Using genomic epidemiology to improve molecular diagnostics for *Streptococcus pneumoniae* and to study streptococcal population structure



Femke Maria Ahlers

Lady Margaret Hall

Nuffield Department of Population Health

Big Data Institute

University of Oxford

United Kingdom

This thesis is submitted for the degree of

Doctor of Philosophy in Population Health

Trinity Term 2024

Abstract

Using genomic epidemiology to improve molecular diagnostics for *Streptococcus pneumoniae* and to study streptococcal population structure

Thesis submitted in Trinity Term 2024 for the degree of *Doctor of Philosophy*

Femke Ahlers

Various *Streptococcus* species, including *Streptococcus pneumoniae* (pneumococcus), can asymptomatically colonise the human upper respiratory tract. Pneumococci can also cause life-threatening infections such as meningitis and pneumonia, constituting a large burden on global health. Using large bacterial genome datasets, this thesis aimed to improve nucleic acid amplification tests used to identify pneumococci and to enhance our understanding of the streptococcal population structure.

First, I assessed the *in silico* performance of existing PCR-based diagnostics in 7,547 pneumococcal genomes and 1,825 genomes of 55 non-pneumococcal *Streptococcus* (NPS) species. I designed a new multiplex quantitative PCR (qPCR) assay targeting Xisco and SP2020 and validated this assay *in vitro*, finding high specificity, sensitivity, and accurate performance across a range of pneumococcal amplicons.

Building on these results, I assessed the *in silico* performance of existing pneumococcal loop-mediated isothermal amplification (LAMP) assays, revealing cross-reactivity with up to ten NPS species. I designed and validated *in vitro* a

highly sensitive, specific and fast new real-time LAMP assay targeting SP2020, which could be particularly useful in low-resource settings.

Finally, I compiled a diverse *Streptococcus* genome dataset to investigate *in silico* differences across 105 species, including guanine-cytosine content and total assembly length. Using two different methods, I studied the size and function of the streptococcal core genome. Whilst seven single locus typing targets were unable to accurately stratify *S. pneumoniae* from closely related viridans streptococci, the phylogeny based on 412 streptococcal core genes provided a robust framework for mapping the *in silico* presence of antimicrobial resistance genes and virulence factors across the genus.

This thesis harnessed the power of genomic data to improve pneumococcal diagnostic assays and to study differences between streptococcal species. The new multiplex qPCR and real-time LAMP assays can facilitate improved pneumococcal identification in clinical and research microbiology laboratories and the *in silico* plus *in vitro* methodology can be applied to improve diagnostics for other pathogens. The insights gained into streptococcal population structure contribute to an enhanced understanding of streptococcal evolution, epidemiology and virulence factor distribution.

Contents

Abstract **2**

Contents **4**

List of Tables **10**

List of Figures **12**

Acknowledgements **15**

Funding **17**

Declaration of contribution **18**

Chapters 2 and 3 18

Chapter 4 19

Planned publications **20**

Abbreviations **21**

Measurement Units **23**

1 General Introduction **24**

1.1 The *Streptococcus* genus 25

1.1.1 Historical perspective 25

1.1.2 Phenotypic and biochemical classifications 26

1.1.2.1 Haemolysis types 26

1.1.2.2 Lancefield serogroups 26

1.1.3 Streptococcal disease 27

1.1.3.1 Human disease 27

1.1.3.2 Animal disease 28

1.2 *Streptococcus pneumoniae* 30

1.2.1 Historical perspective 30

1.2.2 Epidemiology of pneumococcal carriage and disease 30

1.2.2.1 Pneumococcal carriage 31

1.2.2.2 Pneumococcal disease 32

1.2.3 Pneumococcal serotypes 33

1.2.4 Pneumococcal vaccines 34

1.2.5 Horizontal gene transfer 35

1.2.6 Pneumococcal diagnostics 35
1.3 Sequencing bacterial genomes 37
1.3.1 Historical perspective 37
1.3.2 High-throughput next-generation sequencing methods 38
1.3.3 Single-molecule sequencing 38
1.3.4 Genome quality control metrics 39
1.3.5 Repurposing the wealth of WGS data 39
1.4 Bacterial genotyping 41
1.4.1 Historical perspective 41
1.4.2 WGS-informed diagnostics 41
1.4.3 WGS-informed vaccines 42
1.4.4 Bacterial genomic epidemiology 43
1.4.4.1 SNP-based methods 43
1.4.4.2 *k*-mer-based methods 44
1.4.4.3 Gene-by-gene typing methods 44
1.4.5 Phylogenetic reconstruction methods 52
1.5 Thesis structure and aims 53

2 Improving PCR-based diagnostics used for the identification of *Streptococcus pneumoniae* through *in silico* analysis of large bacterial genome datasets 56

2.1 Introduction 57
2.1.1 Pneumococcal identification 57
2.1.2 PCR-based identification 58
2.1.3 Pneumococcal PCR targets 60
2.1.4 Chapter aims 61
2.2 Methods 62
2.2.1 Compilation of publicly available genomes for *in silico* analyses 62
2.2.2 Assessment of the *in silico* performance of the PCR assays 63
2.2.3 Genome sequencing and assembly of UKHSA's streptococci ... 65
2.2.4 Compilation of panels for *in vitro* testing 66
2.2.5 Purification of DNA extracts for *in vitro* panels 68
2.2.6 qPCR primer and probe design 68
2.2.7 Multiplex qPCR protocol used to test the three *in vitro* panels ... 69

2.2.8 Limit of detection experiments 70
2.2.9 Data analysis..... 70
2.3 Results 72
2.3.1 Identification of published assays and design of new assays 72
2.3.2 Design of *in silico* datasets and *in vitro* panels..... 72
2.3.3 *In silico* assessment in the pneumococcal study dataset..... 74
2.3.4 *In silico* assessment in the NPS dataset 78
2.3.5 Investigating genomes with discrepant results..... 79
2.3.6 Youden's J statistics of published assays 81
2.3.7 Limit of detection experiments 82
2.3.8 Testing the *in vitro* sensitivity panel..... 85
2.3.9 Testing the *in vitro* specificity panel..... 85
2.3.10 Design of the pneumococcal amplicon diversity panel..... 86
2.3.11 Testing the *in vitro* pneumococcal amplicon diversity panel..... 88
2.4 Discussion..... 92

3 Development of a real-time loop-mediated isothermal amplification (LAMP) assay targeting SP2020 for the identification of *Streptococcus pneumoniae* informed by large genomic datasets 98
3.1 Introduction 99
3.1.1 Loop-mediated isothermal amplification (LAMP)..... 99
3.1.2 LAMP detection methods..... 100
3.1.3 Published pneumococcal LAMP targets..... 102
3.1.4 Chapter aims..... 103
3.2 Methods 104
3.2.1 Design of new real-time LAMP assay targeting SP2020..... 104
3.2.2 *In vitro* protocol for LAMP assays..... 106
3.2.3 Assessment of the *in silico* performance of LAMP assays..... 106
3.2.4 Phylogenetic clustering of NPS genomes testing *in silico* positive. 109
3.2.5 Selection of isolates for *in vitro* testing..... 109
3.2.6 qPCR protocol of published SP2020 assay 110
3.2.7 Data analysis..... 110
3.3 Results 112

3.3.1 Development of a real-time LAMP assay targeting SP2020 112
3.3.2 Identification of published pneumococcal LAMP assays..... 113
3.3.3 *In silico* assessment in the pneumococcal study dataset..... 114
3.3.4 *In silico* assessment in the NPS dataset 117
3.3.5 *In silico* Youden's J statistic of LAMP assays..... 118
3.3.6 Investigating the *in silico* discrepant results 119
3.3.7 Phylogenetic clustering of NPS genomes testing *in silico* positive. 121
3.3.8 Limit of detection of SP2020_3+ 123
3.3.9 *In vitro* validation of the LAMP assay 124
3.3.10 SP2020 diversity panel 124
3.4 Discussion..... 126

4 Understanding differences between species in the *Streptococcus* genus using *in silico* methods 130

4.1 Introduction 131
4.1.1 The *Streptococcus* genus 131
4.1.2 Phenotypic methods to differentiate streptococci..... 131
4.1.3 Genotypic methods to differentiate streptococci 133
4.1.3.1 Single gene approaches 133
4.1.3.2 Core genome approaches 135
4.1.4 Virulence factors 139
4.1.4.1 Antimicrobial resistance (AMR)..... 139
4.1.4.2 Pneumococcal capsular locus..... 140
4.1.5 Chapter aims..... 141
4.2 Methods 142
4.2.1 Compilation of the *Streptococcus* genome dataset..... 142
4.2.2 rMLST species identification and clustering..... 143
4.2.3 Single locus sequence typing clustering 144
4.2.4 Investigation of the *Streptococcus* core genome..... 145
4.2.4.1 Unidirectional BLASTN search..... 145
4.2.4.2 PROKKA and Panaroo 146
4.2.5 Construction of phylogenetic trees..... 147

4.2.6 Investigation of the presence of pneumococcal virulence factors 148
4.2.6.1 qPCR assays for *S. pneumoniae* AMR genes 148
4.2.6.2 PCR assays for *S. pneumoniae* serotypes/serogroups.... 149
4.2.7 Data analysis and visualisation 153
4.3 Results 154
4.3.1 Compilation of a *Streptococcus* genome dataset 154
4.3.2 rMLST clustering 157
4.3.3 Single locus sequence typing clustering 164
4.3.4 Investigation of the *Streptococcus* core genome..... 171
4.3.4.1 Unidirectional BLASTN search..... 171
4.3.4.2 Panaroo 177
4.3.5 Investigation of the presence of pneumococcal virulence factors 181
4.3.5.1 In silico amplicons of AMR genes 181
4.3.5.2 In silico amplicons of *S. pneumoniae* serotypes 184
4.4 Discussion 187
5 General Discussion..... 193
5.1 Summary of findings..... 194
5.1.1 A new multiplex qPCR assay for pneumococcal identification . 194
5.1.2 A new real-time LAMP assay for pneumococcal identification . 195
5.1.3 A streptococcal core genome phylogeny to map virulence factors. 196
5.2 Strengths..... 198
5.3 Limitations 200
5.4 Future work 202
5.4.1 Validation of diagnostic assays to directly test clinical samples202
5.4.2 Diagnostic surveillance in the African Meningitis Belt 203
5.4.3 Applications of NAATs to monitor AMR..... 204
5.4.4 Differentiation of viridans group streptococci 204
5.4.5 Application of WGS in pneumococcal vaccine design and beyond 205

5.5 Overall conclusions 206

6 References 207

7 Appendices 240

7.1 Appendices to Chapter 2 240

7.1.1 Nucleotide sequence alignments of *in silico* PCR amplicons ... 240

7.2 Appendices to Chapter 3 245

7.2.1 Biophysical metrics of pneumococcal LAMP primers 245

7.2.2 Pneumococcal genomes used for phylogenetic analysis of *in silico* positive LAMP assays in NPS genomes 246

7.3 International conference abstracts 247

7.3.1 Abstract poster presentation ISPPD 2022 247

7.3.2 Abstract oral presentation ECCMID 2023 249

7.3.3 Abstract poster presentation ECCMID 2023 251

List of Tables

Table 1.1 Clinically relevant *Streptococcus* species in humans and animals. 29

Table 1.2 Pneumococcal multilocus sequence typing (MLST) scheme. 47

Table 1.3 Phylogenetic reconstruction methods 52

Table 2.1 Targets of pneumococcal PCR-based identification assays. 61

Table 2.2 Published and newly designed primers and probes used in PCR and qPCR assays and the length of their amplicons. 64

Table 2.3 *In silico* assessment of published and newly designed primers and probes within the pneumococcal genome dataset (n=7,547). 77

Table 2.4 Summary of the *in silico* assessment of published and newly designed primers and probes within the pneumococcal genome dataset (n=7,547). 77

Table 2.5 Summary of the *in silico* assessment of published and new primers and probes within the non-pneumococcal *Streptococcus* genome dataset (n=1,825). 79

Table 2.6 Biophysical metrics of primers and probes of pneumococcal PCR assays. 81

Table 2.7 C_q-values of limit of detection dilution series of multiplex qPCR assays. 84

Table 2.8 Non-pneumococcal *Streptococcus* species in the *in vitro* specificity panel (n=277). 86

Table 2.9 Results of the *in silico* analyses of pneumococcal genomes from UKHSA (n=6,090). 87

Table 2.10 Summary of *in silico* results for UKHSA pneumococci of the amplicon diversity panel (n=248). 88

Table 2.11 Summary of the *in silico* and *in vitro* test results of the amplicon diversity panel (n=248). 89

Table 3.1 Newly designed loop-mediated isothermal amplification primers targeting the pneumococcal SP2020 gene..... 105
Table 3.2 Published pneumococcal loop-mediated isothermal amplification primers. 108
Table 3.3 *In silico* assessment of published and newly designed LAMP primers within the pneumococcal genome dataset (n=7,547). 114
Table 3.4 Summary of the *in silico* assessment of published and newly designed LAMP primers within the pneumococcal genome dataset (n=7,547). 115
Table 3.5 *In silico* assessment of published and newly designed LAMP primers within the non-pneumococcal *Streptococcus* dataset (n=1,825). 117
Table 3.6 Summary of the *in silico* assessment of published and newly designed LAMP primers within the non-pneumococcal *Streptococcus* dataset (n=1,825). 118
Table 4.1 Published single gene typing targets [219–223]. 134
Table 4.2 Protein-encoding single gene typing targets [69,219–221]. 144
Table 4.3 Settings used for Genome Comparator and Panaroo [73,226]..... 146
Table 4.4 Primers and probes recommended by CDC for pneumococcal AMR detection. 149
Table 4.5 Standard PCR primers recommended by the CDC for pneumococcal serotype/serogroup determination. 151
Table 4.6 Presence of protein-encoding single locus typing genes in *Streptococcus* dataset (n=1,919). 164
Table 4.7 *In silico* presence of qPCR amplicons for AMR genes in *Streptococcus* dataset (n=1,919) 181
Table 4.8 Pneumococcal serotypes/serogroups PCR assays resulting in one or more positive *in silico* hits in NPS genomes (n=1,819)..... 185

List of Figures

Figure 1.1 Morphology of streptococcal cells [7]. 25

Figure 1.2 *Streptococcus pneumoniae* carriage and disease [24]. 31

Figure 1.3 Increase in genomic data available at GenBank from 2002 to 2024. .. 40

Figure 1.4 Position of seven pneumococcal multilocus sequence typing (MLST) genes on the pneumococcal reference genome TIGR4 [69,73,114]. 46

Figure 1.5 Schematic overview of pneumococcal multilocus sequence typing. ... 49

Figure 1.6 Position of 53 pneumococcal ribosomal multilocus sequence typing (rMLST) loci on *Streptococcus pneumoniae* genome TIGR4 [69,70,73,89]. 50

Figure 2.1 Schematic overview of qPCR using fluorescent probes. 59

Figure 2.2 Schematic overview of the genome datasets and bacterial isolates.... 73

Figure 2.3 Genome metrics of the *in silico* pneumococcal study dataset (n=7,547). 74

Figure 2.4 Summary of the *in silico* pneumococcal study dataset (n=7,547)..... 75

Figure 2.5 Description of the non-pneumococcal *Streptococcus* (NPS) species genome dataset (n=1,825). 78

Figure 2.6 Summary of pneumococcal genomes (A) and non-pneumococcal *Streptococcus* (B) genomes causing discrepancies after *in silico* assessments of pneumococcal PCR assays..... 80

Figure 2.7 Limit of detection dilution series of Xisco_1, Xisco_2, Xisco_3 and SP2020_new singleplex assays. 83

Figure 2.8 Multiplex qPCR assays tested on the *in vitro* amplicon diversity panel. 90

Figure 2.9 Relationship between the quantification cycles (C_q-values) of the pneumococcal targets and 16S rRNA gene positive control..... 91

Figure 3.1 Schematic overview of loop-mediated isothermal amplification (LAMP). 101
Figure 3.2 Assessment of newly designed LAMP primer combinations. 112
Figure 3.3 Alignment of qPCR primer/probe and LAMP primer binding regions to pneumococcal target genes. 113
Figure 3.4 Amplicon diversity of the new SP2020_3+ LAMP assay in the *in silico* pneumococcal study dataset of 7,547 genomes. 116
Figure 3.5 Summary of pneumococcal genomes (A) and non-pneumococcal *Streptococcus* (NPS) (B) genomes causing discrepancies after *in silico* assessments of pneumococcal LAMP assays..... 120
Figure 3.6 Clustering of NPS genomes testing *in silico* positive for LAMP assays. 122
Figure 3.7 Limit of detection experiments of the SP2020_3+ LAMP assay. 123
Figure 3.8 Fifteen unique SP2020_3+ LAMP amplicon sequences of the *in vitro* SP2020 diversity panel..... 125
Figure 4.1 Schematic overview of the compilation of the *Streptococcus* genome dataset and application of genome quality control metrics. 155
Figure 4.2 Descriptive metrics of the *Streptococcus* dataset..... 156
Figure 4.3 Maximum likelihood phylogeny based on nucleotide sequences of 51 non-paralogous rMLST loci of 1,919 *Streptococcus* genomes. 159
Figure 4.4 Distribution of GC-content and total assembly size of *Streptococcus* species. 161
Figure 4.5 Differences in GC-content across *Streptococcus* species. 163
Figure 4.6 Maximum likelihood trees based on the *groEL* and *gyrB* gene sequences in the *Streptococcus* dataset (n=1,919). 167

Figure 4.7 Maximum likelihood trees based on the *recA* and *rpoB* gene sequences in the *Streptococcus* dataset (n=1,919). 168
Figure 4.8 Maximum likelihood trees based on the *sodA* and *tuf* gene sequences in the *Streptococcus* dataset (n=1,919). 169
Figure 4.9 Maximum likelihood trees based on the *rpsB* gene and *rpsB* amplicon fragment of the *Streptococcus* dataset (n=1,919). 170
Figure 4.10 Different settings result in different streptococcal core genome sizes. 172
Figure 4.11 Mean *p*-distances of the inner core, soft core, shell and cloud genes. 174
Figure 4.12 Functional classifications of the core genomes defined using different Genome Comparator settings compared to the coding sequences of *Streptococcus pneumoniae* reference genome TIGR4. 175
Figure 4.13 Closely related streptococcal genomes share more of the *S. pneumoniae* TIGR4 coding genes. 176
Figure 4.14 Different settings of Panaroo result in different numbers of inner core, soft core, shell and cloud genes. 178
Figure 4.15 Functional classifications of the 412 core genes defined using Panaroo setting 5. 179
Figure 4.16 Maximum likelihood tree based on the nucleotide alignment of 412 core genes of 1,919 *Streptococcus* genomes. 180
Figure 4.17 Clustering of AMR genes among *Streptococcus* genomes. 183
Figure 4.18 Clustering of *in silico* amplicon presence of pneumococcal serotypes/serogroups in NPS species on a core genome phylogeny..... 186

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my primary supervisor Professor David Eyre for his invaluable support and mentorship. Throughout my DPhil, I have been in awe of David's vast clinical knowledge and technical skillset and am very grateful for his kindness and patience when explaining new concepts to me. I would also like to express my gratitude to my second supervisor Professor Derrick Crook for being a great source of encouragement and for helping me with the timely completion of this thesis. This thesis would not have been possible without the fantastic academic and personal guidance I received from my supervisors, for which I am truly grateful.

I would like to express my deepest thanks to Dr David Litt and Dr Norman Fry, with whom I collaborated at the United Kingdom Health Security Agency (UKHSA). I could not have been welcomed more warmly by them and I am grateful for their kindness whilst answering my questions and teaching me new skills in their laboratory. I also thank Dr Seyi Eletu, Dr Carmen Clark, Dr Juliana Coelho, Dr Joshua D'Aeth, Dr Bruno Pichon and Samuel Rose for making my time at UKHSA both very enjoyable and productive.

Research in the field of genomic epidemiology is highly collaborative, and I gratefully acknowledge the countless number of researchers that generated the bacterial genome data and created the software I used in this thesis. This thesis builds on a foundation of work conducted by others at the University of Oxford, including Professor Angela Brueggemann, who also supervised the first two results chapters of this thesis, and Dr Melissa Jansen van Rensburg. I am grateful for the technical expertise and scientific discussions they provided. I thank Dr

Keith Jolley and Dr James Bray for their work on the PubMLST databases and for always being willing to answer my questions. I thank Professor Daniel Wilson and Dr Bernadette Young for their insightful feedback and the scientific discussions during my Transfer of Status and Confirmation of Status vivas.

I would like to thank the postdocs and students – past and present – that I met in the Big Data Institute for their encouragement and friendship throughout this DPhil. Special thanks to Kevin, Jane, Qingze, Chang, Kumeren, Jiandong, Anders, Alex, Maddy, Duncan, Kaveeta, Sofia and Jacqui for making lunches in the Big Data Institute and dinners outside of the office so much fun. Thanks also to the post-graduate administration team, especially Marie-Alice Frappat and Directors of Graduate Studies Professor Maria Quigley and Professor Mark Sheehan for their support and guidance throughout my studies. Many thanks to Dr Fiona Spensley, my Tutor for Graduates at Lady Margaret Hall. I am thankful for the funding I received from the Oxford Medical Research Council Doctoral Training Partnership and from the Nuffield Department of Population Health and I am grateful for the fantastic research community they provided.

Finally, a heartfelt thank you to my wonderful friends and family for their continuous encouragement. Words cannot suffice to thank Alvar Paris for his love, warmth and support, and for always standing by my side. I cannot imagine completing this DPhil without you. To my sisters, Fleur and Florien, thank you so much for always being there for me, even when we are hundreds of miles apart. This thesis is dedicated to my parents, Otto and Ineke, whom I will never be able to thank enough for their love, for giving me the freedom to pursue my dreams and supporting me every step of the way.

Funding

This DPhil was funded by the Oxford Medical Research Council Doctoral Training Partnership [grant number MR/N013468/1, start date 01/Oct/2020] and a Nuffield Department of Population Health studentship, and I am grateful for their support throughout my studies.

For Chapters 2 and 3, I performed *in vitro* assays with reagents that were funded by the Molecular Epidemiology for Vaccination Policy (MEVacP) project. The MEVacP project is managed as part of the UK Vaccine Network by NIHR and funded by UK Aid from the Department of Health and Social Care. The views expressed in Chapter 2 and Chapter 3 are not necessarily those of the Department of Health and Social Care.

For Chapter 4, I performed the computational analyses using Biomedical Research Computing (BMRC) from the University of Oxford, which is supported by the Wellcome Trust Core Award [grant number 203141/Z/16/Z] and the NIHR Oxford Biomedical Research Centre. The views expressed in Chapter 4 are not necessarily those of the Department of Health and Social Care, NIHR or NHS.

Declaration of contribution

I, Femke Ahlers, conducted all of the analyses presented in this thesis, except where otherwise stated or referenced. I gratefully acknowledge and declare the supervision, guidance and assistance I received in relation to this thesis.

Chapters 2 and 3

The research and/or writing of Chapters 2 and 3 were supervised by David Eyre, Derrick Crook and Angela Brueggemann at the University of Oxford and by Norman Fry and David Litt at the United Kingdom Health Security Agency (UKHSA). Whole-genome sequencing data were generated by others prior to these studies. Genomes were stored in the PubMLST databases using $BIGS_{DB}$ software, which was developed and maintained by Keith Jolley. Melissa Jansen van Rensburg compiled the pneumococcal study dataset prior to this project. Melissa provided guidance when compiling the non-pneumococcal *Streptococcus* (NPS) genome dataset and re-ran her Python scripts to apply genome quality control metrics. *Streptococcus* genome assemblies were compiled from the public species identification database in PubMLST (https://pubmlst.org/species-id, copyright © 2010-2019, University of Oxford), which is maintained and curated by James Bray and Keith Jolley. James Bray curated the ribosomal multilocus sequence typing (rMLST) loci and Angela Brueggemann curated the pneumococcal MLST loci. Madeleine Butler used the NPS genome dataset for her PhD research and thesis [1]. Kaveeta Malhi studied *ply, psaA* and SP2020 *in silico* for her Final Honours Scheme Research Project and thesis at the University of Oxford. I repeated Kaveeta's analyses using a different analysis tool (the ipcress *in silico* PCR tool [2]), using the same pneumococcal study dataset and the newly compiled NPS genome dataset.

For the *in vitro* specificity panel, NPS isolates were cultured and DNA purified by others at UKHSA under the supervision of Juliana Coelho and Bruno Pichon. Pneumococcal isolates included in the *in vitro* sensitivity panel were cultured and DNA was purified by others at UKHSA. Sequencing and assembly of genomes of UKHSA's streptococci was done by others at UKHSA. I performed *in vitro* work under the day-to-day supervision of David Litt.

Chapter 4

The writing of Chapter 4 was supervised by David Eyre and Derrick Crook. The genomes included in the *Streptococcus* dataset were compiled from the publicly available assemblies found in the species identification database in PubMLST (https://pubmlst.org/species-id, copyright © 2010-2019, University of Oxford), which is maintained and curated by James Bray and Keith Jolley and the rMLST loci were curated by James Bray. Kaveeta Malhi studied *rpsB* for her Final Honours Scheme Research Project and thesis at the University of Oxford. I discussed and conceptualised the dataset design and research objectives with David Eyre, Angela Brueggemann and James Bray. Duncan Berger showed me how to run PROKKA and Panaroo. Kevin Yuan helped writing scripts to construct phylogenetic trees. David Eyre and I discussed and interpreted all results.

Planned publications

Some of the results, including some of the figures, tables and text, presented in this thesis have been written up as publications, but have not been submitted to date.

Femke M. Ahlers, David Litt, Melissa J. Jansen van Rensburg, James E. Bray, Keith A. Jolley, Seyi Eletu, Juliana Coelho, Bruno Pichon, Odile B. Harrison, Martin C. J. Maiden, David W. Eyre, Norman K. Fry, Angela B. Brueggemann. Utilising large and diverse bacterial genome datasets to improve the identification of *Streptococcus pneumoniae* via PCR-based diagnostics. *In preparation for submission.* (2024).

Femke M. Ahlers, David Litt, Melissa J. Jansen van Rensburg, James E. Bray, Keith A. Jolley, Juliana Coelho, Bruno Pichon, Odile B. Harrison, Martin C. J. Maiden, David W. Eyre, Norman K. Fry, Angela B. Brueggemann. Exploiting large genomic datasets for the development of a real-time loop-mediated isothermal amplification (LAMP) assay targeting SP2020 for the identification of *Streptococcus pneumoniae. In preparation for submission.* (2024).

Abbreviations

AL	alignment length
AMR	antimicrobial resistance
ANI	average nucleotide identity
B3	reverse outer primer
BeTs	triangles of pairwise best hits
$BIGS_{DB}$	Bacterial Isolate Genome Sequence Database
BIP	reverse inner primer
BLAST	basic local alignment search tool
BLASTN	nucleotide BLAST
CC	clonal complex
CDC	Centers for Disease Control and Prevention
cgMLST	core genome multilocus sequence typing
$CI_{95\%}$	confidence interval
COGs	Clusters of Orthologous Genes
CPS	capsular polysaccharide
C_q-value	quantification cycle
CSF	cerebrospinal fluid
DDH	DNA-DNA hybridisation
ddNTP	dideoxynucleotide
DNA	deoxyribonucleic acid
ENA	European Nucleotide Archive
F3	forward outer primer
FIP	forward inner primer
FP	forward primer
FRET	fluorescence resonance energy transfer
GAS	Group A *Streptococcus*
GBS	Group B *Streptococcus*
GC	guanine-cytosine
GPS	Global Pneumococcal Sequence
HGT	horizontal gene transfer
ID	identity
ipcress	In-silico PCR Experiment Simulation System
IPD	invasive pneumococcal disease
IQR	interquartile range
iTOL	interactive tree of life
LAMP	loop-mediated isothermal amplification
LB	reverse loop primer
LCTN	lowest common taxonomic node
LF	forward loop primer
LMIC	low- and middle-income countries
LOA	limits of agreement
$LOA_{95\%}$	95% limits of agreement

LPSN	List of Prokaryotic Names with Standing in Nomenclature
LRTIs	lower respiratory tract infections
MAFFT	multiple sequence alignment based on fast Fourier transform
MALDI-TOF MS	matrix-assisted laser desorption ionisation time-of-flight mass spectrometry
MCL	Markov Clustering algorithm
MIC	minimum inhibitory concentration
MIQE	Minimum Information for Publication of Quantitative Real-Time PCR Experiments
MLEE	multilocus enzyme electrophoresis
MLST	multilocus sequence typing
MUSCLE	multiple sequence comparison by log expectation
n	number
NAAT	nucleic acid amplification test
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
NPS	non-pneumococcal *Streptococcus*
PCR	polymerase chain reaction
PCV	pneumococcal conjugate vaccine
p-distance	pairwise evolutionary distance
PopPUNK	Population Partitioning Using Nucleotide *k*-mers
PPV	pneumococcal polysaccharide vaccine
QC	quality control
qPCR	quantitative polymerase chain reaction
rMLST	ribosomal multilocus sequence typing
RNA	ribonucleic acid
RP	reverse primer
rpm	revolutions per minute
rRNA	ribosomal ribonucleic acid
rST	ribosomal sequence type
SNP	single nucleotide polymorphism
ST	sequence type
T_m	melting temperature
$UI_{95\%}$	95% uncertainty interval
UK	United Kingdom
UKHSA	United Kingdom Health Security Agency
UPGMA	unweighted pair-group method using arithmetic averages
WGS	whole genome sequencing
WHO	World Health Organisation

Measurement Units

°C	degrees Celsius
bp	base pairs
g	gram
h	hour
Mb	mega base pairs
rpm	revolutions per minute
sec	seconds
µL	microlitre
µM	micromolar

Chapter 1

General Introduction

1.1 The *Streptococcus* genus

1.1.1 Historical perspective

In recorded history, diseases likely to be streptococcal have been observed for millennia, with the Greek physician Hippocrates first describing erysipelas in the 4th century BC [3]. The name *Streptococcus* (from Greek στρεπτός meaning chain and κόκκος meaning berry) was first formulated in 1874 by surgeon Theodor Billroth when describing bacteria isolated from wound infections that resembled a chain of berries [4].

The *Streptococcus* genus is part of the family of *Streptococcaceae*, the order of Lactobacillales, class of Bacilli and phylum Bacillota [5]. Species of the *Streptococcus* genus are Gram-positive, catalase negative, non-motile bacteria, which do not form endospores, with a spherical or ovoid cell shape and a diameter of less than 2 μM [6,7]. Chains of over 50 cells can be produced in liquid cultures [6]. This formation of chains or pairs of streptococci can be explained by the axis of bacterial cell division occurring in one plane (Figure 1.1) [7].

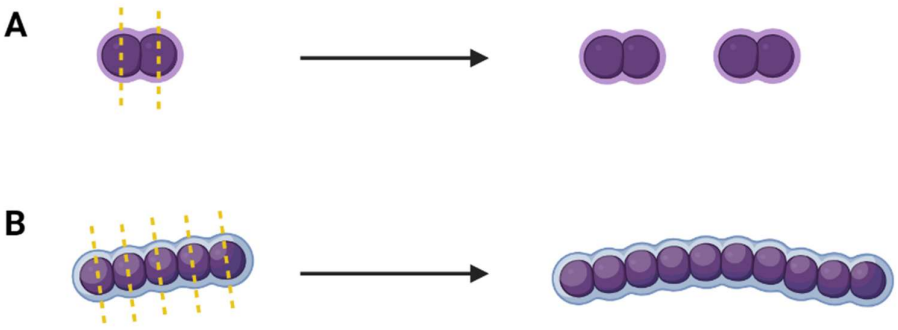


Figure 1.1 Morphology of streptococcal cells [7].
The yellow dashed lines indicate the planes of cell division.
(A) *Streptococcus pneumoniae* divides in a single plane and forms diplococci (pairs). The lancet-shaped cells are encapsulated with a polysaccharide capsule.
(B) *Streptococcus pyogenes*, which has a capsule made of hyaluronic acid, forms a chain of streptococci, due to parallel planes of division.
Adapted from [7], image created with Biorender.com.

1.1.2 Phenotypic and biochemical classifications

In the 20th century, various biochemical and phenotypic methods were developed to classify streptococci [3,7]. In the subsequent sections, streptococcal stratification based on haemolysis types and Lancefield serogrouping will be described. Aside from those two classification systems, various standard biochemical assays have been used to type streptococci, including testing for acidification, fermentation and hydrolysis of various substrates [8]. Furthermore, streptococci have also been differentiated based on their capacity to grow in different broth conditions, by ranging the pH, temperature or salt concentration [8].

1.1.2.1 Haemolysis types

Based on their haemolysis pattern on blood agar plates, streptococci can be differentiated into three different groups: the alpha (α), the beta (β) and the gamma (γ) haemolytic streptococci [8]. The α-haemolytic streptococci perform partial lysis of red blood cells, resulting in a green halo around the colonies, which is the feature that viridans streptococci are named after [8]. The β-haemolytic streptococci have a clear halo around their colony indicative of their complete lysis of red blood cells. The γ-haemolytic streptococci do not perform haemolysis [7]. Whilst typing based on haemolysis provided a useful step forward in stratifying streptococci, haemolysis types can alter to some degree depending on the culture conditions [9,10].

1.1.2.2 Lancefield serogroups

In 1933, Rebecca Lancefield described a strategy to serologically differentiate streptococci based on the polysaccharides and teichoic acid antigens present on the streptococcal cell wall [10,11]. This typing method allows classification into twenty Lancefield serogroups, each one represented by a letter [12]. Lancefield

serogrouping presented a breakthrough in streptococcal stratification [12]; however, many streptococcal species do not have a Lancefield serogroup and some species can have multiple Lancefield serogroups [8]. Of the streptococci that do express Lancefield serogroups, the groups most commonly causing disease in humans are A, B, C and G [10], as will be described in more detail in the next section. Reflecting the enduring significance of these designations, Group A and Group B *Streptococcus* are terms that remain in common use [8].

1.1.3 Streptococcal disease

Many streptococcal species are part of the normal microbial flora of humans and animals; however, some streptococcal species can cause disease, representing important causes of mortality and morbidity worldwide [6].

1.1.3.1 Human disease

Important streptococcal causes of invasive disease in humans are *Streptococcus pyogenes,* which is commonly called Group A *Streptococcus* (GAS); *Streptococcus agalactiae,* which is commonly called Group B *Streptococcus* (GBS); and *Streptococcus pneumoniae* which is commonly called the pneumococcus [6,7,10]. GAS can asymptomatically colonise various sites of the human body in a carriage state, including the oropharynx, skin, rectum and mucous membranes of genitalia [13]. Human diseases caused by GAS range from superficial infections such as pharyngitis to scarlet fever and rheumatic heart disease [13]. GBS can be carried asymptomatically in the vaginal tract, but is also an important cause of infections in neonates, pregnant and elderly people [14]. In particular, neonatal GBS disease includes severe invasive infections such as meningitis [14]. Further to causing human disease, GBS can also cause mastitis in

cows and camels [15]. Pneumococcal diseases are described in depth in section 1.2.

The viridans streptococci are common commensals of the human upper respiratory tract that cause invasive infections relatively rarely, including *Streptococcus infantis, Streptococcus mitis, Streptococcus oralis* and *Streptococcus pseudopneumoniae* [8,9]. However, if viridans streptococci enter normally sterile sites, they can be responsible for severe invasive diseases such as endocarditis [16]. Furthermore, *Streptococcus mutans* is a common cause of dental carries and *Streptococcus anginosus* infections include various types of abscesses [16].

1.1.3.2 Animal disease

Several streptococci can cause disease in animals, including *Streptococcus suis, Streptococcus dysgalactiae, Streptococcus equi, Streptococcus uberis* and *Streptococcus canis* [15]. The natural habitat of *S. suis* is the microbiome of the nasal cavities of swine; however, it is also an important cause of morbidity in piglets, with *S. suis* serotype 2 being particularly virulent [17]. *S. suis* can also cause zoonotic infections in humans, including meningitis or streptococcal toxic shock-like syndrome [17].

S. dysgalactiae includes two subspecies: *S. dysgalactiae* subsp. *dysgalactiae* (which is α-haemolytic and has Lancefield group C) and *S. dysgalactiae* subsp. *equisimilis* (which is β-haemolytic and has Lancefield group G, A, C or L) [12,15]. *S. dysgalactiae* subsp. *dysgalactiae* can cause bovine mastitis, whilst *S. dysgalactiae* subsp. *equisimilis* is a human commensal which can cause diseases such as pharyngitis [15]. *S. equi* can cause strangles in horses [15], and *S. uberis*

can infect the bovine udder resulting in mastitis [18]. *S. canis* is a commensal of the microbiome of dogs and cats and can also opportunistically cause disease [15]. Table 1.1 gives an overview of streptococcal species causing disease in humans and animals [7,8,10,12,15,19].

Table 1.1 Clinically relevant *Streptococcus* species in humans and animals. Adapted from [7,8,10,12,15,19].

***Streptococcus* species/type**	**Disease**	**Host**	**Lancefield serogroup**	**Haemolysis group**
S. pneumoniae	Pneumonia, meningitis, bloodstream infections, sinusitis, otitis media	Human	None	α
Viridans streptococci	Endocarditis, caries, gum disease, abscesses	Human	None	α
S. pyogenes	Scarlet fever, pharyngitis, skin and soft tissue infections, toxic shock syndrome, bloodstream infections, acute rheumatic fever, rheumatic heart disease	Human	A	β
S. agalactiae	Bloodstream infections, pneumonia and meningitis in neonates, infections in adults, bovine mastitis	Human, cattle	B	β
S. dysgalactiae	Septic arthritis in lambs, bovine mastitis, skin infections, pharyngitis	Animal, human	C (G, A, L)	β/α
S. equi	Strangles in horses	Horses, human	C	β
S. canis	Urinary tract infections, bloodstream infections, pneumonia, bone or soft tissue infections	Dog, cat, human	G	β
S. iniae	Bloodstream infections and meningoencephalitis in fish	Fish	None	β
S. suis	Meningitis, bloodstream infections and arthritis in pigs, meningitis and toxic shock-like syndrome in humans	Swine, human	R, S, T	α/γ
S. uberis	Bovine mastitis	Cattle	None	α/γ

Note: *S.*, *Streptococcus*.

1.2 *Streptococcus pneumoniae*

1.2.1 Historical perspective

The ancestor of *S. pneumoniae* and its closely related mitis group likely colonised the nasopharynx of hominoids six to seven million years ago [20,21]. In recorded history, the pneumococcus was first identified and described in 1881 independently by Louis Pasteur and George Sternberg following the injection of rabbits with human saliva, which resulted in pneumococcal septicaemia [22]. Since then, pneumococci have been central to multiple landmark discoveries in science, including being one of the first infections to show the efficacy of treatment with penicillin and being used in the experiments by Oswald Avery, Colin Macleod and Maclyn McCarty that demonstrated DNA is the source of genetic information in 1944 [22,23].

1.2.2 Epidemiology of pneumococcal carriage and disease

S. pneumoniae is an encapsulated, extracellular bacterium and opportunistic human pathogen [24,25]. *S. pneumoniae* typically resides in the human upper respiratory tract, particularly in the nasopharynx [24,26]. Here, pneumococci asymptomatically colonise the mucous layer lining the nasopharyngeal membranes in what is often referred to as a “carriage” state [24,26]. Whilst this carriage state is common and benign, it is also the precursor of pneumococcal disease and facilitates person-to-person transmission via respiratory droplets (Figure 1.2A) [26]. Diseases caused by *S. pneumoniae* range from milder infections including otitis media and sinusitis to severe invasive infections such as meningitis, community-acquired pneumonia and bloodstream infections (Figure 1.2B) [25].

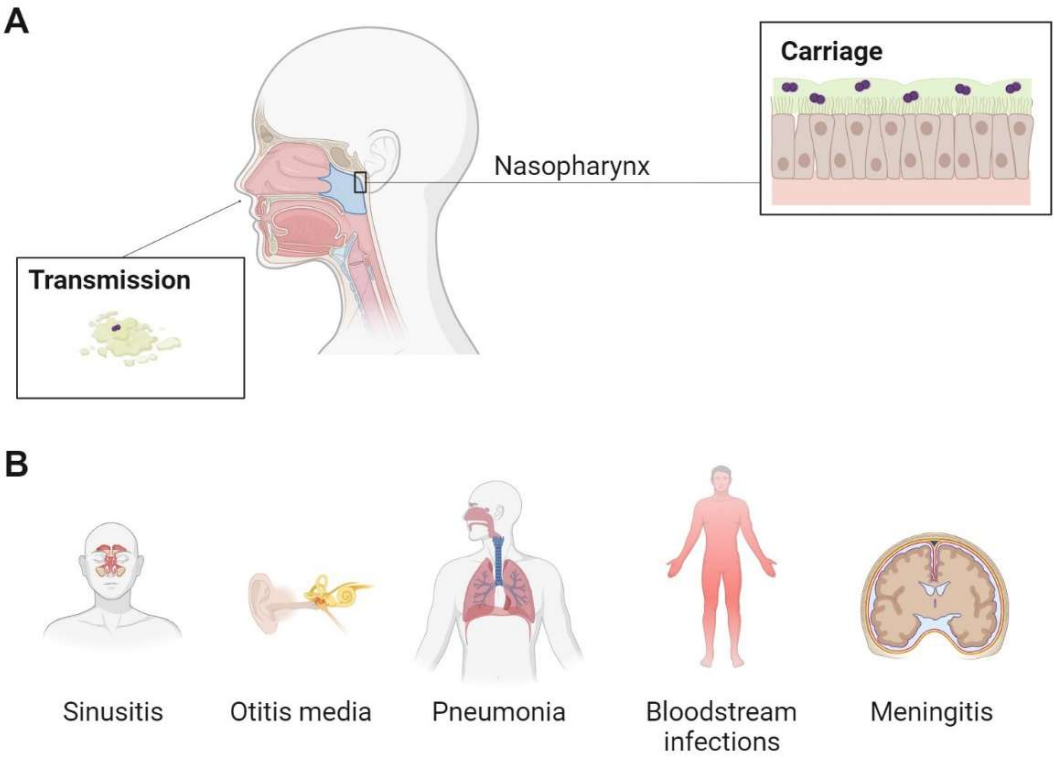


Figure 1.2 *Streptococcus pneumoniae* carriage and disease [24].
(A) *S. pneumoniae* can asymptomatically colonise the mucosal lining of the nasopharynx facilitating transmission.
(B) Pneumococcal diseases include sinusitis, otitis media, pneumonia, bloodstream infections and meningitis.
Figure adapted from [24] and created with Biorender.com.

