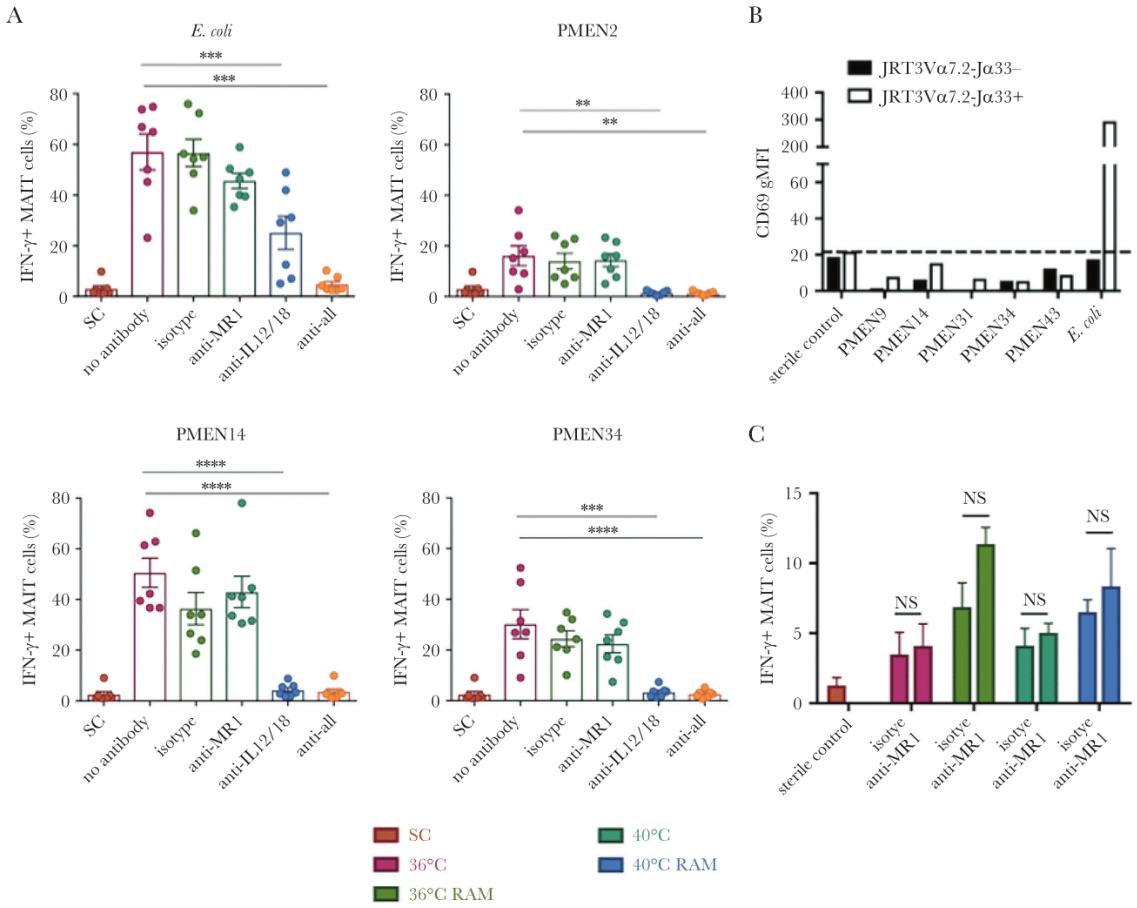


Figure 3. Mucosal-associated invariant T (MAIT) cell activation by pneumococci in presence of monocytes is not MR1-dependent. *A*, Paraformaldehyde-fixed Pneumococcal Molecular Epidemiological Network (PMEN) reference strains or *Escherichia coli* were cultured with peripheral blood mononuclear cells and THP1 cells in the presence of anti-MR1 blocking, anti–interleukin 12 (IL-12), and anti–interleukin 18 (IL-18) blocking antibodies. Interferon γ (IFN-γ) expression from MAIT cells is shown. $^{****}P < .0001$, $^{***}P < .001$, and $^{**}P < .01$, by repeated measures 1-way analysis of variance (ANOVA) with the Dunnett multiple comparisons test, compared with the no antibody control (n = 7). *B*, Jurkat cells expressing the MAIT cell T-cell receptor (TCR; white bars) or control cells not expressing the MAIT cell TCR (black bars) were cultured with THP1 cells overnight with the indicated PMEN strains or *E. coli* as a positive control. Activation was measured as the geometric mean fluorescence intensity (gMFI) of CD69 expressed by Jurkat cells. The dotted line indicates value for the sterile control (SC). Data are representative of 3 independent experiments. *C*, The PMEN34 strain was grown overnight at 36°C or 40°C in Todd Hewitt broth with 0.5% yeast extract (THB-Y) and either cultured in THB-Y for the last 4 hours or transferred to riboflavin-free medium (RAM). The bacteria were then fixed and cultured with PBMCs and THP1 cells overnight. The frequency of IFN-γ–expressing MAIT cells is shown in the presence or absence of anti-MR1 blocking antibody. NS, nonsignificant by 2-way ANOVA with the Sidak multiple comparisons test (n = 3).

