University of Southampton

Faculty of Medicine

Clinical and Experimental Sciences

Genome wide association study in antimicrobial resistance of disease-causing *Streptococcus pneumoniae* in Singapore

DOI https://doi.org/10.5258/SOTON/D1944

by

Rachael Wallis

Thesis for the degree of Doctor of Philosophy

September 2021

University of Southampton

Abstract

Faculty of Medicine

Clinical and Experimental Sciences

Thesis for the degree of Doctor of Philosophy

Genome wide association study in antimicrobial resistance of disease-causing *Streptococcus pneumoniae* in Singapore

by

Rachael Wallis

Pneumococcal disease caused by the bacterium *Streptococcus pneumoniae* is responsible for substantial mortality and morbidity worldwide, making it a global public health concern. *S. pneumoniae* has the exceptional ability to adapt to its surroundings to overcome selection pressures and maintain survival; this includes adaptation in response to host immune defences and antibiotics.

Increasing levels of resistance to commonly used antimicrobials have been identified in many countries and threaten the effectiveness of present and future treatment options. A longitudinal collection of *S. pneumoniae* isolates in Singapore offered the opportunity to track evolution. The serotypes most associated with invasive pneumococcal disease were serotypes 4, 8, 20 7A, 19A and 3, and many of these serotypes predominate in adult infection; serotype 6B predominates in paediatric infection. Resistance of isolates was highest to cotrimoxazole (63%), erythromycin (58%), tetracycline (58%) and doxycycline (58%).

While basic molecular epidemiology is a cornerstone of this analysis, the collection of ~2000 isolates, accompanied with some basic clinical information and bacterial antibiotic resistance phenotypes, provided the opportunity to perform a genome wide association study (GWAS) on antimicrobial non-susceptibility. High levels of recombination were identified in this dataset therefore the impact of this on population structure was assessed to ensure an accurate correction for the GWAS. A comparison in the ability of Gubbins and ClonalFrameML to identify recombination from this dataset showed ClonalFrameML was more conservative and concise in recombination calls and could process larger numbers of isolates. These findings showed how it was possible to perform a complete recombination analysis on large, diverse datasets. The recombination sites identified by ClonalFrameML were not a major contributor in the population clustering of isolates.

GWAS' were performed to identify single nucleotide polymorphisms (SNPs) associated with non-susceptibility to penicillin, cotrimoxazole, erythromycin, clindamycin, chloramphenicol, tetracycline and doxycycline. This showed good association with drug resistance, with SNPs identified in genes involved in the peptidoglycan pathway, folate metabolism, protein synthesis and DNA synthesis. Many of these confirmed previous findings in the field. In addition, SNPs were associated with erythromycin resistance in new loci involved in DNA synthesis (*dnaG, sigA*). Although validation is required, these additional results provide a new opportunity to develop insights into the biological mechanisms that underlie the important clinical outcomes of drug resistance.

Table of Contents

Table of Contents i

Table of Tables vii

Table of Figures ix

Research Thesis: Declaration of Authorship xv

Acknowledgements xix

Definitions and Abbreviations xxi

Chapter 1 Introduction 25

1.1.1 Genomic analysis of bacteria 25

1.2 Streptococcus pneumoniae 26

1.3 Pneumococcal disease 26

1.4 Epidemiology 26

1.5 Prevention and treatment 27

1.5.1 Vaccination 27

1.5.1.1 Pneumococcal Polysaccharide Vaccine (PPV) 28

1.5.1.2 Pneumococcal Conjugate Vaccine (PCV) 28

1.5.1.3 Effect of vaccination on circulating pneumococcus 29

1.5.2 Antimicrobials 29

1.6 Challenges arising from *S. pneumoniae* adaptation 30

1.6.1 Evasion of Pneumococcal Conjugate Vaccine 31

1.6.2 Antimicrobial resistance 31

1.6.2.1 Contribution of point mutation in generating resistance to antibiotics 32

1.6.2.2 Contribution of horizontal transfer of genetic material in generating resistance to antibiotics 34

1.7 Identification of resistance mechanisms in pneumococcal populations 36

1.8 Hypothesis 37

1.9 Aims and Objectives 38

Chapter 2 Methods 39

2.1 Singapore Pneumococcal dataset 39

2.1.1 Participating hospitals 39

2.1.2 Ethics 39

2.1.3 Study isolates 39

2.1.4 Laboratory methods 40

2.1.4.1 Bacterial cultures 40

2.1.4.2 DNA extraction and sequencing library preparation 40

2.1.5 Genome sequence analysis 40

2.1.6 Assembly statistics and sample inclusion criteria 42

2.1.7 Classification of metadata 44

2.2 Analysis of recombination programs 44

2.2.1 Generation of sequences containing artificial recombination events and/or diversity 46

2.2.2 Datasets for recombination analysis using simulated sequence with a single recombination 47

2.2.3 Datasets for recombination analysis using simulated sequence with two recombination events 47

2.2.4 Calculating divergence value 48

2.2.5 Recombination analysis 48

2.2.6 False positivity rate of programs 48

2.3 Background diversity in clinical isolates 49

2.4 Construction of phylogenetic trees 49

2.5 Recombination analysis on Singapore dataset of 2,059 *S pneumoniae* isolates..... 51

2.6 Comparing population structure of 1,828 *S. pneumoniae* isolates. 51

2.7 Genome wide association studies for an antimicrobial resistant phenotype. 51

2.8 Identification of core genome from 2,059 *S. pneumoniae* isolates 52

Chapter 3 Pneumococcal disease in Singapore 53

3.1 Introduction 53

3.1.1 Difference in disease capabilities between serotypes 53

3.1.2 Pneumococcal disease in Singapore 54

3.2 Results 57

3.2.1 Descriptive Epidemiology of Singapore Pneumococcal dataset 57

3.2.2 Serotypes associated with disease 63

3.2.3 Change in serotype distribution over time 66

3.2.4 Proportion of vaccine serotypes responsible for disease 68

3.2.5 Levels of antimicrobial resistance in *S. pneumoniae* 70

3.2.6 Level of diversity in the dataset 73

3.3 Discussion 75

3.3.1 Analysis of sample collection 75

3.3.2 Epidemiological interpretation based on pneumococcal dataset of 2,059 75

3.4 Conclusion and Future work 77

Chapter 4 Detection of recombination from clinical isolates 79

4.1 Introduction 79

4.2 Recombination in bacteria 80

4.2.1 Recombination in *S. pneumoniae* 81

4.3 Detection of recombination 82

4.3.1 Programs available for recombination analysis 82

4.3.1.1 The Genealogies Unbiased by Recombination in Nucleotide Sequences (GUBBINS) 83

4.3.1.2 ClonalFrameML 83

4.4 Application of recombination programs on 2,059 *S. pneumoniae* isolates 84

4.5 Results 86

4.5.1 Determining dataset size to compare Gubbins and ClonalFrameML 86

4.5.2 Comparison of program output 87

4.5.2.1 Detection of artificial recombination from datasets consisting of clinical isolates 87

4.5.2.2 Detection of artificial recombination in different mutation backgrounds 89

4.5.2.3 Internal consistency in recombination calls between Gubbins and ClonalFrameML 92

4.5.2.4 Determining the distance at which recombination programs distinctly identify two adjacent recombination events....................................... 94

4.6 Discussion.. 97

4.6.1 Capacity of program to detect recombination from a diverse dataset........... 97

4.6.2 Comparison of internal consistency between recombination programs........ 99

4.7 Conclusion and future work... 101

Chapter 5 Antimicrobial resistance phenotypes analysed by Genome Wide Association study 103

5.1 Introduction 103

5.1.1 Application of genome wide association studies outside a human host 103

5.1.2 Considerations of microbial GWAS.. 105

5.1.3 GWAS studies of pneumococcal antimicrobial resistance 106

5.2 Results... 109

5.2.1 Defining population structure for genome wide association study 109

5.2.1.1 Recombination analysis of 2,059 clinical *S. pneumoniae* isolates...... 109

5.2.1.2 Application of programs to compare population structure between whole genome and recombination-free genome datasets............... 111

5.2.2 Genome wide association study of 1,828 isolates 117

5.2.2.1 Identification of SNPs associated with penicillin non-susceptibility from whole genome dataset 117

5.2.2.2 Comparison of penicillin non-susceptibility between association studies.. 124

5.2.3 Identification of SNPs associated with cotrimoxazole non-susceptibility from whole genome dataset 127

5.2.4 Identification of SNPs associated with erythromycin non-susceptibility from whole genome dataset 129

5.2.5 Identification of SNPs associated with clindamycin non-susceptibility from whole genome dataset 131

5.2.6 Identification of SNPs associated with chloramphenicol, doxycycline and tetracycline non-susceptibility from whole genome dataset....................... 133

5.2.7 GWAS of penicillin non-susceptibility using recombination-free genomes ..135

5.3 Discussion 137

5.3.1 The effect of recombination on population structure 137

5.3.2 Genome wide association studies to identify SNPs associated with antimicrobial resistance 138

5.3.2.1 Penicillin 139

5.3.2.2 Cotrimoxazole....142

5.3.2.3 Macrolides 143

5.3.2.4 Clindamycin 144

5.3.2.5 Chloramphenicol 144

5.3.2.6 The role of recombination in the genetic association of penicillin non-susceptibility....145

5.4 Conclusions and future work 146

Chapter 6 Discussion 149

6.1 Introduction....149

6.2 Key findings 149

6.2.1 Epidemiology....150

6.2.2 Identification of recombination from clinical isolates 151

6.2.3 GWAS to identify SNPs association with antimicrobial resistance 152

6.2.4 Limitations....153

6.2.5 Concluding remarks....154

Appendix A Constructing simulated sequence containing one recombination....157

Appendix B Phylogenetic tree of 2,059 *S. pneumoniae* isolates specifying ‘inner’ and ‘outer’ clades 159

Appendix C Basic dataset to investigate the effect of background mutation on the detection of artificial recombination 161

Appendix D Dataset containing a similar sequence to investigate the effect of background mutation on recombination detection....163

Appendix E Dataset to investigate the effect of background mutation on recombination detection when the recombination event has a low mutation rate compared to the rest of the chromosome 165

Appendix F Construction of sequence containing two recombination events with no background mutation 167

Appendix G Construction of sequence containing two recombination events with incorporated background mutation......169

Appendix H Basic dataset for analysis of simulated sequence with two recombination events......171

Appendix I QQ plots for the GWAS of penicillin non-susceptibility 173

Appendix J Summary of SNP frequencies in genes associated with penicillin resistance between studies 175

Appendix K QQ plots for the GWAS of cotrimoxazole, erythromycin and clindamycin resistance 179

Appendix L Summary of SNPs in common genes identified in this GWAS for cotrimoxazole resistance and the GWAS for cotrimoxazole resistance by Mobegi *et al* (2017) (255)......181

Appendix MSummary of SNPs in common genes identified in this GWAS for erythromycin resistance and the GWAS for erythromycin resistance by Mobegi *et al* (2017) (255)......183

Appendix N QQ plots for the GWAS of chloramphenicol resistance 185

Appendix O QQ plots for the GWAS of doxycycline and tetracycline resistance......187

Appendix P QQ plots for the GWAS of penicillin non-susceptibility using the SvIR classification from recombination-free genomes 189

Appendix Q Isolate laboratory data for 313 *S. pneumoniae* isolates collected 2013-2016 191

Appendix R Isolate antimicrobial susceptibility data for 313 *S. pneumoniae* isolates collected 2013-2016......209

List of References 221

Table of Tables

Table 1 Number of contigs and genome size identified by assemblers VelvetOptimiser and SPAdes 43

Table 2 Proportion of resistance identified from pneumococcal populations in Singapore. n=number of isolates tested for susceptibility to antimicrobials...................56

Table 3 Distribution of isolates across patient age groups collected by participating hospitals (KKH, NUH, SGH, TTSH and NHPL) between 1997 – 2016..57

Table 4 Comparison in the number of IPD cases notified to Ministry of Health Singapore with the number of IPD isolates collected in the present study...................................61

Table 5 Site of infection and frequency of isolate collection in IPD (n=1089) and non-IPD (n=846) 62

Table 6 Serotype distribution of 2,059 *S. pneumoniae* isolates (IPD=898) and (non-IPD=733) for the most frequently identified serotypes. Rank ordered by serotype specific odds ratio (OR). Serotypes associated with specific allocation to a disease classification highlighted in bold. 64

Table 7 Serotypes associated with specific allocation to an age group rank ordered by odds ratio (OR). Significant OR where 95% CI do not span 1 are highlighted in bold. 65

Table 8 Total number of isolates with antimicrobial susceptibility data showing the proportions classified as sensitive, intermediately resistant or resistant..........................71

Table 9 Summary of recombination detection by Gubbins and ClonalFrameML from simulated and clinical datasets...90

Table 10 Summary statistics following recombination analysis by Gubbins and ClonalFrameML of 100 subsets consisting of 10 randomly selected clinical *S. pneumoniae* isolates and the artificial recombination sequence. Gubbins results from a comparative study by Croucher *et al* (2011) (73) also included for reference....................93

Table 11 Quantitative assessment of similarity between phylogenies using the Robinson Foulds (RF) metric calculated by the Environment for Tree Exploration toolkit (210)109

Table 12 Frequency of genome wide association SNPs associated with penicillin non-susceptibility identified in SvIR analysis (section 5.2.2.1.1), Mobegi *et al* (2017) (255) and in Chewapreecha *et al* (2014) (179) 125

Table 13 Description of six SNPs co-detected in this study, in Chewapreecha *et al* (2014) (179) and Mobegi *et al* (2017) (255).. 126

Table of Figures

Figure 1 Site and mechanism of action for antibiotics used to treat *S. pneumoniae* infection. PABA (para-aminobenzoic acid) DHF (Dihydrofolic acid) THF (tetrahydrofolic acid) DHPS (dihydropteroate synthase) DHR (dihydrofolate reductase) 30

Figure 2 Flow diagram summarising the workflow involved in the recombination analysis of Gubbins and ClonalFrameML 45

Figure 3 Maximum-likelihood phylogenetic trees constructed by FastTree using whole genomes of 2,059 isolates of *S. pneumoniae* and 5 artificial sequences with varying levels of background mutation incorporated (not including known recombination region). The red arrows show the position of sequences with incorporated background mutation; (a) 0.001% mutation; (b) 0.01% mutation; (c) 0.1% mutation; (d) 1% mutation; (e) 2% mutation. 50

Figure 4 Total number of *S. pneumoniae* isolates collected and stored by all participating hospitals (KKH, NUH, SGH, TTSH and NHPL) each year between 1997 – 2016 (n=1,756). The red line represents isolates causing IPD (n=1,062) and the blue represents isolates causing non-IPD (n=833) 58

Figure 5 Total number of *S. pneumoniae* isolates collected and stored by specific hospitals; KKH 824 isolates (IPD=311, non-IPD=513), NUH 204 isolates (IPD=78, non-IPD=126), SGH 317 isolates (IPD=279, non-IPD=38) and TTSH 411 isolates (IPD=256, non-IPD=155) that had both year and site of infection information between 1997 – 2016. The red line represents isolates causing IPD and the blue represents isolates causing non-IPD 59

Figure 6 Proportion of total disease caused by the seven serotypes most isolated from 2,059 *S. pneumoniae* isolates. Of these serotypes, (a) shows serotypes 23F, 19F, 14 present in PCV7, (b) shows serotypes 3, 19A, 6A present in PCV13, (c) shows serotype 6E not included in either PCV7/13. The dashed vertical line illustrated the respective years PCV7 and PCV13 was implemented into national immunisation programs. 67

Figure 7 Proportion of total isolates that contain a serotype present in the vaccines PCV7 (top), PCV13 (middle) and PPV23 (bottom). The dashed vertical line illustrated the respective years PCV7 and PCV13 was implemented into national immunisation programs...69

Table of Figures

Figure 8 Change in the proportion of total isolates with resistance to the described antibiotics between 1997 – 2016. Total number of isolates with supporting antibiotic resistance data for the specified year was (a) n=939, (b) n=1045, (c) n=1491, (d) n=177, (e) n=1751, (f) n=356. 72

Figure 9 A maximum-likelihood phylogenetic tree constructed by FastTree using whole genomes of 2,059 isolates of *S. pneumoniae* 73

Figure 10 goeBURST analysis of 2,059 *S. pneumoniae* isolates showing 250 clonal complexes. The largest CC group 0 (n=264) is showing ST81 as the predominant strain. The other main group CC1 (n=249) has the predominant strain ST236. CC69 (n= 261) encompassed all the NF strains. 74

Figure 11 Duration of recombination analysis for subset sizes of 11, 21 and 31 isolates for Gubbins (red) and ClonalFrameML (blue). All ClonalFrameML results are representative of 100 runs for each subset size. Subsets of 11 isolates for Gubbins are representative of 100 runs. Subsets of 21 isolates for Gubbins is only representative of four runs, the rest did not complete. No runs of 31 completed with Gubbins 86

Figure 12 Minimum dissimilarity value of the closest isolate to wildtype in 100 subsets of 11 processed by ClonalFrameML and Gubbins. Symbol represents the outcome of recombination detection. Same result seen in both Gubbins and ClonalFrameML 88

Figure 13 Minimum dissimilarity value of the closest isolate to wildtype in 100 subsets of 21 isolates (red) and 31 isolates (green). Outcome of recombination detection represented by symbol. Datasets of 21 and 31 isolates were selected independently of one another and results presented are only for ClonalFrameML. 88

Figure 14 Minimum dissimilarity values in subsets containing simulated sequences of increasing background mutation rate. Blue values are from subsets that do not contain a similar sequence in the recombination analysis. Red values are subsets with a similar sequence to the recombination sequence. Blue data points are present under the red in 0%, 0.001% and 0.1% background mutations. 91

Figure 15 Venn diagram illustrating level of internal consistency in recombination calls (bp) between Gubbins and ClonalFrameML from 100 subsets consisting of 10 randomly selected clinical *S. pneumoniae* isolates and the artificial recombination sequence. A total of 1,911,701 bp was identified as recombination in Gubbins and 336,930 bp in ClonalFrameML......................93

Figure 16 Example of recombination called across the genome by Gubbins and ClonalFrameML. The total number of recombination bp called by ClonalFrameML for all six recombination events was 4,633bp. The total number of recombination bp called by Gubbins in the recombination even shown was 24,820bp. 94

Figure 17 Illustration of the distances required by programs to distinguish between two recombination events when the size of the recombination event ranged from 1kbp to 10kbp..96

Figure 18 Distance between recombination events required by programs to identify two recombination events in the presence of background mutation. Size of recombination event in all subsets is 10kb. Gubbins was not able to identify recombination events in a 1% or 2% mutation rate (not plotted). ClonalFrameML could not separate recombination events within the tested ranges at 2% (not plotted) 96

Figure 19 A maximum-likelihood phylogenetic tree constructed by FastTree using recombination-free genomes (14% of the genome) of 2,059 isolates of *S. pneumoniae*......110

Figure 20 A maximum-likelihood phylogenetic tree constructed by FastTree using core genomes (366 genes) of 2,059 isolates of *S. pneumoniae*. 110

Figure 21 Principal component analysis showing PC1 and PC2 of 1,828 isolates of *S. pneumoniae*. Clusters of isolates are labelled 1-6 for the whole genome dataset and 1-4 for the recombination-free dataset. In whole genome data, cluster 1 (n=119), cluster 2 (n=171), cluster 3 (n=13), cluster 4 (n=1113), cluster 5 (n=164), cluster 6 (n=248). In recombination-free genome data, cluster 1 (n=249), cluster 2 (n=85), cluster 3 (n=1190), cluster 4 (n=304)..112

Figure 22 Cumulative scree plots of the proportion of variation for the first 10 principal components in the PCA analysis of whole and recombination-free genome datasets 113

Figure 23 Principal component analysis showing PC1 and PC2 of 1,828 isolates of *S. pneumoniae*. The plot is coloured by age groups and corresponding tables show the proportion of isolates within each of the clusters described in Figure 21. ...114

Figure 24 Principal component analysis showing PC1 and PC2 of 1,828 isolates of *S. pneumoniae*. The plot is coloured by disease type and corresponding tables show the proportion of isolates within each of the clusters described in Figure 21.... 115

Figure 25 Principal component analysis showing PC1 and PC2 of 1,828 isolates of S. *pneumoniae*. The plot is coloured by the three main clonal clusters (CC) found within the dataset. Corresponding tables show the proportion of isolates within each of the clusters described in Figure 21 116

Figure 26 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvIR analysis. Genes that are specifically associated with penicillin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction 119

Figure 27 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvI analysis. Genes that are specifically associated with penicillin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction 121

Figure 28 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvR analysis. Genes that are specifically associated with penicillin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction 123

Figure 29 A Venn diagram summarising the number of significant association SNPs in each of the studies, and the number of those that are co detected in this study, in Chewapreecha *et al* (2014) (179) and Mobegi *et al* (2017) (255) 126

Figure 30 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and cotrimoxazole non-susceptibility. Genes that are specifically associated with cotrimoxazole resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction. 128

Figure 31 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and erythromycin non-susceptibility. Genes that are specifically associated with erythromycin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction 130

Figure 32 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and clindamycin non-susceptibility. Genes that are specifically associated with clindamycin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction....................132

Figure 33 Manhattan plot showing the position of significant associations in the genome and their relative p-value in chloramphenicol GWAS. Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and chloramphenicol non-susceptibility. Genes that are specifically associated with chloramphenicol resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction. 134

Figure 34 Manhattan plot showing significant association SNP in SvIR classification of GWAS from recombination-free datasets. Manhattan plot showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvIR analysis of recombination-free genomes. The gene specifically associated with penicillin resistance is labelled and its position represented by the vertical line. The significant association SNP is coloured red and the dotted horizontal line represents the significance cut-off after Bonferroni correction. 136

List of Accompanying Materials

DOI https://doi.org/10.5258/SOTON/D1944

Research Thesis: Declaration of Authorship

Print name: Rachael Wallis

Title of thesis: Genome wide association study in antimicrobial resistance of disease-causing *Streptococcus pneumoniae* in Singapore

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signature: ....

Date: 03/09/2021

Acknowledgements

Undertaking this PhD has allowed me to grow and challenge myself both academically and personally in ways that I had never imagined, and for that I will be forever grateful to have been given the opportunity. I have been lucky enough to work with amazing research groups and phenomenal scientists in both the UK and Singapore.

I wish to express my thanks for the guidance and support of my supervisory team; Dr. Stuart Clarke, Dr. David Cleary, Professor Martin Hibberd, Associate Professor Swaine Chen, and Dr Elita Jauneikaite. Working with you, and your incredible groups, has been a privilege and has taught me so many things. Many thanks to the collaborators working in the study hospitals and to the sequencing team at the Genome Institute of Singapore. I would additionally like to thank Dr Paola De Sessions and Dr Yaa Oppong for their help, and an additional thank you to Professor Peter Friedmann for all your time and suggestions offered in the final few months.

My friends and family have been amazing through this journey, and I want to thank you all. A special thank you to Susie and Ly for the endless support and encouragement you offer to our family. To Jax for always being able to cheer me up, for being my motivation, and yet, a wonderful distraction. To Zac, for forever being my companion in our adventures, the biggest thank you of all goes to you.

Through hard work and determination, it is always possible to achieve your goals in life, and together, we finally did it!

Definitions and Abbreviations

A	Adenine nucleobase of DNA
AIDS	Acquired immunodeficiency syndrome
ATP	Adenosine triphosphate
BA	Blood agar plate
BAPS	Bayesian Analysis of Population Structure
Bp	Basepairs
C	Cytosine nucleobase of DNA
CC	Clonal complex clusters as determined by goeBURST analysis
CI	Confidence intervals
CMH	Cochran-Mantel-Haenszel
Contig	Overlapping segments of DNA representative of a consensus region of DNA
Core genome	Genes present in all isolates within the group analysed
CPS	Capsular polysaccharides
DHF	Dihydrofolic acid
DHPS	Dihydropteroate synthase
DHR	Dihydrofolate reductase
DNA	Deoxyribonucleic acid
ERGO	University of Southampton Ethics and Research Governance Online
G	Guanine nucleobase of DNA
GC	Genomic control
GEMMA	Genome-wide efficient mixed model analysis for association studies
GFF	General feature format
GERMS	GIS efficient rapid microbial sequencing
GIS	Genome Institute of Singapore
GWAS	Genome wide association study

HIV	Human immunodeficiency virus
IgG	Immunoglobulin G
IPD	Invasive pneumococcal disease
Kb	Kilobase
KKH	KK Women’s and Children’s Hospital, Singapore
LD	Linkage disequilibrium
MDR	Multidrug resistant
MEGA	Macrolide efflux genetic assembly
MGE	Mobile genetic elements
MLSB phenotype	Resistance to macrolides, lincosamides and streptogramin B
MLST	Multi-locus sequence type
MOH	Ministry of Health, Singapore
MTB	*Mycobacterium tuberculosis*
N50	Shortest contig length necessary to cover 50% of the genome
NGS	Next generation sequencing
n-mers	DNA words of ‘n’ length
NPHL	National Public Health Laboratory, Singapore
NUH	National University Hospital, Singapore
OR	Odds ratio
PABA	Para-aminobenzoic acid
PBPs	Penicillin binding proteins
PC	Principal component
PCA	Principal component analysis
PCV	Pneumococcal conjugate vaccine
PCV7	PCV which contained purified capsular polysaccharide of seven disease-causing serotypes of *S. pneumoniae* (Prevenar$^{\text{TM}}$, Wyeth/Pfizer)

PCV10	PCV which contained purified capsular polysaccharide of 10 disease-causing serotypes of *S. pneumoniae* conjugated to a non-typable *Haemophilus influenzae* protein D (Synflorix^{TM}, GSK)
PCV13	PCV which contained purified capsular polysaccharide of 13 disease-causing serotypes of *S. pneumoniae* (Prevenar 13^{TM}, Pfizer)
PPV	Pneumococcal polysaccharide vaccine
PPV23	Pneumococcal polysaccharide vaccine containing 23 serotypes
QQ	quantile-quantile
Recombination	Bacterial DNA transfer from one organism (donor) to another organism (recipient)
r/m	Recombination/mutation
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SEER	Sequence element enrichment analysis
SGH	Singapore General Hospital
SNP	Single nucleotide polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SRST2	Short Read Sequence Typing for Bacterial Pathogens
SvI	Sensitive vs intermediate resistance classification
SvIR	Sensitive vs Intermediate resistance and resistant classification
SvR	Sensitive vs resistant classification
ST	Sequence types
T	Thymine nucleobase of DNA
THF	Tetrahydrofolic acid
TSA	Tryptic soy agar
TTSH	Tan Tock Seng Hospital, Singapore
USA	United States of America
WHO	World Health Organisation

Chapter 1 Introduction

Streptococcus pneumoniae is a major cause of disease and the main approaches to disease control are via vaccination and treatment of symptomatic people with antibiotics. Treatment with antibiotics is compromised by the spread of resistance. Mechanisms of resistance involve genetic mutations which modify enzymes or proteins involved in vital bacterial functions such as cell wall synthesis, metabolic performance and proliferation, and DNA replication. Mutations are acquired both through replication and repair of DNA generating vertical transmission, but also via horizontal passing of DNA as plasmids or through recombination. It therefore becomes important to understand the epidemiology of resistance by detailed analysis of the mutations associated with resistance in the microbial populations. This can be done by harnessing the power of genome wide association studies (GWAS). Once the kinetics and geography of trends in phenotypes and resistance are mapped, strategies can be developed to modify both vaccine and antibiotic usage in large geographic areas.

1.1 Genomic analysis of bacteria

Prokaryotic taxonomy has historically been classified by phenotypic characteristics alone, and the addition of genetics has only come into use since the 1960's. This began with the DNA-DNA hybridisation technique which measured genetic relatedness between bacteria (1). The development of the polymerise chain reaction and successful sequencing of the 16S rRNA gene which encodes the small subunit of ribosomal RNA in the 1980s transformed taxonomic classification. Following this, bacterial operational taxonomic units could be generated and compared with reference databases to infer taxonomy (2). This gene is relatively conserved therefore limits taxonomic determination at species level. A landmark in genomic analysis was the sequencing of the first bacterial genome in 1995. At this time bacterial genome sequencing was reserved to specialist centres as it was expensive and time consuming. The development of next generation sequencing (NGS) technologies introduced from 2005 allowed this approach to be more accessible and economic and therefore saw a rapid increase in the number of prokaryotic genomes being sequenced. There are now a number of NGS platforms developed including; Illumina/Solexa platform, SOLiD (sequencing by Oligonucleotide Ligation and Detection), pyrosequencing, Ion Torrent technology and single molecule real-time (SMRT) sequencing that have revolutionised DNA sequencing (3). This transformed the way in which microbiology was studied and has enabled researchers to accurately track the emergence and spread of pathogens as well as investigate virulence and antimicrobial resistance (4, 5).

1.2 Streptococcus pneumoniae

The diverse genus *Streptococcus* consists of 159 species (6), some of which are commensals and others pathogens in a human host. One of the species clinically relevant in human disease is *Streptococcus pneumoniae* (*S. pneumoniae*) which causes pneumococcal disease. It is transmitted through respiratory secretions and microscopically, is a Gram-positive diplococcus.

1.3 Pneumococcal disease

S. pneumoniae is one of many organisms transiently carried as normal flora in the upper respiratory tract of humans (7). Colonisation, or carriage as it is often referred, can be high in certain populations (8) and children are known to exhibit higher carriage rates than adults (9). Although a precursor for disease, carriage is usually symptomless. Infection from *S. pneumoniae* occurs following colonisation (10) and there are well-investigated factors that influence transition from carriage to invasion within the nasopharynx, such as the microbiome composition of the host (11), cigarette smoking (12) and asthma (13). A number of proteins present on the bacterial surface enhance colonisation and aid in avoiding the immune response of the host (14). Children under the age of five years, the elderly, as well as immunocompromised individuals, are at highest risk of pneumococcal infection (15, 16). These groups may not be able to effectively mediate antibody-initiated complement opsonisation to activate the classic complement pathway to protect against infection (8, 17). Progression to non-invasive pneumococcal disease (non-IPD) occurs by contiguous spread from the nasopharynx to cause otitis media, sinusitis, bronchitis, or conjunctivitis (8). In invasive pneumococcal disease (IPD) the organism is able to enter the bloodstream leading to bacteraemia and disseminate through the body to infect an otherwise sterile site such as the lungs (pneumonia) or meninges (meningitis) (18). Pneumonia is a leading cause of death, responsible for 1.3 million child deaths in 2013 (19) and unfortunately the pneumococcus is the leading causative organism for both pneumonia and bacteraemia. In cases of pneumococcal meningitis, case fatality rates are as high as 20% and unfavourable outcomes occur in 38-50% of cases (20). This makes pneumococcal disease a major public health concern globally.

1.4 Epidemiology

S. pneumoniae normally has a capsule composed of chains of complex polysaccharide subunits (13) which, depending on differences in the capsule operon, are differentiated into immunologically distinct serotypes. To date, approximately 100 circulating serotypes have been described (21), which can be further differentiated into 46 immunologically similar serogroups

(22). Large collections of pneumococcus isolates from around the world show there to be differences in the proportion of circulating serotypes and those associated with infection differ with time, geographical location, age, and clinical disease (13, 22-28). Incidence of disease is highest in resource poor settings such as Sub-Saharan Africa, Asia and Latin America (29). This could be due to a number ofl contributory factors such as climate and overcrowding. Since the introduction of capsular serotypes as vaccine antigens, there has been the need to monitor the serotypes internationally. The case carrier ratio of serotypes has been compared to identify some strains associated with higher invasive potential, for example serotypes; 1, 5, 7, 14, 18C, 19A and 6B (30, 31) and these serotypes are therefore commonly associated with IPD. Other serotypes, particularly those that lack the polysaccharide capsule, form normal commensal flora of the nasopharynx and rarely cause invasive disease in humans (32). These non-encapsulated forms however have been reported to be the most common cause of conjunctivitis (33).

1.5 Prevention and treatment

Pneumococcal infection is diagnosed by the presentation of clinical symptoms associated with disease and isolation of *S. pneumoniae* from the site of infection. Direct treatment against pneumococcal disease is with antimicrobials but, in order to control disease burden, prevention methods such as vaccination has also been implemented in many countries across the world.

1.5.1 Vaccination

It is well known that a major virulence determinant of *S. pneumoniae* is its anti-phagocytic polysaccharide capsule. The capsule locus (CPS) exhibits similar organisation in almost all strains in that the genes coding for the capsule are flanked by *dexB* and *aliA* and are transcribed by a single operon (34, 35). This capsule not only determines serotype but contributes significantly to invasive propensity (22, 32). Production of this capsule is essential for both virulence and colonisation, resulting in it becoming an early target antigen for vaccines. Vaccines were based on biochemical differences of the CPS from the most clinically relevant serotypes responsible for causing invasive disease. The pneumococcal capsular polysaccharide subunits are highly immunogenic (36) therefore the aim of vaccination is to induce polysaccharide specific IgG opsonophagocytic antibodies, required for bacterial clearance should infection occur (37, 38). Vaccines are of vital importance in reducing the worldwide burden of disease but are not available in all countries. The World Health Organisation (WHO) recommended the use of pneumococcal conjugate vaccine (PCV) in all countries, with the emphasis being on implementation into national immunization programmes of countries with mortality rates of >50/1000 live births, annual death rate in children of >50,000, and countries with high prevalence of HIV and sickle cell disease (39). Organisations such as The Vaccine Alliance (GAVI), the Advance Market Commitment funding mechanisms and other international donors are providing the resources necessary to make these vaccines available and accessible in low-income countries (40). GAVI has supported the rollout of PCV in more than 25 countries since 2010 and, encouragingly, more than 50 countries have been approved for GAVI support to introduce PCV into their national childhood immunisation program. Increased education to the general public on pneumococcal disease and the risks and benefits of vaccines would greatly help reduce the burden of disease and increase uptake of this vaccine (18).

1.5.1.1 Pneumococcal Polysaccharide Vaccine (PPV)

PPV was the first vaccine against pneumococcal disease in the USA, licenced in 1977 and consisted of purified capsular polysaccharide antigens from 14 pneumococcal serotypes (13). In 1983, this was replaced by PPV23 (Pneumovax23^{TM}, Merck), containing purified capsular polysaccharide antigens from 23 pneumococcal serotypes which at the time were responsible for 60-76% of invasive disease in the USA (13). In some countries, the USA for example, this vaccine is routinely given to immunocompetent adults over 65 years who also have cardiovascular or pulmonary disease, chronic liver disease, cirrhosis, alcoholism, diabetes, cochlear implant, leak of cerebral spinal fluid, or any other chronic illness (13). Since 2008 the Advisory Committee on Immunization Practices expanded this recommendation to include smokers and asthmatics. This vaccine generates a T-cell independent antibody response by the capsular polysaccharide directly stimulating an immune response in B-cells (37). It is not given to children <2 years as the T-cell receptor repertoires in the very young differs from that of older children and adults (41), and their immune system is too immature to respond to this vaccination (42). As no T-cell dependent immune response is induced by this vaccine, no memory B-cells are generated resulting in immunity of moderate duration, 5-10 years (43). In addition, PPV does not have the ability to extend protection benefits beyond the directly targeted population by interrupting pneumococcal carriage. This is referred to as ‘herd immunity’ where there is the indirect protection of unvaccinated individuals by decreasing overall circulation of infective agents within susceptible populations (44).

1.5.1.2 Pneumococcal Conjugate Vaccine (PCV)

The first PCV licenced in the USA since 2000 was PCV7 (Prevenar^{TM}, Wyeth/Pfizer) which contained purified capsular polysaccharide of seven disease-causing serotypes of *S. pneumoniae*. Each polysaccharide is individually conjugated to the non-toxic diphtheria CRM197 carrier protein, which limits the number of serotypes able to be included in the vaccine (45). Due to the presence of the carrier protein, PCV is highly immunogenic in all age groups and able to induce a protective antibody response, recruit the help of T-cells, and stimulate immunological memory and mucosal immunity (39, 46). The target population is children <2 years and the immunocompromised. Similar to the PPV vaccine, there have been later modifications to increase coverage of the vaccine by the inclusion of additional serotypes. The additional vaccines that became available included the 10-valent pneumococcal non-typable *Haemophilus influenzae* protein D vaccine (Synflorix^{TM}, GSK) in 2008 (47), and a 13-valent PCV (Prevenar 13^{TM}, Pfizer) in 2010 (48) providing coverage for 10 or 13 pneumococcal serotypes respectively.

1.5.1.3 Effect of vaccination on circulating pneumococcus

Vaccination has been successful at reducing pneumococcal colonisation of disease-causing serotypes in children (49), in adults (50), and at inducing herd immunity (51, 52). After implementation of PCV7 in the USA in 2000, the incidence of overall IPD declined by 45% and incidence of PCV7 serotypes declined by 94% (53). Reduction in PCV serotypes following vaccine implementation was also reported in Australia (54), Africa (55), Europe (56) and Asia (57), as well as significant change to the population of carriage isolates (58). The expansion of serotypes contained in the pneumococcal polysaccharide vaccine (PPV) and pneumococcal conjugate vaccine (PCV) was necessary to protect against important serotypes frequently isolated in areas of the world other than the United States, particularly Sub-Saharan Africa (59, 60) and Asia where 95% of all pneumococcal infections occur (29).The elimination of previously dominant serotypes by the vaccine allowed non-vaccine serotypes in circulation to expand and fill the now vacant niche to cause disease, a process known as serotype replacement. This increase in non-vaccine serotypes was identified in both carriage (61, 62) and invasive disease (63-65), and of particular concern is the non-PCV13 serotype 35B in which incidence of IPD is increasing (66). This phenomenon of serotype replacement or emergence of non-vaccine serotypes may limit the overall benefit of the vaccine, and this could be particularly evident in low-income countries where generally there is a broader spectrum of serotypes causing disease (40).

1.5.2 Antimicrobials

Primary choices for treatment are dependent on clinical disease and the guidelines of the prescribing country. A range of antimicrobials are available to treat infection and mechanisms of action include interference with cell wall synthesis, protein synthesis or nucleic acid synthesis (Figure 1). In respiratory disease macrolides are prescribed which have a broad spectrum of activity against common and atypical respiratory pathogens. Macrolides are generally well tolerated by the patient and easily administered (67). Beta-lactam antibiotics are used for non-IPD cases and chloramphenicol is often used for eye or ear infections. Serious IPD is treated with ceftriaxone and additional antibiotics such as vancomycin can be given in addition depending on the nature of infection. Alternative options in cases of disease from resistant strains may include imipenem, cefoperazone or cefotaxime (68).

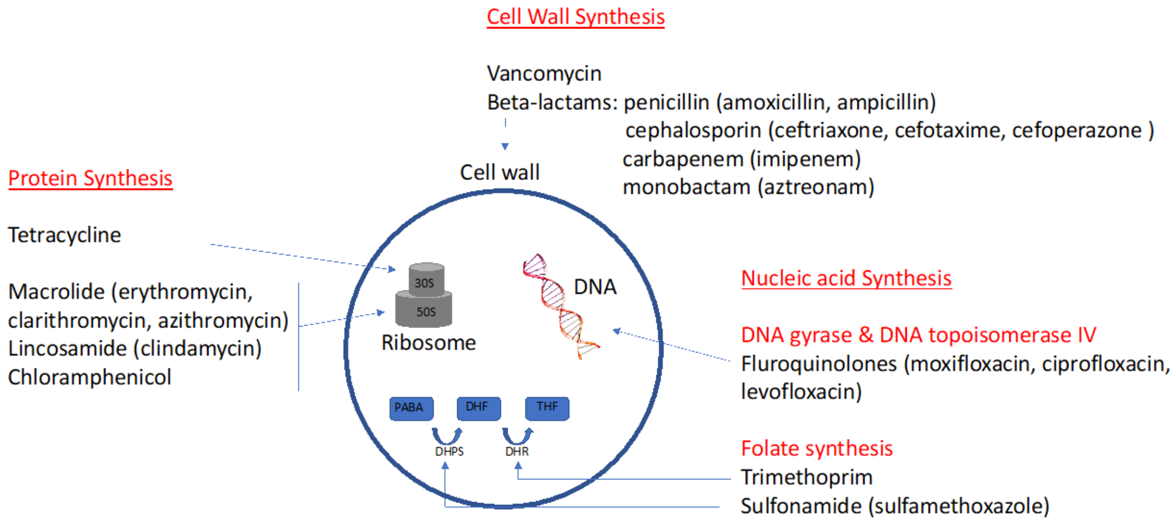

Cell Wall Synthesis
Vancomycin
Beta-lactams: penicillin (amoxicillin, ampicillin)
cephalosporin (ceftriaxone, cefotaxime, cefoperazone)
carbapenem (imipenem)
monobactam (aztreonam)
Cell wall
Protein Synthesis
Tetracycline
Macrolide (erythromycin, clarithromycin, azithromycin)
Lincosamide (clindamycin)
Chloramphenicol
30S
50S
Ribosome
DNA
Nucleic acid Synthesis
DNA gyrase & DNA topoisomerase IV
Fluroquinolones (moxifloxacin, ciprofloxacin, levofloxacin)
PABA
DHF
THF
DHPS
DHR
Folate synthesis
Trimethoprim
Sulfonamide (sulfamethoxazole)

Figure 1 Site and mechanism of action for antibiotics used to treat *S. pneumoniae* infection. PABA (para-aminobenzoic acid) DHF (Dihydrofolic acid) THF (tetrahydrofolic acid) DHPS (dihydropteroate synthase) DHR (dihydrofolate reductase)

1.6 Challenges arising from *S. pneumoniae* adaptation

From a population genomics standpoint, the pneumococcus has been well studied through large sample studies (69-72). This has provided in-depth knowledge on how the organism has been able to respond to selection pressures, such as host immune system, vaccine introduction and treatment with antimicrobials (73). Studies like this allow the monitoring of how these pressures affect epidemiology and evolution. Despite this, the pneumococcus continues to thrive and remains the greatest causative organism of pneumonia, the infectious disease which results in the highest number of deaths every year. Variations in bacterial genomes are generated through point mutation and genetic rearrangement, the introduction of phages and plasmids, and by recombination. Challenges that remain in preventing and treating pneumococcal disease are dependent on the location and extent of these changes made to the genome. Although many serotypes of *S. pneumoniae* associated with invasive disease are currently covered by PCV13, the organism has the potential to carry out gene switching events between isolates with enhanced antibiotic resistance and capsule polysaccharide genes that are not targeted by the vaccine creating challenging treatment options for the future.

1.6.1 Evasion of Pneumococcal Conjugate Vaccine

In the pneumococcal genome, the polysaccharide capsule locus (CPS) has a serotype specific region that encodes the capsule and is under strong selective pressure because of its high immunogenicity (74). As the CPS is flanked by conserved genes *dexB* and *aliA*, homologous recombination encompassing these flanking regions can lead to the replacement of vaccine targeted CPS in one genome with non-vaccine CPS from a different lineage, termed 'capsule switching' (75). Vaccines are, and will remain critical in controlling pneumococcal disease, however, they are complicated and expensive to manufacture (76). As vaccines only target a specific number of serotypes, this provides the potential for vaccine-escape variants and has been shown to occur in several populations (71, 77-80). This is particularly important when considering vaccine efficiency as the events result in invasive strains gaining a capsule that is no longer targeted by the vaccine. An aim of the present study was to assess the circulating serotypes in Singapore in relation to available vaccines. It had been previously identified that 30% of pneumococcal disease and 18.6% fatal cases in adults were infections by serotypes not covered by any of the currently available vaccine.

1.6.2 Antimicrobial resistance

The introduction of penicillin in the 1940s rapidly reduced the morbidity and mortality of pneumococcal disease as the organism was highly susceptible to the agent at this time (81). The first report of pneumococcal resistance to penicillin alone was described in 1967 in an adult patient in Australia (82). This was followed by the first multidrug resistant (MDR) strain, classified as showing resistance to 3 or more different antibiotics from a child in South Africa in 1977 (83). Since the 1970s, resistance to penicillin and other commonly used antibiotics has emerged and been reported in a large number of countries including Spain (84, 85), Hungary (86), South Africa (87), Iceland (88), France (89), United States (90, 91) and a number of Asian countries (40, 92). The availability and widespread use of antimicrobials are important in driving resistance in pneumococcus and resistance to a variety of classes of antimicrobials have now been described such as beta-lactams, chloramphenicol, tetracycline, cotrimoxazole (trimethoprim-sulfamethoxazole), clindamycin, and erythromycin (93-95). Many of these are first line therapy options which creates concern about future treatment. Among the various types of antimicrobial resistance seen, macrolide resistance in pneumococcal isolates can range from 10-90% in populations (57, 96, 97) and has emerged to have the most marked increase in resistance in many parts of the world.

Antibiotic resistant organisms threaten modern medicine and burden healthcare systems as they often result in longer hospitalisation , are costlier to treat and have higher morbidity and mortality. Patterns of resistance change temporally and geographically in accordance with selective pressures (98). Antimicrobial resistance has been described in 43 clones of *S. pneumoniae* which include both vaccine and non-vaccine serotypes (38, 79, 90, 93, 99-103). Resistance occurs in pneumococcus by the transfer of genetic material between cells and/or through mutations occurring in single cells. Acquisition of drug resistance can affect the viability of the bacteria and can contribute to overall fitness of the cell (104). Compensatory mutations are secondary mutations that may reverse this effect therefore it is necessary to consider the overall phenotype of resistance as an evolutionary process occurring as a sequence of steps (105).

1.6.2.1 Contribution of point mutation in generating resistance to antibiotics

Alterations in nucleotide base sequence can arise following exposure to toxin or ultraviolet light, or alternatively by random mutations made by DNA polymerase during DNA replication and in the DNA mismatch repair system. Random fluctuations in the number of gene variants present in a population over time occurs naturally via a process known as genetic drift (106). Generally, those that are favoured by Darwinian positive selection will be maintained within the population and each year an average of 2-4 mutations are introduced into the genome of *S. pneumoniae* through this process (69). A popular choice of therapy to treat non-IPD is using a macrolide, examples include erythromycin, azithromycin and clarithromycin. The mode of action for the macrolide class of antibiotic is to inhibit bacterial protein synthesis. It does this by binding to the bacterial 50S ribosomal subunit and causes dissociation of peptides from transfer RNA, disrupting bacterial protein elongation (107). In *S. pneumoniae* mutations in the ribosomal subunit 23S rRNA and in the ribosomal proteins L4 and L22 have been shown to confer resistance to macrolides (108-110). The mutation prevents effective macrolide binding to its ribosomal target site allowing protein synthesis to proceed.

Another popular choice of therapy is beta-lactam antibiotics. Within this class there are four major subgroups to which specific antimicrobials belong: penicillins, cephalosporins, monobactams and carbapenems (Figure 1). *S. pneumoniae* has six penicillin binding proteins (PBPs), a family of enzymes involved in peptidoglycan metabolism. These enzymes catalyse the transpeptidation reaction that cross links cell wall peptidoglycan and several are essential to cell survival (111). The mode of action for beta-lactam antibiotics is to covalently bind to the serine active site of PBPs and inactivate enzyme activity and hence, cross linking. Resistance to beta-lactams is mediated predominantly by variants in three of these PBPs; *pbp2x*, *pbp2b* and *pbp1a* where alteration of the antibiotic target site results in lower antibiotic binding affinities than in the native versions (112). Bacteria with these can still adequately perform peptidoglycan synthesis required for cell wall construction. Penicillin non-susceptibility in pneumococcus has been identified from point mutations that lead to amino acid substitutions within the transpeptidase domain of the PBPs (112). In particular, it is alterations in the amino acids around the serine residue that substantially reduce the affinity for penicillin (113). The primary determinant of resistance is through the acquisition of a series of stepwise mutations in the PBP genes that reduce the amount of drug binding (114-122). As each of the PBPs have differences in their affinity for individual beta-lactams, the relationship between which PBP contain the variants is relevant for the level of resistance. The acquisition of variants in *pbp2x* and *pbp2b* modulate low level resistance and, once acquired, are a pre-requisite first step in the acquisition of *pbp1a* variants which then lead to high level Β-lactam resistance (117, 123, 124). Penicillin resistance is primarily due to mutations in *pbp1a*, *pbp2b* and *pbp2x*, oxacillin resistance from mutations in *pbp2x* and *pbp2b*, and cephalosporin resistance from mutations in *pbp1a* and *pbp2x* (125).

Resistance to fluroquinolones can also occur following chromosomal mutations in the pneumococcus (126). The main targets for this class of antibiotic are the type II topoisomerase enzymes DNA topoisomerase IV and DNA gyrase (127). Both of these enzymes are considered essential for bacterial growth as together they regulate chromosome supercoiling and decatenation (128). Gyrase introduces negative supercoils into DNA and controls super-helical tension by relieving torsional stress before transcription and replication (127). Topoisomerase IV performs unlinking activity of interlinked chromosomes during replication, enabling the segregation of chromosomes to daughter cells at cell division (127, 128). There are slight differences in the enzymes targeted by the fluroquinolone antibiotics, for example moxifloxacin mainly targets DNA gyrase (129) whereas ciprofloxacin and levofloxacin target both but with a preference of topoisomerase IV (130). The effect is the formation of a complex between the drug, enzyme and DNA which is converted into a lethal double stranded DNA break by collision with replication forks, resulting in cell death (128, 131). Both gyrase and topoisomerase IV are composed of two subunits respectively; *gyrA*, *gyrB*, *parC*, *parE*. Mutations that confer resistance can occur in any of these proteins but predominantly they occur in *gyrA* and *parC* (132, 133). As was seen for beta-lactams, the development of resistance requires a combination of at least two mutations with the genes described located in the quinolone resistance determining region.

The antibiotic cotrimoxazole is composed of trimethoprim and sulfamethoxazole and can be routinely administered for the treatment of childhood respiratory infections (134). Consumption can drive resistance and high levels of resistance have been reported in several countries (135-137). Its additional use as a prophylactic antibiotic for patients with HIV/AIDS can contribute to this. Both components of this antibiotic act in different stages of the tetrahydrofolate synthesis pathway which reduces folate cofactors for the production of essential amino acid and nucleic acid metabolites such as purines, vitamins, amino acids and thymidylate (94, 138). The target for the sulphonamide is dihydropteroate synthase (DHPS) which is otherwise responsible for the conversion of para-aminobenzoic acid (PABA) and dihydropteroate diphosphate into dihydrofolic acid (DHF) (139, 140) (Figure 1). Sulphonamides competitively inhibit DHPS by acting as alternative substrates for the precursor PABA (138) and resistance has been attributed by both horizontal transfer of plasmid genes such as *sul*I and *sul*III (141, 142), as well as chromosomal mutations in the DHPS gene (*sul/folP*) (143). The role of trimethoprim is to block the reduction of DHF to tetrahydrofolic acid (THF) which is the final step in the tetrahydrofolate synthesis pathway (Figure 1). It does this by inhibiting the enzyme dihydrofolate reductase (DHR) and trimethoprim resistance has been identified through mutations that lead to amino acid substitution in the *folA* gene which encodes dihydrofolate reductase (94, 144).

1.6.2.2 Contribution of horizontal transfer of genetic material in generating resistance to antibiotics

Horizontal transfer of genetic material can occur from chromosomal homologous recombination or through the transfer of mobile genetic elements (MGE) such as insertion sequences, bacteriophages, plasmids, transposons and integrons (145). *S. pneumoniae* is naturally competent and autonomously mobile entities can be excised and integrated into genomes, either at a different position in the same cell, or within neighbouring cells. Resistance is largely driven by the expansion of clones that have acquired the resistant phenotype by horizontal genetic transfer (69) and the focus of this work addresses recombination of genetic material from another strain as the mode of horizontal transfer. In the pneumococcus, recombination occurs frequently in genes associated with antibiotic resistance for example *pbp1a*, *pbp2b*, *pbp2x*, *folA* (69, 146-148). Recombination events between *S. pneumoniae* and other species of the mitis group that co-inhabit the nasopharynx, for example *S. mitis*, *S.oralis*, *S. pseuodpneumoniae*, are common and have been shown to be a major contributor in dispersing resistant determinants among pneumococcal genotypes (146, 149, 150). The mosaic structure of the PBPs is evidence of the dynamic interchange of genetic material in these genes. Recombination of the genetic material encompassing resistance mutations or, of genes associated with resistance can rapidly disperse resistance in populations. It is possible that multiple genes can be acquired in a single step and studies have shown penicillin binding protein genes *pbp2x* and *pbp1a* crucial in determining beta-lactam susceptibility, have been acquired alongside a new capsule as they closely flank the CPS locus (151, 152). Recombination events such as these can have drastic effects as they simultaneously lead to antimicrobial non-susceptibility and vaccine escape making current treatment options no longer viable (153).

The erythromycin ribosomal methylase (*erm*) family of genes are resistance genes that are frequently transferred. In *S. pneumoniae* the *ermB* gene is the most common which provides high level resistance to macrolides, but there are others that are rarely found; *ermA* and *ermTR* (154, 155). Although there have been reports of this gene only resulting in resistance to macrolides (M phenotype) (156), its presence usually results in a resistance phenotype characterised as the MLSB phenotype as it confers resistance not only to macrolides, but also lincosamides and streptogramin B by its ribosomal methylation activity (157, 158). The adenine specific N-methyltransferase encoded by *ermB* dimethylates the target site of the 23S rRNA which prevents antibiotic binding (157). Another common mechanism of macrolide resistance in *S. pneumoniae* is through antibiotic efflux. For this, acquisition of the genes *mefE* (also known as *mefA* due to 90% sequence similarity with *mefA* in S. pyogenes) and *mel* are required which are carried on the macrolide efflux genetic assembly (MEGA) element (159). The MEGA element can be widely disseminated within populations through horizontal DNA exchange. There have been reports of *mefE* homologs such as *mefI* being identified in *S. pneumoniae,* but this is rare (160). The two genes operate as a two-component efflux pump that synergistically provides resistance. The gene *mefE* encodes a protein belonging to the major facilitator superfamily which expels molecules from cell by the utilisation of proton motive force, rather than ATP, as part of the pump mechanism (156). The gene *mel* encodes an ATP-binding cassette transporter protein that is predicted to interact with chromosomally encoded transmembrane complexes as the protein alone lacks typical hydrophobic membrane-binding domains (159). It may be that *mel* is able to displace ribosome bound macrolide and transfer molecules to *mefE* for efflux out of the cell (107). Both the methylation and efflux mechanisms of resistance are inducible by the presence of macrolides (161, 162).

Conjugative transposons such as Tn916/Tn1545 commonly contain the gene *tetM* which confers resistance to tetracycline. These are distributed through horizontal transfer in pneumococcus and other Gram-positive organisms. Tetracycline resistance through the acquisition of *tetM*, or occasionally *tetO*, occurs as these genes encode proteins that offer ribosomal protection and GTPase activity that aids in the displacement of antibiotic from the bacterial ribosome (163). The conjugative transposons may also incorporate additional resistance determinants such as *ermB* and/or *mefE* as previously discussed to form a number of larger Tn916-like composite elements (164). Tn916 commonly inserts into a Tn5252 mobile element forming a composite TN5253 -like element (165). The Tn5252 can carry the *cat* gene which, if acquired, produces the enzyme chloramphenicol acetyltransferase resulting in resistance to chloramphenicol by acetylation of the antibiotic preventing its binding to the bacterial ribosome (165).

1.7 Identification of resistance mechanisms in pneumococcal populations

In order to control disease, there is a need to be able to identify ways in which the organism has responded to interventions and use this information to anticipate future changes in circulating pneumococcal populations. It is known that the continued use of conjugate vaccines is likely to lead to on-going successive rounds of serotype replacement unless a vaccine that contains all circulating serotypes is created. In addition, it has been observed that high levels of genomic plasticity within the species means virulence and resistance determinants can be easily widespread throughout populations. Incidence of pneumococcal diseases across the world are still alarming and highlight that current methods of vaccine and antibiotic treatment alone are not a sustainable way to control infection. Horizontal transfer plays an important role in the acquisition of antibiotic non-susceptibility genes. Advances in technology facilitate the potential to look directly for these known genes and make subsequent changes to treatment based on this additional information.

Significant health threats in the future are primarily from strains that are not covered by current vaccines, but which carry novel resistance determinants that result in treatment failure. The importance of mutation has been identified in the adaptation and evolution of lineages which may be slower and subtler than those seen from horizontal transfer. Research into non-susceptibility testing has traditionally been carried out using techniques such as laboratory mutagenesis and genome sequence analysis (124, 166-168), however resolution of these methods is limited. Investigation using whole genome sequencing data may be able to identify novel resistance determinants other than those attributed to recombination which could influence prescribing policies or identify novel target treatments for the future. Resistance phenotypes may emerge through single or multiple mechanisms and are likely to be unique depending on pressures experienced at different geographical locations. This is the area yet to be investigated in many bacterial species and is extremely important considering how quickly this organism is able to respond to selection pressures.

A new and more comprehensive way to identify novel alleles associated with resistance from populations is through a genome wide association study. This approach can assess the entire genome to identify the genetic basis for bacterial traits in an unbiased manner, without prior characterisation of candidate genes. They offer the potential to build a detailed understanding of causative variations attributed to phenotypes of interest and identify evolutionary responses to dynamic environmental conditions. Association studies have been frontiers in the identification of causal variants relating to a broad range of conditions for many years, for example conditions in human health including height (169), diabetes (170) and inflammatory bowel disease (171); host factors contributing to infectious diseases (172, 173); and in agriculture (174). This approach has now been applied to bacterial datasets and has yielded some interesting and biologically relevant findings relating to virulence and antimicrobial resistance (175-181).

1.8 Hypothesis

S. pneumoniae continues to be a major cause of death worldwide even though there are both vaccines and antimicrobial therapies available for use. A major contributor to treatment failure is increasing levels of resistance to commonly used antimicrobials. This organism has the exceptional ability to diversify its genome by the uptake of DNA, and therefore, the exchange of genes continues to be a challenge for current and future treatment regimes. Many of the genes contributing to antimicrobial resistance are transferred by recombination in this highly competent organism, however, it is hypothesised that areas of the genome have also evolved over time that contribute to antimicrobial resistance. It could be that these mutations result in novel or secondary mechanisms of resistance but are overlooked as they may not contribute high levels of resistance as, for example, the *erm(B)* gene would. In addition, if a trend towards antimicrobial resistance emerges because of a single, or the accumulation of mutations, this may influence antimicrobial prescribing policies for the future. The hypothesis is that whole genome sequencing from a large dataset of clinically relevant *S. pneumoniae* can be used in a genome wide association study to describe the variations involved in generating resistance. Identification of these mutations would help to gain a more detailed understanding of antimicrobial resistance mechanisms and will provide insight into modification of the genome in response to antimicrobial pressure over time. This would help develop clear strategies for antibiotic regimes to stay ahead in disease management.

Specific hypotheses:

- There will be changes over time in epidemiological data regarding disease causing serotypes in Singapore.
- There will be changes over time in the resistance profiles of disease-causing isolates in Singapore.
- GWAS of these isolates will reveal significant associations between antibiotic resistance and single nucleotide polymorphisms that could contribute to antibiotic resistance to clinically used antibiotics.