1.2.2.1 Pneumococcal carriage

Pneumococcal nasopharyngeal carriage rates differ depending on various factors, including the age of the host and geographic location, with carriage rates correlating inversely with the country's income status [27]. Children less than five years old in sub-Saharan Africa have particularly high pneumococcal carriage rates of 63.2% [95% confidence interval ($CI_{95\%}$) 55.6-70.8] [28]. Increased carriage rates are associated with day-care centres and having young siblings [29]. Further, high carriage rates have been reported in orphanages and among indigenous populations [30]. In an Australian Aboriginal population, pneumococcal

carriage rates of 67% [$CI_{95\%}$ 62-72] were found in children and 26% [$CI_{95\%}$ 22-30] in adults [31]. Whilst children are the main reservoir for pneumococcal carriage, adult carriage also plays a role in transmission [32].

Pneumococcal transmission is affected by seasonal climate dynamics, with higher transmission rates recorded during drier and cooler months [33,34]. Co-colonisation with other pathogens, exposure to smoke and viral respiratory tract infections are also positively associated with higher carriage rates [27]. Overall, the asymptomatic pneumococcal carriage state provides benefits to pneumococci and (to some extent) also to its human host: the nasopharynx provides an opportunity for pneumococci to exchange DNA, but it also allows humans to acquire B-cell mediated immunity against the colonising pneumococcal serotype [30].

1.2.2.2 Pneumococcal disease

Pneumococci residing asymptomatically in the nasopharynx can spread to other organs, such as the sinuses and ears, where they can cause local infections [25]. Furthermore, aspiration of pneumococci can cause lower respiratory tract infections (LRTIs), which can be followed by seeding into the bloodstream and/or crossing of pneumococci into the blood-brain barrier in severe cases [25]. Pneumococcal pneumonia, meningitis and bloodstream infections are commonly referred to as invasive pneumococcal disease (IPD) and can be life-threatening [30].

IPD has several risk factors that closely mirror risk factors for pneumococcal carriage; rates of IPD are highest at the extreme ages (i.e. in young children and elderly adults), in low- and middle- income countries (LMICs) and in

immunocompromised people [25]. Risk factors of IPD include being unvaccinated, smoking and chronic cardiac or respiratory disease [30,35]. Further risk factors for pneumococcal LRTIs include malnutrition and exposure to air pollution [36].

Pneumococcal pneumonia is a major cause of death worldwide, particularly in children <5 years old in India, Pakistan, Nigeria and the Democratic Republic of the Congo [37]. In 2015, pneumococci were the leading aetiology of deaths caused by LRTIs in children <5 years old, with 393,000 deaths globally [95% uncertainty interval ($UI_{95\%}$) 228,000-532,000] [38]. Similarly, across all ages, *S. pneumoniae* was the leading aetiology of deaths caused by LRTIs in 2016, with 1,189,937 deaths globally [$UI_{95\%}$ 690,445-1,770,660] [39]. In 2019, *S. pneumoniae* was predicted to be the deadliest single aetiological agent of meningitis worldwide, responsible for an estimated 18.1% [$UI_{95\%}$ 17.1-19.2] of a total predicted 236,000 deaths [$UI_{95\%}$ 204,000-277,000] [40]. The incidence of IPD was reduced during the COVID-19 pandemic due to non-pharmaceutical containment measures lessening transmission [41,42].

1.2.3 Pneumococcal serotypes

Historically, the pneumococcal capsular polysaccharide (CPS) was pivotal to the discovery of bacterial transformation by Frederick Griffith in 1928 [25,43]. In his landmark experiments, mice died after being co-infected with heat-killed encapsulated pneumococci and live non-virulent non-encapsulated pneumococci, due to the acquisition of the *cps* locus by the live pneumococci [22,25,43]. The pneumococcal CPS is important for evasion of the host immune response, particularly through limiting opsonisation by the complement system and phagocytosis [44].

The *S. pneumoniae* CPS surrounds the bacterial cell wall of most pneumococci and determines the pneumococcal serotype [45,46]. The CPS is a major pneumococcal virulence factor and the proteins that facilitate the biosynthesis and export of the CPS are encoded by the capsular polysaccharide synthesis (*cps*) locus [45,46]. In 2020, Ganaie et al. documented the 100th pneumococcal serotype [47]. Each pneumococcal serotype evokes a distinct host immune response [45,46]. In addition, pneumococcal serotypes are associated with their potential for invasiveness, i.e. the capacity to cause IPD [48].

Various phenotypic and genotypic methods can be used to determine pneumococcal serotypes [45]. The gold standard serotyping method remains the Quellung reaction, but other methods such as latex agglutination or nucleic acid amplification tests (NAATs) can also be used [45]. Furthermore, bioinformatics approaches have been developed to predict the pneumococcal serotype, including PneumoCaT and SeroBA [49,50]. Presence of the *cps* genes is typically considered limited to *S. pneumoniae*; however, in Chapter 4 of this thesis, the presence of the pneumococcal *cps* genes in non-pneumococcal *Streptococcus* genomes will be investigated.

1.2.4 Pneumococcal vaccines

All currently licensed pneumococcal vaccines target the pneumococcal CPS surface antigens [51]. Two types of pneumococcal vaccines can be differentiated: the pneumococcal polysaccharide vaccine (PPV) and the pneumococcal conjugate vaccines (PCV) [51]. In the PCV vaccines, the pneumococcal CPS antigen is conjugated to the diphtheria CRM_{197} protein, eliciting a larger immune response [52]. The introduction of the PCVs resulted in a decrease in both IPD and pneumococcal carriage by the serotypes included in the vaccine, but an

increase in the serotypes not included in the vaccine [53,54]. This vaccine escape is facilitated by pneumococcal capsular switching and by replacement of strains [53–55].

1.2.5 Horizontal gene transfer

In bacteria, three main methods of horizontal gene transfer (HGT) can be differentiated: conjugation, transduction and transformation [56]. In conjugation, plasmids are exchanged in bacterial cell-to-cell contact, facilitated by a pilus [57]. In transduction, a bacteriophage (a virus that infects bacteria) inserts its nucleic acids into the bacterium [57]. Transformation involves the uptake of DNA from the environment by the bacterium followed by integration of the acquired DNA into the bacterial chromosome [58]. Of these three canonical mechanisms of HGT, *S. pneumoniae* mainly utilises transformation to acquire genetic diversity [59].

The pneumococcus is a highly recombinogenic and naturally competent bacterium [57]. *S. pneumoniae* gains more variation through genetic recombination than through point mutations [57]. Across *S. pneumoniae*, the genes encoding penicillin-binding proteins and the *cps* locus exhibit particularly high rates of recombination [60]. Furthermore, *S. pneumoniae* and closely related mitis group streptococci such as *S. oralis* and *S. mitis* frequently exchange DNA through homologous recombination [20,61]. In addition, transduction occurs in *S. pneumoniae* and a diverse range of prophages have been reported, many of which can also be found in non-pneumococcal streptococci [62,63].

1.2.6 Pneumococcal diagnostics

Various pneumococcal diagnostic methods exist, which can be applied to both carriage and disease isolates [64]. Methods include culturing to study haemolysis,

looking for inhibition by optochin disks on blood agar plates or testing for bile solubility, mass spectrometry, antibody detection tests, antigen detection tests and NAATs [64]. This thesis focusses on the improvement of pneumococcal NAATs. In particular, Chapter 2 will investigate quantitative polymerase chain reaction (qPCR) diagnostics and Chapter 3 will study loop-mediated isothermal amplification (LAMP) assays to detect *S. pneumoniae*. The introductions of these results chapters will outline other diagnostic methods used for *S. pneumoniae*.

1.3 Sequencing bacterial genomes

1.3.1 Historical perspective

In 1977, Sanger et al. described the dideoxynucleotide (ddNTP) chain termination methodology to determine the nucleotide sequence of DNA, a method also called first-generation sequencing or Sanger sequencing [65]. In short, ddNTP terminators are radiolabelled and their incorporation halts synthesis resulting in different lengths of products [66]. Gel electrophoresis allows size separation of these strands, which is performed for each of the four bases (A, G, T or C) in parallel [66]. Improving this method, the addition of different fluorescent labels to the different ddNTPs facilitates automated optical detection of the nucleotides together in one reaction and results in faster readings [66,67]. These methods allowed sequencing of short DNA fragments, which were subsequently assembled *in silico* using overlapping regions into contiguous sequences (known as contigs) through a shotgun approach [66].

Bacterial whole-genome sequencing (WGS) is now nearly thirty years old, following the first complete bacterial genome published in 1995 – the 1.8 Mb chromosome of *Haemophilus influenzae* – created using Sanger sequencing and a shotgun approach [68]. In 2001, Tettelin et al. published the first complete genome of *S. pneumoniae*, which had serotype 4 and is commonly referred to as TIGR4, and has become a major pneumococcal reference genome [69–71]. Since then, the number of bacterial genome sequences generated has increased dramatically; Blackwell et al. extracted over 660,000 bacterial genomes of over 2,300 bacterial species from the European Nucleotide Archive (ENA) in November 2018 – although species frequency was heavily biased with over 90% of genomes belonging to only 20 bacterial species [72]. For *S. pneumoniae,* a collection of

over 30,000 assembled genomes is available on PubMLST [73,74]. This steep increase in bacterial genomes was primarily facilitated by next-generation sequencing (NGS) methods [72,75].

1.3.2 High-throughput next-generation sequencing methods

Several NGS methods have been developed, allowing higher-throughput and more affordable sequencing [76]. NGS methods started with pyrosequencing, often referred to as 454 sequencing (Roche) [77]. However, the steep increase in bacterial WGS data was led by Solexa sequencing (Illumina) technology, which is based on cyclic reversible termination [78]. NGS approaches contain four main steps (preparation of templates, the sequencing itself, imaging and analysis of the data) and the methods of the each steps can differ between the various NGS methods [76].

1.3.3 Single-molecule sequencing

Long-read or single-molecule sequencing methods were first developed in 2007 by Pacific Biosciences (PacBio), providing single-molecule and real-time data generation [79]. This method of sequencing by synthesis allows creation of very long reads, though error rates are high [80]. Another common example of single-molecule sequencing is Oxford Nanopore sequencing [81]. These long-read methods can produce reads over 10,000 bp-long [82]. A bacterial genome can be sequenced within hours with third-generation sequencing [83].

Prior to analysis, attempted reconstruction of a complete genome from short or long reads is undertaken [76,80]. This can be done by assembly, i.e. exploiting overlaps between reads to form contiguous sequences [76,80]. The alternative approach uses a reference genome and maps each read to it [76,80].

1.3.4 Genome quality control metrics

When compiling bacterial genome datasets, applying quality control metrics, for example to ascertain assembly quality, is highly important. Namely, when conducting phylogenetic analyses of bacterial genomes, the single-nucleotide changes that are inherited clonally should underpin analyses [84]. As discussed in section 1.2.5, horizontal gene transfer events in bacteria can interfere with these analyses; hence, bioinformatics tools have been designed to account for these recombination events, as will be described in more detail in section 1.4.5 [85–87]. In addition, errors in sequencing, mapping or assembly can also contribute to issues in downstream analyses [84].

In this thesis, various genome assembly quality control metrics were applied to attempt to ascertain high-quality data was used for analyses. The N50, which is a metric of contiguity, represents the length of the contig at which all the contigs with the same or smaller length make up half of the total assembly length of the genome [88]. Since using the N50 alone to assess genome quality is not reliable [88], other metrics were added to assess genome quality, including the requirement of a low number of contigs. Furthermore, gene-by-gene typing schemes (described in section 1.4.4 below) can be used to assess genome quality. For example, assignment of an allele at every locus of a typing scheme can be used as a metric of genome completeness, whilst assignment of multiple alleles at a single locus could indicate genome contamination [73,89].

1.3.5 Repurposing the wealth of WGS data

The large advances in WGS technology have resulted in a steep increase in the number of genome sequences over the past decades, as exemplified by the genomic data available from GenBank (Figure 1.3) [90,91]. Bioinformatics tools

have been developed to analyse these WGS data; specific approaches have been designed to understand bacterial genomic epidemiology, accounting for recombination [84]. In addition, WGS data can be repurposed in infectious disease research, to better understand pathogen transmission and to inform the design of new diagnostics, antimicrobials and vaccines [92]. The following section will describe applications of bacterial WGS data.

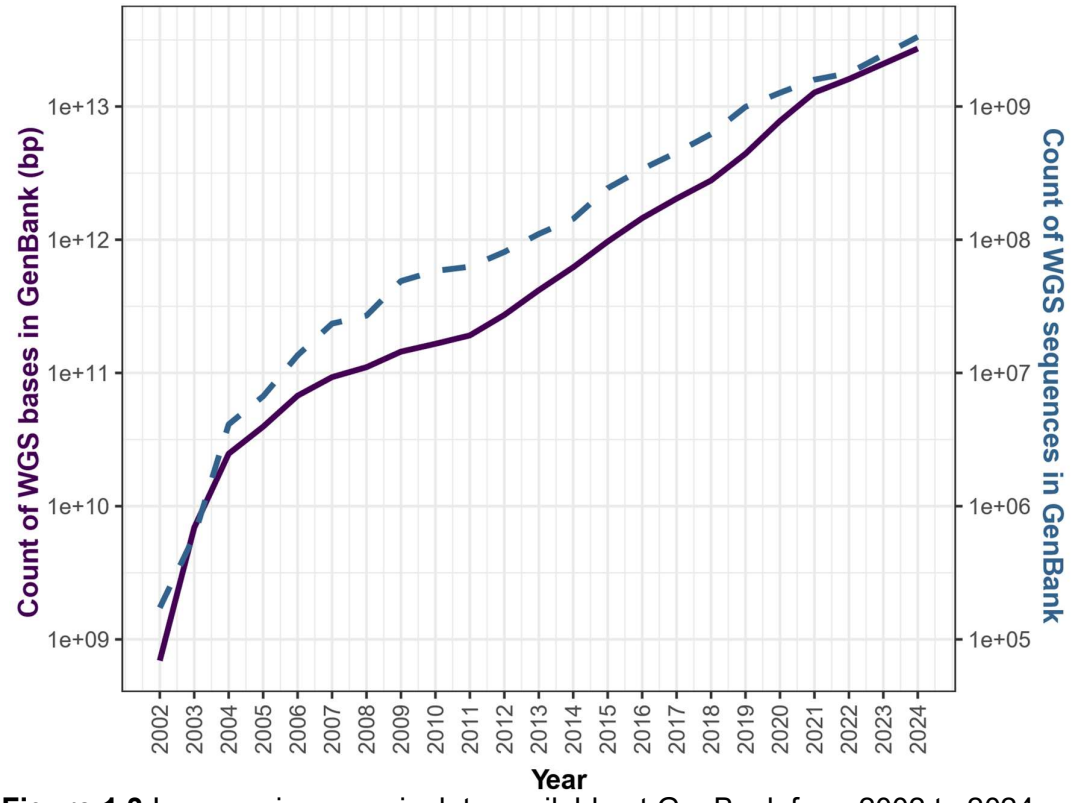


Figure 1.3 Increase in genomic data available at GenBank from 2002 to 2024. The number of nucleotide bases (purple line, left-hand y-axis) and whole-genome sequences (blue dashed line, right-hand y-axis) increased steeply over time [91]. Note the logarithmic scale of both y-axes. Data were accessed in May 2024 from the National Center for Biotechnology Information, using the counts recorded for April of every year (https://www.ncbi.nlm.nih.gov/genbank/statistics/) [91].
Note: bp, base pairs; WGS, whole-genome sequences.

1.4 Bacterial genotyping

1.4.1 Historical perspective

The increased availability of WGS data facilitates high-resolution investigations into the phylogenetic relationships of bacterial isolates [93]. Historically, the term phylogeny was coined in 1866 (prior to genome sequencing) by Ernst Haeckel who also created the first Tree of Life containing microorganisms [94,95]. In 1859, Charles Darwin had already included an evolutionary tree describing relationships between species as the only figure in his world-famous book *On the Origin of Species* [96,97]. In present times, phylogenies have become a common way to describe the evolutionary relatedness of pathogens in infectious disease research and are often constructed using genomic data [98].

Bacterial phylogenetic reconstruction methods provide an important tool to better understand pathogens; however, to fully understand the epidemiology of infections, information on not only the pathogen, but also of the host and the environment should be integrated [99]. Nevertheless, in this thesis, bacterial genomes on their own were utilised to improve diagnostic assays and to study streptococcal genus-wide phylogenetic clustering. Therefore, the WGS applications, genotyping methods and phylogenetic reconstruction approaches described below will focus solely on bacterial genomes.

1.4.2 WGS-informed diagnostics

The decreasing cost and improving speed of bacterial WGS make it an increasingly feasible method for routine identification of pathogens in clinical and research laboratories [90,92,100]. Using bacterial WGS as a diagnostic method has several benefits over traditional microbiological and phenotypic diagnostic methods, including the wealth of information that can be extracted from the whole

genome, e.g. predicting AMR and virulence determinants [99]. Furthermore, bacterial WGS methodology requires a single diagnostic modality independent of the species [99].

Bacterial WGS data can also be applied to assess and improve other diagnostic assays, e.g. to find the most accurate diagnostic targets of NAATs [100,101]. The importance of such analyses is exemplified by the PCR-based assays that were utilised for *Chlamydia trachomatis* targeting a single cryptic plasmid-based target [102]. A 377 bp-long deletion in the diagnostic target resulted in false-negative hits, which in turn meant infections were left untreated and could spread [103]. The risk of such diagnostic escape can be mitigated by targeting a highly conserved gene or by targeting multiple genes in a multiplex NAAT [102]. In this thesis, large and diverse bacterial genome datasets were exploited to assess the *in silico* performance of PCR- and LAMP-based assays and to find the most accurate diagnostic targets for the identification of *S. pneumoniae*.

1.4.3 WGS-informed vaccines

Bacterial WGS data can be used to investigate vaccine candidates in a method often referred to as reverse vaccinology [100]. This strategy uses bioinformatics approaches to look for ubiquitously present secreted or surface-exposed vaccine epitopes and was first applied to the development of a vaccine against *Neisseria meningitidis* serogroup B (MenB) [104]. Reverse vaccinology approaches have also been utilised to search for pneumococcal vaccine candidates [105]. Furthermore, bacterial WGS provides a valuable means of vaccine surveillance; currently licensed *S. pneumoniae* vaccines target a subset of the highly variable serotypes, and monitoring of vaccine evasion and pneumococcal evolution is crucial [55].

1.4.4 Bacterial genomic epidemiology

In bacterial genomic epidemiology, datasets of multiple bacterial isolates are used to cluster together more similar genomes and discern a population structure, furthering understanding of transmission and evolution [84,92]. Different genotyping methods can be distinguished, such as reference-based single nucleotide polymorphism (SNP)-based methods, *k*-mer-based methods and gene-by-gene typing schemes [84]. These methods each have their own advantages and disadvantages [84]. Overall, when studying bacterial comparative genomics, it is crucial that genotyping methods take into account HGT events, as will be discussed further in section 1.4.5 [84,93].

1.4.4.1 SNP-based methods

SNP-based methods capitalise on the full genome sequence and allow high-resolution insight into genetic variation [106]. Two types of SNP calling methods can be differentiated [106]. First, reference-based SNP calling relies on mapping reads to a reference genome to identify nucleotide differences [93]. This approach requires a high-quality reference genome, ideally closely related to the genomes to which it is compared [93]. Reference-based SNP calling allows high-resolution typing with identification of down to single-nucleotide variants and can be used to investigate epidemiological sources and transmission links of infections, as was done for *Clostridioides difficile* [107]. Second, reference-agnostic SNP calling does not require a reference strain, but rather utilises *k*-mer comparisons and is often called a "*k*-mer-based" approach, which will be discussed in the section below [93,106].

1.4.4.2 k-mer-based methods

Using *k*-mer-based bacterial genotyping methods overcomes the requirement for a reference genome; instead, this approach uses pairwise comparisons of *k*-mers, which are chains of nucleotides with length *k* [106]. A further advantage of *k*-mer-based methods is that they can investigate the variation throughout the genome rather than investigating only a set of loci like the gene-by-gene typing methods that will be described in section 1.4.4.3 [93]. Tools such as kSNP3.0 can be utilised to detect SNPs and analyse phylogenetic relatedness based on *k*-mers on raw read files, circumventing both the need for a reference genome and for assembly of the isolates of interest [108].

For *S. pneumoniae,* a commonly adopted *k*-mer-based genotyping method is Population Partitioning Using Nucleotide *k*-mers (PopPUNK) [109]. The pneumococcal lineages assigned using PopPUNK are commonly referred to as Global Pneumococcal Sequence Clusters (GPSCs) [110]. GPSCs have been used to study the geographical differences of *S. pneumoniae*, and to characterise the presence of antibiotic resistance across pneumococcal lineages [110,111].

1.4.4.3 Gene-by-gene typing methods

Gene-by-gene typing methods such as multilocus sequence typing (MLST) schemes contain a constant number of conserved loci which are assigned alleles [73,74,89,112]. The MLST allele designations can be stored in reference databases, such as the PubMLST platform (https://pubmlst.org/) [73,74,89,112]. Alternatively, alleles can be stored as a string of letters (also called a hash) independent of a centralised database, as demonstrated by the hash-based core genome MLST (hash-cgMLST) scheme for *C. difficile* [113].

In this thesis, the ribosomal MLST (rMLST) scheme and pneumococcal MLST scheme were used, which are described in more detail below in sections 1.4.4.3.1 and 1.4.4.3.2, respectively [73,89,112,114]. The pneumococcal MLST scheme was designed to investigate the population structure across *S. pneumoniae* isolates [73,112,114]. In contrast, the rMLST scheme can be used throughout prokaryotes and can be applied to differentiate bacterial species [73,89,112].

Building on MLST and rMLST schemes, the increased availability of WGS data has inspired the development of core genome MLST (cgMLST) and whole genome MLST (wgMLST) schemes, increasing resolution by including a larger number of loci compared to the seven-locus MLST scheme [106,112]. A pneumococcal cgMLST scheme containing 1,222 loci has been developed [74]. Furthermore, within the PubMLST platform, a life identification number barcoding system is implemented to allow rapid stratification of the cgMLST lineages [74].

Recently, the performance of various pneumococcal genotyping methods was compared, including typing using the pneumococcal MLST and cgMLST schemes and *k*-mer-based GPSC assignment using PopPUNK [74,115]. Whilst the methods compared were broadly consistent, King et al. (2024) suggest moving away from the pneumococcal MLST scheme towards higher resolution clustering methods such as cgMLST or GPSC typing [74,115]. These latter two methods had high concordance and the differences present could be attributed to the accessory genome, which is included in GPSC typing, but excluded from the cgMLST scheme [74,115].

1.4.4.3.1 Pneumococcal MLST

In 1998, Maiden et al. proposed MLST as an electronically portable alternative to multilocus enzyme electrophoresis (MLEE) for the characterisation of microorganisms and designed the first MLST scheme for *N. meningitidis* [116]. MLST schemes provided benefits over other typing methods that were used at the time: compared to pulsed-field gel electrophoresis and MLEE, MLST has increased transportability and reproducibility and provides higher resolution clustering than 16S rRNA [112,116]. Following the *N. meningitidis* MLST scheme, Enright and Spratt designed an MLST scheme for *S. pneumoniae*, targeting approximately 450 bp-long fragments within the seven housekeeping genes *aroE, ddl, gdh, gki, recP ,spi* and *xpt* [114]. Figure 1.4 illustrates the location of these seven targets across the chromosome of the pneumococcal reference genome TIGR4 [69,73,114].

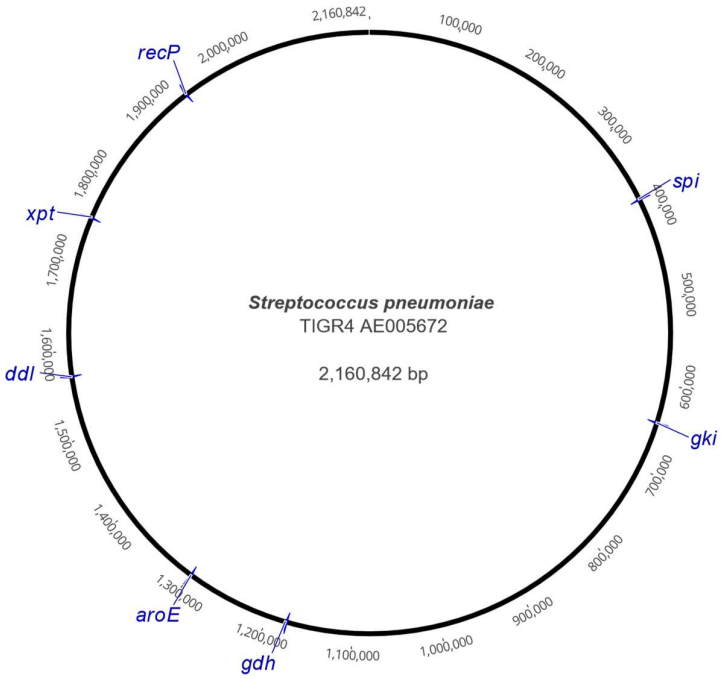


Figure 1.4 Position of seven pneumococcal multilocus sequence typing (MLST) genes on the pneumococcal reference genome TIGR4 [69,73,114].
This figure was created using Geneious Prime Software (v2022.1.1) [117].

The allelic data of MLST schemes can be curated and stored using the PubMLST platform, which is supported by $BIGS_{DB}$ software [73]. PubMLST databases are currently available for 139 organisms, including *S. pneumoniae* (https://pubmlst.org/organisms, accessed May 2024) [73]. For the pneumococcal MLST scheme, hundreds of alleles have been identified and curated for each of the loci (https://pubmlst.org/organisms/streptococcus-pneumoniae, Table 1.2) [69,70,73,114].

Table 1.2 Pneumococcal multilocus sequence typing (MLST) scheme. Adapted from [69,70,73,114].

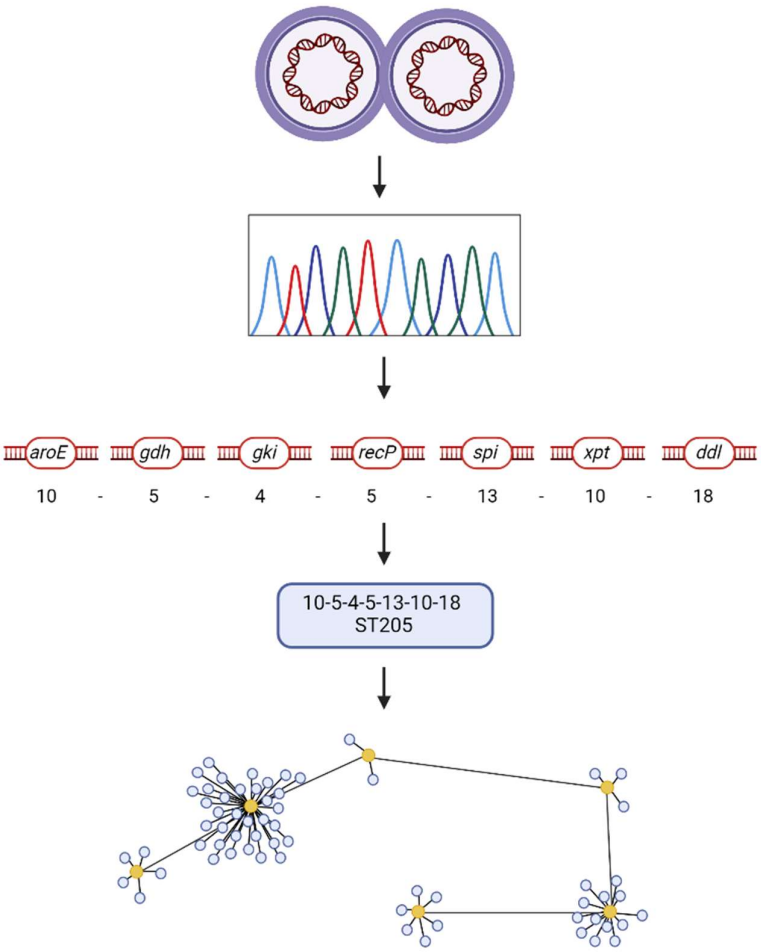
Gene name	*Streptococcus pneumoniae* housekeeping genes[a]		Gene fragment within gene targeted by MLST analysis[b]		Full name/product
	Locus tag	Length (bp)	Number of alleles[b]	Length (bp)[b]	
aroE	SP_1376	855	658	405	Shikimate dehydrogenase
ddl	SP_1671	1,044	1,227	441	D-alanine-D-alanine ligase
gdh	SP_1243	1,488	849	460	Glucose-6-phosphate dehydrogenase
gki	SP_0668	963	872	483	Glucose kinase
recP	SP_2030	1,977	606	450	Transketolase
spi	SP_0402	615	820	474	Signal peptidase I
xpt	SP_1847	582	1,158	369-624	Xanthine phosphoribosyltransferase

[a]The locus tag and length shown are from pneumococcal reference genome TIGR4 [69].
[b]The public *Streptococcus pneumoniae* database on PubMLST was accessed March 20th 2024 (https://pubmlst.org/organisms/streptococcus-pneumoniae) [73,114,118].

MLST compares the sequences at an allelic level rather than a nucleotide level, meaning that HGT events and point mutations are counted as the same change [112]. Such a gene-by-gene allelic typing method is useful when analysing highly recombinogenic pneumococci, as demonstrated by the housekeeping gene *ddl*, which is part of the pneumococcal MLST scheme and has a recombination site [119]. Alternatively, the nucleotide sequences of MLST schemes can be used for nucleotide-based phylogenetic reconstruction, which provides higher resolution insight into the genetic relatedness of isolates than allele-based phylogenetic reconstruction, as will be described in more detail in section 1.4.5 [85].

The *S. pneumoniae* MLST scheme has been used to investigate the contribution of point mutations versus recombination events to acquire genetic diversity [120]. The pneumococcal MLST alleles are ten times more often altered by recombination than by point mutations [120]. Within the pneumococcal MLST loci, at the individual nucleotide level, recombination events contributed to site changes 50 times more often than point mutation events [120].

Traditionally, the MLST alleles were obtained through PCR amplification and Sanger sequencing of the regions of interest; however, the seven housekeeping loci can also be extracted from bacterial WGS data [112]. The allelic profile of the seven fragments determines the sequence type (ST) [114]. STs can be clustered together using a BURST algorithm, which was first replaced by eBURST to allow handling larger datasets, and then by global optimal eBURST (goeBURST) [121,122]. Phyloviz can be utilised to visualise the network of clonal complexes [123,124]. Figure 1.5 displays a schematic overview of the pneumococcal MLST workflow [70,73,112,114].



Pneumococcal DNA

Sequencing of MLST loci
Whilst whole-genome assemblies are increasingly used to extract the selected loci, traditionally MLST sequences were obtained through PCR amplification followed by Sanger sequencing.

Assigning MLST alleles
The seven housekeeping loci are each assigned an allele [114].

Sequence types (STs)
The unique combination of the seven MLST alleles is used to determine the pneumococcal sequence type (ST).

Clonal complexes
The goeBURST algorithm groups STs together into clonal complexes (CCs). The name of the CC is derived from ST of the group founder. The network of clonal complexes can be visualised in Phyloviz and used to understand bacterial population structure.

Figure 1.5 Schematic overview of pneumococcal multilocus sequence typing. In this figure, the sequence type of pneumococcal reference genome TIGR4 is shown [70,73,114]. Figure adapted from [112] and [125], created with Biorender.com.
Note: CC, clonal complex; MLST, multilocus sequence type; ST, sequence type.

1.4.4.3.2 Ribosomal MLST

Jolley et al. (2012) developed a ribosomal MLST (rMLST) scheme, targeting 53 bacterial ribosome protein subunit (*rps*) genes [89]. Since the *rps* genes are required for translation, they are universally present in bacteria and under stabilising evolutionary selection pressure [89]. Figure 1.6 shows the distribution of rMLST loci across the pneumococcal reference genome TIGR4 [69,73,89].

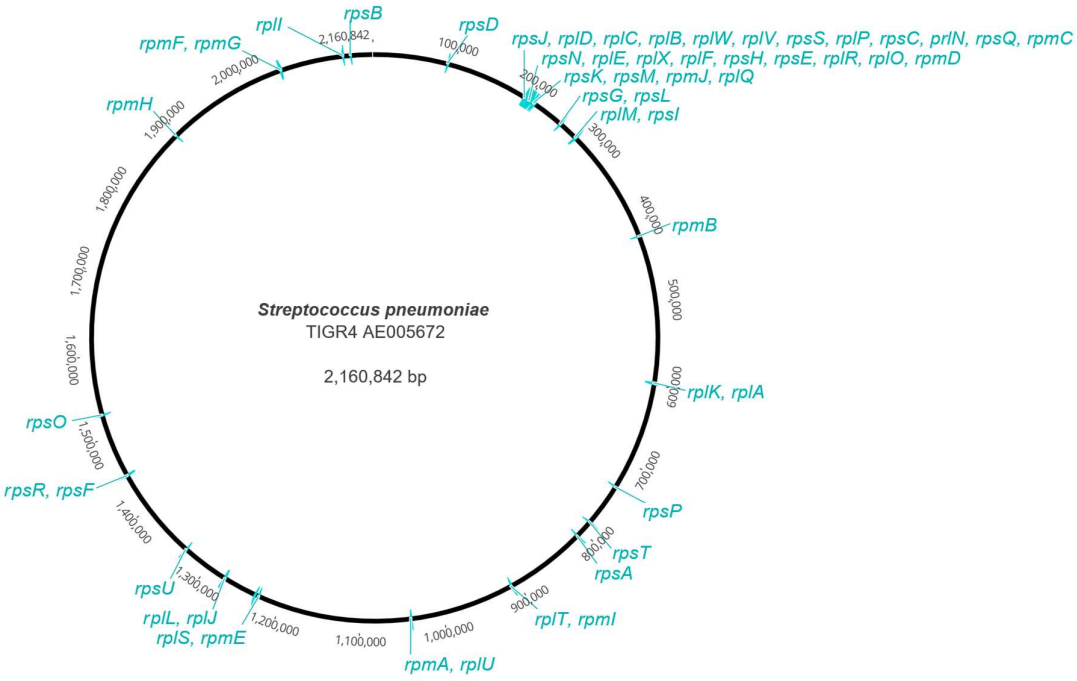


Figure 1.6 Position of 53 pneumococcal ribosomal multilocus sequence typing (rMLST) loci on *Streptococcus pneumoniae* genome TIGR4 [69,70,73,89]. The TIGR4 genome has three copies of the BACT000062 gene, but only one is shown here. This figure was created using Geneious Prime Software (v2022.1.1) [117].

Whilst the rMLST scheme is in the first instance used as a bacterial genotyping and speciation tool, the scheme can also be applied as a metric of genome quality control [73,89]. The absence of one or more rMLST loci can be an indication of low genome coverage and/or missing data [73,89]. In addition, the presence of multiple alleles at non-paralogous loci is an indication of genome contamination

[73,89]. However, in some genera paralogous rMLST loci are common: within the *Streptococcus* genus, ribosomal genes *rpsN* (BACT00014) and *rpmG* (BACT00062) are known to often have multiple copies per isolate [73,89]. In the PubMLST platform, the rMLST genome database houses thousands of bacterial genomes and the typing database stores the allelic profiles (https://pubmlst.org/species-id) [73,89].

1.4.5 Phylogenetic reconstruction methods

Two main types of methods to construct phylogenetic trees using genetic data can be differentiated: distance-based methods and sequence-based methods (the latter are sometimes referred to as character-based methods) [98]. Distance-based methods encompass the unweighted pair-group method using arithmetic averages (UPGMA) and neighbour-joining methods. Sequence-based methods include maximum parsimony, maximum likelihood and Bayesian methods (Table 1.3) [98,126,127].

When creating a neighbour-joining tree based on a gene-by-gene typing scheme such as MLST or rMLST, the allele designations, i.e. the number of allele differences, rather than the underlying nucleotide sequences are used to infer genealogy [85,116]. Whilst this method accounts to some extent for recombination as changes across a whole gene are just counted once, it also means that two alleles that differ by one nucleotide are seen as equally different as two alleles differing in many positions [85]. Sequence-based methods allow higher resolution investigations of phylogenies as more data is used [98]. When using sequence-based methods, bioinformatics tools such as ClonalFrameML or Gubbins can be applied to construct trees taking into account recombination events [85–87].

Table 1.3 Phylogenetic reconstruction methods
Adapted from [98,126,127].

Method type	Method name	Tree type
Distance-based methods	Unweighted pair-group method using arithmetic averages (UPGMA)	Ultrametric tree (molecular clock hypothesis)
	Neighbour-joining (NJ)	Additive tree (minimum evolution principle)
	Least-squares method	Additive tree
Sequence-based methods	Maximum parsimony	Additive tree (minimum evolution principle)
	Maximum likelihood	Additive tree
	Bayesian trees	Additive tree

1.5 Thesis structure and aims

The overarching aim of this thesis was to repurpose large and diverse bacterial genome datasets to improve NAATs to identify *S. pneumoniae* and to better understand differences across the *Streptococcus* genus. In Chapter 2, I used *in silico* analyses to optimise pneumococcal PCR-based diagnostics and validated the new assays *in vitro.* In Chapter 3, I modified this approach to create a new LAMP assay for pneumococcal detection and demonstrated its excellent *in vitro* discriminative performance and speed. In Chapter 4, I compiled a diverse genome dataset of the *Streptococcus* genus, encompassing 105 species and studied the *in silico* differences and similarities across the streptococcal population structure. In Chapter 5, I summarised these findings and discussed avenues for future work.

Chapter 2 aimed to use *in silico* analyses of bacterial genome datasets to improve PCR-based diagnostics for the identification of *S. pneumoniae*. First, I used two large and diverse genome collections of *S. pneumoniae* and non-pneumococcal streptococci to assess the *in silico* performance of seven existing pneumococcal PCR-based assays. I then examined pneumococcal multiple sequence alignments to design new multiplex qPCR assays targeting Xisco and SP2020. Thereafter, I tested the *in vitro* limit of detection, sensitivity and specificity of my newly designed PCR assays at the United Kingdom Health Security Agency (UKHSA). Finally, I created an amplicon diversity panel by analysing pneumococcal genomes of which the bacterial culture was available at UKHSA. This allowed the *in vitro* testing of a panel (n=248) that represented the diversity found in a global pneumococcal genome dataset nearly 25 times its size.

The goal of Chapter 3 was to utilise *in silico* analyses to aid the development of a new, improved, real-time LAMP assay for *S. pneumoniae* detection. Building on

Chapter 2, the design of LAMP assays targeting pneumococcal genes Xisco and SP2020 was attempted; however, the high sequence diversity of Xisco across the global pneumococcal dataset prohibited LAMP assay design. Therefore, SP2020 was targeted and several LAMP primer sets were designed and trialled *in vitro*. Overall, the SP2020_3+ gave the fastest *in vitro* positive result and was taken forward. I identified four existing pneumococcal LAMP-based assays from published literature and compared the *in silico* performance of these four assays to the new SP2020_3+ assay. Thereafter, I tested the *in vitro* limit of detection, sensitivity and specificity of the SP2020_3+ assay at UKHSA and assessed the *in vitro* performance of the SP2020_3+ assay across 15 unique amplicon sequences.

Chapter 4 aimed to use *in silico* approaches to study species diversity across the *Streptococcus* genus. I compiled a large and diverse genome dataset of 1,919 genomes including 105 different species (using the rMLST scheme as a benchmark for speciation) and investigated WGS differences such as GC-content and assembly length. I then examined clustering based on single locus genotyping methods. To attempt higher resolution stratification of the *Streptococcus* genus, I assessed the streptococcal core genome. Two contrasting approaches were used, a unidirectional BLASTN search and a cluster of orthologous genes approach, applying different gene similarity settings for each. A core genome was determined and used as a framework to examine the distribution of virulence factors and AMR determinants.

Finally, in Chapter 5 I summarised and discussed the results and their strengths and limitations. I described future directions, which broadly encompass: applications of the newly developed NAATs for pneumococcal identification;

applications of the *in silico* methods to improve other diagnostics; and applications of *in silico* methods to infectious disease research. In conclusion, this thesis aimed to contribute to improving methods for the differentiation of *S. pneumoniae* from other *Streptococcus* species through the study of genomic epidemiology.