To confirm these results, we also measured degranulation by investigating upregulation of CD107a, which is a further specific marker associated with MAIT activation. Degranulation was also induced by pneumococci grown in THB and was blocked by the anti–MR1-blocking antibody to a varying degree (Figure 4C–D).

Next, bacteria were cultured as above and added to PBMCs and macrophages overnight in the presence of anti-MR1, anti–IL-12, and anti–IL-18 antibodies (Figure 4E). There was a significant effect of blocking MR1 on IFN-γ production from MAIT cells in the presence of pneumococci cultured in THB, regardless of temperature, and full blockade in the presence of anti–IL-12 and anti–IL-18 blocking antibodies.

Finally, to confirm these results, we cultured Jurkat-MAIT cells with fixed pneumococci grown in THB or THB-Y at different temperatures in the presence of monocyte-derived macrophages (Figure 4F). Pneumococci grown in THB significantly increased the expression of CD69 in Jurkat-MAIT cells. Thus, in the presence of monocyte-derived macrophages, pneumococci were able to activate MAIT cells in an MR1-dependent manner.

Riboflavin Operons Are Also Present in Nonpneumococcal *Streptococcus* species

A bioinformatic investigation of 824 genomes of 69 different *Streptococcus* species revealed that the riboflavin operon was also present in other streptococci. Eleven different versions