1.9 Aims and Objectives

Antimicrobial resistance has been emerging worldwide in *S. pneumoniae* and in many cases has resulted in treatment failures. The organism continues to evolve in response to pressures such as vaccines and antimicrobials, and the transfer of DNA between pneumococcal strains and other bacteria facilitates this. Resistance is more prominent in some geographical locations such as Asia. The aim of this project is to identify mutations in disease-causing strains of *S. pneumoniae* that are significantly associated with non-susceptibility to antimicrobials.

1. Identify temporal changes in disease-causing *S. pneumoniae* isolates from Singapore using whole genome sequence data:
    a. Identify serotypes associated with invasive and non-invasive pneumococcal disease.
    b. Highlight changes in serotypes frequency responsible for disease over time.
    c. Assess disease causing serotypes in relation to available vaccines in Singapore.
    d. Identify current levels of antimicrobial resistance and highlight changes that may have occurred over time.
    e. Identify serotypes that are most associated with resistance.

2. Prepare genome data from isolates to accurately account for population structure based on vertical inheritance:
    a. Identify recombination in isolate genomes and determine the influence of this on population structure.

3. Through a GWAS approach mutations in the genome will be identified to highlight:
    a. Possible association with mutation and non-susceptibility to clinically relevant antimicrobials; penicillin, erythromycin, clindamycin, cotrimoxazole, doxycycline, tetracycline and chloramphenicol.

Chapter 2 Methods

2.1 Singapore Pneumococcal dataset

2.1.1 Participating hospitals

The participating hospitals are four of the largest in Singapore; Tan Tock Seng (TTSH) a multi-disciplinary 1,700 bed hospital, National University Hospital (NUH) a 1,239 bed major referral centre and tertiary hospital, Singapore General Hospital (SGH) the largest in Singapore with 1,785 beds and KK Women's and Children's Hospital (KKH) with 830 beds specialising in high risk conditions in women and children.

2.1.2 Ethics

Ethical approval was obtained from Singapore Centralised Institutional Review Board (Ref:2012/614/E) and National Healthcare Group Domain Specific Review Board (Ref: 2012/00339) to collect pneumococcal isolates and, when possible, anonymous corresponding laboratory data described in Appendix Q from the named hospitals from 1970 – March 2016. Material transfer agreements were approved between Genome Institute of Singapore and all collaborating hospitals to transfer pneumococcal isolates and anonymous corresponding clinical data. University of Southampton ERGO approval was granted August 2016. All research was performed in accordance with relevant guidelines and regulations.

2.1.3 Study isolates

Short read Illumina sequencing data and anonymous corresponding laboratory data was available for 1,761 *S. pneumoniae* isolates as described previously (182). An additional 336 *S. pneumoniae* isolates stored as per routine diagnostic practice between July 2013-March 2016 were transferred from National University Hospital (n=68), Tan Tock Seng Hospital (n=135), Singapore General Hospital (n=50) and KK Women's and Children's Hospital (n=83) with anonymous corresponding laboratory data.

2.1.4 Laboratory methods

2.1.4.1 Bacterial cultures

Samples were transported to the Genome Institute Singapore either on 5% blood TSA agar (BD, New Jersey, USA) or in frozen microbank (Pro-Lab Diagnostics, Canada) vials. Upon receipt, *S. pneumoniae* identification was confirmed by characteristic growth on 5% sheep blood Columbia blood agar (BD, New Jersey, USA) and sensitivity to optochin (Thermo Fisher Scientific, Massachusetts, USA) following overnight incubation at 37°C, 5% CO_2. Ten isolates did not grow on receipt or were contaminated and therefore were not used in the study.

2.1.4.2 DNA extraction and sequencing library preparation

Whole genome sequencing was carried out on 326 *S. pneumoniae* isolates. Each isolate was streaked to single colonies on Columbia blood agar agar (BD, New Jersey, USA). A single colony was inoculated into a Todd Hewitt broth (Thermo Fisher Scientific, Massachusetts, USA) and cultured overnight at 37°C, 5% CO_2. After vortexing, cells in 250µl of culture was transferred to a 96x deep well plate and genomic DNA extracted using the Wizard SV 96 Genomic DNA Purification Kit (Promega, Wisconsin, USA). Quantification of DNA was performed using QUBIT 2.0 fluorometer (Invitrogen,USA). Whole genome sequencing libraries were made from 1µl of genomic DNA using the Nextera XT Index Kit v2-dual 8bp (Illumina, California, USA). Finally, 10nM of the sequencing library for each sample was pooled together and sequencing performed by on a Hiseq 4000 (Illumina, California, USA) with a 2 x 151 run by the core sequencing team at the Genome Institute of Singapore.

2.1.5 Genome sequence analysis

All primary sequencing analysis was performed using the Efficient Rapid Microbial Sequencing (GERMS) platform at Genome Institute Singapore (183). Raw FASTQ files for 1,761 *S. pneumoniae* isolates generated in (182) and the 326 generated in section 2.1.4.2 were mapped onto the complete reference genome *S. pneumoniae* ATCC 700669 (European Molecular Laboratory accession FM211187.1 (145) using Burrows-wheeler Aligner (version 0.7.10) (184). Single nucleotide polymorphism calling, and insertion and deletion realignment performed from whole genome data using LoFreq (version 2.1.2) (185) with default parameters. Genomes of optimised kmer length were assembled using VelvetOptimiser (version 1.2.10) (186) and annotated by PROKKA (version 1.13.3) (187).

FASTQ reads, the 1,761 *S. pneumoniae* isolates generated in (182) and the 326 generated in section 2.1.4.2 were used to assign Multi Locus Sequence Types by the Short Read Sequence Typing for Bacterial Pathogens (SRST2) program (188) (version 0.2.0) with default settings. MLST data generated by SRST2 was used in goeBURST (PHYLOVIZ) (version 2.0) (189, 190) to determine allocation of isolates into clonal complexes (CC). The eBURST distance criteria was set as single locus variants which required all sequence types (ST) within the group to share identical alleles at six or seven MLST loci with at least one other ST in the group. FASTQ reads were used to determine serotype using the PneumoCaT (version 1.2.1) (191) tool with default settings.

The odds ratio (OR) calculated in Table 6 estimated the probability of IPD for specific serotypes with reference to all other serotypes in the dataset using the equation (192):

$$OR = \frac{ad}{bc}$$

Where:

a = the number of IPD isolates for the given serotype

b = the number of non-IPD isolates for the given serotype

c = the total number of IPD isolates other than designated serotype

d = the total number of non-IPD isolates other than the designated serotype.

The OR calculation was used to calculate OR described in Table 7. OR statistic and 95% confidence intervals were calculated using the online tool (193), where:

a = the number of specified serotype causing disease in the tested age range

b = the number of specified serotype causing disease in ages other than that tested

c = the number of non-specified serotype causing disease in the tested age range

d = the number of non-specified serotypes causing disease in ages other than that tested.

2.1.6 Assembly statistics and sample inclusion criteria

A quality assessment of assemblies was performed to exclude poor data.

Exclusion criteria was:

- Species identification of <50% by Kraken (version 1.0) (194) underwent a secondary speciation using KmerFinder (version 3.1) (195-197). Confirmation of species other than *S. pneumoniae* resulted in removal from dataset (n=12);
- Minimum assembled genome length of <1.5 Mbp (n=15);
- Total number of contigs >1000 (n=1).

The assembly pipeline for the remaining 2,059 isolates gave on average a total length of 2,055,333bp from 19 - 830 contigs with average number of contigs of 95bp and average N50 of 67,742bp.

A large spread in the number of contigs was identified in assemblies that otherwise did not fulfil the exclusion criteria. To assess whether this was an artefact of VelvetOptimiser (version 1.2.10) (186), or a consequence of sequence data, assemblies from FASTQ files were repeated for a subset of genomes using SPAdes (version 3.15.2) (198, 199) and QUAST (version 5.0.2) (200) used to generate summary statistics for each assembly (Table 1). Assemblies from isolates reported to have very high numbers of contigs by VelvetOptimiser was considerably reduced following the application of SPAdes, particularly for isolate WBB1351 which identified a reduction in the number of contigs from 830 to 241. SPAdes identified a slight increase in contig numbers from assemblies with <30 contigs identified by VelvetOptimiser. This is suggestive that the extremes in range for the number of contigs identified might be an artefact of using the program VelvetOptimiser, which is now considered to be a relatively outdated program, and that the identification of higher numbers of contigs is not an indicator of poor sequence data. Additional factors such as genome size could have affected total number of contigs reported by individual assemblers however, these were similar for both programs. Confirmation of high-quality assemblies from a subset of isolates by SPAdes gives confidence to the sequence data used in downstream analysis.

Table 1 Number of contigs and genome size identified by assemblers VelvetOptimiser and SPAdes

Isolate ID	Contigs identified by VelvetOptimiser	Genome size identified by VelvetOptimiser (bp)	Contigs identified by SPAdes	Genome size identified by SPAdes (bp)
WBB1351	830	1937602	241	2121989
WBB2174	825	2045192	537	1902803
WBB1332	717	1983951	478	1987381
WBB2216	705	2000891	404	2098029
WBB1344	650	1901693	346	2112107
WBB1336	617	1955806	339	1983151
WBB2230	614	1995132	339	1959602
WBB2223	571	1913743	254	2072265
WBB2150	562	1979208	275	2110427
WBB2159	542	1908769	206	2015156
WBB1293	413	1959379	201	2158467
WBB2158	413	2029710	188	2013274
WBB2144	408	1924426	223	2076742
WBB2145	407	1990149	227	2019699
WBB2195	401	2011793	189	2110378
WSB379	264	2124570	149	2050963
WBB2351	227	2068311	124	2079992
WBB1289	215	1754535	149	2124834
WBB2045	209	1923379	116	2058769
WBB2110	201	1920768	145	1958180
WSB1078	28	2024486	87	1993525
WSB2015	28	2140680	77	1980731
WSB1326	27	1954184	68	2057323
WSB1187	25	2105065	91	1976063
WSB1210	25	2045625	68	2052684
WSB1114	25	2071088	66	1958383
WSB1302	23	1998607	70	2026294
WSB1335	23	2069465	68	2041906
WSB1212	23	2053151	65	1978930
WSB1329	19	2027748	112	2022467

2.1.7 Classification of metadata

When classifying the metadata the ‘other’ category in ethnicity, combining individual groups of smaller numbers, included Bangladesh (n=2), Burmese (n=1), Caucasian (n=7), Eurasian (n=3), Filipino (n=1), Indonesian (n=2), Korean (n=1), Nepalese (n=1), Sikh (n=1) and other (n=98). When classifying site of infection into disease types, invasive disease included infection of ascitic fluid (n=1), blood (n=989), bone (n=2), brain (n=1), cerebral spinal fluid (n=19), joint fluid (n=2), knee (n=7), lung (n=2), ovary (n=1), pelvis (n=1), pericardium (n=1), pleura (n=63), submental space (n=1). Non-invasive disease included the following sites of infection; arm (n=2), bronchoalveolar lavage (n=14), back (n=1), bronchus (n=2), chest (n=2), eye (n=36), eyelid (n=1), ear (n=77), endotracheal tube (n=82), finger (n=1), hand (n=1), nasopharynx (n=3), neck (n=2), nipple (n=1), nose (n=35), peritoneum (n=5), portacath (n=1), sputum (n=565), throat (n=2), urine (n=2), vagina (n=11).

2.2 Analysis of recombination programs

Figure 2 summarises the processes undertaken in the analysis of recombination programs.

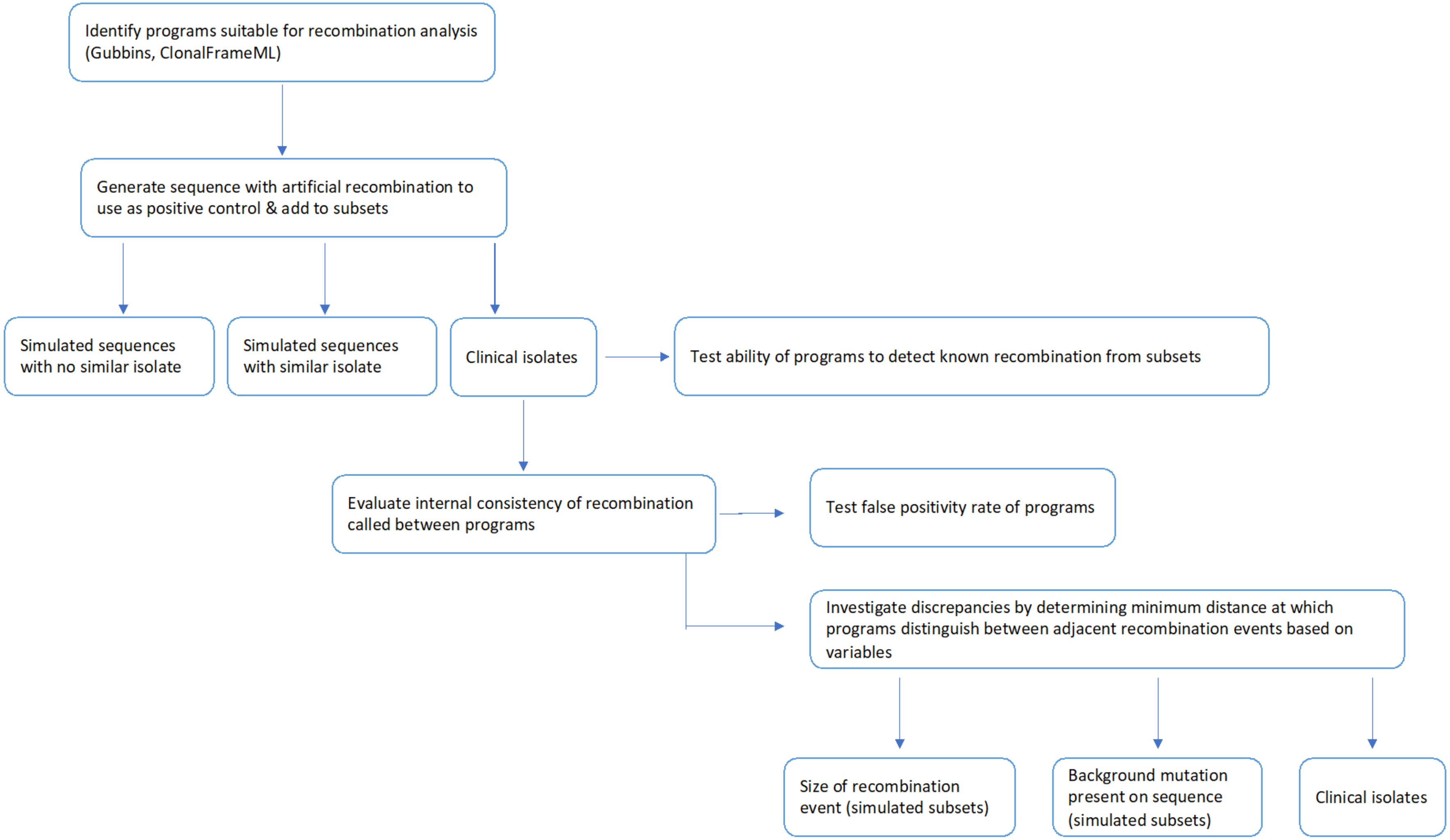

Identify programs suitable for recombination analysis (Gubbins, ClonalFrameML)
Generate sequence with artificial recombination to use as positive control & add to subsets
Simulated sequences with no similar isolate
Simulated sequences with similar isolate
Clinical isolates
Test ability of programs to detect known recombination from subsets
Evaluate internal consistency of recombination called between programs
Test false positivity rate of programs
Investigate discrepancies by determining minimum distance at which programs distinguish between adjacent recombination events based on variables
Size of recombination event (simulated subsets)
Background mutation present on sequence (simulated subsets)
Clinical isolates

Figure 2 Flow diagram summarising the workflow involved in the recombination analysis of Gubbins and ClonalFrameML

2.2.1 Generation of sequences containing artificial recombination events and/or diversity

The areas of true recombination present in the clinical isolates were unknown. Therefore, a simulated sequence was generated containing a known area of recombination to test detection by recombination programs. Previous research into the density of SNPs in recombination events of *S. pneumoniae* have shown a range of 1:150bp (*S .pneumoniae* JJA) – 1:81bp(*S. pneumoniae* Hungary 19A-6) (201). Croucher *et al* (2014) showed each recombination event imported a mean of 104 base substitutions and had a mean length of recombination to be 8.8kb (202),making the density approximately ~1:85bp. Based on this it was decided to create a simulated recombination event with a similar rate of mutation; 10kb genomic region with 100 base substitutions spaced 1:100bp.

The genotype used for the simulation of recombination within the sequence, referred to as 'wildtype' was a clinical isolate with the accession code WSB1573. Single point mutations were uniformly introduced into the whole genome alignment of wildtype at a rate of 1:100bp from positions 101 – 2,221,301 (A=>C, T=>A, C=>G, G=>T) to produce the aligned 'donor sequence' in which recombination will be transferred from. An exchange of corresponding nucleotides between donor and wildtype sequence produced a sequence with known recombination coordinates referred to as 'recombination sequence' (Appendix A).

An exchange of 10,000bp with no additional change to the wildtype sequence tested the performance of recombination programs to detect the simulated recombination when run within a subset of randomly selected clinical isolates. An additional five simulated wildtype sequences were created that incorporated background mutation rates reflecting diversity within the isolates; 0.001%, 0.01%, 0.1%, 1% and 2%. To do this single point mutations were uniformly introduced into wildtype sequence from bp position 102 onwards, so that further mutation did not interfere with the transfer of donor recombination. For 0.01% mutation rate, 222 SNPs are introduced at a ratio of 1:10,000; for 0.1% mutation 2,221 SNPs were introduced at a ratio of 1:1,000; for 1% 22,213 SNPs were introduced at a ratio of 1:100 and for a 2% mutation rate 44,426 SNPs introduced at a ratio 1:50bp. The 10kb recombination region with 100 base substitutions spaced 1:100 was transferred to these sequences and recombination analysis performed on datasets.

Further recombination sequences that contained a second simulated recombination at varying distances upstream were constructed to test the distance required by programs to identify individual recombination events. The position of the first recombination remained constant and the distance to the second ranged from 100bp to 10,000bp, at intervals of 100bp (Appendix F).

Initially sequences were constructed, varying the size of simulated recombination events from 1,000 to 10,000bp at 1000bp intervals to ensure recombination size did not affect the distance required by programs to identify them individually. The dataset in which this was run is shown in Appendix H. After this initial experiment, a series of recombination sequences with two 10kb recombination events of varying distance were constructed with an incorporated background mutation rate ranging from 0% to 2% as previously described (Appendix G).

2.2.2 Datasets for recombination analysis using simulated sequence with a single recombination

Recombination analysis was performed on 100 subsets containing 10, 20 or 30 whole genome alignments of randomly selected clinical isolates and the recombination sequence. These were selected independently of one another for each analysis. Computational resources allocated to run programs included four threads, 16G memory and run time of 100 hours. Detection of simulated recombination for completed runs and run time was documented.

A basic dataset to test the effect of diversity on the detection of recombination was constructed (Appendix C). Five of these datasets were processed, each having a different background mutation rate described in section 2.2.1, and the detection of simulated recombination documented. Next, a dataset designed to test the effect of diversity on the detection of recombination, but which also included a sequence similar to the recombination sequence was constructed (Appendix D) and tested for each mutation background. The difference in the mutation rate between the chromosome and area of recombination within these datasets was shown in Appendix D. To test if recombination was detected as an area of no mutation in a chromosome containing high levels of mutation, a final dataset was constructed (Appendix E). This changed the mutation relationship of segments on the chromosome (Appendix E). To test the effect of background mutation on recombination detection from clinical isolates, 100 subsets containing 10 clinical isolates and the recombination sequence with each of the five mutation backgrounds was created and recombination analysis performed.

2.2.3 Datasets for recombination analysis using simulated sequence with two recombination events

Sequences containing two artificially introduced recombination events of varying distances apart (described in Section 2.2.1) were added to 100 subsets of 10 randomly selected clinical isolates. This was performed for each mutation background and recombination analysis performed with a

100 hour cut-off processing window. The distance at which programs can identify individual recombination events was recorded for each dataset.

2.2.4 Calculating divergence value

A dissimilarity matrix containing whole genomes of 2,059 *S. pneumoniae* isolates and the recombination sequence (section 2.2.1) was constructed using SNPRelate (203) in R (version 3.4.1) (204). The divergence value for a subset was determined from the dissimilarity matrix by calculating the pairwise distance between each isolate in the subset and the recombination sequence. The minimum divergence score was the smallest of these values.

2.2.5 Recombination analysis

An aligned genome sequence for each isolate included in recombination analysis was reconstructed as described in section 2.1.5. Gubbins (version 2.0.0) was used to call recombination on the set of aligned genomes using default settings (205). For ClonalFrameML, FastTree (version 2.1.10) was initially used to generate a maximum likelihood phylogenetic tree from the aligned genomes with the –gtr, and –nt command line options (206, 207), and then ClonalFrameML (version 1.0), used to call recombination using default parameters (208). The performance of individual programs was determined by successful detection of simulated recombination coordinates in output files. To compare total recombination between programs, the union of all combined segments called on the recombination sequence by individual programs was taken.

2.2.6 False positivity rate of programs

A major contributor of the signal for recombination is a high density of SNPs clustered in the genome relative to the level of background SNPs. It is these clusters that need to be removed in order to identify the false positivity rate of individual programs. This was achieved for each of the clinical isolates by scrambling the genome of all isolates so that each bp is randomly assigned a new position. The new positions of the bp remained constant for all isolates which maintained allele frequency in the population but broke up clusters of SNPs.

2.3 Background diversity in clinical isolates

Diversity between clinical isolates (n=2,059) and the reference strain *S. pneumoniae* ATCC 700669 (European Molecular Laboratory accession FM211187.1 (145) was calculated by counting the number of SNPs present between the two sequences. Based on the number of SNPs, the percentage of diversity can be calculated by ((number of SNPs/total number of bp)x100) for each clinical isolate.

2.4 Construction of phylogenetic trees

FastTree (version 2.1.10) was used to generate a maximum likelihood phylogenetic tree from concatenated whole genome alignments of the 2,059 *S. pneumoniae* isolates (generated in section 2.1.5) with the –gtr, –nt and –boot 100 command line options (206, 207), Figure 9. FastTree (version 2.1.10) was used to generate a maximum likelihood phylogenetic tree from concatenated whole genome alignments of the 2,059 *S. pneumoniae* isolates (generated in section 2.1.5) and the alignments of five additional wildtype sequences with mutation backgrounds of 0.001%, 0.01%, 0.1%, 1% and 2% incorporated. The –gtr, –nt and –boot 100 command line options (206, 207) were used, Figure 3. FastTree (version 2.1.10) was used to generate a maximum likelihood phylogenetic tree from concatenated genome alignments that had all areas of recombination identified by ClonalFrameML removed (as described in Section 2.5). The –gtr, –nt and –boot100 command line options (206, 207) were used, Figure 19. FastTree (version 2.1.10) was used to generate a maximum likelihood phylogenetic tree from core gene alignments (generated in section 2.8) with the –gtr, –nt and –boot100 command line options (206, 207), Figure 20. All phylogenetic trees were visualised using Interactive Tree of Life (209). A quantitative assessment of similarity between phylogenies of whole genome and recombination free genome, and between whole genome and core gene phylogenies was performed using the Environment for Tree Exploration toolkit (version 3.0) (210) which calculated the Robinson Foulds metric (Table 11).

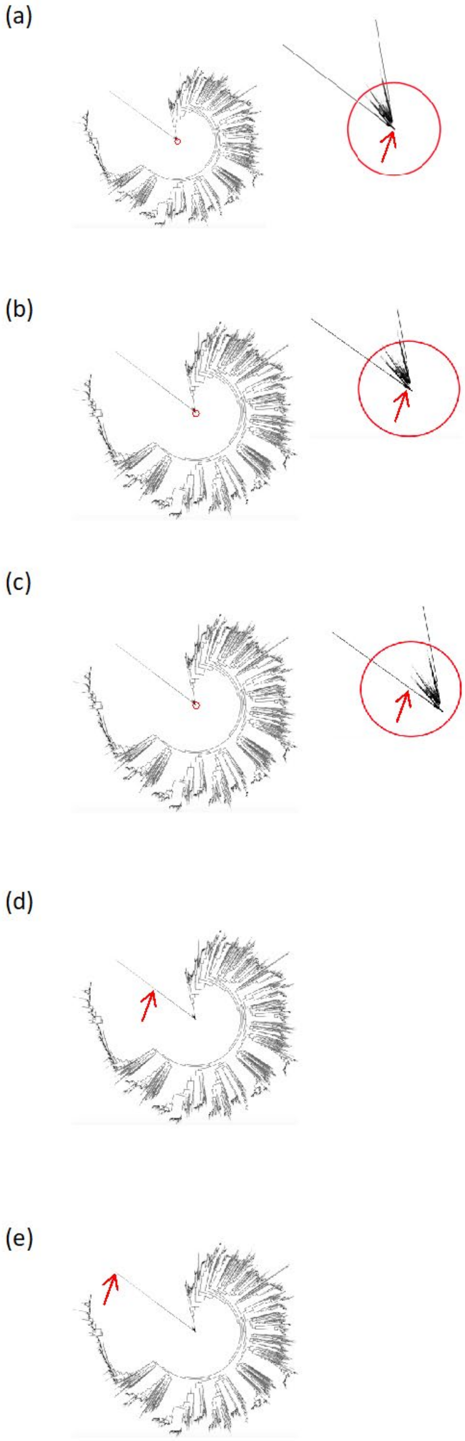


Figure 3 Maximum-likelihood phylogenetic trees constructed by FastTree using whole genomes of 2,059 isolates of *S. pneumoniae* and 5 artificial sequences with varying levels of background mutation incorporated (not including known recombination region). The red arrows show the position of sequences with incorporated background mutation; (a) 0.001% mutation; (b) 0.01% mutation; (c) 0.1% mutation; (d) 1% mutation; (e) 2% mutation.

2.5 Recombination analysis on Singapore dataset of 2,059 *S pneumoniae* isolates

A total of 400 datasets consisting of 100 randomly selected isolates were generated and recombination analysis using ClonalFrameML was performed (as described in Section 2.2.5). Recombination coordinates from individual datasets were merged using Bedtools merge (Version 2.29.2) (211). The bp in these recombinogenic areas of the aligned genome were replaced with 'N' using Bedtools maskfasta (Version 2.29.2) (211) to create recombination-free genomes.

2.6 Comparing population structure of 1,828 *S. pneumoniae* isolates.

Isolates that did not have antimicrobial susceptibility data for at least one antibiotic were removed from the dataset intended for GWAS analysis leaving 1,828 isolates. PopPUNK (version 1.2.2) (212) was used to determine the population structure from sequence assemblies of whole genomes generated in section 2.1.5 and assemblies of recombination-free genomes generated in section 2.5 using the --easy-run function. Quality checks showed the '--min-k' was appropriately set as no random probabilities were greater than 0.05 and there was a good network score of 0.9.

To perform the PCA, a binary matrix was constructed for all 1,828 isolates. For this, bases were called from mapped sequences as described in section 2.1.5, each base was denoted by a single numeric value, 1 or 0 to designate the presence or absence of a SNP respectively when compared to the reference sequence. Areas of recombination removed in the recombination-free dataset were denoted by NA. Distance data was calculated and PCA plots were constructed using cmdscale and ggplot2 in R (version 3.5.2). Corresponding scree plots showing the cumulative variation for the first 10 principal components (PC) was generated using barplot in R (version 3.5.2).

2.7 Genome wide association studies for an antimicrobial resistant phenotype.

Phenotypes of antimicrobial susceptibility were determined in the microbiology laboratories for 1828/2059 pneumococcus isolates during routine clinical care and were described as sensitive (S), intermediately resistant (I) or resistant (R). The disk diffusion method was used by the hospital laboratories to determine antibiotic susceptibility (182). Here, a disk impregnated with antibiotic is placed on an inoculated agar plate. After overnight incubation, the radius of the zone of inhibition is determined to classify susceptibility status. Confirmation of resistant isolates was performed with Etest which uses predefined gradients of antimicrobials on a test strip to determine the minimum inhibitory concentration of antibiotic for that isolate. A sensitive phenotype means bacterial growth is inhibited by the antibiotic, and a resistant phenotype would result in bacterial growth in the presence of antibiotic. Intermediate resistance would result in some, but not all inhibition of growth. The intermediate resistance profile from GWAS analysis was excluded for all but penicillin as numbers were small. There were suitable numbers of cases and controls to perform a GWAS on penicillin, cotrimoxazole, erythromycin and chloramphenicol, clindamycin, doxycycline and tetracycline. A kinship matrix was initially created in GEMMA (genome-wide efficient mixed model analysis for association studies) (version 0.98.1) (213) which was used to account for population structure in the following test of association. A minor allele frequency cut-off of 0.01 was used for all analyses and reported SNPs with a p value <0.01. The threshold for significance was set after incorporating a Bonferroni adjustment for multiple comparisons. A Manhattan plot was created for each analysis to visualise the P-value (-log10) of each variant against its position in the genome using plot in R (version 3.5.2) and the significance threshold applied. Manhattan plots showing individual gene regions and significant SNP were created using ggplot2 in R (version 3.5.2). The quantile-quantile (QQ) plots were created using qqman in R (version 3.5.2) and the lambda GC value (genomic inflation factor) was derived to give a measure of the inflation within the sample by dividing the median value of the observed chi-squared statistic by the median expected chi-squared statistic (p=0.5). This will be one in the case of the null.

2.8 Identification of core genome from 2,059 *S. pneumoniae* isolates

Annotated assemblies of 2,059 pneumococcal isolates in GFF3 format produced by PROKKA (version 1.11) (187) (section 2.1.5) were inputted into the pan genome pipeline ROARY (version 3.11.2) (214) to calculate the pan genome. Here a multi-FASTA alignment of all the core genes was generated using PRANK (version 140603) (215).

Chapter 3 Pneumococcal disease in Singapore

3.1 Introduction

The pneumococcal polysaccharide capsule is an important pathogenic factor that acts as a barrier to inhibit binding of host's complement to the bacterial surface (216). This allows bacterial escape from opsonisation and contributes to the varying pathogenicity of different serotypes by influencing invasiveness, drug resistance profile, severity of disease and colony forming abilities of the cell. There are known differences between countries in both the global distribution of serotypes and the levels of antimicrobial resistance. Two contributors that impact on this are the effect of vaccine implementation and antimicrobial use.

3.1.1 Difference in disease capabilities between serotypes

Capsular polysaccharides (CPS) are composed of repeating units of simple saccharides that are polymerised into a polysaccharide chain. Diversity of serotypes is due to variation in the chemical structure of CPS and this could be in the oligosaccharide units themselves, or in the attached side groups (34). The designated serotype of the pneumococcus can profoundly affect certain abilities such as its capacity to cause invasive disease (32). A considerable amount of research has been carried out to identify serotypes that are more associated with invasive pneumococcal disease (IPD). Serotypes that have been statistically implicated in invasive disease, and are rarely seen in carriage, include serotypes 1, 3, 4, 5, 6B, 7F, 8, 14, 18C, 19A, and there has been much crossover and consistency in findings between studies that use populations from various geographical locations (30-32, 217-221). Many of these studies also consistently find serotypes 6A, 11A, 19F, 23F, 35F to be less invasive (30-32, 217, 219).

In addition to invasive potential some serotypes are associated with increased mortality. Serotype 3 appears to dominate in fatal cases in (24, 222) and these, along with the additional serotypes 6B, 19F (19) and 1, 7F (223, 224) are all associated with higher mortality rates. Others implicated with a high mortality rate include serotypes 23A, 17F, 9N and 18C (225), 11A (222). The impact of the exact serotype is likely to be based on geography; for example in the African meningitis belt pneumococcal meningitis caused by serotype 1 has been shown to have a high fatality rate (226). Also serotypes 1, 5 and 7 have been shown across the world to result in serious or complicated disease, however these serotypes did not have a higher prevalence in IPD cases than serotypes 6B, 23F or 19 in South East Asia (28).

Prolonged carriage duration and high acquisition rate are other factors which may contribute to the invasive potential of the pneumococcus (192). The expectation is that more invasive serotypes cause disease shortly after acquisition and the less invasive serotypes require a longer duration of carriage to result in disease (32). This was investigated using data from nasopharyngeal carriage and IPD isolates from children (32). Serotype 3 isolates did not have a high odds ratio for invasive disease, however, neither did it have a long carriage duration. Isolates were only recovered at 1/3 of the timepoints sampled, suggesting it is the potential of the serotype or genotype to cause disease at the point of acquisition that is important, rather than carriage duration (32).

Some additional serotypes to those commonly associated in IPD have been described to cause IPD in single studies, for example, serotype 12F (218), 2, 9, 16 (192), 20, 9N, 9L, 12B (217), 9V (219), 18 (220), 22F and 33F (221) showing there is some variability in the ability of serotypes to cause IPD. Of particular concern is the finding from two independent studies that show serotypes 1, 5, and 7 with a high invasive potential in children (225) and serotypes 1 and 7F (222) have a high potential to act in an opportunistic manner and cause IPD in otherwise healthy individuals. There has however been discrepancies between studies, for example (217) found serotype 3 to have a high odds ratio (OR) for invasive disease whereas (219) found the serotype to be statistically associated with disease that is less invasive. Both studies use data that encompasses a range of age groups, and both use IPD isolates. The additional inclusion of carriage isolates in (217) and the different geographical locations of Portugal and Switzerland between studies may be affecting the invasive potential exhibited by serotype 3 between datasets.

3.1.2 Pneumococcal disease in Singapore

Pneumococcal disease contributes considerably to the overall burden of disease in Singapore as it is the predominant cause of pneumonia which is the second most common cause of death (227). Between 1995 – 2004 the mean annual hospitalisation rate for pneumococcal disease was 10.9 per 100,000 population, in elderly patients this was between 16 – 61 per 100,000 for those aged 65 – 75 and between 53 – 173 per 100,000 population for patients aged ≥ 75 years (228). More recently, incidence of disease in children under five has been described as slightly higher than the mean at 13.6 per 100,000 population (229). Locally reported case fatality rates ranged from 13.1 – 21.4% (230) and are highest in the senior age group (228).

IPD was made a legally notifiable disease in 2010 (231) and pneumococcal vaccines are available and intended to prevent IPD and pneumonia. The PCV7 vaccine has been available since 2005 within the private market (232) but was not added to the National Childhood Immunisation program until 2009. This was superseded by PCV13 in December 2011 and is recommended for use in children over six weeks old (233). Despite the availability of PPV23 since 1988, and the national recommendation from Singapore's Ministry of Health to vaccinate all adults over 65 years, vaccination rates among seniors remains low at <8% (234, 235). The identification of serotypes causing disease allows estimation of coverage provided by the respective vaccinations. This was performed in 2014 and described coverage in adult populations as 34.6%, 58.5% and 69.1% for PCV7, PCV13 and PPV23 respectively (182). From child populations coverage by PCV7 and PCV13 was much higher at 64.5% and 79.1% respectively (182).

Before PCV vaccine was available the predominant circulating serotypes causing disease were 19F, 6B, 23F and 14 (236). Now, the main serotypes causing disease in the over 65 years are serotypes 3, 14, and 19A , in the 19-64 years age range are 3, 6B, 7F, 8, 19A, 14 19F and 23F (72, 224, 237), and finally in paediatrics are 14, 23F, 19F and 6B (28, 224).

The prevalence of antimicrobial resistance in *S. pneumoniae* is monitored in Asia due to the concern of increasing resistance (238). In Singapore, early studies in 1990 revealed penicillin resistance was as low as 0.5% (239). More recently, studies have described resistance to numerous antimicrobials in carriage and disease isolates (Table 2) and show the variability in data. Higher prevalence of resistance was identified in carriage rather than disease isolates, however the dramatic increase described between 1997 (240) and 2007 (232) is concerning because of the potential of transferring resistance determinants during colonisation to isolates that then go on to cause disease. Encouragingly, the most recent of these studies described lower proportions of resistance in disease isolates (237). Serotypes that have been associated with resistant phenotypes in Singapore include 14, 19F and 19A phenotypes (224, 229).

Levels of resistance described in Table 2 are not from population-based studies and some were based on only small numbers of isolates. Due to this they may not accurately reflect the national status of antimicrobial resistance. Genome sequences of 2,059 *S. pneumoniae* isolates from Singapore generated for this study were used to obtain a detailed understanding of the changing epidemiology of pneumococci between 1997 – 2016. The determination of current disease-causing serotypes and the identification of changes after vaccine implementation would help to assess the efficacy of current vaccines. The resistance profiles associated with disease isolates would also provide valuable data regarding current levels of antimicrobial resistance in the country and identify potential areas where antimicrobial prescribing policies might need to be reviewed.

Table 2 Proportion of resistance identified from pneumococcal populations in Singapore.

n=number of isolates tested for susceptibility to antimicrobials

Study period (year)	Study isolates	Antibiotic resistance penicillin (%)	Antibiotic resistance ceftriaxone (%)	Antibiotic resistance clindamycin (%)	Antibiotic resistance tetracycline (%)	Antibiotic resistance erythromycin (%)	Antibiotic resistance cotrimoxazole (%)	Reference
1997	Carriage children n=102	27.4	-	24.5	48	38.4	-	(240)
1997 - 2004	Disease children n=147	44	15	-	-	62	-	(229)
1997 - 2013	Disease children adult n=472	79 35.4	8 3	- -	62.5 55.1	40.6 83.2	- -	(224)
2000 - 2001	Disease Exact ages not described n=35	17.1	0	-	-	40	67	(238)
2007 – 2008	Carriage children n=59	69.5	-	45.8	67.8	78	-	(232)
2012-2017	Disease >50 years n=77	6%	-	22.4	-	47	34.3	(237)

3.2 Results

3.2.1 Descriptive Epidemiology of Singapore Pneumococcal dataset

Epidemiological analysis of pneumococcal isolates helped to identify change in characteristics of disease over time. A total of 2,059 viable *S. pneumoniae* isolates that were successfully sequenced in the present study and in (182) were combined for analysis. The metadata associated with isolates was used to compare the distribution across hospital settings and age groups.

The total number of isolates collected from each of the four participating hospitals varied due to differences in local policies for isolate storage. Of the 2,059 isolates suitable for downstream processing, 910 were from KKH (44%), 141 from NPHL (7%), 231 from NUH (11%), 330 from SGH (16%) and 447 were from TTSH (22%). The specialism of KKH is in women and children, therefore, isolates from this institution were predominantly from children. Similarly, TTSH is mainly an adult hospital and this is reflected by a large collection of adult isolates (Table 3).

Table 3 Distribution of isolates across patient age groups collected by participating hospitals (KKH, NUH, SGH, TTSH and NHPL) between 1997 – 2016

Age (years)	Frequency of isolates KKH (n=910) (%)	Frequency of isolates NUH (n=231) (%)	Frequency of isolates SGH (n=330) (%)	Frequency of isolates TTSH (n=447) (%)	Total (%) (n=2059)
<1	275 (30%)	4 (2%)	6 (2%)	1 (1%)	289 (14%)
2-5	368 (40%)	12 (5%)	2 (1%)	0	391 (19%)
6-17	122 (13%)	7 (3%)	5 (2%)	2 (1%)	138 (7%)
18-64	28 (3%)	93 (40%)	162 (48%)	250 (57%)	605 (29%)
65+	1 (1%)	54 (23%)	145 (44%)	183 (41%)	435 (21%)
N/A	116 (13%)	61 (27%)	10 (3%)	0	201 (10%)

The specimen type was used to classify disease into IPD and non-IPD based on whether isolate collection was from a normally sterile site as described in section 2.1.6. Based on this classification, 1,089 isolates were associated with IPD (53%), 846 were associated with non-IPD (41%), and information was unknown for 123 samples (6%). From these, information on the corresponding year of isolation was available for a total 1,756 isolates. The variation in the frequency of isolate collection across the years and how they were distributed between IPD and non IPD was shown for all collected isolates (Figure 4) and for isolates collected at specific hospitals (Figure 5). Isolates from National Public Health Laboratory (NPHL) were not included as all but one isolate was from IPD and samples only from the years 2009 and 2010 obtained.

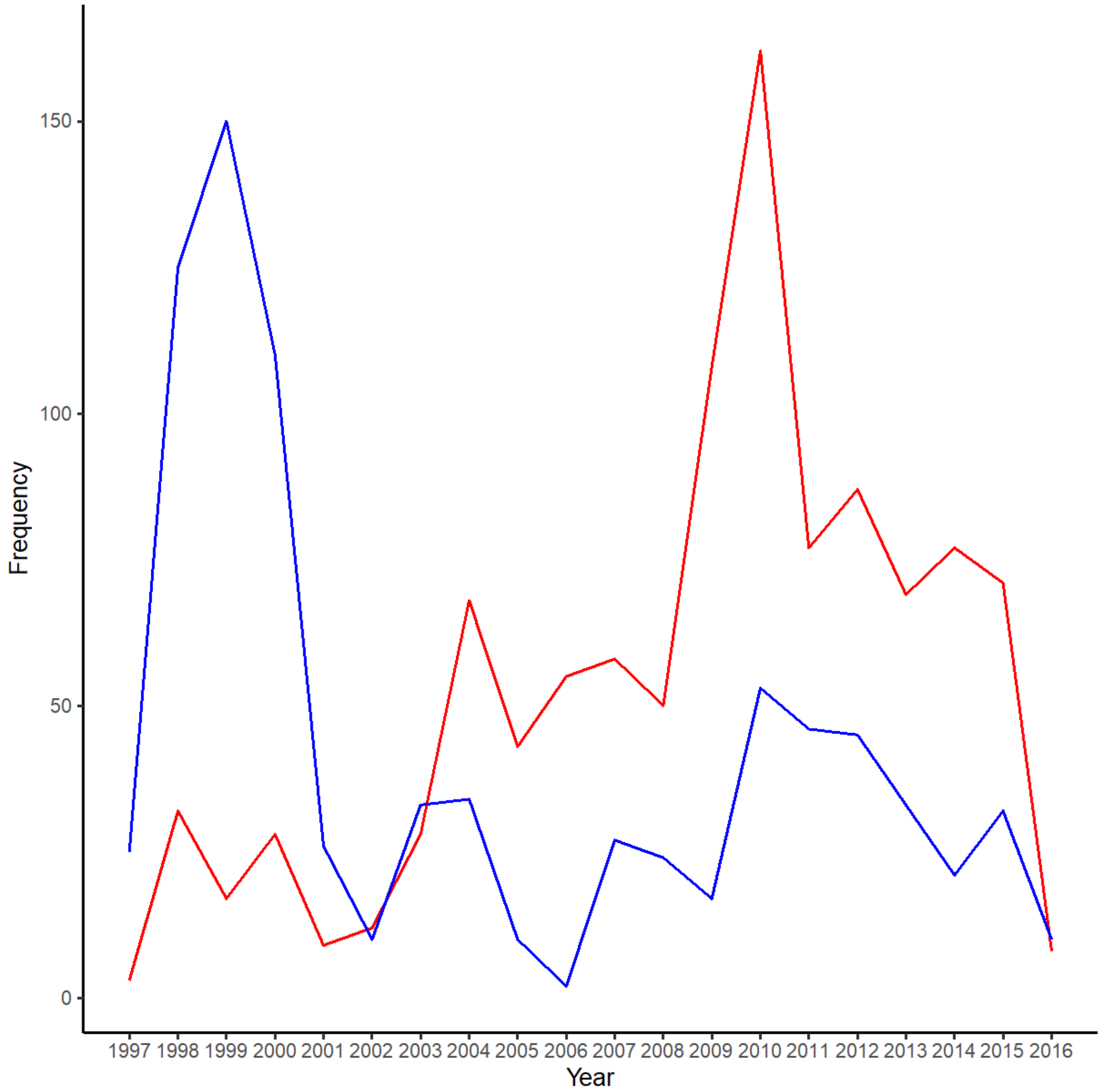


Figure 4 Total number of *S. pneumoniae* isolates collected and stored by all participating hospitals (KKH, NUH, SGH, TTSH and NHPL) each year between 1997 – 2016 (n=1,756). The red line represents isolates causing IPD (n=1,062) and the blue represents isolates causing non-IPD (n=833)

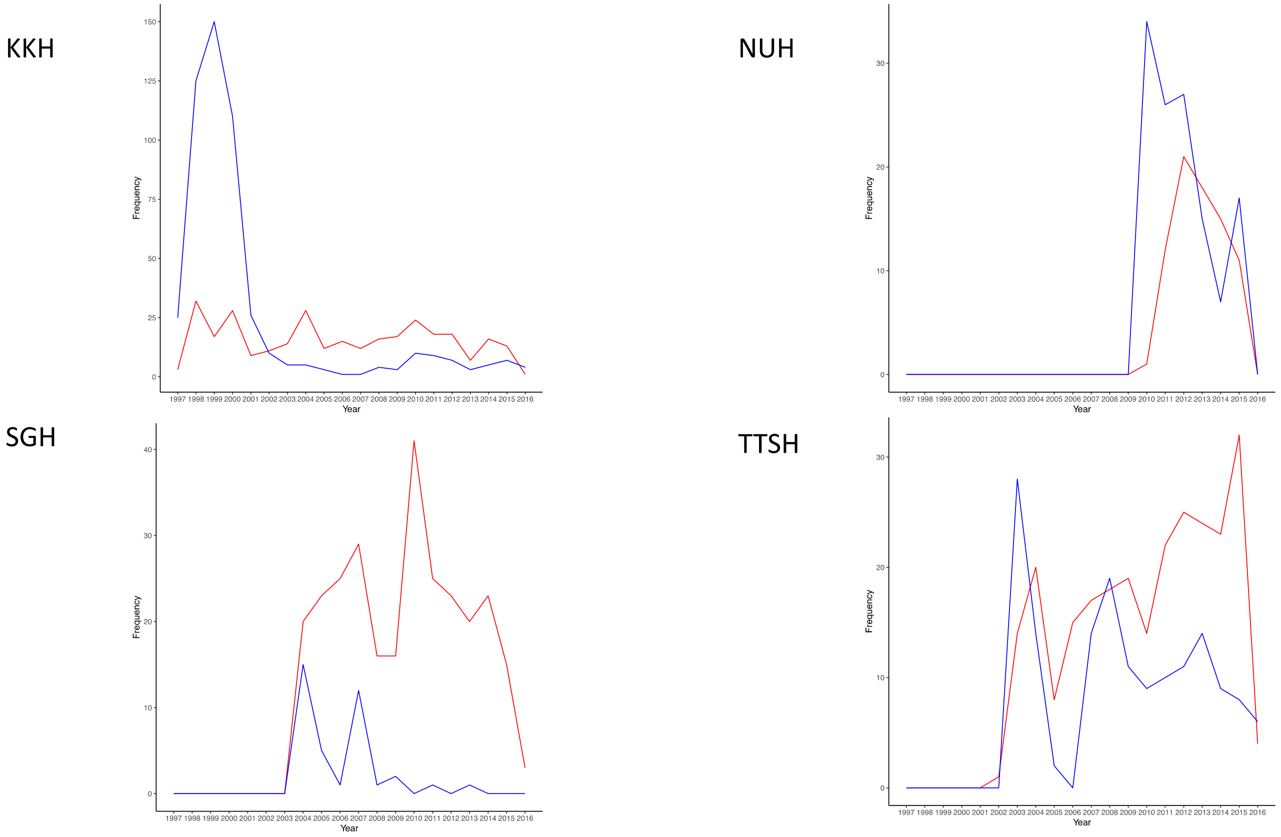


Figure 5 Total number of *S. pneumoniae* isolates collected and stored by specific hospitals; KKH 824 isolates (IPD=311, non-IPD=513), NUH 204 isolates (IPD=78, non-IPD=126), SGH 317 isolates (IPD=279, non-IPD=38) and TTSH 411 isolates (IPD=256, non-IPD=155) that had both year and site of infection information between 1997 – 2016. The red line represents isolates causing IPD and the blue represents isolates causing non-IPD

Figure 5 shows KKH contributed the highest number of IPD and non-IPD isolates from all hospitals and isolates were collected over a much longer timeframe 1997-2016. Although there is some fluctuation in the number of IPD specimens collected yearly this generally remained stable. For non-IPD isolates there was a surge in numbers collected between 1997-2001, then after this point numbers considerably decreased and thereafter remained stable and lower than that of IPD. No isolates were obtained by NUH until 2010 and rapid increase in both IPD and non-IPD in 2010 reflects the time at which the hospital began collecting the pneumococcal isolates. Between 2010 – 2012 numbers of IPD isolates collected continued to increase, then between 2012 – 2016 numbers rapidly declined. The decline in non-IPD isolates fluctuated more between the years 2010 – 2014 and this was followed by another sudden surge in isolate collection in 2015 which was not maintained. For NUH there is a gradual but constant decline in the numbers of IPD however, the opposite is seen in TTSH. Data from TTSH fluctuates far more than the other hospitals throughout the years. Isolates begin to be stored for IPD and non-IPD in 2002 and 2003 respectively and the initial surge is reflective of study participation. For IPD isolates, there are three further spikes in isolate collection, but overall, there is an upward trend in collection over the period 2002 – 2015. This trend is not mirrored in non-IPD isolates where although there are a further two peaks in isolate collection in 2008 and 2013 , there is a general decrease in isolates collected. Isolate collection in SGH began in 2004 and this was the only hospital that reported a higher number of IPD than non-IPD isolates in the first year. The increase in IPD isolates continued until 2007 followed by a rapid decline in 2008 and plateau in 2009. A second peak in IPD isolates collected occurred in 2010 and after this point a gradual decline in numbers was witnessed. For non-IPD isolates, a second peak was identified in 2007 and after this point very little non-IPD was collected.

The data from each of the hospitals was merged to identify trends of disease types across the years. Figure 4 shows the initial peak in non-IPD isolates in 1999 when sample collection was initiated by KKH followed by a dramatic drop when their sample collection waned. The second increase in numbers in 2003 and 2004 reflects the addition of samples from other hospitals but is considerably smaller as after the initial peak KKH contributes a very small number of isolates. After this point, levels of non-IPD consistently remained lower than IPD. Numbers of IPD isolates begin smaller than non-IPD however a sharp increase in numbers in 2004 changed this and it remained the case for the remainder of the collection period. A sharp increase in isolate frequency is present in 2009 and 2010 due to the addition of IPD isolates from NHPL but levels remained elevated even after this period between 2011 – 2015. For all isolates, values for the year 2016 is not directly representative of the year or comparable to the other years because data collection was only performed until March 2016.

Once IPD became notifiable in 2010, it became possible to determine how reflective the collected isolates were of total IPD rates in Singapore. Table 4 shows 98% of the total IPD was represented from the data in 2010, however thereafter the isolates collected represented only ~50% of total IPD.

Table 4 Comparison in the number of IPD cases notified to Ministry of Health Singapore with the number of IPD isolates collected in the present study

Year	Number of IPD isolates collected from collaborating hospitals	Number of IPD cases notified (241, 242)	Proportion of total IPD cases represented by dataset (%)
2010	162	166	98
2011	77	148	52
2012	87	163	53
2013	69	167	41
2014	77	147	52
2015	71	146	49

A total of 1,267 isolates (61%) were from male patients, 694 from female patients (34%) and the patient gender was not known for 98 isolates (5%). The ethnicity of the patients with pneumococcal disease was varied, 1,058 isolates were from Chinese patients (51%), 337 isolates from Malay patients (16%), 194 from Indian patients (10%), 117 isolates grouped in an ‘other’ classification (6%) as described in section 2.1.6, ethnicity metadata was not available for 353 isolates (17%). The most common syndrome in IPD was bacteraemia (91%) and in non-IPD was pneumonia (79%) Table 5.

Table 5 Site of infection and frequency of isolate collection in IPD (n=1089) and non-IPD (n=846)

IPD site of infection	**Frequency (n)**	**Non-IPD site of infection**	**Frequency (n)**
Blood	988	**Chest**	665
Pleura	63	**Ear**	77
Cerebral spinal fluid	19	**Nose**	38
Knee	7	**Eye**	36
Lung	2	**Skin**	15
Bone	2	**Vagina**	11
Joint fluid	2	**Urine**	2
Ascitic fluid	1	**Throat**	2
Brain	1		
Ovary	1		
Pelvis	1		
Pericardium	1		
Submental space	1		

3.2.2 Serotypes associated with disease

There were 64 different serotypes identified within the dataset, the most common being 19F (n=365, 18%) followed by 23F (n=244, 12%), 14 (n=196, 10%), 19A (n=161, 8%), 3 (n=158, 8%), 6E (n=143, 7%) and 6A (n=97, 5%). Of those tested, 17 isolates did not result in serotype allocation using PneumoCaT. An empirical OR was calculated for the most common serotypes to compare the probability of disease outcome due to the serotype. An OR of 1 indicated an equal probability that the serotype will be identified from IPD or non-IPD. Serotypes in which the 95% confidence intervals (CI) spanned 1 were not associated with just one disease group. Serotypes with an OR >1 indicated it had an increased probability of being isolated from IPD and examples of serotypes attributed to this were serotypes 14 (OR 3.29, CI 2.28 – 4.76), 19A (OR 2.49, CI 1.71 – 3.62), 3 (OR 1.63, CI 1.15 – 2.31), 8 (OR 5.65, CI 2.54 – 12.53), 7A (OR 4.2, CI 1.96 – 9), 4 (OR 5.89, CI 2.5 – 13.9) and 20 (OR 4.54, CI 1.56 – 13.18). An OR <1 indicated an increased probability of it being associated with non-IPD and examples include serotypes 19F (OR 0.11, CI 0.08 – 0.15), 23F (OR 0.53, CI 0.4 – 0.7) and 15A (OR 0.37, CI 0.2 – 0.66) (Table 6).

The serotypes that appeared to be specifically associated with a distinct age group were summarised in Table 7. An OR greater than 1, whose 95% CO were also greater than 1 indicated an increased probability of infection within the age range. Serotypes that showed increased probability of infection in the 18-64 years age range included serotype 1 (OR 10.69, CI 3.08-37.07), 8 (OR 4.51, CI 1.62-12.85), 7A (OR 3.1, CI), 12F (OR 2.71, CI), and serotype 4 (OR 2.54, 1.42-4.51). Serotypes that showed increased probability of infection in the ≥65 years age group were serotype 6D (OR4.0, CI1.71-9.32), 7A (OR 2.31, CI 1.29-4.29), 4 (OR 2.39, CI 1.33-4.29) and serotype 20 (OR 2.43, CI 1.11-5.33). An increased probability for infection in the ≤ 5 years group was seen for serotype 6B (OR 4.56, CI 1.62-12.85). An OR that was <1 and where the 95% CI remained <1 indicated a reduction in the probability of infection within that age group. Examples of these were present in the ≤ 5 years group for serotype 6D (OR 0.27, CI 0.08-0.91), 4 (OR 0.005-0.26) and serotype 20 (OR 0.07, CI 0.009-0.5).

Table 6 Serotype distribution of 2,059 *S. pneumoniae* isolates (IPD=898) and (non-IPD=733) for the most frequently identified serotypes. Rank ordered by serotype specific odds ratio (OR). Serotypes associated with specific allocation to a disease classification highlighted in bold.

Serotype (n)	Proportion of total disease caused by serotype (%)	Total IPD (n=898)	Total non-IPD (n=733)	OR (95% CI)
4 (51)	2	44	6	**5.89 (2.5 – 13.9)**
8 (57)	3	49	7	**5.65 (2.54 – 12.53)**
20 (27)	1	23	4	**4.54 (1.56 – 13.18)**
7A (53)	3	42	8	**4.2 (1.96 – 9)**
14 (196)	10	146	38	**3.29 (2.28 – 4.76)**
19A (161)	8	117	39	**2.49 (1.71 – 3.62)**
6D (22)	1	15	6	1.96 (0.76 – 5.06)
23A (32)	2	19	8	1.86 (0.81 – 4.27)
3 (158)	8	101	50	**1.63 (1.15 – 2.31)**
6E (143)	7	85	49	1.38 (0.96 – 1.98)
6A (97)	5	53	39	1.06 (0.69 – 1.62)
15B (31)	2	16	12	1.04 (0.49 – 2.2)
6C (21)	1	11	9	0.95 (0.39 – 2.3)
11D (23)	1	9	13	0.53 (0.23 – 1.26)
23F (244)	12	95	130	**0.53 (0.4 – 0.7)**
15A (57)	3	17	35	**0.37 (0.2 – 0.66)**
19F (365)	18	56	280	**0.11 (0.08 – 0.15**)

Table 7 Serotypes associated with specific allocation to an age group rank ordered by odds ratio (OR). Significant OR where 95% CI do not span 1 are highlighted in bold.

Serotype (n)	**Frequency of serotype in ≤ 5 years (%)** **OR (95% CI)**	**Frequency of serotype in 6-17 years (%)** **OR (95% CI)**	**Frequency of serotype in 18-64 years (%)** **OR (95% CI)**	**Frequency of serotype in ≥65 years (%)** **OR (95% CI)**	**Frequency of serotype in unknown age group (%)**
1 (20)	0	2 (10) 1.55 (0.35-6.81)	15 (75) **10.69 (3.08-37.07)**	1 (5) 0.19 (0.03-1.43)	2 (10)
6B (19)	13 (68) **4.56 (1.62-12.85)**	0	4 (21) 0.59 (0.19-1.81)	1 (5) 0.19 (0.03-1.43)	1 (5)
8 (57)	1 (2) 0.03 (0.004-0.22)	0	37 (65) **4.51 (2.55-7.99)**	17 (30) 1.48 (0.83-2.65)	2 (4)
6D (22)	3 (14) **0.27 (0.08-0.91)**	1 (5) 0.58 (0.08-4.38)	6 (27) 0.78 (0.3-2.0)	12 (55) **4.0 (1.71-9.32)**	0
7A (53)	0	0	29 (55) **3.1 (1.76-5.59)**	20 (38) **2.31 (1.29-4.13)**	4 (8)
12F (18)	0	1 (6) 0.82 (0.11-6.27)	9 (50) **2.71 (1.00-7.32)**	6 (33) 1.97 (0.71-5.45)	2 (11)
4 (51)	1 (2) **0.04 (0.005-0.26**)	1 (2) 0.26 (0.04-1.88)	26 (51) **2.54 (1.42-4.51)**	20 (39) **2.39 (1.33-4.29**)	3 (6)
20 (27)	1 (4) **0.07 (0.009-0.5)**	2 (7) 1.03 (0.24-4.4)	12 (44) 1.81 (0.83-3.93)	11 (41) **2.43 (1.11-5.33)**	1 (4)

3.2.3 Change in serotype distribution over time

The seven most common serotypes associated with disease were described in section 3.2.2. As the total number of samples was highly variable across the study period, serotypes were expressed as a proportion of total isolates each year to identify change over the period 1997 – 2016 (Figure 6). Although fluctuations are present, the serotypes covered by the PCV7 vaccine 23F, 19F and 14 saw a decreasing trend over the entire study period. This began before PCV7 vaccine was introduced. Two serotypes, 3 and 19A that were not covered by PCV7, but which were later included in PCV13 saw an opposite trend. Initially, the proportion of infection was low from these serotypes, and increased over time, even after implementation of PCV13. The proportion of disease caused by the third additional serotype present in PCV13 was maintained at similar levels throughout. The final serotype commonly identified in the dataset, serotype 6E, showed a constant decline in the proportion of overall disease it caused.