Chapter 2

Improving PCR-based diagnostics used for the identification of *Streptococcus pneumoniae* through *in silico* analysis of large bacterial genome datasets

Publication

Some of the figures, tables and text in this chapter were written up as a publication, but this manuscript has not been submitted to date.

> Femke M. Ahlers, David Litt, Melissa J. Jansen van Rensburg, James E. Bray, Keith A. Jolley, Seyi Eletu, Juliana Coelho, Bruno Pichon, Odile B. Harrison, Martin C. J. Maiden, David W. Eyre, Norman K. Fry, Angela B. Brueggemann. Utilising large and diverse bacterial genome datasets to improve the identification of *Streptococcus pneumoniae* via PCR-based diagnostics. *In preparation for submission* (2024).

International conferences

Parts of the figures, tables and text described in this chapter were presented at two international conferences: I presented a poster at the 12th International Symposium on Pneumococci and Pneumococcal Diseases (the abstract can be found in section 7.3.1) and I presented an oral presentation at the 33rd European Congress of Clinical Microbiology & Infectious Diseases (the abstract can be found in section 7.3.2).

2.1 Introduction

2.1.1 Pneumococcal identification

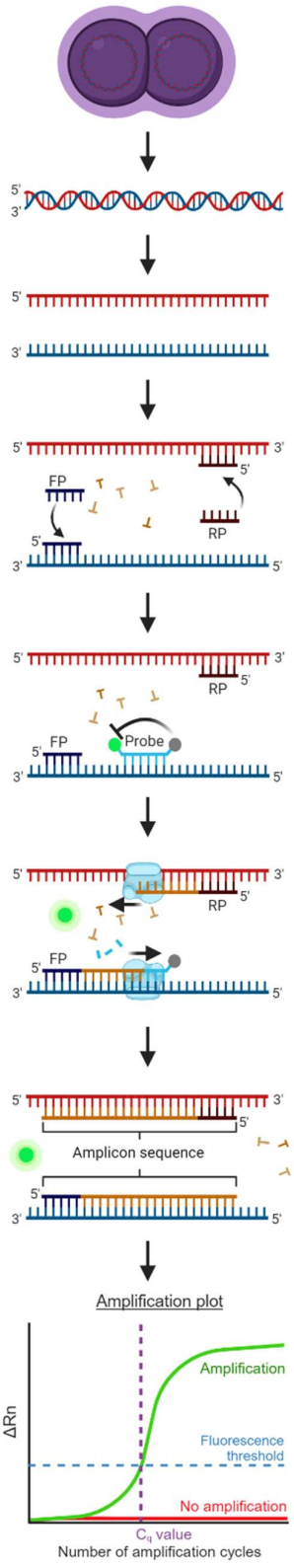
The accurate identification of *S. pneumoniae* is important for monitoring carriage rates and global disease surveillance, as well as in routine clinical practice, but is challenged by the misidentification of closely related non-pneumococcal *Streptococcus* (NPS) species [128]. Similar to *S. pneumoniae,* many NPS mitis group species can asymptomatically colonise the human nasopharynx, but unlike pneumococci they only rarely cause disease [129]. The culture-based speciation method recommended by the World Health Organisation (WHO) for the detection of carriage isolates of pneumococci includes testing for optochin susceptibility and bile solubility [130]. However, anomalous results have been reported for these phenotypic techniques [131,132]. As a result, the misidentification of NPS species as pneumococci can lead to erroneous estimations of pneumococcal disease burden, vaccine impact and antimicrobial resistance (AMR) levels [132–134]. Similarly, routine clinical diagnostics based on the analysis of bacterial protein compositions using matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS) are susceptible to misidentifying *S. pneumoniae* and mitis group NPS species, due to similarities between these species [135,136].

Culture-based diagnostics require viable bacteria, which might no longer be present in clinical samples once they reach microbiology laboratories after long transit times. Additionally, Roberts et al. (2022) found that the recoverability of *S. pneumoniae* from blood cultures is affected by delays (>46-50h) in sub-culturing at warm temperatures (35°C), whilst *S. agalactiae, Escherichia coli, Haemophilus influenzae* and *Staphylococcus aureus* are detectable after storage in these conditions [137]. This reduced recoverability of *S. pneumoniae* after delayed

processing of blood cultures can be explained by its autolysis activity [137]. Hence, recovering viable pneumococci in resource-limited tropical settings is particularly challenging [137]. Further, in the case of invasive infections, collecting culturable bacteria after the initiation of clinically urgent antimicrobial treatment, for example for suspected meningitis, is often impossible [138].

2.1.2 PCR-based identification

Polymerase chain reaction (PCR) assays can provide a culture-independent, rapid and accurate method to identify pneumococci [139]. In conventional PCR assays, the presence of the amplification product is visualised using agarose gel electrophoresis, which separates DNA fragments based on charge and size [140]. In quantitative PCR (qPCR) assays, amplification products are detected in real-time by measuring the fluorescence emitted by a reporter dye using a fluorimeter machine [141]. Different types of reporter dyes exist, with varying levels of complexity. The qPCR assays that were developed in this chapter use dual labelled probes, which have a quencher bound to their 3' end that absorbs the fluorescence of the reporter dye bound to their 5' end through fluorescence resonance energy transfer (FRET) [141]. Upon cleavage of the 5' reporter dye by *Taq* polymerase, the fluorophore is released and emits its fluorescent signal (Figure 2.1). Assays can be multiplexed to allow detection of multiple DNA targets within the same reaction mixture, by using probes with different fluorophores that emit fluorescence at different wavelengths [142]. Multiplex assays can be used to test for the presence of multiple pathogens simultaneously, or to detect multiple genes within one pathogen [142].



Streptococcus pneumoniae

Pneumococcal DNA

DNA denaturation
During the denaturation phase of the PCR assay, the temperature increase results in separation of the DNA strands.

Primer annealing
The forward primer (FP) and reverse primer (RP) anneal to the single-stranded DNA.

Probe annealing
A dual labelled probe anneals to the single-stranded DNA: the 5' end of the probe contains a fluorophore which is inhibited by a quencher attached to the 3' end of the probe.

DNA amplification
During the amplification phase, the *Taq* polymerase enzyme extends the primers. The 5'-to-3' exonuclease activity of the polymerase results in cleavage of the probe, releasing the fluorescent reporter dye, which now emits fluorescence as it is no longer in the proximity of the quencher.

DNA amplicon formation
The DNA "amplicon" encompasses the region of DNA from the forward to reverse primer sequence. The number of amplicons is doubled each amplification cycle (2^n).

Quantification of fluorescence
The fluorescent signal from the free reporter dye can be measured in real-time using a fluorimeter machine to create an amplification plot. The red line indicates a negative qPCR assay, whilst the green S-shaped curve indicates a positive qPCR assay. The quantification cycle (C_q-value; indicated by vertical purple dashed line) indicates the number of amplification cycles at which the fluorescence curve crosses the threshold (indicated by horizontal blue dashed line).

Figure 2.1 Schematic overview of qPCR using fluorescent probes.
Figure was adapted from [143] and [144] and created using Biorender.com.

2.1.3 Pneumococcal PCR targets

For a pneumococcal PCR assay, the ideal amplification target is present in all pneumococci whilst absent from any other bacterial species [130]. However, such genes are rare due to high rates of HGT within mitis group streptococci [128,133,139,145]. Several pneumococcal virulence-associated genes have long been suggested as pneumococcal PCR targets, including the major pneumococcal autolysin gene (*lytA*) which is recommended by the Centers for Disease Control and Prevention (CDC) [146,147], the pneumolysin gene (*ply*) [146], the pneumococcal surface antigen A (*psaA*) [146] and an iron uptake ABC transporter (*piaB*) [148]. In addition, genomic studies have put forward gene targets for pneumococcal identification: a genomic subtractive hybridisation study identified Spn9802 [149–151]; a pangenomic comparison of *S. pneumoniae*, *S. mitis, S. oralis* and *S. infantis* genomes identified Xisco as a specific pneumococcal protein [145,152]; and a recent study using diverse WGS datasets identified SP2020 as a target to differentiate pneumococci from *S. pseudopneumoniae* (Table 2.1) [153,128]. Nevertheless, concerns remain around whether or not these targets falsely identify NPS species. Further, prophages, which are viral genomes that have integrated their DNA into the bacterial chromosome, could cause false positive results [62,63]. In particular, prophages carrying a pneumococcal *lytA* homologue are widespread in NPS species [62]. Finally, the pneumococcal population is diverse and pneumococcal PCR targets should be present throughout pneumococcal lineages to ensure high sensitivity across different contexts. This is particularly important because the increased global sampling of pneumococcal isolates has shown differences in dominant lineages by geographic location [111,154].

Table 2.1 Targets of pneumococcal PCR-based identification assays.

Name of target gene	TIGR4[a] locus tag	Protein length in TIGR4[a] (number of amino acids)
lytA	SP_1937	319
piaB	SP_1033	340
ply	SP_1923	472
psaA	SP_1650	310
Spn9802	SP_0634	114
SP2020 or SPN0001	SP_2020	243
Xisco	SP_1992	222

[a]Pneumococcal reference genome TIGR4 [69].

2.1.4 Chapter aims

The increased availability of WGS data can be exploited to reassess the *in silico* performance of PCR targets and to design improved assays before *in vitro* validation [101,155,156]. Using large bacterial genome datasets, the diversity of nucleic acid signatures targeted by primers and probes in a bacterial population can be evaluated [156]. Therefore, the aims of this chapter were to:

- use large and diverse datasets of pneumococcal genomes (n=7,547) and NPS genomes (n=1,825) to assess the *in silico* performance of existing PCR targets,
- use WGS data to design a new qPCR assay with improved *in silico* performance,
- exploit a dataset of >6,000 pneumococci, for which both WGS data and the physical isolates were available to design an *in vitro* panel with maximised amplicon diversity,
- validate the new multiplex qPCR assay *in vitro*.

2.2 Methods

2.2.1 Compilation of publicly available genomes for *in silico* analyses

Two genomic datasets were used for the initial *in silico* assessment of PCR targets, containing a total of 9,372 publicly available assembled genomes of 56 *Streptococcus* species. All genomes were stored and analysed using the PubMLST platform supported by $BIGS_{DB}$ software [73]. The pneumococcal *in silico* study dataset contained 7,547 pneumococcal genomes compiled prior to this project, as part of the Pneumococcal Genome Library [74]. The NPS *in silico* dataset contained 1,825 genomes of 55 *Streptococcus* species.

To create the NPS *in silico* dataset, publicly available genomes of streptococcal species of interest – those that can cause human disease and/or colonise the oral cavity and/or are genetically closely related to *S. pneumoniae* – were selected from the ribosomal multilocus sequence typing (rMLST) database (https://pubmlst.org/rmlst, accessed March 2021) [73,89,157]. Five genome assembly quality control metrics were applied as follows. First, the number of contigs had to be ≤500. Second, the N50, which represents the length of the contig at which the sum of bases in all contigs of the same or smaller size constitutes half of the total assembly length, had to be ≥20,000 [158]. Third, a complete rMLST allelic profile had to be present for each genome, meaning an allele had to be defined for each of the 53 loci. Fourth, only a single allele was allowed to be present at non-paralogous rMLST loci, since the assignment of multiple alleles at non-paralogous loci could be an indication of contamination [73,89,157]. Finally, rMLST speciation had to return the same species as had been submitted initially to the database [73,89,157]. Of genomes that passed all five of these quality control metrics, one isolate was included per unique ribosomal

sequence type (rST), selecting a maximum of 180 genomes per species. If more than 180 unique rSTs were present for one species, neighbour-joining trees from concatenated rMLST sequences were constructed for that species and visualised in Interactive Tree of Life (iTOL) [159]. Outliers were removed and 180 diverse rSTs were chosen, prioritising rSTs with a higher frequency in the database and selecting the genome with the lowest number of contigs per rST.

2.2.2 Assessment of the *in silico* performance of the PCR assays

In silico analyses were performed using the In-silico PCR Experiment Simulation System (ipcress) software built into the $BIGS_{DB}$ platform [2,73]. Using the forward and reverse primers of the 11 assays (Table 2.2), the ipcress software scanned the bacterial genomes one-by-one to see if an *in silico* amplicon sequence could be formed. Ipcress was run permissively, allowing a maximum of four mismatches per primer binding region, which amounted to up to 20% of the primer length given that the shortest primer was 20 bp-long. In addition, the maximum product length was set permissively to three times the amplicon size found in pneumococcal reference genome TIGR4 (Table 2.2) [70]. Unlike in an *in vitro* PCR reaction, where the primers are incorporated into the amplicon product rather than the original sequence, the ipcress software included the primer binding regions found in the genome creating an *in silico* amplicon sequence in which mismatches from the primers could be counted [2]. The *in silico* amplicon sequences were loaded into Geneious Prime Software (v2022.1.1), where duplicates were removed and unique amplicons aligned using the MUSCLE algorithm with eight iterations [117,160]. Probe binding regions were annotated and amplicons with more than four mismatches per probe binding region were excluded, and counted as a negative *in silico* result.

Table 2.2 Published and newly designed primers and probes used in PCR and qPCR assays and the length of their amplicons.

Target[a] [Study]	Primer/ probe	Nucleotide sequence (5' to 3')	Size (bp)[b]
lytA [146]	FP	ACGCAATCTAGCAGATGAAGCA	75
	RP	TCGTGCGTTTTAATTCCAGCT	
	Probe	TGCCGAAAACGCTTGATACAGGGAG	
piaB [148]	FP	CATTGGTGGCTTAGTAAGTGCAA	104
	RP	TACTAACACAAGTTCCTGATAAGGCAAGT	
	Probe	TGTAAGCGGAAAAGCAGGCCTTACCC	
psaA [146]	FP	GCCCTAATAAATTGGAGGATCTAATGA	114
	RP	GACCAGAAGTTGTATCTTTTTTTCCG	
	Probe	CTAGCACATGCTACAAGAATGATTGCAGAAAGAAA	
ply [146]	FP	GCTTATGGGCGCCAAGTCTA	78
	RP	CAAAGCTTCAAAAGCAGCCTCTA	
	Probe	CTCAAGTTGGAAACCACGAGTAAGAGTGATGAA	
Spn9802 [150]	FP	AGTCGTTCCAAGGTAACAAGTCTAG	157
	RP	ACCAACTCGACCACCTCTTTC	
	Probe	ATCAGATTGAAGCTGATAAAACGATAC	
SP2020 [128]	FP	TAAACAGTTTGCCTGTAGTCG	155
	RP	CCCGGATATCTCTTTCTGGA	
	Probe	AACCTTTGTTCTCTCTCGTGGCAGCTCAA	
Xisco [152]	FP	TGACGATTCTAGGAAAAGATACAG	548
	RP	AGCAGGTGACTGGTAGGTAAC	
SP2020_new [This study]	FP	GCCTGTAGTCGAATGACGATCA	143
	RP	CGGATATCTCTTTCTGGAACGATT	
	Probe[c]	AACCTTTGTTCTCTCTCGTGGCAGCTCAA	
Xisco_1 [This study]	FP	CGATTGGAGTGCAAGCCATT	135
	RP	CAGCAGGTGACTGGTAGGTAACAT	
	Probe	TTGCATTGTTCCTGCAGTTGCTTGCG	
Xisco_2 [This study]	FP	GTGACGATTCTAGGAAAAGATACAGTTC	125
	RP	CTTCCAGTATTGCCTCCCTTATCT	
	Probe	CAATCTGCGAAAGGTGAATCTGTAACTCAAGAA	
Xisco_3 [This study]	FP	TCACAAGCTTCTAAGCAATTAGCTACTG	222
	RP	CACTCCAATCGTTTCTGCTTCC	
	Probe	AGCAGGATGAAATCAAAGGCGCACC	

Note: bp, base pairs; FP, Forward primer; RP, Reverse primer.
[a]All assays were quantitative PCR assays except for the published Xisco assay [152], which was a standard PCR assay and therefore had no probe.
[b]The amplicon length was determined for the pneumococcal reference genome TIGR4 [69].
[c]The probe used in the SP2020_new assay was unchanged from the published SP2020 probe [128].

A presence-absence matrix was created to investigate the genomes causing discrepancies after the *in silico* assessment. In this matrix, any genome with one or more *in silico* positive amplicon sequence(s) was counted as positive. In addition, for each PCR assay, the overall *in silico* performance was assessed calculating Youden's J statistic [161].

$$Youden's\ J\ statistic\ = \frac{True\ positive\ pneumococcal\ genomes}{Total\ pneumococcal\ study\ dataset} + \frac{True\ negative\ NPS\ genomes}{Total\ NPS\ dataset} - 1$$

2.2.3 Genome sequencing and assembly of UKHSA's streptococci

A total of 2,957 unpublished genomes sequenced by the United Kingdom Health Security Agency (UKHSA) from 22 *Streptococcus* species also had isolates available. These sequences were used to identify isolates for the design of the *in vitro* panels (the selection of isolates for *in vitro* panels is described below in section 2.2.4). These 2,957 genomes included 277 NPS isolates used in an *in vitro* specificity panel and 2,680 surveillance pneumococcal isolates, which were used in either an *in vitro* sensitivity panel (n=246) or for the design of an amplicon diversity panel (n=2,434).

These genomes were sequenced and assembled by the UKHSA Central Sequencing Laboratory prior to this study using the Illumina HiSeq 2500 platform. Trimmomatic was used to trim the deplexed fastq files [162]. An internal UKHSA bioinformatics pipeline was applied to the trimmed reads, and reads passing the quality control steps were assembled using SPAdes (v3.8.0) [163]. *Streptococcus* species identification was performed at UKHSA by using a KmerID tool. Reads identified as *S. pneumoniae* were assigned a pneumococcal MLST allelic profile using MOST and isolates were sent for re-sequencing if there was a non-

consensus base score of >15% at any position within the seven MLST loci [73,114,164]. Pneumococcal serotyping was performed through the *in silico* PneumoCaT tool [165]. All unpublished genome assemblies were uploaded to private projects in the PubMLST databases [73].

2.2.4 Compilation of panels for *in vitro* testing

Three panels were created for *in vitro* testing. First, the *in vitro* sensitivity panel consisted of 246 pneumococcal isolates which had caused invasive disease in the United Kingdom consecutively, and were submitted to UKHSA in April 2022 and were part of UKHSA's surveillance collection. WGS data for each of these isolates was generated by UKHSA, as described in section 2.2.3 and genomes passed the following genome quality control metrics. For each of the 246 pneumococcal genomes, the number of contigs was <500 and the N50 >20,000 bp. In addition, all 53 rMLST loci and all seven MLST loci had been assigned an allele [73,89,114]. Only one allele had been assigned per rMLST and MLST locus, with the exception of the two rMLST loci which are known to be paralogous in *S. pneumoniae* (BACT000014 and BACT000062) [73,89,114].

Second, the *in vitro* specificity panel consisted of 277 isolates of 21 NPS species from a laboratory collection held by UKHSA. As described in section 2.2.3, WGS data of these isolates was generated by UKHSA, and the 277 genomes passed the genome quality controls as above; however, a complete allelic profile at the pneumococcal MLST loci was not required since these are non-pneumococcal genomes.

Third, the *in vitro* pneumococcal amplicon diversity panel consisted of isolates selected to maximise the diversity of amplicons of *lytA, piaB,* SP2020_new and

the three new Xisco qPCR assays found in 6,090 pneumococci of which both WGS data and the bacterial isolates were available at UKHSA [146,148,128]. To create this amplicon diversity panel, unpublished pneumococcal genomes of the UKHSA surveillance collection, dating October 2020 - February 2022 were uploaded to a private pneumococcal PubMLST database and combined with two previously published genome datasets from Sheppard et al. and Kapatai et al. [165,166]. The following genome quality control metrics were applied. The number of contigs had to be ≤500 and the total genome assembly length had to be within 1.9-2.3 Mb. A complete rMLST allelic profile had to be present at all 53 loci and a complete MLST profile had to be present at all seven loci [89,114]. From the 6,090 pneumococcal genomes that passed all quality control metrics, the *in silico* amplicon sequences of *piaB, lytA,* SP2020_new and the three Xisco qPCR assays were investigated (as described in section 2.2.2) and isolates were selected to optimise amplicon diversity. When multiple isolates with the same amplicon were present in the *in silico* dataset, isolates were selected to maximise pneumococcal serotype diversity, and isolates with a low number of contigs were preferred.

In addition, example isolates that were *in silico* negative for one of the assays were included to test *in vitro*. To further investigate these *in silico* negative hits, a BLAST search for the reference amplicon was performed using $BIGS_{DB}$ software, outputting five hits per genome [73,167]. The hits and flanking regions were uploaded into Geneious Prime Software (v2022.1.1) and the primers and probe were aligned [117]. This allowed stratification into three categories: 1) genomes in which the amplicon was absent; 2) genomes which had an incomplete amplicon due to a contig break; 3) genomes where the probe binding region had a deletion or more than four mismatches.

2.2.5 Purification of DNA extracts for *in vitro* panels

Frozen pneumococcal glycerol stocks were cultured onto Columbia blood agar plates and incubated overnight at 36.5 °C with 5% CO_2. Pneumococcal DNA extractions were performed using the QiaSymphony DNA Mini Kit (QIAGEN) on a QiaSymphony machine, following manufacturer's guidelines. NPS isolates were cultured and their DNA was extracted by others at UKHSA following UKHSA's standard operating procedures prior to this project. DNA extracts of the pneumococcal amplicon diversity panel were quantified using the Qubit dsDNA BR Assay Kit on a Qubit 4 Fluorometer, following Invitrogen's guidelines and DNA concentrations were standardised to 0.1 ng/µL in TE buffer [Tris-HCl 10 mM (pH 8.0), EDTA 1 mM]. For DNA samples in the *in vitro* sensitivity and specificity panels, 1:100 dilutions of DNA in elution buffer were made in TE buffer [Tris-HCl 10 mM (pH 8.0), EDTA 1 mM] for use in PCR testing.

2.2.6 qPCR primer and probe design

Three assays targeting Xisco and primers for SP2020 were designed using MUSCLE alignments (eight iterations) of unique gene sequences found in the *in silico* pneumococcal study dataset (n=7,547) [73,160]. The biophysical metrics of the oligonucleotides, such as the melting temperature (T_m), percentage guanine-cytosine content, likelihood of formation of hairpin secondary structure and self-dimers, were predicted using Geneious Prime Software (v2022.1.1). Primers and probes were ordered from Eurofins Genomics. For SP2020_new, two probes with two different fluorophores were ordered: 1) a FAM-labelled probe that allowed fair comparison in limit of detection experiments against the FAM-labelled Xisco assays; 2) a CY3-labelled probe that allowed multiplexing the assay. As a positive control for the presence of bacterial DNA and the absence of PCR inhibitors, a

published qPCR assay targeting the universal 16S ribosomal RNA gene was used, changing only the fluorophore on the probe to CY5 to allow multiplexing [168].

2.2.7 Multiplex qPCR protocol used to test the three *in vitro* panels

The Minimum Information for publication of Quantitative real-time PCR Experiments (MIQE) guidelines were followed to test the three *in vitro* panels [169]. Each Xisco assay was run in multiplex with the SP2020_new and 16S rRNA (hereafter 16S) assays. The reaction mixture (20 µL) contained 2 µL of target DNA, 1X TaqPath qPCR Master Mix (Applied Biosystems), Xisco_1 or Xisco_2 or Xisco_3 primers at 0.3 µM and FAM-labelled probe at 0.3 µM, SP2020_new primers at 0.3 µM and CY3-labelled probe at 0.3 µM and 16S primers at 0.3 µM and 16S CY5-labelled probe at 0.1 µM. For comparison, a multiplex assay of published primers and probes of *lytA*, *piaB* and 16S was run (Table 2.2). The reaction mixture (20 µL) contained 2 µL of target DNA, 1X TaqMan Fast Advanced Master Mix (Applied Biosystems), *lytA* primers at 0.2 µM, FAM-labelled probe at 0.2 µM, *piaB* primers at 0.2 µM, JOE-labelled probe at 0.2 µM, 16S primers at 0.3 µM and 16S CY5-labelled probe at 0.1 µM. Assays were run in 96-well plates, which also contained a pneumococcal dilution series as the positive control and water as the negative control. After centrifuging plates at 3,000 rpm, DNA was amplified in the QuantStudio 5 system (Applied Biosystems) using the following cycling conditions: 50 °C for 2 minutes, 95 °C for 20 sec, followed by 40 cycles of 95 °C for 3 seconds, 60 °C for 30 seconds.

2.2.8 Limit of detection experiments

A pneumococcal sample of which the genome sequence had no mismatches in the primer/probe binding regions of any of the assays tested *in vitro* (i.e. Xisco_1, Xisco_2, Xisco_3, SP2020_new, *lytA* and *piaB*) was used to determine the limit of detection. Three independent serial dilutions (DNA per reaction: 2E-10; 2E-11; 2E-12; 2E-13; 2E-14, 1E-14, 5E-15 and 3E-15 g) were made in TE buffer [Tris-HCl 10 mM (pH 8.0), EDTA 1 mM]. The mean total genome length of the 38 complete genomes in the pneumococcal *in silico* study dataset (2,115,978 bp), together with the mass of DNA, was used to convert concentrations of DNA to an estimate of the number of pneumococcal genome copies per reaction volume. To fairly compare the limit of detection of singleplex qPCR assays, FAM-labelled probes were used for all assays. For Xisco_1, Xisco_2, Xisco_3 and SP2020_new, the reaction mixture (20 µL) contained 1X TaqPath qPCR Master Mix (Applied Biosystems), Xisco_1 or Xisco_2 or Xisco_3 or SP2020_new primers and FAM-labelled probe at 0.3 µM and 2 µL target DNA. The limit of detection was measured for the multiplex assays using the same qPCR protocol as described previously when testing the *in vitro* panels (section 2.2.7).

2.2.9 Data analysis

R statistical software (v4.2.1; [170]) and various packages, including binom [171], ComplexUpset [172], cowplot [173], dplyr [174], ggplot2 [175], readr [176], tibble [177], and tidyverse [178] were used for data analyses, to calculate exact binomial 95% confidence intervals and to visualise results. *In vitro* PCR data were analysed using the Applied Biosystems Design & Analysis Software 2.6.0 (ThermoFisher) and datasets were downloaded as .xslx files and opened in Microsoft Excel (v2016) for initial analysis and to create data tables. R statistical software was

used to create all figures, except from schematic overviews which were created using Biorender.com and nucleotide alignments which were visualised using Geneious Prime Software (v2022.1.1) [117,179]. Where necessary, images were arranged and modified in Inkscape. A linear model was fitted of the C_q-values and the logarithm of pneumococcal DNA copy number per well to calculate the qPCR efficiency using the formula below [144].

$$Efficiency = 10^{-\frac{1}{slope}} - 1$$

In the Bland-Altman plots, 95% limits of agreement were calculated by adding or subtracting 1.96 times the standard deviation from the mean of the differences in C_q-values of pneumococcal targets [180].

2.3 Results

2.3.1 Identification of published assays and design of new assays

Six published qPCR assays targeting *lytA, piaB, ply, psaA,* Spn9802 and SP2020 and one published standard PCR assay targeting Xisco were identified from literature for the *in silico* analyses (for recent reviews on targets for pneumococcal identification see [133] and [139]). In addition, three new qPCR assays targeting Xisco were designed: Xisco_1, Xisco_2, Xisco_3. A new pair of primers for SP2020 was also designed, which was combined with the published SP2020 probe [128]: SP2020_new (Table 2.2).

2.3.2 Design of *in silico* datasets and *in vitro* panels

All 11 assays were analysed *in silico* using two large and diverse datasets of publicly available genomes from the PubMLST database (Figure 2.2A) [73]. These two genome datasets were the pneumococcal study dataset (n=7,547) and the NPS dataset (n=1,825) [73]. The *in vitro* testing was done using the four newly designed qPCR assays (Xisco_1, Xisco_2, Xisco_3 and SP2020_new) as well as published assays targeting *lytA* and *piaB* on three panels of isolates stored at UKHSA (Figure 2.2B). These three *in vitro* panels were the sensitivity panel (n=246), the specificity panel (n=277) and the amplicon diversity panel (n=248).

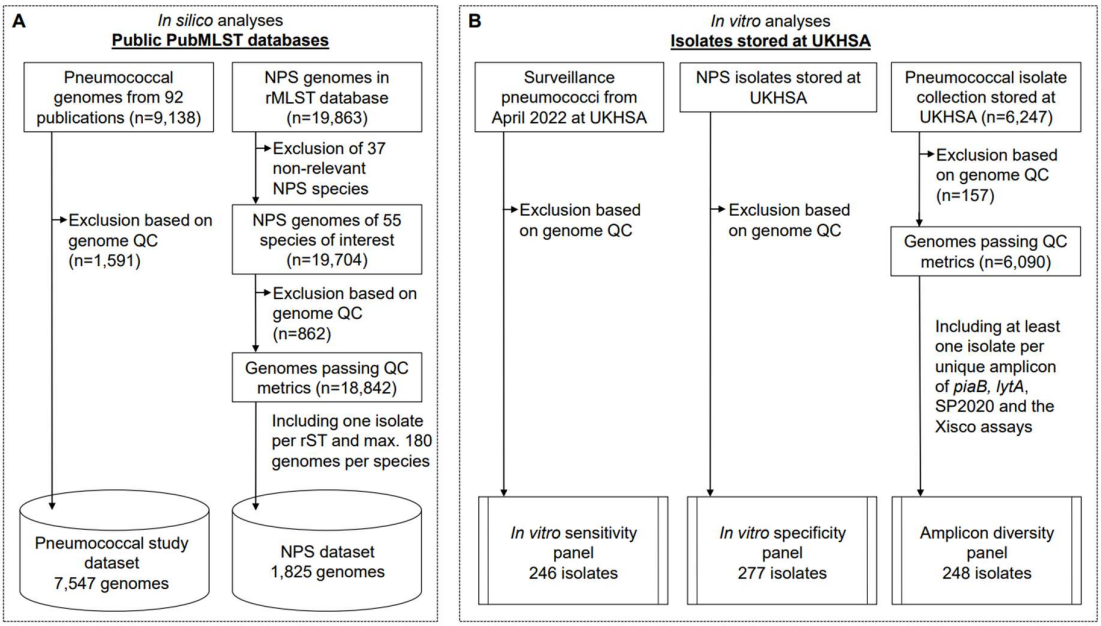


Figure 2.2 Schematic overview of the genome datasets and bacterial isolates. (A) Description of the two genome datasets used for *in silico* analyses. (B) Pneumococcal and NPS isolates tested in the *in vitro* analyses.

Note: NPS, non-pneumococcal *Streptococcus*; QC, quality control; rMLST, ribosomal multilocus sequence typing; rST, ribosomal sequence type; UKHSA, United Kingdom Health Security Agency.

2.3.3 *In silico* assessment in the pneumococcal study dataset

The genomes in the pneumococcal study dataset (n=7,547) passed stringent quality control metrics, having a low number of contigs (mean=139, [min=1; max=694]) and a high N50 (80,485 bp, [8,849 bp; 2,287,774 bp]). The total genome length (2.09 Mb, [1.94 Mb; 2.32 Mb]) and %GC (39.6%, [39.0%; 40.9%]) aligned with what is expected for a pneumococcal genome (Figure 2.3) [74]. The isolates originated from 43 countries across six continents, dated from 1916-2018, contained both carriage and disease isolates, and included 94 different serotypes, as well as nontypable strains (Figure 2.4).

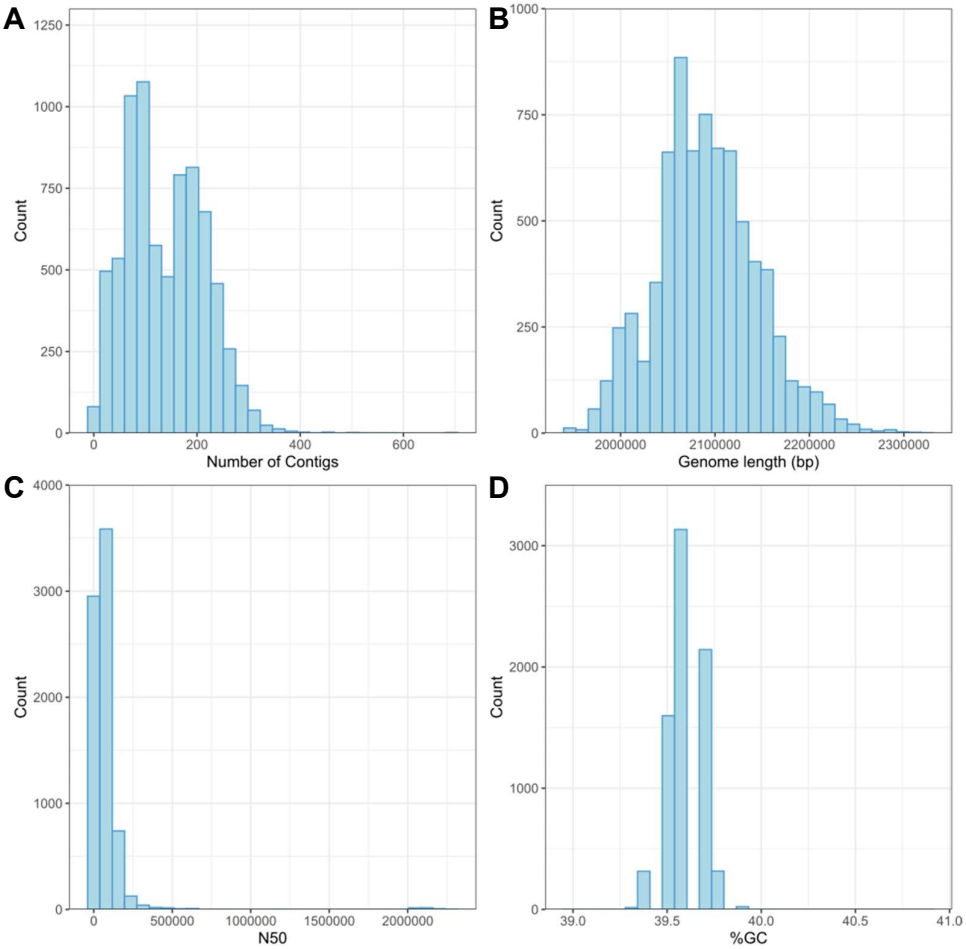


Figure 2.3 Genome metrics of the *in silico* pneumococcal study dataset (n=7,547). (A) Distribution of the number of contigs per genome. (B) Distribution of the total genome length in base pairs. (C) Distribution of the N50 of the genomes. (D) Distribution of the %GC content.

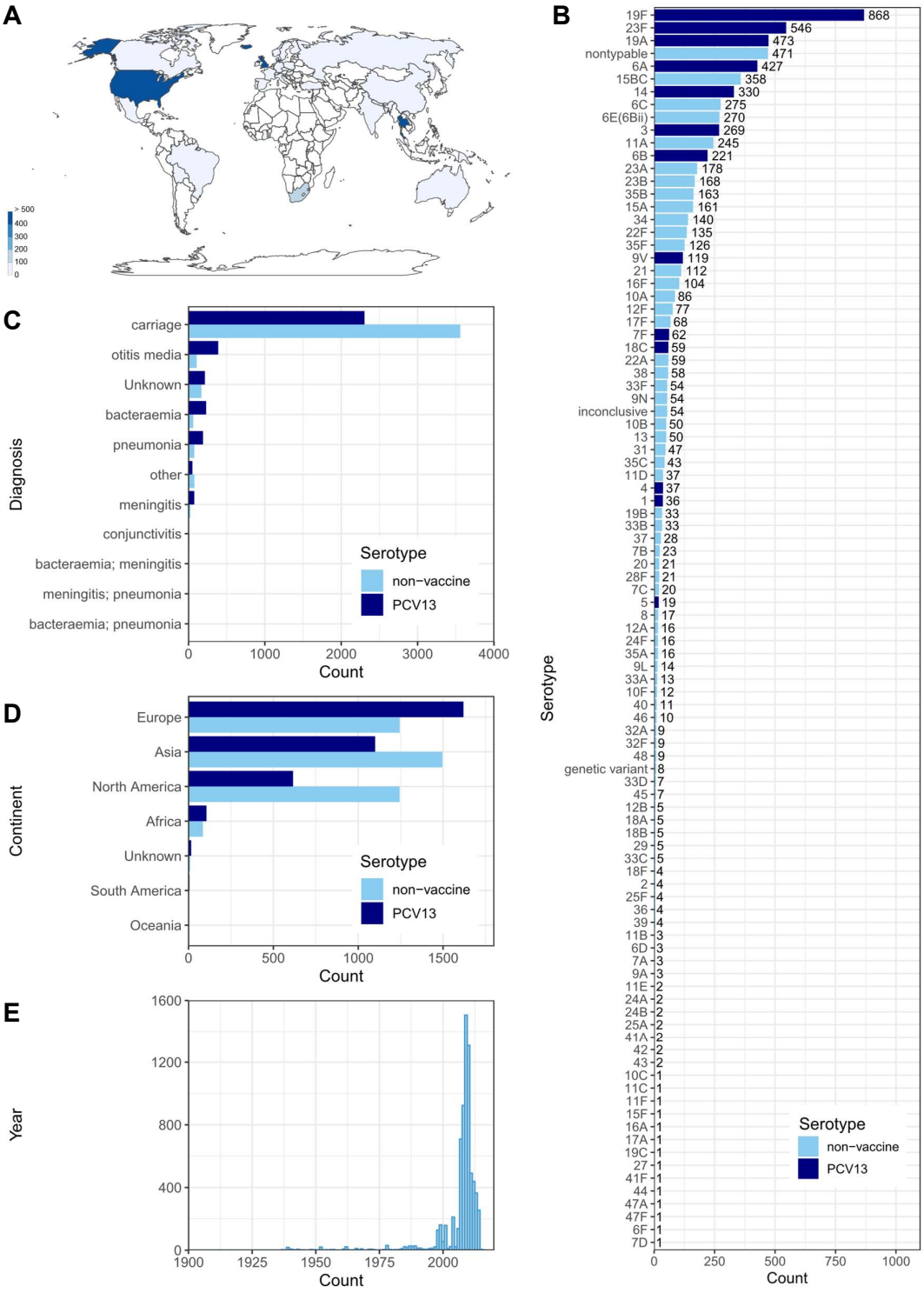


Figure 2.4 Summary of the *in silico* pneumococcal study dataset (n=7,547). (A) World map showing the countries of origin of the isolates in the dataset, created using BIGS$_{DB}$ software in PubMLST [73]. (B-D) Isolates are coloured based on whether or not their serotypes are included in the 13-valent pneumococcal conjugate vaccine (PCV13) [181]. (B) Pneumococcal serotypes. (C) Diagnosis of isolates. (D) Continent of origin of the samples. (E) Isolation year of the isolates.

The presence of the amplicons of the 11 PCR-based diagnostics was assessed *in silico* in the pneumococcal study dataset (Table 2.3). For most pneumococcal genomes, one *in silico* amplicon was found for each assay. Nevertheless, genome(s) with more than one amplicon were found for *lytA, piaB, ply,* Spn9802, SP2020 and SP2020_new. Notably, 469 genomes had two *lytA in silico* amplicons, 57 genomes had three *lytA* amplicons and nine genomes had four *lytA* amplicons per genome. The number of unique *in silico* amplicon sequences (i.e. the sequence between the forward and reverse PCR primers, including the primer binding sites) found was highest for Xisco (n=159), which also had the largest range in amplicon length. The number of unique amplicons was lowest for *ply*, SP2020 and SP2020_new (n=13 for each assay). *In silico* qPCR amplicon sequences were aligned and primer and probe binding regions annotated, allowing detailed investigation of mismatches (Appendix Figure 7.1-7.5). For example, in the *piaB* amplicons, a common (n=2,601) T>A mutation in the reverse primer binding region (amplicon nucleotide position 79) was observed.

The *in silico* positivity rates and 95% exact binomial confidence intervals were calculated (Table 2.4), finding that *lytA* had the lowest positivity rate (96.8%) of all diagnostics assessed in this study. Specifically, 243 pneumococcal genomes did not have a *lytA* amplicon *in silico*; however, this number might be an overestimation, caused by difficulty assembling the lytic amidase gene due to the presence of bacteriophage homologues in pneumococcal genomes. Of the newly designed assays, the *in silico* amplicon positivity rate ranged from 99.6-99.9%.

Table 2.3 *In silico* assessment of published and newly designed primers and probes within the pneumococcal genome dataset (n=7,547).

Target	Genomes with x amplicons (n)					Unique amplicons (n)	Range of amplicon length (bp)
	x=0	x=1	x=2	x=3	x=4		
lytA[a]	243	6,769	469	57	9	17	74-75
piaB[a]	120	7,426	1	0	0	19	104
ply[a]	3	7,523	15	6	0	13	78
psaA[a]	27	7,520	0	0	0	18	113-115
Spn9802[a]	59	7,483	5	0	0	30	154-158
SP2020[a]	29	7,512	5	1	0	13	155
Xisco[a]	6	7,541	0	0	0	159	431-581
SP2020_new[b]	29	7,512	5	1	0	13	143
Xisco_1[b]	6	7,541	0	0	0	41	123-153
Xisco_2[b]	4	7,543	0	0	0	38	101-125
Xisco_3[b]	6	7,541	0	0	0	38	159-255

[a]Previously published assay (see Table 2.2).
[b]Newly designed assay (see Table 2.2).