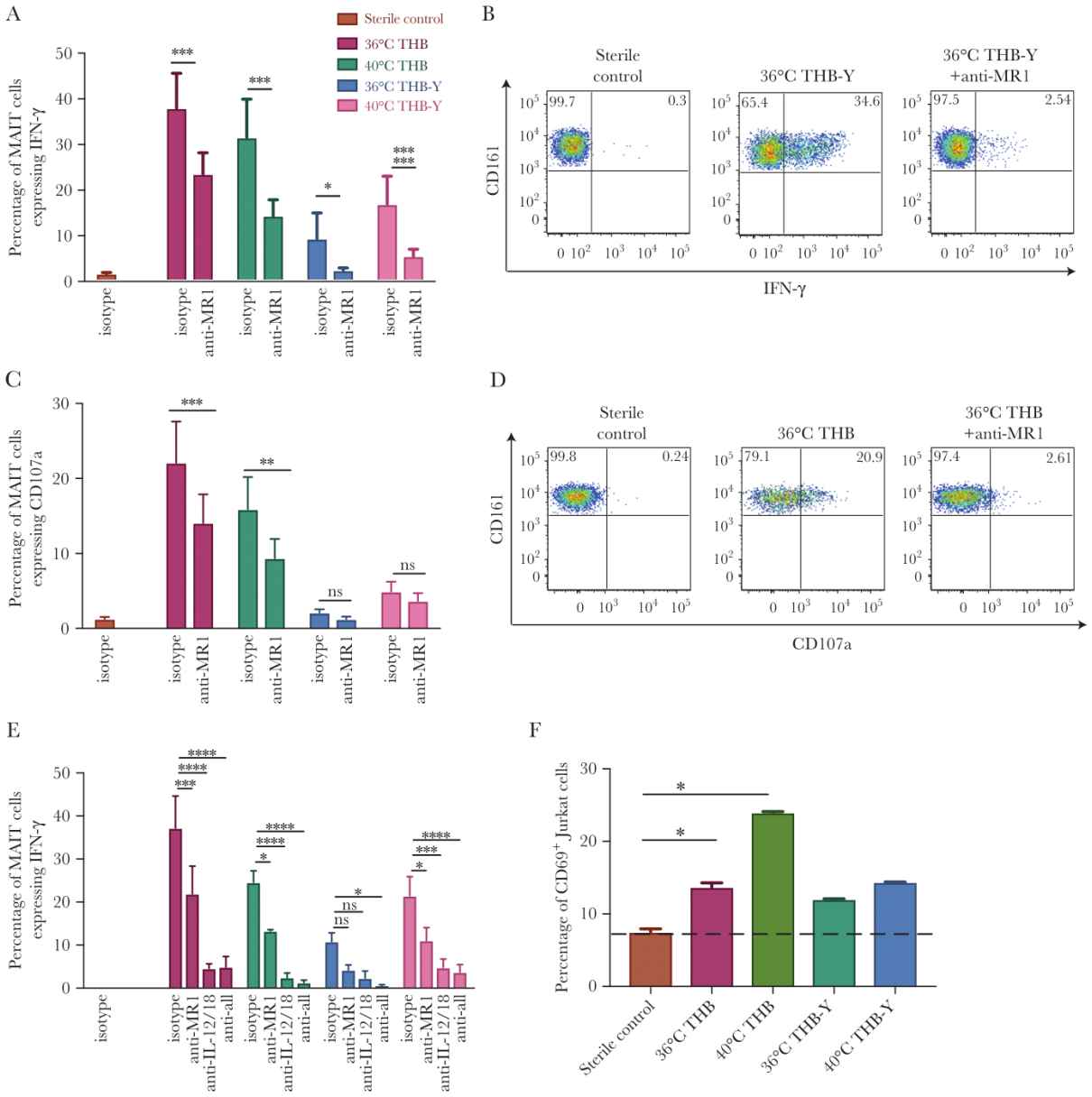
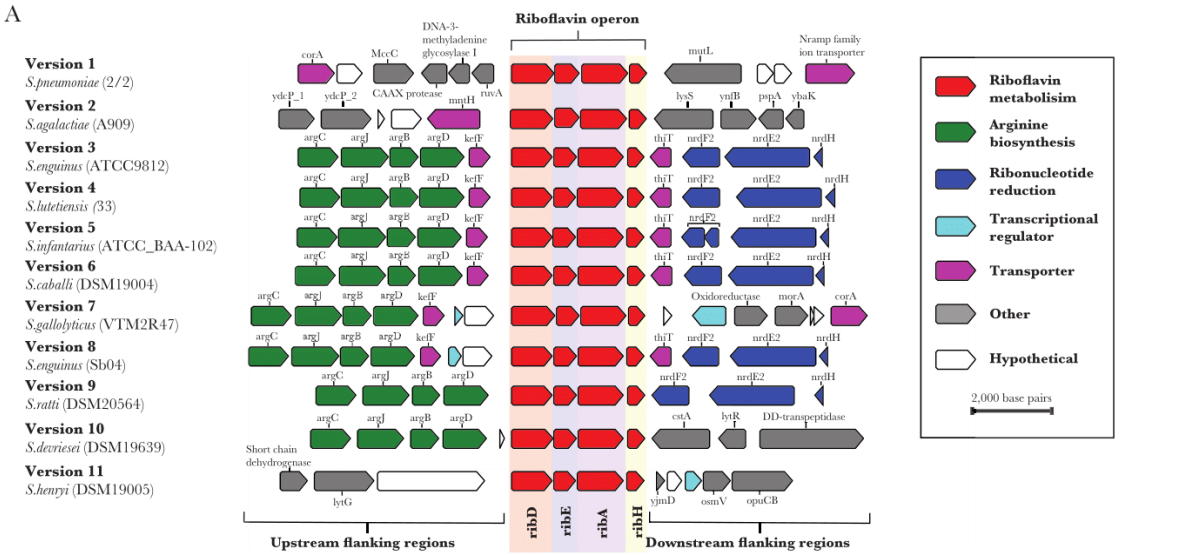