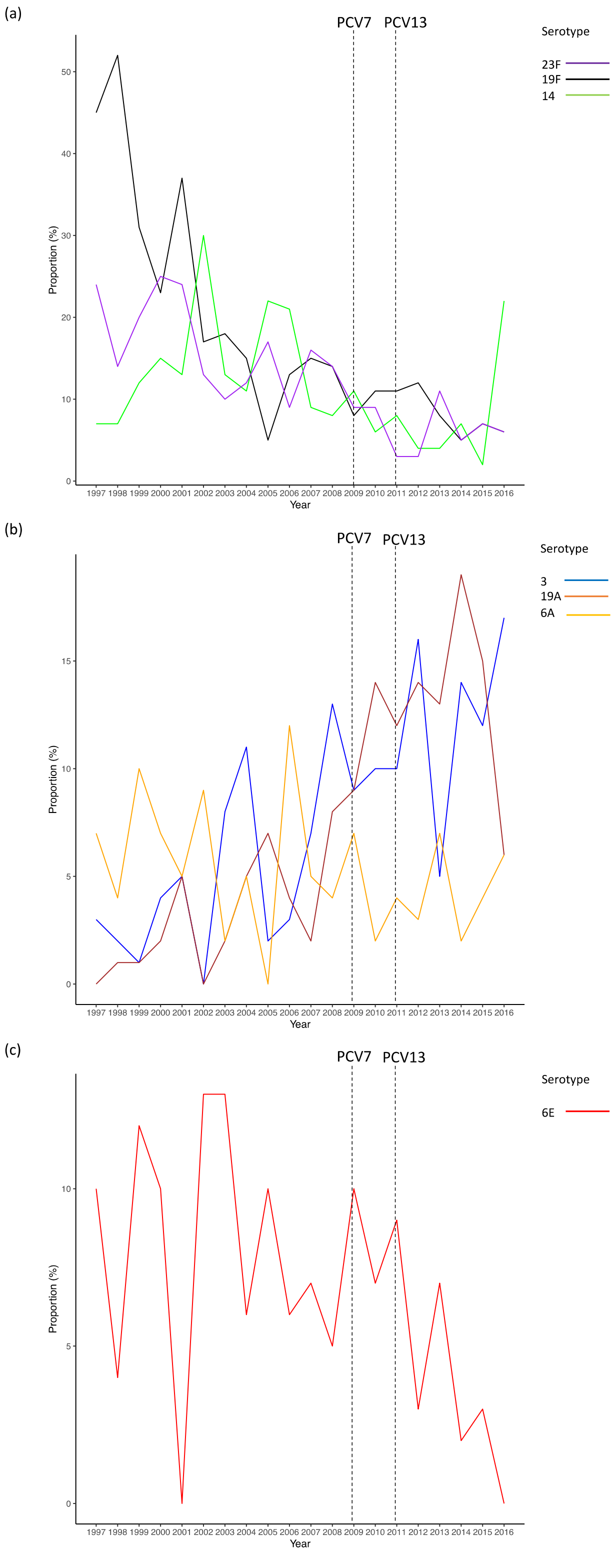


Figure 6 Proportion of total disease caused by the seven serotypes most isolated from 2,059 *S. pneumoniae* isolates. Of these serotypes, (a) shows serotypes 23F, 19F, 14 present in PCV7, (b) shows serotypes 3, 19A, 6A present in PCV13, (c) shows serotype 6E not included in either PCV7/13. The dashed vertical line illustrated the respective years PCV7 and PCV13 was implemented into national immunisation programs.

3.2.4 Proportion of vaccine serotypes responsible for disease

Pneumococcal vaccines offer protection against disease organisms whose serotype is encompassed by the vaccine. Changes in the proportion of disease from some of these serotypes has been shown in Figure 6 however, to assess efficacy of current and future vaccines, the proportion of disease from isolates containing all serotypes targeted by the vaccine must be determined. Figure 7 shows vaccine efficacy by describing the proportion of disease attributed by isolates that contain the serotype covered by the PCV7, PCV13 and PPV23 vaccines.

Vaccine coverage of PCV7 between 1997 – 2016 ranged from 21% - 76%, and across the study period, the proportion of disease caused by vaccine serotypes decreased. To replace this, non-PVC7 serotypes caused increasing proportions of infection, and after 2004, were responsible for the predominance of disease. At the time of PCV7 implementation into vaccination schedules, 34% of the disease isolates collected contained serotypes present in the vaccine. The range of vaccine coverage seen from PCV13 was 50% - 86% and, at the time of implementation, 50% of disease isolates contained serotypes covered by PCV13. Coverage offered by PPV23 ranged from 59% - 89% across the study period (Figure 7).

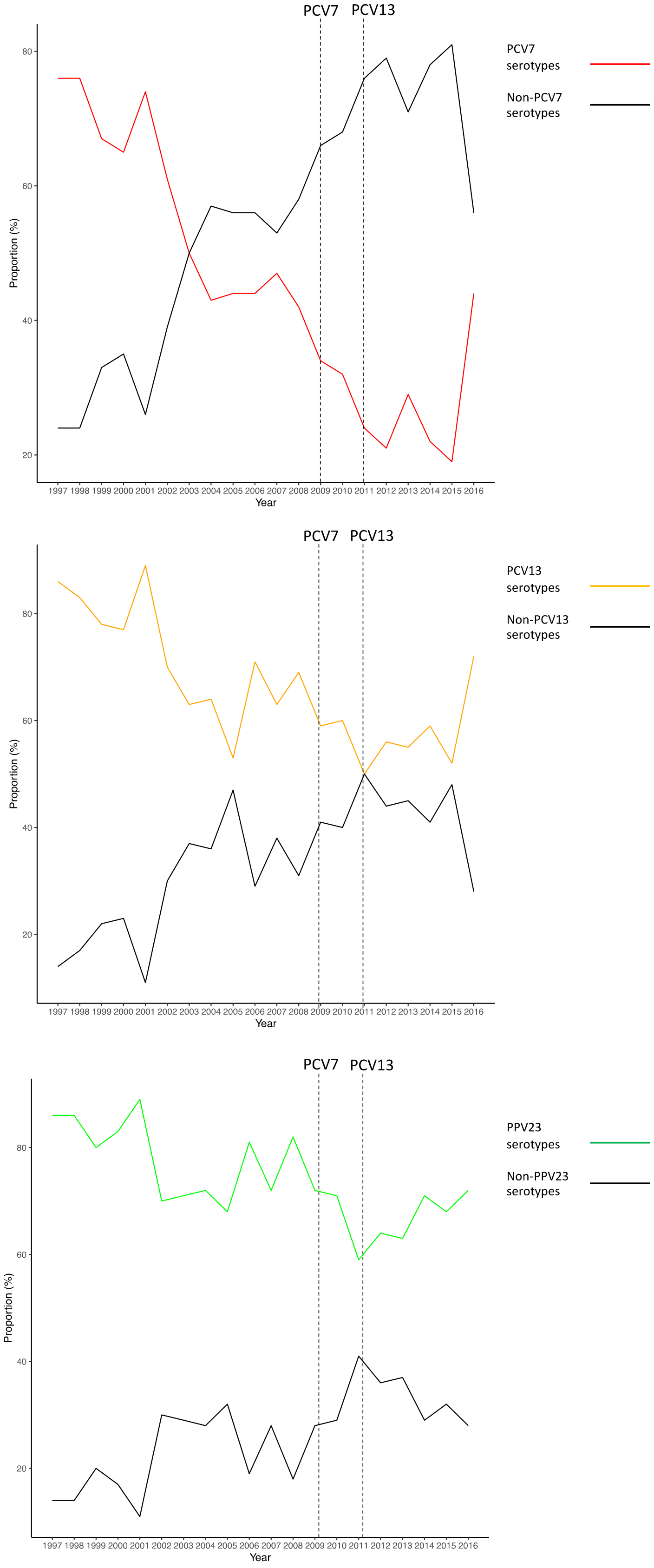


Figure 7 Proportion of total isolates that contain a serotype present in the vaccines PCV7 (top), PCV13 (middle) and PPV23 (bottom). The dashed vertical line illustrated the respective years PCV7 and PCV13 was implemented into national immunisation programs.

3.2.5 Levels of antimicrobial resistance in *S. pneumoniae*

Antimicrobial data coupled with the *S. pneumoniae* isolate was used to estimate levels of resistance within the dataset (Table 8). The largest proportion of isolates had phenotypic resistance to cotrimoxazole (63%, 690/1095), followed closely by erythromycin (58%, 1045/1808), doxycycline (58%, 63/109) and tetracycline (58%,212/367). No resistance to vancomycin (n=744) or linezolid (n=16) was present in any isolates however only a small number of isolates were tested for the latter antibiotic. Of the isolates with penicillin resistance, 88% were also resistant to erythromycin (462/523), 82% to cotrimoxazole (429/523), and 74% were resistant to both erythromycin and cotrimoxazole (389/523). Serotype 19F had the highest proportion of resistance than any other serotype to penicillin (41%, 212/523), cotrimoxazole (34%, 232/690), oxacillin (32%, 8/25), erythromycin (29%, 299/1045), tetracycline (26%, 56/212) and clindamycin (15%, 40/266). Serotype 19A had the highest proportion of resistance to doxycycline (16%, 10/63).

Due to the fluctuations in isolate collection across the years, isolates with data regarding year of collection and antimicrobial susceptibility were used to calculate the proportion of isolates with a resistant phenotype across the study period (Figure 8). Although overall higher numbers of isolates had cotrimoxazole resistance (63%), the decrease in the proportion of isolates with resistance across the years was identified. A decrease in resistance across the years was also identified for penicillin and chloramphenicol. The proportion of isolates with resistance to erythromycin declined initially between 2001 – 2004 however after this point it was maintained at similar levels. Overall similar proportions of tetracycline resistance were identified across the time period. An increase in the proportion of collected isolates with clindamycin resistance was identified between 2002 and 2016, with approximately 30- 40% of the collected isolates showing resistance by the end of the study period (Figure 8).

Table 8 Total number of isolates with antimicrobial susceptibility data showing the proportions classified as sensitive, intermediately resistant or resistant

Antibiotic (n)	Resistant (%)	Intermediate resistance (%)	Sensitive (%)
Cotrimoxazole (1095)	690 (63)	22 (22)	383 (35)
Erythromycin (1808)	1045 (58)	17 (1)	746 (41)
Doxycycline (109)	63 (58)	2 (2)	44 (40)
Tetracycline (367)	212 (58)	11 (3)	144 (39)
Oxacillin (81)	32 (40)	0	49 (60)
Penicillin (1670)	523 (31)	145 (9)	1002 (60)
Clindamycin (958)	266 (28)	3 (30)	689 (72)
Chloramphenicol (180)	45 (25)	2 (1)	133 (74)

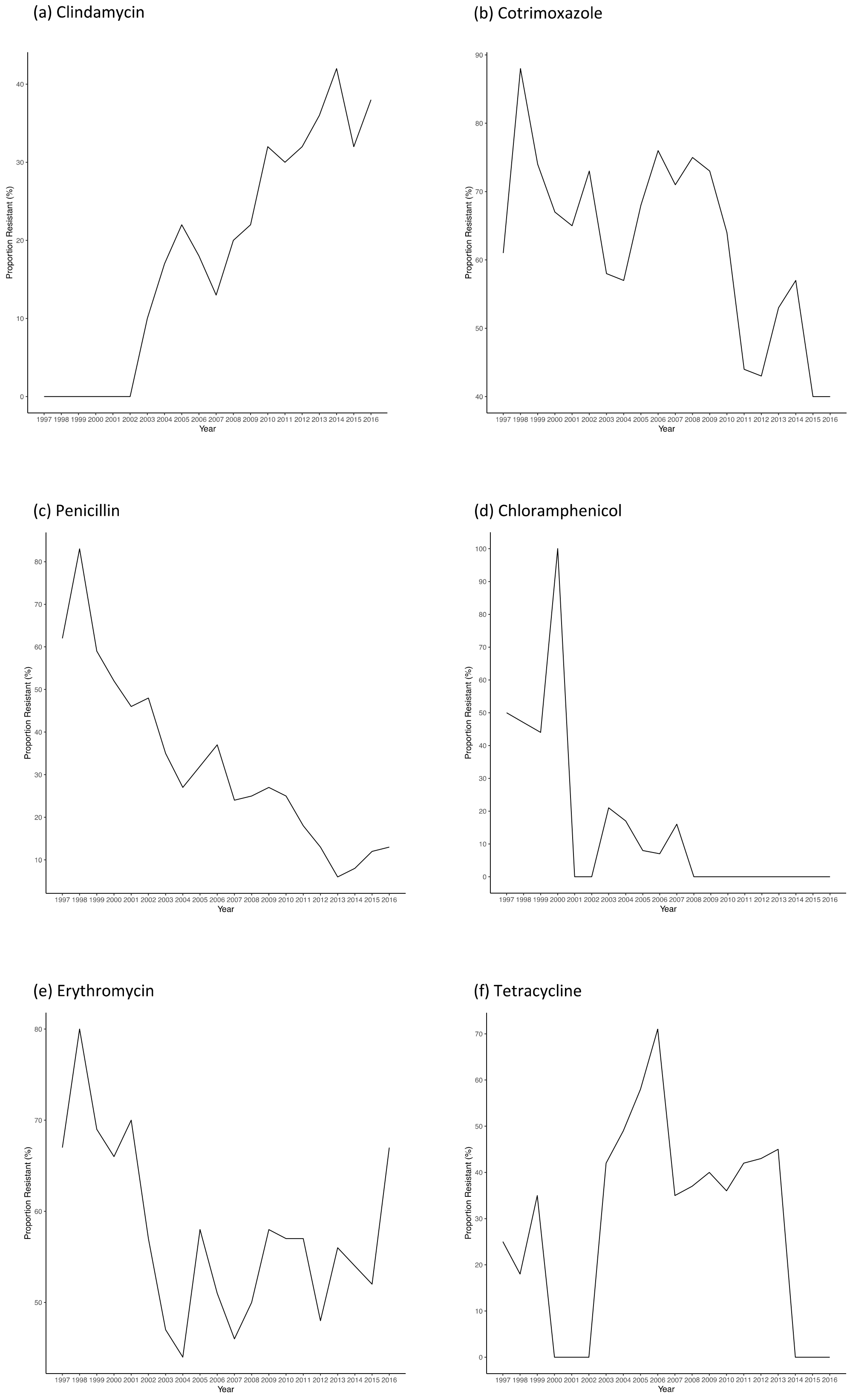


Figure 8 Change in the proportion of total isolates with resistance to the described antibiotics between 1997 – 2016. Total number of isolates with supporting antibiotic resistance data for the specified year was (a) n=939, (b) n=1045, (c) n=1491, (d) n=177, (e) n=1751, (f) n=356.

3.2.6 Level of diversity in the dataset

Whole genome sequencing data from 2,059 *S. pneumoniae* isolates was used to construct a phylogenetic tree (section 2.4) to show the relationship within the species in the data (Figure 9). The pattern of branching within the tree is complex. Many branch points represent the divergence event for a large number of descendant groups. In some cases, these go on to form many more internal branches representing the vast diversification of isolates, and others result in a much smaller number of descendants. The branch lengths between some nodes can be quite long showing there to be a high number of changes occurring in the sequences prior to the next level of separation and this is indicative of high diversity.

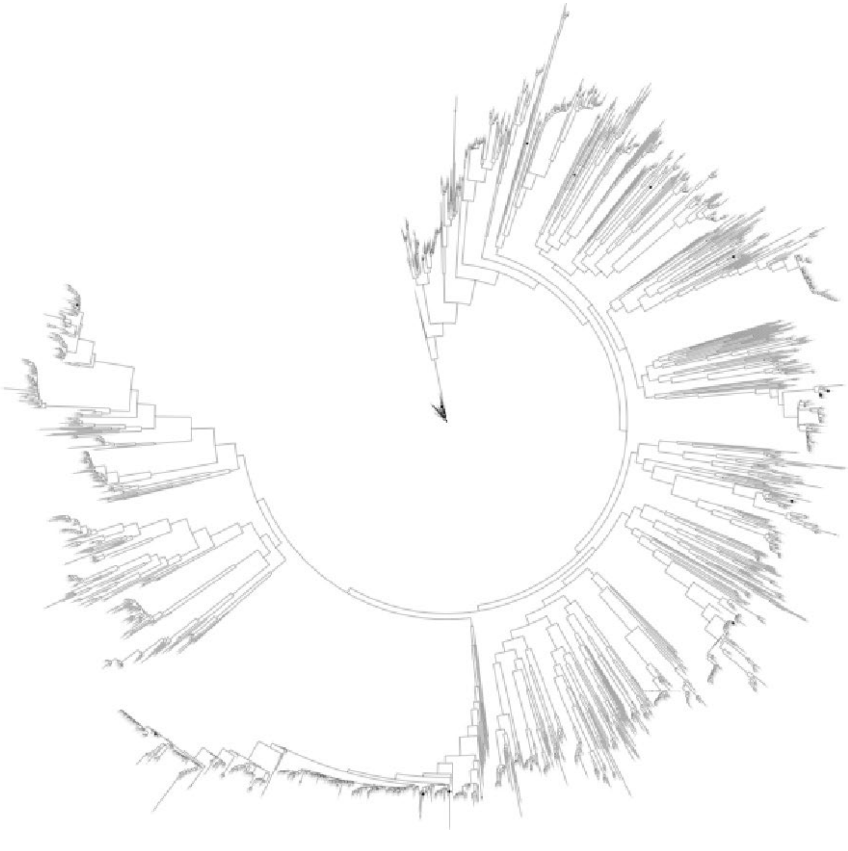


Figure 9 A maximum-likelihood phylogenetic tree constructed by FastTree using whole genomes of 2,059 isolates of *S. pneumoniae*

Multi-locus sequence typing (MLST) identified 385 individual sequence types from 2,059 isolates, the most common of which was ST81 (10%, 208/2059). Sequence types were not found for 261 isolates (13%) as they were novel combinations of alleles absent in the MLST database. Analysis performed by goeBURST identified 250 clonal clusters within the dataset. Of these, 64 clonal clusters (CC) consisted of more than one sequence type (ST) and encompassed 1,336 of the study isolates (65%). The largest of these was made up of 264 isolates from 11 STs, the predominant being ST81 (79%). The remaining 186 CC consisted of singleton groups that were made up of only one ST.

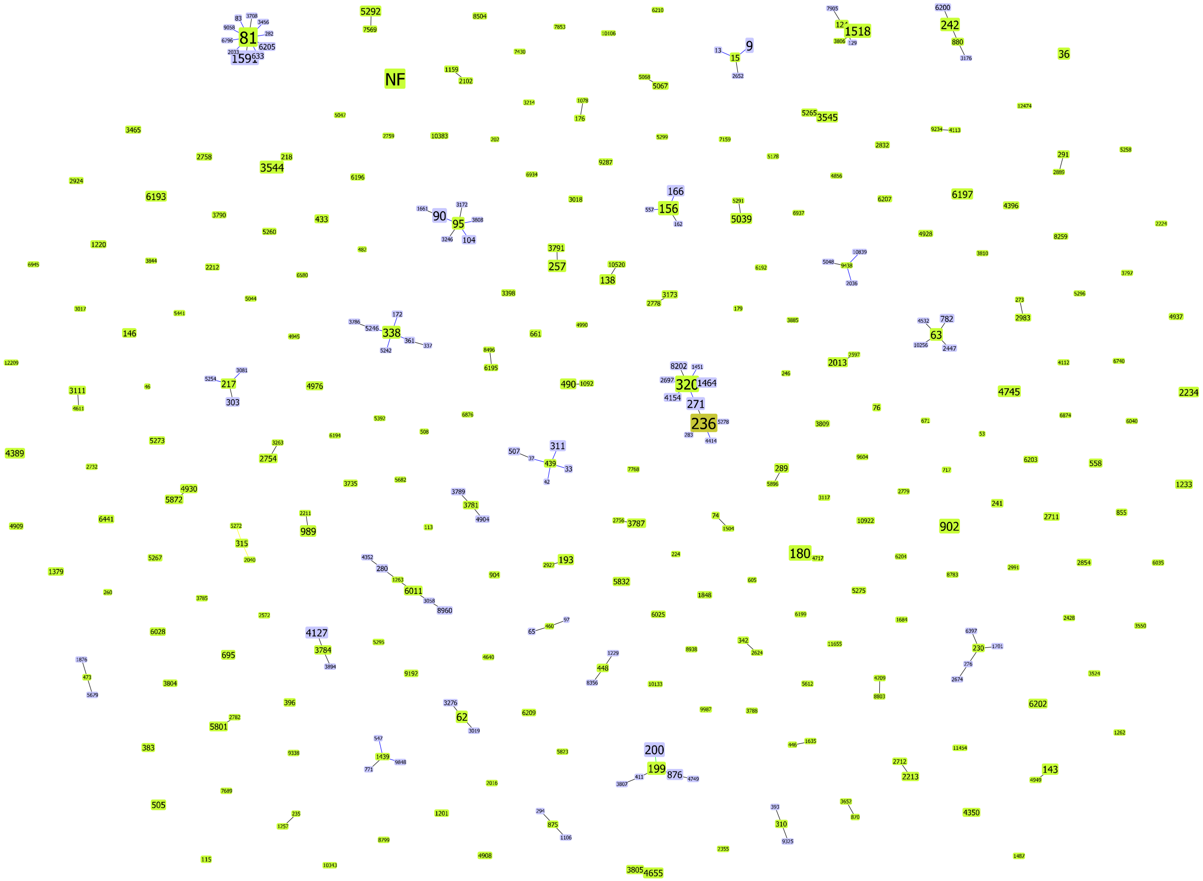


Figure 10 goeBURST analysis of 2,059 *S. pneumoniae* isolates showing 250 clonal complexes. The largest CC group 0 (n=264) is showing ST81 as the predominant strain. The other main group CC1 (n=249) has the predominant strain ST236. CC69 (n= 261) encompassed all the NF strains.

3.3 Discussion

The isolates included within the study were from the four main hospitals of Singapore. Obtaining samples from a range of institutions not only provides the opportunity to collect the greatest number of samples associated with pneumococcal disease, but it also generates a more accurate perspective of the bacterial population than if samples were collected from a single location.

3.3.1 Analysis of sample collection

Demographic analysis was performed to determine whether one group was overrepresented in the dataset which could lead to interpretive biases between clinical and patient data, antimicrobial resistance, and strain. The identification of isolates contributed by each of the individual hospitals showed KKH overall provided the highest number of samples and that these were mostly representative of child isolates < 5 years. The other participating hospitals TTSH, SGH and NUH did not commence data collection until later, in 2002, 2004 and 2010 respectively, and therefore early IPD infection is more reflective of childhood infections presenting at KKH.

The breakdown of isolates by disease type provided some insight into how rates of both IPD and non-IPD have changed over the years, however due to notable fluctuations in numbers between the years, this is likely to be more of a reflection of hospital storage policies. Comparisons in the numbers of IPD isolates collected by hospitals and those notified to MOH Singapore show that a large proportion of IPD was not accounted for by the collected isolates. Due to this, it was not possible to correctly assess disease incidence or prevalence across the study years using this data. Despite this, there was still a high number of both IPD and non-IPD isolates available for analysis which provided a reasonable representation of pneumococcal disease and specific characteristics of disease within Singapore.

3.3.2 Epidemiological interpretation based on pneumococcal dataset of 2,059

All of the isolates in this study have some level of invasive potential as they have transitioned from colonisation to cause disease. Many of the earlier studies that identified serotypes to have an increased invasive potential in IPD compared them to carriage isolates however, this study uses non-IPD as a reference. The serotypes identified to be more associated with IPD rather than non IPD agreed with other studies in the field (28, 72, 224, 237) when carriage was the baseline reference. Serotypes with a high odds ratio for disease in certain age groups were also identified from the dataset and of these, serotype 8 associated with adult disease and serotype 6B associated paediatric disease was also seen in other studies (72, 224).

Proportions of isolates causing disease with serotypes that are covered by PCV7 vaccine decreased over the time period, and interestingly this occurred before implementation of PCV7. The expansion of serotypes 3 and 19A which are later included in the protection of PCV13, showed justification for increasing the serotype coverage of this vaccine. Vaccine implementation in 2011 did not cease or reduce the proportion of disease caused by the additional three serotypes absent in PCV7, however, as more of the population are receiving the PCV13 rather than PCV7, it would be expected for the frequency of disease caused by these serotypes to gradually decrease.

The proportion of disease covered by PCV7 serotypes was 76% at the beginning of the study period in 1997. However, by 2015 serotype coverage of disease isolates was as little as ~20%. Low level coverage of PCV7 vaccine was also identified in Singapore by (224) which covered only 37% of adult IPD. The decreasing coverage offered by PCV7 evidenced the need to expand the number of serotypes covered in the vaccine. This was confirmed by the increase in coverage of disease-causing serotypes by PCV13 which continues to predominate infection. Serotype replacement has been described in other countries (243) but as an increase in disease caused by non-PCV7 serotypes was identified before vaccine implementation, there was no evidence of this. Similar proportions of disease caused by PCV13 and non-PCV13 serotypes were described after vaccine implementation, again showing no evidence of serotype replacement. The proportion of infection that was not covered by any of the vaccines was ~30%, similar to the findings in (224), however this did rise to 41% in 2011. This is cause for concern because it had occurred despite poor vaccine uptake being described in the elderly population of Singapore largely brought about through misconceptions in the benefits and effects of the pneumococcal vaccination and in the cost (234). This highlights the limitations of current vaccines and the need to monitor and improve upon the coverage capacity of current vaccines.

A range of resistance rates have been described in previous studies from collections of *S. pneumoniae* isolates in Singapore (Table 2). Penicillin resistance is commonly tested and, between 2004 – 2020 resistance has been identified in 6 – 44% of the disease isolates tested (229, 237, 238). The most recent of these studies only identified 6% of bacteria as resistant (237) and this along with a decrease in the proportion of isolates with resistance in the present study, implies penicillin resistance is not increasing. Resistance to erythromycin has been reported to range from 40 – 62% in disease isolates (229, 237, 238) and the resistance level of 58% seen in this study again falls within these levels and is generally maintained across the study period. The highest proportion of resistance seen in this dataset and (238) was to cotrimoxazole (63%) however, the most recent findings by (237) described much lower resistance levels of 34.5%. This reduced resistance may be reflective of the adult population tested in (237). The proportion of isolates with cotrimoxazole resistance has been shown to be decreasing over time in this study however, as resistance to both erythromycin and cotrimoxazole remain relatively high, treatment with these should be used with caution in suspected pneumococcal disease. Levels of clindamycin resistance between this study and (237) are similar 28% and 22% respectively, however, the proportion of isolates with resistance has been identified as increasing over the study period and therefore this should continue to be monitored over time. Higher levels of resistance have been described for pneumococcal isolates associated with carriage rather than disease for penicillin (70%), clindamycin (46%), and erythromycin (78%) (240) and could be due to the increased repertoire of resistance determinants transferred between organisms during asymptomatic colonisation. Antimicrobial resistance data was not complete for all isolates in this dataset, therefore the proportion of isolates with the resistant phenotypes are only estimations of the whole pneumococcal population. For some antimicrobials such as oxacillin where there are only a small number of isolates with corresponding data, the proportion of resistant and sensitive phenotypes might change significantly. Serotype 19F was responsible for the highest proportion of resistance for all antimicrobials, with the exception of doxycycline, where 19A had higher proportions. The high frequency of non-susceptibility in serotype 19F was also present in (224). A major limitation in the interpretation of the observed change in resistance patterns over time is there is no data available to allow analysis of how antimicrobial prescribing rates in Singapore relate to these observed changes.

Within this dataset, a considerable level of diversity has been shown by large numbers of serotypes, sequence types and novel MLSTs. This reflects findings shown previously (72) and as isolates are collected over many years and levels of recombination are high in the species, some degree of variation was expected. In addition, Singapore is an area of high urbanisation with a dense population existing in close proximity. There are many ethnic backgrounds and easy mobility of individuals between countries. All these factors could assist in the genetic exchange of DNA between organisms and contribute to the high levels of diversity within this specific ecological area.

3.4 Conclusion and Future work

The large collection of *S. pneumoniae* isolates and supporting laboratory data provided the opportunity to perform a comprehensive study into changing epidemiology and resistance. As isolate collection was limited by differences in hospital policies, the serotype distribution represented here cannot be representative of total prevalence. In addition, resistance to antimicrobials can only be based on the available corresponding data. Despite these limitations it nevertheless remained one of the most representative studies performed from a single country.

Some serotypes showed an increased OR to IPD when compared to non-IPD and many of these, for example serotype 8, 4, 20 and 7A also appear to predominate in adult IPD. Of particular concern are serotypes 19A and 3 as, in addition to being more associated with IPD, they also appear to be increasing in the proportion of disease despite their inclusion in PCV13. Vaccine coverage of PCV13 was maintained at ~60% and PPV23 at ~70%. The majority of resistance was from serotype 19A which has shown a decreasing trend in the proportion of infection over time. This serotype however, was identified to be more associated with non-IPD therefore, the decrease could be affected by a reduction in non-IPD isolates collected during the latter period of the study. In the Singapore study (224) serotype 19A was identified to be largely responsible for drug non-susceptibility and has also been identified in other countries with predominantly penicillin non-susceptibility (223). Although not the case here, serotype 19A was shown to be increasing in the proportion of disease it was causing across the study period therefore, resistance in this serotype must be monitored.

Chapter 4 Detection of recombination from clinical isolates

4.1 Introduction

Tests for association between specific genetic features and particular phenotypes are based on differences in genetic variation which can be introduced through mutation or by horizontal transfer of genetic material, for example recombination. Specific genotypes are causally and statistically associated with a phenotypic trait when samples taken from a natural population are analysed by association mapping (244). Genome wide association studies (GWAS) have been used since 2004 in human association studies however, it is a relatively new method in molecular bacterial research. The GWAS approach to identify bacterial resistance determinants is limited because of inherent factors attributed to bacteria rather than humans; bacterial genomes are clonal as a result of mitosis at every generation, which results in substantial linkage across the genome referred to as linkage disequilibrium. Other than an increase in alleles maintained by positive selection, some genotypes present in populations increase due to their linkage to these loci, a process known as genetic hitchhiking (245).

The pneumococcus has shown its success in being able to perform rapid and considerable adaptation through horizontal transfer of genetic material, and this can occur through transformation, transduction and/or conjugation. Genetic transformation is the phenomenon in which cells are able to take up DNA from the environment and incorporate it into their genome (201), and was first observed in the bacterium *S. pneumoniae* (246). It involves the acquisition of exogenous DNA from the surroundings, followed by integration into the host genome (247). Recombination in the form of acquisition and incorporation of genetic elements is not restricted to its own species in the naturally competent *S. pneumoniae*, and differences in DNA sequence between donor and recipient can be as much as 25-30% (247), substantially amplifying the heterogeneity of the common gene pool. During recombination, double-stranded exogenous DNA binds to the cell membrane of the competent cell where it undergoes single stranded nicks (248). Using the endonuclease *EndA*, one of the single stranded DNA molecules is transported across the membrane with 3’ – 5’ polarity while the other strand is degraded (249, 250). Fragments of internalised DNA are then embedded in a nucleoprotein complex (251) and become integrated into the host chromosome at regions of similar sequence.

Some species of bacteria such as helicobacter, have high levels of pan-genomic variation (252) which can lead to the sub-structuring of distinct strains within the population (244, 247). In pneumococcal populations, there is variation in the population structure of individual genes across the genome, including the most essential genes (253, 254). Although these variable population structures reflect their evolutionary histories, it supports the observations that high rates of recombination are present in all genes including the most conserved, and therefore, recombination may dilute or eliminate the phylogenetic signal necessary to identify relationships between strains of a population (253). Although many phylogenies rely on estimations from core alignments, these observations show that at any one genome region, it is not clonal descent that is represented, it is the average of highly variable population structures at specific loci (253). A GWAS identifies association with a phenotype by the comparison of cases which lack the phenotype, with controls that exhibit the phenotype. Population structures can influence the identification of variation attributed to phenotypes in GWAS because allele frequencies occurring naturally between cases and controls could be due to systematic ancestry differences. In these cases, if a GWAS was performed, they would be identified as spurious associations. Minimisation of the spurious associations by controlling for population stratification, maximises power to detect true associations. Previous GWAS studies have tried to reduce the effect of spurious associations by using a hierarchical and spatial clustering model to identify clusters within datasets and then account for this in the association test (179, 255). This study aims to identify SNPs associated with antimicrobial resistance in a population of *S. pneumoniae* isolates from Singapore, however the influence of recombination in doing this remains unknown. To investigate the effect of recombination on population structures critical in GWAS, areas of the genome in which recombination has occurred must first be identified.

4.2 Recombination in bacteria

Recombination occurs despite the potential risk of disrupting existing regulatory and protein interaction networks in the recipient cell (256). Recombination rate is in part attributed to the level of environmental stress encountered by the organism as transformation of DNA has been shown as beneficial in stressful environments, but costly in otherwise benign environments (257). The inherent differences in the rate of recombination influence the overall effect that recombination has on the population structure of bacterial species. The ‘r/m’ (recombination/mutation) value is the proportion of polymorphisms accumulated from the import of sequence by recombination relative to natural mutation (71). This can be calculated and used to compare differences in bacteria. *Mycobacterium tuberculosis* (MTB) have a low r/m value for example, 0.486 as recombination occurs rarely (258). In organisms where recombination is more frequent, the r/m value for an isolate can encompass a wider range as some lineages within the species are known to have higher rates of recombination than others (71, 259-261). In *Helicobacter pylori* this has been described as 0.3 – 109.7 (259) and in *S. pneumoniae* 0.06 – 34.06 (71, 202, 262). Unlike MTB, which is contained within the macrophage for much of the infection cycle, *S. pneumoniae* is a commensal of the nasopharynx and therefore inhabits the same niche as many other taxa facilitating inter and intra species mobilization of genes (263). Recombination in *S. pneumoniae* can occur not only during colonisation of the respiratory tract, but also during polyclonal infection (264) or during biofilm formation (265). This has led to *S. pneumoniae* having the potential for highly variable genomes and diverse populations (266). This is thought to be advantageous in some organisms as it allows long-term survival of clones and rapid adaptation in response to environmental changes (77).

4.2.1 **Recombination in *S. pneumoniae***

In the pneumococcus, recombination is the main method for horizontal transfer of genetic material. Recombination occurs whilst the organism is exhibiting a 'competent' state early in logarithmic growth (267). Proteins involved in this process are regulated by quorum sensing, responding to changes in pneumococcal population density (268). The extracellular hormone competence stimulating peptide initiates cells to coordinate a number of processes including; their entry into competence which can last up to 40 minutes; differential regulation of genes responsible for uptake and integration of DNA; and finally the production and release of autolytic enzymes to kill neighbouring cells, a process known as fratricide (269, 270). The dedicated system responsible for the acquisition of environmental DNA in the pneumococcus is highly coordinated and regulated.

Within subpopulations of the same species, the rate of recombination is variable (69, 247), and does not seem to correlate with genetic relatedness of isolates (271). Information from a large Massachusetts, USA dataset of *S. pneumoniae* has shown sequence type ST320 to have the highest recombination (77). Some lineages have the propensity to either donate or receive DNA more than others, and generally organisms that lack a capsule show a higher rate of both compared to encapsulated isolates (69). Although there is variation between pneumococcal lineages (69), an average of 72 mutations are introduced into the genome in every recombination event (73). The size of the recombination fragments imported into cells have been shown to differ through the literature, but generally shorter fragments of similar sequence are optimal for transformation (201). These micro-recombination events ranging from 0.03-3kb (201, 270, 272) up to ~6-8kb (73, 202, 273-275) are transferred more frequently. The pneumococcus can also carry out macro-recombination involving much larger fragments of DNA ranging from ~30kb up to

235kb (73, 78, 151, 264, 276, 277). The occurrence of these events are rarer in comparison but can be associated with major phenotypic change such as serotype switching to evade vaccine (272).

Recombination does not occur uniformly across the genome but instead occurs in hotspots around genes involved in responses to selection pressures such as antibiotic utilisation, host immune responses and vaccine evasion (69, 73, 146, 149, 150, 278). Examples of genes include those that encode cell surface antigens; *pspA*, *pspC*, those associated with increased pathogenicity; *LytA*, *Ply*, *nanA* (279, 280), and those that are associated with antibiotic resistance; *pbp1a, pbp2b*, *pbp2x*, *folA* (69, 146-148). These can have long-term evolutionary consequences that alter resistance profiles. *S. pneumoniae* isolates have been shown to be capable of hyper-recombination (150), and it is believed the significantly higher levels of recombination in the 368 isolates tested was responsible for elevated levels of resistance to penicillin, erythromycin, tetracycline, chloramphenicol and cefotaxime (150). It is likely the transfer of short fragments from *S. mitis* and *S. oralis* are vital in facilitating the generation of mosaic genes seen in pneumococci, such as in the genes encoding penicillin binding proteins (281, 282). With high rates of recombination present in the pneumococcus, recombination is likely to be present in most pneumococcal populations, however, the extent of change to the genome may vary. The genomes of six isolates collected from a single patient over a seven month period had recombined 7.8% of its genome (264) whereas 74% of the genome had been altered by recombination in a global sample of 240 PMEN1 (ST81) isolates (73).

4.3 Detection of recombination

Developments in next generation sequencing now allow rapid generation of whole genome sequence data, which can be used to reconstruct phylogenetic patterns and provide insight into the mechanisms of evolutionary change. Recombination is a dominant force of genetic variation and therefore, it is crucial to have bioinformatic approaches with sufficient ability to detect genomic regions affected by recombination.

4.3.1 Programs available for recombination analysis

There are a number of publicly available software programs that identify areas of recombination in bacterial genomes. When the current study commenced (2015), popular methods included ClonalFrameML (208) and Gubbins (205). None of these programs were able to overcome fundamental problems with recombination detection. For example, it is likely that true levels of recombination will be underestimated because some events occur between highly similar loci making them undetectable using such techniques. In addition ancient recombination events that occurred before divergence to species level will be underestimated due to the accumulation of additional point mutations over time (283).

4.3.1.1 The Genealogies Unbiased by Recombination in Nucleotide Sequences (GUBBINS)

This software is designed to identify recombination events in closely related species, for example strains of *S. pneumoniae* belonging to the same sequence cluster that have been densely sampled (205), but its use has extended to successfully study other bacteria (284). Isolates within the same sequence cluster would normally result in isolates sharing serogroup or clones as defined by MLST. The program estimates the 'background SNP density' as the probability that a SNP occurs at a single genomic location within such similar organisms. This is the total number of SNPs identified in whole genome sequencing, divided by the overall size of the genome. Following this, a sliding window approach is employed to scan and evaluate nucleotides across the whole genome. The SNP density in each sliding window is compared to background SNP density to identify regions containing a significantly higher number than is expected by chance (205). These regions will be described as recombination regions in the program output as the elevation in SNPs would not have been generated by spontaneous mutations. Phylogenetic and sequence reconstruction methods such as FastTree and/or RAxML can be incorporated and further flexibility of the algorithm is that it can be applied to full genome alignments without the need to remove accessory loci (205).

4.3.1.2 ClonalFrameML

This software performs inference of recombination within bacterial genomes. The initial step is to construct a maximum likelihood tree from the dataset which is taken to be the initial clonal genealogy (208). At internal nodes generated from clonal genealogy, the ancestral sequences and any missing base calls in the observed sequences are reconstructed by maximum likelihood (285). Following this, to obtain maximum likelihood estimates of recombination parameters and branch lengths of clonal genealogy, a Baum-Welch Expectation –Maximisation algorithms is used. At all sites the maximum likelihood importation status is inferred using a Viterbi algorithm, and finally bootstrap methods quantify any uncertainty in the parameters (208).

4.4 Application of recombination programs on 2,059 *S. pneumoniae* isolates

The majority of datasets in which isolates undergo recombination analysis in the literature are small and/or consist of isolates that are closely related which requires no prior consideration regarding processing. For species that are naturally more diverse, or for larger datasets of bacterial populations, many studies adopt the method of subgrouping the dataset into closely related isolates and then perform recombination analysis individually on these groups (69, 286). This is because none of the described programs can handle very large or highly diverse datasets. This methodology does give useful insight into what recombination is present in the dataset, however, may not necessarily capture all recombination events if present between subgroups analysed separately. The study performed here aimed to identify all recombination within the dataset. As none of the described programs can perform recombination analysis from all isolates it was instead necessary to analyse subsets and then combine results to assess all recombination. Isolates were randomly assigned to subgroups to capture and analyse different combinations of isolates.

Although the Gubbins algorithm recommends the processing of similar isolates, the program claims to remain specific in the detection of recombination even when these conditions are not met (205). This was proved by similar recombination events being found by both ClonalFrame and Gubbins following the analysis of eight *H. pylori* isolates, an organism well known for its diversity, despite no prior processing into individual groups (205). A comparison of recombination analysis of 11 *S. pneumoniae* PMEN1 (Spain23F ST81) isolates was also performed and found only slight discrepancies in the number of identified recombination events, and a very similar phylogeny reconstruction between Gubbins and ClonalFrame (205). Both examples suggest no real difference in recombination events identified between programs. The capacity of the program may be linked to the test organism and the level of diversity within the dataset. Neither of the previous comparisons are reflective of this dataset in terms of size or diversity, therefore it remains unknown whether one program may perform better or is more suited to perform recombination analysis on the dataset of Singapore isolates. The success of Gubbins being able to process the *H. plylori* isolates (205) gives confidence that the methodology implemented will allow the comparison of programs in the detection of recombination from a large, and diverse dataset.

One way that genome wide association studies are used is to identify associations between mutation and phenotypes such as antimicrobial resistance, with phylogeny being used to account for the effect of population structure. As evolution of pneumococcal populations are dominated by recombination, phylogenies using whole genome data will be heavily distorted in favour of representing the effect of recombination rather than mutation. Deep rooted phylogenetic signal present in *S. pneumoniae* datasets can be eliminated when recombination levels are high (247) and this may affect associations identified during the GWAS. The aim of this study is to generate a phylogeny based on mutation alone, therefore it is necessary to remove this confounder from all isolates to accurately cluster the population based on vertical inheritance alone. No detailed comparison of recombination detection has been performed on a clinical dataset as large and diverse as this thus far and this will form the initial investigation.

4.5 Results

4.5.1 Determining dataset size to compare Gubbins and ClonalFrameML

In order to use either program to run recombination analysis, it was necessary to divide the dataset of 2,059 *S. pneumoniae* isolates into smaller subsets. Once the upper limit in subset size was determined, output between programs was compared. A simulated sequence containing a region of artificially introduced recombination was constructed for inclusion into subsets. This was included in all subsets and detection of the known region indicated program performance. Recombination analysis by Gubbins and ClonalFrameML was performed on identical subsets of increasing size until a difference was seen in the processing capacity of one of the programs (see section 2.2.2). Figure 11 shows ClonalFrameML recombination analysis completed for all subset sizes within 10 hours. Gubbins on the other hand required a much longer run time to carry out recombination analysis on the same datasets. It was only the smallest subset size of 11 isolates that completed in the allocated run time. Only 4/100 runs containing subsets of 21 isolates completed and no subsets of 31 isolates completed.

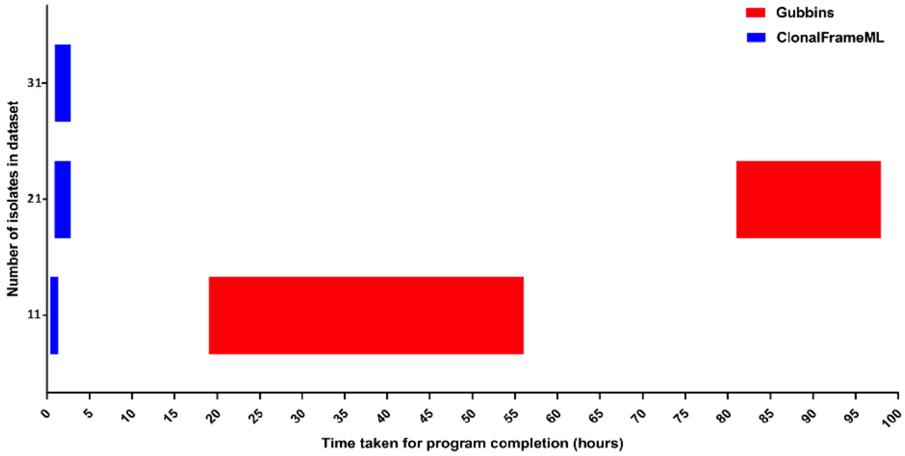


Figure 11 Duration of recombination analysis for subset sizes of 11, 21 and 31 isolates for Gubbins (red) and ClonalFrameML (blue). All ClonalFrameML results are representative of 100 runs for each subset size. Subsets of 11 isolates for Gubbins are representative of 100 runs. Subsets of 21 isolates for Gubbins is only representative of four runs, the rest did not complete. No runs of 31 completed with Gubbins

4.5.2 Comparison of program output

Reliable comparison of recombination programs can be performed by testing the ability to detect a known region of artificial recombination located on a sequence present in all subsets. Subsets of 11 randomly selected isolates completed recombination analysis in both programs therefore, this was the subset size used for all future analysis in the comparison of Gubbins and ClonalFrameML.

4.5.2.1 Detection of artificial recombination from datasets consisting of clinical isolates

Of the 100 subsets processed by each program (section 2.2.2), both ClonalFrameML and Gubbins were able to identify recombination in 77/100 subsets. In most cases, when recombination was recognised, the exact position was correctly identified. The first SNP of the recombination region was missed in three subsets, but this was consistent between programs. Investigation of the 23 subsets in which recombination was not found revealed that these consisted only of isolates located in the outer clade of the phylogenetic tree highlighted in Appendix B.

The level of divergence between each clinical isolate in the subset and the recombination sequence was calculated (section 2.2.4). This value represented how similar the isolates in the subset were to the recombination sequence, and the most similar sequence with the lowest divergence score was the representative minimum dissimilarity value for each subset plotted (Figure 12). In subsets with a high divergence score of >0.6, both programs were not able to identify the known area of recombination, whereas when the divergence score was low <0.2 recombination was found (Figure 12). The minimum dissimilarity score was also plotted against recombination outcome following analysis of 21 and 31 isolates by ClonalFrameML. The same relationship of recombination being detected in subsets with a low divergence score of <0.2, and not detected in subsets with a comparatively high divergence score of >0.7 was identified (Figure 13).

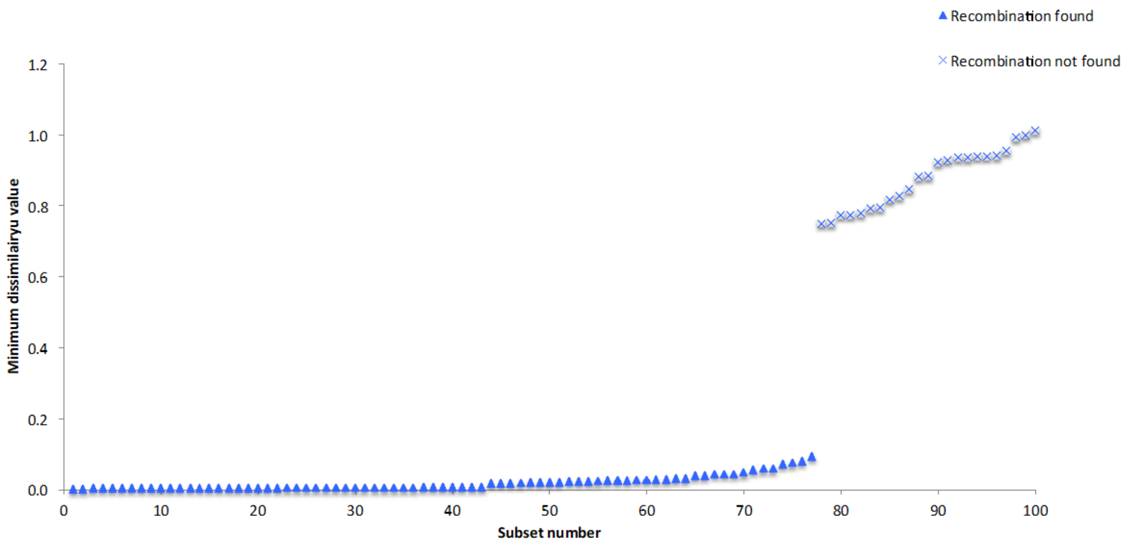


Figure 12 Minimum dissimilarity value of the closest isolate to wildtype in 100 subsets of 11 processed by ClonalFrameML and Gubbins. Symbol represents the outcome of recombination detection. Same result seen in both Gubbins and ClonalFrameML

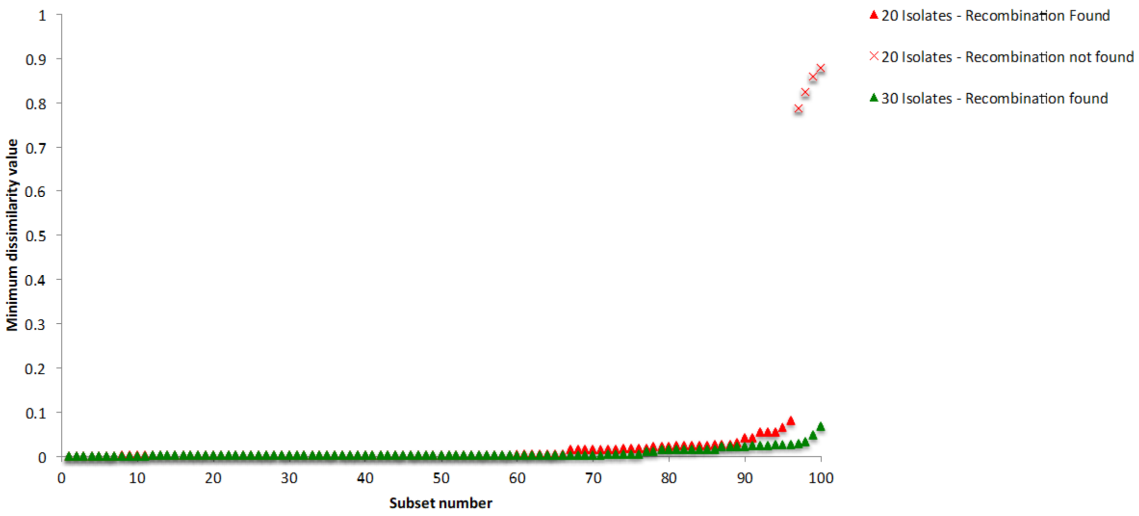


Figure 13 Minimum dissimilarity value of the closest isolate to wildtype in 100 subsets of 21 isolates (red) and 31 isolates (green). Outcome of recombination detection represented by symbol. Datasets of 21 and 31 isolates were selected independently of one another and results presented are only for ClonalFrameML.

4.5.2.2 Detection of artificial recombination in different mutation backgrounds

To test whether the diversity of isolates within subsets affected the ability to detect recombination, a background mutation rate was incorporated into the recombination sequence (section 2.2.1), and the ability of programs to detect this region assessed. The degree of diversity between each isolate in the dataset and the reference sequence was calculated (section 2.3) and showed a difference of 0.005% - 1.3% (average = 0.75%). To encompass this, the range of mutations incorporated in simulations was between 0.001% and 2%. The position of the newly mutated sequences in relation to other isolates within the phylogenetic tree is shown in Figure 3. The sequence with 0.001% mutation rate was located in the very centre of the inner clade of the phylogenetic tree. As the mutation rate of the sequences increased further to 0.01% and 0.1%, sequences became positioned at a greater distance from the centre of the inner clade but remain within it. Sequences with a mutation rate of 1% and 2% are different enough to result in them diversifying on another branch from the rest of the isolates in the dataset. The sequence with 1% mutation rate was located halfway up this branch, and the sequence with 2% mutation was located at the very tip (Figure 3).

Table 9 shows ClonalFrameML successfully detects recombination in all mutation backgrounds, but that Gubbins loses the ability to detect recombination in a 1% and 2% mutation background. To better understand why Gubbins was not able to detect recombination when this level of mutation was present, minimum dissimilarity values of subsets were calculated for all mutation backgrounds. Figure 14 illustrated at 1% and 2% mutation there is no longer a similar sequence present in the subset, shown by the rise in dissimilarity values from 0.001 to 0.01 and 0.02 respectively.

A similar sequence was added to the simulated dataset to test whether recombination became detectable in a sequence of 1% and 2% diversity when this was included. This proved to be the case and both programs detected recombination in all mutation backgrounds (Table 9). The dissimilarity values for the subsets were calculated and found that the presence of a similar sequence maintained low minimum dissimilarity scores at 1% and 2% of 0.001 (Figure 14).

In all subsets tested so far, the area of recombination had a higher mutation rate than what was present on the rest of the genome (Appendix D). To test if programs were also able to detect recombination as an area of low mutation relative to the chromosome, recombination analysis was performed on a final simulated dataset (Appendix E). It was known that a similar sequence was required in the dataset for recombination to be detected in the higher background mutation rates therefore, this was accounted for in dataset construction. Recombination was detected from all mutation backgrounds by both programs (Table 9).

Table 9 Summary of recombination detection by Gubbins and ClonalFrameML from simulated and clinical datasets. Simulated datasets[1] have no similar sequence present. Simulated datasets[2] have a similar sequence present and recombination is seen as high mutation in a chromosome of no mutation (Appendix D). Simulated datasets[3] have similar sequence present and recombination is seen as no mutation in chromosome of high mutation (Appendix E)

Background mutation rate (%)	**Detection of artificial recombination from simulated datasets[1] (n=100) by Gubbins**	**Detection of artificial recombination from simulated datasets[1] (n=100) by ClonalFrameML**	**Detection of artificial recombination from simulated datasets[2] (n=100) by Gubbins**	**Detection of artificial recombination from simulated datasets[2] (n=100) by ClonalFrameML**	**Detection of artificial recombination from simulated datasets[3] (n=100) by Gubbins**	**Detection of artificial recombination from simulated datasets[3] (n=100) by ClonalFrameML**	**Detection of artificial recombination from datasets composed of clinical isolates (n=100) by Gubbins**	**Detection of artificial recombination from datasets composed of clinical isolates (n=100) by ClonalFrameML**
0.001	Yes	Yes	Yes	Yes	Yes	Yes	100	77
0.01	Yes	Yes	Yes	Yes	Yes	Yes	100	77
0.1	Yes	Yes	Yes	Yes	Yes	Yes	92	0
1	No	Yes	Yes	Yes	Yes	Yes	0	0
2	No	Yes	Yes	Yes	Yes	Yes	N/A	0

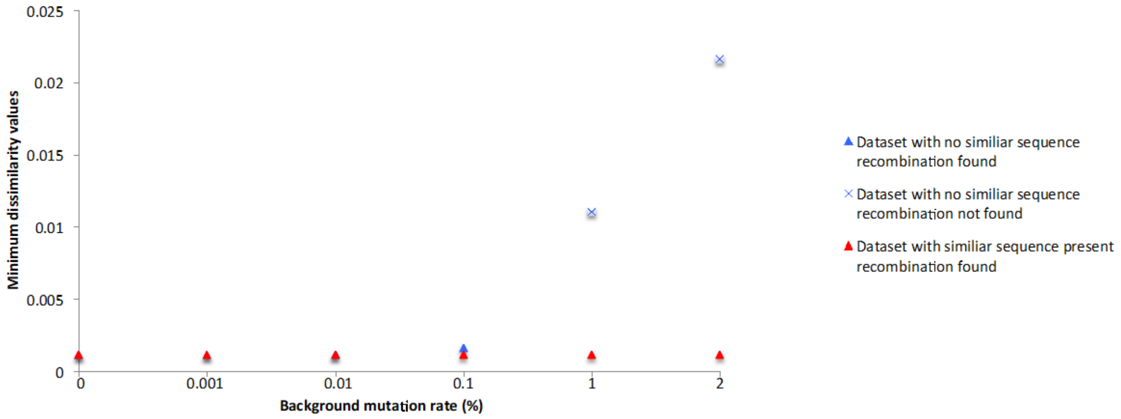


Figure 14 Minimum dissimilarity values in subsets containing simulated sequences of increasing background mutation rate. Blue values are from subsets that do not contain a similar sequence in the recombination analysis. Red values are subsets with a similar sequence to the recombination sequence. Blue data points are present under the red in 0%, 0.001% and 0.1% background mutations.

The final test which determined whether there was an effect on recombination detection when mutation was present in the sequence was to extend recombination analysis to datasets of real rather than simulated isolates (section 2.2.2). Table 9 showed ClonalFrameML was able to detect recombination in 77/100 subsets in the lower mutation backgrounds of 0.001% and 0.01%. This was the same 77/100 subsets that positively identified recombination when there was no mutation present in section 4.5.2.1. Further increase in the mutation background resulted in recombination not being detected. Gubbins was able to identify the artificial recombination in all subsets when there was a mutation rate of 0.001% or 0.01% present. The complete recombination was identified in 77/100 runs, which was also true for ClonalFrameML, and the remaining 23/100 runs were able to identify approximately 8.4kb of the 10kb area. In a 0.1% background Gubbins identified recombination in 91/100 subsets. Again, the complete block is identified in 77/91 subsets, and the remaining 14 found between 5kb and 7kb of the recombination. Like ClonalFrameML, Gubbins was not able identify recombination from any subset when the level of background mutation reached 1%. At 2% no results could be obtained as processing was not completed within the designated 100 hour run time.

4.5.2.3 Internal consistency in recombination calls between Gubbins and ClonalFrameML

Internal consistency between programs can be assessed by comparing output of recombination analysis from identical datasets. The data in Table 10 described the combined recombination analysis of all recombination from 100 subsets processed by both Gubbins and ClonalFrameML and compared them to a study that previously used Gubbins for recombination analysis on *S. pneumoniae* (73). Gubbins identified recombination in 99% of the reference genome which was considerably higher than the 71% identified by ClonalFrameML. In the comparative study (73), Gubbins called 74% of the genome in recombination which was much closer to the proportion ClonalFrameML identified in these combined subsets. The average size of recombination also notably differed by roughly 10-fold between programs. ClonalFrameML identified a higher number of smaller recombination events averaging ~300bp whereas Gubbins identified fewer recombination overall, but they were much larger at ~3500bp (Table 10).

Internal consistency was further investigated by specifically identifying recombination called on the artificial recombination sequence. This was the internal control and was present in the analysis of all subsets. Combined data from the analysis of 100 subsets identified recombination in 86% (1,911,701bp) of the sequence from Gubbins whereas recombination was only identified in 15% (336,930bp) from ClonalFrameML. Of the genome proportion called in recombination by ClonalFrameML, only 555bp were not also called by Gubbins showing a very good overlap between programs (Figure 15). The recombination identified included the 10kb region artificially introduced. To assess the false positivity rate of programs all-natural recombination was removed (as described in section 2.2.6), which left only the artificial region, and analysis repeated. ClonalFrameML saw a reduction in bp called in recombination from 336,930bp to 10,003bp and Gubbins a reduction from 1,911,701bp to 11,840bp. Excluding the artificial region of recombination, this meant ClonalFrameML incorrectly identified only 3bp in recombination and Gubbins 1,840bp.