Table 2.4 Summary of the *in silico* assessment of published and newly designed primers and probes within the pneumococcal genome dataset (n=7,547).

Target	Predicted amplicon positive (%) [$CI_{95\%}$]
lytA[a]	96.8 [96.4-97.2]
piaB[a]	98.4 [98.1-98.7]
ply[a]	100 [99.9-100]
psaA[a]	99.6 [99.5-99.8]
Spn9802[a]	99.2 [99.0-99.4]
SP2020[a]	99.6 [99.4-99.7]
Xisco[a]	99.9 [99.8-100]
SP2020_new[b]	99.6 [99.4-99.7]
Xisco_1[b]	99.9 [99.8-100]
Xisco_2[b]	99.9 [99.9-100]
Xisco_3[b]	99.9 [99.8-100]

[a]Previously published assay (see Table 2.2).
[b]Newly designed assay (see Table 2.2).

2.3.4 *In silico* assessment in the NPS dataset

An *in silico* NPS dataset was compiled containing 1,825 assembled genomes of 55 *Streptococcus* species. *S. agalactiae, S. pyogenes* and *S. suis* were present most often (n=180 genomes for each species), followed by *S. mutans* (n=177) and *S. dysgalactiae* (n=153). All NPS genomes had a unique rST and passed stringent quality control metrics, including a low number of contigs (48, [min=1; max=500]), and a high N50 (627,484 bp, [20,370 bp; 2,719,245 bp]) (Figure 2.5).

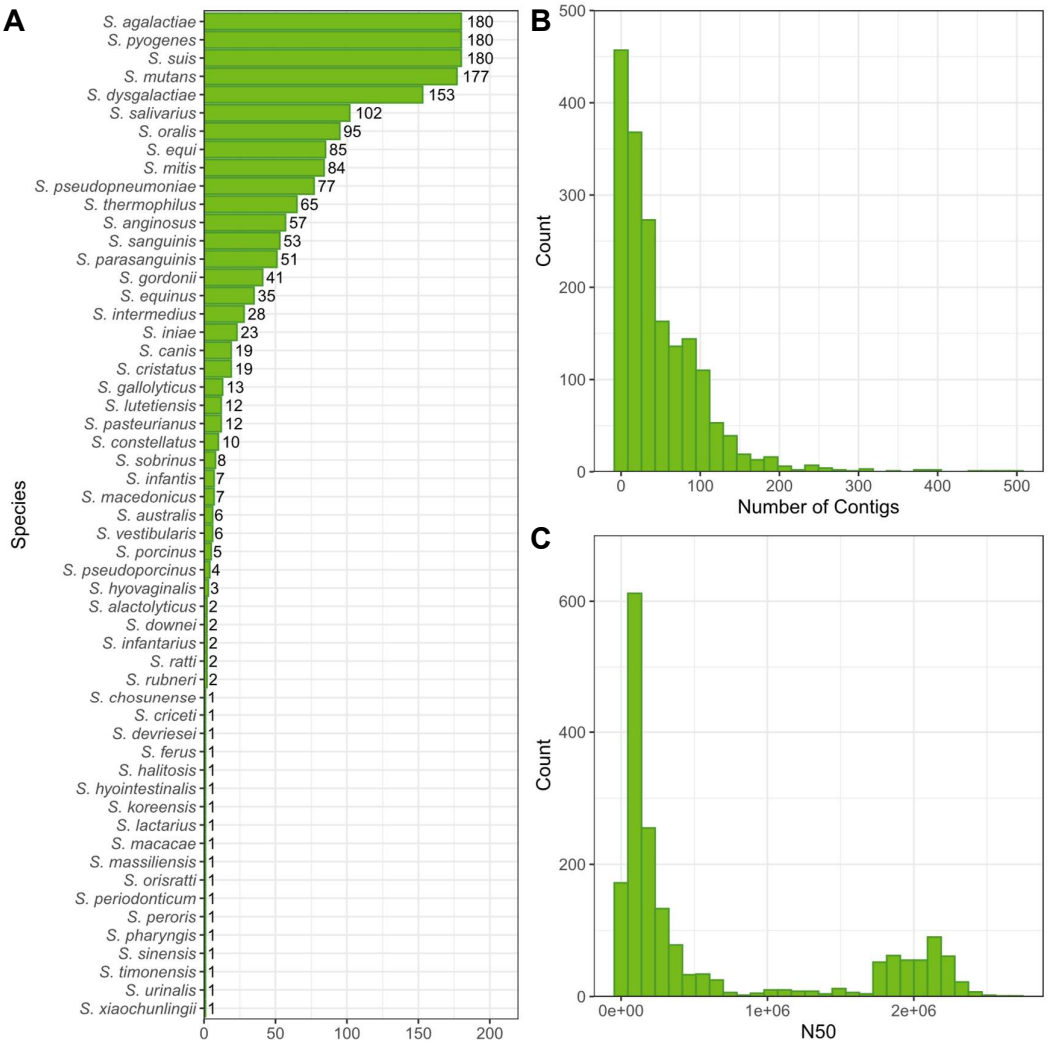


Figure 2.5 Description of the non-pneumococcal *Streptococcus* (NPS) species genome dataset (n=1,825).
(A) Frequency of each of the 55 different *Streptococcus* species. (B) Distribution of the number of contigs per NPS genome. (C) Distribution of the N50 values of the genome dataset.

The presence of the 11 PCR-based diagnostics was assessed in the *in silico* NPS dataset (Table 2.5). Of the published assays, the predicted amplicon negativity rate ranged from 95.5-100% and the only target that did not cause any positive hits in NPS genomes was Xisco. Of the newly designed assays, the predicted amplicon negativity rate ranged from 99.9-100%.

Table 2.5 Summary of the *in silico* assessment of published and new primers and probes within the non-pneumococcal *Streptococcus* genome dataset (n=1,825).

Target	Genomes with x amplicons (n)			Predicted amplicon negative (%) [$CI_{95\%}$]
	x=0	x=1	x=2	
lytA[a]	1,822	3	0	99.8 [99.5-100]
piaB[a]	1,822	3	0	99.8 [99.5-100]
ply[a]	1,743	79	3	95.5 [94.5-96.4]
psaA[a]	1,808	17	0	99.1 [98.5-99.5]
Spn9802[a]	1,752	73	0	96.0 [95.0-96.9]
SP2020[a]	1,824	1	0	99.9 [99.7-100]
Xisco[a]	1,825	0	0	100 [99.8-100]
SP2020_new[b]	1,824	1	0	99.9 [99.7-100]
Xisco_1[b]	1,825	0	0	100 [99.8-100]
Xisco_2[b]	1,825	0	0	100 [99.8-100]
Xisco_3[b]	1,825	0	0	100 [99.8-100]

[a]Previously published assay (see Table 2.2).
[b]Newly designed assay (see Table 2.2).

2.3.5 Investigating genomes with discrepant results

In total, 94.0% (7,091/7,547) of genomes in the pneumococcal study dataset were positive for all of the 11 assays. The 456 pneumococcal genomes that caused *in silico* discrepancies were stratified by the combination of assays that were *in silico* negative (Figure 2.6A). In the NPS genome dataset, 95.2% (1,738/1,825) of genomes were *in silico* negative for all assays and the 87 genomes that were *in silico* positive for one or more assays were *S. pseudopneumoniae* and *S. mitis* (Figure 2.6B). All 77 *S. pseudopneumoniae* genomes included in this dataset were

positive for one or more of the published qPCR assays and 10 of 84 *S. mitis* genomes were positive for either *ply* or *psaA.*

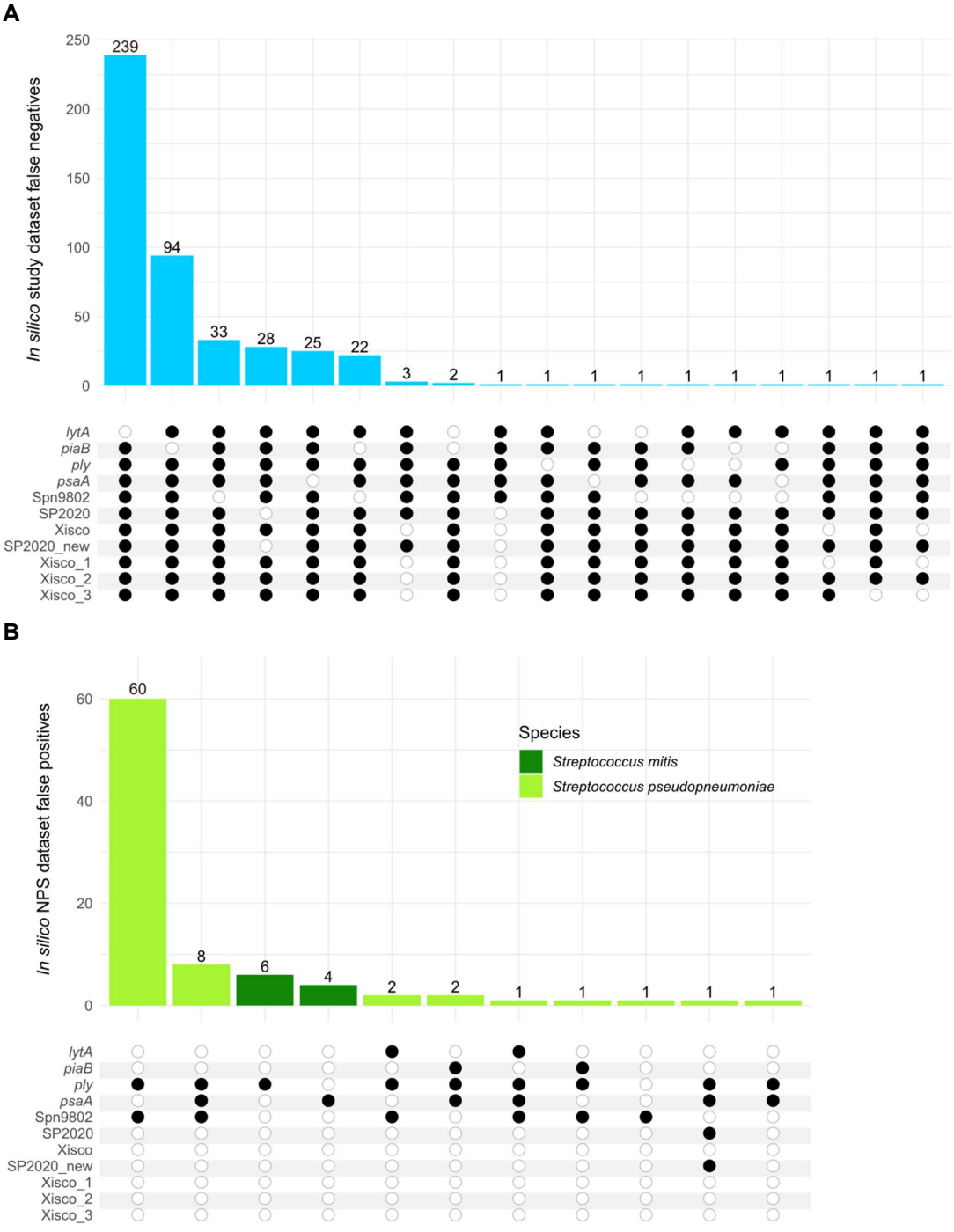


Figure 2.6 Summary of pneumococcal genomes (A) and non-pneumococcal *Streptococcus* (B) genomes causing discrepancies after *in silico* assessments of pneumococcal PCR assays.
Black circles indicate that an *in silico* PCR amplicon was detected, white circles indicate *in silico* negative tests.

2.3.6 Youden's J statistics of published assays

Youden's J statistics were calculated to assess the overall *in silico* performance of the seven published assays. Xisco and SP2020 had the highest *in silico* Youden's J statistic (>0.99), followed by *psaA* (0.99), *piaB* (0.98), *lytA* (0.97), *ply* and Spn9802 (0.95). Three new qPCR assays for Xisco were designed targeting the most conserved DNA regions of aligned genes of Xisco found in the pneumococcal study dataset (Xisco_1, Xisco_2 and Xisco_3). Multiple sequence alignments of unique SP2020 gene sequences found in the same dataset were used to design a new set of primers for SP2020 (SP2020_new) to allow multiplexing with Xisco, since the melting temperatures of the published SP2020 primers were not compatible [128]. Various biophysical metrics were assessed *in silico*, such as the hairpin T_m and the self-dimer T_m, to allow the selection of oligonucleotides with a low likelihood of forming secondary structures (Table 2.6).

Table 2.6 Biophysical metrics of primers and probes of pneumococcal PCR assays.
Calculated using Geneious Prime [117].

Gene target	Primer type	Length (bp)	T_m (°C)	%GC	Hairpin T_m (°C)	Self-dimer T_m (°C)
SP2020_new	FP	22	60.2	50.0	52.2	None
	RP	24	58.5	41.7	None	2.1
	Probe	29	68.1	48.3	None	None
Xisco_1	FP	20	59.2	50.0	33.5	7.1
	RP	24	62.1	50.0	65.9	None
	Probe	26	67.5	50.0	49.0	13.8
Xisco_2	FP	28	60.0	39.3	None	None
	RP	24	59.7	45.8	None	None
	Probe	33	65.6	39.4	None	None
Xisco_3	FP	28	61.7	39.3	58.1	16.5
	RP	22	59.8	50.0	None	None
	Probe	25	66.3	52.0	48.0	10.8

Note: FP, Forward primer; RP, Reverse primer; T_m, melting temperature; %GC, percentage guanine-cytosine content.

2.3.7 Limit of detection experiments

Three independent serial dilutions with pneumococcal DNA copy numbers per reaction ranging from one to 87,568 were used to determine the limit of detection of the assays designed in this study: Xisco_1, Xisco_2, Xisco_3, and SP2020_new (Figure 2.7). All singleplex assays had a FAM-labelled probe and the fluorescence threshold was set to 0.1 dRn to determine the C_q-values. All assays consistently detected down to two pneumococcal DNA copies per reaction; therefore, a linear model was fitted to determine qPCR efficiency through the data points ranging from two to 87,568 pneumococcal DNA copies per reaction. The qPCR efficiency of all assays was >0.95, confirming that the optimisation of amplification conditions had been successful.

When testing the *in vitro* panels, qPCR reactions were performed in multiplex. The limits of detection of the multiplex assays were assessed using the following multiplex assay combinations: Xisco_1, SP2020_new and 16S; Xisco_2, SP2020_new and 16S; Xisco_3, SP2020_new and 16S; and *lytA*, *piaB* and 16S. Each multiplex assay was able to detect at least down to nine pneumococcal genome copies per well in each of the three runs consistently (Table 2.7).

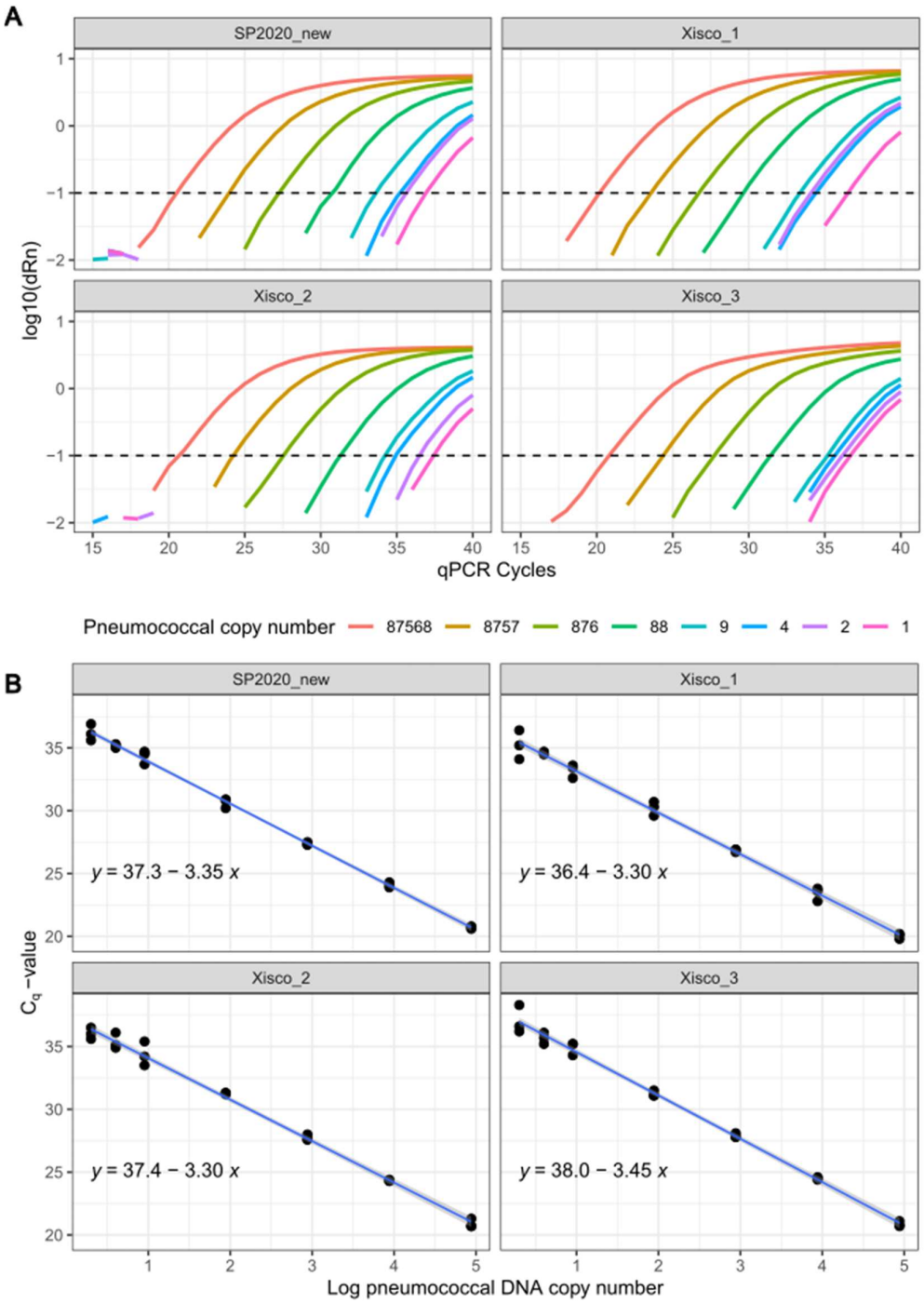


Figure 2.7 Limit of detection dilution series of Xisco_1, Xisco_2, Xisco_3 and SP2020_new singleplex assays.
(A) Representative results for one of the three dilution series, depicting qPCR curves coloured by pneumococcal copy number per reaction. Horizontal dashed lines indicate the fluorescence threshold at 0.1 dRn. (B) A linear model fitted to the results for all three dilution series (24 data points in total), with 95% confidence intervals in grey.

Table 2.7 C_q-values of limit of detection dilution series of multiplex qPCR assays.

Serial dilution		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
DNA per reaction (g)		**2E-10**	**2E-11**	**2E-12**	**2E-13**	**2E-14**	**1E-14**	**5E-15**	**3E-15**
DNA copies per reaction (n)		**87,568**	**8,757**	**876**	**88**	**9**	**4**	**2**	**1**
Xisco_1	Run 1	21.0	24.4	27.7	31.3	35.5	neg	neg	neg
Xisco_1	Run 2	21.0	24.5	27.9	31.3	35.9	36.8	neg	neg
Xisco_1	Run 3	20.8	24.2	27.7	30.9	35.8	neg	37.8	neg
SP2020_new (Xisco_1)	Run 1	21.2	24.5	27.9	31.2	34.2	35.2	neg	neg
SP2020_new (Xisco_1)	Run 2	21.3	24.7	28.1	31.4	34.8	neg	neg	neg
SP2020_new (Xisco_1)	Run 3	21.1	24.6	28.0	31.3	34.8	neg	38.4	neg
Xisco_2	Run 1	19.8	22.8	26.4	29.9	33.6	34.6	37.1	35.7
Xisco_2	Run 2	20.0	23.4	26.9	29.7	34.7	35.0	36.1	neg
Xisco_2	Run 3	19.7	23.3	26.9	30.2	34.6	35.1	36.9	37.0
SP2020_new (Xisco_2)	Run 1	18.9	21.4	25.6	29.4	33.0	36.6	36.2	35.7
SP2020_new (Xisco_2)	Run 2	19.2	22.6	26.2	29.4	35.4	neg	38.0	neg
SP2020_new (Xisco_2)	Run 3	19.1	22.6	26.2	29.5	34.6	34.0	35.5	37.0
Xisco_3	Run 1	20.6	24.2	27.5	31.3	34.4	35.2	37.7	neg
Xisco_3	Run 2	20.9	24.6	28.0	31.3	34.4	36.0	neg	neg
Xisco_3	Run 3	21.0	23.8	27.7	31.2	35.0	35.0	36.0	neg
SP2020_new (Xisco_3)	Run 1	21.1	24.7	28.0	31.4	34.7	35.8	neg	39.6
SP2020_new (Xisco_3)	Run 2	21.4	24.9	28.2	31.7	34.7	35.6	36.2	neg
SP2020_new (Xisco_3)	Run 3	21.4	24.6	28.0	31.4	35.5	35.0	neg	neg
lytA	Run 1	20.5	23.9	27.4	30.4	34.5	34.3	35.6	36.8
lytA	Run 2	20.4	24.0	27.3	30.9	34.5	35.6	neg	37.6
lytA	Run 3	20.6	23.7	27.3	30.9	34.1	35.4	36.0	37.2
piaB	Run 1	21.7	25.0	28.3	31.6	35.1	36.0	36.5	neg
piaB	Run 2	21.7	25.0	28.3	31.5	36.3	neg	36.1	neg
piaB	Run 3	21.6	24.9	28.4	31.7	35.3	37.2	36.0	36.7

Note: C_q-value, quantification cycle; neg, negative result.
[a]The following fluorescence thresholds were used: Xisco_1, Xisco_2, Xisco_3 and *lytA*, 0.1; *piaB,* 0.01; and SP2020_new, 0.01.

2.3.8 Testing the *in vitro* sensitivity panel

To create the *in vitro* sensitivity panel, the genomes of consecutive pneumococcal isolates that caused IPD in the United Kingdom in April 2022 and were sent to the national reference laboratory were subjected to quality control metrics (Figure 2.2B). The 246 pneumococcal isolates that passed all metrics had a low number of contigs (mean=91, [min=42; max=375]), a high N50 (86,981 bp, [26,965 bp; 439,878 bp]) and had a %GC content as expected for pneumococcal genomes (39.6%, [39.1%; 39.8%]). The *in vitro* sensitivity panel included 75 STs, 122 rSTs and 25 pneumococcal serotypes; the five serotypes with the highest frequency were serotype 8 (n=42), followed by 3 and 22F (n=29 each), 19A (n=21) and 9N (n=13). All of the 246 pneumococci tested positive for *lytA*, SP2020_new and the three new Xisco assays *in vitro*, giving a sensitivity of 100% [$CI_{95\%}$ 98.5-100%]. A single isolate was negative for *piaB,* giving a sensitivity of 99.6% [$CI_{95\%}$ 97.8-100%]. The *in vitro* results for these 246 isolates exactly mirrored the presence of *in silico* amplicons for the corresponding genomes.

2.3.9 Testing the *in vitro* specificity panel

The *in vitro* specificity panel contained 277 isolates from 21 different NPS species and included 257 unique rSTs (Figure 2.2B, Table 2.8). All isolates of the specificity panel tested negative *in vitro* for *lytA, piaB,* SP2020_new and the three new Xisco assays, giving a specificity of 100% [$CI_{95\%}$ 98.7-100%]. *In silico* assessment predicted that there would be a single positive isolate (*S. mitis*) for *lytA*; however, the alignment of the *lytA* primers and probe to the predicted amplicon displayed four nucleotide mismatches in both the probe and reverse primer binding regions and a single mismatch in the forward primer binding region, explaining why the *in vitro* result was negative.

Table 2.8 Non-pneumococcal *Streptococcus* species in the *in vitro* specificity panel (n=277).

Non-pneumococcal *Streptococcus* species	Isolates included (n)
Streptococcus constellatus	8
Streptococcus cristatus	3
Streptococcus downei	2
Streptococcus gallolyticus	23
Streptococcus gordonii	21
Streptococcus infantarius	6
Streptococcus intermedius	29
Streptococcus lutetiensis	24
Streptococcus massiliensis	1
Streptococcus minor	1
Streptococcus mitis	20
Streptococcus mutans	31
Streptococcus oralis	17
Streptococcus parasanguinis	26
Streptococcus pasteurianus	21
Streptococcus peroris	1
Streptococcus pseudopneumoniae	11
Streptococcus pseudoporcinus	1
Streptococcus salivarius	1
Streptococcus sanguinis	23
Streptococcus sobrinus	7

2.3.10 Design of the pneumococcal amplicon diversity panel

After applying genome quality control metrics, 6,090 pneumococcal genomes of which the physical isolates were available at UKHSA were investigated to assess the amplicon sequences of *lytA*, *piaB,* SP2020_new and the three new Xisco assays (Table 2.9). The number of unique amplicons was the highest for the three Xisco assays (range 31-45) compared to *lytA*, *piaB* and SP2020_new (range 9-15). The Xisco assays also had the highest positivity rate of 100% compared to SP2020_new, *lytA* and *piaB* (range 98.6-99.8%).

Table 2.9 Results of the *in silico* analyses of pneumococcal genomes from UKHSA (n=6,090).

Target	Genomes with x amplicons (n)				Unique amplicons (n)	Range of amplicon length (bp)	Predicted amplicon positive (%) [$CI_{95\%}$]
	x=0	x=1	x=2	x=3			
lytA[a]	71	5,544	448	27	11	75	98.8 [98.5-99.1]
piaB[a]	84	6,006	0	0	15	104	98.6 [98.3-98.9]
SP2020_new[b]	12	6,078	0	0	9	143	99.8 [99.7-99.9]
Xisco_1[b]	0	6,090	0	0	45	105-165	100 [99.9-100]
Xisco_2[b]	0	6,090	0	0	31	101-125	100 [99.9-100]
Xisco_3[b]	1	6,089	0	0	44	159-222	100 [99.9-100]

[a]Previously published assay (see Table 2.2).
[b]Newly designed assay (see Table 2.2).

The amplicon diversity panel was selected to maximise the diversity of amplicon sequences for *lytA, piaB,* SP2020_new and the three new Xisco assays. In addition, pneumococcal isolates that had an *in silico* negative result were included to test *in vitro.* Due to the absence of isolates of certain rare amplicons in the freezers at UKHSA, the final amplicon diversity panel did not include all unique amplicons found for *piaB* (14/15 included), Xisco_1 (42/45 included) and Xisco_3 (43/44 included); nevertheless, a panel of 248 pneumococci was selected that covered the vast majority of unique amplicons found in 6,090 pneumococcal genomes (Table 2.10).

Table 2.10 Summary of *in silico* results for UKHSA pneumococci of the amplicon diversity panel (n=248).

Target	Genomes with x *in silico* amplicons (n)			Unique amplicons (n)
	x=0	x=1	x=2	
lytA[a]	31	197	20	11
piaB[a]	22	226	0	14
SP2020_new[b]	12	236	0	9
Xisco_1[b]	0	248	0	42
Xisco_2[b]	0	248	0	31
Xisco_3[b]	1	247	0	43

[a]Previously published assay (see Table 2.2).
[b]Newly designed assay (see Table 2.2).

2.3.11 Testing the *in vitro* pneumococcal amplicon diversity panel

The amplicon diversity panel (n=248) was tested *in vitro* and results were compared to the presence of amplicons found through *in silico* analyses (Table 2.11). All assays that tested positive *in silico* also had an *in vitro* positive result. In contrast, assays that tested negative *in silico*, sometimes tested positive *in vitro*, which could be explained by missing genome sequence data such as end of contig hits. Notably, whilst only 87.5% (217/248) of pneumococcal genomes tested positive *in silico* for *lytA*, 99.6% (247/248) of the corresponding pneumococcal isolates tested positive *in vitro* for *lytA* (Table 2.11).

The amplicon diversity panel was tested *in vitro* using four multiplex qPCR assays, which each contained two pneumococcal targets as well as the 16S positive control assay (Figure 2.8 panels A, E, I and M). None of the pneumococcal isolates tested negative for both of the qPCRs targeting pneumococcal genes within the multiplex assays: there were no isolates testing negative for both SP2020_new and one of the three Xisco assays, nor for both *lytA* and *piaB*

(Figure 2.8 panels B, F, J and N). Bland-Altman plots were utilised to investigate the agreement between the C_q-values of the two pneumococcal targets present in each multiplex assay, by plotting the differences in C_q-values (on the y-axis) against the means of C_q-values (on the x-axis) for each isolate (Figure 2.8 panels C, G, K and O) [180]. To calculate the mean of the differences in C_q-values and 95% limits of agreement ($LOA_{95\%}$), only the isolates that tested positive *in vitro* for both of the pneumococcal targets in each multiplex assay were included: 239 isolates for SP2020_new and Xisco_1; 239 isolates for SP2020_new and Xisco_2; 238 isolates for SP2020_new and Xisco_3; and 277 isolates for *lytA* and *piaB* (Figure 2.8 panels D, H, L and P). For the SP2020_new and Xisco_1 assay the mean of differences was 0.94 [$LOA_{95\%}$ 0.19-1.69]; for SP2020_new and Xisco_2 0.43 [$LOA_{95\%}$ −0.13-0.99]; for SP2020_new and Xisco_3 0.30 [$LOA_{95\%}$ −0.09-0.69]; and for *lytA* and *piaB* 0.79 [$LOA_{95\%}$ −2.19-3.77]. Overall, the SP2020_new and Xisco_3 multiplex assay had the narrowest range between the limits of agreement, whilst the *lytA* and *piaB* multiplex assay had the widest range.

Table 2.11 Summary of the *in silico* and *in vitro* test results of the amplicon diversity panel (n=248).

Target	*In silico* positive (n)	*In vitro* positive (n)	*In silico* negative		
			Deletion in probe binding region (n)	End of contig (n)	No amplicon (n)
lytA[a]	217	247	0	25	6
piaB[a]	226	228	0	0	22
SP2020_new[b]	236	239	0	3	9
Xisco_1[b]	248	248	--	--	--
Xisco_2[b]	248	248	--	--	--
Xisco_3[b]	247	247	1	0	0

[a]Previously published assay (see Table 2.2).
[b]Newly designed assay (see Table 2.2).

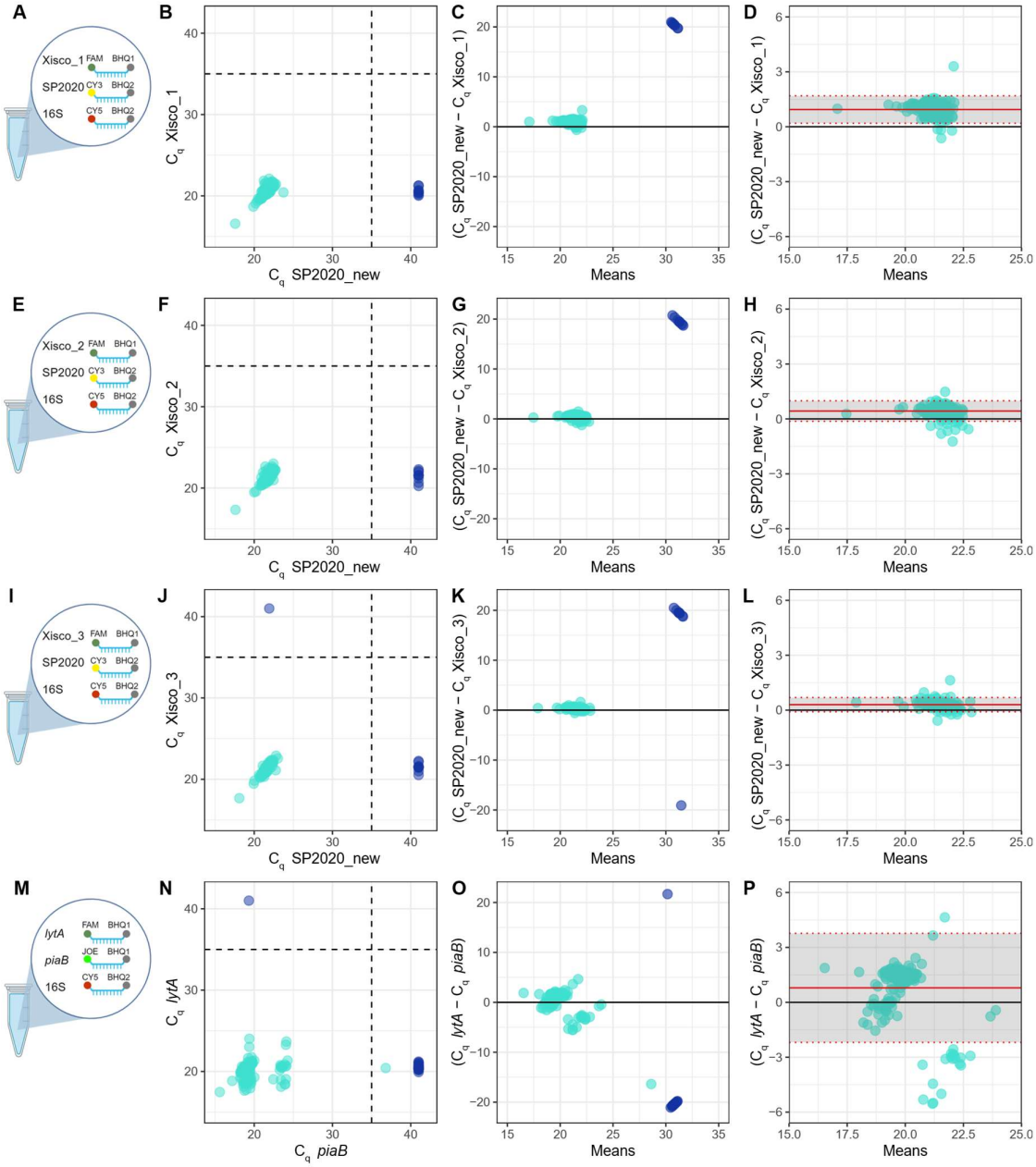


Figure 2.8 Multiplex qPCR assays tested on the *in vitro* amplicon diversity panel. Dark blue circles represent negative assays at a hypothetical C_q-value of 41.
(A, E, I, M) Schematic overview of the dual labelled probes and their fluorophores for the four multiplex assays, created using Biorender.com. Panels arranged in the same row represent the results of that multiplex qPCR assay.
(B, F, J, N) Relationship of C_q-values of multiplexed pneumococcal targets. Black dashed lines indicate the C_q-threshold of 35 cycles and a C_q>35 is a negative test.
(C, G, K, O) Bland-Altman plots of the C_q-values of multiplexed pneumococcal targets, including all 248 isolates in each of the four panels.
(D, H, L, P) Bland-Altman plots of C_q-values of isolates that were *in vitro* positive for both of the multiplexed pneumococcal targets. Red horizontal lines are the means of the differences in C_q-values. The red dotted lines represent the 95% limits of agreement.
Note: C_q-value, quantification cycle.

Aside from functioning as a positive control for the presence of bacterial DNA, the 16S assay included in each reaction was used to compare the C_q-values of assays of interest. In the amplicon diversity panel, there were more unique amplicons for the three Xisco assays (range 31-43; Table 2.10) than for the *lytA, piaB* and SP2020_new assays (range 9-14). Despite this, the C_q-values of the Xisco assays against the C_q-values of 16S clustered together closely in comparison to the C_q-values for the *lytA* and *piaB* assays, which were less strongly correlated with 16S C_q-values (Figure 2.9). Whilst qPCR assays were run for 40 cycles, assays in the amplicon diversity panel were considered negative if no amplification was detected within 35 cycles.

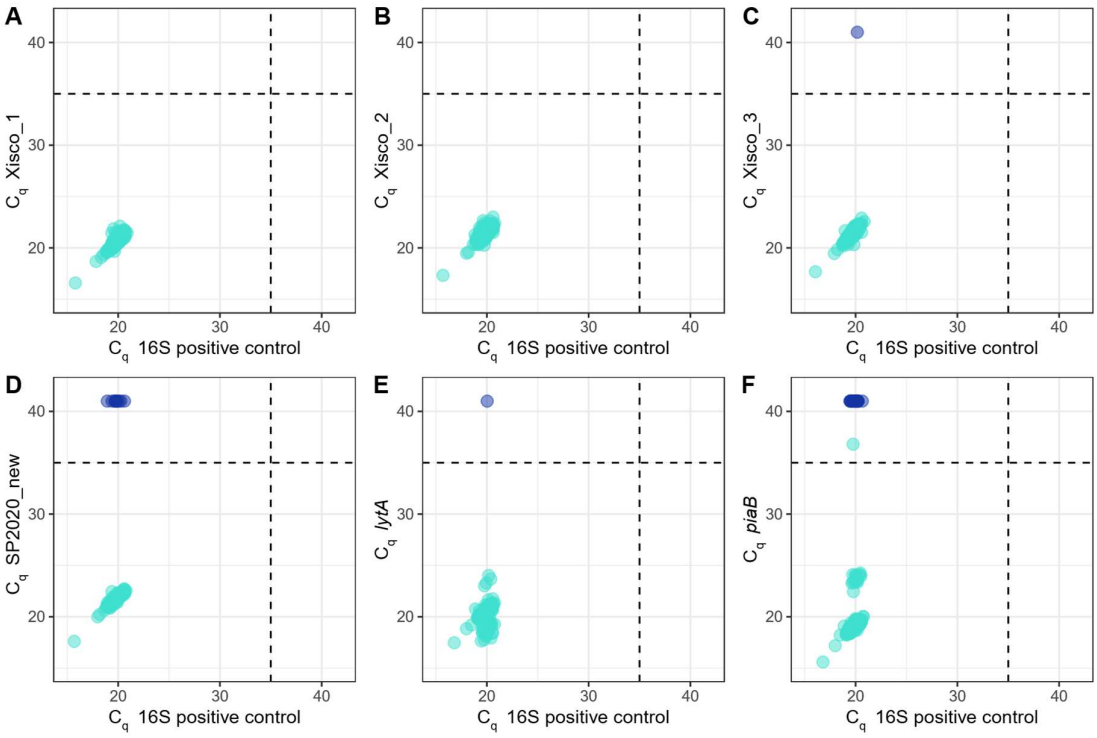


Figure 2.9 Relationship between the quantification cycles (C_q-values) of the pneumococcal targets and 16S rRNA gene positive control.
Horizontal and vertical dashed lines indicate the C_q-value of 35 cycles, which was used a threshold below which an assay is positive and above which the assay is counted as a negative result; only one pneumococcal isolate that had a C_q-value within 35-40 cycles, which was for *piaB* (panel F). Dark blue circles are placeholders of negative assays at a hypothetical C_q-value of 41.

2.4 Discussion

PCR-based diagnostics provide a rapid method for the accurate characterisation of pathogens, which can support effective clinical microbiology and public health surveillance. However, many of the PCR assays which are currently in use were designed before large WGS datasets became widely available and thus the population-level diversity of PCR primer and probe binding regions was not examined during their development [101]. Enhancements in the affordability, accuracy and speed of WGS methodology have facilitated a large increase in the number of publicly available genomes [90,182]. Here, the abundance of WGS data was exploited to assess and improve PCR-based assays for pneumococcal identification.

In silico analyses were used to evaluate seven existing pneumococcal PCR-based assays targeting *lytA, piaB, ply, psaA*, Spn9802, SP2020 and Xisco in large and diverse genome datasets. Xisco and SP2020 were the most accurate pneumococcal targets, with Youden's J statistics of >0.99 for each. Only one of the 7,547 genomes of the pneumococcal study dataset was negative *in silico* for both the new Xisco and SP2020_new assays (Figure 2.6A), whilst none of the *in vitro* tested pneumococcal isolates were negative for both targets.

Xisco was the only target for which none of the genomes in the NPS dataset tested *in silico* positive. Although the NPS genome dataset contained 55 different species, all *in silico* positive hits found for *lytA, piaB, ply, psaA,* Spn9802, SP2020 and SP2020_new were caused by *S. pseudopneumoniae* and/or *S. mitis* genomes. The frequency of this finding might be biased by *S. mitis* and *S. pseudopneumoniae* being in the top 10 most frequent species of the dataset: 84 genomes of *S. mitis* and 77 genomes of *S. pseudopneumoniae* were present in

the NPS genome dataset totalling 1,825 genomes of 55 *Streptococcus* species (Figure 2.5). Nevertheless, this result is in alignment with the close phylogenetic clustering of *S. mitis* and *S. pseudopneumoniae* to pneumococci [129]. Whilst the *in silico* positive *S. mitis* genomes were positive for only one of the qPCR assays (Figure 2.6B; six *S. mitis* genomes were positive for only *ply* and four *S. mitis* genomes were positive for only *psaA*), 76/77 *S. pseudopneumoniae* genomes were *in silico* positive for at least two pneumococcal PCR targets. *S. pseudopneumoniae* was coined as a separate species from *S. pneumoniae* based on DNA-DNA homology studies [132], and shares many, but not all, pneumococcal virulence factors [183].