Figure 4. MR1-dependent activation of mucosal-associated invariant T (MAIT) cells by pneumococci in the presence of macrophages. *A–D*, Pneumococcal Molecular Epidemiological Network (PMEN) strain 34 was grown for 16 hours in Todd Hewitt broth with (THB) or without (THB-Y) yeast extract at either 36°C or 40°C (overnight). The bacteria were fixed and added to peripheral blood mononuclear cells (PBMCs) and monocyte-derived macrophages overnight in the presence or absence of anti-MR1 blocking antibody. The frequency of MAIT cells expressing interferon γ (IFN-γ; *A*) and representative example of IFN-γ expression from MAIT cells by fluorescence-activated cell-sorting analysis (*B*) are shown with isotype control or anti-MR1 blocking antibody. ****$P < .0001$, ***$P < .001$, and *$P < .05$ by 2-way analysis of variance (ANOVA) with the Sidak multiple comparisons test (n = 6). The frequency of MAIT cells expressing CD107a (*C*) and a representative example of CD107a expression from MAIT cells by fluorescence-activated cell-sorting analysis (*D*) are shown with isotype control or anti-MR1 blocking antibody. ***$P < .001$, **$P < .01$, or nonsignificant (NS) by 2-way ANOVA with the Sidak multiple comparisons test (n = 4). *E*, The PMEN34 strain was grown for 16 hours in THB-Y or THB at 36°C or 40°C. The bacteria were fixed immediately and added to PBMCs and monocyte-derived macrophages overnight in the presence or absence of indicated combinations of anti-MR1, anti–interleukin 12 (IL-12), and anti–interleukin 18 (IL-18) blocking antibodies. Frequencies of MAIT cells expressing IFN-γ are shown. ****$P < .0001$, ***$P < .001$, **$P < .01$, *$P < .05$, or NS by 2-way ANOVA with the Dunnett multiple comparisons test (n = 3). *F*, Jurkat cells expressing the MAIT cell T-cell receptor were cultured with monocyte-derived macrophages overnight with the PMEN34 strain. Activation was measured as the frequency of Jurkat cells expressing CD69. The dotted line indicates CD69 expression by Jurkat cells in the presence of sterile control. *$P < .05$ by repeated measures 1-way ANOVA with the Dunnett multiple comparisons test. All experiments were performed in duplicate, and data are representative of 2 independent experiments.

of the riboflavin operon were identified among 13 nonpneumococcal *Streptococcus* species (Supplementary Table 2). The majority of these riboflavin operons were located between genes involved in arginine biosynthesis (*argC*, *argJ*, *argB*, and *argD*) and those involved in ribonucleotide reduction (*nrdF2*, *nrdE2*, and *nrdH*; Figure 5A). Despite identical gene synteny between different versions of the riboflavin operon (Figure 5A), they differed greatly in nucleotide sequence identity (Figure 5B).

The pneumococcal version of the riboflavin operon was found among all 16 genomes of *Streptococcus pseudopneumoniae* and 2 genomes each of *Streptococcus mitis* and *Streptococcus oralis* (Figure 5C–D). No riboflavin operon genes were identified among the remaining 48 *S. mitis* and 48 *S. oralis* genomes. Pneumococci, *S. pseudopneumoniae*, *S. mitis*, and *S. oralis* are all closely related commensal streptococcal species that can exchange DNA with one another; therefore, the limited numbers of riboflavin operons present in *S. mitis* and *S. oralis* suggest that the examples identified here were the result of horizontal genetic exchange [32]. Some versions of riboflavin operons identified among other *Streptococcus* species were exclusively present in only one species. For example, version 2 was identified in all 50 genomes of *S. agalactiae* but no other species, whereas *Streptococcus equinus* contained 3 different versions of the riboflavin operon, one of which (version 5) was also found in *Streptococcus infantarius* (Figure 5D and Supplementary Table 2). Furthermore, >2400 *Streptococcus pyogenes* genomes



B

	Version 1	Version 2	Version 3	Version 4	Version 5	Version 6	Version 7	Version 8	Version 9	Version 10	Version 11
Version 1		60%	52%	52%	53%	53%	53%	53%	51%	49%	52%
Version 2	60%		53%	53%	53%	52%	53%	53%	52%	50%	53%
Version 3	52%	53%		88%	81%	80%	80%	80%	63%	61%	65%
Version 4	52%	53%	88%		88%	81%	80%	81%	63%	61%	65%
Version 5	53%	53%	81%	88%		81%	81%	83%	64%	61%	65%
Version 6	53%	52%	80%	81%	81%		88%	86%	63%	61%	65%
Version 7	53%	53%	80%	80%	81%	88%		83%	63%	61%	65%
Version 8	53%	53%	80%	81%	83%	86%	83%		64%	62%	66%
Version 9	51%	52%	63%	63%	64%	63%	63%	64%		72%	62%
Version 10	49%	50%	61%	61%	61%	61%	61%	62%	72%		59%
Version 11	52%	53%	65%	65%	65%	65%	65%	66%	62%	59%	