Table 10 Summary statistics following recombination analysis by Gubbins and ClonalFrameML of 100 subsets consisting of 10 randomly selected clinical *S. pneumoniae* isolates and the artificial recombination sequence. Gubbins results from a comparative study by Croucher *et al* (2011) (73) also included for reference

	ClonalFrameML recombination analysis	**Gubbins recombination analysis**	**Gubbins recombination analysis of 240 *S. pneumoniae* PMEN1 strains (73)**
Proportion of reference genome that has undergone recombination in at least one isolate	71%	99%	74%
Number of recombination events	275,636 (average per run = 2,756)	83,636 (average per run = 836)	702
Size of recombination events (bp)	1-19,080 (mean = 372)	3-119,300 (mean = 3,695)	3 - 72,038

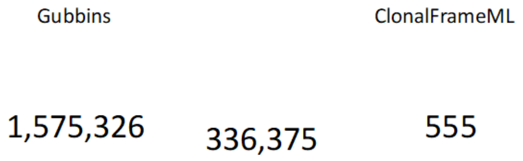


Figure 15 Venn diagram illustrating level of internal consistency in recombination calls (bp) between Gubbins and ClonalFrameML from 100 subsets consisting of 10 randomly selected clinical *S. pneumoniae* isolates and the artificial recombination sequence. A total of 1,911,701bp was identified as recombination in Gubbins and 336,930bp in ClonalFrameML

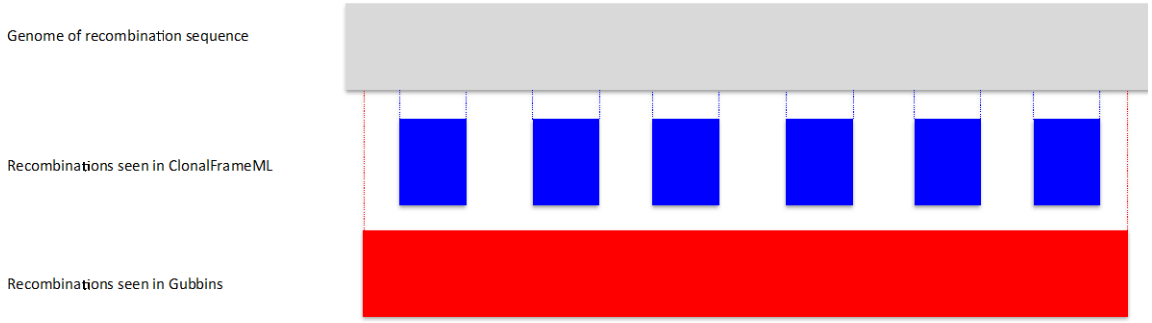

Genome of recombination sequence
Recombinations seen in ClonalFrameML
Recombinations seen in Gubbins

Figure 16 Example of recombination called across the genome by Gubbins and ClonalFrameML. The total number of recombination bp called by ClonalFrameML for all six recombination events was 4,633bp. The total number of recombination bp called by Gubbins in the recombination even shown was 24,820bp.

4.5.2.4 Determining the distance at which recombination programs distinctly identify two adjacent recombination events

The pattern of recombination calls between programs is illustrated by Figure 16. ClonalFrameML identified multiple smaller recombination events positioned across the genome, whereas Gubbins identified a single larger recombination event. To test program ability to distinguish between adjacent recombination events, a new recombination sequence was created that systematically varied the distance between two recombination events and assessed the distance at which programs identified them individually (section 2.2.1). Simulations were performed using a range of recombination block sizes and Figure 17 shows that the size of recombination does not affect the distance in either program. Gubbins overall required a much greater distance of ~9kbp to separate adjacent recombination events whereas ClonalFrameML required only ~2.6kbp. To further this, analysis was carried out on simulated datasets that had varying levels of background mutation to test whether diversity in the sequence affected the distance required between recombination events (section 2.2.1). As it had already been confirmed that the size of recombination had no effect on distance, only recombination events of 10kbp were tested and the range of diversity matched that previously tested. Figure 18 showed ClonalFrameML continued to require ~3kbp to identify adjacent recombination blocks until the background mutation rate reached 1%. At 2%, ClonalFrameML was not able to separate recombination events within the distances tested. The distance required by Gubbins also remained at approximately

9kbp in background mutation rates up to 0.1%. There was a slight reduction in distance in the latter mutation background to ~8kbp and any further mutation resulted in no recombination being detected by Gubbins.

The final scenario in which to test the distance required to distinguish between recombination events was to use a subset of clinical isolates rather than simulated sequences (section 2.2.3). Recombination analysis by ClonalFrameML identified recombination events in 77/100 subsets and a distance of 300bp was required by all subsets to distinguish between individual recombination events. Gubbins was only able to complete processing in 70/100 subsets within the allocated time. Recombination was identified in 56/70 subsets and identified distinct recombination events at a distance ranging from 8.8-9.3kbp (mean=9.2). The 14/70 subsets in which recombination was not identified was consistent with those not identified in ClonalFrameML.

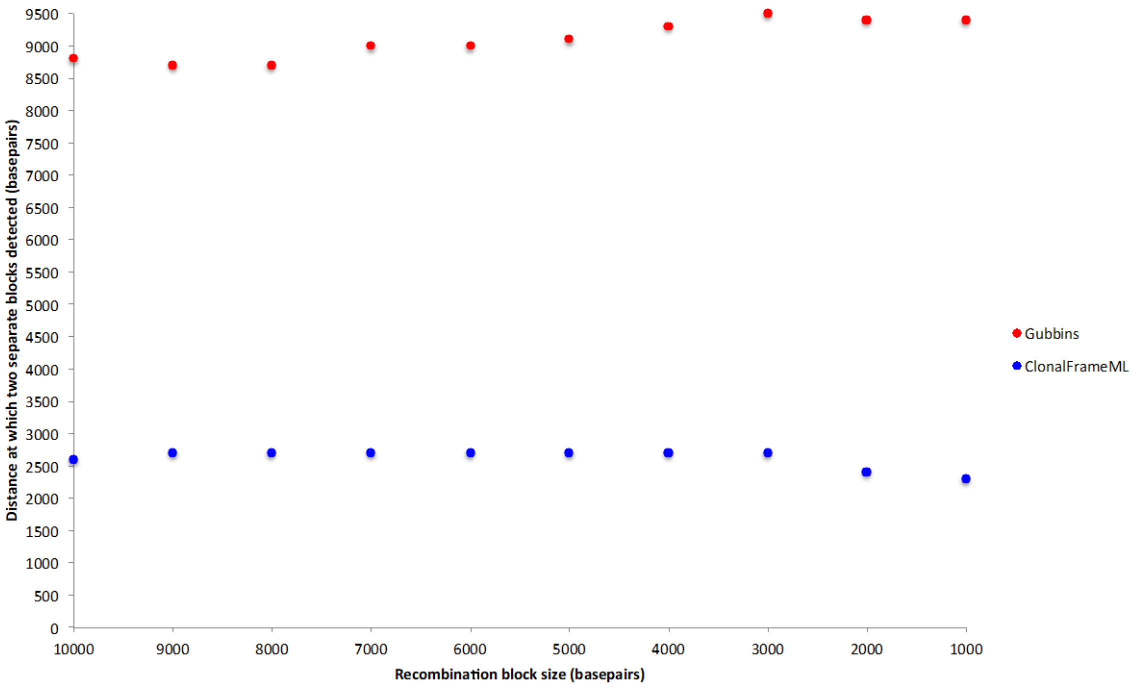


Figure 17 Illustration of the distances required by programs to distinguish between two recombination events when the size of the recombination event ranged from 1kbp to 10kbp

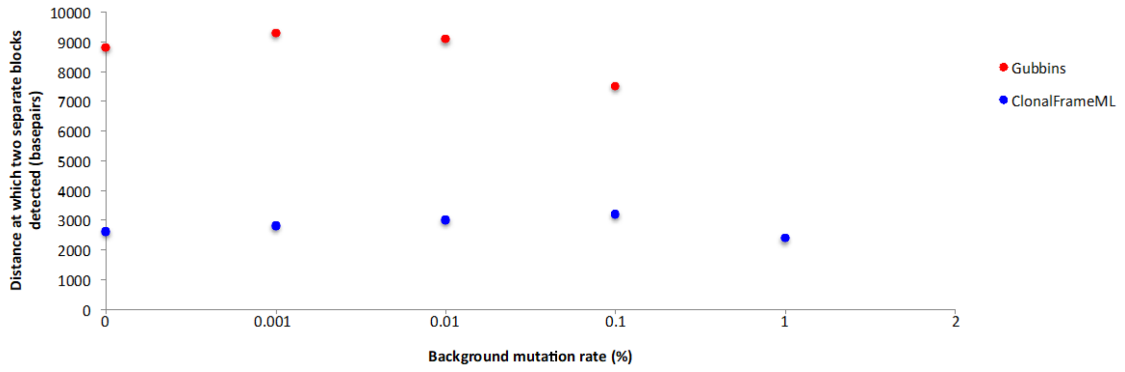


Figure 18 Distance between recombination events required by programs to identify two recombination events in the presence of background mutation. Size of recombination event in all subsets is 10kb. Gubbins was not able to identify recombination events in a 1% or 2% mutation rate (not plotted). ClonalFrameML could not separate recombination events within the tested ranges at 2% (not plotted)

4.6 Discussion

Diverse sets of bacterial isolates pose issues for recombination programs, which have been previously overcome by splitting datasets into related subgroups prior to analysis. In this study, isolates were randomly allocated into subsets prior to recombination analysis so that analysis was not limited within specified groups and maximum recombination could be identified between all isolates. However, to do this it was necessary to compare the utility of widely used tools for the detection of recombination, namely Gubbins and ClonalFrameML under the parameters that would eventually be used. The program BratNextGen (287) was not included in the comparison of recombination programs as it required some manual processing steps and therefore complete automation was not possible. To compare these tools, both simulated isolate sets with and without seeded known recombination, and clinical datasets were used. ClonalFrameML was able to process larger subsets of isolates, and this is preferable as the presence of a similar sequence to that containing recombination is required for detection. Results presented here also show that Gubbins connects recombination events over much larger proportions of the genome, which results in more recombination being identified in comparison to ClonalFrameML. If Gubbins was used to identify recombination, and recombination was removed to estimate the effect on population structure, then a greater proportion of the genome will unnecessarily be excluded. This could include important information regarding phylogeny required to determining population structure.

4.6.1 Capacity of program to detect recombination from a diverse dataset

A simulated sequence containing a region of known recombination was used as a positive control to compare programs and validate the method as the regions of ‘true’ recombination remained unknown in the clinical isolates. The sequence was a good representation of isolates within the dataset as it was the most frequently isolated serotype/MLST (23F/ST81). After subsets of various sizes were tested (section 4.5.1), the only size capable of tandem analysis for program comparison was subsets of 11 isolates, as Gubbins was not able to complete processing of larger subsets in the allocated run time. Previous comparison of recombination analysis run-time does not find this same limitation, in fact the opposite was seen. In Croucher *et al* (2014), Gubbins completed in 48.4 seconds, whereas it took ClonalFrame 705.4 hours to complete analysis (205). The dataset tested was drastically different to the dataset used in the current study in that it analysed closely related isolates. This clearly demonstrated that the processing capabilities of recombination programs can be greatly affected by the level of diversity within the dataset.

Initial recombination analysis on clinical isolates showed each program performed equally well with both identifying recombination in 77/100 subsets. When recombination was not detected in a subset of 11 strains, the 10 clinical strains were not closely related to the 11th strain carrying the recombination. This was evidenced by the 10 strains being placed in the outer clade of the phylogenetic tree, whereas the 11th strain was placed in the inner clade. This observation led to the hypothesis that a similar sequence to that containing the recombination was required to be present within the subset for the artificial recombination to be found. To test this, values of the dissimilarity matrix used to make the phylogenetic tree were used as a covariant of recombination called, against recombination not called. These dissimilarity values represented divergence as greater distance on the tree is because of a higher divergence. The smallest of these values, termed the minimum dissimilarity value, was plotted in Figure 12 and Figure 13. This showed a clear distinction in the minimum dissimilarity value between subsets in which neither program was able to identify recombination and those in which both programs were able to identify recombination. The probability that a similar sequence is included in a subset is related with dataset size. The more strains present in the subset, the more likely a similar sequence would be present, and this in turn could lead to the detection of recombination. Figure 13 supports this theory as when ClonalFrameML was run on subsets of 30 isolates rather than 10, the minimum dissimilarity value remained low and recombination was detected in all 100 subsets.

The natural diversity of sequences within the dataset was known to be as much as 1.3% therefore, simulations on subsets with differing degrees of diversity was run to test its effect on recombination detection. In a basic test dataset Gubbins lost the ability to detect recombination in a 1% and 2% mutation background and it was hypothesised this was because the sequence containing recombination became too different to others in the subset. The location of the mutated sequences supported this theory as they formed a separate branch and were no longer positioned in the inner clade (Figure 3). Once a similar sequence was added to the subset, and a low dissimilarity score maintained, Gubbins was then able to detect recombination in these mutation backgrounds. ClonalFrameML performed better in simulated datasets as the level of background mutation did not affect the ability to detect recombination. Due to the range of diversity within isolates, it is biologically plausible for DNA to be transferred from a sequence with high diversity to a sequence of low diversity, or from a sequence of low diversity to a sequence that is more diverse. The ability of programs to detect recombination events from both scenarios were tested and programs performed equally well.

Until this point the ability of recombination programs to detect known recombination from datasets containing clinical isolates (section 4.5.2.1), and in simulated datasets with an incorporated background mutation rate (section 4.5.2.2) had been tested. These were then combined to ensure recombination could be detected in sequences with a background mutation rate when run with subsets of randomly selected clinical isolates. This was reflective of the true relationship between isolates in subsets when recombination analysis was performed on the Singapore dataset. In this scenario ClonalFrameML was only able to detect recombination when present in a mutation background of 0.001% and 0.01%. Gubbins was able to detect recombination from datasets in a 0.001%, 0.01% and 0.1% mutation background. The complexity in the relationship between clinical isolates within a subset was greater than that of simulated sequences as each clinical isolate had unique and varying levels of diversity from one another. Due to this, the minimum dissimilarity parameter cannot be used alone to predict recombination detection in these datasets.

4.6.2 Comparison of internal consistency between recombination programs

It was shown that there was high internal consistency of recombination calls between ClonalFrameML and Gubbins, which gave more confidence to the recombination events being true. Far more recombination was called by Gubbins than ClonalFrameML and this occurs not just in a single subset but in multiple subsets. In Gubbins the identification of recombination as regions of elevated densities of base substitutions can be confounded in datasets containing a high diversity of sequences, and this could be the reason for the comparably elevated levels of recombination in Gubbins. It has been shown that diversity can affect recombination detection, and that a similar sequence was required for detection. Therefore, specificity was tested by randomising base positions to remove clustering of mutations, while still maintaining the overall tree topology, overall nucleotide sequence divergence and overall allele frequencies at each polymorphic position. Such a decrease in recombination found outside of the known area showed that the signal for recombination was successfully removed. There may have been a slight reduction of specificity in the calls made by Gubbins rather than ClonalFrameML because it did still identify some excess recombination (1,840bp). Results from both programs showed <0.1% recombination called outside the artificially introduced area proving the increase in calls by Gubbins was not a consequence of it having a higher false positivity rate.

Further investigation comparing recombination output by Gubbins and ClonalFrameML illustrated major differences in size and number of recombination events between programs (Table 10). Gubbins data produced here was compared to that following the analysis by Croucher *et al* (2011) of a different pneumococcal dataset (73) to understand whether results are simply reflective of Gubbins analysis or if they are dependent on the dataset tested (Table 10). The comparative dataset (73) contained 240 isolates PMEN1 isolates, which were likely to be more similar to one another than the 10 randomly selected isolates in each of the subsets tested in this study. A much higher level of recombination was witnessed from this dataset (99%) than in (73) (74%), and the higher diversity between isolates could be responsible for this. In addition, the frequency and size of recombination events encompassed a much larger range in this dataset than described in (73). A study that directly compared recombination output in a similar fashion to that performed here showed Gubbins identified a smaller number of recombination events than ClonalFrame; 28 rather than 48 (205) and a stark difference in the frequency of recombination events was not identified. The isolates in which recombination analysis was performed in (205) were closely related to one another which was very different to that studied here. These differences imply it is likely recombination statistics are related to the diversity of the dataset tested, which makes accurate comparison of recombination data between studies difficult.

The difference in recombination block size between Gubbins and ClonalFrameML in this analysis prompted the exploration into the positions of recombination events in relation to one another. The example in Figure 16 represented findings throughout the data and showed overcall could be from the connection of multiple adjacent blocks from Gubbins, whereas ClonalFrameML identified these distinctly. Gubbins has been previously shown to identify irregular mosaic recombination events rather than the individual segments (205) and this may be the reason for this finding within the dataset. Incorrect definition of recombination boundaries is an important source of error and could account for why the size of recombination events in Gubbins were so much larger than those seen in ClonalFrameML. This hypothesis was tested in simulated datasets by identifying the number of bp required by each program to identify individual recombination events when they are closely positioned in the genome (section 4.5.2.4). Figure 17 confirmed Gubbins required a much larger distance to distinguish between recombination events, and until a distance of ~9kbp was reached Gubbins inaccurately called a much larger area of recombination than was present in the genome. ClonalFrameML identified the separation of recombination events across a much shorter distance and was therefore more accurate in recombination calls. Analysis was extended to include simulated datasets with varying levels of background mutation as it was a better representation of the clinical isolates within the dataset and results would be more applicable to real recombination analysis. The effect of mutation on distance required to identify individual recombination events was similar to the dataset with no similar sequence present. Recombination was identified and there was very little difference in distances for either program as mutation rate increased to 0.1%. As was seen previously, a further increase in mutation lead to recombination not being identified in Gubbins and this was likely due to the lack of a similar sequence being present. ClonalFrameML was able to detect individual recombination events in a 1% mutation background, however at 2% mutation recombination it was no longer possible for ClonalFrameML to identify the two separate recombination events.

It was identified that Gubbins called more of the genome in recombination than ClonalFrameML and confirmed using simulated datasets that this could be due to Gubbins connecting areas of recombination over much larger distances. It has also been shown that the detection of recombination can be influenced by mutation rate, and the similarity of other sequences within the dataset. Although the distance required to distinguish between recombination events in the range of mutation backgrounds present within our dataset was tested, results from this may not be applicable to datasets containing only clinical isolates. In real rather than test datasets, each isolate will be unique in terms of mutation and this could affect the distances at which recombination events are connected. Analysis on clinical isolates saw that once again ClonalFrameML identified recombination in the same 77/100 subsets identified in Section 4.5.2.1 but that the distance required to call individual recombination events in these clinical subsets reduced from ~2,800bp seen in simulated subsets to 300bp. This reduction was consistent in all 77 subsets showing an improvement in the ability to accurately distinguish between adjacent recombination events when processing real datasets. Gubbins was only able to identify recombination in 56/70 subsets because analysis for 30 subsets did not complete in the designated 100 hour run time. The reduction in distance required by ClonalFrameML following the processing of clinical samples was not mirrored in Gubbins. As was the case in simulated datasets, the distance required to separate adjacent recombination events in clinical datasets remained at ~9kbp. The findings support the theory that if natural recombination events are positioned in close proximity in the genome, it is far more likely that Gubbins will connect these over much greater distances than ClonalFrameML, resulting in more of the genome being incorrectly called in recombination.

4.7 Conclusion and future work

The practicalities and suitability of using two well described recombination programs to perform analysis on a large clinical dataset of diverse isolates was investigated. The method chosen to perform this comparison is novel in that it randomly selected isolates for analysis rather than relying on prior classification from MLST, serotype or phylogeny data. It is the first in-depth comparison of Gubbins and ClonalFrameML using this method and encompasses both simulated and clinical sequences.

The performance of recombination analysis was investigated in relation to diversity in both simulated and real datasets and analysis highlighted the requirement of a similar sequence to successfully identify recombination in ClonalFrameML and Gubbins. If at least one similar sequence was not present in a randomly selected clinical dataset, recombination may be missed. In some instances, Gubbins could use low level mutation in clinical datasets to increase levels of recombination detection, however, the program may not be as accurate in its recombination calls from these datasets. It has also been demonstrated that ClonalFrameML can overcome this limitation by its ability to perform recombination analysis on larger, randomly selected datasets, increasing the likelihood of sequence inclusion. For Gubbins, dataset size remained a limitation when isolates were selected in this way.

In comparable datasets Gubbins calls recombination across more of the genome than ClonalFrameML. It has been shown that this was not due to a lack of specificity in Gubbins, but instead was due to the connection of adjacently positioned recombination events over larger areas in both simulated and clinical datasets. Results presented here show that if Gubbins is used to generate a recombination-free phylogeny, then a greater proportion of the genome will be unnecessarily excluded. This could include important information regarding phylogeny. Consequently, ClonalFrameML was deemed the more suitable program for application to the Singapore *S. pneumoniae* clinical isolate dataset.

Since the comparison between Gubbins and ClonalFrameML was initiated the program FastGEAR was released which described the ability to detect recombination from both external origins and between the inferred lineages (253). Although this was not included in program analysis, Mostowy *et al* (2017) compared recombination analysis between FastGEAR (253), Gubbins, ClonalFrameML and STRUCTURE. Results showed FastGEAR detected recombination events in simulated datasets well, particularly from full alignments and that it was able to detect ancestral recombination well (253). There was a simulated range of diversity tested in the comparison, however, the performance of individual programs was not compared following analysis of clinical samples. The diversity of some isolates may contribute to errors in recombination calls therefore it cannot be said with certainty that the same level of accuracy would be seen in real datasets. Also, the simulated datasets in (253) consisted of 30 sequences, however when Gubbins was used to perform recombination analysis on a dataset of that size in this work, it became too computationally demanding. This is suggestive that the diversity within the Singapore dataset is considerably higher than that tested in (253) and therefore the ability of recombination detection may not be comparable. FastGEAR (253) would need to be performed on both simulated datasets and subsets of clinical isolates to test its ability to detect recombination events. This additional comparative analysis would provide information regarding its suitability of detecting recombination from large datasets of diverse clinical isolates.

Chapter 5 Antimicrobial resistance phenotypes analysed by Genome Wide Association study

5.1 Introduction

Studies of phylogeny often suggest horizontal transfer of genetic material to be the main way that clinically relevant antimicrobial resistance determinants are acquired, a theory supported by much experimental evidence (69, 73, 288). However, the study by Lehtinen *et a*l (2020) showed that this was not the only contributor that determined the frequency of antimicrobial resistance, and the origin of some mechanisms of resistance in *S. pneumoniae* have been demonstrated from one, or the accumulation of a number, of spontaneous mutations likely to have arisen as a consequence of environmental selection (289). Mutations that result in variants of proteins can remain in populations, providing the amino acid substitution does not negatively affect survival. The use of antimicrobials provides a selective pressure for mutations associated with resistance and this phenomenon highlights dangers in antimicrobial overuse. The work in this chapter aimed to investigate the genetic signatures in the populations of *S. pneumoniae* isolates from Singapore to map the patterns and distributions of single nucleotide polymorphisms (SNPs) that generate the antimicrobial resistance phenotype, by performance of a genome wide association study. To do this accurately, areas of the genome in which recombination occurred were identified and their influence in defining population structure prior to GWAS determined.

5.1.1 Application of genome wide association studies outside a human host

Methods adopted in human GWAS have been adapted for application in bacterial genomes to facilitate the discovery of novel mutations associated with phenotypes. Results from such studies have helped to inform disease management and treatment for many human pathogens. For example, much research has been performed to identify pathogen genetic determinants related to clinical phenotypes of pathogens. Virulence is a complex trait and studies have used a genome wide approach to identify loci that directly affects toxicity, as well as epistatically interacting loci, in *S. aureus* (178). Specific loci such as Panton-Valentine leucocidin locus have been associated with increasing the odds of generating the specific type of disease known as pyomyositis from *S. aureus* (290). Associations with disease traits in other organisms include extra-intestinal virulence determinants in *Escherichia coli* (291), and novel factors associated with invasiveness in *Streptococcus pyogenes* (292). Another complex trait investigated from bacterial genotypes is antimicrobial resistance. Again this has been performed for a number of different clinically important organisms in which resistance is a cause for concern such as *Mycobacterium tuberculosis* (177, 180, 181, 293-296), *S. aureus* (175, 294), *Plasmodium falciparum* (297), *E. coli* (294, 298), *Klebsiella pneumoniae* (294) and HIV (299).

In addition to bacteria, variation in the host genome that contributes to disease has been investigated through the application of a genome wide approach. Host variation that effects the host-pathogen interaction has been investigated in HIV infection to identify significant SNPs associated with HIV-1 variants and mapped to the human leukocyte antigen regions (172, 300, 301). Application of GWAS has been useful in understanding farm animal host specificity of *Campylobacter jejuni* infection and highlighted association genes involved in the biosynthesis of vitamin B5 (176). Finally the host genetic factors that contribute to *Neisseria meningitis* have been identified such as variants in complement factor H, that offers protection in childhood infection (302), and in carbonic anhydrase X (303).

Previous GWAS applied specifically to pneumococci have highlighted a variety of factors contributing to disease. Investigation of both bacterial and human variation in pneumococcal meningitis identified pneumococcal genetic variation contributed 70% to the invasive potential of the organism, and variations in the genes *pspC* and *zmpD* contributed significantly to this (173). Identification of genes or gene variants that are involved in invasiveness is extremely useful as this reveals alternative candidates for future pneumococcal vaccines. Interestingly, it was shown that the bacterial genotype had no effect on the severity of disease and that, instead, human genetics contributed one third in determining disease severity and half to the susceptibility to pneumococcal meningitis. (173). Additional studies identified human host-associated variants in long intergenic non-coding RNAs have been associated with pneumococcal bacteraemia (304). It is well established that pneumococcal carriage is a prerequisite for disease (305) and therefore, a detailed understanding of the factors that affect carriage and/or duration of carriage has the potential to provide opportunities to interrupt this stage in infection and reduce the burden of disease. A study into this highlighted 63% of the variation in carriage duration in the nasopharynx is dependent on bacterial genomic variation rather than previous carriage by the host or age of the host (306). Pneumococcal associations which impact mortality as a clinical endpoint include arginine biosynthesis genes (307), and the phage derived gene *pblB,* which is involved in platelet binding (308). Finally, GWAS have identified genetic variation associated with serotype 1 pneumococci to modulate tropism to central nervous system tissues, increasing virulence for meningitis (309).

5.1.2 Considerations of microbial GWAS

The aim of association studies is to identify significant associations between single nucleotide polymorphisms (SNPs) and a measured phenotype, and GWAS does this across the entire genome. These studies require three elements: a sufficiently large study population that effectively provides genetic information regarding the research question (310); single polymorphic alleles which can be genotyped to adequately cover the whole genome (311); and appropriate and accurate analytical methods which have sufficient power to allow identification of the statistically significant genetic associations in an unbiased fashion (312). Population structures within bacterial populations of interest may contain subgroups of isolates that are on average more related to each other than to other members of the wider population (58). Bacterial population structure has the potential to increase the false positivity rate in tests of association and to decrease power of GWAS (313). This is because bacterial population structure and kinship indicate covariance between individuals based on genetic similarities and heritability of the phenotype. However, tests of association assume statistical independence between individuals. Ignoring this covariance results in a deflated p-value that does not form a uniform distribution under the null hypothesis in the test for association (314). Standard practice to account for these confounding factors is to infer population structure and kinship based on genome wide SNP data, and then, either to account for the effect in the test of association, or alternatively, to remove problematic individuals from the analysis (313). This stratification for population structure should minimise falsely positive associations that could be obtained by chance.

Generally, SNPs in core genome are used as units of measurement in GWAS (175, 177-179), however other methods available use gene presence or absence or the measurement of ‘n-mers’ to study both the core and flexible genome simultaneously (176, 315-317). Association techniques are broadly broken down into allele counting or homoplasy counting. Allele counting methods generate association signals from over-representation of an allele at a particular site in cases relative to controls. This method can be limited due to strong population structure and linkage disequilibrium (LD) in bacterial populations, as overall phylogeny is not considered when generating the association. There are a number of ways to correct for population structure. One method is to use an analytical test such as that used in the program Genomic Control to normalise all inflated p-values by single inflation factor λ (314). However, this approach may overcorrect, resulting in removal of all statically significant GWAS hits (180). Other less conservative methods identify subpopulations present within the overall population and then test for associations conditional on these defined subpopulations. This method of using ancestry to correct for population stratification is a popular approach and a variety of programs can infer subpopulations. Such programs include; BAPS (318), principle component analysis in EIGENSTRAT

(319) or multi-dimensional scaling in PLINK (320). These epi-clusters are then used as covariates in association testing, for example Cochran-Mantel-Haenszel test. If this approach is adopted, thousands of genomes are required as, with small sample size, this correction can reduce GWAS power significantly (180). A homoplasy is a shared characteristic between isolates that did not arise from a common ancestor and results in convergent evolution. Homoplasy counting methods count repeated and independent convergent mutations emerging at a higher rate on branches of cases relative to controls to generate evidence of association (180). Software that implements this method that is available to use in association studies includes PhyC (177) and ROADTRIPS (321). Population stratification, and linkage disequilibrium to some extent, is intrinsically accounted for using this method by its phylogenetic convergence criterion. Homoplasy counting requires a smaller number of events than allele counting to reach statistical significance, therefore, the decision of which to use could be based on available sample size. Homoplasy counting for example, would produce a much stronger signal than allele counting from a small sample size that contained a strong phylogenetic structure. Allele-counting methods theoretically are able to detect all convergent sites that would be detected from homoplasy counting, as well as non-convergent sites with a sufficiently large sample size. The review by San *et al* (2020) details well the wide range of tools used in bacterial GWAS and highlights there is not yet a standardised methodology (322). There is variation in phenotype classification, statistical tests to detect associations, methods to account for population structure and within the input for analysis.

5.1.3 GWAS studies of pneumococcal antimicrobial resistance

GWAS that have been performed in *S. pneumoniae* in order to identify areas of the genome associated with antimicrobial resistance have previously been performed (179, 255, 292). Results obtained offer a platform against which findings from this study can be compared. The study by Chewapreecha *et al* (2014) performed two independent association analyses on datasets of carriage isolates to identify SNPs associated with beta-lactam resistance (179). The datasets analysed comprised 3,085 and 616 isolates collected from Maela, Thailand and Massachusetts, USA respectively. Population structures within datasets were defined using a Bayesian clustering approach; Bayesian Analysis of Population Structure (BAPS) software (318, 323). The test of association was performed using the Cochran-Mantel-Haenszel (CMH) statistic to identify associations between specific variants and beta-lactam non-susceptibility conditional on the population structure clusters. Analysis was performed independently for the two datasets and associations commonly identified from both (301 SNPs) were selected for further analysis. Although these attempts were made to reduce genomic inflation factor and control for intrinsic population structure of bacteria, this was not fully achieved because inflation values of 2.56 and 3.76 were identified after the reduction (179) rather than the desired value of 1.

The study by Mobegi *et al* (2017) analysed four datasets: 1,680 disease and carriage isolates, 1,013 isolates from (179), a dataset from Nijmegen, the Netherlands (n=349) and the final one from children with sickle cell anaemia, USA (n=318) (255). The population clusters used to control for population stratification were again determined using BAPS software (318, 323) and the CMH correction statistic to test for associations between SNP and resistance phenotype based on these clusters. Non-susceptibility to a wider range of antimicrobials was tested including penicillin, cotrimoxazole, erythromycin, trimethoprim, ofloxacin and ciprofloxacin. Both studies (179, 255) used a threshold of >0.01 for the minor allele frequency, reported associations with p-value <0.01 and incorporated the Bonferroni correction for multiple comparisons. Both (179, 255) also use a single reference genome *S. pneumoniae* ATCC 700669.

The final study that previously performed a GWAS to determine resistance determinants in *S. pneumoniae* was by Lees *et al* (2016) (292). However, the methods adopted vary significantly to those described for (179, 255). Instead of adopting a SNP-based method which uses clustering algorithms based on core alignments and then stratify association tests based on groups of samples, it used sequence element enrichment analysis (SEER) using n-mers to discover associations with antibiotic resistance (292). The same isolates analysed in (179) were used for application of this method to measure mechanisms associated with resistance to beta lactams, tetracycline, trimethoprim, erythromycin and chloramphenicol. Population structure was corrected for by the generation of a distance matrix from a random subsample of n-mers. Metric multi-dimensional scaling was then performed which is equivalent to using principal components of the SNP matrix. This removed the need for SNP calling or core genome alignment (319, 324) and gave the same results as clustering core alignments SNPs using hierBAPS (292).

High levels of recombination have been shown to occur in *S. pneumoniae* which can be beneficial in GWAS as it may result in areas of the genome that are in linkage disequilibrium being broken up, potentially reducing false positive associations and boosting the potential power to discover causal variants. However, such high proportions of the genome being affected by recombination might also change how isolates are clustered within population structures and therefore the genetic associations identified may arise from recombination rather than mutation. Accurately accounting for bacterial population structures is critical in reducing spurious associations in GWAS, therefore, prior to performing the GWAS on antimicrobial resistance, this study aims to perform an accurate population correction independent of recombination, to cluster isolates based only on vertical inheritance. This can then be compared with population structures determined from whole genome data to determine the effect recombination had on the population structure of the Singapore dataset. Some programs like Gubbins claim to be able to generate a correct phylogeny taking into account the recombination, but it has been shown through the comparison of programs described in Chapter 4, that Gubbins was not able to process large or diverse datasets and results may contain some inaccuracies depending on the dataset analysed.

Following this, the present study aimed to perform an accurate GWAS to determine SNPs associated with antimicrobial resistance using a large dataset that should provide sufficient power to detect associations if they are present. This GWAS to identify associations to a range of clinically relevant antimicrobials was performed on a dataset of ~2,000 *S. pneumoniae* isolates, only collected from disease from a single geographical location. To date, analysis such as this has not been performed. Previous studies have used carriage only isolates, combined carriage and disease isolates, and combined isolates from a range of geographical locations in their analysis . Many of the same genes were identified to contain associations with antimicrobial resistance in the previous studies (179, 255, 292), and offer a platform against which findings from this study can be compared. Analysis of datasets from different geographical locations or from predominantly disease rather than carriage isolates could highlight rare or unique SNPs determined only in these populations.

5.2 Results

5.2.1 Defining population structure for genome wide association study

5.2.1.1 Recombination analysis of 2,059 clinical *S. pneumoniae* isolates

Of the recombination programs tested in Chapter 4, ClonalFrameML was deemed to be the most suitable program to carry out recombination analysis on this dataset. Recombination analysis was performed on 400 subsets of 100 randomly selected *S. pneumoniae* as described in section 2.5. The frequency at which individual isolates were present in the subsets analysed ranged from 7-35 and within this dataset 86% of the genome was called as recombinogenic within at least one isolate. The mean size of recombination was 359bp (range 1-21,058bp). Regions of recombination identified by ClonalFrameML were removed from the genomes of all isolates as described in section 2.5 and the phylogeny for the recombination-free genomes is shown in Figure 19. Comparison of this phylogeny to that from whole genome sequences (Figure 9) showed the two phylogenies are very similar in their structure and pattern of branching. Both show the right side to have far more branch points diverging into many descendants, and less branching and shorter branches on the left side of the phylogeny. The axis of one of the branches had been flipped in Figure 19 but both showed a single section of tightly structured decedents. The distance between the recombination-free and the whole genome phylogenetic trees were calculated using the Robinson Foulds (RF) metric which showed a similarity of 71%. A total of 366 core genes were identified from the dataset (section 2.8) and the subsequent phylogeny (Figure 20) shows a similarity of 70% to the whole genome phylogenies (Table 11).

Table 11 Quantitative assessment of similarity between phylogenies using the Robinson Foulds (RF) metric calculated by the Environment for Tree Exploration toolkit (210)

	Whole genome phylogeny	**Whole genome phylogeny**	**Whole genome phylogeny**
Recombination free phylogeny	Normalised RF distance 0.61	RF symmetric distance 2245	Frequency of edges found in both phylogenies (%) 71
Core gene phylogeny	Normalised RF distance 0.61	RF symmetric distance 2491	Frequency of edges found in both phylogenies (%) 70

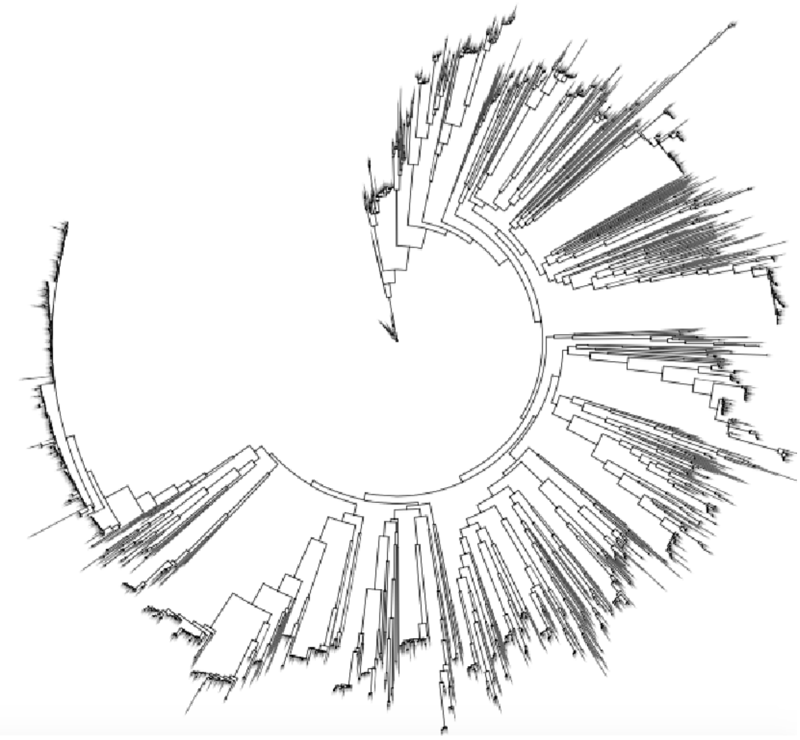


Figure 19 A maximum-likelihood phylogenetic tree constructed by FastTree using recombination-free genomes (14% of the genome) of 2,059 isolates of *S. pneumoniae*.

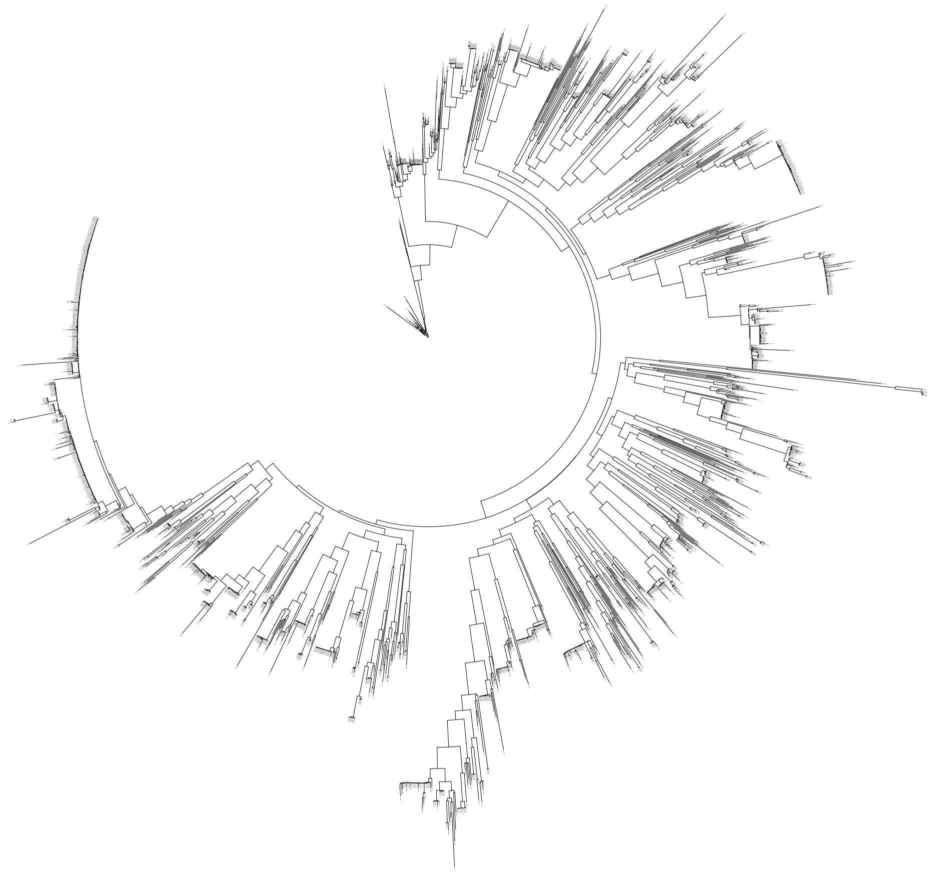


Figure 20 A maximum-likelihood phylogenetic tree constructed by FastTree using core genomes (366 genes) of 2,059 isolates of *S. pneumoniae.*

5.2.1.2 Application of programs to compare population structure between whole genome and recombination-free genome datasets

Supporting antimicrobial susceptibility data for a total of 1,828 isolates was available for use in the association study. The population structure of these isolates, based independently on both whole genomes and recombination-free genomes, was determined using popPUNK which clustered datasets into 226 components and 247 components respectively. The largest group present in both datasets consisted of 298 isolates.

Principal component analysis (PCA) was performed independently on datasets consisting of whole genome data and recombination-free genomes (section 2.6) to obtain larger clusters of isolates than identified by PopPUNK in which to compare isolate cluster allocation. Six clusters were identified from whole genome data and four clusters from recombination-free genome data (Figure 21). The isolates present in specific clusters were compared between datasets to identify the influence of recombination and high levels of consistency was seen in how they were grouped. Isolates present in individual clusters one, two and three of the recombination-free PCA correspond to clusters six, five and four of the whole genome PCA respectively. Isolates present in cluster four of the recombination-free PCA correspond to isolates present in clusters one, two and three of the whole genome PCA. Scree plots associated with individual PCAs (Figure 22) showed in both datasets the first 10 principal components (PCs) account for less than 60% of the variation in the dataset.

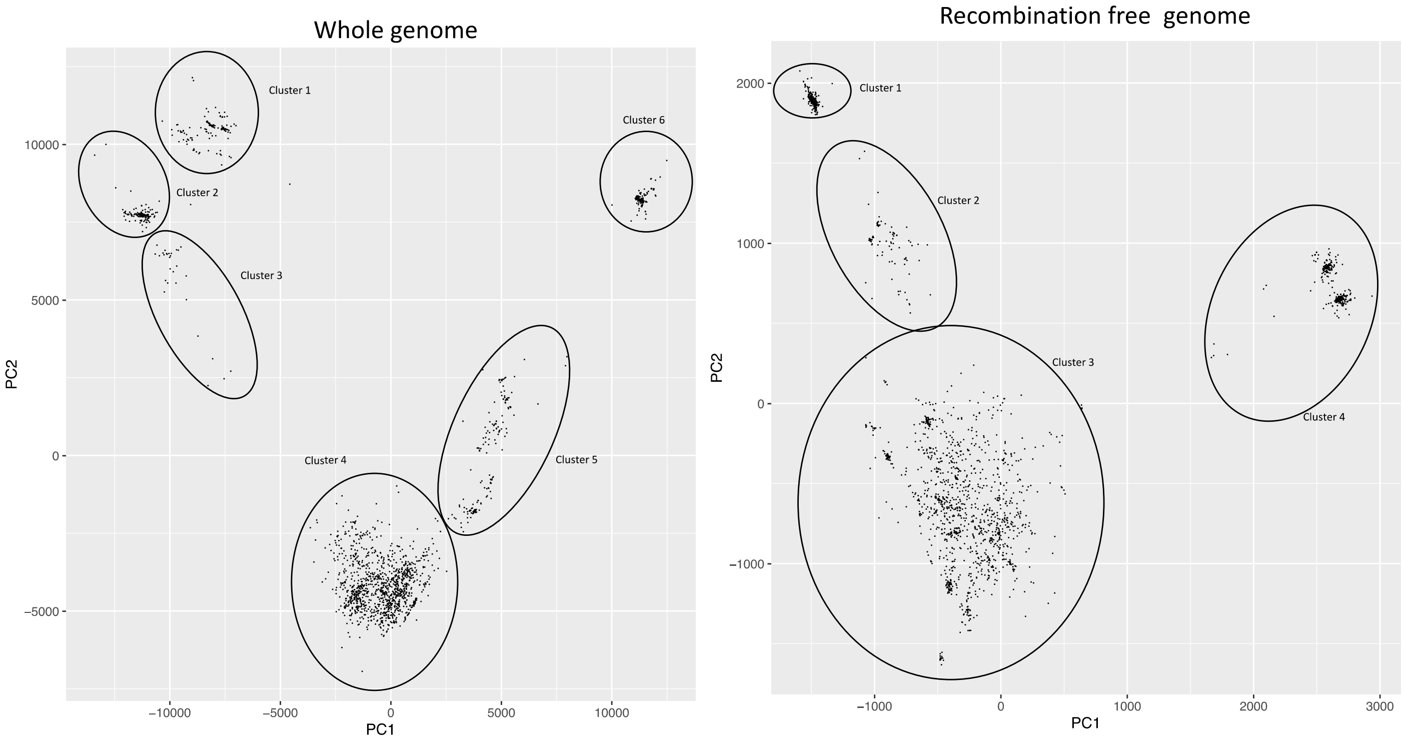


Figure 21 Principal component analysis showing PC1 and PC2 of 1,828 isolates of *S. pneumoniae*. Clusters of isolates are labelled 1-6 for the whole genome dataset and 1-4 for the recombination-free dataset. In whole genome data, cluster 1 (n=119), cluster 2 (n=171), cluster 3 (n=13), cluster 4 (n=1113), cluster 5 (n=164), cluster 6 (n=248). In recombination-free genome data, cluster 1 (n=249), cluster 2 (n=85), cluster 3 (n=1190), cluster 4 (n=304).

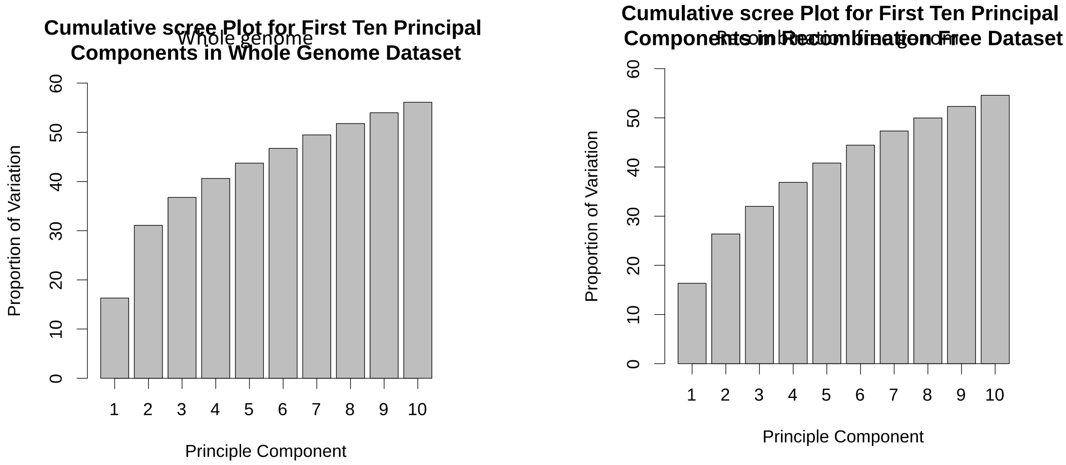
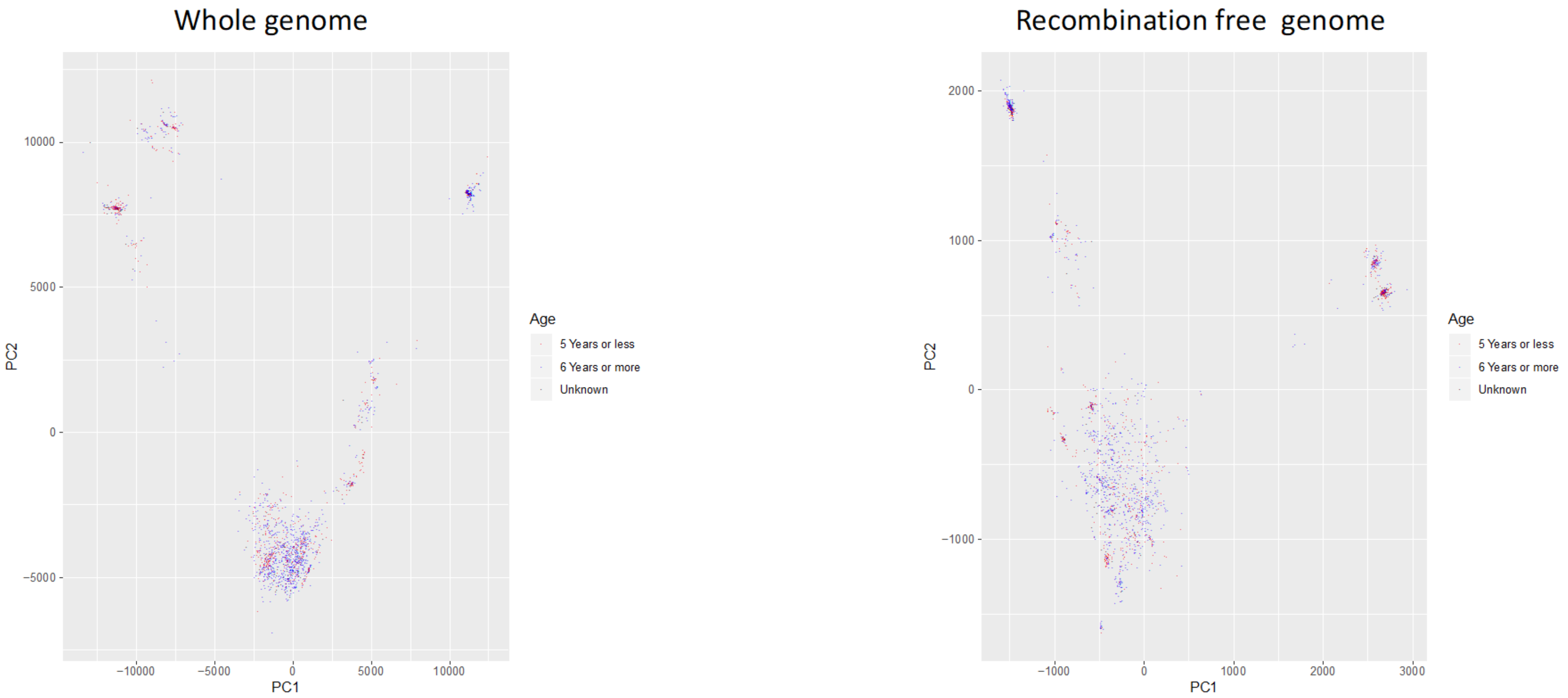


Figure 22 Cumulative scree plots of the proportion of variation for the first 10 principal components in the PCA analysis of whole and recombination-free genome datasets

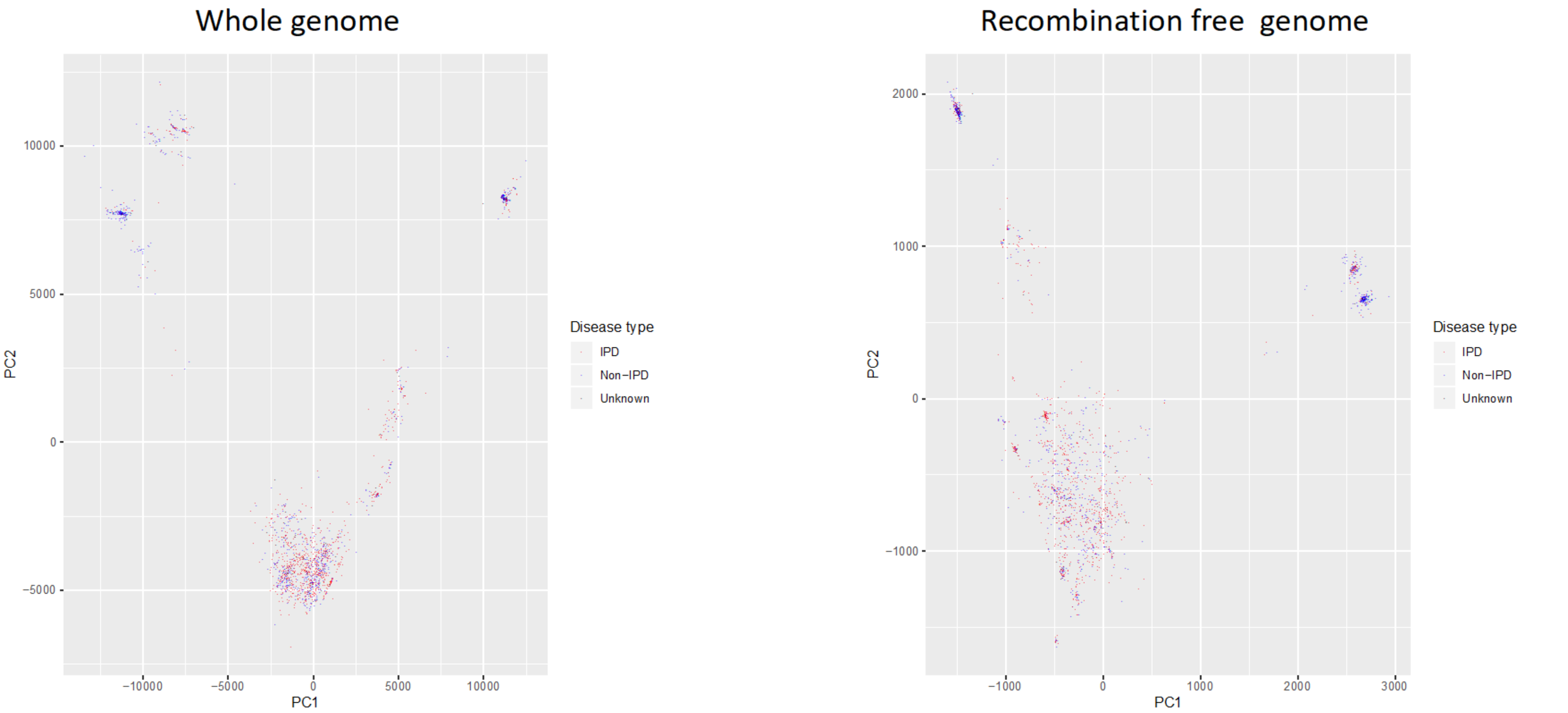
Metadata associated with isolates was used to colour PCA plots to highlight factors that could attribute to cluster allocation. Variables in age (Figure 23) and disease type (Figure 24) were well distributed amongst clusters. Figure 25 highlighted the position of isolates corresponding to the three main clonal cluster (CC) groups as designated by goeBURST and showed these are localised within specific clusters of the PCA. Isolates allocated to CC0 were present only in a single cluster of the PCA in both datasets. Similarly isolates of CC1 were separated from the main cluster and it is likely that differences in MLST loci were contributing to population clustering in the PCA.



Age group	Number of isolates (%)					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
≤5 years (n=658)	53 (8)	103 (16)	3	314 (48)	82 (12)	103 (16)
≥6 years (n=1037)	61 (6)	40 (4)	9 (1)	735 (70)	70 (7)	122 (12)

Age group	Number of isolates (%)			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
≤5 years (n=658)	104 (16)	33 (5)	362 (55)	159 (24)
≥6 years (n=1037)	122 (12)	47 (4)	757 (73)	111 (11)

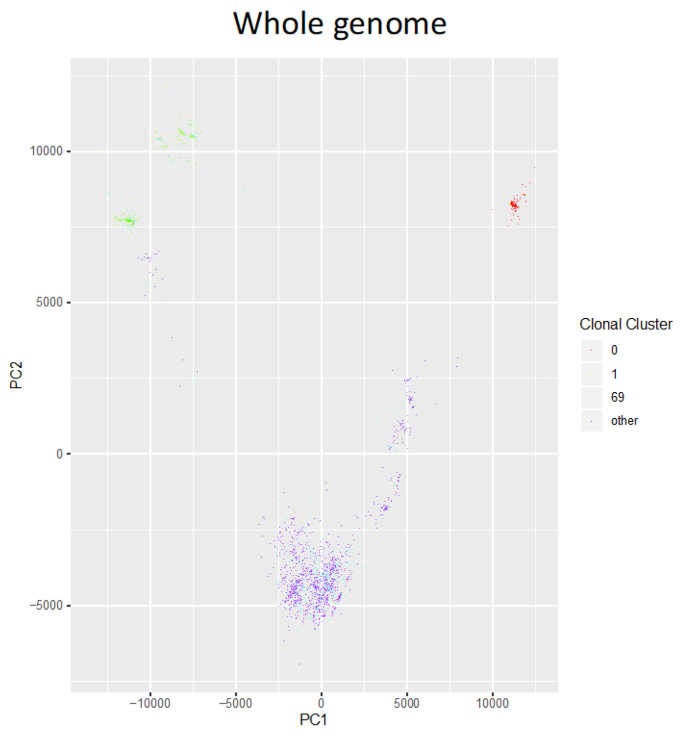
Figure 23 Principal component analysis showing PC1 and PC2 of 1,828 isolates of *S. pneumoniae*. The plot is coloured by age groups and corresponding tables show the proportion of isolates within each of the clusters described in Figure 21.



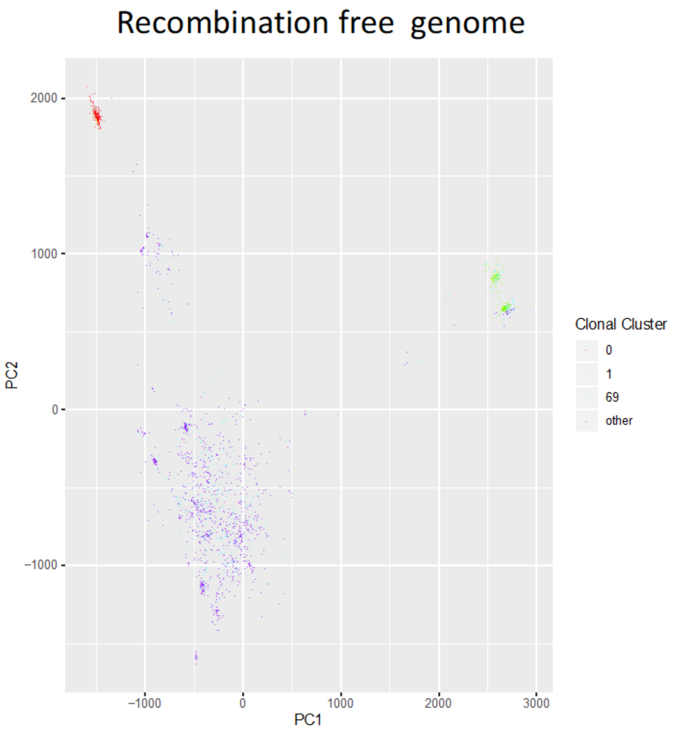
Disease type	Number of isolates (%)					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
IPD (n=930)	51 (5)	18 (2)	5 (1)	691 (74)	105 (11)	60 (7)
Non-IPD (n=819)	62 (8)	144 (18)	7 (1)	381 (46)	53 (6)	172 (21)

Disease type	Number of isolates (%)			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
IPD (n=930)	60 (7)	51 (5)	745 (80)	74 (8)
Non-IPD (n=819)	173 (21)	24 (3)	408 (50)	214 (26)

Figure 24 Principal component analysis showing PC1 and PC2 of 1,828 isolates of *S. pneumoniae*. The plot is coloured by disease type and corresponding tables show the proportion of isolates within each of the clusters described in Figure 21.