In this study, *in silico* investigations were used as a rapid and inexpensive method to predict the positivity of PCR assays, allowing four mismatches per primer or probe binding region for a predicted positive result. Nevertheless, the amount of mismatches per primer binding region that still constitutes an *in vitro* positive result is dependent on a multitude of factors, including the assay protocol and annealing temperatures. In addition, the location of the mismatch within the primer binding region is crucial and mismatches near the 3' end from which the polymerase will extend the primer are known to have a much larger effect on qPCR efficiency than mismatches near the 5' end [184]. In this study, all *S. pneumoniae* genomes with an *in silico* predicted positive result that were tested *in vitro* also tested positive *in vitro*. However, in the *in vitro* specificity panel, there was a single *S. mitis* isolate that tested positive *in silico* and negative *in vitro* for *lytA.* For this *S. mitis* isolate, the *in silico* alignment of the amplicon and *lytA* primers and probes revealed that there were four primer mismatches in the probe and reverse primer each, and one mismatch in the forward primer binding region. Thus, allowing four mismatches

per primer/probe binding region to predict an *in silico* positive result might be too permissive in some cases, especially if the maximum allowed number of mismatches occurs in more than one oligonucleotide.

The reliability of the *in silico* analyses is dependent upon the quality of the genomes assessed. In this study, several metrics were applied to assure high-quality genome assemblies; however, concerns remain that amplicon sequences were not detected due to missing data, sequencing errors and/or contig breaks. Overall, 243 genomes of the pneumococcal study dataset were *in silico* negative for the *lytA* assay (Table 2.3), which might be an artefact of *lytA* being particularly difficult to assemble due to the presence of homologous prophage sequences [62,63]. Where multiple copies of similar sequences are present in short read data without different surrounding contexts, then the correct location for these reads may not be successfully determined and this may result in a contig break. In the pneumococcal amplicon diversity panel, of the 31 pneumococcal genomes which were negative *in silico* for *lytA,* 30 tested positive *in vitro* and only one tested negative *in vitro* (Table 2.11). Since prophages carrying homologues of *lytA* are known to infect NPS species, the *in silico* assessment might also have missed NPS isolates that would be positive *in vitro* for *lytA* [62]. Using long-read genome sequencing or hybrid assemblies that combine short-read with long-read data could provide more accurate and complete bacterial genomes [185,186], and might overcome these challenges. This would address this current shortcoming of *in silico* approaches; however, as current read archives are predominantly still of short read data, it is likely to remain relevant for the near future.

In this study, highly sensitive and specific new multiplex qPCR assays targeting Xisco and SP2020 were developed, through investigation of conserved regions

found in a large and diverse pneumococcal genome dataset (n=7,547). The additional targeting of 16S provided a useful internal positive control for the presence of bacterial DNA and the absence of PCR inhibitors, and allowed side-by-side comparisons of the C_q-values of different pneumococcal targets (Figure 2.9). Targeting two pneumococcal genes in one assay reduced overall false negative results in pneumococci (as shown for the amplicon diversity panel in Figure 2.8B, F and J), which might be useful when testing clinical samples from normally sterile sites such as cerebrospinal fluid (CSF) to determine whether or not disease is caused by pneumococci. In contrast, in pneumococcal carriage studies the pre-test probability of the presence of other closely related *Streptococcus* species is high. In that case, a multiplex pneumococcal qPCR assay could increase the specificity of pneumococcal detection in poly-microbial samples compared to measuring one target by requiring both targets to be positive for the identification of pneumococci [187].

The limit of detection experiments showed a consistent qPCR cycle interval between the ten-fold dilutions, suggesting that the new assay would not only be useful to determine whether or not pneumococcal DNA was present, but might also be utilised to quantify pneumococcal DNA across a wide range of copy numbers. Quantification of pneumococcal DNA can in turn be applied clinically to estimate the carriage or infection burden and to track the clinical response during treatment or in controlled infection challenge studies. In line with this, for *S. pyogenes,* a qPCR assay able to reliably quantify the bacterial load of human pharyngeal swabs has been shown to provide a reliable method to track controlled human infection models used in vaccine development [188].

Limitations of this research included that not all published PCR-based pneumococcal diagnostics were assessed *in silico*; instead, some gene targets were excluded (e.g. *recA*, *ComX* and the *cps* gene [133,189]) and only one published assay was selected for each of the seven included gene targets. Some selection bias was present in the *in silico* and *in vitro* datasets; for example, the NPS genome dataset contained 55 different species of which some are much more likely than others to co-colonise the pneumococcal nasopharyngeal niche and thus cause false positives in a carriage study [129,133]. Furthermore, the *in silico* positivity in the pneumococcal study dataset might have been inflated for some loci; this dataset was compiled prior to this project and used the presence of pneumococcal core genes as a quality control metric for the inclusion of genes [74]. In short, pneumococcal genomes that lacked too many of the pneumococcal core genes were excluded [74]. However, some of these core genes overlapped with the PCR targets tested here, therefore this original quality control metric for the pneumococcal study dataset could have inflated performance [74]. Both the *in vitro* sensitivity panel and amplicon diversity panel used high-concentration pneumococcal DNA samples extracted from pure bacterial cultures and future work should investigate if the new qPCR assays can also detect pneumococcal DNA directly from clinical samples, such as from CSF. In addition, future work could use the new qPCR assays to test poly-microbial specimens, such as nasopharyngeal or saliva samples.

Strengths of this research included the usage of large genomic datasets to assess the performance of existing PCR targets and to design new assays. Both the ipcress tool and the genomes used for the initial *in silico* analyses are publicly available through the PubMLST platform, providing an easily accessible, widely

applicable and highly reproducible method [2,73]. The selection of an amplicon diversity panel allowed testing a small panel *in vitro* that encompassed the amplicon diversity found in >6,000 pneumococci, saving costly and labour-intensive *in vitro* experiments. The *in vitro* testing itself was also optimised using a fast-cycling qPCR protocol, which took less than 25 minutes to reach 40 cycles, creating a rapid assay.

In the coming years, WGS and new applications like metagenomics sequencing are expected to become more widely implemented for direct culture-independent identification and characterisation of pathogens [90,92]. However, until WGS has become more accessible and widespread, the increased wealth of existing WGS data can be repurposed to provide improved PCR-based identification assays. Here, the combination of complementary *in silico* and *in vitro* assessments provided a rapid method to design and develop new highly sensitive and specific multiplex qPCR assays. This strategy of combining *in silico* and *in vitro* analyses could be applied to any pathogen of which sufficient genomes are available [101]. Similarly, this approach could be adapted to assess and design PCR-based assays used for sub-species characterisation, targeting AMR or pathogenicity genes. Finally, these methods could be applied to other types of nucleic acid amplification tests [101]. Indeed, the next chapter will use large and diverse bacterial genome datasets to design and develop a new pneumococcal loop-mediated isothermal amplification assay.

Chapter 3

Development of a real-time loop-mediated isothermal amplification (LAMP) assay targeting SP2020 for the identification of *Streptococcus pneumoniae* informed by large genomic datasets

Publication

Some of the figures, tables and text in this chapter were written up as a publication, but this manuscript has not been submitted to date.

> Femke M. Ahlers, David Litt, Melissa J. Jansen van Rensburg, James E. Bray, Keith A. Jolley, Juliana Coelho, Bruno Pichon, Odile B. Harrison, Martin C. J. Maiden, David W. Eyre, Norman K. Fry, Angela B. Brueggemann. Exploiting large genomic datasets for the development of a real-time loop-mediated isothermal amplification (LAMP) assay targeting SP2020 for the identification of *Streptococcus pneumoniae. In preparation for submission.* (2024).

International conferences

Parts of the figures, tables and text described in this chapter were presented as a poster at the 33rd European Congress of Clinical Microbiology & Infectious Diseases (ECCMID). The abstract of the first author poster presentation can be found in section 7.3.3.

3.1 Introduction

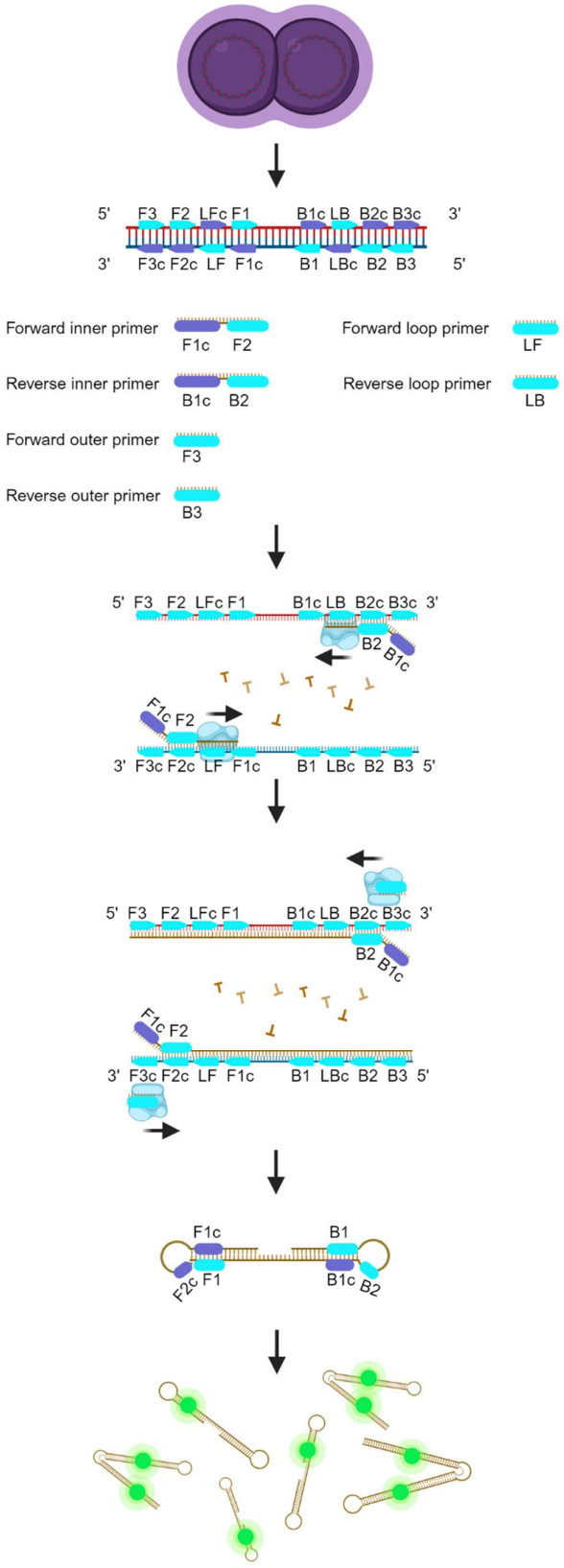
3.1.1 Loop-mediated isothermal amplification (LAMP)

The roadmap towards eliminating meningitis globally by 2030, designed by the World Health Organisation (WHO), highlights the importance of the development of affordable, accessible and rapid diagnostic assays which can be deployed into the field at scale [190]. Loop-mediated isothermal amplification (LAMP) assays provide a fast and low-cost alternative to PCR assays for the detection of nucleic acids, making LAMP assays particularly useful in low-resource settings [139,191]. Unlike PCR assays, which require heat denaturation for DNA strand separation, LAMP assays are performed at a constant temperature, for example on a heat block or incubator [192]. Since LAMP assays do not require the thermocycler machine used in PCR assays, they are a more accessible method than PCR in terms of equipment requirements [191].

LAMP technology was first developed in the year 2000, as a novel isothermal method to amplify DNA through strand displacement activity of *Bst* polymerase I or its homologues [193]. LAMP assays require four primers that target six primer binding regions: the forward inner primer (FIP) and reverse inner primer (BIP) target two DNA binding regions each and the forward outer primer (F3) and reverse outer primer (B3) target one DNA binding region each [193]. The addition of an optional pair of loop primers, called the forward loop primer (LF) and reverse loop primer (LB), has been shown to increase the reaction speed [193,194]. LAMP reactions produce stem-loop structured DNA, to which the forward and reverse inner primers can hybridise again to create secondary stem-loop structures [193].

3.1.2 LAMP detection methods

There are several types of *in vitro* detection methods that can be used to determine the result of a LAMP assay [195]. When using agarose gel electrophoresis to visualise the end-point reaction products, positive LAMP assays will produce a ladder-like pattern due to the variable length of stem-loop structured amplification products [195]. Real-time detection of LAMP reactions can be facilitated by measuring the turbidity resulting from the precipitation of magnesium pyrophosphate ($Mg_2P_2O_7$) [196]. This white precipitate is formed when pyrophosphate ions released from dNTP upon DNA polymerisation react with magnesium ions added to the LAMP reaction mixture [196]. If a turbidimeter is not available, centrifuging samples will spin down the precipitate which is then observable by the naked eye [197]. Alternatively, the addition of calcein and manganese ion complex or the addition of hydroxynaphthol blue dye to reaction mixtures produces a visible colour change when a LAMP assay is positive [197]. In this chapter, a real-time LAMP assay was created through the addition of intercalating dye SYBR Green, which binds double-stranded DNA (dsDNA) non-specifically (Figure 3.1). The fluorescence emitted by the SYBR Green was dye measured using a fluorimeter machine.



Streptococcus pneumoniae

LAMP primer design
LAMP reactions require four primers: the forward inner primer (FIP), reverse inner primer (BIP), forward outer primer (F3) and reverse outer primer (B3). FIP and BIP each contain two binding regions, one of which binds to the template strand (F2 and B2, highlighted in blue) and one of which binds to the complementary strand (F1c and B1c, highlighted in purple). An optional loop primer pair can sometimes be designed to accelerate the reaction speed.

LAMP inner primer extension
The forward and reverse inner primers anneal. Amplification then starts via extension by the polymerase. The polymerase enzyme has strand displacement activity.

LAMP outer primer extension
Extension of the outer primers (F3 and B3 which bind upstream of FIP and BIP, respectively) results in displacement of the amplicons by the polymerase. The displaced strands will undergo further amplification: they will anneal to forward and reverse inner primers and are amplified again, after which the outer primers will anneal to displace the amplicons, etc.

Formation of loop products
This process results in stem-loop structured amplification products, which in turn serve as templates for further amplification.

SYBR Green dye binding to dsDNA
Upon binding of SYBR Green dye to double-stranded DNA, it emits a fluorescent signal which can be measured in real-time using a fluorimeter machine.

Figure 3.1 Schematic overview of loop-mediated isothermal amplification (LAMP). Figure adapted from [142], [193] and [143], created using Biorender.com.

3.1.3 Published pneumococcal LAMP targets

Pneumococcal LAMP assays exist targeting various genes, including *lytA* [198–201], *psaA* [201], *ply* [202–204] and a heparinise II/III-like protein (SPNA45_01710) [205]. Comparison of a LAMP assay targeting *lytA* to a PCR assay targeting *lytA* showed a higher specificity for the LAMP assay when testing a panel of clinical isolates of oral streptococci [198]. Furthermore, a pneumococcal LAMP assay targeting *lytA,* which includes one loop primer, can detect down to 10 pneumococcal genome copies in spiked cerebrospinal fluid samples [199,200]. Du et al. (2022) designed five LAMP primer sets targeting *lytA* and five LAMP assays targeting *psaA,* selecting the primer set with the fastest reaction time for each target [201]. They measured turbidity changes in real-time at 650 nm every six seconds and added calcein dye to confirm the end-point positivity with the naked eye [201].

The *ply* gene is also targeted by several real-time pneumococcal LAMP assays [202,203]. Wang et al. (2019) developed a LAMP assay for point-of-care testing using a microchip platform to detect *S. pneumoniae* and *Mycobacterium pneumoniae*, targeting *ply* for pneumococcal identification [206]. Higgins et al. (2018) created a multiplex LAMP assay for the detection of *Haemophilus influenzae, Neisseria meningitidis* and *S. pneumoniae* for the diagnosis of bacterial meningitis [205]. For pneumococcal detection, the SPNA45_01710 gene was targeted [205]. Multiplexing was accomplished by labelling the F1c part of the forward inner primer with a fluorophore and quencher and incorporating the *Tth* endonuclease IV (TEC) enzyme in the reaction mixture which cleaves abasic sites within double-stranded DNA (dsDNA) [205]. When a loop structured amplification product is formed, the F1c binding region is part of dsDNA and TEC will release

the fluorophore [205]. Using this mechanism, different fluorophores can be used to label the forward inner primer of different targets, similar to multiplex qPCR assays having probes with different fluorophores [205].

3.1.4 Chapter aims

Whilst several pneumococcal LAMP assays exist, none target Xisco and SP2020, which were found to have the highest Youden's J statistics after *in silico* analyses of different pneumococcal PCR-based diagnostics in the previous chapter. A LAMP assay targeting the Xisco gene could not be designed since LAMP assays require more and larger primer binding regions than PCR assays and the high sequence diversity of the Xisco gene precluded the creation of LAMP primer targeting conserved regions. Therefore, a new LAMP assay targeting solely SP2020 was developed in this chapter. Overall, the aims of this chapter were to:

- design a real-time pneumococcal LAMP assay targeting SP2020,
- compare the *in silico* performance of the new pneumococcal LAMP assay targeting SP2020 to published pneumococcal LAMP assays using large and diverse genome datasets,
- test the *in vitro* sensitivity and specificity of the new pneumococcal LAMP assay and compare this LAMP assay to a published qPCR assay targeting SP2020 [128],
- utilise a dataset of >6,000 pneumococci of which the physical isolate was present at UKHSA and the WGS was available to design an *in vitro* panel with maximised SP2020 allelic diversity.

3.2 Methods

3.2.1 Design of new real-time LAMP assay targeting SP2020

PrimerExplorer Software (v5; accessed via https://primerexplorer.jp/e/) was used to design five LAMP primer sets targeting the SP2020 gene sequence of the pneumococcal reference genome TIGR4 [207,69]. Each LAMP primer set contained primers B3, BIP, F3 and FIP (Table 3.1). For each LAMP assay, the FIP was made up of DNA binding regions F1c and F2 and the BIP was made up of DNA binding regions B1c and B2. The default settings were used in PrimerExplorer Software (v5), which were as follows [207]. The length of F1c and B1c had to be 20-25 nucleotides each, with a T_m between 60-63 °C [207]. The length of F2, B2, F3 and B3 had to be 18-25 nucleotides, with a T_m between 55-58 °C (Appendix Table 7.1) [207]. In addition, the distance between F2 and B2 had to be between 120-180 bp and the distance between F1c and B1c had to be a maximum of 100 bp [207]. Using PrimerExplorer Software (v5), an additional loop primer pair could only be generated for assay SP2020_3, which was called SP2020_3+ when loop primers were included in the reaction mixture [207].

Table 3.1 Newly designed loop-mediated isothermal amplification primers targeting the pneumococcal SP2020 gene.

Assay name	Primer name	Oligonucleotide sequence 5'-3'[a]	Size (bp)[b]
SP2020_1	B3	GTTTGTTCCAGAACATAAGGTT	220
	BIP	TTAGCAGAAAAACTACAGGTCGC**CAATAATACGCA GGCGGTA**	
	F3	GCTCAAAAAGAAAATTAATCGTTCC	
	FIP	CCTCGAGTCAATTGTAGAATGAGC**AGAAAGAGATA TCCGGGGA**	
SP2020_2	B3	CGGTAAATATTATAAACTGGACTCT	231
	BIP	AGATGCTCATTCTACAATTGACTCG**GACCTGTAGT TTTTCTGCTAA**	
	F3	GGGCTTAATTTTTAGAAAACGT	
	FIP	CGGATATCTCTTTCTGGAACGATT**GCAGGGAACCT TTGTTCTC**	
SP2020_3	B3	AGTCCCCGGATATCTCTT	197
or SP2020_3+	BIP	TAGAAAACGTGGGCAGGGAA**GGAACGATTAATTTT CTTTTTGAGC**	
	F3	ACCATGAACGAACAATTGC	
	FIP	GCCCTCAGAAACTAACAAGTCTAAA**TGATGAAATG ACATTAGCTAAACAG**	
	LB	CCTTTGTTCTCTCTCGTGGC	
	LF	CGTCATTCGACTACAGGCAAA	
SP2020_4	B3	AGTCCCCGGATATCTCTT	179
	BIP	TAGAAAACGTGGGCAGGGAA**GGAACGATTAATTTT CTTTTTGAGC**	
	F3	CCTGATGAAATGACATTAGCT	
	FIP	AAGCCCTCAGAAACTAACAAGTC**ACAGTTTGCCTG TAGTCGA**	
SP2020_5	B3	AATCATAGCGATGACGACTA	194
	BIP	CCTTCCAACGGAACCGGTAT**CGAGTACTCAAACG GAGTTC**	
	F3	CAATTACATTGAAGGAAAGTTAGG	
	FIP	GCAAGTAATGTTGCTCATTTTCTGA**ATTGCATATTG CCAGTGC**	

[a]The BIP is made up of B1c and B2, of which B2 is in bold; the FIP is made up of regions F1c and F2, of which F2 is in bold.
[b]The amplicon size was determined for the pneumococcal reference genome TIGR4 [69].
Note: BIP, reverse inner primer; B3, reverse outer primer; FIP, forward inner primer; F3, forward outer primer; LB, reverse loop primer; LF, forward loop primer.

3.2.2 *In vitro* protocol for LAMP assays

For each of the LAMP assays, 10X primer stocks were made following the manufacturer's recommendations, containing 16 µM FIP, 14 µM BIP (14 µM was used rather than 16 µM to ensure sufficient stock from the volume available), 2 µM F3 and 2 µM B3 in water [208]. In the case of SP2020_3+, the 10X primer stock also contained 4 µM FL and 4 µM BL. The LAMP reaction mixture (25 µL) consisted of 1X WarmStart LAMP Master Mix ([208]; New England BioLabs), 1X SYBR Green fluorescent dye ([208]; New England BioLabs), 1X LAMP primers and 2 µL target DNA. LAMP reactions were run in 96-well plates, which included a dilution series of pneumococcal DNA derived from a known pneumococcal isolate held by UKHSA as the positive control and water as the negative control in each plate. Plates were centrifuged for 5 seconds at 3,000 rpm and incubated at 65 °C, imaging fluorescence every 20 seconds on a QuantStudio 5 machine (Applied Biosystems). The incubation time was 60 minutes for the initial experiments testing the six different LAMP primer combinations and for the SP2020_3+ limit of detection experiments. When testing the three *in vitro* panels, the incubation time was 30 minutes. After the incubation phase, the polymerase was inactivated by incubating the plates at 90 °C for 6 minutes. When determining the limit of detection of the SP2020_3+ LAMP assay, three independent dilution series were tested of a pneumococcal strain held by UKHSA which had the SP2020 TIGR4 allele, in TE buffer [Tris-HCl 10 mM (pH 8.0), EDTA 1 mM] [69].

3.2.3 Assessment of the *in silico* performance of LAMP assays

The four published pneumococcal LAMP assays (Table 3.2) and the newly designed SP2020_3+ LAMP assay (Table 3.1) were assessed in the *in silico* pneumococcal study dataset and NPS dataset previously described in sections

2.2.1 and 2.2.2. The ipcress tool was run with the LAMP forward and reverse outer primer for each assay (F3 and B3, Table 3.1 and Table 3.2), permissively allowing four mismatches per primer, with the maximum amplicon product length set to three times the amplicon size found in the pneumococcal reference genome TIGR4 [2,69,73]. *In silico* predicted amplicon sequences were uploaded into Geneious Prime Software (v2022.1.1) where duplicates were removed and sequences aligned using the MUSCLE algorithm with eight iterations [117,160]. The other LAMP binding regions (F1c, F2, B1c and B2) were annotated and amplicons with more than four mismatches per binding region were counted as *in silico* negative results. The loop primer binding regions were also investigated; however, since these primers are not required for a positive LAMP assay result they were not used to determine if the *in silico* result was positive or negative. For each of the LAMP assays, a genome was counted as *in silico* positive if one or more amplicons were present to create a presence-absence matrix. *In silico* negative pneumococcal genomes and *in silico* positive NPS genomes were investigated further. Youden's J statistics were calculated for each LAMP assay, as described in section 2.2.2.

Table 3.2 Published pneumococcal loop-mediated isothermal amplification primers.

Assay name [Study]	Primer name	Oligonucleotide sequence 5'-3'[a]	Size (bp)[b]
lytA[c] [201]	B3	ACTGGTACTGGTTCGACAAC	221
	BIP	CCTTGTACTTGACCCAGCCTGT**CAGGCTGGAA GAAAATCGCT**	
	F3	TGTTCCGTCTGGTTTGAGGT	
	FIP	CGCTAAAGAAGGCGCCATGGTA**TACCAGCCT GTTCCGTCC**	
	LB	CTTCATGGCACCTTCTTCGTTG	
	LF	TCAAATGCCTTTATCCAGTCAGC	
ply [204]	B3	AGTATCTTGAAACACATCTCCT	223
	BIP	TCCATTCAGGCGAAAAGCAGAT**GGATTTTTAA CAGCGTCTACG**	
	F3	TGTCCCAGCTAGAATGCA	
	FIP	GTCTTTTCAAAGTCAGAACCAAACT**ATGAAAA AATCATGGCTCACA**	
	LB	AGCAGATTTATTATACAGTCAG	
	LF	TGACCTTGAGTTGTTCCATG	
psaA[d] [201]	B3	CTTGTTCTGCGATAGAGTCAGT	225
	BIP	AGAATCAAGTGTGGATGACCGTCC**CGTAGATT GGGATGTTTGTGTC**	
	F3	TCTCTAAAGCCTATGGTGTTCC	
	FIP	GGCGAAGTTTTTCAACCAAGGTCT**AGTGCCTA CATCTGGGAAAT**	
	LF	GTTCAGGAGTTCCTTCTTCTTCAG	
SPNA45_01710[e] [205]	B3	AGCAAACTCACCAAGCGC	222
	BIP	TCTGTCTGGTAGACAGAATGACGGA**TCTTTGA GAATCAGATGCTGGA**	
	F3	TCCGTCAACGAGGCACAA	
	FIP	TGGAAAATGCTCTGGCTTTTGAAGTGA**CCTAC ACCAATATCCTCGCT**	
	LB	GCGCAATGATGGTATAATCCAG	
	LF	TGATGAAACAGACAAGCTGATTCT	

[a]BIP and FIP are each made up of two binding regions: BIP is made up of B1c and B2, of which B2 is in bold; FIP is made up of regions F1c and F2, of which F2 is in bold.
[b]The amplicon size was determined for the pneumococcal reference genome TIGR4 [69].
[c]This study developed five LAMP assays targeting *lytA*, of which *lytA-1* had the fastest amplification speed and was taken forward. Therefore, the *lytA-1* primers are used for *in silico* analysis in this chapter.
[d]This study developed five LAMP assays targeting *psaA*, of which *psaA-2* had the fastest amplification speed and was taken forward. Therefore, the *psaA-2* primers are used for *in silico* analysis in this chapter. This assay only has a forward and not a reverse loop primer.
[e]This study developed a multiplex LAMP assay using a *Tth* endonuclease cleavage (TEC) primer/probe, where a fluorophore, spacer and quencher are added to the FIP sequence.
Note: BIP, reverse inner primer; B3, reverse outer primer; FIP, forward inner primer; F3, forward outer primer; LB, reverse loop primer; LF, forward loop primer.

3.2.4 Phylogenetic clustering of NPS genomes testing *in silico* positive

The phylogenetic clustering of the genomes in the NPS dataset that tested *in silico* positive for one or more of the four published LAMP assays and/or the newly designed SP2020_3+ LAMP assay were investigated in relation to *S. pneumoniae*. To do this, 12 publicly available and previously published pneumococcal genomes were selected (Appendix Table 7.2), which passed the following quality control metrics. The number of contigs had to be ≤500 and the N50 had to be >20,000 [73]. There had to be a complete rMLST allelic profile at all 53 loci, whilst only a single allele was present per non-paralogous rMLST locus [73,89,157]. Finally, there had to be a complete MLST profile at all seven loci [89,114]. The rMLST allelic profiles of the 12 *S. pneumoniae* genomes and of all of the genomes in the NPS dataset (1,825 genomes representing 55 NPS species, described in section 2.2.1) were used to construct a phylogenetic tree. Three of the 53 rMLST loci (*rpsN*, *rpmD* and *rpmG*) were excluded since they were paralogous in one or more of the 1,837 genomes, and the nucleotide sequences of the remaining 50 rMLST loci were concatenated and used to create a neighbour-joining tree using $BIGS_{DB}$ software, which was visualised and midpoint rooted in iTOL (v6; https://itol.embl.de/) [73,159]. Annotation files were created using the *in silico* results of testing the LAMP assays in the NPS dataset (described in section 3.2.3).

3.2.5 Selection of isolates for *in vitro* testing

Three panels were used for *in vitro* testing, including the *in vitro* sensitivity panel (n=246) and *in vitro* specificity panel (n=277) described in section 2.2.3. The third *in vitro* panel consisted of isolates selected to maximise the number of unique SP2020 alleles. In brief, the allelic sequences of the SP2020 gene were assessed

in the 6,090 isolates of which both the WGS data and the physical isolate were available at UKHSA as described in section 2.2.3 using $BIGS_{DB}$ software in PubMLST [73,165,166]. Isolates were selected to include all unique alleles of SP2020 found in this dataset; where multiple isolates with the same allele were present, pneumococcal serotype diversity was optimised and the isolate with the lowest number of contigs was selected.

3.2.6 qPCR protocol of published SP2020 assay

For comparison, a published qPCR assay targeting SP2020 was run on the three *in vitro* panels using primers and a FAM-labelled probe as previously described by Tavares et al. [128] (Table 2.2). As a positive control for the presence of bacterial DNA, a published 16S rRNA qPCR assay was run in duplex with the SP2020 assay, using a CY5-label on the 16S rRNA probe [168]. The qPCR assay mixture (20 µL) contained 1X TaqPath qPCR Master Mix (Applied Biosystems), the SP2020 probe, forward and reverse primer and the 16S forward and reverse primer all at 0.3 µM, the 16S probe at 0.1 µM and 2 µL of purified bacterial DNA extract. Similarly to the LAMP plates, each of the qPCR plates contained a dilution series of a pneumococcal sample as the positive control and water as the negative control. After centrifuging the plates for 5 seconds at 3,000 rpm, they were incubated using the QuantStudio 5 system (Applied Biosystems) using the following conditions: 50 °C for 2 minutes, 95 °C for 20 sec, followed by 40 cycles of 95 °C for 3 seconds, 55 °C for 30 sec.

3.2.7 Data analysis

The fluorescence signal measured by the QuantStudio 5 system (Applied Biosystems) was analysed for each qPCR and LAMP reaction using the Applied Biosystems Design & Analysis Software 2.6.0 (ThermoFisher). For the

SP2020_3+ LAMP assay, settings in the software were as follows: no passive reference dye was present and fluorescence curves were viewed in the multicomponent plot. For the SP2020 qPCR assay, settings were as follows: ROX passive reference dye was present, and the amplification plot data frames were analysed. Datasets were exported as .xlsx files and opened in Microsoft Excel (v2016) for exploratory analysis and to create data tables. Confidence intervals were calculated and results were visualised using R statistical software (v4.2.1; [170]) as described in section 2.2.9. Schematic overviews were created using Biorender.com and nucleotide alignments using Geneious Prime Software (v2022.1.1) and if needed, figures were arranged and annotated in Inkscape [117,179].

3.3 Results

3.3.1 Development of a real-time LAMP assay targeting SP2020

Since the large sequence variation within the Xisco gene precluded the design of a LAMP assay, the more conserved 729 bp-long pneumococcal SP2020 gene sequence was targeted (Figure 3.2A). A loop primer pair could only be designed for SP2020_3 (called SP2020_3+ when loop primers were included). When testing the six LAMP primer combinations *in vitro,* SP2020_2 and SP2020_5 showed background drift in the non-template control (Figure 3.2B). Of the remaining assays, SP2020_3+ had the fastest reaction time (12 minutes); therefore, SP2020_3+ was selected for further *in silico* investigation and *in vitro* validation.

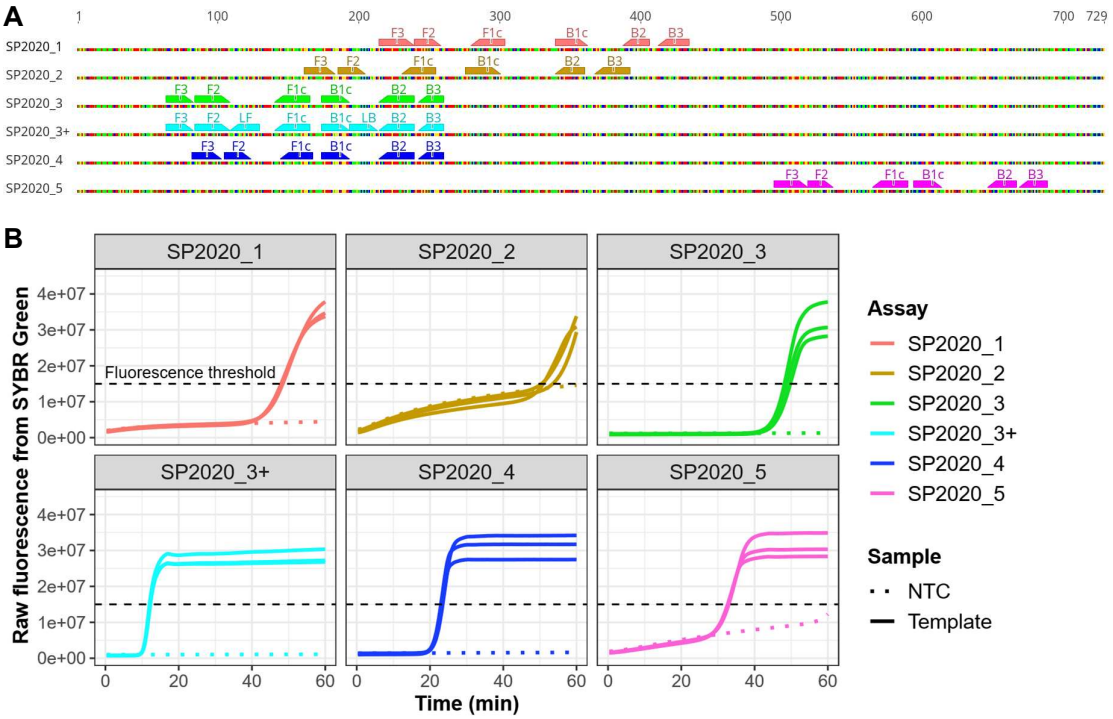


Figure 3.2 Assessment of newly designed LAMP primer combinations.
(A) Different LAMP primer binding regions aligned to the pneumococcal SP2020 gene from TIGR4 [69]. SP2020_3 is the only assay for which a loop primer pair could be designed (called SP2020_3+ when loop primers are added).
(B) Diagrams show the *in vitro* results of the six LAMP primer combinations (all reactions were run in triplicate) and their non-template controls (dotted line). The horizontal black dashed line indicates a threshold set to 15 million raw fluorescence units.
Note: LAMP, loop-mediated isothermal amplification; NTC, non-template control.

3.3.2 Identification of published pneumococcal LAMP assays

In total, the *in silico* performance of five LAMP assays was assessed. Four published pneumococcal LAMP assays targeting *lytA, ply, psaA* and SPNA45_01710 were identified from literature [201,204,205] and the SP2020_3+ assay was designed in this study. Four of the pneumococcal genes (*lytA, ply, psaA* and SP2020) targeted by the LAMP assays have previously been described as targets for the PCR-based identification of pneumococci (in section 2.1.3). However, the separate assessment of the *in silico* performance of LAMP primers was required since the LAMP assays have more primer binding regions than PCR assays and because – with the exception of the PCR and LAMP amplicons targeting SP2020 – the LAMP and PCR amplicons did not overlap (Figure 3.3).

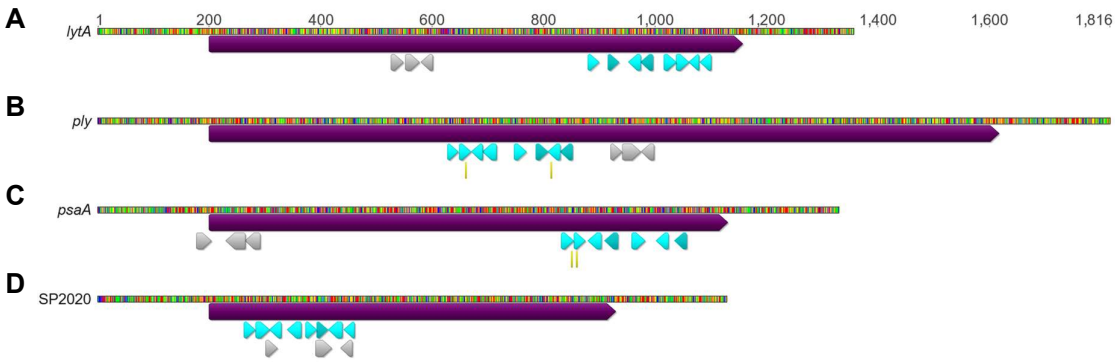


Figure 3.3 Alignment of qPCR primer/probe and LAMP primer binding regions to pneumococcal target genes.
(A-D) Genes (purple) with 200 base pairs of flanking region on either side from reference genome TIGR4 [69] are shown, with annotations highlighting mismatches (yellow), qPCR primer/probe (grey) and LAMP primer binding regions (blue). All qPCR primers/probe and LAMP primers were identified from literature, apart from the newly designed SP2020_3+ LAMP assay.
(A) The *lytA* gene, with qPCR primer/probe [146] and LAMP [201] primer binding regions annotated.
(B) The *ply* gene, with qPCR primer/probe [146] and LAMP [204] primer binding regions annotated.
(C) The *psaA* gene, with qPCR primer/probe [146] and LAMP [201] primer binding regions annotated. The qPCR forward primer binds partially outside of the gene.
(D) The SP2020 gene, with qPCR primer/probe [128] and newly designed SP2020_3+ LAMP primer binding regions annotated.
Note: LAMP, loop-mediated isothermal amplification; qPCR, quantitative polymerase chain reaction.

3.3.3 *In silico* assessment in the pneumococcal study dataset

The *in silico* performance of the four published LAMP assays and the new SP2020_3+ LAMP assay was assessed in the pneumococcal study dataset (Table 3.3). Most of the pneumococcal genomes had one copy of each of the LAMP amplicons. However, multiple amplicons per genome were found for LAMP assays targeting *lytA, ply* and SP2020_3+. Up to four *lytA* LAMP amplicons were found per genome: six pneumococcal genomes had four amplicons each; 77 genomes had three amplicons each; and 937 genomes had two amplicons each. For 1,709 pneumococcal genomes, no *in silico lytA* amplicons were found, which is likely an overestimation of the real number of pneumococci without a *lytA* LAMP amplicon, due to assembly issues at the *lytA* locus caused by prophage homologous of the *lytA* gene as described above [62,63]. Consequently, the *lytA* target might be detectable *in vitro,* whilst the *lytA* amplicon is absent *in silico* due to assembly issues. The number of unique amplicons was highest for the *lytA* LAMP assay (n=166), followed by SPNA45_01710 (n=84). The newly designed SP2020_3+ LAMP assay had the lowest number of unique amplicon sequences (n=17).

Table 3.3 *In silico* assessment of published and newly designed LAMP primers within the pneumococcal genome dataset (n=7,547).

Target	Genomes with x amplicons (n)					Unique amplicons (n)	Amplicon length (bp)
	x=0	x=1	x=2	x=3	x=4		
lytA[a]	1,709	4,818	937	77	6	166	221-404[c]
ply[a]	16	7,530	1	0	0	35	223
psaA[a]	0	7,547	0	0	0	37	225
SPNA45_01710[a]	43	7,504	0	0	0	84	221-222
SP2020_3+[b]	29	7,512	5	1	0	17	197

[a]Previously published assay (see Table 3.2).
[b]Newly designed assay (see Table 3.1).
[c]The length of 165/166 amplicons of the *lytA* LAMP assay ranged between 221-233 bp and only 1/166 amplicons was 404 bp-long due to a scaffold within that amplicon sequence.

The predicted amplicon positivity rate and 95% exact binomial confidence intervals [$CI_{95\%}$] were calculated for the published and newly designed LAMP primers (Table 3.4). For the published LAMP assays, the predicted amplicon positivity rate ranged from 77.4-100%. Notably, the *psaA* LAMP assay had an *in silico* predicted positivity rate of 100% and was the only LAMP assay without any *in silico* negative pneumococcal genomes (Table 3.3 and Table 3.4).

Table 3.4 Summary of the *in silico* assessment of published and newly designed LAMP primers within the pneumococcal genome dataset (n=7,547).

Target	Predicted amplicon positive (%) [$CI_{95\%}$]
lytA[a]	77.4 [76.4-78.3]
ply[a]	99.8 [99.7-99.9]
psaA[a]	100 [100-100]
SPNA45_01710[a]	99.4 [99.2-99.6]
SP2020_3+[b]	99.6 [99.4-99.7]

[a]Previously published assay (see Table 3.2).
[b]Newly designed assay (see Table 3.1).

For the newly designed SP2020_3+ LAMP assay, the predicted amplicon positivity rate was 99.6%. In total, 7,525 amplicons for the SP2020_3+ LAMP assay were found in 7,518/7,545 genomes of the *in silico* pneumococcal study dataset (Table 3.3). Of the 29 pneumococcal genomes that tested negative *in silico* for the SP2020_3+ LAMP assay, 25/29 had pneumococcal serotype 19F, 2/29 were nontypable and 2/29 had an inconclusive serotype. There were 17 unique SP2020_3+ amplicon sequences; however, the three amplicons with the highest frequency constituted the vast majority of the amplicons found, with proportions of 68.44% (5,150/7,525), 18.07% (1,360/7,525) and 13.02% (980/7,525) (Figure 3.4).