C

Streptococus species	*Number of genomes in the data set*	*Number of genomes with a riboflavin operon*	*Percentage of genomes with a riboflavin operon*
S.pseudopneumoniae	*16*	*16*	*100%*
S.mitis	*50*	*2*	*4%*
S.oralis	*50*	*2*	*4%*
S.agalactiae	*50*	*50*	*100%*
S.equinus	*27*	*27*	*100%*
S.lutetiensis	*2*	*2*	*100%*
S.infantarius	*2*	*2*	*100%*
S.gallolyticus	*17*	*17*	*100%*
S.pasteurianus	*5*	*5*	*100%*
S.henryi	*2*	*2*	*100%*
S.caballi	*1*	*1*	*100%*
S.ratti	*1*	*1*	*100%*
S.devriesei	*1*	*1*	*100%*

Figure 5. Evidence for different versions of riboflavin operons in other *Streptococcus* species. *A*, The riboflavin operon found in pneumococci (version 1) and its flanking genes are depicted and compared to 10 additional representative versions of the riboflavin operon found among other *Streptococcus* species. *B*, Matrix of pair-wise comparisons of nucleotide similarity among the 11 different versions of the riboflavin operon. *C*, Summary of the riboflavin operons found in 13 nonpneumococcal *Streptococcus* species. *D*, Phylogenetic tree constructed on the basis of the concatenated sequences of 53 ribosomal multilocus sequence type loci among 571 pneumococci and 824 *Streptococcus* species genomes. Branches of the tree were colored gray if no riboflavin operon was identified within the genome (eg, as seen for *Streptococcus pyogenes*), whereas other colors represent genomes in bacterial species that did possess a version of a riboflavin operon. The colored outer ring indicates the version of riboflavin operon that was identified in each genome or set of genomes. The rectangular box contains an expanded view of the circled area of the phylogenetic tree.

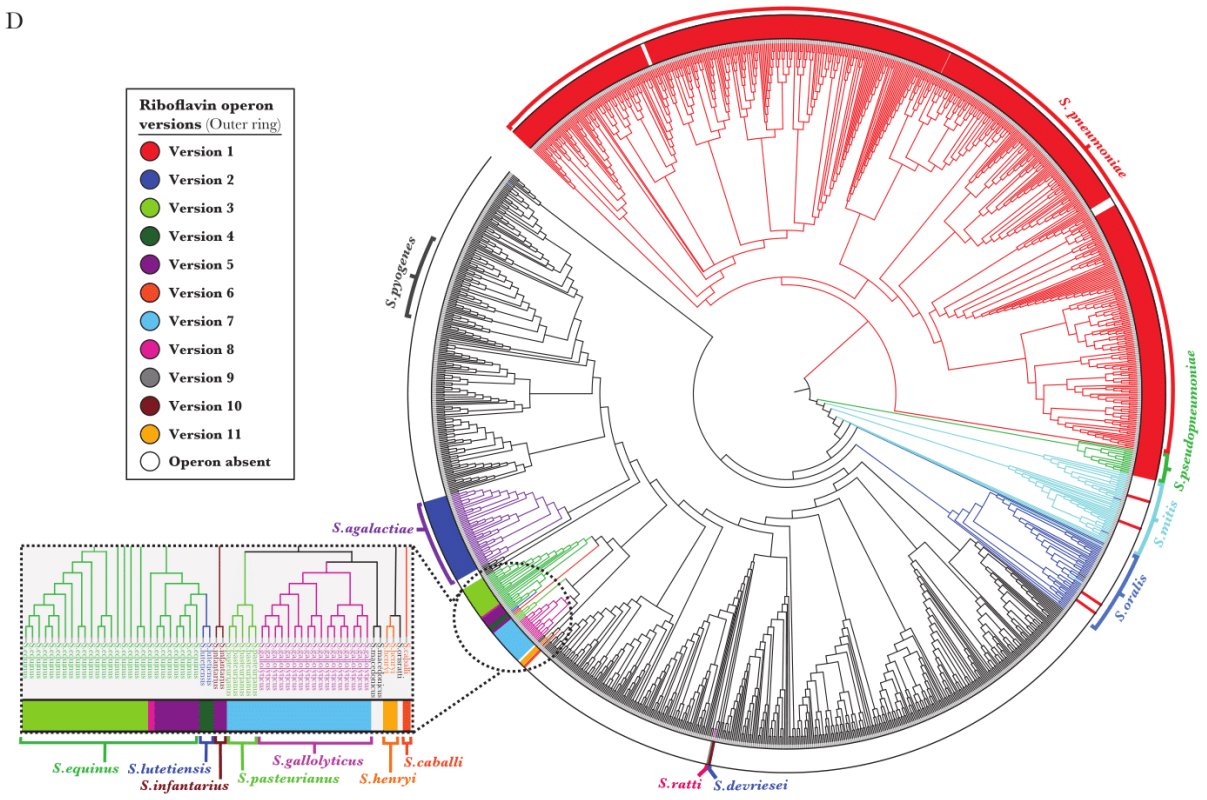


Fig 5. *Continued*

were investigated for the presence of *Rib* genes, but the riboflavin operon could not be found (Figure 5).