Clonal cluster group	Number of isolates					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
0 (n=241)	0	0	0	0	0	241 (100)
1 (n=234)	103 (44)	130 (56)	1	0	0	0
69 (n=242)	16 (7)	25 (10)	3 (1)	179 (74)	12 (5)	7 (3)



Clonal cluster group	Number of isolates (%)			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0 (n=241)	240 (100)	0	0	1
1 (n=234)	0	0	0	234 (100)
69 (n=242)	7 (3)	6 (3)	185 (76)	44 (18)

Figure 25 Principal component analysis showing PC1 and PC2 of 1,828 isolates of S. *pneumoniae*. The plot is coloured by the three main clonal clusters (CC) found within the dataset. Corresponding tables show the proportion of isolates within each of the clusters described in Figure 21

5.2.2 Genome wide association study of 1,828 isolates

The number of isolates present in clusters determined by popPUNK were too small to perform a meaningful GWAS. The grouping of isolates based on population structures using PCA provided larger clusters and allocation of isolates to the main cluster did not dramatically differ between whole genome and recombination-free datasets. The first few PCs however, did not account for the majority of variation within the dataset therefore, the use of these to correct for population structure was not suitable.

Given the failure of both PCA and popPUNK in cluster differentiation, a different approach was undertaken. Here, a genome-wide efficient mixed model analysis for association studies (GEMMA) was used to look for associations between SNPs and a non-susceptible phenotype which accounts for population structure by the production of a kinship matrix within the program (section 2.7). The inclusion of all isolates with supporting antimicrobial data, rather than individual analysis on smaller clusters identified in the PCA, was performed to increase power and to identify variants truly causative of the phenotype rather than those that are uniquely isolated with a single clonal lineage or cluster. Supporting antimicrobial susceptibility data for a total of 1,828 isolates was available and antibiotics with suitable numbers of cases and controls included penicillin, erythromycin, cotrimoxazole, clindamycin, chloramphenicol, doxycycline, and tetracycline.

5.2.2.1 Identification of SNPs associated with penicillin non-susceptibility from whole genome dataset

Penicillin susceptibility data was available for 1,670 isolates broken down into 1,002 sensitive (60%), 145 intermediate (8%) and 523 resistant (32%). This was the largest group of isolates with the intermediate phenotype therefore a GWAS using a range of phenotype classifications was conducted. The sensitive vs resistant classification (SvR) would identify SNPs with an association to the resistant antimicrobial phenotype, and the sensitive vs intermediate classification (SvI) would identify SNPs associated with an intermediate resistance phenotype. Combining the data for isolates with either an intermediate resistance or resistant phenotype in the classification sensitive vs Intermediate and resistant (SvIR) would identify SNPs associated with either an intermediate or resistant phenotype, and the increased numbers would maximise statistical power on the analysis.

5.2.2.1.1 Penicillin non-susceptibility associations from a sensitive vs intermediate and resistant (SvIR) GWAS classification

A GWAS to identify associations from the SvIR classification consisted of 1,002 controls with a sensitive phenotype (60%) and 677 cases (40%) with either an intermediate or resistant phenotype. A total of 63,570 SNPs was identified, and 120 significant SNPs associated with an intermediate or resistance phenotype. These were located in genes involved in peptidoglycan biosynthesis pathway (*pbpX*, *ponA*, *penA*), genes associated with the recombination pathway (*recU*), genes of the cell division pathway (*gpsB*), and transferases for cell wall biogenesis (*rlmL*, *rsmH*) (Figure 26). A QQ plot of p-values (Appendix I) showed the population stratification was successful using the kinship matrix as the p-values mostly follow the reference line and inflation only occurs for SNPs with high -log10 (p-values). The associated lambda GC score of 0.72 (Appendix I) however, revealed the study was underpowered, therefore it might not be possible to identify all associations present in the dataset.

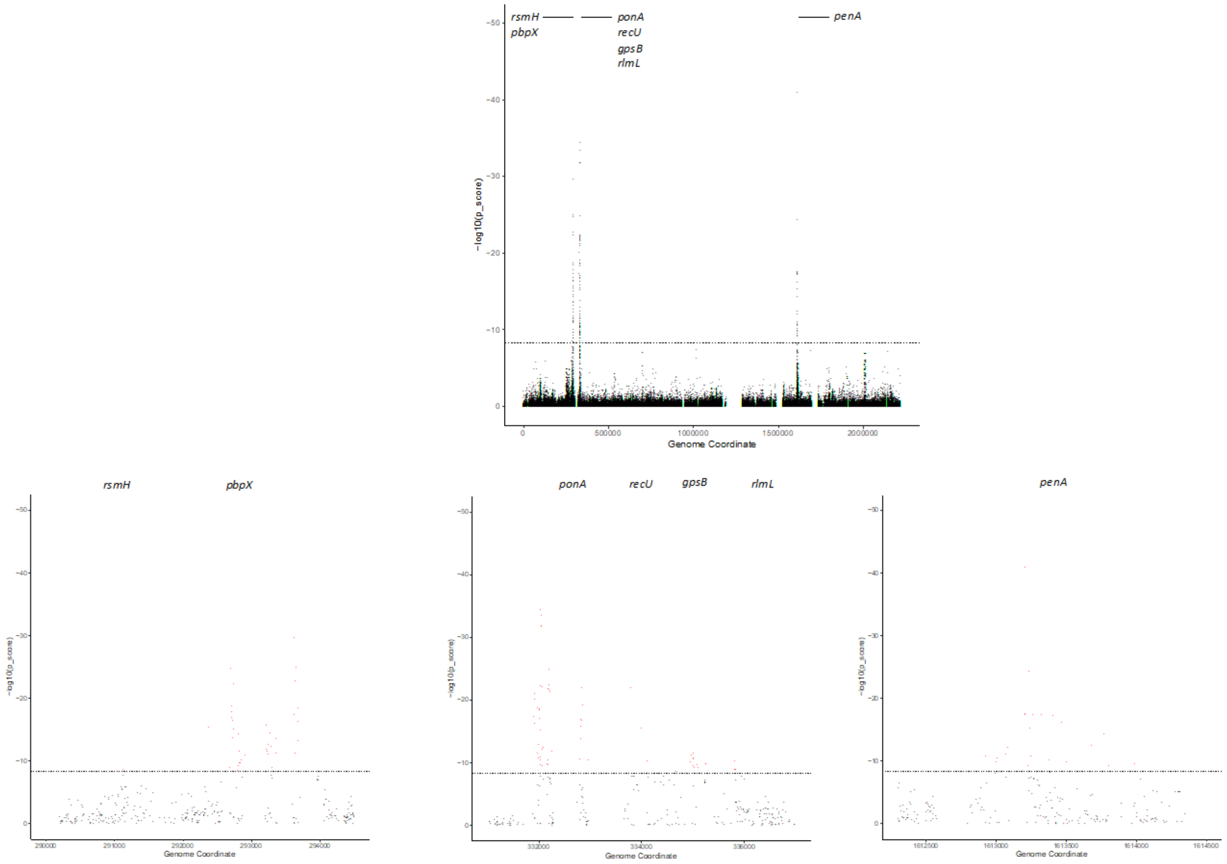


Figure 26 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvIR analysis. Genes that are specifically associated with penicillin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction

5.2.2.1.2 Penicillin non-susceptibility associations from a sensitive vs intermediate (SvI) GWAS classification

To identify SNPs only associated with an intermediate resistance phenotype to penicillin, a GWAS using the SvI classification was performed consisting of 1,002 controls with a sensitive phenotype (87%) and 145 cases with an intermediate phenotype (13%). In total 67,699 SNPs were identified, of which 138 were significant. The QQ plot (Appendix I) showed good control of population structure and even though the number of cases was smaller, the lambda GC of 0.75 was slightly better than seen for the SvIR classification (Appendix I). Like in the SvIR classification, association SNPs were located in genes involved in peptidoglycan biosynthesis pathway (*pbpX*, *ponA*, *penA*), two genes associated in the recombination pathway were identified (*recU* and *recR*), along with transferases for cell wall biogenesis (*rsmH* and *mraY)*. Additional SNPs not found in the SvIR or the SvR classification were identified in the genes *mraY* and *recR* (Figure 27).

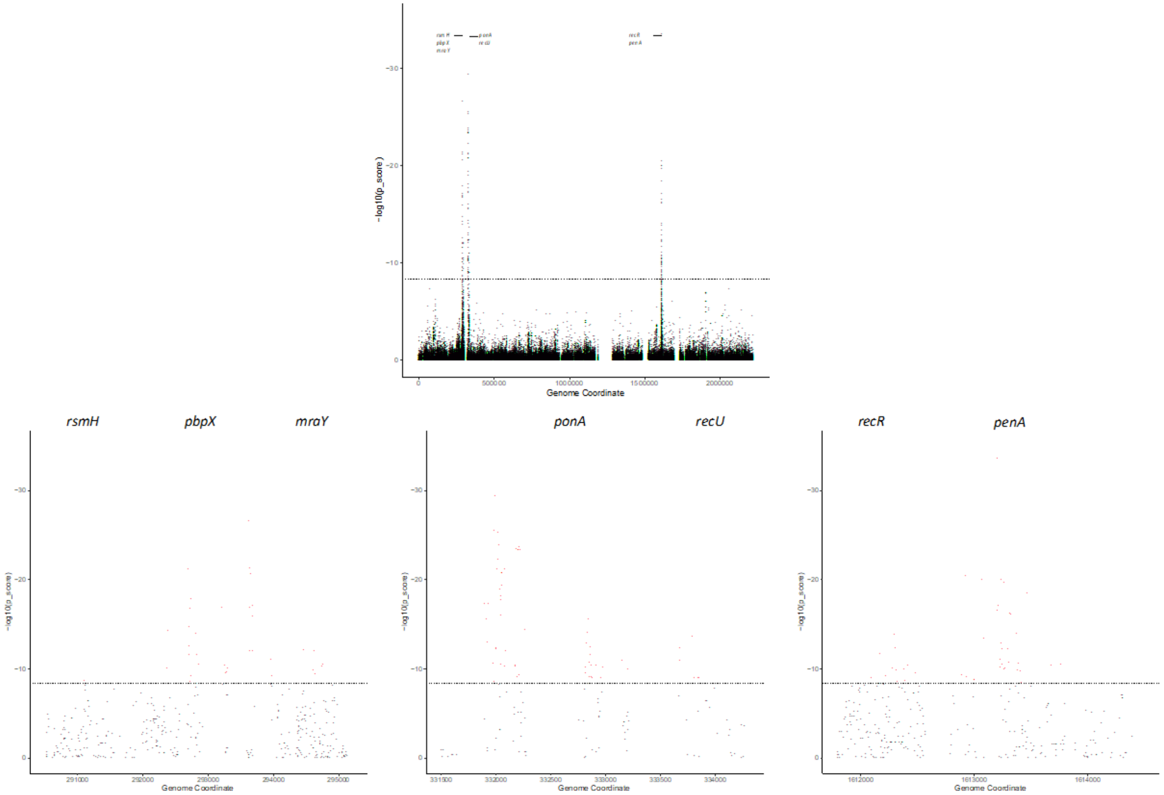


Figure 27 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvI analysis. Genes that are specifically associated with penicillin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction

5.2.2.1.3 Penicillin non-susceptibility associations from a sensitive vs resistant (SvR) GWAS classification

To get a further understanding of whether some SNPs were associated only with the resistant phenotype independent of the intermediate resistant phenotype, a GWAS was performed using the SvR classification. For this, 1,002 sensitive isolates were used as controls (66%) and 523 resistant isolates were cases (34%). The analysis identified 63,307 SNPs of which 152 were significant. The genes in which the SNPs were identified is shown in Figure 28 and with two exceptions they are the same genes identified in the SvIR GWAS classification. Significant association SNPs are present in genes involved in peptidoglycan biosynthesis pathway (*pbpX*, *ponA*, *penA*), genes associated with the recombination pathway (*recU*), the cell division pathway (*gpsB*), and transferases for cell wall biogenesis (*rlmL*). An association SNP identified in the gene *rhaB* unique to this SvR analysis is involved in carbohydrate metabolism (*rhaB*). The QQ plot (Appendix I) showed a good control for population structure however the lambda GC score of 0.62 revealed analysis was further underpowered than the SvIR classification.

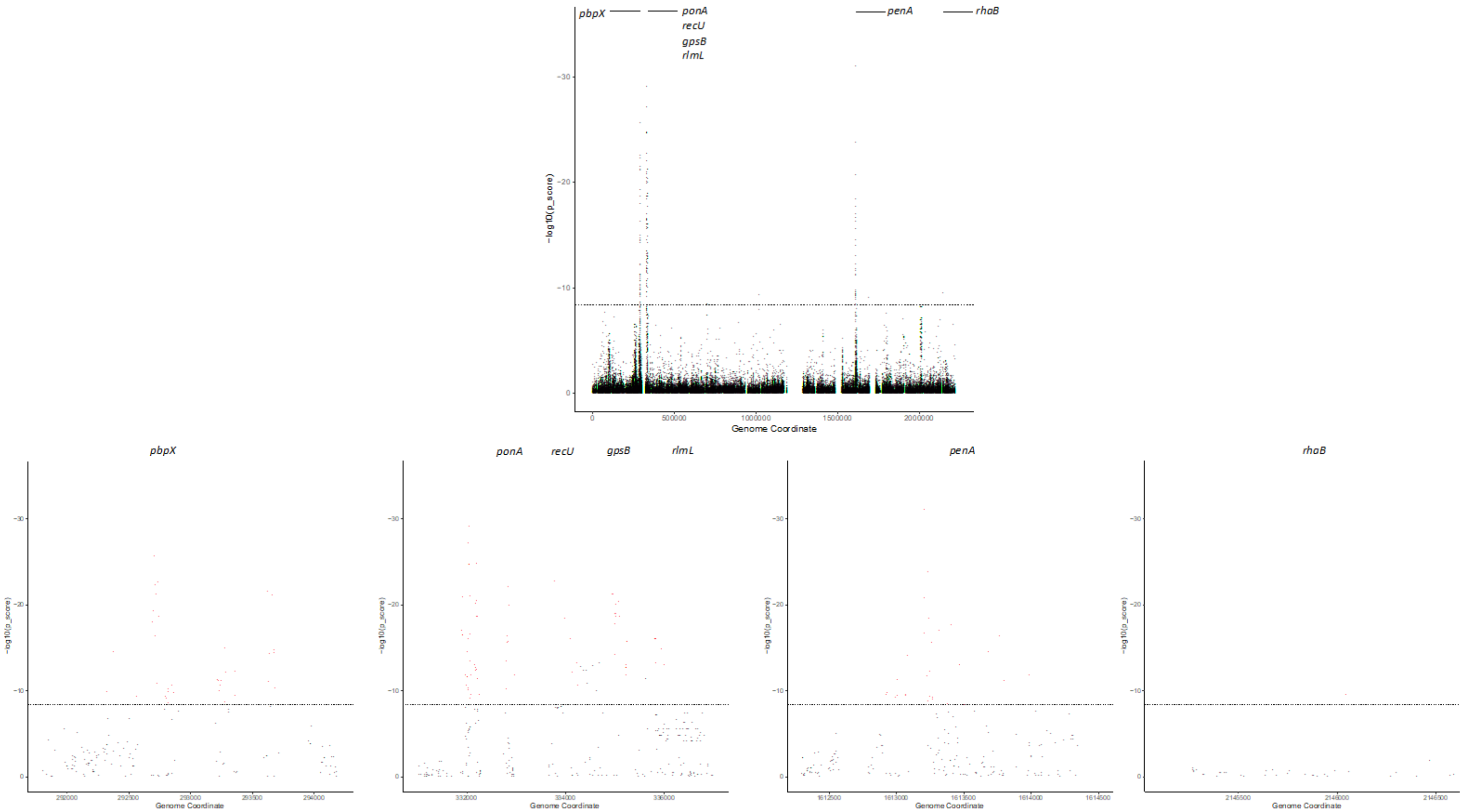


Figure 28 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvR analysis. Genes that are specifically associated with penicillin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction

5.2.2.2 Comparison of penicillin non-susceptibility between association studies

It has been shown that the inclusion of the intermediate resistance phenotype in GWAS can identify additional genes associated with non-susceptibility in this dataset (*rsmH*, *mraY*, r*ecR*). A comparison of results from the SvIR classification with others in the field (179, 255) was performed to identify consistency and give additional strength to findings. Differences in methodology by Lees *et al* (2016) (292) excluded it from the comparison as the exact association SNPs were not generated in the same manner. The 120 significant association SNPs generated in the SvIR analysis was compared with the 301 and 426 association SNPs identified by Mobegi *et al* (2017) and Chewapreecha *et a*l (2014) (179, 255) respectively. Appendix J showed that between all three studies, association SNPs were identified in a total of 163 genes, in hypothetical genes or in intergenic regions. Of these 163 genes, 150 were unique to the Mobegi *et al* (2017((255) study with a range of 1-14 SNPs present in each association gene. One gene *dexB* with five association SNPs was only identified in Chewapreecha *et al* (2014) (179).

A total of four and five SNPs respectively were identified in the gene *rlmL* common between this study and that of Mobegi *et al* (2017) (255) (Appendix J). A single SNP in the gene *rsmH* identified in this analysis was only otherwise present in Chewapreecha *et al* (2014) (179) who described ten association SNPs (Appendix J). An additional five genes were common only in Mobegi *et al* (2017) and Chewapreecha *et a*l (2014) (179, 255); *mraY*, *clpX, dhfR*, *clpC_1* and *ftsL*, and for all but one of these genes (179) identified a higher frequency of SNPs (Table 12). A total of five genes associated with penicillin non-susceptibility were common to all studies *gpsB*, *pbpX, polA, recU* and *penA* and the frequency of association SNPs between studies described in Table 12.

Table 12 Frequency of genome wide association SNPs associated with penicillin non-susceptibility identified in SvIR analysis (section 5.2.2.1.1), Mobegi *et al* (2017) (255) and in Chewapreecha *et al* (2014) (179)

Gene	**Number of association SNPs identified in this analysis**	**Number of association SNPs identified by Mobegi *et al* (2017)**	**Number of association SNPs identified by Chewapreecha *et al* (2014)**
polA	42	6	23
pbpX	36	4	134
penA	22	9	50
gpsB	11	3	6
recU	3	7	5
mraY	0	1	37
clpC_1	0	12	19
dhfR	0	3	5
ftsL	0	1	2
clpX	0	2	1

Consistency in many genes associated with penicillin resistance has been shown by three independent studies (Table 12). A total of 774 SNPs were identified amongst the three studies and Figure 29 highlighted the proportion of SNPs that overlap between them. There were a large number of unique SNPs called in each study, 408 in Mobegi *et al* (2017) (255), 237 in Chewapreecha *et al* (2014) (179) and 62 in this SvIR analysis. Despite some SNPs being located in common genes, there were only six SNPs identified in all three (Table 13). The Mobegi *et al* (2017) and Chewapreecha *et a*l (2014) studies (179, 255) have nine common SNPs not identified in this analysis, and there were nine and three SNPs shared between analysis performed here and that of Chewapreecha *et al* and Mobegi *et al* (179, 255) respectively (Figure 29).

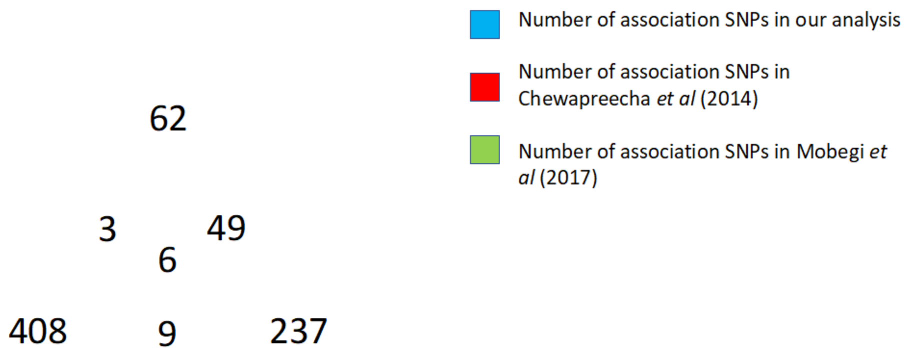

Number of association SNPs in our analysis
Number of association SNPs in Chewapreecha et al (2014)
Number of association SNPs in Mobegi et al (2017)
62
3
49
6
408
9
237

Figure 29 A Venn diagram summarising the number of significant association SNPs in each of the studies, and the number of those that are co detected in this study, in Chewapreecha *et al* (2014) (179) and Mobegi *et al* (2017) (255)

Table 13 Description of six SNPs co-detected in this study, in Chewapreecha *et al* (2014) (179) and Mobegi *et al* (2017) (255)

SNP position	Gene	Product
293661	*pbpX*	Penicillin binding protein 2X
333792	*recU*	Holliday junction specific endonuclease
334107	*recU*	Holliday junction specific endonuclease
335104	*gpsB*	Cell cycle protein
1613503	*penA*	Penicillin binding protein 2b
1613770	*penA*	Penicillin binding protein 2b

5.2.3 Identification of SNPs associated with cotrimoxazole non-susceptibility from whole genome dataset

A GWAS for cotrimoxazole resistance was conducted using 383 isolates with a sensitive phenotype as controls (36%) and 690 isolates with a resistant phenotype as cases (64%). In total 61,443 SNPs were identified with 44 SNPs significantly found to be associated with cotrimoxazole resistance. Of these, 36 SNPs were located in genes encoding enzymes with roles in folate metabolism (*dhfR*, *fpgS* and *sulA)* (Figure 30). The remaining SNPs were identified in intergenic regions (n=3) and in a single hypothetical protein (n=5). The associated QQ plot (Appendix K) mostly followed the expected line showing the GWAS accounted for population structure, and the lambda GC score of 0.86 revealed the analysis was marginally underpowered (Appendix K). A GWAS to identify SNPs associated with cotrimoxazole resistance was performed by Mobegi *et al* (2017) (255) and SNPs in common genes between studies is described in Appendix L.

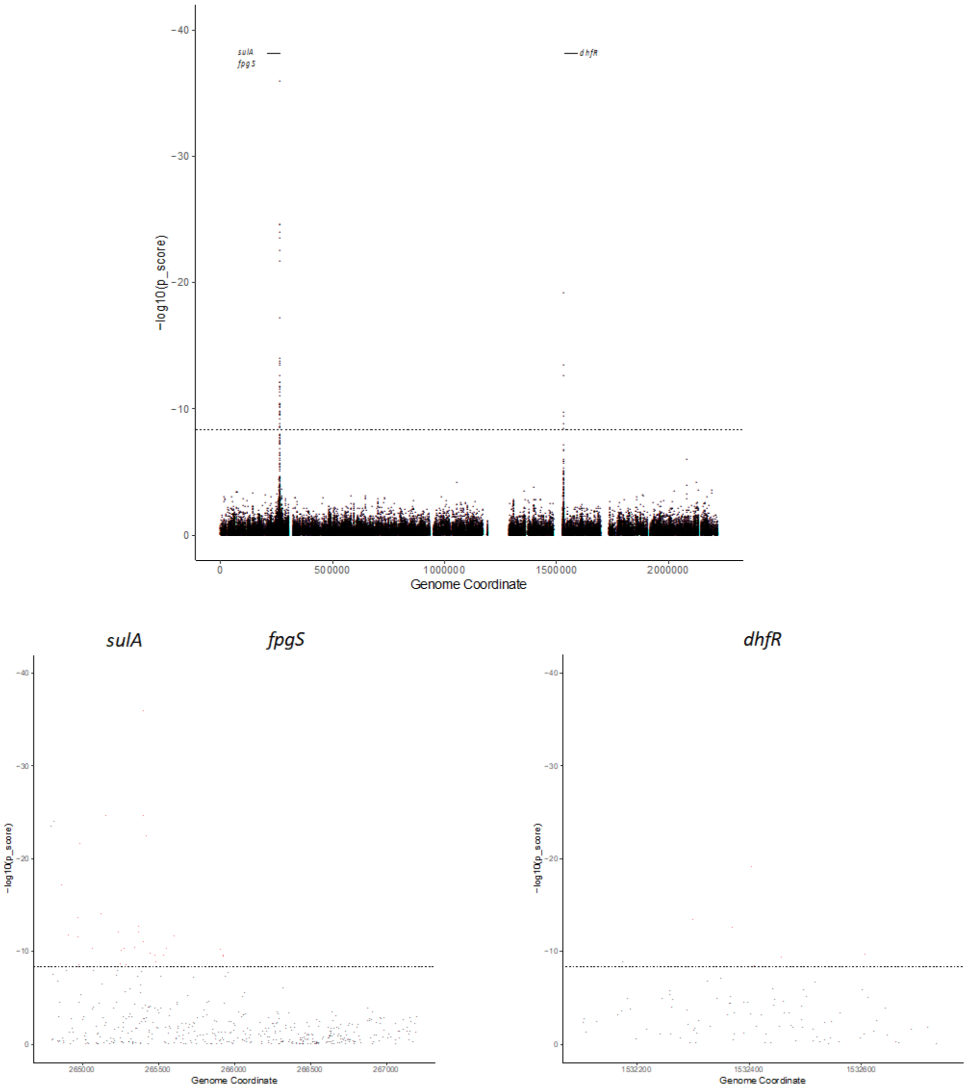


Figure 30 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and cotrimoxazole non-susceptibility. Genes that are specifically associated with cotrimoxazole resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction.

5.2.4 Identification of SNPs associated with erythromycin non-susceptibility from whole genome dataset

Erythromycin resistance data was available for 1,791 *S. pneumoniae* isolates, 746 were classified as controls with a sensitive phenotype (42%) and the remaining 1,045 isolates with a resistant phenotype were cases (58%). In total 64,129 SNPs were identified but only ten were significant. Significant SNPs were located in genes associated with peptidoglycan synthesis (*pbpX*, *glmU*), thiamine biosynthesis (*ykoD*, *apbE*), and in genes associated with DNA synthesis (*sigA*, *dnaG*). An additional association was seen in the insertion sequence ISSpn2 (Figure 31). The QQ plot shown in Appendix K showed normal distribution of p-values and a lambda GC value of 0.85 again reflects an underpowered analysis. A GWAS to identify SNPs associated with erythromycin resistance was performed by Mobegi *et al* (2017) (255) and a comparison of SNPs detected in common genes between studies (*pbpX, ykoD, glmU, apbE*) demonstrated that although there was an overlap in association genes detected, no common SNPs were identified (Appendix M).

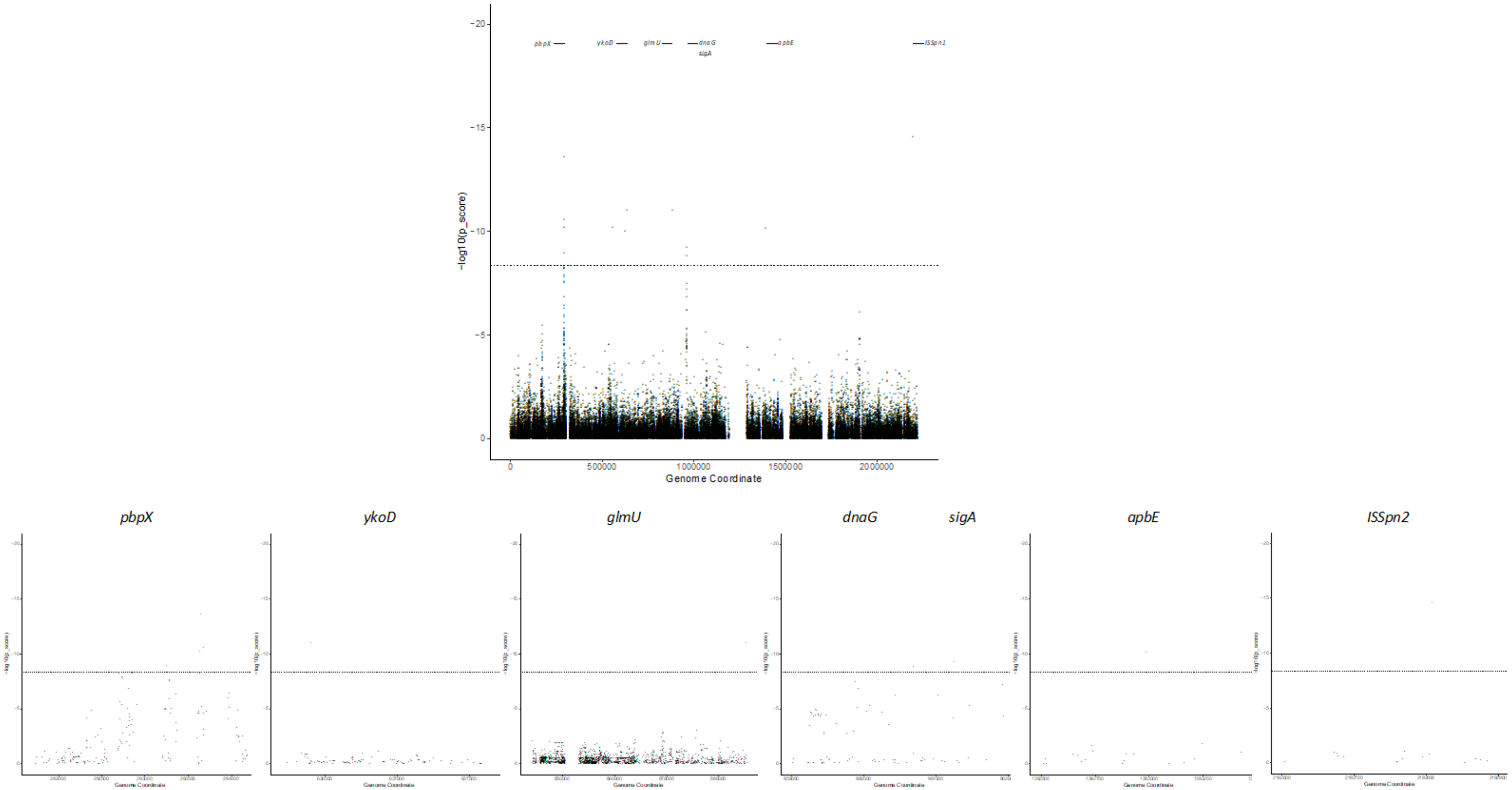


Figure 31 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and erythromycin non-susceptibility. Genes that are specifically associated with erythromycin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction

5.2.5 Identification of SNPs associated with clindamycin non-susceptibility from whole genome dataset

The association test of non-susceptibility for clindamycin consisted of 689 sensitive isolates as controls (72%) and 266 resistant isolates as cases (28%). Analysis identified 66,015 SNPs but only six SNPs reached significance. Association SNPs were located in the genes *pbpX* and *mraY* involved in peptidoglycan synthesis and in the gene *lepA* involved in translation (Figure 32). The QQ plot of p-values (Appendix K) showed population structure was accounted for successfully and the lambda GC score of 0.86 showed analysis was slightly underpowered.

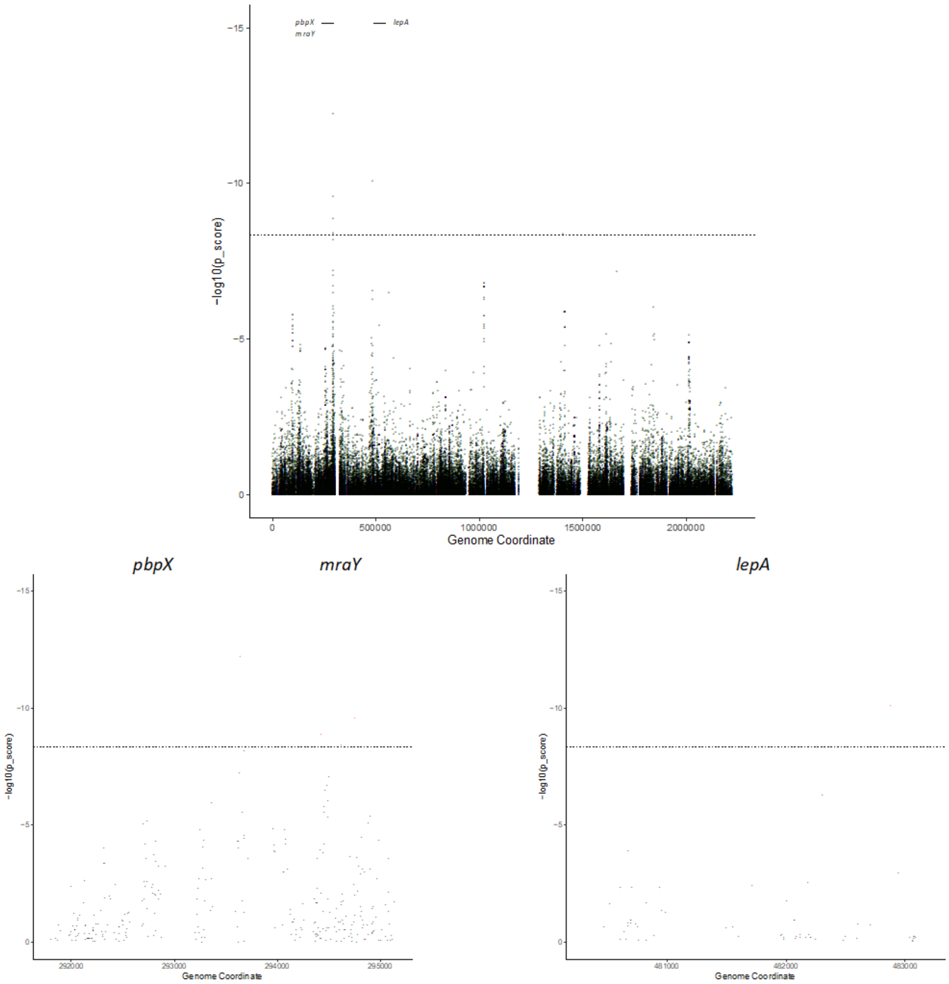


Figure 32 Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and clindamycin non-susceptibility. Genes that are specifically associated with clindamycin resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction

5.2.6 Identification of SNPs associated with chloramphenicol, doxycycline and tetracycline non-susceptibility from whole genome dataset

The GWAS for chloramphenicol resistance consisted of 133 controls with a sensitive phenotype (75%) and 45 cases exhibiting a resistant phenotype (25%). In total 65,104 SNPs were identified however only four of these were significant. Two SNPs were located in a hypothetical protein, and the remaining two were located in the *clpX* gene encoding for an ATP dependant protease (Figure 33). The step wise pattern seen in the QQ plot and the lambda GC of 1.09 (Appendix N) show that population structure had not been appropriately accounted for in this analysis and remained so even after p-values had been corrected by inflation γ (Appendix N).

The GWAS to identify SNPs associated with doxycycline resistance consisted of 44 controls with a sensitive phenotype (41%) and 63 cases exhibiting a resistant phenotype (59%). Following analysis 63,105 SNPs were identified however none exceeded the significance threshold. QQ plots (Appendix O) showed p-values in the doxycycline GWAS deviate under the line of best fit slightly, and that there is slight inflation as the lambda GC score is 1.12 indicating the population structure may not have been fully corrected.

The GWAS for tetracycline resistance consisted of 144 sensitive controls (40%) and 212 (60%) resistant cases. An initial 64,416 SNPs were identified however no SNPs were identified to be significantly associated. Tetracycline population structure was correctly accounted for, it strongly adhered to the line of best fit and had a good lambda GC score of 1.02 (Appendix O).

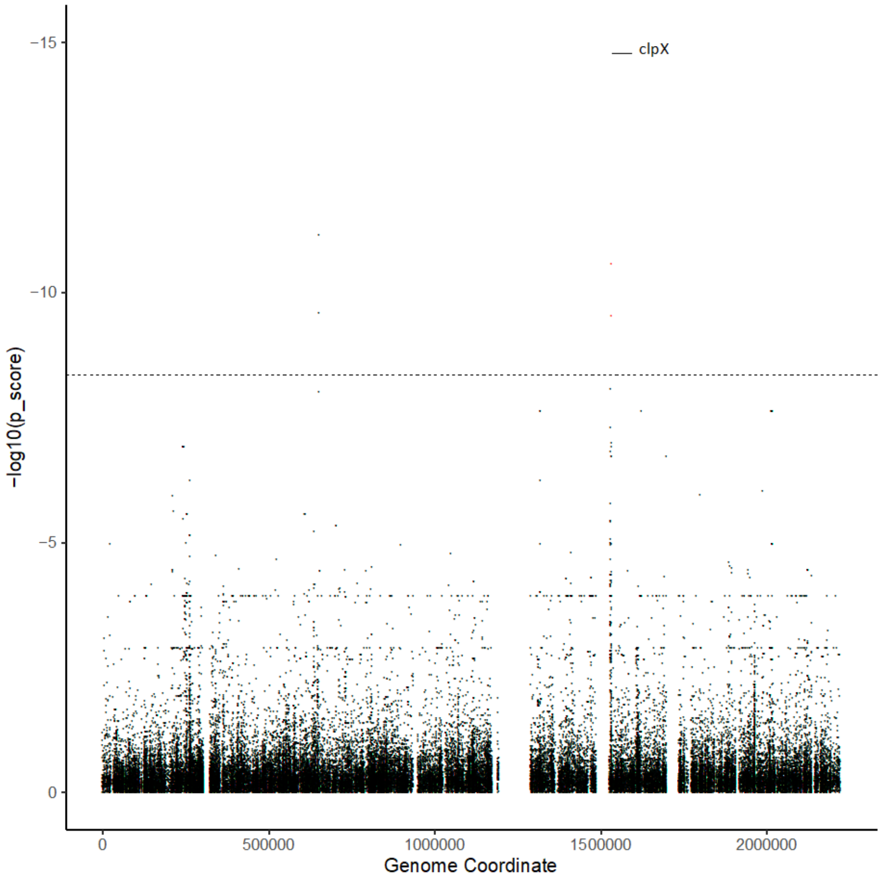


Figure 33 Manhattan plot showing the position of significant associations in the genome and their relative p-value in chloramphenicol GWAS. Manhattan plots showing the statistical significance and genome coordinate of genome wide associations between SNPs and chloramphenicol non-susceptibility. Genes that are specifically associated with chloramphenicol resistance are labelled and their position represented by vertical lines in the top panel. The bottom panels show the genes in more detail, gene areas are shaded in grey show the position and significance of association SNPs in red. The dotted horizontal line in both panels represent the significance cut-off after Bonferroni correction.

5.2.7 GWAS of penicillin non-susceptibility using recombination-free genomes

To determine whether the SNP changes in genes implicated in non-susceptibility to penicillin occurred as a result of recombination or mutation, the GWAS using the SvIR classification was performed using the recombination-free dataset generated in section 5.2.1.1. Following analysis, 1,798 association SNPs were identified only a single significant SNP located in the *ftsK* gene was identified (Figure 34). The QQ plot (Appendix P) showed successful correction of population structure and the lambda GC score of 0.89 revealed analysis was slightly underpowered.

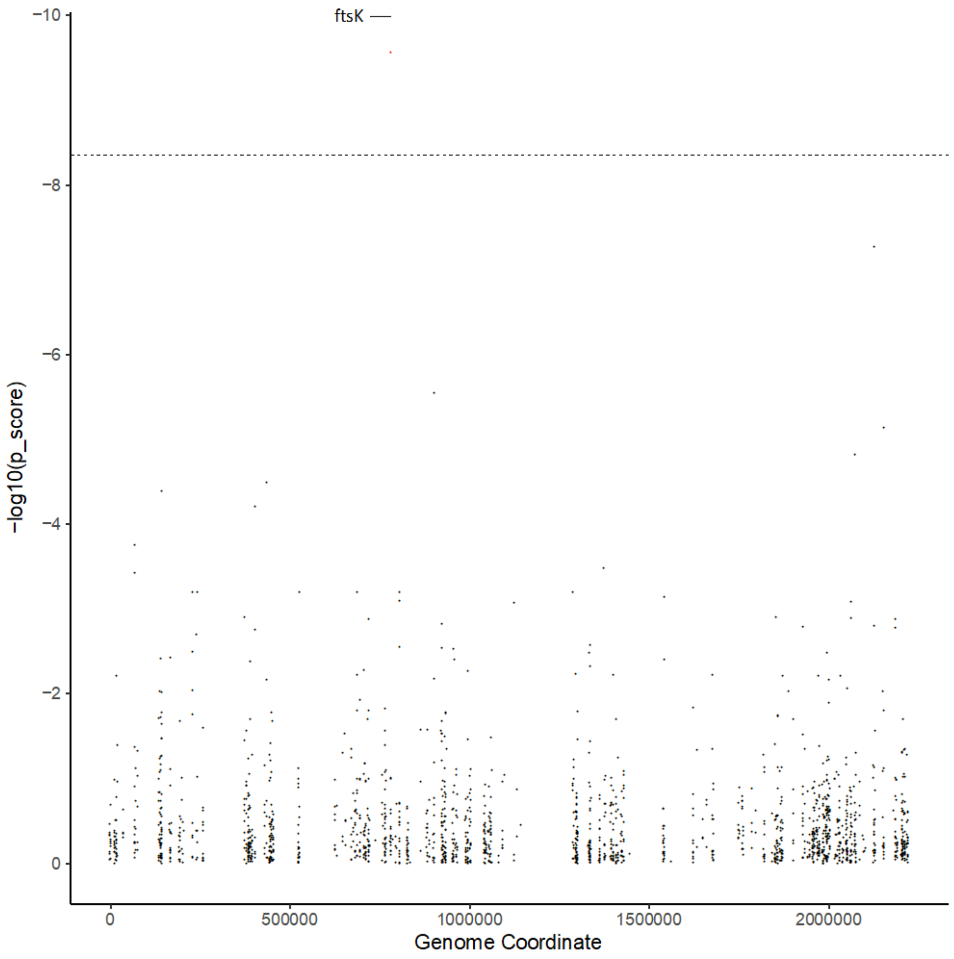


Figure 34 Manhattan plot showing significant association SNP in SvIR classification of GWAS from recombination-free datasets. Manhattan plot showing the statistical significance and genome coordinate of genome wide associations between SNPs and penicillin non-susceptibility in SvIR analysis of recombination-free genomes. The gene specifically associated with penicillin resistance is labelled and its position represented by the vertical line. The significant association SNP is coloured red and the dotted horizontal line represents the significance cut-off after Bonferroni correction.

5.3 Discussion

It has been shown that high proportions of the *S. pneumoniae* genome can be altered by recombination. This work aimed to identify how recombination impacted the population structure of *S. pneumoniae* isolates from Singapore. The purpose of this was to ensure population structure was correctly determined prior to performing a GWAS on antimicrobial resistance. This was done through the initial identification and removal of recombination from 2,059 *S. pneumoniae* isolates to generate a recombination-free dataset. The population structures between this, and the *S. pneumoniae* dataset consisting of whole genome data, was compared using phylogeny and PCA. Phylogenetic trees created for both datasets visually looked very similar and their similarity to a phylogeny representative of core genes suggests core genes remained the predominant influence to determine phylogeny. In addition, the same isolates composed the main population cluster of PCA in both datasets suggesting areas of recombination were not used in population clustering by PCA.

Sufficient laboratory data was available to independently identify SNPs associated with the phenotypes of antimicrobial sensitivity, intermediate resistance, and resistance to penicillin. Results revealed SNPs associated with an intermediate resistant only phenotype were present in genes involved in recombination and cell wall synthesis. SNPs associated with a resistant phenotype to penicillin, clindamycin, and erythromycin were identified in genes involved in peptidoglycan and cell wall synthesis. Additional resistant determinants to erythromycin were identified in genes involved in thiamine biosynthesis and DNA synthesis, and, in clindamycin resistance, in genes involved in translocation.

The GWAS identifying penicillin non-susceptibility determinants using the recombination-free dataset saw the removal of all previous significant associations. This is conclusive that all these SNPs were located in areas of the genome which at some point had recombined in at least one of the 2,059 *S. pneumoniae* isolates.

5.3.1 The effect of recombination on population structure

Analysis in Chapter 4 showed ClonalFrameML was the most suitable recombination program to apply to this dataset of *S. pneumoniae* isolates. Interpretation following recombination analysis showed a large proportion of the genome was identified as recombination in at least one isolate (86%). The large number of subsets processed provided good representation of each isolate in the analysis.

Phylogenies from whole genomes, recombination-free genomes and core genes were similar by approximately 70%, which showed construction of all three must be largely based on the core genome. Despite these similarities, it was hypothesised that population structure may differ between whole genome and recombination-free genomes which may influence the downstream association test. The program popPUNK was specifically designed to determine population structure of *S. pneumoniae* isolates and hence, was used initially to define population structure from datasets. Isolates were clustered into components based on similarities in both core and accessory genomes relative to the rest of the population (212). PopPUNK clustered both whole genome and recombination-free datasets into a large number of components, too many to use in the population structure corrections and with too few isolates within the component to provide power to discover associations. Due to this, principal component analysis (PCA) was used to identify sub-structures in the datasets. PCA captures the inferred genetic ancestry of individuals into principal components which can then be used as fixed effects in a regression based test for association to account for population structure (319). The PCA plots showed there was far more divergence in the whole genome dataset as the scale of the PC1 and PC2 axes was much greater than that of the recombination-free dataset (Figure 21). Although there were far fewer SNPs to represent variance in the recombination-free dataset, it showed that much of the variance within the population is described by recombination sites. There was a large amount of similarity in the isolates that were grouped together to make up PCA clusters between datasets. This shows that removal of recombination areas did not affect the inclusion of isolates within clusters and further reinforces initial observations seen in the phylogeny comparison, that it was largely core genes that informed clustering.

5.3.2 Genome wide association studies to identify SNPs associated with antimicrobial resistance

Optimally, in PCA the first few principal components (PCs) encompass most of the variation in the data however, in these datasets the first 10 PCs only accounted for <60% of total variance (Figure 22). Rather than including a large number of principle components as fixed effects in the regression model, the linear mixed model software GEMMA (213) was used to test for association. This method has shown good performance in modelling the dependence structure of the dataset as it is able to capture both population structure and kinship (313, 325, 326). It explicitly models pairwise relatedness between all individuals to determine covariance and regresses the phenotype on principal components of the genotype matrix as random rather than a fixed effect (327). QQ plots were generated following each analysis to check the correction for population stratification as deviation from the reference line can reflect systemic inflation of the test statistic (328). QQ plots of p-values from the penicillin, cotrimoxazole, erythromycin, and clindamycin association studies showed population stratification was accounted for as there was an excess of low p-values across all SNPs which have not reached significance and largely follow the reference line. The point at which the p-values depart from this distribution showed inflation only of the SNPs with high p-values, which were considered to be associated with the phenotype of interest. The lambda GC calculation gave a direct measure of the inflation in the sample and should be 1 in the case of the null. Inflation would be indicated for values >1.05 (328) but the majority of association studies performed here proved to be underpowered. Selection of an appropriate statistical significance threshold to differentiate between true positives, false positives or false negatives is critical in GWAS. Although there have been many statistical tests proposed to account for multiple testing (329-331), the Bonferroni correction is considered the most conservative in its selection of a threshold p-value. This is because it maintains the assumption that every genetic variant tested is independent of one another (332). Correcting for multiple testing does limit type 1 errors however, it is not without limitation as it also inflates type 2 errors (333).

5.3.2.1 Penicillin

Penicillin susceptibility data was available for 1,670 pneumococcal isolates and phenotypes were classified as sensitive, intermediate or resistant by hospital providers. As reasonable numbers of each category were present, a novel approach was implemented which performed the GWAS using different combinations of phenotypes to classify cases and controls. This provided the opportunity to identify SNPs associated with an independent intermediate resistant phenotype as well as the resistant phenotype.

The mode of action for penicillin is inhibition of cell wall biosynthesis, therefore it was encouraging to identify significant associations in genes that contribute to this pathway such as *pbpX*, *penA*, *ponA* (also referred to as *pbp2x*, *pbp2b*, *pbp1a*) in all classifications of GWAS performed (SvR, SvIR and SvI). Many penicillin binding proteins (PBPs) are present in *S. pneumoniae*. A higher number of association SNPs were identified in these genes than any other, which could signify them to be the primary determinant of penicillin resistance (115-117). As penicillin has historically been given as first line treatment, SNPs associated with resistance in these genes could be reflective of prescribing policies. Associations in these genes were also identified by the independent studies (179, 255, 292) giving strong evidence to their role in resistance. Mutations in PBP that result in structural changes may impact drug binding and subsequently be beneficial to the organism by resulting in varying levels of resistance.

Association SNPs were identified in the gene *rsmH* which functions upstream of the PBPs to encode transferases of the peptidoglycan biosynthesis pathway during cell wall biosynthesis. Associations within this gene were only identified in the SvIR and SvI classification of cases and controls, and therefore could be indicative of an association to an intermediate resistance phenotype within this dataset. Chewapreecha *et al* (2014) also identified associations in this gene (referred to as *mraW*) (179). However, they were associated with a resistant rather than intermediate resistant phenotype. It may be that the modification to the pathway was slightly different between pneumococcal populations and that SNPs within this gene could result in either an intermediate or a resistant phenotype. Further research would be required to support this theory. The identification of significant SNPs in additional genes that encode transferases of peptidoglycan synthesis were *mraY* and *rlmL,* and these reinforce SNPs in the same genes also identified in (179, 255) respectively. The present study identified association with *mraY* only in the SvI classification of cases and controls. The reason why the SNPs were not also identified in the SvIR analysis could be that the SNPs from resistant phenotypes are potentially reducing the p-value of SNPs in the SvIR classification resulting in them not exceeding the threshold for significance. Associations in the gene *rlmL* were only present above significant thresholds in the SvR and SvIR classification of cases and controls and not the SvI classification, indicating it could have been associated only with a resistant rather than an intermediate resistant phenotype. Association SNPs were also found in a range of genes involved in other processes including the recombination pathway, cell division and in carbohydrate metabolism.

The gene *gpsB* is important in cell wall growth and viability (334). Like with *rlmL,* a considerable number of SNPs were identified in this gene in the SvIR and SvR classification of cases and controls, and the absence of SNPs in the SvI classification suggested an association with a resistant rather than intermediate phenotype. Associations were identified in this study and in the studies by Chewapreecha *et al* (2014) and Mobegei *et al* (2017) (179, 255), giving additional confidence to the association seen in this gene. *GpsB* is putatively essential in *S. pneumoniae* and experimental evidence has shown division defects following gene depletion, and significant cell elongation and cessation of growth similar to cell deformation following methicillin use (335). Associations in the gene *ftsL* which has similar function to *gpsB,* was also identified in the studies by Chewapreecha *et al* (2014) and Mobegei *et al* (2017) (179, 255). There may be some additional interactions occurring between the genes *gpsB, ftsL* and the PBPs that have a direct or an indirect effect on penicillin susceptibility.

The link between recombination and the acquisition of non-susceptibility to antimicrobials has been well investigated and described, and the associations present in genes of the recombination pathway could be modifications that enhance or facilitate this process. Significant association

SNPs were identified in the *recU* gene of the recombination pathway in this study and by Chewapreecha *et al* (2014) and Mobegei *et al* (2017) (179, 255) which again gives confidence to the validity of this association with penicillin non-susceptibility. A number of SNPs were identified in an additional gene of the recombination pathway *recR* but only from the SvI analysis. This could indicate an association to an intermediate phenotype only, which does not reach significance in the SvIR analysis. Another gene with significant associations novel to the SvR classification of cases and controls is in *rhaB.* The gene is involved in carbohydrate degradation but in this case as only a single SNP was identified, which barely reached the threshold for significance, validity would need to be confirmed.

5.3.2.1.1 Further comparison between the present study and Chewapreecha et al and Mobegi et al

Both studies by Chewapreecha *et al* (2014) and Mobegei *et al* (2017) performed a GWAS on penicillin non-susceptibility and found associations in other genes outside the peptidoglycan pathway not detected in this analysis (Table 12) (179, 255). The genes *clpC_1* and *clpX* (*clpC_1* referred to as *clpL* in original studies) are a major heat shock protein and chaperone which have the capacity to interact and stabilize *pbp2x.* This could aid in resistance and experimental evidence showed mutants lacking this gene can be more susceptible to penicillin (166). Associations in the metabolic gene *dhfR* (referred to as *dyr*) involved in resistance to trimethoprim and cotrimoxazole was identified from both studies (179, 255). This gene affects the DNA synthesis pathway rather than cell wall synthesis which is the mode of action for penicillin. Association identified in this gene could be from the misuse of antimicrobials in Thailand and the USA where the pneumococcal populations originated. The most common antimicrobials for treating upper respiratory tract infections in Thailand are beta lactams and cotrimoxazole, and if both are used in a short time frame there is the potential to drive the co-selection of resistance to different classes of antimicrobials (179). The same findings were not identified from this study however, it was slightly under powered and therefore possible additional associations may be present but not captured with the applied level of significance.

The overlap in genes and SNPs between independent studies validates and gives confidence in results. There were only 6 SNPs that were common in all studies even though associations in many of the same genes were identified. As the same isolates were used in the Mobegi *et al* (2017) analysis (255) and Chewapreecha *et al* (2014) (179) it was un expected to see only 15 SNPs commonly identified between them and 55 SNPs commonly identified between this analysis and Chewapreecha *et al* (2014) (179) (Figure 29). The total number of SNPs associated with penicillin resistance in this analysis was much higher than that reported from other studies (179, 255).

Before significance thresholds were applied >60,000 SNPs were identified in this analysis whereas in the Massachusetts cohort (255) 4,317 SNPs were reported and in the Maela cohorts (179) 1,721 and 858 SNPs were reported. The differences might be reflective of the differences in methodology as (179, 255) use the Cochran-Mantel-Haenszel to test for associations between antibiotic resistance phenotypes and SNPs whereas the genome-wide efficient mixed-model for association was used in this study. Alternatively, it could be due to inherent characteristics of the isolates tested as cohorts comprise of carriage, disease, or a mixture of these from different geographical locations which are likely to have differing levels of genotypic variance. In all three studies it is likely that some of the association SNPs identified are not causative of antimicrobial non-susceptibility but rather they are linked to causative SNPs.

It has been shown that the classification of phenotypes that constitute cases and controls can alter the subsequent association SNPs that are identified. Using different classifications in this penicillin GWAS has shown consistencies in association genes but has also highlighted the addition or loss of associations in genes between the classifications. In this analysis, there is no evidence that the phenotype is dependent on the number of SNPs as in both the *ponA* and *penA* genes there was a greater number of SNPs present in the SvI classification than the SvR. When an association SNP is identified only in the SvI classification of this analysis and the SvR classification of the work of others (179, 255), it could be due to the independent evolution of individual datasets. Depending on the specific mechanisms of resistance that have evolved, what might result in full resistance in one pneumococcal population may constitute only intermediate resistance in another. Over time it may be these SNPs are associated only with the resistant rather than the intermediate phenotype. Depending on what antimicrobial data is available for populations, different classifications of cases and controls might be beneficial in determining novel or emerging resistance mechanisms within specific populations.

For the remaining antimicrobials tested, the genome wide association analysis did not include isolates with an intermediate phenotype because numbers were small and the effect of these on the p-values for the resistant phenotype was not fully understood.

5.3.2.2 Cotrimoxazole

Cotrimoxazole targets bacterial DNA synthesis by sequential blockade of folic acid enzymes in the synthesis pathway. SNPs associated with cotrimoxazole resistance were predominantly present in the genes *sulA*, *fpgS* and *dhfR* that produces enzymes dihydropteroate synthase, folypolyglutamate synthase and dihydrofolate reductase respectively; all of these are involved in folate metabolism. Associations in these genes were also identified by the independent study performed by Mobegi *et al* (2017) (255) (Appendix L) and changes on target enzymes *dhfR* (also referred to as *dyr*) and *sulA* (also referred to as *folP*) have been shown to enhance resistance to cotrimoxazole (336, 337) giving confidence to results.

5.3.2.3 Macrolides

Macrolides act by inhibiting protein synthesis of bacteria. Therefore, common mechanisms of resistance are to alter the ribosomal target site to prevent macrolide binding or to export antibiotic via an efflux pump. The analysis performed in this study identified significant associations in genes involved in the biosynthesis of thiamine *ykoD* and *apbE*. The *ykoD* cistron of the ykoFEDC operon putatively encodes the ATPase component of a unique thiamine-related ABC transporter (338) and a*pbE* encodes the thiamine biosynthesis lipoprotein. SNPs in genes affecting this pathway could be associated with resistance as thiamine is an important nutrient for the synthesis of bacterial capsular polysaccharide. Changes could affect the ability of the enzyme to penetrate the bacterial cell membrane and therefore block subsequent binding to its internal ribosome. The identification of associations in genes outside this pathway and in genes of peptidoglycan synthesis were also observed. They included the *glmU* gene which encodes for a bifunctional enzyme with acetyltransferase and uridyltransferase activity (339), and in the gene *pbpX* previously described to be associated with penicillin resistance. The mode of action for erythromycin differs to that of penicillin and does not rely on disrupting the peptidoglycan pathway. Therefore, the effect of these mutations on erythromycin resistance was not fully understood. Mutations in these genes have however, been independently identified by both this study and in Mobegi *et al* (2017) (255) suggesting an association to the resistant phenotype might be true. Significant association SNPs in genes involved in DNA replication (*sigA, dnaG*) and in the insertion sequence ISSpn2 were identified, and associations in these genes are unique to this study. *SigA* encodes RNA polymerase major sigma factor involved in bacterial transcription, and mutations in this could affect the binding of the macrolide to its target site on the ribosome. The gene *dnaG* encodes DNA primase, which if inhibited by mutation, is expected to halt DNA replication and, as a result, cell proliferation (340). Although erythromycin targets protein synthesis, there may be some benefit in the modification of genes associated in the DNA synthesis for the organism. Another explanation could be that it is a result of an epistatic relationship with another antibiotic, for example fluroquinolones that have a direct effect on bacterial DNA synthesis. An insertion sequence is a bacterial mobile DNA element and pneumococcal genomes often have an over distribution of these elements which can affect antimicrobial resistance (341, 342).

5.3.2.4 Clindamycin

Clindamycin is an antibiotic active against *Streptococci* and indicated for use in a range of infections including septicaemia, peritonitis and intra-abdominal infection. It is part of the lincomycin class of antibiotics and binds to bacterial 50s ribosomal subunit to interfere with bacterial protein synthesis (343). This was the first association study performed which identified associations with clindamycin resistance and genes such as *pbpx* and *mraY*. These genes are not involved in the protein synthesis pathway but instead are involved in the peptidoglycan biosynthesis pathway and a transferase respectively. Associations in these genes were seen for other antibiotics such as penicillin. Therefore, they could be linked or have some indirect effect in clindamycin resistance. The association was identified with a novel gene *lepA* encoding elongation factor 4 (344), a ribosomal dependent GTPase essential in translation elongation. The gene is required for protein synthesis and therefore, adaptation may support its association with a non-susceptible phenotype. With only one association SNP present in the gene, and no reinforcement from other studies, it is difficult to determine the validity of this association with the resistance phenotype.