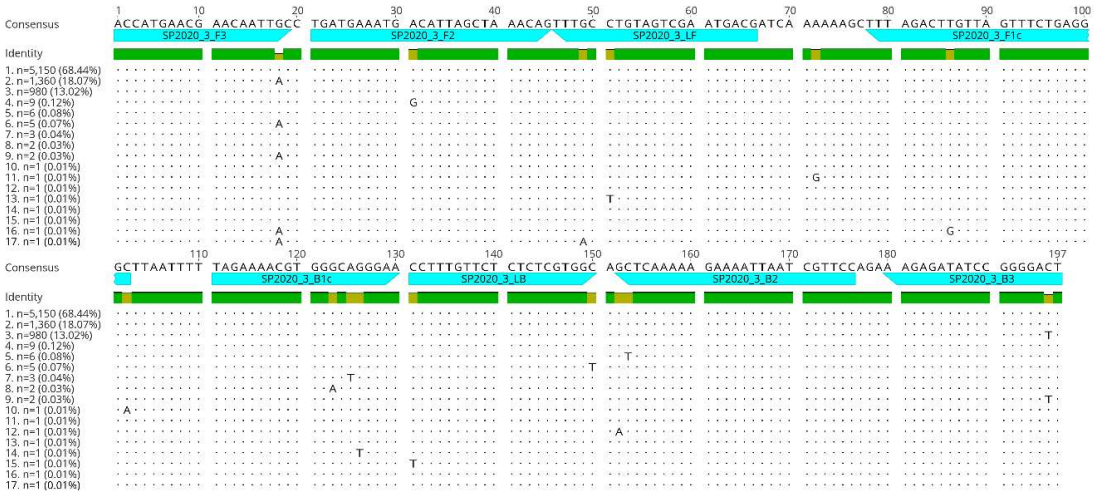


Figure 3.4 Amplicon diversity of the new SP2020_3+ LAMP assay in the *in silico* pneumococcal study dataset of 7,547 genomes.
Percentages are out of 7,525 positive amplicons found in the dataset. The consensus sequence was determined based upon the nucleotide in each position that represents at least 50% of nucleotides in that position and dots are used to represent nucleotides that are identical to the consensus sequence. The 'identity' row depicts the nucleotide sequence pairwise identity at each position: green, 100%; brown, 30-100%.
Note: BIP, reverse inner primer; B3, reverse outer primer; FIP, forward inner primer; F3, forward outer primer; LB, reverse loop primer; LF, forward loop primer; n, number of amplicons.

3.3.4 *In silico* assessment in the NPS dataset

The *in silico* performance of the new SP2020_3+ LAMP assay and the four published LAMP assays was assessed in the NPS dataset (Table 3.5). Up to three amplicons per genome were found for the *lytA* assay, up to two amplicons per genome were found for the *ply* and *psaA* assays each and up to one amplicon per genome was found for SPNA45_01710 and SP2020_3+ each. The number of unique amplicons was highest for the *psaA* assay (n=161), followed by *lytA* (n=44) and *ply* (n=16). Unique amplicon sequences were caused by nucleotide mismatches and not by insertions or deletions, resulting in a constant amplicon length for each of the LAMP assays.

Table 3.5 *In silico* assessment of published and newly designed LAMP primers within the non-pneumococcal *Streptococcus* dataset (n=1,825).

Target	Genomes with x amplicons (n)				Unique amplicons (n)	Amplicon length (bp)
	x=0	x=1	x=2	x=3		
lytA[a]	1,759	48	16	2	44	221
ply[a]	1,742	80	3	0	16	223
psaA[a]	1,555	269	1	0	161	225
SPNA45_01710[a]	1,806	19	0	0	10	222
SP2020_3+[b]	1,824	1	0	0	1	197

[a]Previously published assay (see Table 3.2).
[b]Newly designed assay (see Table 3.1).

The predicted amplicon negativity rate was calculated in the NPS dataset for each LAMP assay (Table 3.6) and ranged from 85.2-99.0% for the published LAMP assays. The *psaA* LAMP assay had the lowest predicted amplicon negativity rate (85.2%), followed by *ply* (95.5%). For each of the published LAMP assays, the *in silico* positive hits were caused by multiple *Streptococcus* species: the *psaA* LAMP assay was *in silico* positive in 10 different NPS species and the *lytA, ply* and

SPNA45_01710 assays were *in silico* positive for two different NPS species each. The newly designed SP2020_3+ LAMP assay had the highest *in silico* predicted amplicon negativity rate (99.9%) with only a single *S. pseudopneumoniae* genome testing *in silico* positive.

Table 3.6 Summary of the *in silico* assessment of published and newly designed LAMP primers within the non-pneumococcal *Streptococcus* dataset (n=1,825).

Target	Predicted amplicon negative (%) [$CI_{95\%}$]	*In silico* positive *Streptococcus* species
lytA[a]	96.4 [95.4 – 97.2]	*S. mitis* (n=1) and *S. pseudopneumoniae* (n=65)
ply[a]	95.5 [94.4 – 96.4]	*S. mitis* (n=6) and *S. pseudopneumoniae* (n=77)
psaA[a]	85.2 [83.5 – 86.8]	*S. chosunense* (n=1), *S. cristatus* (n=2), *S. halitosis* (n=1), *S. infantis* (n=5), *S. mitis* (n=84), *S. oralis* (n=95), *S. peroris* (n=1), *S. pseudopneumoniae* (n=77), *S. sanguinis* (n=3) and *S. timonensis* (n=1)
SPNA45_01710[a]	99.0 [98.4 – 99.4]	*S. halitosis* (n=1) and *S. oralis* (n=18)
SP2020_3+[b]	99.9 [99.7 – 100.0]	*S. pseudopneumoniae* (n=1)

[a]Previously published assay (see Table 3.2).
[b]Newly designed assay (see Table 3.1).

3.3.5 *In silico* Youden's J statistic of LAMP assays

Youden's J statistics were calculated to assess the overall *in silico* performance of the four published LAMP assays and the newly designed SP2020_3+ LAMP assay. The *in silico* Youden's J statistics for the published LAMP assays ranged from 0.74-0.98: the Youden's J statistic was 0.74 for *lytA;* 0.95 for *ply;* 0.85 for *psaA*; and 0.98 for SPNA45_01710. Overall, the newly designed SP2020_3+ LAMP assay had the highest *in silico* Youden's J statistic of >0.99.

3.3.6 Investigating the *in silico* discrepant results

In the pneumococcal study dataset, 76.5% (5,772/7,547) of genomes were *in silico* positive for all five of the LAMP assays. The 1,775 pneumococcal genomes that were *in silico* negative for one or more LAMP assays were stratified by the different combinations of *in silico* negative targets (Figure 3.5A). In total, 1,753 pneumococcal genomes were *in silico* negative for only one of the LAMP targets: 1,687 genomes were *in silico* negative for only *lytA*; 27 genomes were negative for only SP2020_3+; 25 genomes were negative for only SPNA45_01710; and 14 were negative for only *ply*. However, there were also pneumococcal genomes that tested *in silico* negative for multiple LAMP targets, with all three combinations of *in silico* negative targets including the *lytA* LAMP assay (Figure 3.5A). The *in silico* negative results of the *lytA* assay might be attributable to genome assembly problems caused by the presence of *lytA* homologues carried by prophages within the genome [62,63].

In total, 85.2% (1,555/1,825) of the NPS genomes were *in silico* negative for all five of the LAMP assays. The 270 genomes that were *in silico* positive for one or more the LAMP assays were caused by 10 *Streptococcus* species: *S. oralis* (n=95); *S. mitis* (n=84)*; S. pseudopneumoniae* (n=77)*; S. infantis* (n=5)*; S. sanguinis* (n=3)*; S. cristatus* (n=2)*; S. chosunense* (n=1)*; S. halitosis* (n=1)*; S. peroris* (n=1)*; S. timonensis* (n=1). These 270 NPS genomes were stratified by the different combinations of *in silico* positive targets (Figure 3.5B). Most of the genomes were *in silico* positive for only the *psaA* LAMP assay (n=168). The other 102/270 NPS genomes causing *in silico* positive pneumococcal LAMP assays were positive for more than one assay, with all combinations including the *psaA* target (Figure 3.5B).

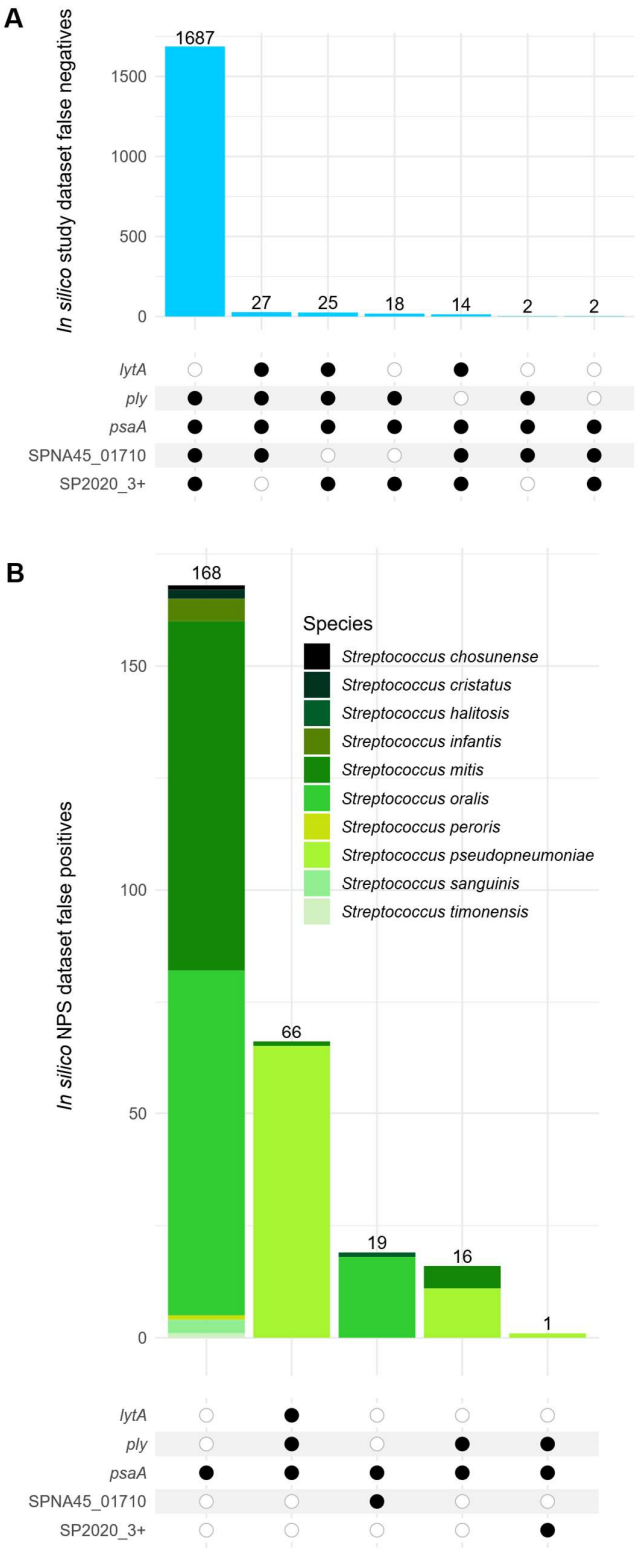


Figure 3.5 Summary of pneumococcal genomes (A) and non-pneumococcal *Streptococcus* (NPS) (B) genomes causing discrepancies after *in silico* assessments of pneumococcal LAMP assays.
Black circles indicate that an *in silico* LAMP amplicon was detected, white circles indicate *in silico* negative tests.

3.3.7 Phylogenetic clustering of NPS genomes testing *in silico* positive

A neighbour-joining midpoint rooted rMLST tree was created to explore the clustering of the NPS genomes testing *in silico* positive for one or more LAMP assays in relation to *S. pneumoniae.* Apart from *S. cristatus* and *S. sanguinis*, all NPS species that tested *in silico* positive for one or more of the pneumococcal LAMP assay(s) clustered together phylogenetically (Figure 3.6). In the NPS genome dataset, the frequency of species varies between one to 180 genomes (Figure 2.5A). For some species, all of the genomes included in the dataset were *in silico* positive for one or more of the LAMP assays: *S. oralis* (95/95); *S. mitis* (84/84)*; S. pseudopneumoniae* (77/77)*; S. chosunense* (1/1)*; S. halitosis* (1/1)*; S. peroris* (1/1) and *S. timonensis* (1/1). However, for other species, only a subset of the genomes present for that species tested positive for one or more of the LAMP assays: *S. infantis* (5/7)*; S. sanguinis* (3/53) and *S. cristatus* (2/19). Based on the rMLST alleles, the 12 *S. pneumoniae* genomes clustered most closely together to *S. pseudopneumoniae,* followed by *S. mitis* and *S. chosunense*. *S. pseudopneumoniae* was also the species with the highest number of genomes *in silico* positive for three of the pneumococcal LAMP assays (n=66; 65 *S. pseudopneumoniae* genomes were *in silico* positive for *lytA, ply* and *psaA,* and one *S. pseudopneumoniae* genome was *in silico* positive for *ply, psaA* and SP2020_3+) followed by *S. mitis* (n=1; one *S. mitis* genome was positive for *lytA, ply* and *psaA*).

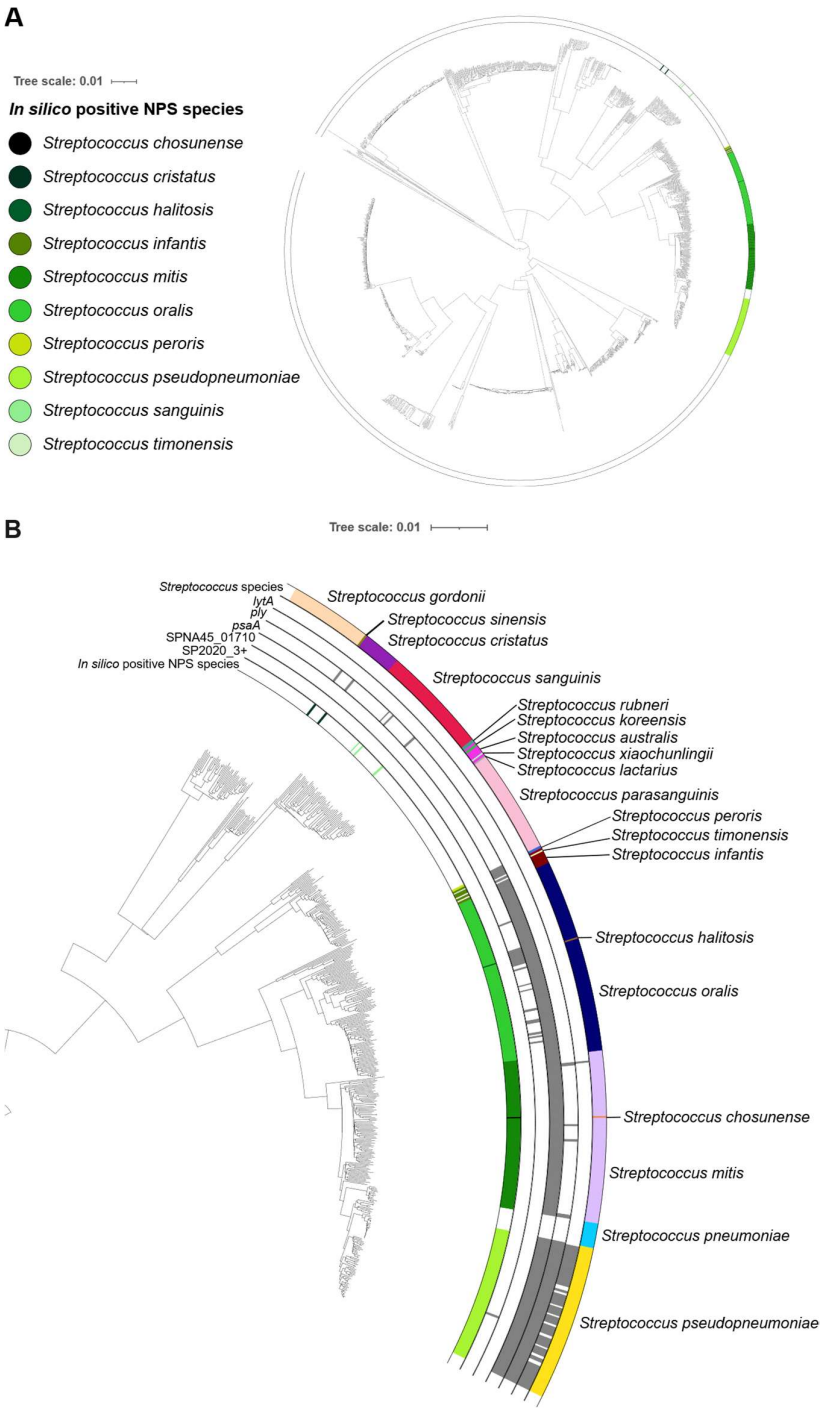


Figure 3.6 Clustering of NPS genomes testing *in silico* positive for LAMP assays. (A) Neighbour-joining tree of rMLST loci of the NPS genome dataset (n=1,825) and *S. pneumoniae* genomes (n=12), with annotations indicating NPS genomes testing *in silico* positive for pneumococcal LAMP assays coloured by species. (B) Zoomed-in branches from panel A, featuring all *in silico* positive NPS genomes. The inner ring of annotations is as in panel A. Five subsequent rings mark *in silico* positivity (in grey) of SP2020_3+, SPNA45_01710, *psaA, ply* and *lytA,* respectively. The outermost ring highlights the *Streptococcus* species.
Note: NPS, non-pneumococcal *Streptococcus*; LAMP, loop-mediated isothermal amplification; rMLST, ribosomal multilocus sequence typing.

3.3.8 Limit of detection of SP2020_3+

The limit of detection of the new SP2020_3+ LAMP assay was determined experimentally using three independent serial dilutions of pneumococcal DNA (Figure 3.7). The new SP2020_3+ LAMP assay could consistently detect down to 65 pneumococcal DNA copies in a 25 µL reaction volume (2.6 copies/µL). Lower pneumococcal DNA copy numbers were only detected in some but not in all of the three repeats of the LAMP assay.

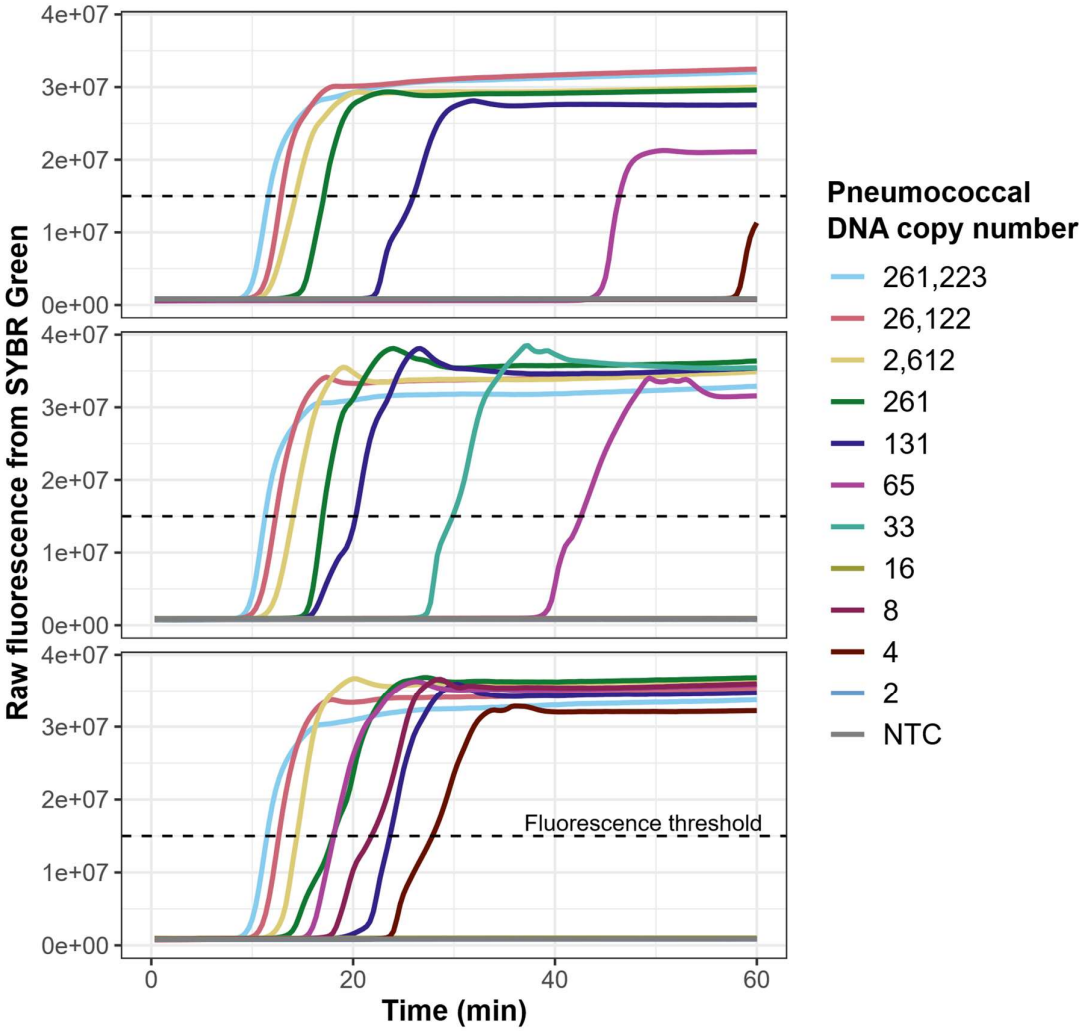


Figure 3.7 Limit of detection experiments of the SP2020_3+ LAMP assay. Three independent serial dilutions of pneumococcal DNA were utilised to determine the limit of detection of the new LAMP assay SP2020_3+. The horizontal black dashed line indicates a threshold set to 15 million raw fluorescence units.
Note: NTC; non-template control.

3.3.9 *In vitro* validation of the LAMP assay

The *in vitro* sensitivity of the newly designed SP2020_3+ LAMP assay was assessed using the *in vitro* panel of 246 pneumococci (described in section 2.2.3) and compared to the performance of a published qPCR assay targeting SP2020 [128]. All of the pneumococci tested positive for both the SP2020_3+ LAMP assay and the SP2020 qPCR assay (246/246; *in vitro* sensitivity 100% [$CI_{95\%}$ 98.5-100%]). The *in vitro* specificity of the newly designed SP2020_3+ LAMP assay was assessed using the *in vitro* panel of 277 isolates from 21 different NPS species (Table 2.8, described in section 2.2.3) and compared to the performance of a published qPCR assay targeting SP2020 [128]. None of these NPS isolates caused cross-reactivity with the SP2020_3+ LAMP assay nor with the qPCR assay (0/277; *in vitro* specificity 100% [$CI_{95\%}$ 98.7-100%]).

3.3.10 SP2020 diversity panel

A diversity panel was selected to include isolates representing the 71 unique SP2020 gene alleles that were found amongst the 6,090 pneumococcal genomes of which isolates were physically present at the UKHSA. Within these 71 unique SP2020 alleles, there were 15 unique SP2020_3+ LAMP amplicon sequences found *in silico*. The 15 amplicons were highly similar: there was a maximum of two nucleotide deviations from the consensus sequence per amplicon and none of the amplicons had insertions nor deletions (Figure 3.8A). Testing this panel *in vitro* showed that the SP2020_3+ assay was positive for all unique amplicons and that a fast reaction time was achieved for all isolates despite mutations in primer binding regions (Figure 3.8B). The published SP2020 qPCR, for which there were nine unique amplicons represented in the 71 SP2020 alleles of the diversity panel, also tested positive *in vitro* for all isolates [128].

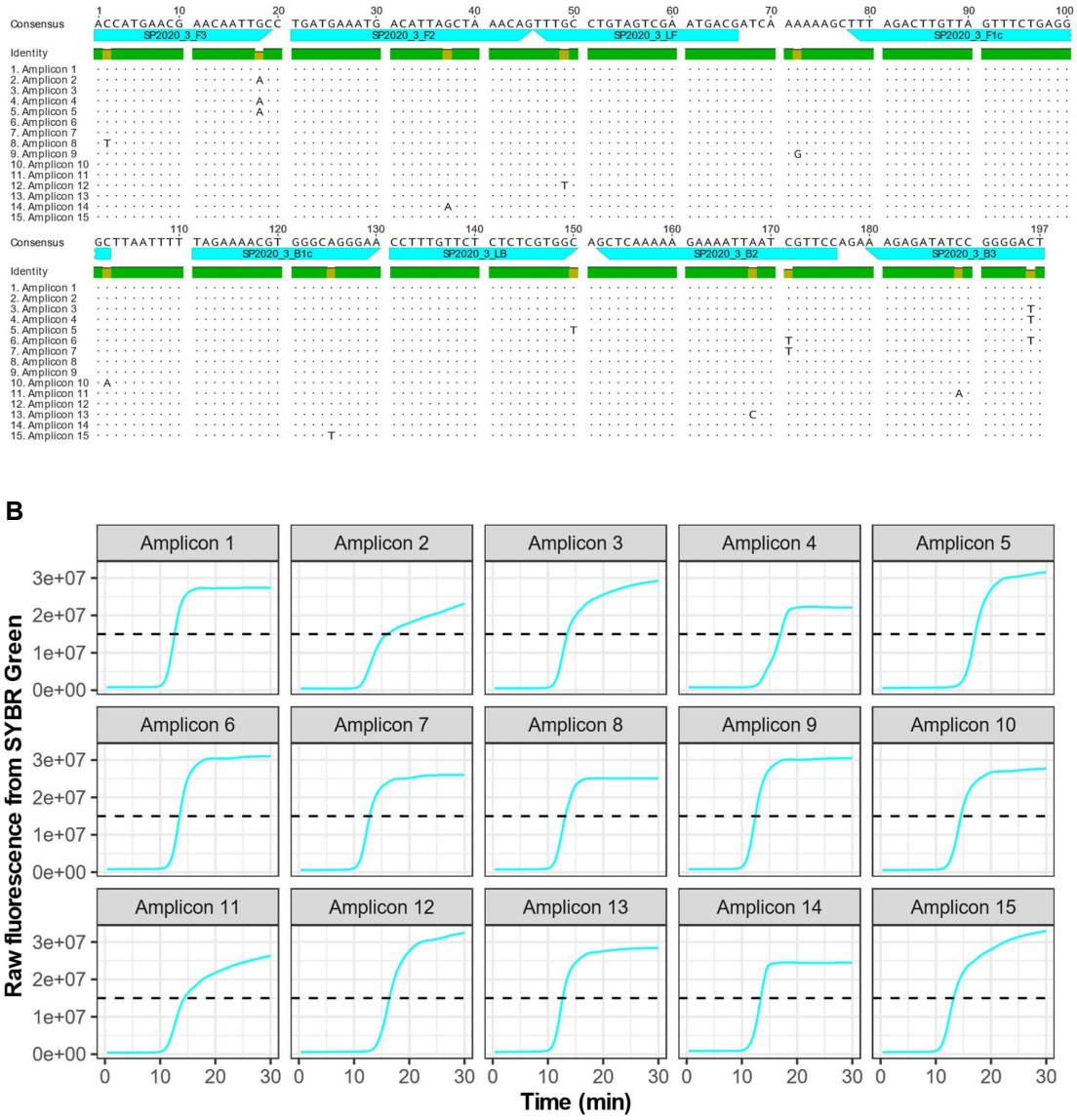


Figure 3.8 Fifteen unique SP2020_3+ LAMP amplicon sequences of the *in vitro* SP2020 diversity panel.
(A) Fifteen unique LAMP amplicon sequences were found for the new SP2020_3+ LAMP assay in >6,000 pneumococcal genomes. Dots are used to represent nucleotides that are identical to the consensus sequence.
(B) *In vitro* results of the isolates representing the fifteen unique SP2020_3+ amplicon sequences. The horizontal black dashed line indicates a threshold set to 15 million raw fluorescence units.

3.4 Discussion

When developing diagnostic assays for infectious disease surveillance in low-resource settings, the WHO emphasises the importance of the ASSURED characteristics. This acronym encompasses the following eight metrics: Affordable, Sensitive, Specific, User-friendly, Rapid, Robust, Equipment-free and Delivered [209]. LAMP assays align excellently with these characteristics by virtue of needing only simple and minimal equipment [191]. Therefore, this chapter developed a LAMP assay for the identification of pneumococci, using a combination of *in silico* and *in vitro* approaches.

In the previous chapter, pneumococcal targets SP2020 and Xisco were found to have the highest Youden's J statistics out of all of the PCR-based assays tested *in silico*. A LAMP assay targeting Xisco could not be designed due to the high sequence variability within the Xisco gene. Therefore, attention was redirected towards the SP2020 target and six LAMP primer combinations for SP2020 were tested *in vitro.* The presence of four to six primers per LAMP reaction as well as the large length of the FIP and BIP primers of a LAMP assay are known to enhance the likelihood of primer-primer hybridisation causing false positive LAMP assays [195]. Indeed, two assays (SP2020_2 and SP2020_5) showed amplification in the non-template control when incubating the assay for 60 minutes (Figure 3.2). The SP2020_3+ LAMP assay had a faster reaction time than the SP2020_3 assay, which is consistent with previous literature finding that the addition of loop primers accelerates a LAMP reaction [193,194].

The performance of the SP2020_3+ LAMP assay was compared to published pneumococcal LAMP assays *in silico* using two large and diverse bacterial genome datasets. *In silico* analyses showed that 99.6% (7,518/7,547) of genomes

in the pneumococcal study dataset were positive for the newly designed SP2020_3+ LAMP assay, whereas only 0.1% (1/1,825) of NPS genomes were positive. In the pneumococcal genome dataset, the predicted amplicon positivity of the published assays *psaA* (100%) and *ply* (99.8%) was higher than for SP2020_3+ and the predicted amplicon positivity for SPNA45_01710 (99.4%) and *lytA* (77.4%) was lower than SP2020_3+ (Table 3.4). However, the large number of pneumococcal genomes testing *in silico* negative for *lytA* might be an artefact of genome assembly errors due to *lytA* homologues. All of the four published pneumococcal LAMP assays assessed *in silico* were positive in more NPS genomes than SP2020_3+ (range 19-270 positive NPS genomes per published LAMP assay). Notably, the *psaA* LAMP assay was positive *in silico* for 270 NPS genomes of 10 *Streptococcus* species (Table 3.6), whilst the *psaA* PCR assay was positive for 17 genomes of two *Streptococcus* species (Figure 2.6). This difference could have arisen due to the absence of overlap between the *psaA* PCR primer/probe and LAMP primer binding regions; instead they bind to opposite ends of the *psaA* gene (Figure 3.3).

The new SP2020_3+ LAMP assay had a detection limit of 65 pneumococcal copies per reaction (Figure 3.7, reaction volume 25 µL); however, in chapter 2, the SP2020_new qPCR assay could consistently detect down to two pneumococcal copies per reaction (Figure 2.7, reaction volume 20 µL). This contrasts with existing literature, which suggests that LAMP assays can detect lower pneumococcal genome copy numbers than PCR assays [199,200]. For example, the published *lytA* LAMP assay had a reported detection limit of 10 genome copies per reaction (25 µL) [199,200].

The performance of the SP2020_3+ LAMP assay was compared to a published pneumococcal PCR assay targeting SP2020 *in vitro* using three panels of bacterial isolates. When testing the *in vitro* sensitivity and specificity panels, the SP2020_3+ LAMP assay performed as well as the published PCR assay, finding a sensitivity and specificity of 100% for each. The amplicon diversity panel tested the performance of SP2020_3+ across 15 different amplicons, finding the LAMP assay gave fast positive results for all unique amplicon sequences (Figure 3.7), which aligns with the objective of the WHO to create rapid and robust assays [209].

LAMP assays might provide benefits over PCR assays when testing clinical samples directly, since LAMP assays have been shown to tolerate biological substances better than PCR assays [210]. Therefore, future work could focus on using the SP2020_3+ primers to test CSF, serum or saliva directly. In addition, future work could integrate the SP2020_3+ primers into a point-of-care microchip platform, as was done for a pneumococcal *ply* assay [204].

Limitations of this research included that, despite the premise that LAMP assays are more accessible and equipment-free than PCR assays, in this study the *in vitro* testing of LAMP assays was done using a QuantStudio 5 machine. A QuantStudio 5 machine is equipped with the capability to perform thermocycling and measure fluorescence in real-time, of which only the fluorimeter functionality was employed for the LAMP assays. Future work should test the newly designed SP2020_3+ primers using more accessible laboratory equipment such as a simple heat block.

Strengths of this research included the use of large-scale genomic datasets to assess the performance of the new SP2020_3+ LAMP primers *in silico,* and to create a small SP2020 amplicon diversity panel for *in vitro* testing that represented the pneumococcal sequence diversity found in thousands of genomes, which saved time and reagents in the laboratory. *In vitro* real-time detection was facilitated through the addition of SYBR Green and showed that the new LAMP assay SP2020_3+ could rapidly detect pneumococci. Overall, a highly sensitive, specific and rapid LAMP assay for the identification of *S. pneumoniae* was developed, which could be particularly beneficial in low-resource settings.

Chapter 4

Understanding differences between species in the *Streptococcus* genus using *in silico* methods

Publication

The figures, tables and text in this chapter will form the basis of a future publication, but have not been submitted to date.

4.1 Introduction

4.1.1 The *Streptococcus* genus

The List of Prokaryotic Names with Standing in Nomenclature (LPSN) records 121 validly published distinct *Streptococcus* species (https://lpsn.dsmz.de/genus/streptococcus, accessed in June 2024) and accurate differentiation of these species is necessary for a range of applications [211]. Streptococcal species vary in their ability to cause disease [19]. Many streptococcal species are commensals of the physiological microbiome of humans and animals [8,19]. Whilst most *Streptococcus* species rarely cause infections, *S. agalactiae, S. pneumoniae* and *S. pyogenes* are important human pathogens [8,19]. Defining streptococci to a species level is key to understanding their epidemiology, designing public health interventions and monitoring AMR and vaccine impact [8]. Furthermore, some streptococcal species have industrial applications, such as *S. thermophilus* in the production of fermented dairy products and faecal streptococci as contamination indicators, making correct identification critical to ensure food safety and for environmental management [7]. Differentiating streptococci is also of academic interest to study evolution, horizontal gene transfer and interspecies differences across the streptococcal population structure [84].

4.1.2 Phenotypic methods to differentiate streptococci

The definitions of streptococcal species boundaries have been dependent upon the typing methods available. Phenotypic typing of bacterial isolates includes examining morphological differences of cells under a microscope or of bacterial colonies on growth media; using biochemical assays to investigate metabolic and enzymatic activity; and conducting serological tests to detect surface-exposed

antigens [8]. Various phenotypic assays have been designed to categorise streptococci, but these often cannot differentiate to a species level [10]. Blood agar culturing allows stratification of streptococci into three haemolytic groups: α (partial haemolysis), β (complete haemolysis) and γ (no haemolysis) [9,10]. However, differentiation by haemolysis is limited and can vary to some degree based on different culturing conditions [9,10].

The non-pneumococcal α-haemolytic streptococci are commonly referred to as the viridans group streptococci, as described in section 1.1.2 [9]. Viridans streptococci are typically commensals of the human upper respiratory tract and only occasionally cause invasive disease such as endocarditis, with immunocompromised persons at greater risk of infections [8,9]. Viridans streptococci can be distinguished from *S. pneumoniae* through investigation of optochin susceptibility and bile solubility, although anomalous results have been reported [8]. Further differentiation of viridans streptococci to a species-level has historically relied on biochemical assays, including testing for the hydrolysis of arginine, esculin and hippurate and fermentation of substrates such as arabinose, mannitol and sorbitol [8,16]. However, the speciation of viridans streptococci is challenged by the complexity of testing and the large amount of phenotypic similarity that these species exhibit [134]. Hence, viridans streptococci are often regarded as a single entity for clinical purposes, despite substantial diversity between species [212].

Many of the β-haemolytic streptococci can be stratified using Lancefield serotyping, but the carbohydrate antigen groups on which this typing is based do not always correlate to one species and some species can have multiple Lancefield serogroups [8]. Further biochemical assays have been developed to

differentiate streptococci, including carbohydrate fermentation tests and tolerance tests such as pH and salt concentration [8,213]. Additionally, MALDI-TOF MS, which has become the modality of choice for differentiating species in general in clinical microbiology, can accurately identify *S. agalactiae* [214], but gives highly similar protein mass spectra for *S. pneumoniae, S. mitis* and other viridans streptococci, leading to misidentification of these species [136].

4.1.3 Genotypic methods to differentiate streptococci

A range of *in silico* approaches have been used to define species boundaries. In 1987, it was suggested that bacteria are part of the same species if DNA-DNA hybridisation (DDH) relatedness is ≥70% and if the difference in DNA melting temperatures is <5 °C [215]. Later, a similarity threshold of <98.7% of the small subunit (16S) rRNA gene was suggested to be the cut-off for the definition of a novel bacterial species [216]. As WGS became more widely available, sequence-derived metrics such as the average nucleotide identity (ANI) were shown to be able to replace DDH to accurately investigate species relationships, with species often defined as sharing ≥95% ANI [217,218]. In 2012, Jolley et al. created the rMLST scheme, targeting 53 ribosomal protein subunit (*rps*) genes present throughout bacteria, to differentiate species [89].

4.1.3.1 Single gene approaches

To overcome some of this complexity in defining species in general, investigators have sought to develop single gene assays to differentiate streptococci [219–223]. The ideal single locus typing target should be present throughout the genus in a single copy [219,224]. Whilst nucleotide sequences should be conserved within a species, interspecies variability should be sufficiently large to allow differentiation

[219,224]. Various single gene targets have been suggested in literature for the differentiation of streptococcal species (Table 4.1) [219–223].

Table 4.1 Published single gene typing targets [219–223].

Target gene	Full name/product
groEL[a]	Heat shock protein
gyrB[a]	B subunit of DNA gyrase
recA[a]	Subunit of bacterial recombinase
rnpB[b]	RNA subunit of endonuclease P
rpoB[a]	Beta subunit of RNA polymerase
rpsB[a]	Ribosomal protein S2
sodA[a]	Manganese-dependent superoxide dismutase
tuf[a]	Elongation factor
16S rRNA[b]	Small subunit of ribosomal RNA

[a]Protein-encoding genes.
[b]The *rnpB* and 16S rRNA genes encode RNA molecules, not proteins.

The 16S rRNA gene was amongst the first genes to be used to differentiate species in the *Streptococcus* genus [225]. Kawamura et al. (1995) used 16S rRNA typing to classify the viridans streptococci into five phylogenetic groups, named anginosus, bovis, mitis, mutans and salivarius [222]. Facklam (2002) built on this method and defined the sanguinis group, which includes species initially defined as part of the mitis group, as a separate group within the viridans streptococci, whilst excluding the bovis group from the viridans streptococci and instead regarding these streptococci as a distinct entity [8,9]. Pyogenic streptococci form another group separate from the viridans streptococci based on 16S rRNA sequences [222].

The >99% sequence homology across the 16S rRNA gene alleles found in various streptococci (*S. mitis, S. pneumoniae, S. pseudopneumoniae* and *S. oralis*) meant other gene targets were explored to allow improved differentiation [9]. By

pyrosequencing two regions of the *rnpB* gene in 113 blood cultures, Innings et al. (2005) showed that the variable *rnpB* regions can be used to differentiate several *Streptococcus* species, but not the anginosus group of viridans streptococci [223]. Glazunova et al. (2009) compared the ability of 16S rRNA, *groEL, gyrB, rpoB* and *sodA* to differentiate 65 strains of 58 *Streptococcus* species and concluded that *groEL* could best identify species [219]. Using a panel of 21 isolates of 15 *Streptococcus* species and five isolates of different *Enterococcus* species, Kosecka-Strojek et al. (2020) compared the ability of 16S rRNA, *recA, rpoB, sodA,* and *tuf* to discriminate species, finding that 16S rRNA and *rpoB* performed best, whilst *recA* was able to differentiate viridans streptococci [220]. These earlier small-scale studies highlight some of the challenges researchers faced before the widespread availability of WGS. Wyllie et al. (2017) repurposed streptococcal genomes and found that the *rpsB* gene, which is one of the 53 targets of the rMLST scheme, could be used to differentiate streptococcal species [221]. Whilst single locus typing methods provided a major advance in the classification and stratification of *Streptococcus* species, they are vulnerable to the potential for distortion of results due to events such as gene loss, horizontal gene transfer and gene duplication [219]. As WGS methodology has become faster, more accurate and less expensive, we can now characterise and differentiate bacteria with higher resolution by increasing the number of loci used to type [112].

4.1.3.2 Core genome approaches

The definition of a core genome is the set of genes present in all, or nearly all, genomes of a species or genus, often applying a threshold of ≥95% presence. There are different ways in which core genes can be identified from a genome dataset. The Genome Comparator tool built into the $BIGS_{DB}$ platform performs a

unidirectional nucleotide BLAST (BLASTN) search of a set of reference genes against the genomes of interest, investigating their presence or absence [73]. Other bioinformatics tools infer clusters of orthologous genes, using triangles of pairwise best hits (BeTs) or a Markov Clustering algorithm (MCL) [226]. Examples of bioinformatics tools that define clusters of orthologous genes in prokaryotic genomes are chewBBACA [227], OrthoMCL [228], panX [229], Panaroo [226], PIRATE [230], PPanGGoLiN [231], PEPPAN [232] and Roary [233].