DISCUSSION

Our study is the first to demonstrate that MAIT cells can respond to and recognize pneumococci. Given the urgent global need to tackle antimicrobial-resistant strains and pneumococci not covered by the currently available vaccines, understanding the mechanism by which MAIT cells are activated by pneumococci provides a method of targeting a metabolic pathway that is highly conserved among pneumococci.

We found that MAIT cell activation by pneumococci through the MR1-restricted pathway was dependent on the type of antigen-presenting cell. There was a significant effect of blocking MR1 recognition by MAIT cells when the pneumococci were presented by macrophages but not when presented by monocytes. It has been suggested that monocytes are poor antigen presenters [33, 34]. Alveolar macrophages have been shown to be crucial for bacterial clearance in vivo [30, 31] and may be able to efficiently phagocytose the bacteria or provide more costimulatory signals. Even in the presence of macrophages, the MAIT cell response to pneumococci was highly codependent on cytokines. This is most likely due to the weak TCR signal induced by pneumococci, as seen by the significant but weak activation of Jurkat-MAIT cells even when using macrophages. The pneumococcus is a bacterium highly specialized to evade the host immune system by circumventing phagocytosis and antigen presentation, most notably by possessing a thick polysaccharide capsule, as well as through autolysis, which reduces the production of phagocyte-activating cytokines, such as tumor necrosis factor α and IL-12 [35]. The uniquely high sensitivity of MAIT cells to cytokines such as IL-12 [14] will allow these cells to boost the immune response and provide early IFN-γ production.

The pneumococcal disease with the biggest burden is pneumonia, which is the leading infectious cause of mortality in young children [1], as well as community-acquired pneumonia in elderly individuals [2]. Given that we show the mechanism by which MAIT cells respond to pneumococci and the critical role that MAIT cells play in vivo against lung infections [18, 20], it would be reasonable to suggest that MAIT cells may play a role in pneumococcal pneumonia. These cells may also be a critical factor in secondary pneumococcal pneumonia that follows influenza virus infection [36]. MAIT cells are reduced in frequency in the blood of patients with acute influenza virus infection, particularly in those who died of this disease [15, 37]. Whether the low numbers of MAIT cells in neonates [38,

39] or the decline in MAIT cell numbers in elderly individuals [40] and during influenza [15] affects the susceptibility of these patients to pneumococcal pneumonia will be important to investigate in in vivo models [41].

We used a population genomics approach to assess the prevalence and diversity of the riboflavin operon among a large and diverse collection of pneumococci. This revealed that riboflavin genes are nearly ubiquitous and highly conserved at a nucleotide level among pneumococci recovered over the past century. We also found that a number of nonpneumococcal *Streptococcus* species possess these genes, including other commensal streptococci, such as *S. agalactiae*, presenting opportunities for future studies. For example, given that MAIT cells reside in the female genital mucosa [42], it would be important to explore whether there is a MAIT cell response in the context of vaginal colonization of *S. agalactiae* among pregnant women and invasive neonatal infections [43]. Of note, not all streptococcal genomes investigated possessed a riboflavin operon (notably *S. pyogenes*, which is a major invasive pathogen), while there was also evidence of *S. mitis* and *S. oralis* acquiring the riboflavin operon through horizontal genetic exchange. Hence, caution must be exercised when extrapolating findings based on a small number of bacterial strains to the population as a whole, since they may not be representative.

Overall, these data show a robust response of MAIT cells to pneumococci and conservation of the relevant biosynthetic pathway in this organism and other closely related *Streptococcus* species. Given the low levels of MAIT cells among individuals in early life and their decline in older individuals—the highest-risk populations for invasive pneumococcal disease—further understanding of the functional role of MAIT cells in vivo in host defense against this major pathogen is of interest.