5.3.2.5 Chloramphenicol

The mode of action of Chloramphenicol is to inhibit the peptide transferase activity of the bacterial ribosome which prevents protein chain elongation and therefore protein synthesis. It is a widely prescribed antimicrobial used to treat ear and eye infections. Only two significant associations in the gene *clpX*, a class of ATPases that aid in tolerance to environmental stresses and cell survival (345), were identified from the GWAS. Interpretation of the QQ plot generated from association p-values highlighted that population structure had not successfully been accounted for in the analysis, therefore it is not possible to determine whether these were real associations with resistance. This remained the case even after the attempt to rectify results by correction of the inflamed population statistic. Further investigation as to why the population stratification was not successful only for this antibiotic showed that 30/45 isolates with the resistant phenotype were clustered together in cluster 6 of the whole genome PCA (Figure 21). This is one of the caveats to performing a GWAS, sometimes cases intrinsically group together in the population which makes it very difficult to correct for in the methodology. No significant SNPs associated with doxycycline or tetracycline were identified in this study. These two analyses had the smallest number of cases for all those tested however, neither study showed it to be underpowered by a lambda GC score <1. It is likely that based on the isolates included in the cases no SNPs present had a strong enough association with the non-susceptible phenotype.

5.3.2.6 The role of recombination in the genetic association of penicillin non-susceptibility

Performing genome wide association analysis to determine associations to a range of clinically relevant antimicrobials resulted in the identification of a number of significant SNPs in genes essential to bacterial pathways. Many of these have the potential for contributing to the non-susceptible phenotype and these observations confirm findings of previous work in the field. To further the knowledge of whether the association SNPs identified are present as a result of mutation or recombination, a final GWAS on penicillin non-susceptibility was performed on the recombination-free genomes generated in section 5.2.1.1. Results from the analysis showed removal of the majority of signal previously identified in the penicillin GWAS as only 1,798 rather than 63,570 associations were identified. After significance thresholds were applied, only a single SNP remained in the gene *ftsK* which is known to produce a multifunctional protein that acts in cell division and chromosome segregation in *E. coli* (346). The direct involvement of this gene in penicillin resistance was questionable as it was a single SNP with a relatively low p-value and therefore would need validation. Results following this final analysis showed most of the genes found to have associations with a non-susceptible phenotype to penicillin are also sites of recombination. This implicates the importance of recombination in determining the resistance profile of pneumococcus because mechanisms of resistance can be transferred as whole genes, mosaic segments within genes, and as we have seen from this analysis can also transfer single mutations in a larger area of recombination that could be implicated in novel mechanisms of resistance. Some resistance-causing mutations were previously known and result in ‘mosaic genes’ such as the PBP. This research has highlighted the potential of a wider range of genes that exhibit a mosaic appearance through recombination, and which potentially contain resistance conferring mutations.

The true effect of some of these resistance conferring mutations is hard to determine because through epistasis, cell fitness is dependent on the genetic background of the individual strain and additional mutations or compensatory adaptations present within the cell could be affecting overall fitness (347). It is possible that combinations of some of the mutations are commonly detected and required for the non-susceptible phenotype. Other mutations might be not associated with the phenotype at all and instead have some compensatory role affecting cellular fitness.

5.4 Conclusions and future work

A comparison of population structures generated from whole genome data and recombination-free genome data showed that both cluster similar isolates together. Phylogenies of these datasets and that of the core genome shared 70% similarity showing that the areas of the genome affected by recombination are not the drivers that determine phylogenies, and rather its variation in core genes.

Genome wide associations were performed using GEMMA to identify significant SNPs associated with a non-susceptible phenotype to a variety of antibiotics including penicillin, cotrimoxazole, erythromycin, doxycycline, clindamycin, chloramphenicol and tetracycline. SNPs associated with a resistant phenotype were identified for all these antimicrobials except doxycycline and tetracycline. It was not possible to correctly account for the population structure in the analysis for chloramphenicol resistance due to the intrinsic similarity of cases. Separate analysis for penicillin using case classifications as intermediate resistance, resistant, and a combination of intermediate resistance and resistant, highlighted significant association SNPs in slightly different genes, and this might be a consideration for future analysis in bacterial GWAS if the necessary antimicrobial data is available. For penicillin, the genes, when mutated, that are suggested to convey full resistance are *gpsB* and *rlmL* as they were not detected in the SvI GWAS classification. Those that could convey an intermediate resistance phenotype include *mraY*, *recR* and *rsmH*. Genes associated with either an intermediate resistant, or resistant phenotype were *pbpX*, *penA*, *ponA*, and *recU*.

Associations in genes associated with folate metabolism *sulA*, *fpgS* and *dhfR* were identified following the GWAS for cotrimoxazole resistance and modifications in this essential pathway are logical as in order for bacteria to survive they still need to perform these processes even in the presence of the antibiotic. GWAS for both clindamycin and erythromycin identified association SNPs in genes that are not directly involved in the mode of action of the antimicrobial, for example in genes of the peptidoglycan pathway or in DNA synthesis, and these could be representative of compensatory mutations acting indirectly on the phenotype. It is hard to determine complex interactions that occur between genes directly from genome sequences as association SNPs identified could be interacting epistatically with the resistant phenotype. In addition epistasis is known to occur between the PBP genes and that the effect of a single mutation, for example the resistant phenotype, is dependent on the presence or absence of mutations in other PBP (348).

The final aim of the analysis was to identify whether the SNPs associated with the non-susceptible phenotype originated purely from evolution rather than recombination. The additional GWAS for penicillin non susceptibility on recombination-free genomes showed previously identified associations were lost. From this it could be concluded that although specific mutations can contribute to the resistance phenotype, they are present on areas of the genome that have also undergone recombination by at least one isolate within the dataset. Weighting the associations by the frequency that the loci are also seen within a recombination block will be useful to determine whether the SNP is associated more with natural mutation or recombination.

Growing knowledge in the field has led to the development of new programs and tools that claim to be able to detect associations whilst correctly accounting for bacterial population structures such as PYSEER (349) and BUGWAS (294). In addition to using SNPs to identify associations, these can use Kmers which allows the identification of multiple forms of genetic variation such as genes, INDELS, copy number variants and sequence insertions. This method would also capture multi-allelic SNPs/genes that might be responsible for a particular phenotype. Methodologies that use kmers however may require additional computational resources as they are less compact than SNPs. Longer kmer length increases sensitivity of the tests, but in turn significantly increases the requirement of memory and processor usage. Results following Kmer based analysis can be complex and hard to interpret, therefore an alternative option which also captures multiple types of variation could be the program DBGWAS. This implements De Bruijn graphs in its GWAS and claims easy interpretation and visualisation of results. Further work applying such programs to large datasets of clinical isolates like this one could prove useful in further confirming or identifying new mechanisms associated with resistance or virulence. A kmer based rather than SNP based approach performed in future work would allow the comparison of results from different methodologies on the same dataset. If kmers incorporated a number of SNPs this may collectively enhance power of previously individual SNPs and highlight areas most suitable for further investigation. This could give more confidence, or see the removal of, some of the single borderline significant SNPs identified, such as *rsmH*, *lepA* and *rhaB*. As all the GWAS with significant associations to antimicrobial non-susceptibility were underpowered, another way more power could be generated would be to use a larger dataset, with Singapore data combined with datasets used by other authors Chewapreecha *et al* (2014) and Mobegi *et al* (2017) (179, 255). It would be necessary to ensure population structure was correctly accounted for in what is likely to be an even more diverse dataset, and whether analysis should be extended only to disease-causing isolates would need to be considered. Estimations as to whether the effect size of a single variant would be strong enough to change the phenotype could be made from this analysis to identify which candidate SNPs/loci should be prioritised for future functional characterisation. Six SNPs were identified to be associated with penicillin resistance from critical processes such as the cell cycle pathway, recombination pathway and cell wall biosynthesis from this study and two independent studies (179, 255), which gives a high degree of confidence to these results. Functional verification of candidates could be performed by knock-out experiments to establish causality or by using transposon insertion mutations and then testing the susceptibility profile of the mutants. This would identify the false positivity rate of the methodology and functionally characterise candidate SNPs/loci.

Chapter 6 Discussion

6.1 Introduction

Pneumococcal disease caused by the bacterium *S. pneumoniae* is a major cause of morbidity and mortality, especially in the young and elderly. It remains a leading cause of death worldwide even with the availability of antibiotics and vaccines. This highlights the need to monitor vaccine efficacy within specific populations and to explore potential new mechanisms of drug resistance to effectively control future disease. Geographic heterogenicity in serotype and resistance is known, therefore an investigation of these from a local collection is very valuable. This work used whole genome data from a large collection of disease-causing *S. pneumoniae* isolates (n=2,059) from Singapore to obtain insight into the serotype distribution responsible for disease, and to estimate levels of vaccine efficacy. In addition, data regarding susceptibility to antimicrobials for the *S. pneumoniae* isolates allowed levels of resistance to be determined within this collection. Data generated proved the first and second specific hypotheses to be true as there were both epidemiological changes, and changes in the resistance profile of disease-causing isolates over time in Singapore. The paired whole genome data and antimicrobial susceptibility data provided the opportunity to perform a genome wide association study to identify single nucleotide polymorphisms (SNPs) associated with a phenotype of resistance to clinically relevant antimicrobials. To date, a study such as this that utilises a large collection of disease only isolates, from a single geographic location has not yet been performed. The third specific hypothesis was also proved to be true as this work identified SNPs with significant associations to antimicrobial resistance to numerous clinically used antibiotics.

6.2 Key findings

Isolates of disease-causing *S. pneumoniae* were collected retrospectively from major hospitals in Singapore and sequenced as described in (182) and in Section 2.1. This resulted in a dataset consisting of 2,059 isolates responsible for causing pneumococcal disease between 1997-2016 which included isolates that infected an otherwise sterile site, resulting in invasive pneumococcal disease (IPD), and those that caused infection in a non-sterile site, resulting in non-invasive pneumococcal disease (non-IPD).

6.2.1 Epidemiology

Differences in the polysaccharide capsule of *S. pneumoniae* enable its differentiation into serotypes. This study identified a total of 64 different serotypes from the disease-causing isolates; the most frequently identified serotypes were 19F (n=365, 18%), 23F (n=244, 12%), and 14 (n=196, 10%), which is consistent with findings from other studies from Singapore (224). These serotypes, covered by the PVC7 vaccine, all followed a decreasing trend in the proportion of infection they caused across the study period. Other commonly isolated serotypes were serotype 3 (n=158, 8%) and 19A (n=161, 8%) which are present in the PCV13 vaccine but not the PCV7 vaccine. These showed an increasing trend across the study years and justify the need to extend coverage from PCV7 to PCV13. Serotype 6E (n=143, 7%), also frequently identified from this dataset, is not included in the PCV vaccines. Over the study period the proportion of disease caused by this serotype decreases, however, as it is still frequently associated with infection this should continue to be monitored. Serotypes found to have a high odds ratio (OR) in causing disease in the adult population relative to the paediatric population included serotypes 1, 8, 7A, 12F, 4, 6D, and 20. Serotype 6B had a higher OR for disease in paediatric relative to adult populations, and similar findings for the age differentiation in serotype 6B and serotype 8 were also reported globally from (224) and (72) respectively.

Certain serotypes are known to be associated with an increased invasive potential, and this study showed that serotypes 4, 8, 20, 7A, 14, and 19A had a high OR for IPD relative to non-IPD. These findings are consistent with those from datasets of other geographical locations including England, Portugal, Switzerland, USA, and Finland (31, 32, 217, 219, 221). In contrast, serotypes 23F, 15A, and 19F were associated with not causing IPD. Generally, the proportion of total infections caused by serotypes with a higher invasive potential was low (≤3%); in contrast, serotypes 14 and 19A caused 10% and 8% of total infections, respectively. Vaccine coverage provided by PCV7 ranged across the study duration from 21% - 76%, for PCV13 from 50 – 86%, and for PPV23 at 56% - 89%. This shows that these vaccines, which were originally designed for the US and European markets, are at present predominantly targeting the correct serotypes responsible for causing a large proportion of disease in Singapore.

Resistance data to a number of antimicrobials was provided by hospital laboratories which was consistent with previous reports of resistance in Singapore (Table 2). Overall resistance to cotrimoxazole was highest (63%), followed by erythromycin (58%), tetracycline (58%), and doxycycline (58%). Although resistance to cotrimoxazole and erythromycin show comparatively higher proportions of isolates with resistance than clindamycin (28%), when the proportion of isolates was compared between the study years, it was shown that resistance was decreasing for cotrimoxazole and erythromycin but increasing for clindamycin (Figure 8). Of the 31% of isolates with penicillin resistance, 74% exhibited additional resistance to both cotrimoxazole and erythromycin. The serotype that is most commonly identified with a resistance phenotype to all but doxycycline was serotype 19F. This again reinforced findings from Chong et al (2008) who also identified serotype 19F as a predominantly resistant serotype (229). Antibiotic use is known to drive antibiotic resistance and therefore antibiotic prescribing data for the same period in Singapore could have provided context for the changes in resistance over time. If a reduction in prescribing of cotrimoxazole or erythromycin for example was observed, this could highlight the success of effective antibiotic stewardship efforts.

6.2.2 Identification of recombination from clinical isolates

One of the main ways that this organism remains such a considerable burden to healthcare is due to the genomic plasticity of *S. pneumoniae*. Being naturally competent, it can acquire exogenous DNA by horizontal transfer and incorporate genetic material into its genome. The acquisition of such material, particularly of genes associated with drug resistance or capsular polysaccharide genes, can have a direct effect on the current preventative measures used to control disease such as vaccines and/or antimicrobials. Recombination is a major contributor to this and can result in capsule switching events between organisms, leading to a loss of protection from the vaccine or the transfer of resistance genes, such as *ermB,* which result in resistance to antimicrobials used to treat infection. Other resistance determinants have been identified from single, or the accumulation of, mutations. Such mutations associated with resistance to antimicrobials can be identified through a genome wide association study. However, to avoid spurious results, it is necessary to separate mutations that cause a phenotype from non-causal linked mutations. To do this, inherent similarities and differences between isolates resulting from clonal expansion must be accounted for by determination of and controlling for the population structure. It was hypothesised that as these mutations are examples of vertical inheritance, the best way to identify SNPs associated with antimicrobial resistance was to base the population structure on vertical inheritance. High levels of recombination described in *S. pneumoniae* may influence estimations of population structures, therefore, the initial investigation was to determine the best way of identifying areas of recombination from genomes, and then compare allocation of isolates to population clusters based on recombination-free genome data.

One of the main aims was to compare the output following recombination analysis by Gubbins and ClonalFrameML to select the program most suitable to identify all areas of recombination in the dataset of 2,059 clinical *S. pneumoniae* isolates. Once identified, these areas were to be removed to compare population structures of isolates based only on vertical inheritance, with that of whole genome data which included horizontal inheritance from recombination. Results following analysis showed ClonalFrameML to be the most appropriate program to use. Firstly, it was shown that a similar sequence to that containing recombination was required to be present in the dataset to successfully detect an artificially introduced recombination event. As there was a range of diversity in the isolates this may pose some limitation to recombination detection. ClonalFrameML was able to process larger datasets which increased the likelihood of the inclusion of a similar sequence and successful recombination detection. In addition, ClonalFrameML was shown to be able to distinguish between smaller recombination sites as opposed to Gubbins where this was grouped into one recombination event, which consequently reduces the proportion of the genome incorrectly identified as recombination. This larger proportion of the genome that would remain after the removal of recombination events may contain valuable data which influences determination of population structure.

6.2.3 GWAS to identify SNPs association with antimicrobial resistance

Following recombination analysis of the clinical isolates, recombination events were removed, and these recombination-free genomes used to perform a principal component analysis. The population structures between isolates in the dataset could then be visualised and compared with clustering generated from whole genome data. The same isolates were predominantly clustered together in both datasets which suggested data present in the sites of recombination was not used to inform isolate allocation a specific cluster. Although isolate allocation to the cluster remained the same between both the whole genome and the recombination-free datasets, a large proportion of the total variance was not captured in the first ten principal components in either, therefore it was not appropriate to perform the GWAS based on these clusters. Instead, a kinship matrix was generated using GEMMA, a genome wide efficient mixed model analysis program to identify and account for population structure in its subsequent test for association and this was possible from the complete dataset. GWAS to identify SNPs associated with non-susceptibility to penicillin, erythromycin, cotrimoxazole, doxycycline, clindamycin, tetracycline, and chloramphenicol was performed using GEMMA. Additional classifications of cases and controls were used in the penicillin GWAS to identify SNPs specifically associated with an intermediate resistant, resistant phenotype or both. Many of the SNPs identified were located in genes involved in the target site for the specific antimicrobial. For penicillin this included genes in the cell wall and peptidoglycan biosynthesis *(pbpX*, *ponA*, *penA*, *gpsB, rlmL, rsmH*, *mraY*) however, SNPs were also present in other genes such as those involved in the recombination pathway (*recU*, *recR*). SNPs in some of the genes involved in peptidoglycan synthesis were also present for other antimicrobials such as erythromycin and clindamycin. Reasons for this could be due to linkage within the genome, co-evolution of resistance mutations or alternatively could be examples of compensatory mutations. The GWAS for cotrimoxazole resistance identified SNPs in genes involved in folate metabolism *(dhfR, fpgS, sulA*) which coincides with the antimicrobial target of action on DNA synthesis. Additional SNPs identified in the erythromycin GWAS were present in genes associated with thiamine biosynthesis (*ykod, abpE*) and DNA synthesis (*sigA, dnaG*). Further investigation into these could identify them as contributors to the resistant phenotype in the dataset of pneumococcal isolates from Singapore.

6.2.4 Limitations

The isolates used in this study are only from clinical cases of disease in Singapore. Due to this, the findings and conclusions presented are only representative of pneumococcus in this geographical location. Isolates were obtained retrospectively from archived stores from the four major hospitals in Singapore. Comparison of the number of IPD cases collected with figures notified to the Ministry of Health Singapore after mandatory reporting was implemented in 2010 showed that, after 2011, this dataset captured approximately half of the total number of IPD cases (Table 4). In addition, the large fluctuation in numbers of non-IPD isolates collected in different study years was also indicative of discrepancies in the storage of pneumococcal isolates for inclusion in this study. Due to this, a full epidemiological investigation of disease-causing isolates from Singapore was not possible, and instead data was used to estimate characteristics of disease-causing isolates in the wider population. This limitation in the number of isolates available also limited power for analysis. The program used for genome assembly, VelvetOptimiser, was suitable for use at the time but is now considered to be outdated. It was recognised that an artefact of this program resulted in a large spread of contigs which can be an indicator of poor sequencing data. The large spread of contigs previously observed was reduced after processing a subset of isolates using the more up-to-date assembler SPAdes (198, 199). The recombination analysis performed was limited to the two programs available at the time: Gubbins and ClonalFrameML. The programs used were not capable of processing the dataset in its entirety, therefore recombination analysis had to be performed on subsets. Additional programs such as FasTGEAR (253) are now available for inclusion in future analysis. This study used SNPs in the test for association which excludes other forms of genetic variation. Future work could implement recently developed tools that use kmer based (349) or De Bruijn graph methodology (350) to encompass these additional forms which may increase power in the GWAS.

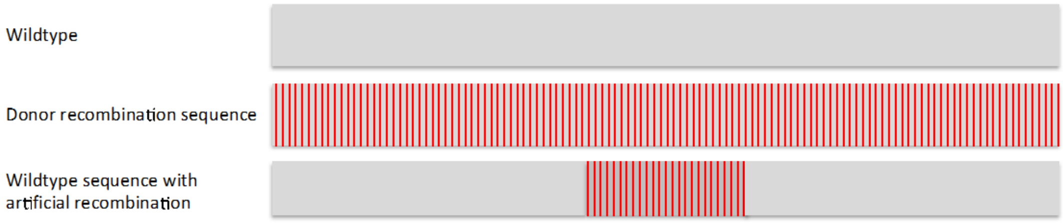
6.2.5 Concluding remarks

Of considerable concern to the management and prevention of future *S. pneumoniae* infection is new emerging variants that have novel or enhanced virulence properties and/or the emergence of isolates with resistance mechanisms not currently combated by the present repertoire of antimicrobials. Data presented has proved that very high levels of recombination are present within pneumococcal populations and these affect areas of the genome associated with drug resistance. Recombination cannot be prevented but increased understanding and application of detection methods to clinical datasets is a vital step that has been addressed through this work. Accurate identification of recombination is of considerable importance as serotype switching events, as well as the emergence of non-vaccine serotypes, have been described in pneumococcal populations. Until there are new vaccine targets, the current vaccine needs to adequately cover disease from emerging variants. In addition to the need to identify new vaccine targets for the future, it is of paramount importance to continue to monitor local and international levels of vaccine coverage. This study identifies serotypes associated with disease and follows resistance levels to a range of antimicrobials across the study duration. This provided the data to map characteristics of disease to this geographical location which could be used to inform local policies. At present, reasonable levels of vaccine coverage was shown from this data. Although the findings here are only representative of Singapore, they can be compared with those found from different geographical locations. Consistency in findings from independent datasets regarding the serotypes responsible for causing invasive disease, the serotypes responsible for disease in distinct age groups, and serotypes most associated with antimicrobial resistance phenotypes adds confidence to the data.

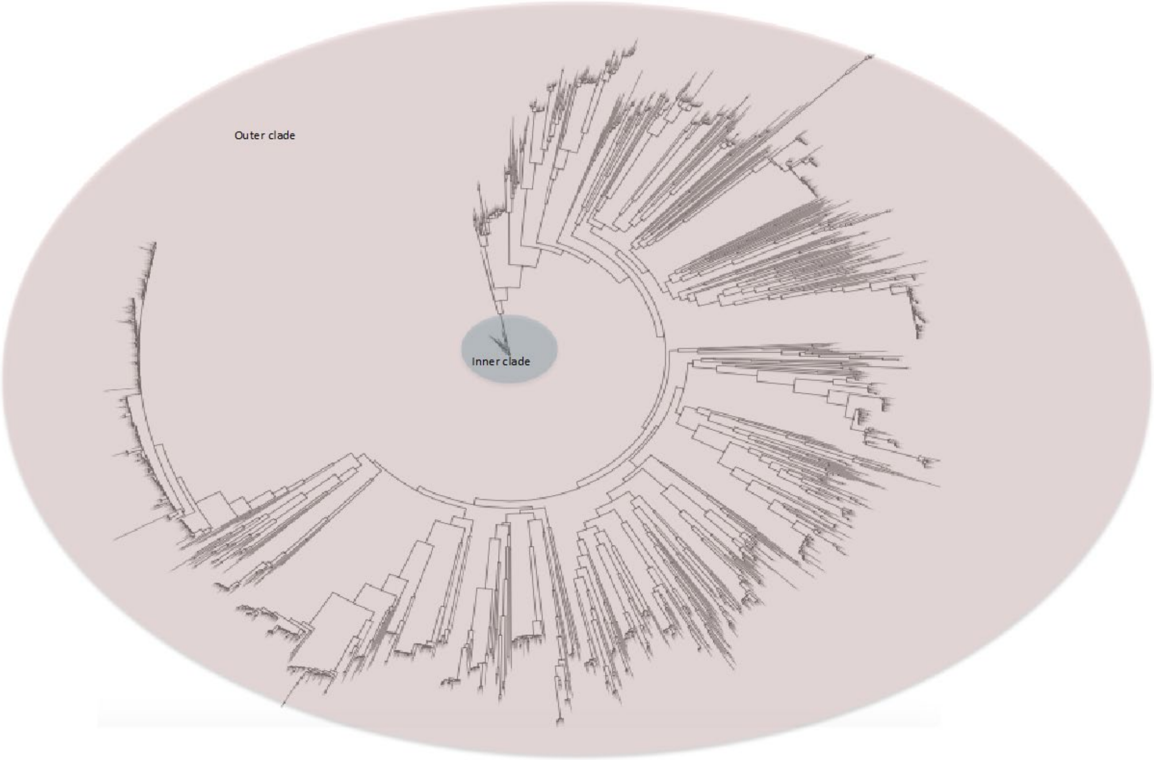
As evidenced through the recent SARS-CoV-2 pandemic, techniques such as those used in this study to identify mutations and understand their implications are vital in counterbalancing threats from emerging variants (351-353). Better still, predicting the effects of potential mutations before they occur could result in a significant reduction in the morbidity and mortality caused by this organism. This study showed utilisation of a GWAS was successful in identifying SNPs that are potentially associated with antimicrobial resistance to a number of clinically relevant antimicrobials. Findings from this confirmed associations in genes also identified in other pneumococcal populations, as well as identifying SNPs in some new, and potentially novel genes. These findings are exciting as they could be examples of new mechanisms of resistance not identified by previous methods. Future GWAS could help to prospectively search for mechanisms of resistance to additional antibiotics used in infectious diseases. This would be particularly beneficial for the beta-lactam class of antibiotic to closely monitor the mutations as it is known that the accumulation of these can result in differing levels of resistance. This emerging field of molecular medicine can have a direct impact on clinical cases of disease by directing treatment offered to patients by the identification of specific molecular changes as well as providing useful information to help anticipate resistance in the future.

Appendix A Constructing simulated sequence containing one recombination

Illustration of how the recombination sequence was generated, as described in section 2.2.1. The red lines represent artificially incorporated SNPs within the sequences (not to scale). The recombination sequence is wildtype with a 10kb region transferred from donor. Recombination programs should detect the central region with increased diversity as recombination.

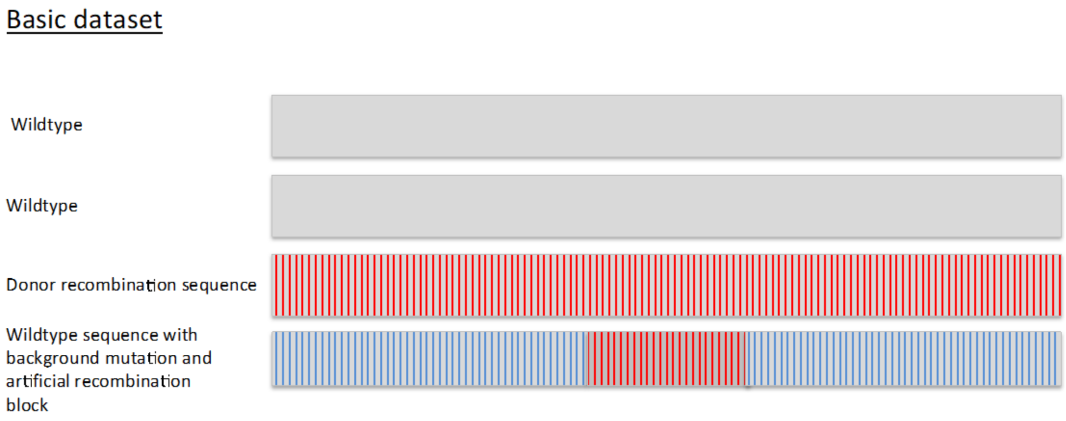

Wildtype
Donor recombination sequence
Wildtype sequence with artificial recombination

Appendix B Phylogenetic tree of 2,059 *S. pneumoniae* isolates specifying ‘inner’ and ‘outer’ clades



Appendix C Basic dataset to investigate the effect of background mutation on the detection of artificial recombination

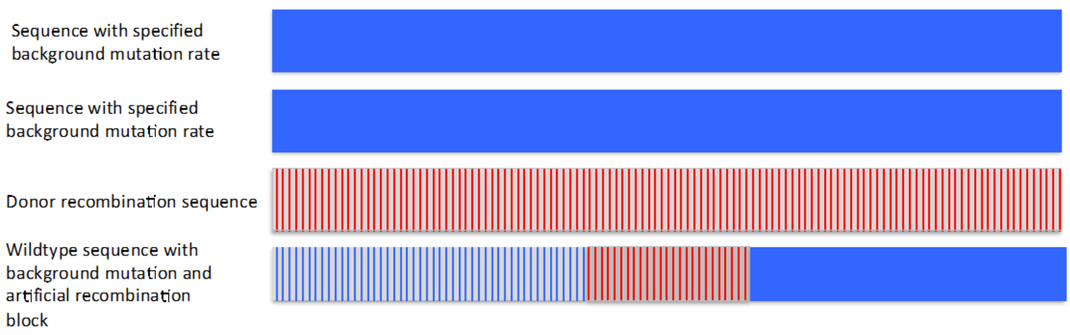
Illustration of the sequences included in the basic dataset.


Basic dataset
Wildtype
Wildtype
Donor recombination sequence
Wildtype sequence with background mutation and artificial recombination block

Appendix D Dataset containing a similar sequence to investigate the effect of background mutation on recombination detection

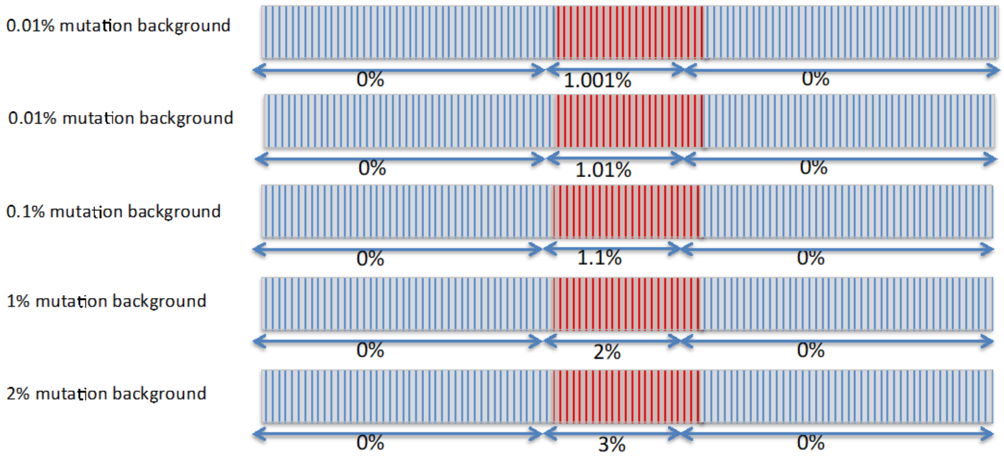
Illustration of the sequences included in dataset.

Dataset with similar sequence present


Sequence with specified background mutation rate
Sequence with specified background mutation rate
Donor recombination sequence
Wildtype sequence with background mutation and artificial recombination block

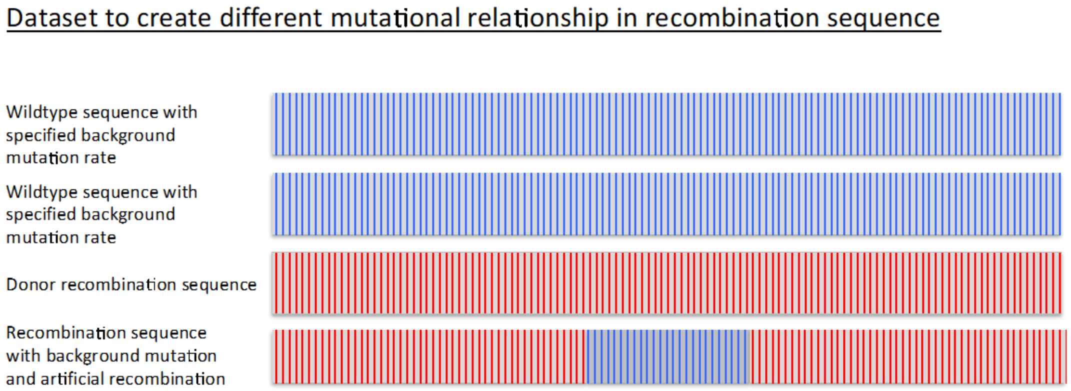
The genomic difference between chromosome with specified mutation and the recombination event.

Difference in mutation across the genome of recombination sequence

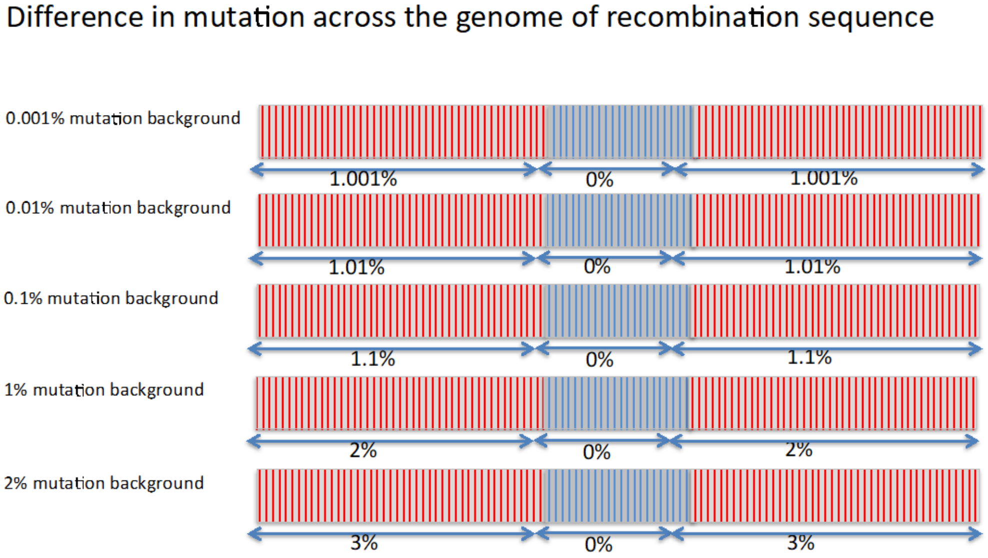


Appendix E Dataset to investigate the effect of background mutation on recombination detection when the recombination event has a low mutation rate compared to the rest of the chromosome

Illustration of the sequences included in dataset.

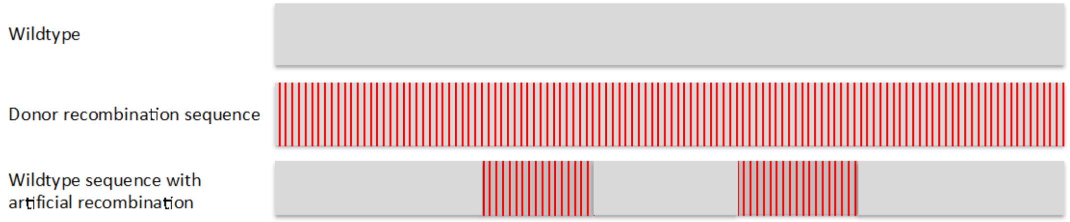

Dataset to create different mutational relationship in recombination sequence
Wildtype sequence with specified background mutation rate
Wildtype sequence with specified background mutation rate
Donor recombination sequence
Recombination sequence with background mutation and artificial recombination

The genomic difference between chromosome with specified mutation and the recombination event.


Difference in mutation across the genome of recombination sequence
0.001% mutation background
1.001%
0%
1.001%
0.01% mutation background
1.01%
0%
1.01%
0.1% mutation background
1.1%
0%
1.1%
1% mutation background
2%
0%
2%
2% mutation background
3%
0%
3%

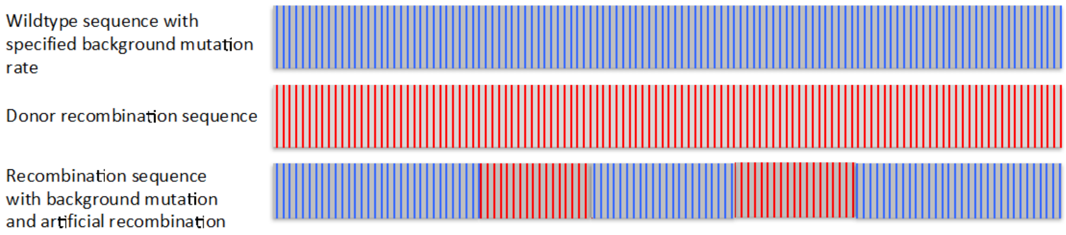
Appendix F Construction of sequence containing two recombination events with no background mutation

Illustration of how the sequence containing two recombination events was generated as described in section 2.2.1. The red lines represent artificially incorporated SNPs within the sequences (not to scale). The recombination sequence is wildtype with recombination events transferred from donor. The distance between the two recombination events ranges from 100 – 10,000bp.


Wildtype
Donor recombination sequence
Wildtype sequence with artificial recombination

Appendix G Construction of sequence containing two recombination events with incorporated background mutation

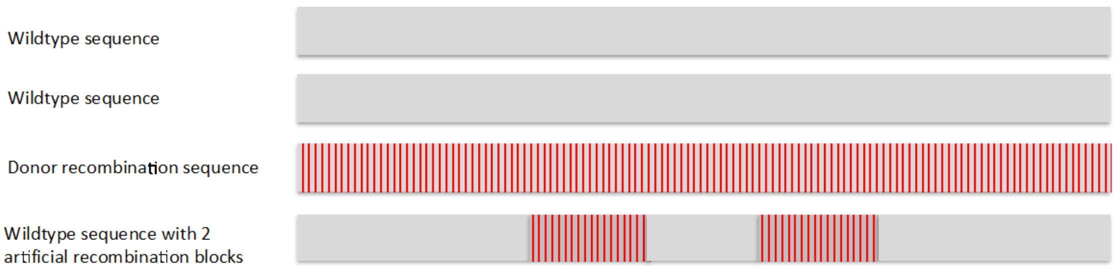
Illustration of how the sequence containing two recombination events were generated as described in section 2.2.1. The red lines represent artificially incorporated SNPs within the sequences (not to scale). The blue lines represent artificially incorporated SNPs to achieve the desired background mutation rate to the wildtype sequence. Mutation rates tested were 0.001%, 0.01%, 0.1%, 1% and 2%. Recombination events are transferred from donor sequence and the distance between the two recombination events ranges from 100 – 10,000bp.


Wildtype sequence with specified background mutation rate
Donor recombination sequence
Recombination sequence with background mutation and artificial recombination

Appendix H Basic dataset for analysis of simulated sequence with two recombination events

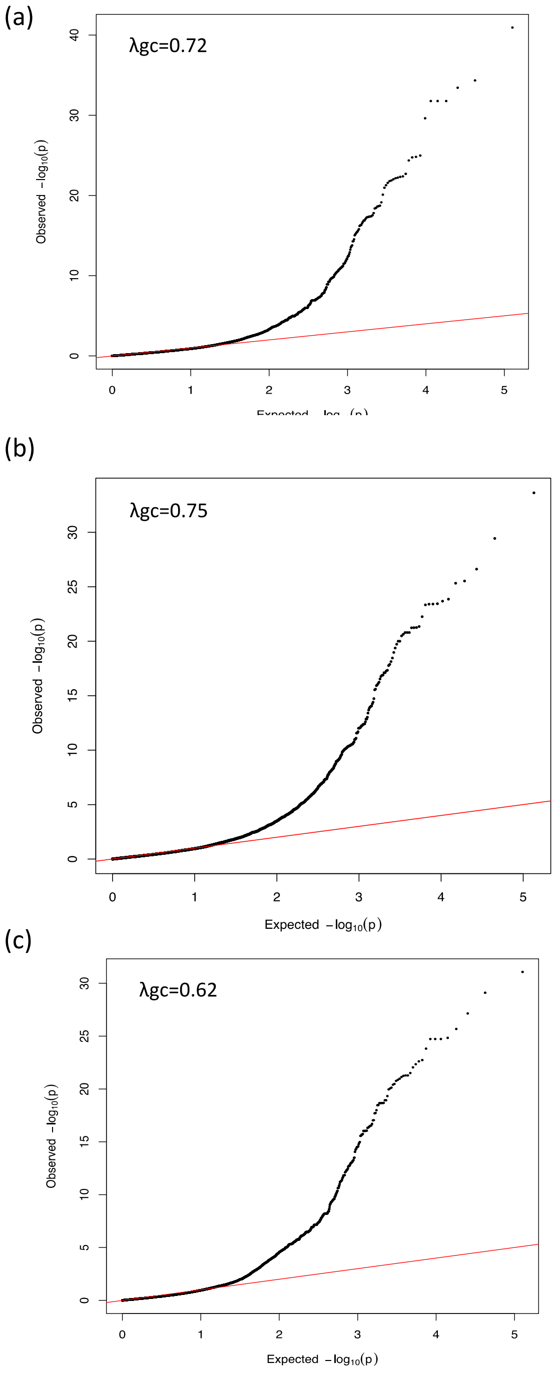
Sequences included in the basic dataset testing the effect of recombination size on distance required by programs to identify individual recombination events

Basic dataset


Wildtype sequence
Wildtype sequence
Donor recombination sequence
Wildtype sequence with 2 artificial recombination blocks

Appendix I QQ plots for the GWAS of penicillin non-susceptibility

(a) QQ plot for SvIR analysis (b) QQ plot for SvI GWAS analysis (c) QQ plot for GWAS SvR analysis. Each plot compares the distribution of -log(p-values) observed in the analysis with the expected distribution under the null hypothesis. The red line is the y=x reference line and lambda GC values (λgc) shown for each.



Appendix J Summary of SNP frequencies in genes associated with penicillin resistance between studies

Summary table of SNPs in genes combing the results of SvIR GWAS analysis with that of Mobegi *et al* (2017) (255) and Chewapreecha *et al* (2014) (179).

Gene	Total SNPs	SvIR analysis	Mobegi *et al* (2017)	Chewapreecha *et al* (2014)
pbpX	150	36	4	134
Hypothetical protein	75	1	73	1
Intergenic SNP	69	0	67	3
penA	65	22	9	50
polA	60	42	6	23
mraY	37	0	1	37
clpC_1	28	0	12	19
tsf	14	0	14	0
gpsB	11	11	3	6
rsmH	11	1	0	10
recU	10	3	7	5
rlmL	9	4	5	0
dhfR	6	0	3	5
mreC	6	0	6	0
rpsB	6	0	6	0
dexB	5	0	0	5
luxS	5	0	5	0
priA	5	0	5	0
xseA	5	0	5	0
metE	4	0	4	0
prfB	4	0	4	0
rpoC	4	0	4	0
ybbH_2	4	0	4	0
clpX	3	0	2	1
ecsA_1	3	0	3	0
gabR	3	0	3	0
lacG	3	0	3	0
nagB	3	0	3	0
udk_2	3	0	3	0
agaA_1	2	0	2	0
amyS	2	0	2	0
bglK_2	2	0	2	0
btuD_1	2	0	2	0
ddl	2	0	2	0

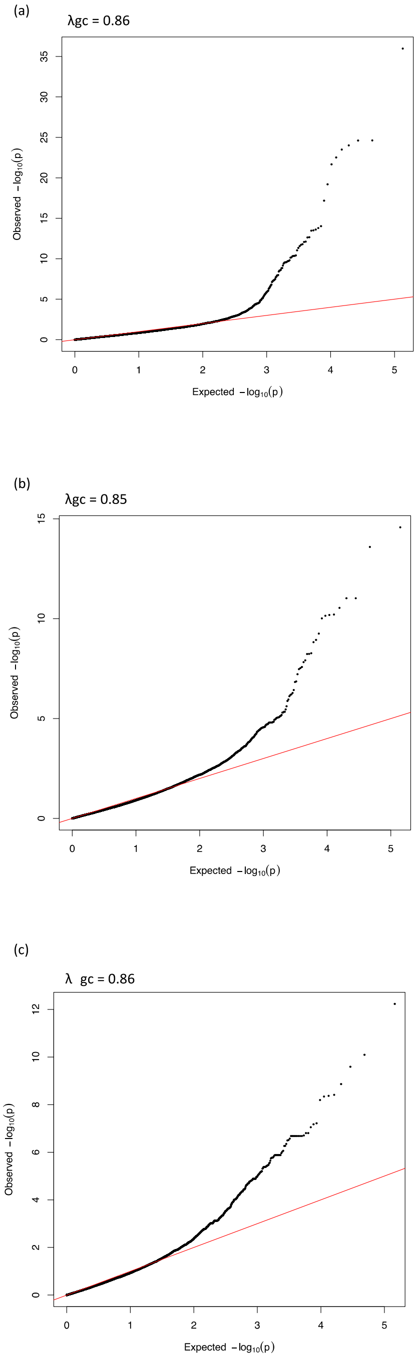
dnaN	2	0	2	0
ftsL	2	0	1	2
glgB	2	0	2	0
glgX_1	2	0	2	0
glmM	2	0	2	0
ileS	2	0	2	0
lepA_1	2	0	2	0
licT_2	2	0	2	0
lptB	2	0	2	0
manZ_1	2	0	2	0
map	2	0	2	0
mapZ	2	0	2	0
metF	2	0	2	0
mshD_1	2	0	2	0
nrdE1	2	0	2	0
nrdF	2	0	2	0
nrdH	2	0	2	0
pox5	2	0	2	0
thrS	2	0	2	0
truB	2	0	2	0
yccU	2	0	2	0
yesO	2	0	2	0
addA	1	0	1	0
aguA	1	0	1	0
apaH	1	0	1	0
araQ_1	1	0	1	0
arcB	1	0	1	0
artM_1	1	0	1	0
bbmA	1	0	1	0
bfmBAB	1	0	1	0
bglF	1	0	1	0
btuD_4	1	0	1	0
btuD_5	1	0	1	0
clcA	1	0	1	0
comEA	1	0	1	0
cpoA	1	0	1	0
ctpE	1	0	1	0
cysM	1	0	1	0
cysS	1	0	1	0
dapA	1	0	1	0
dnaA	1	0	1	0
dnaG	1	0	1	0
dnaK	1	0	1	0
ettA_1	1	0	1	0

exoA	1	0	1	0
femA	1	0	1	0
fmt	1	0	1	0
fruA	1	0	1	0
fucI	1	0	1	0
gap	1	0	1	0
gatA	1	0	1	0
glgD	1	0	1	0
glmS	1	0	1	0
glpO	1	0	1	0
gmuE	1	0	1	0
gmuR	1	0	1	0
gnd	1	0	1	0
gph_1	1	0	1	0
graS	1	0	1	0
gyrB	1	0	1	0
hisS	1	0	1	0
hit	1	0	1	0
ilvD	1	0	1	0
lacX	1	0	1	0
lacZ	1	0	1	0
licA_1	1	0	1	0
lrp	1	0	1	0
malQ	1	0	1	0
mepA	1	0	1	0
metI	1	0	1	0
miaA	1	0	1	0
mro	1	0	1	0
msmE	1	0	1	0
mtcA1	1	0	1	0
murD	1	0	1	0
murF	1	0	1	0
murI	1	0	1	0
niaR	1	0	1	0
nimT	1	0	1	0
nusG	1	0	1	0
pcrA	1	0	1	0
pdg	1	0	1	0
pepF1_2	1	0	1	0
pepN	1	0	1	0
pepO	1	0	1	0
pepV	1	0	1	0
pepX	1	0	1	0
pfbA	1	0	1	0

pnp	1	0	1	0
potA_2	1	0	1	0
potB	1	0	1	0
psaA	1	0	1	0
pucK	1	0	1	0
pyrB	1	0	1	0
pyrC	1	0	1	0
pyrDA	1	0	1	0
pyrK_1	1	0	1	0
recO	1	0	1	0
relA	1	0	1	0
ribF	1	0	1	0
rplF	1	0	1	0
rplJ	1	0	1	0
rplM	1	0	1	0
rsgA	1	0	1	0
rsmB	1	0	1	0
ruvB_2	1	0	1	0
sacA	1	0	1	0
salL	1	0	1	0
sarA_1	1	0	1	0
sasA_1	1	0	1	0
secY	1	0	1	0
sorB_1	1	0	1	0
spo0J	1	0	1	0
stkP	1	0	1	0
tag	1	0	1	0
tdk	1	0	1	0
trkA	1	0	1	0
trpF	1	0	1	0
trpS2	1	0	1	0
tuf	1	0	1	0
whiA	1	0	1	0
xpt	1	0	1	0
ybhL	1	0	1	0
yhdG	1	0	1	0
yheI	1	0	1	0
yicL_1	1	0	1	0
yqeN	1	0	1	0
yteP_1	1	0	1	0
yumC	1	0	1	0
yxdL_1	1	0	1	0
znuA	1	0	1	0

Appendix K QQ plots for the GWAS of cotrimoxazole, erythromycin and clindamycin resistance

QQ plots showing p-values from respective GWAS analysis (a) cotrimoxazole (b) erythromycin (c) clindamycin. Each plot compares the distribution of -log(p-values) observed in the analysis with the expected distribution under the null hypothesis. The red line is the y=x reference line and lambda GC (λgc) values shown for each.



Appendix L Summary of SNPs in common genes identified in this GWAS for cotrimoxazole resistance and the GWAS for cotrimoxazole resistance by Mobegi *et al* (2017) (255)

SNP position	P value-(log10) in this GWAS	P value(log10) in Mobegi *et al* (2017)	Gene	Product
264869	17.2	NA	*folP/sulA*	dihydropteroate synthase
264912	11.8	75.0	folP/sulA	dihydropteroate synthase
264975	11.6	NA	folP/sulA	dihydropteroate synthase
264978	13.6	NA	*folP/sulA*	dihydropteroate synthase
264981	8.6	NA	*folP/sulA*	dihydropteroate synthase
264987	21.7	NA	folP/sulA	dihydropteroate synthase
265071	10.4	NA	*folP/sulA*	dihydropteroate synthase
265128	14.0	NA	folP/sulA	dihydropteroate synthase
265159	24.6	NA	folP/sulA	dihydropteroate synthase
265242	12.1	NA	folP/sulA	dihydropteroate synthase
265251	8.6	NA	folP/sulA	dihydropteroate synthase
265257	10.1	NA	*folP/sulA*	dihydropteroate synthase
265281	10.4	NA	*folP/sulA*	dihydropteroate synthase
265293	8.5	NA	folP/sulA	dihydropteroate synthase
265349	10.4	NA	folP/sulA	dihydropteroate synthase
265373	12.7	NA	folP/sulA	dihydropteroate synthase
265375	12.1	NA	folP/sulA	dihydropteroate synthase
265404	24.6	NA	folP/sulA	dihydropteroate synthase
265406	36.0	NA	folP/sulA	dihydropteroate synthase
265407	11.0	NA	folP/sulA	dihydropteroate synthase
265424	22.5	NA	folP/sulA	dihydropteroate synthase
265452	9.8	NA	folP/sulA	dihydropteroate synthase
265482	9.6	NA	folP/sulA	dihydropteroate synthase

265485	8.8	NA	folP/sulA	dihydropteroate synthase
265536	9.6	66.4	folP/sulA	dihydropteroate synthase
265554	10.3	NA	folP/sulA	dihydropteroate synthase
265608	11.7	NA	folP/sulA	dihydropteroate synthase
265779	NA	23.2	folP/sulA	dihydropteroate synthase
265910	10.2	NA	*fpgS/folC*	folylpolyglutamate synthase
265927	9.5	NA	*fpgS/folC*	folylpolyglutamate synthase
265930	9.6	NA	*fpgS/folC*	folylpolyglutamate synthase
266035	NA	9.7	*fpgS/folC*	folylpolyglutamate synthase
267011	NA	36.9	*fpgS/folC*	folylpolyglutamate synthase
1532230	NA	20.3	dhfR/dyr	dihydrofolate reductase
1532245	NA	80.6	dhfR/dyr	dihydrofolate reductase
1532301	13.5	NA	dhfR/dyr	dihydrofolate reductase
1532326	NA	131.8	dhfR/dyr	dihydrofolate reductase
1532371	12.6	NA	dhfR/dyr	dihydrofolate reductase
1532406	19.2	NA	dhfR/dyr	dihydrofolate reductase
1532410	8.4	NA	dhfR/dyr	dihydrofolate reductase
1532458	9.4	NA	dhfR/dyr	dihydrofolate reductase
1532608	9.7	NA	dhfR/dyr	dihydrofolate reductase
1532689	NA	23.2	dhfR/dyr	dihydrofolate reductase

Appendix M Summary of SNPs in common genes identified in this GWAS for erythromycin resistance and the GWAS for erythromycin resistance by Mobegi *et al* (2017) (255)

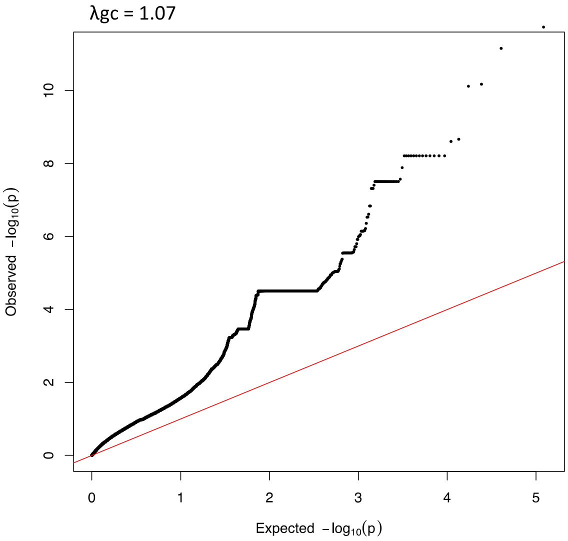
SNP position	P value-(log10) in this GWAS	P value(log10) in Mobegi *et al* (2017)	Gene	Product
291982	NA	15.9	*pbpX*	Penicillin-binding protein
292017	NA	15.5	*pbpX*	Penicillin-binding protein
293251	8.9	NA	*pbpX*	Penicillin-binding protein
293640	10.2	NA	*pbpX*	Penicillin-binding protein
293646	13.6	NA	*pbpX*	Penicillin-binding protein
293685	10.5	NA	*pbpX*	Penicillin-binding protein
636304	NA	8.7	*ykoD*	ABC transporter ATP-binding protein
636402	11.0	NA	*ykoD*	ABC transporter ATP-binding protein
637379	NA	9.9	*ykoD*	ABC transporter ATP-binding protein
884979	NA	11.1	*glmU*	Bifunctional protein
885289	NA	13.2	*glmU*	Bifunctional protein
885387	11.0	NA	*glmU*	Bifunctional protein
1392986	10.1	NA	*apbE*	ApbE family protein
1393408	NA	11.3	*apbE*	ApbE family protein

Appendix N QQ plots for the GWAS of chloramphenicol resistance

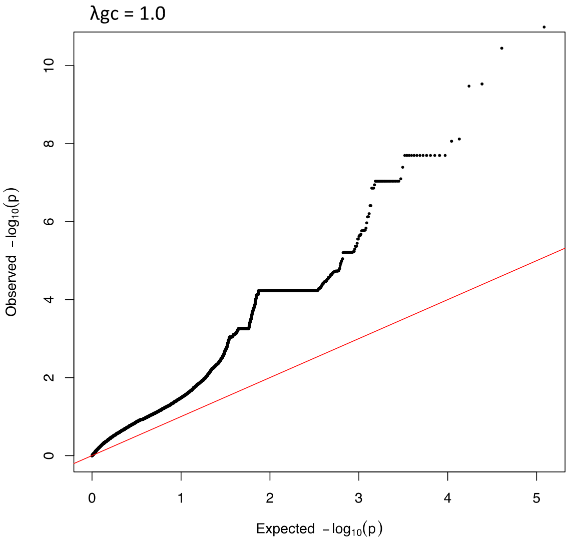
(a) QQ plot showing original p-values following the GWAS analysis of 178 isolates for chloramphenicol resistance. (b) QQ plot showing the p-values after they have been corrected by lambda GC. Each plot compares the distribution of -log(p-values) observed in the analysis with the expected distribution under the null hypothesis. The red line is the y=x reference line and lambda GC (λgc) values shown for each.

)

(a

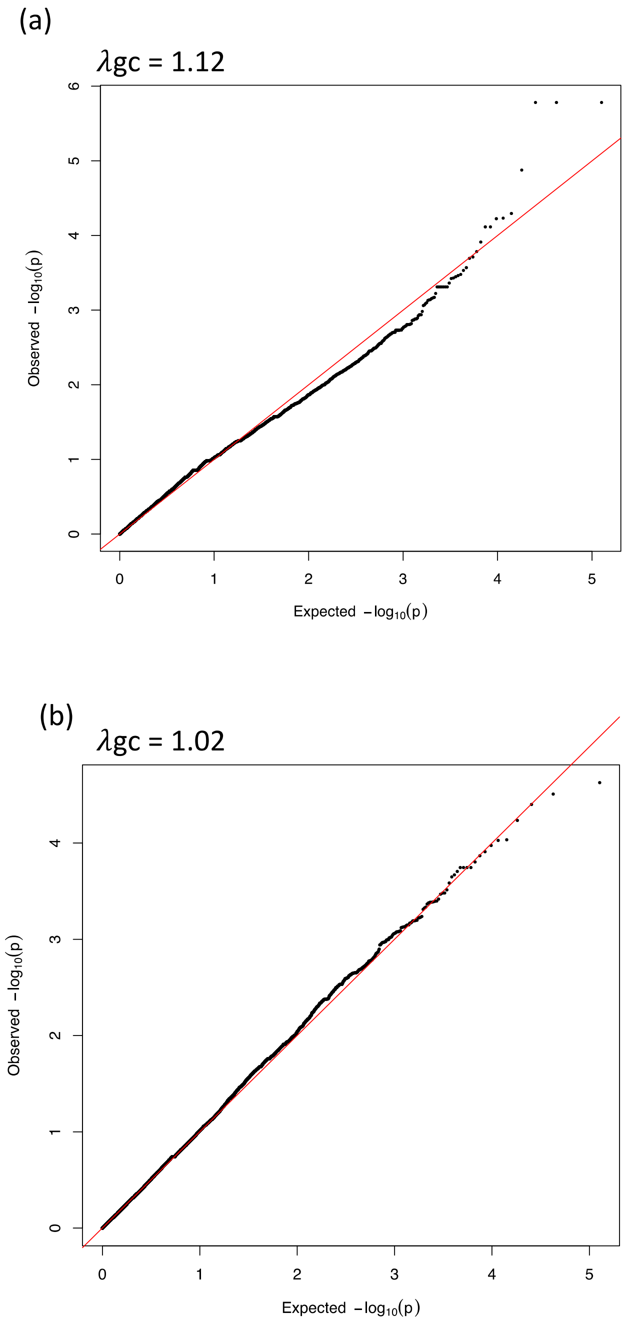


(b)



Appendix O QQ plots for the GWAS of doxycycline and tetracycline resistance

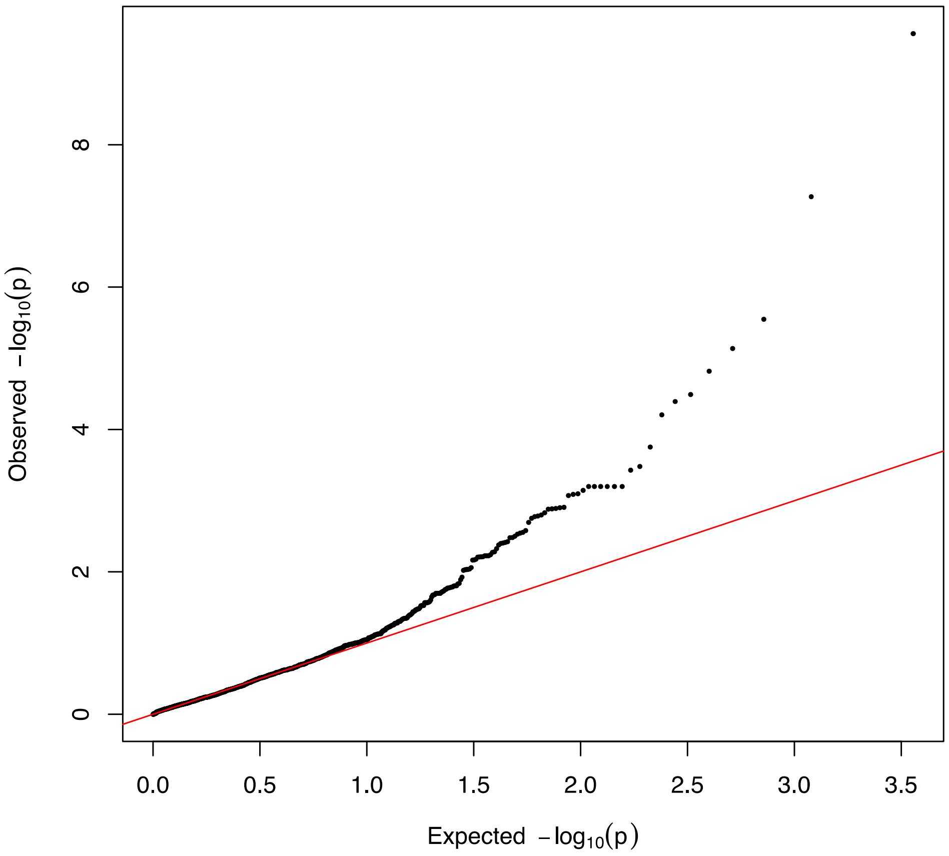
(a) QQ plot showing original p-values following the GWAS analysis of doxycycline resistance. (b) QQ plot showing original p-values following the GWAS analysis of tetracycline resistance. Each plot compares the distribution of -log(p-values) observed in the analysis with the expected distribution under the null hypothesis. The red line is the y=x reference line and lambda GC (λgc) values shown for each.