Streptococcal core genome analyses have allowed new demarcations of species groups within the genus. Gao et al. (2014) clustered 138 streptococcal genomes based on 278 orthologous genes and suggested they can be grouped into seven species groups which form two populations: the pyogenic, bovis, mutans and salivarius groups in one population, and the mitis, anginosus and unknown (containing *S. suis*) group in another population [234]. Richards et al. (2014) used 46 WGS of 44 streptococcal species to investigate their evolution and differentiated the streptococci into eight different groups: anginosus, bovis, downei, mitis, mutans, pyogenic, salivarius and sanguinis [235]. Examining 70 genomes from across the *Streptococcus* genus, Patel and Gupta (2018) suggest the genus can be divided into two major clades and 14 subclades, with six of these subclades – entericus, halotolerans, orisratti, porci, sobrinus and suis – being newly defined [236].

Genus-wide core genome typing schemes have been used to identify individual species, providing higher resolution differentiation than single locus typing. Rasmussen et al. (2016) showed that core genome phylogenetic analyses provided the most distinct clustering allowing differentiation of *S. cristatus, S. gordonii, S. infantis, S. mitis, S. oralis* and *S. sanguinis*, compared to single gene

(based on *recA, sodA* or *gdh)* or multilocus sequence analysis (MLSA, based on *rpoB, sodA, ppaC, pyk, map, pfl* and *tuf*) [237]. They used a dataset of 80 blood culture strains that had caused infective endocarditis, and defined 591 core genes across the six species [237]. In 2016, Jensen et al. investigated the taxonomic classifications of 195 genomes of mitis group streptococci using various genomic methods, including 16S rRNA and ANI - which both were unable to differentiate the various species - and MLSA and core genome clustering, which allowed differentiation [129]. Overall, they found that the core genome phylogenies provided higher resolution into species boundaries and used the scheme to suggest re-evaluated taxonomic definitions of some streptococci within the mitis group [129]. In line with this, Barajas et al. (2019) found that using genome similarity scores based on core genes outperformed both ANI and 16S rRNA gene phylogenies in differentiating 16 species from across the *Streptococcus* genus [238]. Notably, ANI was unable to distinguish the mutans and salivarius groups, and 16S rRNA clustering failed to differentiate *S. pneumoniae* and *S. pseudopneumoniae* [238]. Similarly, Jensen et al. (2021) showed that analyses based on a core genome of 590 genes and single nucleotide polymorphisms were able to differentiate *S. pseudopneumoniae* from other mitis group streptococci, whilst single gene targets could not [239]. In addition, core genome typing holds promise in identifying species within shotgun metagenomics samples, as demonstrated by Chalita et al. (2020) who employed a set of 92 *Streptococcus* core genes [240]. Core genome investigations have also suggested re-evaluation of subspecies definitions, for example within *S. mitis* and *S. oralis* [241]. Furthermore, streptococcal core genome investigations contribute to understanding evolution and interspecies recombination [242].

The increased availability of genomic datasets has led to the development of tools capable of handling thousands of genomes for the investigation of core genomes. Zhou et al. (2020) developed the PEPPAN tool to study bacterial pan and core genomes, applying it to a dataset of over 3,000 streptococcal genomes [232]. Building on this work, the same research group developed a hierarchical clustering (HierCC) method to define core genome clusters in large datasets and applied it to various genera [243]. For the *Streptococcus* genus, they included 84 species and determined a core genome of 372 genes, used as the basis for their cgMLST scheme [243]. Whilst HierCC assignments allowed high-resolution species differentiation within genera *Clostridioides, Escherichia, Salmonella, Vibrio* and *Yersinia*, for the *Streptococcus* genus HierCC did not align with taxonomic classifications [243]. This illustrates that the *Streptococcus* genus poses unique challenges in genomic classification. Recently, Kumar et al. (2023) used PEPPAN to cluster 138 genomes including 115 *Streptococcus* type strains, finding that the PEPPAN core genome phylogeny was better able to delineate species compared to 16S rRNA maximum likelihood phylogenies [244].

Several streptococcal core genome investigations have been performed; however, existing studies used small genomic datasets and/or limited subsets of the currently defined number of *Streptococcus* species. The ideal dataset for studying *Streptococcus* population structure would include as many different species as possible, whilst imposing an upper limit on the number of genomes per species to prevent major biases. Here, I compiled a large and diverse dataset of 1,919 high-quality genomes including 105 different *Streptococcus* species. Using this dataset, I constructed a *Streptococcus* genus-wide core genome phylogeny, providing a framework for mapping the presence of virulence factors.

4.1.4 Virulence factors

4.1.4.1 Antimicrobial resistance (AMR)

AMR determinants have been detected in a range of streptococcal isolates collected from animals and humans from across the world [245]. AMR presents a major challenge to effectively treating infections, which is especially relevant to *Streptococcus* species commonly causing invasive infections [245]. In *S. pneumoniae,* vaccination has led to an overall reduction in pneumococcal AMR [246]. However, an increase in macrolide and tetracycline resistance among non-vaccine serotypes has been observed following vaccine introduction, underscoring the need for ongoing AMR surveillance [246,247].

Whilst monitoring AMR in commonly pathogenic *Streptococcus* species is crucial, understanding AMR in viridans streptococci is important too. Although viridans streptococci are mainly asymptomatically carried in the human oral cavity, they can occasionally cause severe invasive diseases such as endocarditis and bloodstream infections, particularly in immunocompromised patients [8]. Therefore, understanding their antimicrobial susceptibility and resistance is crucial to treat these potentially life-threatening infections effectively. A recent study showed an extremely high presence of AMR genes in viridans streptococci collected from the sputum of patients on prophylactic antibiotics [248].

Viridans streptococci can also act as reservoirs of AMR genes, transmitting AMR determinants to *S. pneumoniae* through HGT whilst co-colonising the same niche [249]. Commensal streptococci, notably *S. mitis,* can provide *pbp* genes that confer beta-lactam resistance to *S. pneumoniae* [250–252]. D'Aeth et al. (2021) showed that inter-species recombination facilitates the acquisition of AMR determinants by *S. pneumoniae* from *S. oralis* and *S. mitis*, e.g. transposon

Tn1207.1 which imparts macrolide resistance or Tn916 which imparts tetracycline resistance [249]. Other studies have identified *S. salivarius* as another reservoir of AMR determinants [253]. In this chapter, I will investigate the presence of *in silico* AMR determinants across the large and diverse genomic *Streptococcus* dataset.

4.1.4.2 Pneumococcal capsular locus

The capsular polysaccharide synthesis (*cps*) locus comprises one of the principle virulence factors of *S. pneumoniae* and determines the pneumococcal serotype, which is targeted by pneumococcal vaccines, as described in more detail in section 1.2.3 [254]. In turn, accurate pneumococcal serotyping is crucial for understanding pneumococcal epidemiology, evolution and monitoring vaccine impact [255]. The Quellung reaction is the gold-standard method for pneumococcal serotyping; however, disadvantages of this method include that antisera are expensive and that substantial expertise is required to interpret the results [254,256]. Therefore, various PCR-based assays have been designed to identify pneumococcal serotypes by targeting serotype-specific central genes of the *cps* locus providing a low-cost, easy and culture-independent method [254–259].

As noted by the CDC, several studies have identified the presence of the *S. pneumoniae cps* genes in NPS species [147,260–265]. This can pose challenges for the specificity of these PCR-based serotyping assays, particularly in nasopharyngeal carriage studies where other oral streptococci are expected to be present. In this chapter, I will investigate the *in silico* presence of *S. pneumoniae* serotype/serogroup PCR amplicons in NPS genomes and map their distribution along a core genome phylogeny.

4.1.5 Chapter aims

The increased availability of WGS data allows the assessment of genotyping methods to study the differences between *Streptococcus* species and to examine taxonomic boundaries. In particular, this chapter applied a core genome clustering approach to study streptococcal population structure and investigate the presence of virulence factors across the phylogeny. Overall, the aims of this chapter were to:

- compile a high-quality dataset of publicly available assembled genomes including many different *Streptococcus* species, using rMLST speciation as the basis for comparison,
- evaluate differences in WGS properties between species such as guanine-cytosine (GC) content and total assembly length,
- evaluate species clustering based on published protein-encoding single locus typing targets,
- define the core genome of the *Streptococcus* genus using two contrasting approaches and a range of different settings,
- investigate the number of *Streptococcus* core genes and their function and evaluate the clustering of 105 different streptococcal species based on a core genome phylogeny,
- study the presence of AMR and pneumococcal *cps* genes across the *Streptococcus* genus, with a focus on the viridans streptococci.

4.2 Methods

4.2.1 Compilation of the *Streptococcus* genome dataset

Overall, the *Streptococcus* genome dataset contained 1,919 genomes from 105 different species. To create this dataset, all of the 75,077 publicly available assembled genome records which had their genus defined as "*Streptococcus"* from the PubMLST species identification database were considered for inclusion (https://pubmlst.org/species-id, accessed in February 2023) [73,89,157]. The 75,077 records were subjected to six genome quality control metrics. First, to exclude poor quality and overly-fragmented assemblies, e.g. arising from low quality sequencing or contamination, the number of contigs had to be ≤500. Similarly, second, the N50 statistic had to be ≥15,000. The N50 statistic represents the length of the contig at which the sum of all contigs of the same or larger length constitutes 50% of the total genome's assembly length [158]. Third, since the presence of multiple alleles per locus can be an indication of genome contamination, only one allele was allowed per rMLST locus, with the exception of rMLST loci BACT000014 (*rpsN*) and BACT000062 (*rpmG*), which are paralogous in streptococci [73,89,157]. Fourth, a complete rMLST allelic profile was required at all 53 loci, since missing alleles could be an indication of poor genome quality which could lead to problems when investigating the core genome. Fifth, the rMLST speciation had to return the same species description as was entered into the database based on the genotypic and/or phenotypic classification performed by the researchers who originally sequenced the genome [73,89,157]. Finally, the taxonomic classification had to be defined to a species level, meaning genomes labelled as "*Streptococcus* sp." were excluded. Of the genomes that passed all of the six quality control metrics, for each species, only one genome per unique

ribosomal sequence type (rST) and one genome per unique genome length in bp was selected, prioritising the genome with the lowest number of contigs to optimise genome assembly quality. This selection of unique rSTs and genome lengths was done to exclude potential duplicate genomes, such as type strains uploaded to the PubMLST platform from different reference databases and genomes from the same bacterial colony sequenced multiple times, e.g. in longitudinal studies. For the 11 species (*S. agalactiae, S. dysgalactiae, S. equi, S. mitis, S. mutans, S. oralis, S. pneumoniae, S. pyogenes, S. salivarius, S. suis* and *S. thermophilus*) which had over 100 genomes present, a maximum of 100 genomes were randomly selected. This maximum of 100 genomes per species was imposed to prevent skewing the downstream analyses, such as determination of the core genome, to one particular species by over-representing it.

4.2.2 rMLST species identification and clustering

Clustering based on the rMLST scheme was used as the benchmark for *Streptococcus* speciation, to which other *in silico* typing methods were compared [73,89,157]. The PubMLST species identification database (https://pubmlst.org/species-id) contains both a genome collection storing WGS assemblies and a typing database storing rMLST alleles linked to the species or genus in which they are found, compiled and curated by James Bray [73,89,157]. The genome collection was used to create the *Streptococcus* dataset (described in section 4.2.1) [73,89,157]. Speciation using rMLST was conducted by James Bray by examining the lowest common taxonomic node (LCTN) of each rMLST allele [73,74,89,157]. For each genome, the allelic designations of the 53 rMLST loci were recorded, and for each allele the LCTN was determined by referencing the typing database to identify which species also possessed the exact allele

[73,74,89,157]. Using a taxonomic tree, it is then verified if the LCTNs of different rMLST loci overlap [73,74,89,157]. Allele support for the genus or species node is calculated, where allele support >90% for one species is considered high confidence [73,74,89,157]. The nucleotide sequences of the 51 non-paralogous rMLST loci were exported from the PubMLST species identification database using allele sequences defined in the typing database. The genomes, sequence bin breakdown and provenance data were exported from the genome collection of the PubMLST species identification database [73,89].

4.2.3 Single locus sequence typing clustering

Seven protein-encoding single locus typing targets were identified from literature [219–221]. For each of the seven genes, the nucleotide sequence of pneumococcal reference genome TIGR4 - from the Nucleotide database of the National Center for Biotechnology Information (NCBI), accession AE005672 - was used to perform a BLASTN search against the 1,919 genomes of the *Streptococcus* dataset within the $BIGS_{DB}$ platform (Table 4.2) [69,266]. The BLASTN word size, which indicates the initial minimum identical BLASTN match from which the search is extended, was set to seven, instead of the standard word size of 11, to increase sensitivity.

Table 4.2 Protein-encoding single gene typing targets [69,219–221].

Target gene	TIGR4 locus tag[a]	Gene length (bp)
groEL	SP_1906	1,623
gyrB	SP_0806	1,947
recA	SP_1940	1,167
rpoB	SP_1961	3,612
rpsB	SP_2215	780
sodA	SP_0766	606
tuf	SP_1489	1,197

[a]Pneumococcal reference genome TIGR4, accession number AE005672, from the Nucleotide database of the National Center for Biotechnology Information (NCBI) [69].

4.2.4 Investigation of the *Streptococcus* core genome

4.2.4.1 Unidirectional BLASTN search

In my first investigation of the core genome, I started with a widely used pneumococcal reference genome and evaluated how many of its genes are shared across the whole *Streptococcus* genus. The nucleotide sequences of the 2,125 coding regions of pneumococcal reference genome TIGR4 were entered into the Genome Comparator module in the PubMLST species identification database supported by $BIGS_{DB}$ software to perform a unidirectional BLASTN search against the 1,919 genomes of the *Streptococcus* dataset [167,73,70,69]. The BLASTN word size was set to seven, and core genome inclusion threshold to 95%. The BLASTN parameters of sequence identity and alignment length were varied and incomplete and paralogous loci were excluded (see Table 4.3 for settings used). The Genome Comparator module also calculated mean *p*-distances [73]. The protein functions of nucleotide sequences of the 2,125 coding regions of pneumococcal reference genome TIGR4 and of the nucleotide sequences of the core genomes defined by settings three to seven were predicted using eggNOG-mapper (v2; against eggNOG database 5.0) [267,268], assigning COG-terms from the National Library of Medicine (https://www.ncbi.nlm.nih.gov/research/cog/) [269–272].

Table 4.3 Settings used for Genome Comparator and Panaroo [73,226].

Setting	Genome Comparator		Panaroo	
	Sequence identity threshold	Length difference threshold	Sequence identity threshold[a]	Length difference threshold[b]
Setting 1	0.98	0.98	0.98	0.98
Setting 2	0.95	0.95	0.95	0.95
Setting 3	0.90	0.90	0.90	0.90
Setting 4	0.85	0.85	0.85	0.85
Setting 5	0.80	0.80	0.80	0.80
Setting 6	0.75	0.75	0.75	0.75
Setting 7	0.70	0.70	0.70	0.70

[a]Panaroo parameter ran as –c
[b]Panaroo parameter ran as –-len_dif_percent

4.2.4.2 PROKKA and Panaroo

Core genes, as defined by the 95% threshold, may not be present in absolutely every species and it is possible that some may be missing from the pneumococcal reference genome used in section 4.2.4.1. Hence, a second approach not dependent on the pneumococcal reference genome was trialled. The protein-coding regions of the 1,919 *Streptococcus* genome assemblies were predicted using prokaryotic genome annotation tool PROKKA (v.1.14.5) [273,274]. PROKKA annotates protein-coding regions in two steps: first it uses feature prediction tools such as Prodigal to identify the coordinates of gene sequences, and then it predicts protein annotations by assessing smaller to larger databases with known protein families hierarchically [273–275]. PROKKA was run with the genus defined as *Streptococcus* [273,274]. The output .gff files of PROKKA were used to investigate the *Streptococcus* pangenome using Panaroo (v.1.3.3) [226,276]. Panaroo is a pangenome clustering tool that uses a graph-based algorithm to create clusters of orthologous genes (COGs) [226,276]. Panaroo parameters were set as follows: “--clean-mode strict --remove-invalid-genes” and the default family

threshold of 0.7 was used. In addition, the Panaroo sequence identity and length difference thresholds were varied to create seven different settings (Table 4.3) [226,276]. Taking one nucleotide allele for each of the 412 core genes (defined by setting 5), the functional classifications of the core genome were predicted by eggNOG-mapper as described in section 4.2.4.1 [269–272].

4.2.5 Construction of phylogenetic trees

In this chapter, maximum likelihood phylogenies were constructed using a series of bioinformatics tools, similar to established pipelines [74]. First, nucleotide alignments were created using MAFFT (v.7.525), a multiple sequence alignment programme based on fast Fourier transform, using the progressive alignment method FFT-NS-2 [277]. For the rMLST trees, complete allele sequences of each of the 51 non-paralogous rMLST loci for all 1,919 *Streptococcus* genomes were aligned locus-by-locus using MAFFT before they were concatenated [277]. For the core genome trees, alignment and concatenation were performed using MAFFT using one of the Panaroo processing scripts [226,276]. In this alignment, if one of the 412 core genes defined by setting 5 of Panaroo was fragmented, the largest fragment was included in the alignment [226,276]. All initial phylogenetic trees were constructed using FastTreeMP (v2.1.11) applying a generalised time-reversible (gtr) model to infer a maximum likelihood phylogeny from the nucleotide multiple sequence alignment [278,279]. For the rMLST trees, ClonalFrameML (v.1.13) was used to reconstruct the trees accounting for recombination events, using the FastTreeMP tree file and MAFFT alignments of the individual genes in multifasta file format as the input files [86]. For the core genome tree, ClonalFrameML (v.1.13) was run using the FastTreeMP tree file and a concatenated MAFFT alignment as input files [86]. Although the latter approach,

used to reconstruct the core genome tree, has the potential to erroneously infer recombination events spanning multiple genes, it was adopted as some core genes were missing from some genomes, which prevented supplying each gene separately in a multifasta file format, as ClonalFrameML is unable to handle missing genes with this setup [86]. For the single locus typing methods, trees created using FastTreeMP were multifurcating since some sequences were shared by multiple genomes. Therefore, trees were not reconstructed using ClonalFrameML, which requires bifurcating trees as input files [86].

4.2.6 Investigation of the presence of pneumococcal virulence factors

4.2.6.1 qPCR assays for S. pneumoniae AMR genes

The qPCR assays recommended by the CDC for the detection of AMR genes in *S. pneumoniae* were obtained from the CDC website (https://www.cdc.gov/streplab/pneumococcus/resources.html, accessed in August 2023; Table 4.4) [147,280]. The ipcress software built into $BIGS_{DB}$ was used to investigate the amplicon presence in the 1,919 genomes of the *Streptococcus* dataset, allowing a maximum of four mismatches per primer binding region and with the maximum amplicon length set to three times the amplicon length published by Velusamy et al. (2020) [2,73,69,280]. Amplicon sequences were imported into Geneious Prime (v.2023.1.2) where duplicates were removed [117]. Unique amplicon sequences were aligned using the MUSCLE algorithm (with eight iterations) before probe binding regions were aligned [117,160]. A maximum of four mismatches per probe binding region was allowed for an *in silico* predicted positive result [117]. The results were used to create a presence-absence matrix.

Table 4.4 Primers and probes recommended by CDC for pneumococcal AMR detection.
Adapted from [147,280].

Target [Study]	Primer/ probe	Nucleotide sequence (5' to 3')	Size (bp)[a]
cat [280]	FP	GTGACAAGGGTGATAAACTCAAATAC	87
	RP	TGGCTCTAACTTATCCCAATAACC	
	Probe	ACTGGTTACAATAGCGACGGAGAGT	
ermB [280]	FP	GCAATTGCTTAAGCTGCCA	97
	RP	ATCTGGAACATCTGTGGTATGG	
	Probe	TGGTTTAGGATGAAAGCATTCCGC	
mef [280]	FP	GAGCTACCTGTCTGGATGATTATG	93
	RP	AAGTGGTGTAACCGCATTGA	
	Probe	TGTTTATCCGTAGCATTGGAACAGCT	
pbp2b[b] [280]	FP	GGCTGTTTGGACCATATAGGTATT	81
	RP	ACTCAGGCTTACGGTTCATTC	
	Probe	CAGAGCTTGGACCGCTGTGATAGG	
tetM [280]	FP	GCTTATACTATAGCCCTGTTAGTACC	104
	RP	TGGCTCTAACAATTCTGTTCCA	
	Probe	TCGGATGCTTGCTCCTATTGTATTGGA	

[a]Amplicon size according to Velusamy et al. (2020) [280].
[b]The *pbp2b* assay is the only assay where a positive result should be interpreted as susceptibility and negative result as resistance.
Note: CDC, Centers for Disease Control and Prevention; FP, forward primer; RP, reverse primer.

4.2.6.2 PCR assays for S. pneumoniae serotypes/serogroups

Forty-one standard PCR assays recommended by the CDC for the detection of *S. pneumoniae* serotypes/serogroups were obtained from the CDC website (https://www.cdc.gov/streplab/pneumococcus/resources.html, accessed in August 2023; Table 4.5) [256,257,255,254,258,259,147]. The *Streptococcus* dataset (n=1,919) contained 100 *S. pneumoniae* genomes, which were excluded for these analyses. The *in silico* presence of the standard PCR amplicons was investigated in all of the remaining NPS genomes in the dataset (n=1,819). The ipcress software, integrated within $BIGS_{DB}$, was used to assess the presence of the *in silico* amplicons, allowing up to four mismatches per primer binding region and

setting the maximum amplicon length to twice the published amplicon length found in *S. pneumoniae* (Table 4.5) [2,73].

All amplicon sequences were exported into Geneious Prime (v.2023.1.2), where duplicates were removed, the number of unique amplicons was counted, and the range of the amplicon lengths was recorded [117]. Assays that were predicted to be *in silico* negative for all 1,819 NPS genomes were excluded from further analyses. For the serotype/serogroup PCR assays with one or more *in silico* positive amplicon(s) among the NPS genomes, the amplicon positivity was used to create a presence-absence matrix. When visualising the results on a phylogenetic core genome tree, branches of NPS species without any *in silico* amplicons for the *S. pneumoniae* serotypes/serogroups were collapsed [281].

Table 4.5 Standard PCR primers recommended by the CDC for pneumococcal serotype/serogroup determination.
Adapted from [147].

Serotype/ serogroup [Study]	Gene	Primer	Nucleotide sequence (5' to 3')	Size (bp)
1 [256]	*wzy*	FP	CTCTATAGAATGGAGTATATAAACTATGGTTA	280
		RP	CCAAAGAAAATACTAACATTATCACAATATTGGC	
2 [257]	*wzy*	FP	TATCCCAGTTCAATATTTCTCCACTACACC	290
		RP	ACACAAAATATAGGCAGAGAGAGACTACT	
3 [256]	*galU*	FP	ATGGTGTGATTTCTCCTAGATTGGAAAGTAG	371
		RP	CTTCTCCAATTGCTTACCAAGTGCAATAACG	
4 [256]	*wzy*	FP	CTGTTACTTGTTCTGGACTCTCGATAATTGG	430
		RP	GCCCACTCCTGTTAAAATCCTACCCGCATTG	
5 [256]	*wzy*	FP	ATACCTACACAACTTCTGATTATGCCTTTGTG	362
		RP	GCTCGATAAACATAATCAATATTTGAAAAAGTATG	
6A/6B/6C/ 6D [256]	*wciP*	FP	AATTTGTATTTTATTCATGCCTATATCTGG	250
		RP	TTAGCGGAGATAATTTAAAATGATGACTA	
6C/6D [255]	*wciN*$_{\beta}$	FP	CATTTTAGTGAAGTTGGCGGTGGAGTT	727
		RP	AGCTTCGAAGCCCATACTCTTCAATTA	
7C/7B/40 [256]	*wcwL*	FP	CTATCTCAGTCATCTATTGTTAAAGTTTACGACGGGA	260
		RP	GAACATAGATGTTGAGACATCTTTTGTAATTTC	
7F/7A [257]	*wzy*	FP	TCCAAACTATTACAGTGGGAATTACGG	599
		RP	ATAGGAATTGAGATTGCCAAAGCGAC	
8 [257]	*wzy*	FP	GAAGAAACGAAACTGTCAGAGCATTTACAT	201
		RP	CTATAGATACTAGTAGAGCTGTTCTAGTCT	
9N/9L [254]	*wzx*	FP	GAACTGAATAAGTCAGATTTAATCAGC	516
		RP	ACCAAGATCTGACGGGCTAATCAAT	
9V/9A [257]	*wzy*	FP	GGGTTCAAAGTCAGACAGTGAATCTTAA	816
		RP	CCATGAATGAAATCAACATTGTCAGTAGC	
10A [256]	*wcrG*	FP	GGTGTAGATTTACCATTAGTGTCGGCAGAC	628
		RP	GAATTTCTTCTTTAAGATTCGGATATTTCTC	
10F/10C/ 33C [257]	*wzx*	FP	GGAGTTTATCGGTAGTGCTCATTTTAGCA	248
		RP	CTAACAAATTCGCAACACGAGGCAACA	
11A/11D [256]	*wzy*	FP	GGACATGTTCAGGTGATTTCCCAATATAGTG	463
		RP	GATTATGAGTGTAATTTATTCCAACTTCTCCC	
12F/12A/ 12B/44/46 [256]	*wzx*	FP	GCAACAAACGGCGTGAAAGTAGTTG	376
		RP	CAAGATGAATATCACTACCAATAACAAAAC	
13 [257]	*wzx*	FP	TACTAAGGTAATCTCTGGAAATCGAAAGG	655
		RP	CTCATGCATTTTATTAACCGCTTTTTGTTC	
14 [254]	*wzy*	FP	GAAATGTTACTTGGCGCAGGTGTCAGAATT	189
		RP	GCCAATACTTCTTAGTCTCTCAGATGAAT	
15A/15F [256]	*wzy*	FP	ATTAGTACAGCTGCTGGAATATCTCTTC	436[a]
		RP	GATCTAGTGAACGTACTATTCCAAAC	
15B/15C [256]	*wzy*	FP	TTGGAATTTTTTAATTAGTGGCTTACCTA	496
		RP	CATCCGCTTATTAATTGAAGTAATCTGAACC	

16F [257]	*wzy*	FP	GAATTTTTCAGGCGTGGGTGTTAAAAG	717
		RP	CAGCATATAGCACCGCTAAGCAAATA	
17F [256]	*wciP*	FP	TTCGTGATGATAATTCCAATGATCAAACAAGAG	693
		RP	GATGTAACAAATTTGTAGCGACTAAGGTCTGC	
18C/18F/ 18B/18A [256]	*wzy*	FP	CTTAATAGCTCTCATTATTCTTTTTTTAAGCC	573
		RP	TTATCTGTAAACCATATCAGCATCTGAAAC	
19A [258]	*wzy*	FP	GAGAGATTCATAATCTTGCACTTAGCCA	566
		RP	CATAATAGCTACAAATGACTCATCGCC	
19F [256]	*wzy*	FP	GTTAAGATTGCTGATCGATTAATTGATATCC	304
		RP	GTAATATGTCTTTAGGGCGTTTATGGCGATAG	
19Fvar [259]	*wzy*	FP	GACAATTCTGGTTGACTTGTTGATTTTG	585
		RP	CTACCAAATACCTCACCAGCTTCC	
20 [256]	*wciL*	FP	GAGCAAGAGTTTTTCACCTGACAGCGAGAAG	514
		RP	CTAAATTCCTGTAATTTAGCTAAAACTCTTATC	
21 [257]	*wzx*	FP	CTATGGTTATTTCAACTCAATCGTCACC	192
		RP	GGCAAACTCAGACATAGTATAGCATAG	
22F/22A [256]	*wcwV*	FP	GAGTATAGCCAGATTATGGCAGTTTTATTGTC	643
		RP	CTCCAGCACTTGCGCTGGAAACAACAGACAAC	
23A [257]	*wzy*	FP	TATTCTAGCAAGTGACGAAGATGCG	722
		RP	CCAACATGCTTAAAAACGCTGCTTTAC	
23B [257]	*wzx*	FP	CCACAATTAGCGCTATATTCATTCAATCG	199
		RP	GTCCACGCTGAATAAAATGAAGCTCCG	
23F [256]	*wzy*	FP	GTAACAGTTGCTGTAGAGGGAATTGGCTTTTC	384
		RP	CACAACACCTAACACTCGATGGCTATATGATTC	
24F/24A/ 24B [257]	*wzy*	FP	GCTCCCTGCTATTGTAATCTTTAAAGAG	99
		RP	GTGTCTTTTATTGACTTTATCATAGGTCGG	
31 [256]	*wzy*	FP	GGAAGTTTTCAAGGATATGATAGTGGTGGTGC	701
		RP	CCGAATAATATATTCAATATATTCCTACTC	
33F/33A/ 37 [256]	*wzy*	FP	GAAGGCAATCAATGTGATTGTGTCGCG	338
		RP	CTTCAAAATGAAGATTATAGTACCCTTCTAC	
34 [256]	*wzy*	FP	GCTTTTGTAAGAGGAGATTATTTTCACCCAAC	408
		RP	CAATCCGACTAAGTCTTCAGTAAAAAACTTTAC	
35A/35C/ 42 [257]	*wzx*	FP	ATTACGACTCCTTATGTGACGCGCATA	280
		RP	CCAATCCCAAGATATATGCAACTAGGTT	
35B [256]	*wcrH*	FP	GATAAGTCTGTTGTGGAGACTTAAAAAGAATG	677
		RP	CTTTCCAGATAATTACAGGTATTCCTGAAGCAAG	
35F/47F [256]	*wzy*	FP	GAACATAGTCGCTATTGTATTTTATTTAAAGCAA	517
		RP	GACTAGGAGCATTATTCCTAGAGCGAGTAAACC	
38/25F/ 25A [256]	*wzy*	FP	CGTTCTTTTATCTCACTGTATAGTATCTTTATG	574
		RP	ATGTTTGAATTAAAGCTAACGTAACAATCC	
39 [257]	*wzy*	FP	TCATTGTATTAACCCTATGCTTTATTGGTG	98
		RP	GAGTATCTCCATTGTATTGAAATCTACCAA	

[a]The CDC website states the 15A/15F assay amplicon length is 434 bp, but Pai et al. report it as 436 bp; thus, the latter was chosen [147,256].

Note: CDC, Centers for Disease Control and Prevention; FP, forward primer; RP, reverse primer.

4.2.7 Data analysis and visualisation

The Oxford Biomedical Research Computing (BMRC) cluster was used to run bash and Python scripts on high-performance batch clusters. Resulting data were analysed and visualised in R statistical software (v4.2.1; [170]) using various packages, including binom [171], ComplexUpset [172], cowplot [173], dplyr [174], forcats [282], ggbreak [283], ggplot2 [175], openxlsx [284], readr [176], tibble [177], tidyverse [178] and tidyr [285]. The unrooted phylogenetic trees were visualised and annotated using Interactive Tree of Life (iTOL; v.6) [159,281]. To annotate the 105 different *Streptococcus* species, visually distinct colours were found using an online colour generator [286] and if necessary Inkscape was used to add annotations and arrange multi-panel plots [287].

4.3 Results

4.3.1 Compilation of a *Streptococcus* genome dataset

The final *Streptococcus* genome dataset contained 1,919 genomes of 105 different species (Figure 4.1). From the initial 75,077 available genomes, 5,032 (6.7%) *Streptococcus* genomes failed one or more of the six genome quality control metrics and were excluded. The rMLST species identification check, which required the rMLST speciation to match the species originally assigned to the WGS data, resulted in the highest number of genomes failing (n=3,399). Of the remaining 70,045 genomes, unique rSTs and genome lengths were selected within species to remove potential duplicates.

Per species, the genome count ranged from one to 100 (Figure 4.2A). Since more than 100 genomes were available after the application of genome quality control metrics for *S. agalactiae, S. dysgalactiae, S. equi, S. mitis, S. mutans, S. oralis, S. pneumoniae, S. pyogenes, S. salivarius, S. suis* and *S. thermophilus*, 100 genomes were selected randomly. Of the other 94 species included, only five were represented by more than 50 genomes: *S. pseudopneumoniae* (n=85), *S. uberis* (n=82), *S. sanguinis* (n=68), *S. parasanguinis* (n=65) and *S. gordonii* (n=59) (Figure 4.2A). The 1,919 *Streptococcus* genomes had a low number of contigs (median=34, interquartile range (IQR) [12-72]) and a high N50 (median=182,831 bp, IQR [78,407-637,966 bp]) (Figure 4.2B-C). Isolates were sampled between 1930 and 2022, with 39.3% collected in recent years (2012-2022; n=754) and for 33.9% of isolates the sampling date was unknown (n=650) (Figure 4.2D). The isolates were sampled from 55 countries on six continents: Africa (n=77), Asia (n=315), Europe (n=446), North America (n=536), Oceania (n=86), South America (n=30) and genomes' location unknown (n=429).

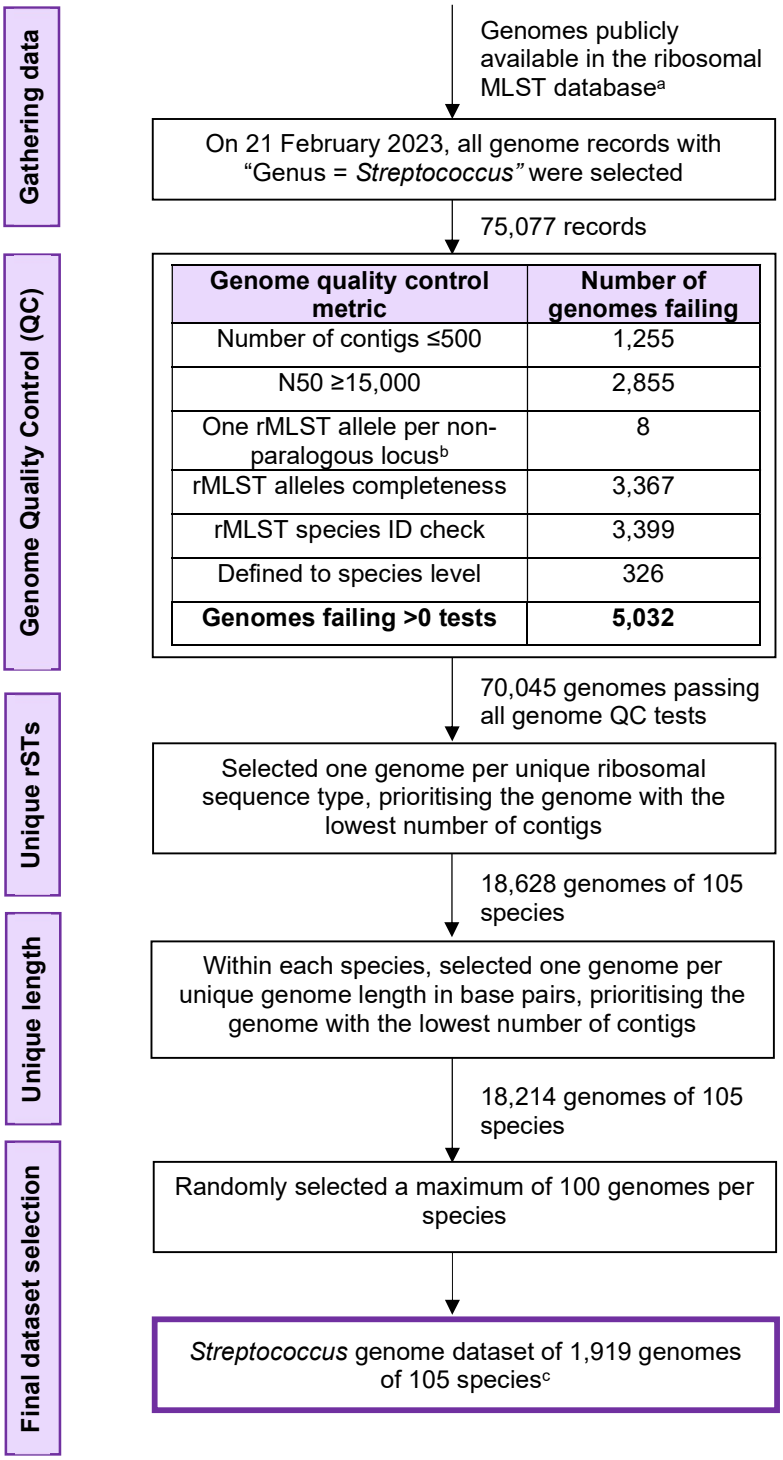


Figure 4.1 Schematic overview of the compilation of the *Streptococcus* genome dataset and application of genome quality control metrics.

[a]Assembled genomes were publicly available via https://pubmlst.org/rmlst [73,89,157].

[b]The two rMLST loci paralogous in streptococci are BACT000014 (*rpsN*) and BACT000062 (*rpmG*).

[c]*Streptococcus ursoris* was reclassified to *Streptococcus ratti* in the PubMLST species identification database in November 2023, resulting in a change in the number of different species from 106 to 105.

Note: ID, identification; QC, quality control; rMLST, ribosomal multilocus sequence typing.

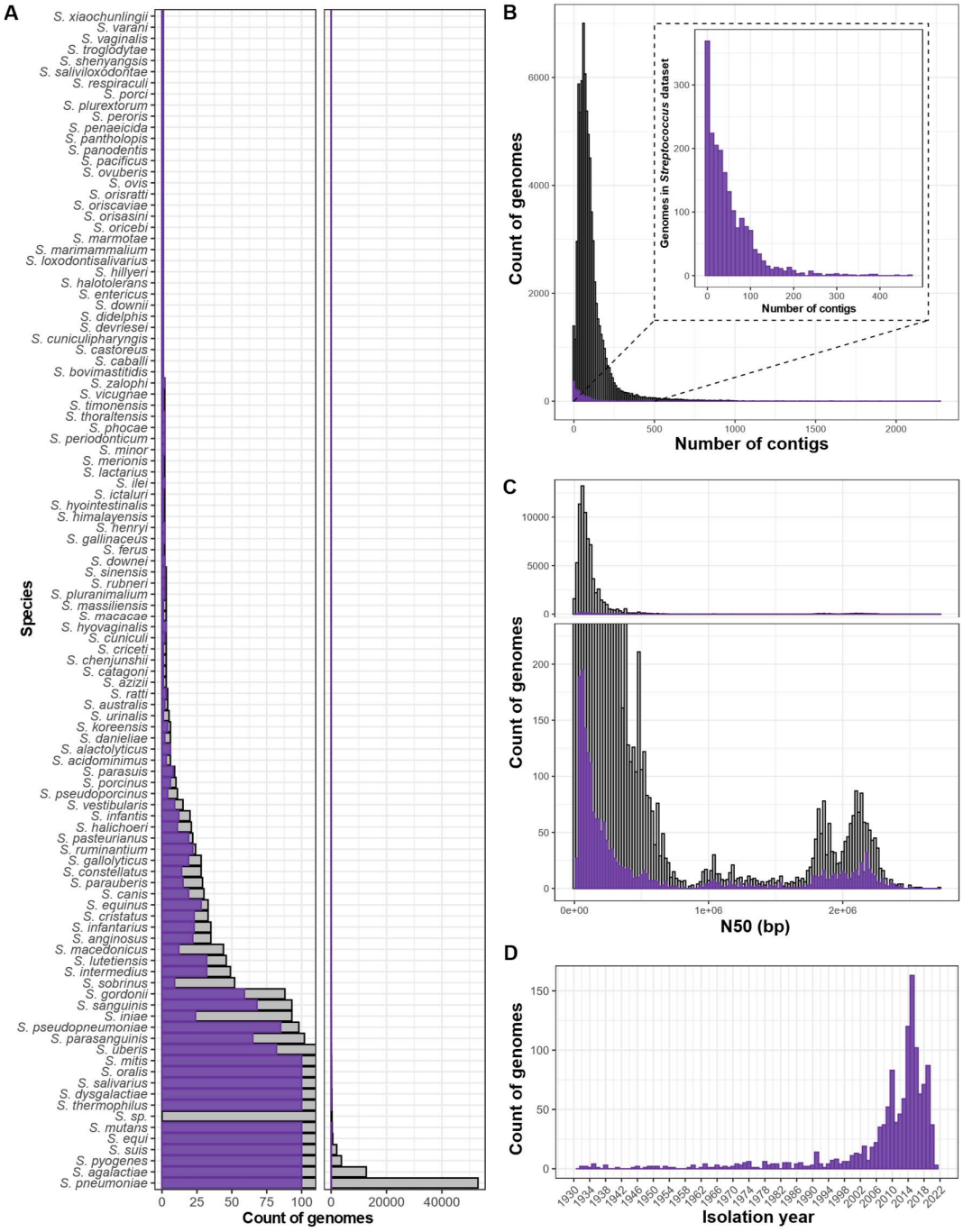


Figure 4.2 Descriptive metrics of the *Streptococcus* dataset.
(A-C) Genomes in the final *Streptococcus* dataset are shown in purple (n=1,919) out of all publicly available *Streptococcus* genomes considered for inclusion in grey (n=75,077). (A) Count of the 105 included *Streptococcus* species. Genomes labelled *"Streptococcus sp."* were not included in the final dataset. (B) Distribution of number of contigs per genome. The magnified panel shows the number of contigs of the genomes included in the *Streptococcus* genome dataset (n=1,919). (C) Distribution of N50 of genomes. The upper panel shows the complete distribution, whilst the lower panel provides a magnified view. (D) Distribution of isolation year of the 1,269/1,919 (66.1%) genomes for which this information was available.