Supplementary Data

Supplementary materials are available at *The Journal of Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Acknowledgments. We thank Prof Regine Hakenbeck at the University of Kaiserslautern, for the stock cultures of streptococci that were newly sequenced in this study; and the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by the Wellcome Trust; grant 090532/Z/09/Z), for the generation of the RNA sequencing data.

Financial support. This work was supported by the Wellcome Trust (grants WT109965MA [to P. K.] and 083511/Z/07/Z [to A. B. B.] and Biomedical Research Fund award 04992/Z/14/Z [to Martin J. C. Maiden, Keith A. Jolley, and A. B. B., for BIGSdb genome database support]); Cancer Research United Kingdom (grant C399/A2291 to V. C.); the National Institutes of Health (grant NIHU19AI082630 to P. K.); the National Institute for Health Research (NIHR) Senior Fellowship (to P. K.); the NIHR Biomedical Research Centre, Oxford (to P. K.); and the University of Oxford John Fell Fund (grant 123/734 to A. B. B.).

Potential conflicts of interest. All authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Liu L, Oza S, Hogan D, et al. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet **2015**; 385:430–40.
2. Drijkoningen JJC, Rohde GGU. Pneumococcal infection in adults: burden of disease. Clin Microbiol Infect **2014**; 20:45–51.
3. Mehr S, Wood N. Streptococcus pneumoniae–a review of carriage, infection, serotype replacement and vaccination. Paediatr Respir Rev **2012**; 13:258–64.
4. World Health Organization (WHO). Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. Geneva: WHO, **2017**. http://www.who.int/medicines/publications/global-priority-list-antibiotic-resistant-bacteria/en/. Accessed 19 December 2017.
5. Huang SS, Platt R, Rifas-Shiman SL, Pelton SI, Goldmann D, Finkelstein JA. Post-PCV7 changes in colonizing pneumococcal serotypes in 16 Massachusetts communities, 2001 and 2004. Pediatrics **2005**; 116:e408–13.
6. Singleton RJ, Hennessy TW, Bulkow LR, et al. Invasive pneumococcal disease caused by nonvaccine serotypes among alaska native children with high levels of 7-valent pneumococcal conjugate vaccine coverage. JAMA **2007**; 297:1784–92.
7. Godfrey DI, Uldrich AP, McCluskey J, Rossjohn J, Moody DB. The burgeoning family of unconventional T cells. Nat Immunol **2015**; 16:1114–23.
8. Treiner E, Duban L, Bahram S, et al. Selection of evolutionarily conserved mucosal-associated invariant T cells by MR1. Nature **2003**; 422:164–9.
9. Dusseaux M, Martin E, Serriari N, et al. Human MAIT cells are xenobiotic-resistant, tissue-targeted, CD161hi IL-17-secreting T cells. Blood **2011**; 117:1250–9.
10. Kjer-Nielsen L, Patel O, Corbett AJ, et al. MR1 presents microbial vitamin B metabolites to MAIT cells. Nature **2012**; 491:717–23.
11. Gherardin NA, Keller AN, Woolley RE, et al. Diversity of T Cells Restricted by the MHC Class I-Related Molecule MR1