Appendix P QQ plots for the GWAS of penicillin non-susceptibility using the SvIR classification from recombination-free genomes

QQ plot compares the distribution of -log(p-values) observed in the analysis with the expected distribution under the null hypothesis. The red line is the y=x reference line and lambda GC (λgc) values shown.

λgc = 0.89



Appendix Q Isolate laboratory data for 313 *S. pneumoniae* isolates collected 2013-2016

Laboratory data for the isolates collected from collaborating hospitals 2013-2016. Data for isolates collected 1997-2013 is located in Jauneikaite (2014) (182).

ID	Hospital	Age(yrs)	Gender	Ethnicity	Site	Disease	Date	Vaccination	Outcome	ST	serotype	goeBURST
WBB1259	TTSH	59	M	Chinese	Blood	IPD	23/05/2015	N/A	Died	320	19A	1
WBB1260	TTSH	65	F	Chinese	Blood	IPD	18/03/2015	N/A	Died	2211	12F	28
WBB1261	TTSH	45	M	Chinese	Blood	IPD	25/02/2015	N/A	Survived	989	12F	28
WBB1262	TTSH	73	M	Chinese	Blood	IPD	13/02/2015	N/A	Survived	4216	8	64
WBB1263	TTSH	70	F	Chinese	Blood	IPD	24/02/2015	N/A	Survived	4216	8	64
WBB1264	TTSH	68	M	Chinese	Knee	IPD	19/05/2015	N/A	Survived	433	22F	65
WBB1265	TTSH	41	M	Chinese	Blood	IPD	29/01/2016	N/A	Survived	9	14	9
WBB1266	TTSH	63	F	Chinese	Blood	IPD	13/02/2016	N/A	Survived	9	14	9
WBB1267	TTSH	63	F	Chinese	Sputum	non-IPD	16/02/2016	N/A	Survived	81	23F	0
WBB1268	TTSH	78	F	Malay	Sputum	non-IPD	08/05/2015	N/A	Survived	3791	19F	32
WBB1269	TTSH	76	M	Chinese	Endotracheal tube	non-IPD	02/12/2015	N/A	Survived	236	19F	1
WBB1270	TTSH	56	M	Chinese	Sputum	non-IPD	21/12/2015	Yes	Survived	81	19F	0

WBB1271	TTSH	66	M	Chinese	Endotracheal tube	non-IPD	N/A	N/A	Died	2758	07C	66
WBB1273	TTSH	52	M	Chinese	Blood	IPD	12/03/2016	N/A	Survived	81	23F	0
WBB1274	TTSH	65	M	Chinese	Blood	IPD	30/08/2015	N/A	Survived	5258	34	67
WBB1275	TTSH	67	M	Chinese	Blood	IPD	11/04/2015	N/A	Survived	1233	18B	68
WBB1276	TTSH	75	M	Chinese	Sputum	non-IPD	02/11/2015	N/A	Survived	81	23F	0
WBB1277	TTSH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	NF	3	69
WBB1278	TTSH	90	M	Chinese	Sputum	non-IPD	14/05/2015	N/A	Survived	320	19A	1
WBB1279	TTSH	80	F	Chinese	Blood	IPD	19/11/2015	N/A	N/A	53	8	70
WBB1280	TTSH	59	M	Chinese	Blood	IPD	06/09/2015	N/A	Died	320	19A	1
WBB1281	TTSH	72	M	Chinese	Sputum	non-IPD	NA	N/A	Died	81	19F	0
WBB1282	TTSH	34	F	Malay	Blood	IPD	28/10/2015	N/A	Survived	289	5	63
WBB1283	TTSH	72	M	Chinese	Blood	IPD	31/10/2015	N/A	Survived	180	3	51
WBB1284	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	383	16F	71
WBB1285	TTSH	70	M	Chinese	Sputum	non-IPD	29/02/2016	N/A	Survived	5832	06D	72
WBB1286	TTSH	58	F	Chinese	Blood	IPD	26/03/2015	N/A	Died	11916	16F	73
WBB1288	TTSH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	5407	38	24
WBB1289	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	383	16F	71
WBB1290	TTSH	66	M	Chinese	Blood	IPD	18/03/2016	N/A	Survived	4745	20	74

WBB1291	TTSH	88	F	Chinese	Blood	IPD	14/09/2015	N/A	Survived	6195	23F	53
WBB1292	TTSH	86	M	Chinese	Blood	IPD	18/08/2015	N/A	Survived	NF	34	69
WBB1293	TTSH	31	F	Other	Nose	non-IPD	08/03/2016	N/A	Survived	NF	19A	69
WBB1294	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	NF	06C	69
WBB1295	TTSH	40	F	Other	CSF	IPD	16/02/2016	N/A	Survived	180	3	51
WBB1296	TTSH	65	M	Chinese	BAL	non-IPD	17/02/2016	N/A	Survived	NF	34	69
WBB1297	TTSH	55	F	Chinese	Nose	non-IPD	25/02/2016	N/A	Survived	2572	23A	75
WBB1298	TTSH	79	M	Chinese	Endotracheal tube	non-IPD	08/03/2016	N/A	Survived	2924	06C	76
WBB1299	TTSH	40	F	Other	Blood	IPD	16/02/2016	N/A	Survived	180	3	51
WBB1300	TTSH	60	M	Chinese	Blood	IPD	26/01/2015	N/A	Survived	146	06E	77
WBB1301	TTSH	79	F	Chinese	Blood	IPD	20/03/2016	Yes	Survived	4745	20	74
WBB1302	TTSH	50	M	Indian	Blood	IPD	28/07/2015	N/A	Survived	NF	23F	69
WBB1303	TTSH	50	M	Bangladesh	Blood	IPD	28/07/2015	No	Survived	8958	40	78
WBB1304	TTSH	54	M	Indian	Blood	IPD	28/12/2015	N/A	Survived	9325	38	15
WBB1305	TTSH	37	M	Malay	Blood	IPD	01/01/2016	N/A	Survived	695	19A	79
WBB1306	TTSH	37	M	Chinese	Blood	IPD	03/12/2015	N/A	Survived	63	14	7
WBB1307	TTSH	53	M	Chinese	Blood	IPD	22/12/2015	N/A	Survived	138	06A	33
WBB1308	TTSH	60	M	Chinese	Blood	IPD	26/11/2015	N/A	Survived	NF	8	69

WBB1309	TTSH	80	M	Chinese	Blood	IPD	26/11/2015	Yes	Survived	2234	8	80
WBB1310	TTSH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	NF	19F	69
WBB1311	TTSH	74	M	Malay	Blood	IPD	10/08/2015	N/A	Died	505	3	81
WBB1313	TTSH	77	M	Chinese	Blood	IPD	22/06/2015	Yes	Survived	2213	4	41
WBB1314	TTSH	65	M	Chinese	CSF	IPD	NA	N/A	Died	3544	07A	29
WBB1315	TTSH	91	M	Indian	Sputum	non-IPD	17/06/2015	N/A	Survived	320	19A	1
WBB1316	TTSH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	NF	23A	69
WBB1317	TTSH	60	M	Chinese	Pleura	IPD	19/05/2015	N/A	Died	7479	15B	30
WBB1318	TTSH	80	F	Indian	Endotracheal tube	non-IPD	29/01/2015	N/A	Survived	6028	15F	82
WBB1319	TTSH	83	M	Chinese	Blood	IPD	11/05/2015	N/A	Survived	NF	34	69
WBB1320	TTSH	84	M	Chinese	Endotracheal tube	non-IPD	27/01/2015	N/A	Died	1591	15A	0
WBB1322	TTSH	70	M	Chinese	Blood	IPD	20/01/2015	N/A	Survived	3500	8	83
WBB1324	TTSH	69	M	Indian	Endotracheal tube	non-IPD	06/01/2016	N/A	Survived	143	14	49
WBB1325	TTSH	72	F	Chinese	Sputum	non-IPD	01/02/2016	Yes	Survived	1591	15A	0
WBB1326	TTSH	26	M	Indian	Blood	IPD	05/01/2015	Yes	Survived	12474	19A	84
WBB1327	TTSH	70	M	Indian	Blood	IPD	11/01/2015	N/A	Died	180	3	51
WBB1328	NUH	N/A	M	N/A	Blood	IPD	11/11/2013	N/A	N/A	6193	3	85
WBB1329	NUH	N/A	F	N/A	Blood	IPD	26/11/2013	N/A	N/A	NF	19A	69

WBB1330	NUH	N/A	F	N/A	Blood	IPD	05/12/2013	N/A	N/A	180	3	51
WBB1331	NUH	N/A	M	N/A	Blood	IPD	10/12/2013	N/A	N/A	NF	39	69
WBB1332	NUH	N/A	F	N/A	Blood	IPD	21/12/2013	N/A	N/A	81	23F	0
WBB1333	NUH	N/A	M	N/A	Blood	IPD	22/04/2014	N/A	N/A	7569	3	43
WBB1334	NUH	N/A	F	N/A	Blood	IPD	24/04/2014	N/A	N/A	63	14	7
WBB1335	NUH	N/A	M	N/A	Blood	IPD	19/07/2014	N/A	N/A	NF	38	69
WBB1336	NUH	N/A	F	N/A	Blood	IPD	07/08/2014	N/A	N/A	558	35B	86
WBB1337	NUH	N/A	M	N/A	Blood	IPD	02/09/2014	N/A	N/A	NF	22F	69
WBB1338	NUH	N/A	M	N/A	Blood	IPD	12/09/2014	N/A	N/A	2854	19F	87
WBB1339	NUH	N/A	M	N/A	Blood	IPD	19/09/2014	N/A	N/A	2234	8	80
WBB1340	NUH	N/A	F	N/A	Portacath	non-IPD	23/09/2014	N/A	N/A	NF	13	69
WBB1341	NUH	N/A	M	N/A	Blood	IPD	07/11/2014	N/A	N/A	458	3	88
WBB1342	NUH	N/A	M	N/A	Blood	IPD	11/11/2014	N/A	N/A	4745	20	74
WBB1343	NUH	N/A	F	N/A	Blood	IPD	01/12/2014	N/A	N/A	193	17F	42
WBB1344	NUH	N/A	F	N/A	Blood	IPD	06/12/2014	N/A	N/A	193	15C	42
WBB1345	NUH	N/A	F	N/A	Blood	IPD	15/12/2014	N/A	N/A	439	23B	4
WBB1346	NUH	N/A	F	N/A	Blood	IPD	16/12/2014	N/A	N/A	81	23F	0
WBB1347	NUH	N/A	M	N/A	Sputum	non-IPD	05/01/2014	N/A	N/A	218	07A	29

WBB1348	NUH	N/A	F	N/A	Unknown	N/A	30/01/2014	N/A	N/A	3398	28F	89
WBB1349	NUH	N/A	F	N/A	Unknown	N/A	30/01/2014	N/A	N/A	3398	28F	89
WBB1350	NUH	N/A	M	N/A	Endotracheal tube	non-IPD	23/04/2014	N/A	N/A	7569	3	43
WBB1351	NUH	N/A	M	N/A	Blood	IPD	16/06/2014	N/A	N/A	3804	11C	90
WBB1352	NUH	N/A	F	N/A	Sputum	non-IPD	25/06/2014	N/A	N/A	320	19A	1
WBB1353	NUH	N/A	F	N/A	Nose	non-IPD	06/08/2014	N/A	N/A	10106	15A	91
WBB1354	NUH	N/A	M	N/A	Sputum	non-IPD	08/09/2014	N/A	N/A	5000	15A	92
WBB2033	TTSH	64	M	Malay	Pleura	IPD	NA	N/A	Survived	5872	4	31
WBB2034	NUH	N/A	F	N/A	Blood	IPD	28/06/2015	N/A	N/A	4127	4	20
WBB2035	NUH	N/A	F	N/A	Sputum	non-IPD	16/06/2015	N/A	N/A	6030	16F	93
WBB2036	NUH	N/A	F	N/A	Sputum	non-IPD	03/12/2015	N/A	N/A	320	19A	1
WBB2037	TTSH	68	M	Chinese	Blood	IPD	07/09/2014	Yes	Survived	63	14	7
WBB2038	TTSH	21	M	Malay	Sputum	non-IPD	13/03/2014	N/A	Survived	271	19F	1
WBB2039	TTSH	67	M	Chinese	Blood	IPD	17/07/2014	Yes	Survived	989	12F	28
WBB2040	TTSH	73	M	Chinese	BAL	non-IPD	06/12/2013	N/A	Survived	1262	15C	94
WBB2041	NUH	N/A	M	N/A	Blood	IPD	20/09/2014	N/A	N/A	NF	15B	69
WBB2042	NUH	N/A	F	N/A	Blood	IPD	01/08/2015	N/A	N/A	63	15A	7
WBB2043	NUH	N/A	M	N/A	Endotracheal tube	non-IPD	06/07/2015	N/A	N/A	NF	4	69

WBB2044	NUH	N/A	M	N/A	Sputum	non-IPD	21/03/2016	N/A	N/A	458	3	88
WBB2045	TTSH	71	F	Chinese	Blood	IPD	06/12/2014	Yes	Survived	876	14	5
WBB2046	TTSH	34	F	Other	Ear	non-IPD	11/08/2014	No	Survived	320	19A	1
WBB2047	TTSH	71	F	Malay	Blood	IPD	23/05/2014	N/A	Survived	989	12F	28
WBB2048	TTSH	43	M	Chinese	Blood	IPD	25/09/2013	N/A	Survived	2674	19A	8
WBB2050	NUH	N/A	M	N/A	Blood	IPD	21/08/2015	N/A	N/A	2754	13	61
WBB2051	NUH	N/A	M	N/A	Endotracheal tube	non-IPD	14/07/2015	N/A	N/A	439	23B	4
WBB2052	NUH	N/A	M	N/A	Sputum	non-IPD	11/12/2015	N/A	N/A	95	06E	2
WBB2053	TTSH	76	M	Chinese	Blood	IPD	21/11/2014	N/A	Survived	338	23A	3
WBB2054	TTSH	35	M	Chinese	Blood	IPD	21/01/2014	Yes	Survived	NF	3	69
WBB2055	TTSH	65	F	Chinese	Blood	IPD	28/12/2013	Yes	Survived	90	06E	2
WBB2056	TTSH	52	M	Malay	Blood	IPD	16/09/2013	No	Died	311	23F	4
WBB2059	NUH	N/A	M	N/A	Sputum	non-IPD	29/07/2015	N/A	N/A	180	3	51
WBB2060	NUH	N/A	M	N/A	Sputum	non-IPD	28/03/2016	N/A	N/A	62	11D	17
WBB2061	TTSH	49	M	Chinese	Blood	IPD	17/11/2014	No	Died	1263	09V	6
WBB2062	TTSH	72	M	Chinese	Sputum	non-IPD	31/07/2014	N/A	Survived	320	19A	1
WBB2063	TTSH	65	F	Chinese	Blood	IPD	11/08/2013	No	Survived	1518	06E	56
WBB2064	TTSH	55	F	Chinese	Sputum	non-IPD	22/11/2013	N/A	Survived	NF	23F	69

WBB2065	NUH	N/A	M	N/A	Sputum	non-IPD	21/12/2014	N/A	N/A	6441	48	95
WBB2066	NUH	N/A	M	N/A	Blood	IPD	13/10/2015	N/A	N/A	NF	06A	69
WBB2067	NUH	N/A	F	N/A	Endotracheal tube	non-IPD	04/08/2015	N/A	N/A	1591	15A	0
WBB2068	TTSH	56	M	Malay	Blood	IPD	14/08/2014	Yes	Survived	1518	06E	56
WBB2069	TTSH	71	M	Chinese	Blood	IPD	31/10/2014	No	Died	6193	3	85
WBB2070	TTSH	N/A	F	Chinese	Arm	non-IPD	28/03/2014	No	Survived	320	19F	1
WBB2071	TTSH	61	M	Chinese	Endotracheal tube	non-IPD	18/11/2013	No	Survived	695	19A	79
WBB2072	TTSH	40	M	Indian	CSF	IPD	16/11/2013	No	Survived	236	19F	1
WBB2073	NUH	N/A	M	N/A	Blood	IPD	01/01/2015	N/A	N/A	3214	19A	96
WBB2074	NUH	N/A	M	N/A	Endotracheal tube	non-IPD	06/01/2015	N/A	N/A	NF	11A	69
WBB2075	NUH	N/A	M	N/A	Nose	non-IPD	06/08/2015	N/A	N/A	902	06A	97
WBB2076	TTSH	29	M	Chinese	Bronchus	non-IPD	05/11/2014	N/A	Survived	320	19A	1
WBB2077	TTSH	30	M	Malay	Blood	IPD	26/10/2014	No	Survived	2754	33B	61
WBB2078	TTSH	63	M	Chinese	Blood	IPD	10/03/2014	No	Survived	12902	3	98
WBB2079	TTSH	64	F	Chinese	Eye	non-IPD	10/09/2013	N/A	Survived	320	19A	1
WBB2080	KKH	5	M	Malay	Pleura	IPD	08/07/2013	N/A	N/A	320	19A	1
WBB2081	NUH	N/A	M	N/A	Blood	IPD	05/01/2015	N/A	N/A	236	19F	1
WBB2083	NUH	N/A	F	N/A	Throat	non-IPD	21/08/2015	N/A	N/A	446	35F	39

WBB2084	TTSH	49	M	Indian	Hand	non-IPD	17/11/2014	N/A	Survived	5407	25A	24
WBB2085	TTSH	55	F	Chinese	Blood	IPD	N/A	No	Survived	876	14	5
WBB2087	TTSH	60	M	Chinese	Blood	IPD	26/10/2013	No	Survived	NF	23F	69
WBB2088	KKH	7	F	Chinese	Blood	IPD	02/08/2013	N/A	N/A	NF	17F	69
WBB2089	NUH	N/A	M	N/A	Blood	IPD	06/01/2015	N/A	N/A	180	3	51
WBB2090	NUH	N/A	M	N/A	Nose	non-IPD	25/02/2015	N/A	N/A	236	19F	1
WBB2091	NUH	N/A	M	N/A	Sputum	non-IPD	22/08/2015	N/A	N/A	1553	23F	99
WBB2092	TTSH	57	F	Chinese	Blood	IPD	14/08/2014	N/A	Survived	310	38	15
WBB2093	TTSH	74	M	Chinese	Blood	IPD	NA	No	Survived	6202	15F	100
WBB2094	TTSH	85	M	Chinese	Sputum	non-IPD	14/05/2014	No	Survived	320	19F	1
WBB2095	TTSH	80	M	Chinese	Nose	non-IPD	08/10/2013	N/A	Survived	338	23A	3
WBB2096	KKH	4	M	Malay	Ear	non-IPD	07/08/2013	N/A	N/A	4908	09V	101
WBB2097	NUH	N/A	M	N/A	Blood	IPD	14/02/2015	N/A	N/A	6202	15F	100
WBB2100	TTSH	80	M	Chinese	Blood	IPD	30/12/2014	N/A	Died	NF	07A	69
WBB2101	TTSH	75	M	Chinese	Blood	IPD	15/10/2014	No	Died	180	3	51
WBB2102	TTSH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	271	19F	1
WBB2104	KKH	4	M	Indian	Ear	non-IPD	07/09/2013	N/A	N/A	1464	19F	1
WBB2105	NUH	N/A	M	N/A	Blood	IPD	21/04/2015	N/A	N/A	6197	8	102

WBB2106	NUH	N/A	F	N/A	Unknown	N/A	10/04/2015	N/A	N/A	6025	11D	103
WBB2107	NUH	N/A	M	N/A	Sputum	non-IPD	04/11/2015	N/A	N/A	11454	3	104
WBB2108	TTSH	67	M	Chinese	Blood	IPD	11/09/2014	N/A	Survived	5246	06D	3
WBB2109	TTSH	86	M	Chinese	Sputum	non-IPD	11/08/2014	Yes	Survived	320	19A	1
WBB2110	TTSH	45	M	Chinese	Blood	IPD	01/02/2014	No	Survived	3544	07A	29
WBB2111	TTSH	68	M	Chinese	Blood	IPD	15/12/2013	No	Survived	NF	4	69
WBB2112	KKH	10	M	Indian	Blood	IPD	18/10/2013	N/A	N/A	81	23F	0
WBB2113	NUH	N/A	F	N/A	Blood	IPD	23/05/2015	N/A	N/A	191	07A	105
WBB2114	NUH	N/A	M	N/A	Sputum	non-IPD	14/04/2015	N/A	N/A	880	23F	11
WBB2115	NUH	N/A	M	N/A	BAL	non-IPD	23/11/2015	N/A	N/A	1379	06C	106
WBB2116	TTSH	56	M	Chinese	Blood	IPD	21/09/2014	N/A	Survived	NF	3	69
WBB2117	TTSH	68	M	Chinese	Blood	IPD	22/01/2014	Yes	Survived	99	11F	107
WBB2118	TTSH	55	F	Malay	Blood	IPD	06/02/2014	No	Survived	9192	19F	108
WBB2119	TTSH	46	M	Malay	Blood	IPD	07/09/2013	N/A	Died	989	12F	28
WBB2120	KKH	8	F	Chinese	Blood	IPD	25/10/2013	N/A	N/A	63	15A	7
WBB2121	NUH	N/A	F	N/A	Blood	IPD	05/06/2015	N/A	N/A	338	23A	3
WBB2122	NUH	N/A	M	N/A	Endotracheal tube	non-IPD	22/04/2015	N/A	N/A	199	15B	5
WBB2123	NUH	N/A	F	N/A	Nose	non-IPD	03/12/2015	N/A	N/A	1464	19F	1

WBB2124	TTSH	68	M	Chinese	Blood	IPD	07/09/2014	Yes	Survived	63	14	7
WBB2125	TTSH	15	M	Chinese	Nose	non-IPD	06/03/2014	No	Survived	320	19A	1
WBB2126	TTSH	63	M	Chinese	Blood	IPD	19/01/2014	Yes	Died	855	06A	109
WBB2127	TTSH	34	M	Other	Blood	IPD	16/08/2013	No	Survived	81	06A	0
WBB2128	KKH	8	F	Chinese	Pleura	IPD	08/11/2013	N/A	N/A	320	19A	1
WBB2144	TTSH	50	M	Indian	Blood	IPD	16/06/2015	N/A	Survived	4216	8	64
WBB2145	KKH	7	M	Indian	Blood	IPD	05/02/2014	N/A	N/A	320	19A	1
WBB2146	KKH	5	M	Chinese	Pleura	IPD	02/07/2014	N/A	N/A	4154	19A	1
WBB2147	KKH	5	M	Chinese	Blood	IPD	17/05/2015	N/A	N/A	338	23A	3
WBB2148	KKH	4	F	Indian	Blood	IPD	27/09/2015	N/A	N/A	193	15B	42
WBB2149	KKH	7	M	Chinese	Blood	IPD	NA	N/A	N/A	62	11D	17
WBB2150	KKH	8	F	Chinese	Blood	IPD	NA	N/A	N/A	386	06C	110
WBB2151	TTSH	51	M	Chinese	Blood	IPD	18/11/2013	No	Died	311	23F	4
WBB2153	KKH	5	F	Chinese	Blood	IPD	09/02/2014	N/A	N/A	63	19A	7
WBB2154	KKH	5	M	Malay	Ear	non-IPD	12/07/2014	N/A	N/A	320	19A	1
WBB2155	KKH	9	F	Malay	BAL	non-IPD	01/06/2015	N/A	N/A	62	11D	17
WBB2156	KKH	5	M	Chinese	Blood	IPD	15/10/2015	N/A	N/A	320	19A	1
WBB2157	KKH	1	F	Chinese	Ear	non-IPD	NA	N/A	N/A	320	19A	1

Appendix Q

WBB2158	KKH	3	M	Malay	Blood	IPD	NA	N/A	N/A	320	19A	1
WBB2159	TTSH	92	M	Chinese	Blood	IPD	19/12/2013	N/A	Died	416	19A	111
WBB2161	KKH	5	M	Malay	Lung	IPD	23/02/2014	N/A	N/A	320	19A	1
WBB2162	KKH	8	M	Chinese	Ear	non-IPD	16/07/2014	N/A	N/A	180	3	51
WBB2163	KKH	10	F	Chinese	BAL	non-IPD	03/06/2015	N/A	N/A	282	06C	0
WBB2164	KKH	7	F	Chinese	Blood	IPD	20/10/2015	N/A	N/A	695	19A	79
WBB2165	KKH	8	M	Chinese	Blood	IPD	NA	N/A	N/A	320	19A	1
WBB2166	KKH	4	F	Indian	CSF	IPD	NA	N/A	N/A	3280	15B	30
WBB2167	TTSH	79	M	Chinese	Blood	IPD	10/09/2013	Yes	Survived	1518	06E	56
WBB2168	TTSH	34	F	Other	Blood	IPD	22/05/2014	No	Survived	289	5	63
WBB2169	KKH	12	M	Indian	Blood	IPD	05/03/2014	N/A	N/A	320	19A	1
WBB2170	KKH	5	F	Chinese	Pleura	IPD	06/08/2014	N/A	N/A	320	19A	1
WBB2171	KKH	4	M	Indian	Blood	IPD	21/06/2015	N/A	N/A	5068	18B	38
WBB2172	KKH	4	M	Chinese	Blood	IPD	03/12/2015	N/A	N/A	NF	3	69
WBB2173	KKH	6	F	Malay	Blood	IPD	NA	N/A	N/A	338	23A	3
WBB2174	KKH	54	F	Other	Unknown	N/A	NA	N/A	N/A	172	23F	3
WBB2175	TTSH	61	M	Malay	Blood	IPD	17/02/2014	N/A	Survived	NF	4	69
WBB2176	KKH	6	M	Chinese	Pleura	IPD	14/03/2014	N/A	N/A	180	3	51

WBB2177	KKH	5	F	Chinese	BAL	non-IPD	24/09/2014	N/A	N/A	7786	N/A	48
WBB2178	KKH	2	F	Malay	Ear	non-IPD	20/06/2015	N/A	N/A	3017	19A	112
WBB2179	KKH	43	F	Chinese	Nipple	non-IPD	16/01/2016	N/A	N/A	558	35B	86
WBB2180	KKH	3	F	Chinese	Blood	IPD	NA	N/A	N/A	9	14	9
WBB2181	KKH	4	F	Chinese	Blood	IPD	NA	N/A	N/A	180	3	51
WBB2182	TTSH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4127	4	20
WBB2183	KKH	3	M	Malay	Ear	non-IPD	10/03/2014	N/A	N/A	15	06B	9
WBB2184	KKH	7	F	Malay	Blood	IPD	10/10/2014	N/A	N/A	8202	19A	1
WBB2185	KKH	15	M	Malay	Blood	IPD	26/06/2015	N/A	N/A	104	06E	2
WBB2186	KKH	1	M	Caucasian	Ear	non-IPD	12/02/2016	N/A	N/A	1451	19A	1
WBB2187	KKH	9	F	Chinese	Blood	IPD	NA	N/A	N/A	NF	06D	69
WBB2188	KKH	7	M	Chinese	Pelvis	IPD	NA	N/A	N/A	7768	15B	113
WBB2189	TTSH	64	M	Chinese	Sputum	non-IPD	11/11/2013	N/A	Survived	386	06C	110
WBB2190	KKH	17	F	Chinese	Unknown	N/A	11/03/2014	N/A	N/A	6945	12F	114
WBB2191	KKH	5	F	Chinese	Blood	IPD	23/10/2014	N/A	N/A	9	14	9
WBB2192	KKH	2	F	Chinese	Ear	non-IPD	07/07/2015	N/A	N/A	6011	3	6
WBB2193	KKH	2	M	Malay	Ear	non-IPD	19/02/2016	N/A	N/A	180	3	51
WBB2194	KKH	39	F	Chinese	CSF	IPD	NA	N/A	N/A	4532	14	7

WBB2195	KKH	3	F	Malay	Blood	IPD	NA	N/A	N/A	2854	19F	87
WBB2196	KKH	7	M	Chinese	Blood	IPD	17/11/2013	N/A	N/A	695	19A	79
WBB2197	KKH	6	M	Malay	Blood	IPD	01/04/2014	N/A	N/A	62	11D	17
WBB2198	KKH	3	M	Caucasian	Blood	IPD	24/12/2014	N/A	N/A	439	23B	4
WBB2199	KKH	7	F	Chinese	Blood	IPD	13/07/2015	N/A	N/A	458	3	88
WBB2200	KKH	6	M	Malay	Ear	non-IPD	20/02/2016	N/A	N/A	320	19F	1
WBB2201	KKH	27	M	Burmese	Blood	IPD	NA	N/A	N/A	338	23A	3
WBB2202	KKH	5	M	Malay	Blood	IPD	NA	N/A	N/A	NF	06E	69
WBB2203	KKH	5	M	Chinese	Blood	IPD	11/12/2013	N/A	N/A	63	15A	7
WBB2204	KKH	4	F	Malay	Ear	non-IPD	09/04/2014	N/A	N/A	9	14	9
WBB2205	KKH	6	F	Chinese	Blood	IPD	06/02/2015	N/A	N/A	180	3	51
WBB2206	KKH	3	M	Chinese	Blood	IPD	21/07/2015	N/A	N/A	338	23A	3
WBB2207	KKH	3	F	Chinese	Blood	IPD	29/02/2016	N/A	N/A	9	14	9
WBB2208	KKH	19	M	Chinese	Blood	IPD	NA	N/A	N/A	NF	19F	69
WBB2209	KKH	12	F	Nepalese	Endotracheal tube	non-IPD	NA	N/A	N/A	6040	39	115
WBB2210	KKH	7	M	Malay	Ear	non-IPD	28/12/2013	N/A	N/A	320	19A	1
WBB2211	KKH	4	M	Malay	Blood	IPD	17/04/2014	N/A	N/A	902	06A	97
WBB2212	KKH	4	M	Malay	Pleura	IPD	18/03/2015	N/A	N/A	8202	19A	1

WBB2213	KKH	4	M	Chinese	Ear	non-IPD	14/08/2015	N/A	N/A	320	19A	1
WBB2214	KKH	5	M	Chinese	Blood	IPD	05/03/2016	N/A	N/A	63	15A	7
WBB2215	KKH	8	F	Chinese	Blood	IPD	NA	N/A	N/A	6011	3	6
WBB2216	KKH	2	F	Malay	Ear	non-IPD	NA	N/A	N/A	193	15C	42
WBB2217	KKH	8	F	Malay	Blood	IPD	28/01/2014	N/A	N/A	695	19A	79
WBB2218	KKH	15	M	Malay	Pleura	IPD	10/05/2014	N/A	N/A	320	19A	1
WBB2219	KKH	6	F	Malay	Blood	IPD	07/05/2015	N/A	N/A	5242	23A	3
WBB2220	KKH	2	M	Malay	Ear	non-IPD	15/09/2015	N/A	N/A	236	19F	1
WBB2221	KKH	3	M	Chinese	Ear	non-IPD	23/03/2016	N/A	N/A	320	19A	1
WBB2222	KKH	34	F	Malay	Blood	IPD	NA	N/A	N/A	7479	15B	30
WBB2223	KKH	2	F	Chinese	Ear	non-IPD	NA	N/A	N/A	62	11D	17
WBB2224	KKH	12	F	Chinese	Pleura	IPD	31/01/2014	N/A	N/A	320	19A	1
WBB2225	KKH	3	F	Chinese	Blood	IPD	28/06/2014	N/A	N/A	97	10A	27
WBB2226	KKH	1	M	Chinese	Ear	non-IPD	06/05/2015	N/A	N/A	62	11D	17
WBB2227	KKH	6	M	Chinese	Blood	IPD	21/09/2015	N/A	N/A	320	19A	1
WBB2228	KKH	7	M	Chinese	CSF	IPD	NA	N/A	N/A	62	11D	17
WBB2230	TTSH	43	F	Chinese	Eye	non-IPD	10/09/2013	N/A	Survived	62	11A	17
WBB2336	SGH	59	M	Chinese	Blood	IPD	02/10/2013	N/A	N/A	8259	8	116

WBB2337	SGH	31	M	Bangladesh	Blood	IPD	29/07/2014	N/A	N/A	4745	20	74
WBB2338	SGH	84	M	Chinese	Blood	IPD	09/11/2014	N/A	N/A	6202	15A	100
WBB2339	SGH	1	M	Malay	Blood	IPD	13/04/2015	N/A	N/A	3111	19A	58
WBB2340	SGH	57	M	Malay	Blood	IPD	23/02/2016	N/A	N/A	NF	4	69
WBB2342	SGH	68	M	Chinese	Blood	IPD	21/10/2013	N/A	N/A	386	06C	110
WBB2343	SGH	75	M	Chinese	Blood	IPD	15/08/2014	N/A	N/A	458	3	88
WBB2344	SGH	72	F	Chinese	Blood	IPD	16/11/2014	N/A	N/A	6011	3	6
WBB2345	SGH	73	M	Chinese	Blood	IPD	24/05/2015	N/A	N/A	386	3	110
WBB2346	SGH	78	M	Chinese	Blood	IPD	05/03/2016	N/A	N/A	62	11D	17
WBB2348	SGH	80	F	Chinese	Blood	IPD	25/10/2013	N/A	N/A	NF	11D	69
WBB2349	SGH	56	F	Chinese	Blood	IPD	19/08/2014	N/A	N/A	8202	19A	1
WBB2350	SGH	73	M	Chinese	Blood	IPD	09/12/2014	N/A	N/A	260	3	117
WBB2351	SGH	76	F	Malay	Blood	IPD	14/05/2015	N/A	N/A	NF	06A	69
WBB2353	SGH	70	M	Chinese	Nose	non-IPD	29/10/2013	N/A	N/A	2697	19F	1
WBB2354	SGH	33	M	Indian	Blood	IPD	20/08/2014	N/A	N/A	303	1	10
WBB2355	SGH	70	M	Indian	Blood	IPD	14/12/2014	N/A	N/A	36	23F	118
WBB2356	SGH	31	M	Indian	Blood	IPD	03/06/2015	N/A	N/A	303	1	10
WBB2358	SGH	79	M	Chinese	Blood	IPD	11/11/2013	N/A	N/A	90	06E	2

WBB2359	SGH	81	F	Chinese	Blood	IPD	29/08/2014	N/A	N/A	62	11D	17
WBB2360	SGH	75	F	Chinese	Blood	IPD	16/12/2014	N/A	N/A	36	23F	118
WBB2361	SGH	63	F	Malay	Blood	IPD	20/06/2015	N/A	N/A	4745	20	74
WBB2363	SGH	61	M	Indian	Blood	IPD	18/11/2013	N/A	N/A	386	06C	110
WBB2364	SGH	78	F	Chinese	Blood	IPD	11/09/2014	N/A	N/A	180	3	51
WBB2365	SGH	73	M	Chinese	Blood	IPD	17/12/2014	N/A	N/A	199	15C	5
WBB2366	SGH	60	F	Indian	Blood	IPD	09/07/2015	N/A	N/A	4908	09V	101
WBB2368	SGH	69	M	Chinese	Blood	IPD	27/11/2013	N/A	N/A	386	06C	110
WBB2369	SGH	63	M	Chinese	Blood	IPD	22/09/2014	N/A	N/A	12209	23F	119
WBB2370	SGH	46	M	Malay	Blood	IPD	28/12/2014	N/A	N/A	74	2	46
WBB2371	SGH	77	M	Chinese	Blood	IPD	13/07/2015	N/A	N/A	6011	3	6
WBB2373	SGH	38	M	Malay	Blood	IPD	04/12/2013	N/A	N/A	3111	19A	58
WBB2374	SGH	80	M	Chinese	Blood	IPD	24/09/2014	N/A	N/A	81	23F	0
WBB2375	SGH	77	M	Chinese	Blood	IPD	12/01/2015	N/A	N/A	1591	23F	0
WBB2376	SGH	60	M	Chinese	Blood	IPD	15/07/2015	N/A	N/A	1553	19A	99
WBB2378	SGH	65	F	Chinese	Blood	IPD	03/02/2014	N/A	N/A	199	15B	5
WBB2379	SGH	35	M	Indonesian	Blood	IPD	07/10/2014	N/A	N/A	217	1	10
WBB2380	SGH	68	F	Chinese	Blood	IPD	02/02/2015	N/A	N/A	7479	15C	30

Appendix Q

WBB2381	SGH	55	F	Chinese	Blood	IPD	21/10/2015	N/A	N/A	386	06D	110
WBB2383	SGH	39	M	Chinese	Blood	IPD	25/02/2014	N/A	N/A	124	14	16
WBB2384	SGH	64	M	Chinese	Blood	IPD	12/10/2014	N/A	N/A	4352	09V	6
WBB2385	SGH	62	F	Chinese	Blood	IPD	17/02/2015	N/A	N/A	199	15C	5
WBB2386	SGH	69	M	Filipino	Blood	IPD	23/12/2015	N/A	N/A	782	14	7
WBB2388	SGH	62	M	Malay	Blood	IPD	03/03/2014	N/A	N/A	4127	4	20
WBB2389	SGH	74	M	Chinese	Blood	IPD	28/10/2014	N/A	N/A	6908	41A	120
WBB2390	SGH	26	M	Indian	Blood	IPD	18/02/2015	N/A	N/A	4745	20	74
WBB2391	SGH	54	M	Chinese	Blood	IPD	14/01/2016	N/A	N/A	3173	06A	62
WBB2393	SGH	87	M	Chinese	Blood	IPD	29/05/2014	N/A	N/A	320	19A	1
WBB2394	SGH	29	M	Indian	Blood	IPD	08/11/2014	N/A	N/A	1518	06E	56
WBB2395	SGH	78	F	Chinese	Blood	IPD	21/02/2015	N/A	N/A	1591	23F	0
WBB2396	SGH	62	F	Indian	Blood	IPD	31/01/2016	N/A	N/A	473	06B	26

Appendix R Isolate antimicrobial susceptibility data for 313 *S. pneumoniae* isolates collected 2013-2016

Laboratory data for the isolates collected from collaborating hospitals 2013-2016. Data for isolates collected 1997-2013 is located in Jauneikaite (2014) (182). The described phenotypes are sensitive (S), resistant (R) and intermediate resistance (I).

ID	Penicillin	Erythromycin	Cotrimoxazole	Clindamycin	Vancomycin	Doxycycline
WBB1259	R	R	N/A	R	S	R
WBB1260	S	S	N/A	S	S	R
WBB1261	S	R	N/A	R	S	R
WBB1262	S	S	N/A	S	S	S
WBB1263	S	S	N/A	S	S	R
WBB1264	S	S	N/A	S	S	S
WBB1265	S	R	N/A	S	S	S
WBB1266	S	R	N/A	R	S	R
WBB1267	S	R	N/A	R	S	R
WBB1268	S	R	N/A	R	S	S
WBB1269	S	R	N/A	S	S	N/A
WBB1270	S	S	N/A	S	S	N/A
WBB1271	S	R	N/A	R	S	R
WBB1273	S	R	N/A	R	S	R
WBB1274	S	S	N/A	S	S	S
WBB1275	S	S	N/A	S	S	S
WBB1276	S	I	N/A	S	S	R
WBB1277	N/A	N/A	N/A	N/A	N/A	N/A
WBB1278	S	R	N/A	R	S	R
WBB1279	S	S	N/A	S	S	S
WBB1280	R	R	N/A	R	S	R
WBB1281	S	R	N/A	S	S	R

WBB1282	S	S	N/A	S	S	R
WBB1283	S	S	N/A	S	S	R
WBB1284	N/A	N/A	N/A	N/A	N/A	N/A
WBB1285	S	S	N/A	S	S	S
WBB1286	S	S	N/A	S	S	R
WBB1288	N/A	N/A	N/A	N/A	N/A	N/A
WBB1289	N/A	N/A	N/A	N/A	N/A	N/A
WBB1290	S	R	N/A	R	S	R
WBB1291	S	S	N/A	S	S	I
WBB1292	S	S	N/A	S	S	S
WBB1293	S	S	N/A	S	S	S
WBB1294	N/A	N/A	N/A	N/A	N/A	N/A
WBB1295	S	S	N/A	S	S	S
WBB1296	S	S	N/A	S	S	R
WBB1297	S	R	N/A	R	S	R
WBB1298	S	R	N/A	R	S	R
WBB1299	S	S	N/A	S	S	N/A
WBB1300	S	S	N/A	S	S	S
WBB1301	S	R	N/A	R	S	R
WBB1302	S	S	N/A	S	S	S
WBB1303	S	S	N/A	S	S	R
WBB1304	s	R	N/A	S	S	R
WBB1305	S	R	N/A	S	S	S
WBB1306	R	S	N/A	S	S	R
WBB1307	S	S	N/A	S	S	S
WBB1308	S	S	N/A	S	S	S
WBB1309	S	S	N/A	S	S	S
WBB1310	N/A	N/A	N/A	N/A	N/A	N/A
WBB1311	S	S	N/A	S	S	S
WBB1313	S	R	N/A	S	S	R
WBB1314	S	S	N/A	S	S	S
WBB1315	S	R	N/A	R	S	S

WBB1316	N/A	N/A	N/A	N/A	N/A	N/A
WBB1317	S	R	N/A	S	S	R
WBB1318	S	S	N/A	S	S	R
WBB1319	S	S	N/A	S	S	S
WBB1320	S	R	N/A	S	S	R
WBB1322	S	S	N/A	S	S	R
WBB1324	S	R	N/A	R	S	R
WBB1325	S	R	N/A	S	S	R
WBB1326	S	R	N/A	S	S	I
WBB1327	S	S	N/A	S	S	S
WBB1328	S	I	R	S	N/A	N/A
WBB1329	I	R	R	R	S	N/A
WBB1330	S	S	S	S	N/A	N/A
WBB1331	S	S	R	S	N/A	N/A
WBB1332	R	R	R	S	S	N/A
WBB1333	S	S	S	S	N/A	N/A
WBB1334	I	R	R	S	N/A	N/A
WBB1335	I	R	I	S	N/A	N/A
WBB1336	I	R	S	S	N/A	N/A
WBB1337	S	S	S	S	N/A	N/A
WBB1338	S	S	R	S	N/A	N/A
WBB1339	S	S	R	S	N/A	N/A
WBB1340	S	S	R	S	N/A	N/A
WBB1341	S	S	S	S	S	N/A
WBB1342	S	R	R	R	N/A	N/A
WBB1343	S	S	R	S	N/A	N/A
WBB1344	S	R	S	R	N/A	N/A
WBB1345	S	S	I	S	N/A	N/A
WBB1346	I	R	R	R	S	N/A
WBB1347	S	S	S	S	N/A	N/A
WBB1348	S	R	R	R	N/A	N/A
WBB1349	S	R	R	R	N/A	N/A

WBB1350	S	S	S	S	N/A	N/A
WBB1351	S	S	S	S	S	N/A
WBB1352	I	R	R	R	N/A	N/A
WBB1353	I	R	R	R	N/A	N/A
WBB1354	S	S	R	S	N/A	N/A
WBB2033	S	R	N/A	R	S	S
WBB2034	S	S	S	S	N/A	N/A
WBB2035	S	S	S	S	N/A	N/A
WBB2036	I	R	R	S	N/A	N/A
WBB2037	S	R	N/A	R	S	R
WBB2038	S	R	N/A	R	S	R
WBB2039	S	S	N/A	S	S	S
WBB2040	S	R	N/A	R	S	S
WBB2041	I	R	R	R	N/A	N/A
WBB2042	S	R	S	R	N/A	N/A
WBB2043	S	S	S	S	N/A	N/A
WBB2044	S	S	S	S	N/A	N/A
WBB2045	S	R	N/A	R	S	R
WBB2046	S	R	N/A	R	S	R
WBB2047	S	S	N/A	S	S	R
WBB2048	S	R	N/A	R	S	R
WBB2050	S	R	S	R	N/A	N/A
WBB2051	S	S	S	S	N/A	N/A
WBB2052	I	R	R	R	N/A	N/A
WBB2053	R	R	N/A	R	S	R
WBB2054	S	S	N/A	S	S	S
WBB2055	S	R	N/A	R	S	R
WBB2056	S	S	N/A	S	S	S
WBB2059	S	S	S	S	N/A	N/A
WBB2060	S	S	S	S	N/A	N/A
WBB2061	S	R	N/A	R	S	R
WBB2062	S	R	N/A	R	S	R

WBB2063	S	R	N/A	S	S	S
WBB2064	S	R	N/A	S	S	S
WBB2065	S	S	R	S	N/A	N/A
WBB2066	I	R	R	R	N/A	N/A
WBB2067	I	R	R	S	N/A	N/A
WBB2068	S	R	N/A	S	S	S
WBB2069	S	S	N/A	S	S	S
WBB2070	S	R	N/A	S	S	S
WBB2071	S	R	N/A	S	S	S
WBB2072	R	R	N/A	S	S	R
WBB2073	S	R	R	S	N/A	N/A
WBB2074	S	S	S	S	N/A	N/A
WBB2075	S	R	I	R	N/A	N/A
WBB2076	R	R	N/A	R	S	R
WBB2077	S	R	N/A	R	S	S
WBB2078	S	R	N/A	R	S	R
WBB2079	S	R	N/A	R	S	R
WBB2080	I	R	R	N/A	N/A	N/A
WBB2081	I	R	R	R	N/A	N/A
WBB2083	S	S	S	S	N/A	N/A
WBB2084	S	S	N/A	S	S	R
WBB2085	S	R	N/A	R	S	R
WBB2087	S	S	N/A	S	S	R
WBB2088	S	S	S	N/A	N/A	N/A
WBB2089	S	S	S	S	N/A	N/A
WBB2090	N/A	R	S	S	N/A	N/A
WBB2091	I	R	R	R	N/A	N/A
WBB2092	S	S	N/A	S	S	R
WBB2093	S	R	N/A	R	S	S
WBB2094	I	R	N/A	R	S	R
WBB2095	S	R	N/A	R	S	R
WBB2096	I	R	R	N/A	N/A	N/A

WBB2097	S	R	S	R	N/A	N/A
WBB2100	S	S	N/A	S	S	R
WBB2101	S	S	N/A	S	S	S
WBB2102	N/A	N/A	N/A	N/A	N/A	N/A
WBB2104	S	S	R	N/A	N/A	N/A
WBB2105	S	S	S	S	N/A	N/A
WBB2106	S	S	R	S	N/A	N/A
WBB2107	S	S	S	S	N/A	N/A
WBB2108	S	R	N/A	R	S	R
WBB2109	I	R	N/A	R	S	R
WBB2110	S	S	N/A	S	S	S
WBB2111	S	S	N/A	S	S	R
WBB2112	I	R	S	N/A	N/A	N/A
WBB2113	S	S	S	S	N/A	N/A
WBB2114	I	R	R	R	N/A	N/A
WBB2115	S	R	R	S	N/A	N/A
WBB2116	S	R	N/A	R	S	S
WBB2117	S	R	N/A	R	S	R
WBB2118	S	S	N/A	S	S	R
WBB2119	S	S	N/A	S	S	R
WBB2120	I	R	S	N/A	N/A	N/A
WBB2121	I	R	S	R	N/A	N/A
WBB2122	S	R	S	R	N/A	N/A
WBB2123	R	R	R	R	N/A	N/A
WBB2124	S	R	N/A	R	S	R
WBB2125	S	R	N/A	R	S	R
WBB2126	S	R	N/A	R	S	R
WBB2127	S	R	N/A	R	S	R
WBB2128	I	R	R	N/A	N/A	N/A
WBB2144	S	S	N/A	S	S	S
WBB2145	I	R	R	N/A	N/A	N/A
WBB2146	I	R	R	N/A	N/A	N/A

WBB2147	I	R	S	N/A	N/A	N/A
WBB2148	S	R	S	N/A	N/A	N/A
WBB2149	S	S	S	N/A	N/A	N/A
WBB2150	S	R	S	N/A	N/A	N/A
WBB2151	S	R	N/A	R	S	R
WBB2153	S	R	S	N/A	N/A	N/A
WBB2154	I	R	R	N/A	N/A	N/A
WBB2155	S	S	S	N/A	N/A	N/A
WBB2156	R	R	R	N/A	N/A	N/A
WBB2157	R	R	R	N/A	N/A	N/A
WBB2158	R	R	R	N/A	N/A	N/A
WBB2159	S	S	N/A	S	S	S
WBB2161	I	R	R	N/A	N/A	N/A
WBB2162	S	S	S	N/A	N/A	N/A
WBB2163	R	R	S	N/A	N/A	N/A
WBB2164	S	S	S	N/A	N/A	N/A
WBB2165	R	R	R	N/A	N/A	N/A
WBB2166	R	R	R	N/A	N/A	N/A
WBB2167	S	R	N/A	S	S	S
WBB2168	S	S	N/A	S	S	S
WBB2169	R	R	R	N/A	N/A	N/A
WBB2170	I	R	R	N/A	N/A	N/A
WBB2171	S	S	R	N/A	N/A	N/A
WBB2172	S	R	S	N/A	N/A	N/A
WBB2173	I	R	S	N/A	N/A	N/A
WBB2174	S	S	S	N/A	N/A	N/A
WBB2175	S	S	N/A	S	S	S
WBB2176	S	S	S	N/A	N/A	N/A
WBB2177	S	R	S	N/A	N/A	N/A
WBB2178	S	S	S	N/A	N/A	N/A
WBB2179	I	R	S	N/A	N/A	N/A
WBB2180	S	R	S	N/A	N/A	N/A

WBB2181	S	R	S	N/A	N/A	N/A
WBB2182	N/A	N/A	N/A	N/A	N/A	N/A
WBB2183	S	S	R	N/A	N/A	N/A
WBB2184	R	R	R	N/A	N/A	N/A
WBB2185	R	R	R	N/A	N/A	N/A
WBB2186	R	R	R	N/A	N/A	N/A
WBB2187	S	S	R	N/A	N/A	N/A
WBB2188	I	R	R	N/A	N/A	N/A
WBB2189	S	R	N/A	R	S	R
WBB2190	S	S	S	N/A	N/A	N/A
WBB2191	S	R	S	N/A	N/A	N/A
WBB2192	S	R	R	N/A	N/A	N/A
WBB2193	S	S	S	N/A	N/A	N/A
WBB2194	S	S	R	N/A	N/A	N/A
WBB2195	S	R	S	N/A	N/A	N/A
WBB2196	S	S	S	N/A	N/A	N/A
WBB2197	S	S	S	N/A	N/A	N/A
WBB2198	S	S	S	N/A	N/A	N/A
WBB2199	S	S	S	N/A	N/A	N/A
WBB2200	I	R	R	N/A	N/A	N/A
WBB2201	I	R	S	N/A	N/A	N/A
WBB2202	S	S	S	N/A	N/A	N/A
WBB2203	S	R	S	N/A	N/A	N/A
WBB2204	S	R	S	N/A	N/A	N/A
WBB2205	S	S	S	N/A	N/A	N/A
WBB2206	I	R	R	N/A	N/A	N/A
WBB2207	S	R	S	N/A	N/A	N/A
WBB2208	S	S	R	N/A	N/A	N/A
WBB2209	S	S	S	N/A	N/A	N/A
WBB2210	I	R	R	N/A	N/A	N/A
WBB2211	I	R	R	N/A	N/A	N/A
WBB2212	I	R	R	N/A	N/A	N/A

WBB2213	I	R	R	N/A	N/A	N/A
WBB2214	I	R	R	N/A	N/A	N/A
WBB2215	S	R	R	N/A	N/A	N/A
WBB2216	S	R	S	N/A	N/A	N/A
WBB2217	S	S	S	N/A	N/A	N/A
WBB2218	I	R	R	N/A	N/A	N/A
WBB2219	I	R	S	N/A	N/A	N/A
WBB2220	I	S	S	N/A	N/A	N/A
WBB2221	I	R	R	N/A	N/A	N/A
WBB2222	I	R	R	N/A	N/A	N/A
WBB2223	S	R	S	N/A	N/A	N/A
WBB2224	R	R	R	N/A	N/A	N/A
WBB2225	S	S	S	N/A	N/A	N/A
WBB2226	S	R	S	N/A	N/A	N/A
WBB2227	I	R	R	N/A	N/A	N/A
WBB2228	S	S	S	N/A	N/A	N/A
WBB2230	S	S	N/A	S	S	S
WBB2336	S	S	N/A	S	S	N/A
WBB2337	S	R	N/A	R	S	N/A
WBB2338	S	R	N/A	R	S	N/A
WBB2339	R	R	N/A	R	S	N/A
WBB2340	S	S	N/A	S	S	N/A
WBB2342	S	R	N/A	R	S	N/A
WBB2343	S	S	N/A	S	S	N/A
WBB2344	S	R	N/A	R	S	N/A
WBB2345	R	R	N/A	R	S	N/A
WBB2346	S	S	N/A	S	S	N/A
WBB2348	S	S	N/A	S	S	N/A
WBB2349	R	R	N/A	R	S	N/A
WBB2350	S	S	N/A	S	S	N/A
WBB2351	S	S	N/A	S	S	N/A
WBB2353	R	R	N/A	R	S	N/A

WBB2354	S	S	N/A	S	S	N/A
WBB2355	S	S	N/A	S	S	N/A
WBB2356	S	S	N/A	S	S	N/A
WBB2358	R	R	N/A	R	S	N/A
WBB2359	S	S	N/A	S	S	N/A
WBB2360	S	S	N/A	S	S	N/A
WBB2361	S	R	N/A	R	S	N/A
WBB2363	S	R	N/A	R	S	N/A
WBB2364	S	R	N/A	R	S	N/A
WBB2365	S	S	N/A	S	S	N/A
WBB2366	S	R	N/A	S	S	N/A
WBB2368	S	R	N/A	R	S	N/A
WBB2369	S	S	N/A	S	S	N/A
WBB2370	S	R	N/A	S	S	N/A
WBB2371	S	R	N/A	R	S	N/A
WBB2373	S	R	N/A	R	S	N/A
WBB2374	R	R	N/A	S	S	N/A
WBB2375	I	R	N/A	R	S	N/A
WBB2376	I	S	N/A	S	S	N/A
WBB2378	S	S	N/A	S	S	N/A
WBB2379	S	S	N/A	S	S	N/A
WBB2380	I	R	N/A	S	S	N/A
WBB2381	S	R	N/A	R	S	N/A
WBB2383	S	S	N/A	S	S	N/A
WBB2384	S	S	N/A	S	S	N/A
WBB2385	S	R	N/A	R	S	N/A
WBB2386	S	S	N/A	S	S	N/A
WBB2388	S	S	N/A	S	S	N/A
WBB2389	S	R	N/A	R	S	N/A
WBB2390	S	R	N/A	R	S	N/A
WBB2391	I	R	N/A	R	S	N/A
WBB2393	S	R	N/A	R	S	N/A

WBB2394	S	R	N/A	S	S	N/A
WBB2395	R	R	N/A	R	S	N/A
WBB2396	R	R	N/A	S	S	N/A

List of References

1. McCarthy BJ, Bolton ET. An approach to the measurement of genetic relatedness among organisms. Proceedings of the National Academy of Sciences of the United States of America. 1963;50(1):156-64.

2. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nature communications. 2019;10(1):5029.

3. Meera Krishna B, Khan MA, Khan ST. Next-Generation Sequencing (NGS) Platforms: An Exciting Era of Genome Sequence Analysis. In: Tripathi V, Kumar P, Tripathi P, Kishore A, Kamle M, editors. Microbial Genomics in Sustainable Agroecosystems: Volume 2. Singapore: Springer Singapore; 2019. p. 89-109.

4. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13.

5. Fournier PE, Raoult D. Prospects for the future using genomics and proteomics in clinical microbiology. Annu Rev Microbiol. 2011;65:169-88.

6. Euzéby JP. List of Bacterial Names with Standing in Nomenclature. International Journal of Systematic Bacteriology. 1997;47:590-2.

7. Bogaert D, Keijser B, Huse S, Rossen J, Veenhoven R, van Gils E, et al. Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. PloS one. 2011;6(2):e17035.

8. Bogaert D, De Groot R, Hermans PW. Streptococcus pneumoniae colonisation: the key to pneumococcal disease. Lancet Infect Dis. 2004;4(3):144-54.

9. Regev-Yochay G, Raz M, Dagan R, Porat N, Shainberg B, Pinco E, et al. Nasopharyngeal carriage of Streptococcus pneumoniae by adults and children in community and family settings. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2004;38(5):632-9.

10. Simell B, Auranen K, Kayhty H, Goldblatt D, Dagan R, O'Brien KL. The fundamental link between pneumococcal carriage and disease. Expert review of vaccines. 2012;11(7):841-55.

11. Pichichero M, Khan M, Xu Q. Next generation protein based Streptococcus pneumoniae vaccines. Human vaccines & immunotherapeutics. 2015:0.

12. Nuorti JP, Butler JC, Farley MM, Harrison LH, McGeer A, Kolczak MS, et al. Cigarette smoking and invasive pneumococcal disease. Active Bacterial Core Surveillance Team. The New England journal of medicine. 2000;342(10):681-9.

13. Centers for Disease Control and Prevention C. The Pink Book: Course Textbook. Epidemiology and Prevention of Vaccine-Preventable Diseases. 2015.

14. Kadioglu A, Weiser JN, Paton JC, Andrew PW. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. Nature Reviews Microbiology. 2008;6(4):288-301.

15. Centers for Disease Control and Prevention C. Prevention of pneumococcal disease: recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Morb Mortal Wkly Rep. 1997;46:1-24.

16. van der Poll T, Opal SM. Pathogenesis, treatment, and prevention of pneumococcal pneumonia. Lancet. 2009;374(9700):1543-56.

17. Paton JC, Andrew PW, Boulnois GJ, Mitchell TJ. Molecular analysis of the pathogenicity of Streptococcus pneumoniae: the role of pneumococcal proteins. Annu Rev Microbiol. 1993;47:89-115.

18. Eng P, Lim LH, Loo CM, Low JA, Tan C, Tan EK, et al. Role of pneumococcal vaccination in prevention of pneumococcal disease among adults in Singapore. International Journal of General Medicine. 2014;7:179-91.

19. Brueggemann AB, Harrold CL, Rezaei Javan R, van Tonder AJ, McDonnell AJ, Edwards BA. Pneumococcal prophages are diverse, but not without structure or history. Scientific reports. 2017;7:42976.

20. van de Beek D, Brouwer MC, Thwaites GE, Tunkel AR. Advances in treatment of bacterial meningitis. The Lancet. 2012;380(9854):1693-702.

21. van Tonder AJ, Bray JE, Quirk SJ, Haraldsson G, Jolley KA, Maiden MC, et al. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. Microbial genomics. 2016;2(10):000090.

22. Hausdorff WP, Feikin DR, Klugman KP. Epidemiological differences among pneumococcal serotypes. Lancet Infect Dis. 2005;5(2):83-93.

23. Hausdorff WP, Siber G, Paradiso PR. Geographical differences in invasive pneumococcal disease rates and serotype frequency in young children. Lancet. 2001;357(9260):950-2.

24. Henriques B, Kalin M, Ortqvist A, Olsson Liljequist B, Almela M, Marrie TJ, et al. Molecular epidemiology of Streptococcus pneumoniae causing invasive disease in 5 countries. The Journal of infectious diseases. 2000;182(3):833-9.

25. Normark BH, Ortqvist A, Kalin M, Olsson-Liljequist B, Hedlund J, Svenson SB, et al. Changes in serotype distribution may hamper efficacy of pneumococcal conjugate vaccines in children. Scandinavian journal of infectious diseases. 2001;33(11):848-50.

26. Foster D, Knox K, Walker AS, Griffiths DT, Moore H, Haworth E, et al. Invasive pneumococcal disease: epidemiology in children and adults prior to implementation of the conjugate vaccine in the Oxfordshire region, England. Journal of medical microbiology. 2008;57(Pt 4):480-7.

27. Johnson HL, Deloria-Knoll M, Levine OS, Stoszek SK, Freimanis Hance L, Reithinger R, et al. Systematic Evaluation of Serotypes Causing Invasive Pneumococcal Disease among Children Under Five: The Pneumococcal Global Serotype Project. PLoS Med. 2010;7(10):e1000348.

28. Jauneikaite E, Jefferies JM, Hibberd ML, Clarke SC. Prevalence of Streptococcus pneumoniae serotypes causing invasive and non-invasive disease in South East Asia: a review. Vaccine. 2012;30(24):3503-14.

29. O'Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, et al. Burden of disease caused by Streptococcus pneumoniae in children younger than 5 years: global estimates. The Lancet. 2009;374.

30. Brueggemann AB, Peto TE, Crook DW, Butler JC, Kristinsson KG, Spratt BG. Temporal and geographic stability of the serogroup-specific invasive disease potential of Streptococcus pneumoniae in children. The Journal of infectious diseases. 2004;190(7):1203-11.

31. Hanage WP, Kaijalainen TH, Syrjanen RK, Auranen K, Leinonen M, Makela PH, et al. Invasiveness of serotypes and clones of Streptococcus pneumoniae among children in Finland. Infect Immun. 2005;73(1):431-5.

32. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. The Journal of infectious diseases. 2003;187(9):1424-32.

33. Haas W, Hesje CK, Sanfilippo CM, Morris TW. High proportion of nontypeable Streptococcus pneumoniae isolates among sporadic, nonoutbreak cases of bacterial conjunctivitis. Current eye research. 2011;36(12):1078-85.

34. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M, et al. Genetic Analysis of the Capsular Biosynthetic Locus from All 90 Pneumococcal Serotypes. PLOS Genetics. 2006;2(3):e31.

35. Garcia E, Llull D, Lopez R. Functional organization of the gene cluster involved in the synthesis of the pneumococcal capsule. International microbiology : the official journal of the Spanish Society for Microbiology. 1999;2(3):169-76.

36. Daniels CC, Rogers PD, Shelton CM. A Review of Pneumococcal Vaccines: Current Polysaccharide Vaccine Recommendations and Future Protein Antigens. The journal of pediatric pharmacology and therapeutics : JPPT : the official journal of PPAG. 2016;21(1):27-35.

37. Leggat DJ, Khaskhely NM, Iyer AS, Mosakowski J, Thompson RS, Weinandy JD, et al. Pneumococcal polysaccharide vaccination induces polysaccharide-specific B cells in adult peripheral blood expressing CD19(+) CD20(+) CD3(-) CD70(-) CD27(+) IgM(+) CD43(+) CD5(+)/(-). Vaccine. 2013;31(41):4632-40.

38. Bogaert D, Hermans PWM, Adrian PV, Rümke HC, de Groot R. Pneumococcal vaccines: an update on current strategies. Vaccine. 2004;22(17):2209-20.

39. Organisation WH. Pneumococcal conjugate vaccine for childhood immunization - WHO position paper. Weekly epidemological record. 2007;12(82):93-104.

40. Song JH. Advances in pneumococcal antibiotic resistance. Expert review of respiratory medicine. 2013;7(5):491-8.

41. George JF, Schroeder HW. Developmental regulation of D beta reading frame and junctional diversity in T cell receptor-beta transcripts from human thymus. The Journal of Immunology. 1992;148(4):1230.

42. Örtqvist Å. Pneumococcal vaccination: current and future issues. European Respiratory Journal. 2001;18(1):184.

43. Fedson DS, Guppy MJ. Pneumococcal vaccination of older adults: conjugate or polysaccharide? Human vaccines & immunotherapeutics. 2013;9(6):1382-4.

44. Kim TH, Johnstone J, Loeb M. Vaccine herd effect. Scandinavian journal of infectious diseases. 2011;43(9):683-9.

45. Westerink MAJ, Schroeder HW, Nahm MH. Immune Responses to pneumococcal vaccines in children and adults: Rationale for age-specific vaccination. Aging and Disease. 2012;3(1):51-67.

46. Stein KE. Thymus-independent and thymus-dependent responses to polysaccharide antigens. The Journal of infectious diseases. 1992;165 Suppl 1:S49-52.

47. Lecrenier N, Marijam A, Olbrecht J, Soumahoro L, Nieto Guevara J, Mungall B. Ten years of experience with the pneumococcal non-typeable Haemophilus influenzae protein D-conjugate vaccine (Synflorix) in children. Expert review of vaccines. 2020;19(3):247-65.

48. Centers for Disease Control and Prevention C. Licensure of a 13-Valent Pneumococcal Conjugate Vaccine (PCV13) and Recommendations for Use Among Children --- Advisory Committee on Immunization Practices (ACIP), 2010. Morbidity and Mortality Weekly Report. 2010;59(09).

49. 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(3):e408-13.

50. Lexau CA, Lynfield R, Danila R, Pilishvili T, Facklam R, Farley MM, et al. Changing epidemiology of invasive pneumococcal disease among older adults in the era of pediatric pneumococcal conjugate vaccine. Jama. 2005;294(16):2043-51.

51. Isaacman DJ, Fletcher MA, Fritzell B, Ciuryla V, Schranz J. Indirect effects associated with widespread vaccination of infants with heptavalent pneumococcal conjugate vaccine (PCV7; Prevnar). Vaccine. 2007;25(13):2420-7.

52. Dagan R, Melamed R, Muallem M, Piglansky L, Greenberg D, Abramson O, et al. Reduction of nasopharyngeal carriage of pneumococci during the second year of life by a heptavalent conjugate pneumococcal vaccine. The Journal of infectious diseases. 1996;174(6):1271-8.

53. Pilishvili T, Lexau C, Farley MM, Hadler J, Harrison LH, Bennett NM, et al. Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine. The Journal of infectious diseases. 2010;201(1):32-41.

54. Lehmann D, Willis J, Moore HC, Giele C, Murphy D, Keil AD, et al. The changing epidemiology of invasive pneumococcal disease in aboriginal and non-aboriginal western Australians from 1997 through 2007 and emergence of nonvaccine serotypes. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2010;50(11):1477-86.

55. Nzenze SA, Shiri T, Nunes MC, Klugman KP, Kahn K, Twine R, et al. Temporal changes in pneumococcal colonization in a rural African community with high HIV prevalence following routine infant pneumococcal immunization. The Pediatric infectious disease journal. 2013;32(11):1270-8.

56. Pichon B, Ladhani SN, Slack MPE, Segonds-Pichon A, Andrews NJ, Waight PA, et al. Changes in Molecular Epidemiology of Streptococcus pneumoniae Causing Meningitis following Introduction of Pneumococcal Conjugate Vaccination in England and Wales. Journal of clinical microbiology. 2013;51(3):820-7.

57. Kim SH, Song JH, Chung DR, Thamlikitkul V, Yang Y, Wang H, 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. Antimicrobial agents and chemotherapy. 2012;56(3):1418-26.

58. Marttinen P, Baldwin A, Hanage WP, Dowson C, Mahenthiralingam E, Corander J. Bayesian modeling of recombination events in bacterial populations. BMC Bioinformatics. 2008;9.

59. Hausdorff WP, Bryant J, Paradiso PR, Siber GR. Which pneumococcal serogroups cause the most invasive disease: implications for conjugate vaccine formulation and use, part I. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2000;30(1):100-21.

60. Hausdorff WP, Bryant J, Kloek C, Paradiso PR, Siber GR. The contribution of specific pneumococcal serogroups to different disease manifestations: implications for conjugate vaccine formulation and use, part II. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2000;30(1):122-40.

61. Hanage WP, Bishop CJ, Lee GM, Lipsitch M, Stevenson A, Rifas-Shiman SL, et al. Clonal replacement among 19A Streptococcus pneumoniae in Massachusetts, prior to 13 valent conjugate vaccination. Vaccine. 2011;29(48):8877-81.

62. Tocheva AS, Jefferies JM, Rubery H, Bennett J, Afimeke G, Garland J, et al. Declining serotype coverage of new pneumococcal conjugate vaccines relating to the carriage of Streptococcus pneumoniae in young children. Vaccine. 2011;29(26):4400-4.

63. von Gottberg A, de Gouveia L, Tempia S, Quan V, Meiring S, von Mollendorf C, et al. Effects of Vaccination on Invasive Pneumococcal Disease in South Africa. New England Journal of Medicine. 2014;371(20):1889-99.

64. Hicks LA, Harrison LH, Flannery B, Hadler JL, Schaffner W, Craig AS, et al. Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998-2004. The Journal of infectious diseases. 2007;196(9):1346-54.

65. Kyaw MH, Lynfield R, Schaffner W, Craig AS, Hadler J, Reingold A, et al. Effect of Introduction of the Pneumococcal Conjugate Vaccine on Drug-Resistant Streptococcus pneumoniae. New England Journal of Medicine. 2006;354(14):1455-63.

66. Richter S, S., Heilmann K, P., Dohrn C, L., Riahi F, Diekema D, J. , Doern G, V. . Pneumococcal Serotypes before and after Introduction of Conjugate Vaccines, United States, 1999–2011. Emerging Infectious Disease journal. 2013;19(7):1074.

67. Bowers JR, Driebe EM, Nibecker JL, Wojack BR, Sarovich DS, Wong AH, et al. Dominance of multidrug resistant CC271 clones in macrolide-resistant streptococcus pneumoniae in Arizona. BMC Microbiol. 2012;12:12.

68. Appelbaum PC. World-Wide development of antibiotic resistance in penumococci. Eur J Clin Microbiol. 1987;6(4):367-77.

69. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, et al. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014;46(3):305-9.

70. Croucher NJ, Chewapreecha C, Hanage WP, Harris SR, McGee L, van der Linden M, et al. Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. Genome Biol Evol. 2014;6(7):1589-602.

71. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet. 2013;45(6):656-63.

72. Jauneikaite E, Mary Carnon Jefferies J, William Vere Churton N, Tzer Pin Lin R, Lloyd Hibberd M, Charles Clarke S. Genetic diversity of Streptococcus pneumoniae causing meningitis and sepsis in Singapore during the first year of PCV7 implementation. Emerging Microbes & Infections. 2014;3(6):e39.

73. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, et al. Rapid Pneumococcal Evolution in Response to Clinical Interventions. Science (New York, NY). 2011;331(6016):430-4.

74. Kelly T, Dillard JP, Yother J. Effect of genetic switching of capsular type on virulence of Streptococcus pneumoniae. Infect Immun. 1994;62(5):1813-9.

75. Coffey TJ, Enright MC, Daniels M, Morona JK, Morona R, Hryniewicz W, et al. Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of Streptococcus pneumoniae. Mol Microbiol. 1998;27(1):73-83.

76. Alderson MR. Status of research and development of pediatric vaccines for Streptococcus pneumoniae. Vaccine. 2016;34(26):2959-61.

77. Andam CP, Hanage WP. Mechanisms of genome evolution of Streptococcus. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2015;33:334-42.

78. Golubchik T, Brueggemann AB, Street T, Gertz RE, Jr., Spencer CCA, Ho T, et al. Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. Nature genetics. 2012;44(3):352-5.

79. Moore MR, Gertz REJ, Woodbury RL, Barkocy-Gallagher GA, Schaffner W, Lexau C, et al. Population snapshot of emergent Streptococcus pneumoniae serotype 19A in the United States, 2005. The Journal of infectious diseases. 2008;197(1):1016 -27.

80. Chochua S, Metcalf BJ, Li Z, Walker H, Tran T, McGee L, et al. Invasive Serotype 35B Pneumococci Including an Expanding Serotype Switch Lineage, United States, 2015-2016. Emerging infectious diseases. 2017;23(6):922-30.

81. Cornick JE, Bentley SD. Streptococcus pneumoniae: the evolution of antimicrobial resistance to beta-lactams, fluoroquinolones and macrolides. Microbes and infection. 2012;14(7-8):573-83.

82. Hansman D, Bullen MM. A resistant pneumococcus. Lancet. 1967;290(7509):264-5.

83. Appelbaum PC, Bhamjee A, Scragg JN, Hallett AF, Bowen AJ, Cooper RC. Streptococcus pneumoniae resistant to penicillin and chloramphenicol. Lancet. 1977;2(8046):995-7.

84. Fenoll A, Jado I, Vicioso D, Perez A, Casal J. Evolution of Streptococcu pneumoniae Serotypes and Antibiotic Resistance in Spain: Update (1990 to 1996). Journal of clinical microbiology. 1998;36(12):3447-54.

85. Fenoll A, Bourgon MC, Munoz R, Vicioso D, Casal J. Serotype distribution and antimicrobial resistance of Streptococcus pneumoniae isolates causing systemic infections in Spain, 1979-1989. Reviews of infectious diseases. 1991;13(1):56-60.

86. Marton A, Gulyas M, Munoz R, Tomasz A. Extremely high incidence of antibiotic resistance in clinical isolates of Streptococcus pneumoniae in Hungary. The Journal of infectious diseases. 1991;163(3):542-8.

87. Friedland IR, Klugman KP. Antibiotic-resistant penumococcal disease in South African children. Am J Dis Child. 1992;146(8):920-3.

88. Kristinsson KG. Epidemiology of Penicillin Resistant Pneumococci in Iceland. Microb Drug Resist. 1995;1(2):121-5.

89. Guillemot D, Carbon C. Antibiotic use and pneumococcal resistance to penicillin: the French experience. Clin Microbiol Infect. 1999;5(4):38-42.

90. Breiman RF, Butler JC, Tenover FC, Elliott JA, Facklam RR. Emergence of drug-resistant pneumococcal infections in the United States. Jama. 1994;271(23):1831-5.

91. Doern GV, Brueggemann AB, Holley HPJ, Rauch AM. Antimicrobial resistance of Streptococcus pneumoniae recovered from outpatients in the United States during the winter months of 1994 to 1995: results of a 30 center national surveillance study. Antimicrobial agents and chemotherapy. 1996;40(5):1208-13.

92. Song JH, Lee NY, Ichiyama S, Yoshida R, Hirakata Y, Fu W, et al. Spread of drug-resistant Streptococcus pneumoniae in Asian countries: Asian Network for Surveillance of Resistant Pathogens (ANSORP) Study. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 1999;28(6):1206-11.

93. McGee L, McDougal L, Zhou J, Spratt BG, Tenover FC, George R, et al. Nomenclature of major antimicrobial-resistant clones of Streptococcus pneumoniae defined by the pneumococcal molecular epidemiology network. Journal of clinical microbiology. 2001;39(7):2565-71.

94. Pikis A, Donkersloot JA, Rodriguez WJ, Keith JM. A Conservative Amino Acid Mutation in the Chromosome-Encoded Dihydrofolate Reductase Confers Trimethoprim Resistance in Streptococcus pneumoniae. The Journal of infectious diseases. 1998;178(3):700-6.

95. Appelbaum PC. Antimicrobial resistance in Streptococcus pneumoniae: an overview. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 1992;15(1):77-83.

96. Xiao Y, Wei Z, Shen P, Ji J, Sun Z, Yu H, et al. Bacterial-resistance among outpatients of country hospitals in China: significant geographic distinctions and minor differences between central cities. Microbes and infection. 2015;17(6):417-25.

97. Farrell DJ, Couturier C, Hryniewicz W. Distribution and antibacterial susceptibility of macrolide resistance genotypes in Streptococcus pneumoniae: PROTEKT Year 5 (2003-2004). International journal of antimicrobial agents. 2008;31(3):245-9.

98. McCormick AW, Whitney CG, Farley MM, Lynfield R, Harrison LH, Bennett NM, et al. Geographic diversity and temporal trends of antimicrobial resistance in Streptococcus pneumoniae in the United States. Nat Med. 2003;9(4):424-30.

99. Pneumococcal Molecular Epidemiology Network [Internet]. 1997 [cited 9th December 2020]. Available from: https://www.pneumogen.net/pmen/index.html.

100. Klugman KP. Pneumococcal resistance to antibiotics. Clin Microbiol Rev. 1990;3(2):171-96.

101. Lynch JP, 3rd, Zhanel GG. Streptococcus pneumoniae: does antimicrobial resistance matter? Seminars in respiratory and critical care medicine. 2009;30(2):210-38.

102. Clarke SC, Lawrie D, Diggle MA. Genetic relatedness of antibiotic-resistant pneumococci isolated during case clusters. Journal of medical microbiology. 2004;53(Pt 11):1097-9.

103. Gertz REJ, Li Z, Pimenta FC, Jackson D, Juni BA, Lynfield R, et al. Increased penicillin nonsusceptibility of nonvaccine-serotype invasive pneumococci other than serotypes 19A and 6A in post 7-valent conjugate vaccine era. The Journal of infectious diseases. 2010;201(5):770-5.

104. Andersson DI, Hughes D. Antibiotic resistance and its cost: is it possible to reverse resistance? Nature Reviews Microbiology. 2010;8(4):260-71.

105. Schulz zur Wiesch P, Engelstädter J, Bonhoeffer S. Compensation of fitness costs and reversibility of antibiotic resistance mutations. Antimicrobial agents and chemotherapy. 2010;54(5):2085-95.

106. Masel J. Genetic drift. Current Biology. 2011;21(20):R837-R8.

107. Schroeder MR, Stephens DS. Macrolide Resistance in Streptococcus pneumoniae. Frontiers in cellular and infection microbiology. 2016;6:98.

108. Tait-Kamradt A, Davies T, Cronan M, Jacobs MR, Appelbaum PC, Sutcliffe J. Mutations in 23S rRNA and ribosomal protein L4 account for resistance in pneumococcal strains selected in vitro by macrolide passage. Antimicrobial agents and chemotherapy. 2000;44(8):2118-25.

109. Canu A, Malbruny B, Coquemont M, Davies TA, Appelbaum PC, Leclercq R. Diversity of ribosomal mutations conferring resistance to macrolides, clindamycin, streptogramin, and telithromycin in Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 2002;46(1):125-31.

110. Franceschi F, Kanyo Z, Sherer EC, Sutcliffe J. Macrolide resistance from the ribosome perspective. Curr Drug Targets Infect Disord. 2004;4(3):177-91.

111. Job V, Di Guilmi AM, Martin L, Vernet T, Dideberg O, Dessen A. Structural studies of the transpeptidase domain of PBP1a from Streptococcus pneumoniae. Acta Crystallogr D Biol Crystallogr. 2003;59(Pt 6):1067-9.

112. Chambers HF. Penicillin-binding protein-mediated resistance in pneumococci and staphylococci. The Journal of infectious diseases. 1999;179 Suppl 2:S353-9.

113. Dowson CG, Barcus V, King S, Pickerill P, Whatmore A, Yeo M. Horizontal gene transfer and the evolution of resistance and virulence determinants in Streptococcus. J Appl Microbiol. 1997;83(S1):42s-51s.

114. Laible G, Hakenbeck R. Five independent combinations of mutations can result in low-affinity penicillin-binding protein 2x of Streptococcus pneumoniae. J Bacteriol. 1991;173(21):6986-90.

115. Grebe T, Hakenbeck R. Penicillin-binding proteins 2b and 2x of Streptococcus pneumoniae are primary resistance determinants for different classes of beta-lactam antibiotics. Antimicrobial agents and chemotherapy. 1996;40(4):829-34.

116. Hakenbeck R, Grebe T, Zahner D, Stock JB. beta-lactam resistance in Streptococcus pneumoniae: penicillin-binding proteins and non-penicillin-binding proteins. Mol Microbiol. 1999;33(4):673-8.

117. Smith AM, Klugman KP. Alterations in PBP 1A Essential for High-Level Penicillin Resistance in Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1998;42(6):1329-33.

118. Chesnel L, Pernot L, Lemaire D, Champelovier D, Croizé J, Dideberg O, et al. The structural modifications induced by the M339F substitution in PBP2x from Streptococcus pneumoniae further decreases the susceptibility to beta-lactams of resistant strains. The Journal of biological chemistry. 2003;278(45):44448-56.

119. Pernot L, Chesnel L, Le Gouellec A, Croizé J, Vernet T, Dideberg O, et al. A PBP2x from a clinical isolate of Streptococcus pneumoniae exhibits an alternative mechanism for reduction of susceptibility to beta-lactam antibiotics. The Journal of biological chemistry. 2004;279(16):16463-70.

120. Dessen A, Mouz N, Gordon E, Hopkins J, Dideberg O. Crystal structure of PBP2x from a highly penicillin-resistant Streptococcus pneumoniae clinical isolate: a mosaic framework containing 83 mutations. The Journal of biological chemistry. 2001;276(48):45106-12.

121. Mouz N, Gordon E, Di Guilmi AM, Petit I, Pétillot Y, Dupont Y, et al. Identification of a structural determinant for resistance to beta-lactam antibiotics in Gram-positive bacteria. Proceedings of the National Academy of Sciences of the United States of America. 1998;95(23):13403-6.

122. Hsieh YC, Su LH, Hsu MH, Chiu CH. Alterations of penicillin-binding proteins in pneumococci with stepwise increase in β-lactam resistance. Pathog Dis. 2013;67(1):84-8.

123. du Plessis M, Bingen E, Klugman KP. Analysis of penicillin-binding protein genes of clinical isolates of Streptococcus pneumoniae with reduced susceptibility to amoxicillin. Antimicrobial agents and chemotherapy. 2002;46(8):2349-57.

124. Smith AM, Klugman KP. Site-Specific Mutagenesis Analysis of PBP 1A from a Penicillin-Cephalosporin-Resistant Pneumococcal Isolate. Antimicrobial agents and chemotherapy. 2003;47(1):387-9.

125. Barcus VA, Ghanekar K, Yeo M, Coffey TJ, Dowson CG. Genetics of high level penicillin resistance in clinical isolates of Streptococcus pneumoniae. FEMS Microbiol Lett. 1995;126(3):299-303.

126. Korzheva N, Davies TA, Goldschmidt R. Novel Ser79Leu and Ser81Ile substitutions in the quinolone resistance-determining regions of ParC topoisomerase IV and GyrA DNA gyrase subunits from recent fluoroquinolone-resistant Streptococcus pneumoniae clinical isolates. Antimicrobial agents and chemotherapy. 2005;49(6):2479-86.

127. Levine C, Hiasa H, Marians KJ. DNA gyrase and topoisomerase IV: biochemical activities, physiological roles during chromosome replication, and drug sensitivities. Biochim Biophys Acta. 1998;1400(1-3):29-43.

128. Drlica K, Zhao X. DNA gyrase, topoisomerase IV, and the 4-quinolones. Microbiol Mol Biol Rev. 1997;61(3):377-92.

129. Pestova E, Millichap JJ, Noskin GA, Peterson LR. Intracellular targets of moxifloxacin: a comparison with other fluoroquinolones. Journal of Antimicrobial Chemotherapy. 2000;45(5):583-90.

130. Morrissey I, George J. Activities of fluoroquinolones against Streptococcus pneumoniae type II topoisomerases purified as recombinant proteins. Antimicrobial agents and chemotherapy. 1999;43(11):2579-85.

131. Drlica K. Mechanism of fluoroquinolone action. Current Opinion in Microbiology. 1999;2(5):504-8.

132. Gillespie SH, Voelker LL, Ambler JE, Traini C, Dickens A. Fluoroquinolone resistance in Streptococcus pneumoniae: evidence that gyrA mutations arise at a lower rate and that mutation in gyrA or parC predisposes to further mutation. Microb Drug Resist. 2003;9(1):17-24.

133. Pan XS, Ambler J, Mehtar S, Fisher LM. Involvement of topoisomerase IV and DNA gyrase as ciprofloxacin targets in Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1996;40(10):2321-6.

134. Straus WL, Qazi SA, Kundi Z, Nomani NK, Schwartz B. Antimicrobial resistance and clinical effectiveness of co-trimoxazole versus amoxycillin for pneumonia among children in Pakistan: randomised controlled trial. Pakistan Co-trimoxazole Study Group. Lancet. 1998;352(9124):270-4.

135. Jacobs MR. Streptococcus pneumoniae: epidemiology and patterns of resistance. Am J Med. 2004;117 Suppl 3A:3s-15s.

136. García-de-Lomas J, Gimeno C, Millas E, Bermejo M, Lázaro MA, Navarro D, et al. Antimicrobial susceptibility of Streptococcus pneumoniae isolated from pediatric carriers in Spain. European Journal of Clinical Microbiology and Infectious Diseases. 1997;16(1):11-3.

137. Koornhof HJ, Wasas A, Klugman K. Antimicrobial resistance in Streptococcus pneumoniae: a South African perspective. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 1992;15(1):84-94.

138. Padayachee T, Klugman KP. Novel Expansions of the Gene Encoding Dihydropteroate Synthase in Trimethoprim-Sulfamethoxazole-Resistant Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1999;43(9):2225.

139. Shiota T, Baugh CM, Jackson RJ, Dillard R. Enzymic synthesis of hydroxymethyldihydropteridine Pyrophosphate and dihydrofolate. Biochemistry. 1969;8(12):5022-8.

140. Bury-Moné S. Antibacterial Therapeutic Agents: Antibiotics and Bacteriophages. Reference Module in Biomedical Sciences: Elsevier; 2014.

141. Rådström P, Swedberg G. RSF1010 and a conjugative plasmid contain sulII, one of two known genes for plasmid-borne sulfonamide resistance dihydropteroate synthase. Antimicrobial agents and chemotherapy. 1988;32(11):1684.

142. Sundström L, Rådström P, Swedberg G, Sköld O. Site-specific recombination promotes linkage between trimethoprim- and sulfonamide resistance genes. Sequence characterization of dhfrV and sulI and a recombination active locus of Tn21. Mol Gen Genet. 1988;213(2-3):191-201.

143. Maskell JP, Sefton AM, Hall LM. Mechanism of sulfonamide resistance in clinical isolates of Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1997;41(10):2121.

144. Adrian PV, Klugman KP. Mutations in the dihydrofolate reductase gene of trimethoprim-resistant isolates of Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1997;41(11):2406-13.

145. Croucher NJ, Walker D, Romero P, Lennard N, Paterson GK, Bason NC, et al. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniaeSpain23F ST81. J Bacteriol. 2009;191.

146. Chi F, Nolte O, Bergmann C, Ip M, Hakenbeck R. Crossing the barrier: evolution and spread of a major class of mosaic pbp2x in Streptococcus pneumoniae, S. mitis and S. oralis. Int J Med Microbiol. 2007;297(7-8):503-12.

147. Dowson CG, Coffey TJ, Spratt BG. Origin and molecular epidemiology of penicillin-binding-protein-mediated resistance to beta-lactam antibiotics. Trends Microbiol. 1994;2(10):361-6.

148. Reichmann P, Konig A, Linares J, Alcaide F, Tenover FC, McDougal L, et al. A global gene pool for high-level cephalosporin resistance in commensal Streptococcus species and Streptococcus pneumoniae. The Journal of infectious diseases. 1997;176(4):1001-12.

149. Hakenbeck R, Balmelle N, Weber B, Gardes C, Keck W, de Saizieu A. Mosaic genes and mosaic chromosomes: intra- and interspecies genomic variation of Streptococcus pneumoniae. Infect Immun. 2001;69(4):2477-86.

150. Hanage WP, Fraser C, Tang J, Connor TR, Corander J. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. Science (New York, NY). 2009;324.

151. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Linares J, et al. Pneumococcal capsular switching: a historical perspective. The Journal of infectious diseases. 2013;207(3):439-49.

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

153. Trzciński K, Thompson CM, Lipsitch M. Single-step capsular transformation and acquisition of penicillin resistance in Streptococcus pneumoniae. Journal of bacteriology. 2004;186(11):3447-52.

154. Syrogiannopoulos GA, Grivea IN, Tait-Kamradt A, Katopodis GD, Beratis NG, Sutcliffe J, et al. Identification of an erm(A) erythromycin resistance methylase gene in Streptococcus pneumoniae isolated in Greece. Antimicrobial agents and chemotherapy. 2001;45(1):342-4.

155. Camilli R, Del Grosso M, Iannelli F, Pantosti A. New genetic element carrying the erythromycin resistance determinant erm(TR) in Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 2008;52(2):619-25.

156. Tait-Kamradt A, Clancy J, Cronan M, Dib-Hajj F, Wondrack L, Yuan W, et al. mefE is necessary for the erythromycin-resistant M phenotype in Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1997;41(10):2251-5.

157. Weisblum B. Erythromycin resistance by ribosome modification. Antimicrobial agents and chemotherapy. 1995;39(3):577-85.

158. Johnston NJ, De Azavedo JC, Kellner JD, Low DE. Prevalence and characterization of the mechanisms of macrolide, lincosamide, and streptogramin resistance in isolates of Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 1998;42(9):2425-6.

159. Ambrose KD, Nisbet R, Stephens DS. Macrolide efflux in Streptococcus pneumoniae is mediated by a dual efflux pump (mel and mef) and is erythromycin inducible. Antimicrobial agents and chemotherapy. 2005;49(10):4203-9.

160. Cochetti I, Vecchi M, Mingoia M, Tili E, Catania MR, Manzin A, et al. Molecular characterization of pneumococci with efflux-mediated erythromycin resistance and identification of a novel mef gene subclass, mef(I). Antimicrobial agents and chemotherapy. 2005;49(12):4999-5006.

161. Chancey ST, Zhou X, Zähner D, Stephens DS. Induction of efflux-mediated macrolide resistance in Streptococcus pneumoniae. Antimicrobial agents and chemotherapy. 2011;55(7):3413-22.

162. Min YH, Kwon AR, Yoon EJ, Shim MJ, Choi EC. Translational attenuation and mRNA stabilization as mechanisms of erm(B) induction by erythromycin. Antimicrobial agents and chemotherapy. 2008;52(5):1782-9.

163. Widdowson CA, Klugman KP. The molecular mechanisms of tetracycline resistance in the pneumococcus. Microb Drug Resist. 1998;4(1):79-84.

164. Roberts AP, Mullany P. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS microbiology reviews. 2011;35(5):856-71.

165. Mingoia M, Tili E, Manso E, Varaldo PE, Montanari MP. Heterogeneity of Tn5253-like composite elements in clinical Streptococcus pneumoniae isolates. Antimicrobial agents and chemotherapy. 2011;55(4):1453-9.

166. Hakenbeck R, Bruckner R, Denapaite D, Maurer P. Molecular mechanisms of beta-lactam resistance in Streptococcus pneumoniae. Future microbiology. 2012;7(3):395-410.

167. Zerfass I, Hakenbeck R, Denapaite D. An important site in PBP2x of penicillin-resistant clinical isolates of Streptococcus pneumoniae: mutational analysis of Thr338. Antimicrobial agents and chemotherapy. 2009;53(3):1107-15.

168. Ding F, Tang P, Hsu MH, Cui P, Hu S, Yu J, et al. Genome evolution driven by host adaptations results in a more virulent and antimicrobial-resistant Streptococcus pneumoniae serotype 14. BMC genomics. 2009;10:158.

169. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9.

170. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41-7.

171. Luo Y, de Lange KM, Jostins L, Moutsianas L, Randall J, Kennedy NA, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nat Genet. 2017;49(2):186-92.

172. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A whole-genome association study of major determinants for host control of HIV-1. Science (New York, NY). 2007;317(5840):944-7.

173. Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Seron MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nature communications. 2019;10(1):2176.

174. Zhu C, Gore M, Buckler ES, Yu J. Status and Prospects of Association Mapping in Plants. The Plant Genome. 2008;1(1):5-20.

175. Alam MT, Petit RA, 3rd, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. Genome Biol Evol. 2014;6(5):1174-85.

176. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proceedings of the National Academy of Sciences. 2013;110(29):11923-7.

177. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. Nat Genet. 2013;45(10):1183-9.

178. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, et al. Predicting the virulence of MRSA from its genome sequence. Genome Research. 2014.

179. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. PLoS Genet. 2014;10(8):e1004547.

180. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Current Opinion in Microbiology. 2015;25:17-24.

181. Oppong YEA, Phelan J, Perdigão J, Machado D, Miranda A, Portugal I, et al. Genome-wide analysis of Mycobacterium tuberculosis polymorphisms reveals lineage-specific associations with drug resistance. BMC genomics. 2019;20(1):252.

182. Jauneikaite E. Population genomics of disease causing Streptococcus pneumoniae in Singapore. Southampton: University of Southampton; 2014.

183. Genome Institute of Singapore G. GERMS platform for Microbial Genomics 2020 [Available from: https://www.a-star.edu.sg/gis/Our-Science/Technology-Platforms/GERMS.

184. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-60.

185. Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40(22):11189-201.

186. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18.

187. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England). 2014;30(14):2068-9.

188. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Medicine. 2014;6(11):1-16.

189. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics (Oxford, England). 2016;33(1):128-9.

190. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics. 2012;13(1):87.

191. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, et al. Whole genome sequencing od Streptococcus pneumoniae: development, evauation and verification of targets for serogroup and serotype predication using an automated pipeline. PeerJ. 2016;10.7717/peerj.2477.

192. Smith T, Lehmann D, Montgomery J, Gratten M, Riley ID, Alpers MP. Acquisition and invasiveness of different serotypes of Streptococcus pneumoniae in young children. Epidemiol Infect. 1993;111(1):27-39.

193. MedCalc. MedCalC Software Ltd 2021 [Available from: https://www.medcalc.org/calc/odds_ratio.php.

194. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014;15(3):R46.

195. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, et al. Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. Journal of clinical microbiology. 2014;52(1):139.

196. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, et al. Benchmarking of Methods for Genomic Taxonomy. Journal of clinical microbiology. 2014;52(5):1529.

197. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018;19(1):307.

198. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol. 2013;20(10):714-37.

199. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19.

200. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics (Oxford, England). 2018;34(13):i142-i50.

201. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. A High-Resolution View of Genome-Wide Pneumococcal Transformation. PLOS Pathogens. 2012;8(6):e1002745.

202. Croucher NJ, Hanage WP, Harris SR, McGee L, van der Linden M, de Lencastre H, et al. Variable recombination dynamics during the emergence, transmission and 'disarming' of a multidrug-resistant pneumococcal clone. BMC Biology. 2014;12(1):49.

203. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics (Oxford, England). 2012;28(24):3326-8.

204. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria2017.

205. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research. 2014.

206. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641-50.

207. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PloS one. 2010;5(3):e9490.

208. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Computational Biology. 2015;11(2):e1004041.

209. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research. 2019;47(W1):W256-W9.

210. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33(6):1635-8.

211. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010;26(6):841-2.

212. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019;29(2):304-16.

213. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012;44(7):821-4.

214. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics (Oxford, England). 2015;31(22):3691-3.

215. Löytynoja A. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 2014;1079:155-70.

216. Hyams C, Camberlein E, Cohen JM, Bax K, Brown JS. The Streptococcus pneumoniae capsule inhibits complement activity and neutrophil phagocytosis by multiple mechanisms. Infection and immunity. 2010;78(2):704-15.

217. Sá-Leão R, Pinto F, Aguiar S, Nunes S, Carriço JA, Frazão N, et al. Analysis of invasiveness of pneumococcal serotypes and clones circulating in Portugal before widespread use of conjugate vaccines reveals heterogeneous behavior of clones expressing the same serotype. Journal of clinical microbiology. 2011;49(4):1369-75.

218. Shouval DS, Greenberg D, Givon-Lavi N, Porat N, Dagan R. Site-specific disease potential of individual Streptococcus pneumoniae serotypes in pediatric invasive disease, acute otitis media and acute conjunctivitis. Pediatr Infect Dis J. 2006;25(7):602-7.

219. Kronenberg A, Zucs P, Droz S, Mühlemann K. Distribution and invasiveness of Streptococcus pneumoniae serotypes in Switzerland, a country with low antibiotic selection pressure, from 2001 to 2004. Journal of clinical microbiology. 2006;44(6):2032-8.

220. Rivera-Olivero IA, del Nogal B, Sisco MC, Bogaert D, Hermans PW, de Waard JH. Carriage and invasive isolates of Streptococcus pneumoniae in Caracas, Venezuela: the relative invasiveness of serotypes and vaccine coverage. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology. 2011;30(12):1489-95.

221. Yildirim I, Hanage WP, Lipsitch M, Shea KM, Stevenson A, Finkelstein J, et al. Serotype specific invasive capacity and persistent reduction in invasive pneumococcal disease. Vaccine. 2010;29(2):283-8.

222. Sjöström K, Spindler C, Ortqvist A, Kalin M, Sandgren A, Kühlmann-Berenzon S, et al. Clonal and Capsular Types Decide Whether Pneumococci Will Act as a Primary or Opportunistic Pathogen. Clinical Infectious Diseases. 2006;42(4):451-9.

223. Verhaegen J, Flamaing J, De Backer W, Delaere B, Van Herck K, Surmont F, et al. Epidemiology and outcome of invasive pneumococcal disease among adults in Belgium, 2009–2011. Eurosurveillance. 2014;19(31):20869.

224. Martinez-Vega R, Jauneikaite E, Thoon KC, Chua HY, Huishi Chua A, Khong WX, et al. Risk factor profiles and clinical outcomes for children and adults with pneumococcal infections in Singapore: A need to expand vaccination policy? PloS one. 2019;14(10):e0220951.

225. Jansen AG, Rodenburg GD, van der Ende A, van Alphen L, Veenhoven RH, Spanjaard L, et al. Invasive pneumococcal disease among adults: associations among serotypes, disease characteristics, and outcome. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2009;49(2):e23-9.

226. Gessner BD, Mueller JE, Yaro S. African meningitis belt pneumococcal disease epidemiology indicates a need for an effective serotype 1 containing vaccine, including for older children and adults. BMC infectious diseases. 2010;10:22.

227. Ministry of Health Singapore. Principle Causes of Death 2019 [Available from: https://www.moh.gov.sg/resources-statistics/singapore-health-facts/principal-causes-of-death.

228. Low S, Chan FL, Cutter J, Ma S, Goh KT, Chew SK. A national study of the epidemiology of pneumococcal disease among hospitalised patients in Singapore: 1995 to 2004. Singapore medical journal. 2007;48(9):824-9.

229. Chong CY, Koh-Cheng T, Yee-Hui M, Nancy TW. Invasive pneumococcal disease in Singapore children. Vaccine. 2008;26(27-28):3427-31.

230. Hsu LY, Lui SW, Lee JL, Hedzlyn HM, Kong DH, Shameen S, et al. Adult invasive pneumococcal disease pre- and peri-pneumococcal conjugate vaccine introduction in a tertiary hospital in Singapore. Journal of medical microbiology. 2009;58(Pt 1):101-4.

231. Ministry of Health Singapore. List of Infectious Diseases legally Notifiable Under the Infectious Diseases Act. 2014.

232. Vasoo S, Singh K, Chow C, Lin RT, Hsu LY, Tambyah PA. Pneumococcal carriage and resistance in children attending day care centers in Singapore in an early era of PCV-7 uptake. J Infect. 2010;60(6):507-9.

233. Society of Infectious Diseases Singapore. Adult Handbook on vaccination in Singapore 2020. 2020.

234. Ho HJ, Chan YY, Ibrahim MAB, Wagle AA, Wong CM, Chow A. A formative research-guided educational intervention to improve the knowledge and attitudes of seniors towards influenza and pneumococcal vaccinations. Vaccine. 2017;35(47):6367-74.

235. Ang LW, Cutter J, James L, Goh KT. Epidemiological characteristics associated with uptake of pneumococcal vaccine among older adults living in the community in Singapore: Results from the National Health Surveillance Survey 2013. Scand J Public Health. 2018;46(2):175-81.

236. Soh SW, Poh CL, Lin RV. Serotype distribution and antimicrobial resistance of Streptococcus pneumoniae isolates from pediatric patients in Singapore. Antimicrobial agents and chemotherapy. 2000;44(8):2193-6.

237. Kim SH, Chung DR, Song JH, Baek JY, Thamlikitkul V, Wang H, et al. Changes in serotype distribution and antimicrobial resistance of Streptococcus pneumoniae isolates from adult patients in Asia: Emergence of drug-resistant non-vaccine serotypes. Vaccine. 2020;38(38):6065-73.

238. Song JH, Jung SI, Ko KS, Kim NY, Son JS, Chang HH, et al. High prevalence of antimicrobial resistance among clinical Streptococcus pneumoniae isolates in Asia (an ANSORP study). Antimicrobial agents and chemotherapy. 2004;48(6):2101-7.

239. Ling ML, Tay L. Epidemiology of pneumococcal infection in Singapore (1977-1986). Annals of the Academy of Medicine, Singapore. 1990;19(6):777-80.

240. Vasoo S, Singh K, Hsu LY, Chiew YF, Chow C, Lin RT, et al. Increasing antibiotic resistance in Streptococcus pneumoniae colonizing children attending day-care centres in Singapore. Respirology (Carlton, Vic). 2011;16(8):1241-8.

241. Ministry of Health Singapore. Communicable Diseases Survellance in Singapore 2014. 2015.

242. Ministiry of Health Singapore. Communicable Diseases Surveillance in Singapore 2018. 2019.

243. Tan TQ. Pediatric invasive pneumococcal disease in the United States in the era of pneumococcal conjugate vaccines. Clin Microbiol Rev. 2012;25(3):409-19.

244. Falush D, Bowden R. Genome-wide association mapping in bacteria? Trends in Microbiology. 2006;14(8):353-5.

245. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155(3):1405-13.

246. Griffith F. The Significance of Pneumococcal Types. The Journal of Hygiene. 1928;27(2):113-59.

247. Feil EJ, Spratt BG. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol. 2001;55:561-90.

248. Lacks S, Greenberg B. Single-strand breakage on binding of DNA to cells in the genetic transformation of Diplococcus pneumoniae. Journal of molecular biology. 1976;101(2):255-75.

249. Mejean V, Claverys JP. DNA processing during entry in transformation of Streptococcus pneumoniae. The Journal of biological chemistry. 1993;268(8):5594-9.

250. Lacks S, Greenberg B, Neuberger M. Role of a deoxyribonuclease in the genetic transformation of Diplococcus pneumoniae. Proceedings of the National Academy of Sciences of the United States of America. 1974;71(6):2305-9.

251. Morrison DA. Transformation in pneumococcus: protein content of eclipse complex. Journal of bacteriology. 1978;136(2):548-57.

252. Blaser MJ, Berg DE. Helicobacter pylori genetic diversity and risk of human disease. J Clin Invest. 2001;107(7):767-73.

253. Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, Marttinen P. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. Mol Biol Evol. 2017;34(5):1167-82.

254. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proceedings of the National Academy of Sciences of the United States of America. 2001;98(1):182-7.

255. Mobegi FM, Cremers AJ, de Jonge MI, Bentley SD, van Hijum SA, Zomer A. Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data. Scientific reports. 2017;7:42808.

256. Baltrus DA. Exploring the costs of horizontal gene transfer. Trends Ecol Evol. 2013;28(8):489-95.

257. Engelmoer DJP, Donaldson I, Rozen DE. Conservative Sex and the Benefits of Transformation in Streptococcus pneumoniae. PLOS Pathogens. 2013;9(11):e1003758.

258. Namouchi A, Didelot X, Schock U, Gicquel B, Rocha EP. After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. Genome Res. 2012;22(4):721-34.

259. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, Suerbaum S. Genomic evolution and transmission of Helicobacter pylori in two South African families. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(34):13880-5.

260. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, et al. Recombination and Population Structure in Salmonella enterica. PLOS Genetics. 2011;7(7):e1002191.

261. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in Escherichia coli: an evolutionary perspective. Mol Microbiol. 2006;60(5):1136-51.

262. Croucher NJ, Mitchell AM, Gould KA, Inverarity D, Barquist L, Feltwell T, et al. Dominant Role of Nucleotide Substitution in the Diversification of Serotype 3 Pneumococci over Decades and during a Single Infection. PLOS Genetics. 2013;9(10):e1003868.

263. Muzzi A, Donati C. Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. Int J Med Microbiol. 2011;301(8):619-22.

264. Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, et al. Generation of Genic Diversity among Streptococcus pneumoniae Strains via Horizontal Gene Transfer during a Chronic Polyclonal Pediatric Infection. PLOS Pathogens. 2010;6(9):e1001108.

265. Marks LR, Reddinger RM, Hakansson AP. High levels of genetic recombination during nasopharyngeal carriage and biofilm formation in Streptococcus pneumoniae. mBio. 2012;3(5).

266. Donkor ES, Bishop CJ, Gould KA, Hinds J, Antonio M, Wren B, et al. High levels of recombination among Streptococcus pneumoniae isolates from the Gambia. mBio. 2011;2(3):e00040-e11.

267. Morrison DA, Baker MF. Competence for genetic transformation in pneumococcus depends on synthesis of a small set of proteins. Nature. 1979;282(5735):215-7.

268. Campbell EA, Choi SY, Masure HR. A competence regulon in Streptococcus pneumoniae revealed by genomic analysis. Molecular Microbiology. 1998;27(5):929-39.

269. Claverys JP, Martin B, Havarstein LS. Competence-induced fratricide in streptococci. Mol Microbiol. 2007;64(6):1423-33.

270. Salvadori G, Junges R, Morrison DA, Petersen FC. Competence in Streptococcus pneumoniae and Close Commensal Relatives: Mechanisms and Implications. Frontiers in cellular and infection microbiology. 2019;9(94).

271. Evans BA, Rozen DE. Significant variation in transformation frequency in Streptococcus pneumoniae. The ISME journal. 2013;7(4):791-9.

272. Mostowy R, Croucher NJ, Hanage WP, Harris SR, Bentley S, Fraser C. Heterogeneity in the Frequency and Characteristics of Homologous Recombination in Pneumococcal Evolution. PLOS Genetics. 2014;10(5):e1004300.

273. Claverys JP, Martin B, Polard P. The genetic transformation machinery: composition, localization, and mechanism. FEMS microbiology reviews. 2009;33(3):643-56.

274. Feil EJ, Smith JM, Enright MC, Spratt BG. Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data. Genetics. 2000;154(4):1439-50.

275. Méjean V, Claverys JP. DNA processing during entry in transformation of Streptococcus pneumoniae. The Journal of biological chemistry. 1993;268(8):5594-9.

276. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Linares J, et al. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. Genome Biol. 2012;13(11):R103.

277. Cowley LA, Petersen FC, Junges R, Jimson DJM, Morrison DA, Hanage WP. Evolution via recombination: Cell-to-cell contact facilitates larger recombination events in Streptococcus pneumoniae. PLoS Genet. 2018;14(6):e1007410.

278. Johnston C, Hinds J, Smith A, van der Linden M, Van Eldere J, Mitchell TJ. Detection of large numbers of pneumococcal virulence genes in streptococci of the mitis group. Journal of clinical microbiology. 2010;48(8):2762-9.

279. King SJ, Whatmore AM, Dowson CG. NanA, a neuraminidase from Streptococcus pneumoniae, shows high levels of sequence diversity, at least in part through recombination with Streptococcus oralis. J Bacteriol. 2005;187(15):5376-86.

280. Neeleman C, Klaassen CH, Klomberg DM, de Valk HA, Mouton JW. Pneumolysin is a key factor in misidentification of macrolide-resistant Streptococcus pneumoniae and is a putative virulence factor of S. mitis and other streptococci. Journal of clinical microbiology. 2004;42(9):4355-7.

281. Sibold C, Henrichsen J, Konig A, Martin C, Chalkley L, Hakenbeck R. Mosaic pbpX genes of major clones of penicillin-resistant Streptococcus pneumoniae have evolved from pbpX genes of a penicillin-sensitive Streptococcus oralis. Mol Microbiol. 1994;12(6):1013-23.

282. Dowson CG, Coffey TJ, Kell C, Whiley RA. Evolution of penicillin resistance in Streptococcus pneumoniae; the role of Streptococcus mitis in the formation of a low affinity PBP2B in S. pneumoniae. Mol Microbiol. 1993;9(3):635-43.

283. Posada DC, Keith A.; Holmes Edward C.;. Recombination in Evolutionary Genomics. Annual Review of Genetics. 2002;36(1):75-97.

284. Harris SR, Clarke IN, Seth-Smith HM, Solomon AW, Cutcliffe LT, Marsh P, et al. Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. Nat Genet. 2012;44(4):413-9, s1.

285. Pupko T, Pe'er I, Shamir R, Graur D. A fast algorithm for joint reconstruction of ancestral amino acid sequences. Mol Biol Evol. 2000;17(6):890-6.

286. Croucher NJ, Finkelstein JA, Pelton SI, Parkhill J, Bentley SD, Lipsitch M, et al. Population genomic datasets describing the post-vaccine evolutionary epidemiology of Streptococcus pneumoniae. Scientific data. 2015;2:150058.

287. Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, et al. Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. 2012;40.

288. MacLean RC, San Millan A. The evolution of antibiotic resistance. Science (New York, NY). 2019;365(6458):1082.

289. Lehtinen S, Chewapreecha C, Lees J, Hanage WP, Lipsitch M, Croucher NJ, et al. Horizontal gene transfer rate is not the primary determinant of observed antibiotic resistance frequencies in Streptococcus pneumoniae. Science Advances. 2020;6(21):eaaz6137.

290. Young BC, Earle SG, Soeng S, Sar P, Kumar V, Hor S, et al. Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. eLife. 2019;8.

291. Galardini M, Clermont O, Baron A, Busby B, Dion S, Schubert S, et al. Major role of the high-pathogenicity island (HPI) in the intrinsic extra-intestinal virulence of Escherichia coli revealed by a genome-wide association study. bioRxiv. 2019:712034.

292. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nature communications. 2016;7:12797.

293. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, et al. Genomic and functional analyses of Mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. Nat Genet. 2016;48(5):544-51.

294. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nature microbiology. 2016;1:16041.

295. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant Mycobacterium tuberculosis. Nat Genet. 2018;50(2):307-16.

296. Phelan J, Coll F, McNerney R, Ascher DB, Pires DE, Furnham N, et al. Mycobacterium tuberculosis whole genome sequencing and protein structure modelling provides insights into anti-tuberculosis drug resistance. BMC medicine. 2016;14:31.

297. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, Amaratunga C, et al. Genetic architecture of artemisinin-resistant Plasmodium falciparum. Nat Genet. 2015;47(3):226-34.

298. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, et al. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Genome Research. 2015;25(1):119-28.

299. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, et al. Genome-Wide Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. PloS one. 2016;11(9):e0163746.

300. Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. eLife. 2013;2:e01123.

301. Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET, et al. Common Genetic Variation and the Control of HIV-1 in Humans. PLOS Genetics. 2009;5(12):e1000791.

302. Davila S, Wright VJ, Khor CC, Sim KS, Binder A, Breunis WB, et al. Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet. 2010;42(9):772-6.

303. Tian C, Hromatka BS, Kiefer AK, Eriksson N, Noble SM, Tung JY, et al. Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nature communications. 2017;8(1):599.

304. Rautanen A, Pirinen M, Mills TC, Rockett KA, Strange A, Ndungu AW, et al. Polymorphism in a lincRNA Associates with a Doubled Risk of Pneumococcal Bacteremia in Kenyan Children. American journal of human genetics. 2016;98(6):1092-100.

305. Song JY, Nahm MH, Moseley MA. Clinical implications of pneumococcal serotypes: invasive disease potential, clinical presentations, and antibiotic resistance. Journal of Korean medical science. 2013;28(1):4-15.

306. Lees JA, Croucher NJ, Goldblatt D, Nosten F, Parkhill J, Turner C, et al. Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. eLife. 2017;6.

307. Piet JR, Geldhoff M, van Schaik BDC, Brouwer MC, Valls Seron M, Jakobs ME, et al. Streptococcus pneumoniae Arginine Synthesis Genes Promote Growth and Virulence in Pneumococcal Meningitis. The Journal of infectious diseases. 2013;209(11):1781-91.

308. Tunjungputri RN, Mobegi FM, Cremers AJ, van der Gaast-de Jongh CE, Ferwerda G, Meis JF, et al. Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease. mBio. 2017;8(1).

309. Chaguza C, Yang M, Cornick JE, du Plessis M, Gladstone RA, Kwambana-Adams BA, et al. Bacterial genome-wide association study of hyper-virulent pneumococcal serotype 1 identifies genetic variation associated with neurotropism. Communications Biology. 2020;3(1):559.

310. Goldstein DB. Common genetic variation and human traits. The New England journal of medicine. 2009;360(17):1696-8.

311. Halperin E, Eskin E. Haplotype reconstruction from genotype data using Imperfect Phylogeny. Bioinformatics (Oxford, England). 2004;20(12):1842-9.

312. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004;36(5):512-7.

313. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459-63.

314. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997-1004.

315. van Hemert S, Meijerink M, Molenaar D, Bron PA, de Vos P, Kleerebezem M, et al. Identification of Lactobacillus plantarum genes modulating the cytokine response of human peripheral blood mononuclear cells. BMC Microbiology. 2010;10(1):1-13.

316. Chaston JM, Newell PD, Douglas AE. Metagenome-Wide Association of Microbial Determinants of Host Phenotype in Drosophila melanogaster. mBio. 2014;5(5).

317. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, et al. Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. Genome Research. 2014.

318. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 2008;9(1):539.

319. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.

320. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81.

321. Thornton T, McPeek MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. American journal of human genetics. 2010;86(2):172-84.

322. San JE, Baichoo S, Kanzi A, Moosa Y, Lessells R, Fonseca V, et al. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. Frontiers in microbiology. 2020;10(3119).

323. Corander J, Tang J. Bayesian analysis of population structure based on linked molecular information. Mathematical Biosciences. 2007;205(1):19-31.

324. Zhu C, Yu J. Nonmetric Multidimensional Scaling Corrects for Population Structure in Association Mapping With Different Sample Types. Genetics. 2009;182(3):875.

325. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010;42(4):348-54.

326. Kenny EE, Kim M, Gusev A, Lowe JK, Salit J, Smith JG, et al. Increased power of mixed models facilitates association mapping of 10 loci for metabolic traits in an isolated population. Hum Mol Genet. 2011;20(4):827-39.

327. Hoffman GE. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PloS one. 2013;8(10):e75707.

328. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017;18(1):41-50.

329. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika. 1988;75(2):383-6.

330. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800-2.

331. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.

332. Kaler AS, Purcell LC. Estimation of a significance threshold for genome-wide association studies. BMC genomics. 2019;20(1):618.

333. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;316(7139):1236-8.

334. Cleverley RM, Rutter ZJ, Rismondo J, Corona F, Tsui H-CT, Alatawi FA, et al. The cell cycle regulator GpsB functions as cytosolic adaptor for multiple cell wall enzymes. Nature communications. 2019;10(1):261-.

335. Land AD, Tsui HC, Kocaoglu O, Vella SA, Shaw SL, Keen SK, et al. Requirement of essential Pbp2x and GpsB for septal ring closure in Streptococcus pneumoniae D39. Mol Microbiol. 2013;90(5):939-55.

336. Huovinen P. Resistance to trimethoprim-sulfamethoxazole. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2001;32(11):1608-14.

337. Buwembo W, Aery S, Rwenyonyi CM, Swedberg G, Kironde F. Point Mutations in the folP Gene Partly Explain Sulfonamide Resistance of Streptococcus mutans. International Journal of Microbiology. 2013;2013:367021.

338. Schyns G, Potot S, Geng Y, Barbosa TM, Henriques A, Perkins JB. Isolation and characterization of new thiamine-deregulated mutants of Bacillus subtilis. J Bacteriol. 2005;187(23):8127-36.

339. Zhang Z, Bulloch EM, Bunker RD, Baker EN, Squire CJ. Structure and function of GlmU from Mycobacterium tuberculosis. Acta Crystallogr D Biol Crystallogr. 2009;65(Pt 3):275-83.

340. Ilic S, Cohen S, Singh M, Tam B, Dayan A, Akabayov B. DnaG Primase-A Target for the Development of Novel Antibacterial Agents. Antibiotics (Basel, Switzerland). 2018;7(3).

341. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, et al. Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science (New York, NY). 2001;293.

342. Hoskins J, Alborn WE, Jr., Arnold J, Blaszczak LC, Burgett S, DeHoff BS, et al. Genome of the bacterium Streptococcus pneumoniae strain R6. J Bacteriol. 2001;183(19):5709-17.

343. Murphy PB BK, Le JK. Clindamycin. StatPearls; Updated 2020 Jun 28.

344. Clancy E, Higgins O, Forrest MS, Boo TW, Cormican M, Barry T, et al. Development of a rapid recombinase polymerase amplification assay for the detection of Streptococcus pneumoniae in whole blood. BMC infectious diseases. 2015;15(1):481.

345. Lemos JAC, Burne RA. Regulation and Physiological Significance of ClpC and ClpP in <em>Streptococcus mutans</em&gt. Journal of Bacteriology. 2002;184(22):6357.

346. Capiaux H, Lesterlin C, Pérals K, Louarn JM, Cornet F. A dual role for the FtsK protein in Escherichia coli chromosome segregation. EMBO reports. 2002;3(6):532-6.

347. Borrell S, Gagneux S. Strain diversity, epistasis and the evolution of drug resistance in Mycobacterium tuberculosis. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2011;17(6):815-20.

348. Weinreich DM, Watson RA, Chao L. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution. 2005;59(6):1165-74.

349. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics (Oxford, England). 2018;34(24):4310-2.

350. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS Genet. 2018;14(11):e1007758.

351. Islam MR, Hoque MN, Rahman MS, Alam ASMRU, Akther M, Puspo JA, et al. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Scientific reports. 2020;10(1):14004.

352. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G-W. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Communications Biology. 2021;4(1):228.

353. Kim J-S, Jang J-H, Kim J-M, Chung Y-S, Yoo C-K, Han M-G. Genome-Wide Identification and Characterization of Point Mutations in the SARS-CoV-2 Genome. Osong Public Health Res Perspect. 2020;11(3):101-11.