Facilitates Differential Antigen Recognition. Immunity **2016**; 44:32–45.
12. Corbett AJ, Eckle SB, Birkinshaw RW, et al. T-cell activation by transitory neo-antigens derived from distinct microbial pathways. Nature **2014**; 509:361–5.
13. Eckle SB, Birkinshaw RW, Kostenko L, et al. A molecular basis underpinning the T cell receptor heterogeneity of mucosal-associated invariant T cells. J Exp Med **2014**; 211:1585–600.
14. Ussher JE, Bilton M, Attwod E, et al. CD161++ CD8+ T cells, including the MAIT cell subset, are specifically activated by IL-12+IL-18 in a TCR-independent manner. Eur J Immunol **2014**; 44:195–203.
15. van Wilgenburg B, Scherwitzl I, Hutchinson EC, et al. MAIT cells are activated during human viral infections. Nat Commun **2016**; 7:11653.
16. Le Bourhis L, Martin E, Péguillet I, et al. Antimicrobial activity of mucosal-associated invariant T cells. Nat Immunol **2010**; 11:701–8.
17. Georgel P, Radosavljevic M, Macquin C, Bahram S. The non-conventional MHC class I MR1 molecule controls infection by Klebsiella pneumoniae in mice. Mol Immunol **2011**; 48:769–75.
18. Meierovics A, Yankelevich WJ, Cowley SC. MAIT cells are critical for optimal mucosal immune responses during in vivo pulmonary bacterial infection. Proc Natl Acad Sci U S A **2013**; 110:E3119–28.
19. Chua WJ, Truscott SM, Eickhoff CS, Blazevic A, Hoft DF, Hansen TH. Polyclonal mucosa-associated invariant T cells have unique innate functions in bacterial infection. Infect Immun **2012**; 80:3256–67.
20. Meierovics AI, Cowley SC. MAIT cells promote inflammatory monocyte differentiation into dendritic cells during pulmonary intracellular infection. J Exp Med **2016**; 213:2793–809.
21. Liu S, Sela S, Cohen G, Jadoun J, Cheung A, Ofek I. Insertional inactivation of streptolysin S expression is associated with altered riboflavin metabolism in Streptococcus pyogenes. Microb Pathog **1997**; 22:227–34.
22. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods **2012**; 9:357–9.
23. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol **2010**; 11:R106.
24. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **2010**; 11:595.
25. 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**; 7:42976.
26. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **2008**; 18:821–9.
27. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics **2014**; 15:211.
28. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol **2012**; 13:R56.
29. Jolley KA, Bliss CM, Bennett JS, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology **2012**; 158:1005–15.
30. Broug-Holub E, Toews GB, van Iwaarden JF, et al. Alveolar macrophages are required for protective pulmonary defenses in murine Klebsiella pneumonia: elimination of alveolar macrophages increases neutrophil recruitment but decreases bacterial clearance and survival. Infect Immun **1997**; 65:1139–46.
31. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet **2004**; 20:44–50.
32. Werno AM, Christner M, Anderson TP, Murdoch DR. Differentiation of Streptococcus pneumoniae from non-pneumococcal streptococci of the Streptococcus mitis group by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol **2012**; 50:2863–7.
33. Kamphorst AO, Guermonprez P, Dudziak D, Nussenzweig MC. Route of antigen uptake differentially impacts presentation by dendritic cells and activated monocytes. J Immunol **2010**; 185:3426–35.
34. Samstein M, Schreiber HA, Leiner IM, Susac B, Glickman MS, Pamer EG. Essential yet limited role for CCR2$^+$ inflammatory monocytes during Mycobacterium tuberculosis-specific T cell priming. Elife **2013**; 2:e01086.
35. Martner A, Skovbjerg S, Paton JC, Wold AE. Streptococcus pneumoniae autolysis prevents phagocytosis and production of phagocyte-activating cytokines. Infect Immun **2009**; 77:3826–37.
36. Morens DM, Taubenberger JK, Fauci AS. Predominant role of bacterial pneumonia as a cause of death in pandemic influenza: implications for pandemic influenza preparedness. J Infect Dis **2008**; 198:962–70.
37. Loh L, Wang Z, Sant S, et al. Human mucosal-associated invariant T cells contribute to antiviral influenza immunity via IL-18-dependent activation. Proc Natl Acad Sci U S A **2016**; 113:10133–8.
38. Walker LJ, Kang YH, Smith MO, et al. Human MAIT and CD8αα cells develop from a pool of type-17 precommitted CD8+ T cells. Blood **2012**; 119:422–33.
39. Koay HF, Gherardin NA, Enders A, et al. A three-stage intrathymic development pathway for the mucosal-associated invariant T cell lineage. Nat Immunol **2016**; 17:1300–11.

40. Novak J, Dobrovolny J, Novakova L, Kozak T. The decrease in number and change in phenotype of mucosal-associated invariant T cells in the elderly and differences in men and women of reproductive age. Scand J Immunol **2014**; 80:271–5.
41. Smith NM, Wasserman GA, Coleman FT, et al. Regionally compartmentalized resident memory T cells mediate naturally acquired protection against pneumococcal pneumonia. Mucosal Immunol **2017**. doi: 10.1038/mi.2017.43.
42. Gibbs A, Leeansyah E, Introini A, et al. MAIT cells reside in the female genital mucosa and are biased towards IL-17 and IL-22 production in response to bacterial stimulation. Mucosal Immunol **2017**; 10:35–45.
43. Martins ER, Pessanha MA, Ramirez M, Melo-Cristino J; Portuguese Group for the Study of Streptococcal Infections. Analysis of group B streptococcal isolates from infants and pregnant women in Portugal revealing two lineages with enhanced invasiveness. J Clin Microbiol **2007**; 45:3224–9.