4.3.2 rMLST clustering

A maximum likelihood phylogenetic tree was constructed using the 51 non-paralogous rMLST loci (Figure 4.3) [73,89,157]. Whilst isolates generally clustered by species, there were several species represented by a single genome that clustered within another species, such as *S. shenyangsis* within *S. mitis*, *S. downii* within *S. oralis* and *S. sinesis* within *S. cristatus*. The longest branch (0.0655 length) connected the *S. entericus* isolate to the tree.

S. pyogenes and *S. agalactiae,* which are the main human pathogens among the β-haemolytic streptococci, clustered separately and distantly on the rMLST tree. Instead, *S. pyogenes* clustered closely to *S. dysgalactiae,* similar to published rMLST phylogenies [89]. The clinical syndromes caused by *S. pyogenes* and *S. dysgalactiae* are also more similar to each other, e.g. skin and soft tissue infections, than those caused by *S. agalactiae* [8]. Notably, some *S. dysgalactiae* isolates clustered more closely with *S. pyogenes* than they did with some other isolates within the same *S. dysgalactiae* species.

Facklam (2002) categorised viridans streptococci into five groups: anginosus, mitis, mutans, salivarius and sanguinis [8]. As expected, on the rMLST tree, the mitis group species *S. pseudopneumoniae,* which rarely causes disease, clustered closely with major human pathogen *S. pneumoniae*. Despite their proximity, a distinct species boundary could be demarcated between *S. pseudopneumoniae* and *S. pneumoniae*. Conversely, *S. peroris* and *S. timonensis* clustered within *S. infantis* isolates on the tree. *S. cristatus,* part of the mitis group, clustered with *S. gordonii,* whilst the latter has been classed in the sanguinis group by Facklam (2002) [8].

Species within the anginosus group – *S. anginosus, S. constellatus* and *S. intermedius* – clustered together closely. These species, previously known as the milleri group, can cause bloodstream infections, empyemas and abscesses [288]. Furthermore, genomes of *S. vaginalis* and *S. periodonticum* clustered within *S. anginosus.* On the rMLST tree, the anginosus group species clustered closer to *S. gordonii* and *S. sanguinis,* than the latter species clustered to *S. parasanguinis* [8].

Within the mutans group, *S. mutans,* which resides in the human oral cavity and contributes to human dental caries [289], clustered together with species that can be found in rats (*S. ferus* and *S. sobrinus*) and in monkeys (*S. downei* and *S. macacae*) [8]. This highlighted close genetic relatedness in rMLST loci despite different host preferences.

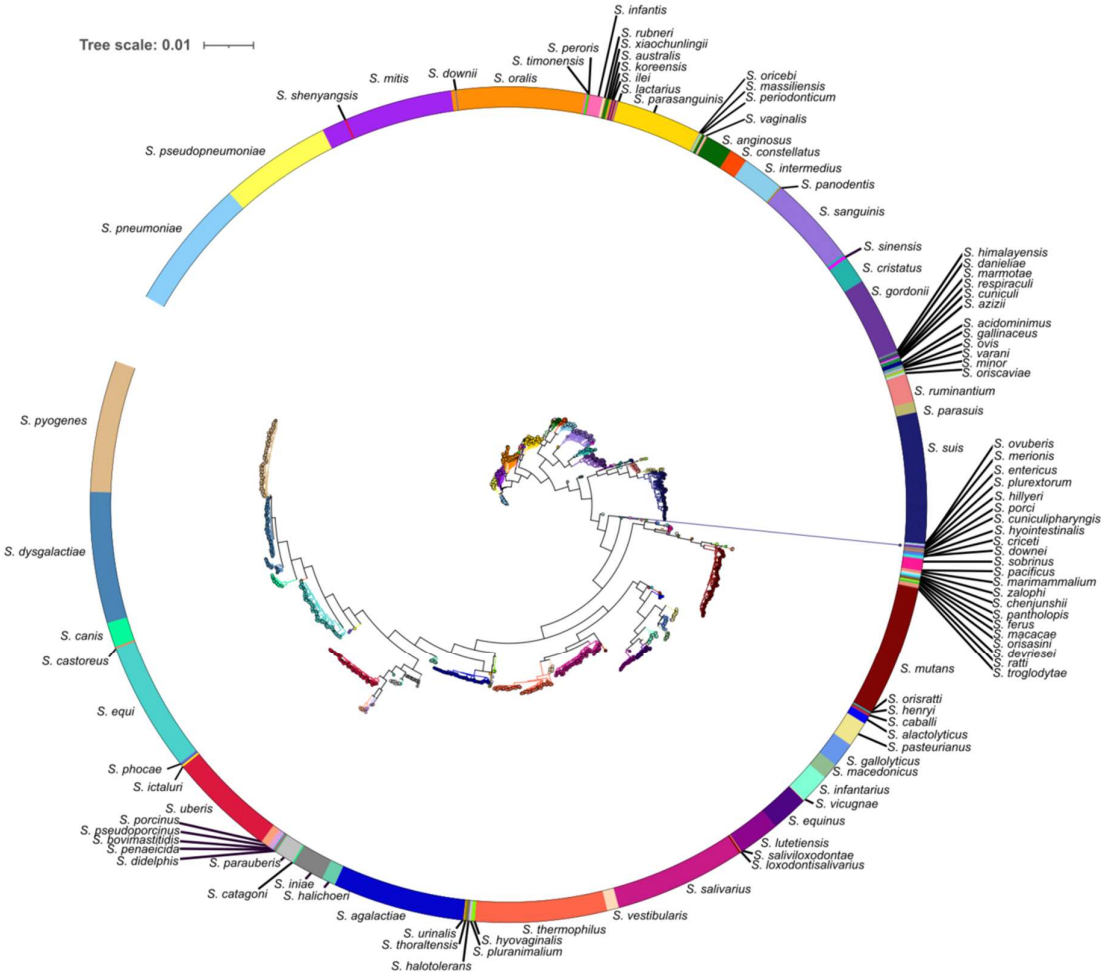


Figure 4.3 Maximum likelihood phylogeny based on nucleotide sequences of 51 non-paralogous rMLST loci of 1,919 *Streptococcus* genomes.
The outer ring and tips of branches were coloured according to the 105 different *Streptococcus* species, based on assignments made by those submitting the genome sequences to public repositories and confirmed by rMLST; all 105 species are also annotated [73,89].
Note: rMLST, ribosomal multilocus sequence typing; *S., Streptococcus.*

The GC-content and the total WGS assembly length were investigated across the *Streptococcus* species. Out of all of the 1,919 genomes, the single-genome representative of *S. marimammalium* had both the lowest GC-content (33.18%) and smallest total assembly length (1.50 Mb). The single-genome representative of *S. panodentis* had both the highest GC-content (47.56%) and the largest total assembly length (2.78 Mb).

For the streptococci of which more than 50 genomes were present per species in the dataset, the distribution of GC-content and total assembly length was investigated further (Figure 4.4). There was fairly minimal variation in GC-content within each species, apart from a single genome outlier for *S. suis*, but there were clear differences between species. Although the 16 species that were compared differed from each other, the numeric range of variation in GC-content was modest. Overall, the mean genome GC-content was lowest for *S. agalactiae* (mean=35.32, [min=35.08; max=35.76%]) and highest for *S. sanguinis* (mean=43.17, [min=42.82; max=43.44%]).

Within each of the 16 species, the total genome assembly length varied by up to 35% (for *S. suis*). This difference in genome length within species was higher than was seen in GC-content. *S. pyogenes* had the smallest mean total assembly length (mean=1.80 Mb, [min=1.69; max=1.93]), while *S. sanguinis* had the largest total assembly length (mean=2.37 Mb, [min=2.26; max=2.50]). Some of the variation in apparent genome length may arise from sequencing and assembly artefacts as well as true variation.

In literature, *S. pneumoniae* is noted to have a larger genome length compared to the typically non-pathogenic mitis group streptococci including *S. mitis*, attributed to *S. pneumoniae* retaining more virulence genes, resulting in greater pathogenic potential [20]. However, the assembly lengths of the 100 *S. pneumoniae* genomes (mean=2,093,435 bp, [min=1,975,277; max=2,238,886]) overlapped with the 100 *S. mitis* genomes (mean=2,005,847 bp, [min=1,819,806; max=2,247,624]) (Figure 4.4). In addition, the 85 *S. pseudopneumoniae* genomes had a higher mean assembly length (mean=2,202,803 bp, [min=2,077,325; max=2,365,106]) than the pneumococcal genomes.

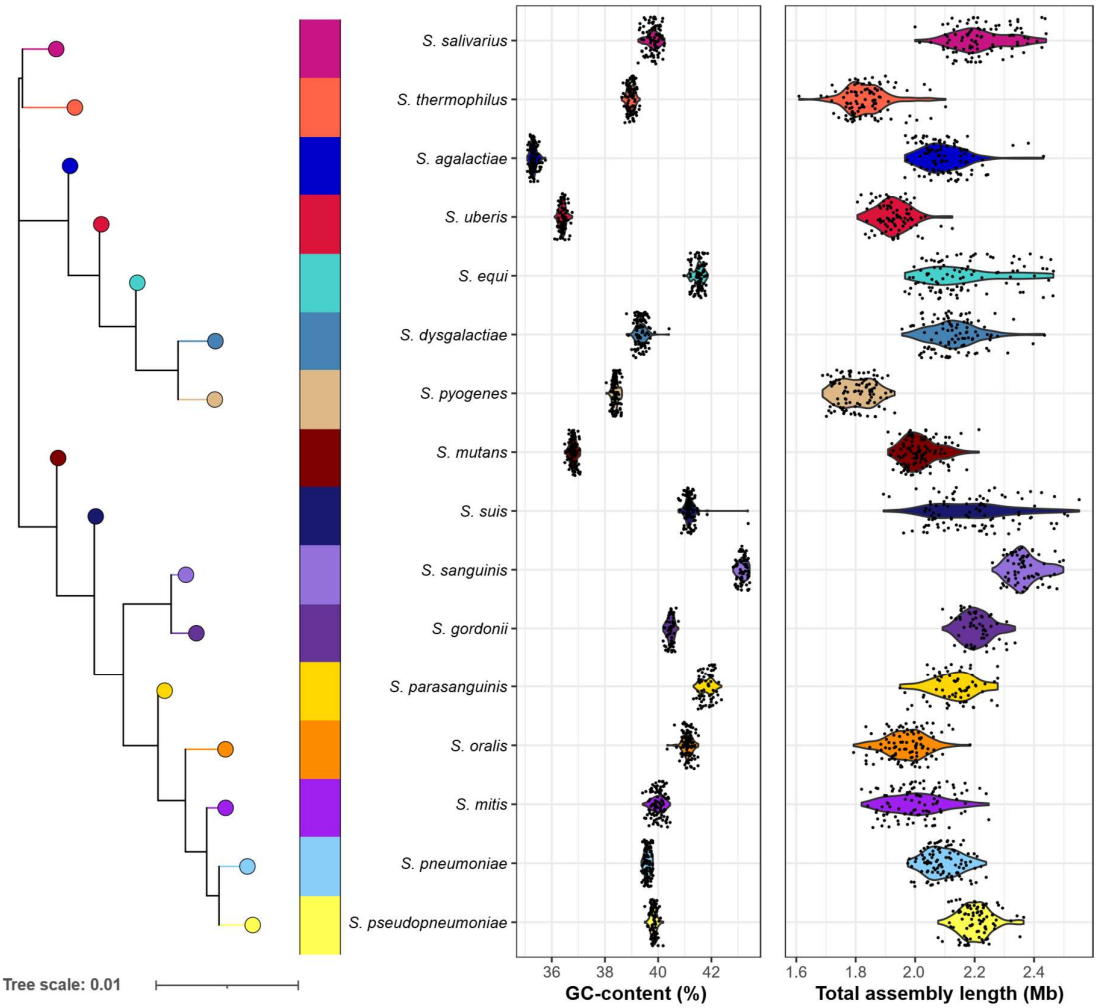


Figure 4.4 Distribution of GC-content and total assembly size of *Streptococcus* species.
Phylogenetic maximum likelihood tree based on 51 non-paralogous ribosomal multilocus sequence typing with one genome per species, for the 16 species which were represented over 50 times in the dataset.
Note: GC, guanine-cytosine; Mb, mega base pairs; *S., Streptococcus*.

To further investigate interspecies patterns of GC-content, in particular to understand if genome-wide differences in mean GC-content between species were driven by specific genes or variation across all genes, a sliding window approach (window size 2,000 bp, step size 500 bp) was used to calculate the GC-content along a single complete genome for *S. agalactiae, S. equi, S. gordonii, S. mitis, S. mutans, S. oralis, S. parasanguinis, S. pyogenes, S. sanguinis, S. suis, S. thermophilus* and *S. uberis* (Figure 4.5). Of these twelve genomes, *S. sanguinis* (2.39 Mb long) had the highest mean GC-content (43.40%) whilst *S. agalactiae* (2.17 Mb long) had the lowest mean GC-content (35.38%). The histograms, depicting the distributions of the GC-content of the sliding windows, displayed one prominent central peak for each species; however, these peaks were not symmetrical about the mean GC-content (Figure 4.5C). Furthermore, whilst the mean GC-content of *S. thermophilus, S. pyogenes, S. mutans, S. uberis* and *S. agalactiae* was below 40%, they each displayed a small increased frequency of fragments around 50% GC-content (Figure 4.5C). Overall, genome-wide variations in GC-content appeared to explain most of the between-species differences.

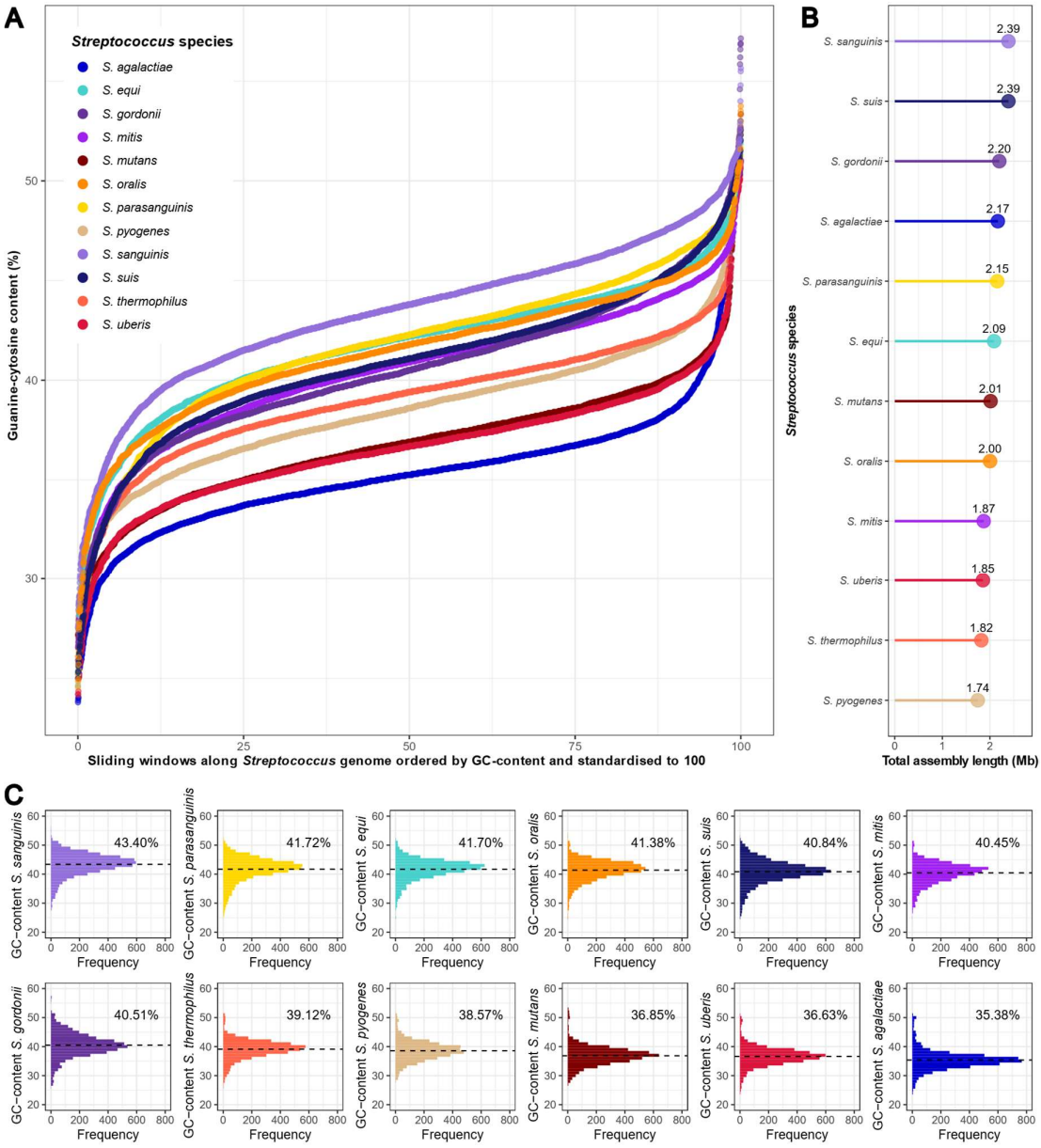


Figure 4.5 Differences in GC-content across *Streptococcus* species.
(A) Percentage of GC-content calculated over a sliding windows of genomic fragments (window size 2,000 bp, step size 500 bp) along one complete genome, for 12 *Streptococcus* species. The number of sliding windows was standardised to 100. The values found for each sliding window were aligned in ascending order of GC-content and coloured by the *Streptococcus* species. (B) Total assembly length of each of the complete *Streptococcus* genomes used to investigate the GC-content in panel A. (C) Histograms displaying the distribution of the GC-content of the 2,000 bp-long fragments. Horizontal dashed lines and annotations indicate the overall GC-content of the assembly.
Note: GC, guanine-cytosine; *S., Streptococcus.*

4.3.3 Single locus sequence typing clustering

For each of the seven published protein-encoding single locus typing genes, the gene sequence of *S. pneumoniae* reference genome TIGR4 was downloaded from NCBI and used for a BLASTN search against the 1,919 *Streptococcus* genomes [69]. The IQRs of the alignment length and percentage identity were narrow around the medians for each of the seven targets (Table 4.6). The median number of mismatches was highest for *rpoB*, which also had the longest median alignment length. Despite *sodA* having the smallest median alignment length, it did not have the lowest number of median mismatches; instead, this was observed for *rpsB* (Table 4.6).

Table 4.6 Presence of protein-encoding single locus typing genes in *Streptococcus* dataset (n=1,919).

	Alignment length (bp)		Identity (%)		Mismatches (bp)	
Target[a]	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
groEL	1,623	1,617-1,624	79.7	78.6-82.7	318	275-338
gyrB	1,950	1,947-1,955	77.9	77.3-80.3	420	378-434
recA	1,052	1,023-1,065	80.5	79.5-82.3	200	183-212
rpoB	3,534	3,530-3,573	84.2	83.1-86.1	556	493-594
rpsB	776	773-780	88.8	87.7-90.6	80	62-83
sodA	606	604-607	79.4	75.2-87.1	122	78-144
tuf	1,197	1,197-1,197	92.6	91.5-95.6	88	53-102

[a]Pneumococcal reference genome TIGR4, accession number AE005672, from the Nucleotide database of the National Center for Biotechnology Information (NCBI) [69].
Note: bp, base pairs; IQR; interquartile range.

Maximum likelihood phylogenetic trees were constructed for each of the seven published protein-encoding single locus typing genes (Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9). ClonalFrameML could not be used to account for recombination due to the multifurcating nature of the trees, resulting from identical

gene nucleotide sequences being found across multiple genomes [86]. In particular, the number of unique nucleotide sequences among the 1,919 genomes varied by gene: 1,117 for *groEL*; 1,168 for *gyrB*; 873 for *recA*; 1,320 for *rpoB*; 766 for *rpsB*; 830 for *sodA*; and 780 for *tuf.*

In each of the seven single locus typing trees, *Streptococcus* species were intertwined in clusters with other species. In particular, the genomes of *S. mitis, S. oralis, S. pneumoniae* and *S. pseudopneumoniae* were often intermixed on the phylogenies, as illustrated by the magnified tree views. Furthermore, the tree based on the *tuf* gene sequence showed two distinct and distant clusters within the *S. dysgalactiae* species (Figure 4.8C).

For three of the seven targets, a small number of genomes (one genome for *recA,* four for *rpoB,* and two for *sodA*) had >20% of the gene missing, likely due to genome assembly errors. However, with the exception of one outlier in an *S. pseudopneumoniae* genome on the *sodA* tree, these isolates clustered with their respective species on the tree. This highlighted the resilience of the FastTreeMP algorithm to clustering partial missing sequences causing gaps, which can occur in assembled short-read sequencing data [278].

In this chapter, BLASTN was used to search for and extract gene sequences from WGS assemblies [73,266]; however, single locus typing traditionally used PCR-based amplification of a part of the target gene before Sanger sequencing of the resulting amplicon. In order to mimic the latter process, the primer sequences designed by Wyllie et al. (2017) were used to extract the *rpsB* amplicon from the pneumococcal reference genome TIGR4 (similarly to described in section 4.2.6.1) [2,69,73,221]. The primer binding regions were removed from this *rpsB* amplicon

resulting in the *rpsB* amplicon fragment (Figure 4.9A). Subsequently, the *rpsB* amplicon fragment was used to perform a BLASTN search against the 1,919 genomes of the *Streptococcus* dataset, using the $BIGS_{DB}$ platform [69,266,73,221]. A maximum likelihood tree was created based on this *rpsB* fragment (Figure 4.9D and E). Overall, neither the full *rpsB* gene nor the *rpsB* amplicon fragment could adequately differentiate *S. mitis* and *S. pseudopneumoniae* genomes.

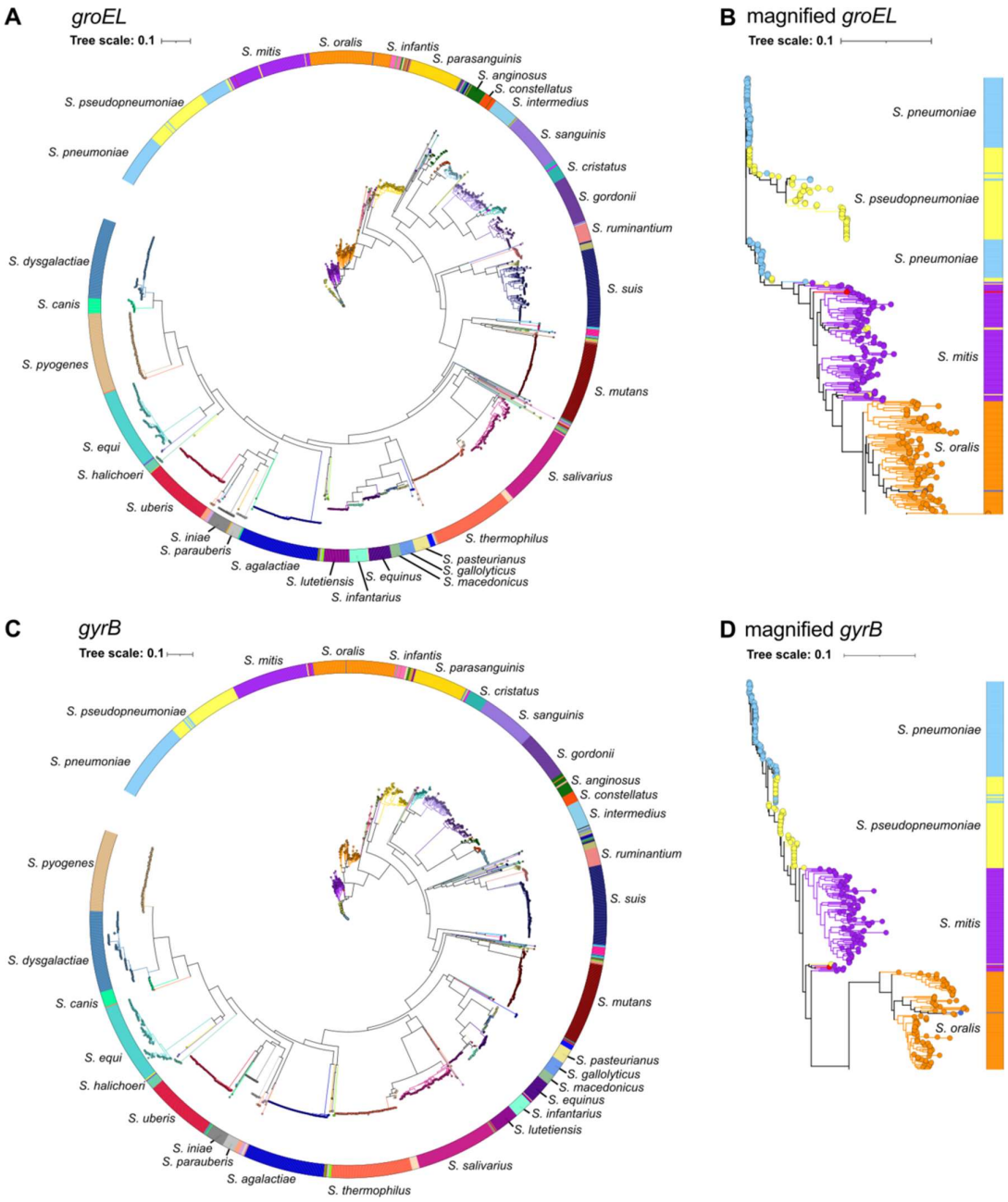


Figure 4.6 Maximum likelihood trees based on the *groEL* and *gyrB* gene sequences in the *Streptococcus* dataset (n=1,919).
The outer ring and tips of branches were coloured according to the 105 different *Streptococcus* species, based on assignments made by those submitting the genome sequences to public repositories and confirmed by rMLST [89,157].
(A, C) The 32 *Streptococcus* species of which more than 10 genomes per species were present in the dataset were annotated.
(A) Tree based on *groEL* gene sequence. (B) Magnified view of *groEL* tree.
(C) Tree based on *gyrB* gene sequence (D) Magnified view of *gyrB* tree.
Note: *S., Streptococcus.*

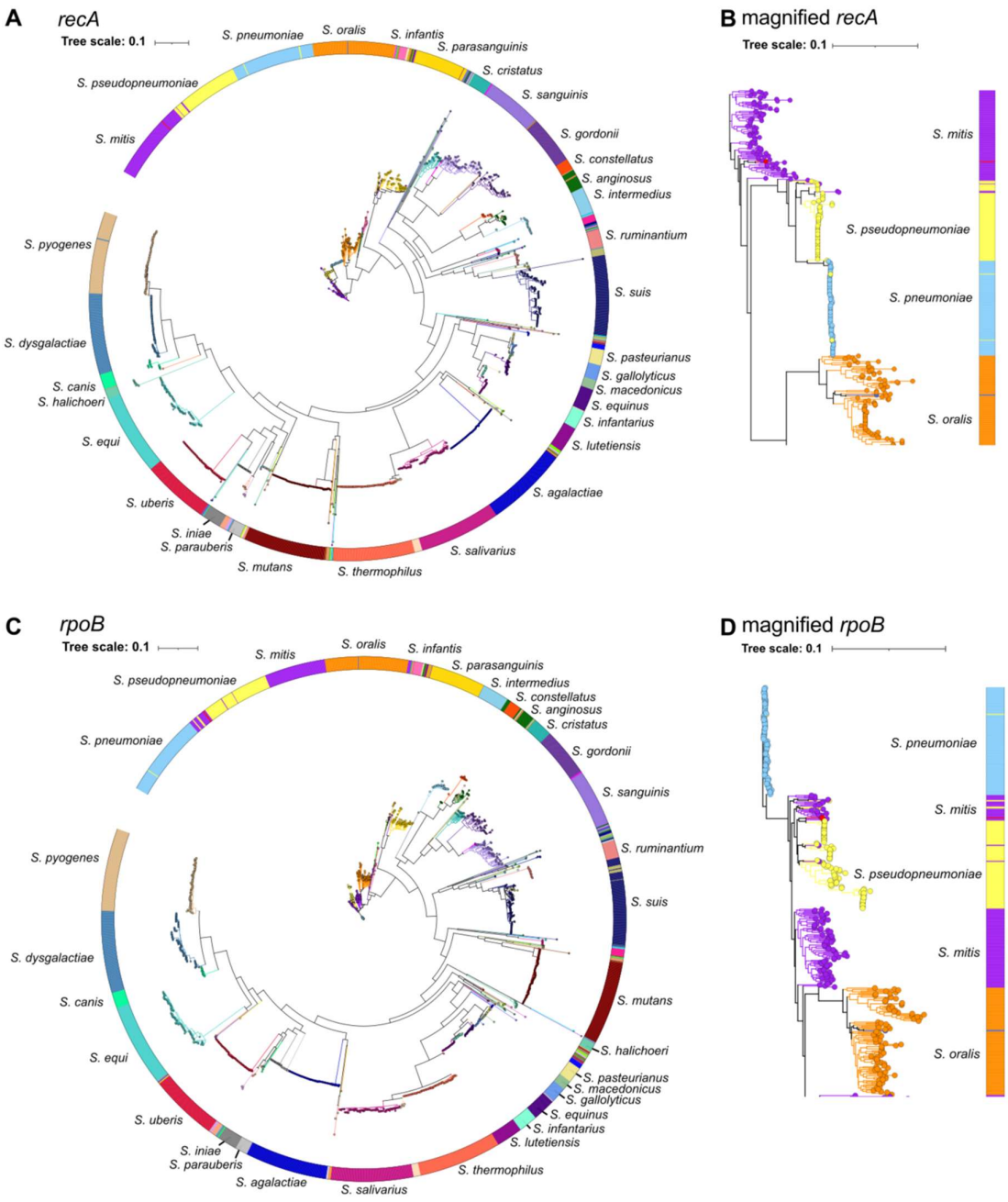


Figure 4.7 Maximum likelihood trees based on the *recA* and *rpoB* gene sequences in the *Streptococcus* dataset (n=1,919).
The outer ring and tips of branches were coloured according to the 105 different *Streptococcus* species, based on assignments made by those submitting the genome sequences to public repositories and confirmed by rMLST [89,157].
(A, C) The 32 *Streptococcus* species of which more than 10 genomes per species were present in the dataset were annotated.
(A) Tree based on *recA* gene sequence. (B) Magnified view of *recA* tree.
(C) Tree based on *rpoB* gene sequence (D) Magnified view of *rpoB* tree.
Note: *S., Streptococcus.*

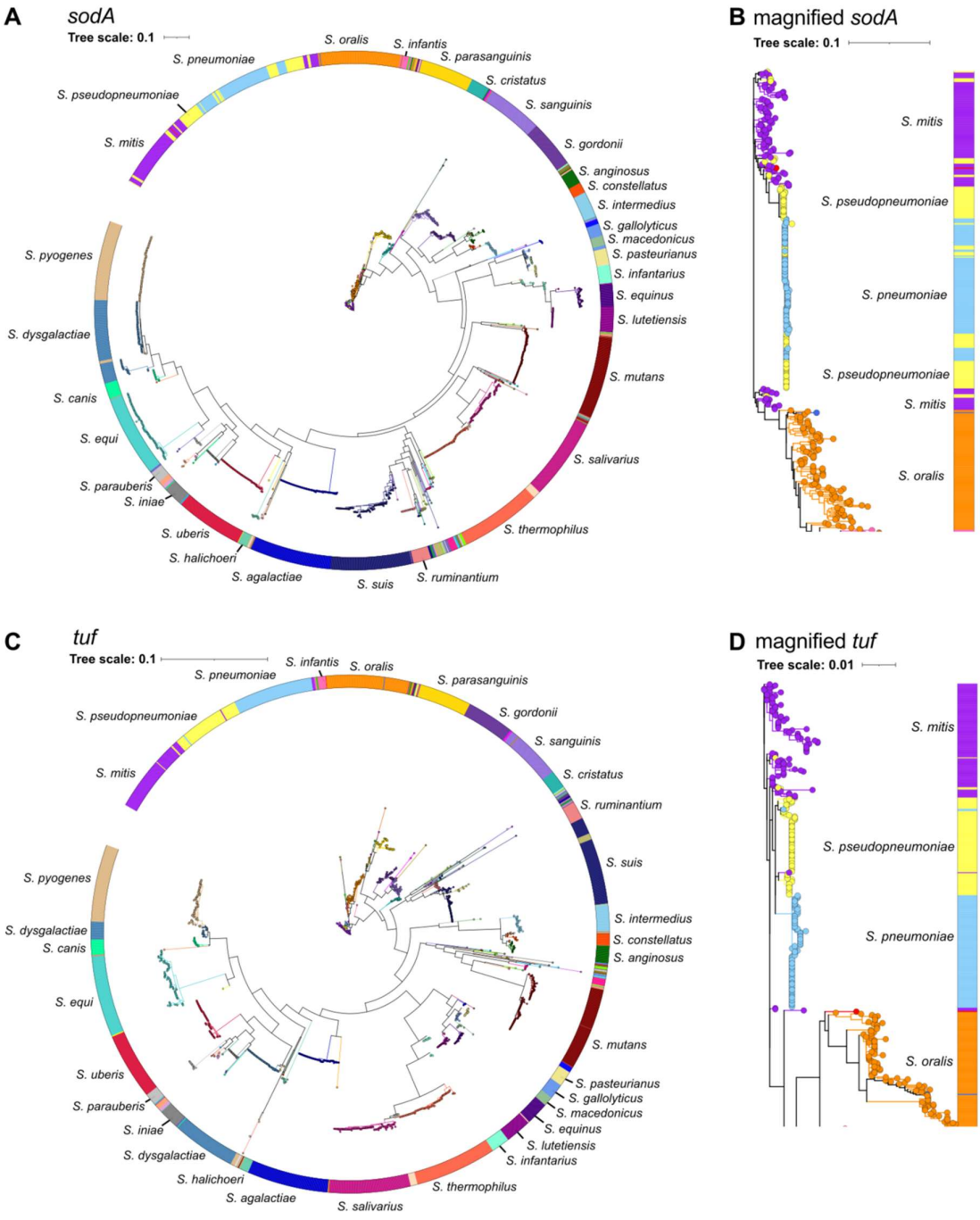


Figure 4.8 Maximum likelihood trees based on the *sodA* and *tuf* gene sequences in the *Streptococcus* dataset (n=1,919).
The outer ring and tips of branches were coloured according to the 105 different *Streptococcus* species, based on assignments made by those submitting the genome sequences to public repositories and confirmed by rMLST [89,157].
(A, C) The 32 *Streptococcus* species of which more than 10 genomes per species were present in the dataset were annotated.
(A) Tree based on *sodA* gene sequence. (B) Magnified view of *sodA* tree.
(C) Tree based on *tuf* gene sequence (D) Magnified view of *tuf* tree.
Note: *S., Streptococcus.*

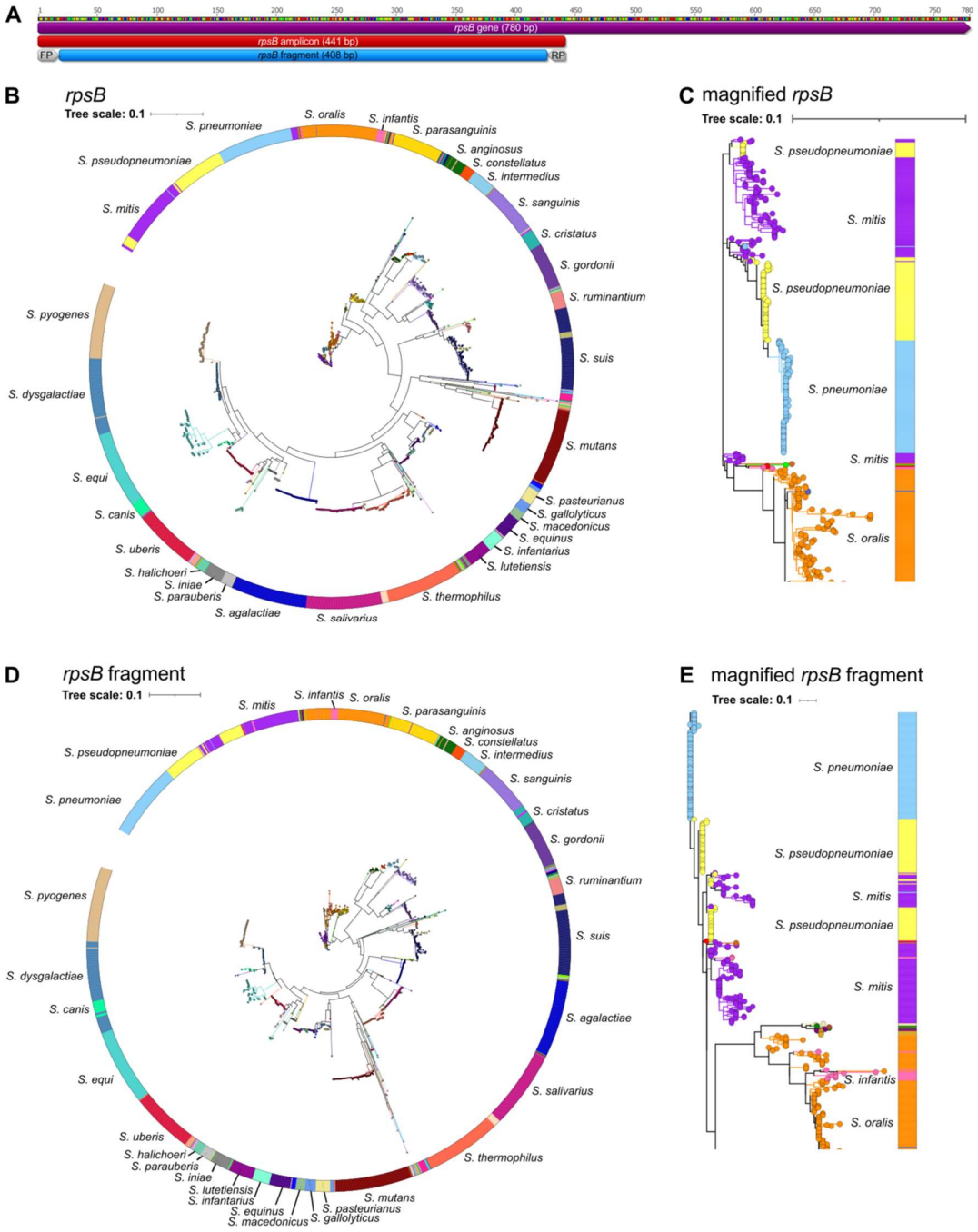


Figure 4.9 Maximum likelihood trees based on the *rpsB* gene and *rpsB* amplicon fragment of the *Streptococcus* dataset (n=1,919).
The outer ring was coloured based on species assignments made by those submitting the genome sequences to public repositories and confirmed by rMLST, and species with more than 10 genomes were annotated on the outer ring [89,157]. (A) Schematic overview of the *rpsB* gene (purple), amplicon (red), primers (grey) and amplicon fragment (blue). (B) Tree based on *rpsB* gene sequence. (C) Magnified view of *rpsB* gene tree. (D) Tree based on *rpsB* amplicon fragment. (E) Magnified view of *rpsB* amplicon fragment tree. Note: FP, forward primer; RP, reverse primer; *S.*, *Streptococcus*. Primer sequences were published by Wyllie et al. (2017) [221].

4.3.4 Investigation of the *Streptococcus* core genome

4.3.4.1 Unidirectional BLASTN search

The 2,215 coding regions of the pneumococcal reference genome TIGR4 had a total concatenated length of 1,808,970 bp, i.e. 83.7% of the complete circular assembly length of 2,160,842 bp [69]. The Genome Comparator module ran unidirectional BLASTN searches of these 2,125 coding sequences against the 1,919 *Streptococcus* genomes, using seven different settings which ranged from stringent requirements for a gene to be considered present (setting 1: 98% identity and 98% alignment length) to more lenient settings (setting 7: 70% identity and 70% alignment length). After defining the presence or absence of genes in all 1,919 *Streptococcus* genomes, the number of genes in the inner core (genes present in ≥99%-100% of genomes), soft core (genes present in ≥95%-<99% of genomes), shell (genes present in ≥15%-<95% of genomes) and cloud (genes present in >0%-<15% of genomes) were determined (Figure 4.10A). Additionally, the core genome was defined as the inner core and soft core genes combined, i.e. the genes present in ≥95% (≥1,824/1,919) of genomes in the *Streptococcus* dataset.

When using settings 1 and 2, the core genome included zero genes; however, as the required sequence similarity and alignment length became more lenient the core genome size increased to 266 genes (Figure 4.10B). As expected, the number of cloud genes decreased from 2,084 to 609 moving from setting 1 to 7. Whilst seven different settings were used to assess the presence of the 2,125 pneumococcal loci, settings 1 and 2 were not included in investigations into mean *p*-distances and functional classifications of loci, as they resulted in zero inner core and soft core genes.

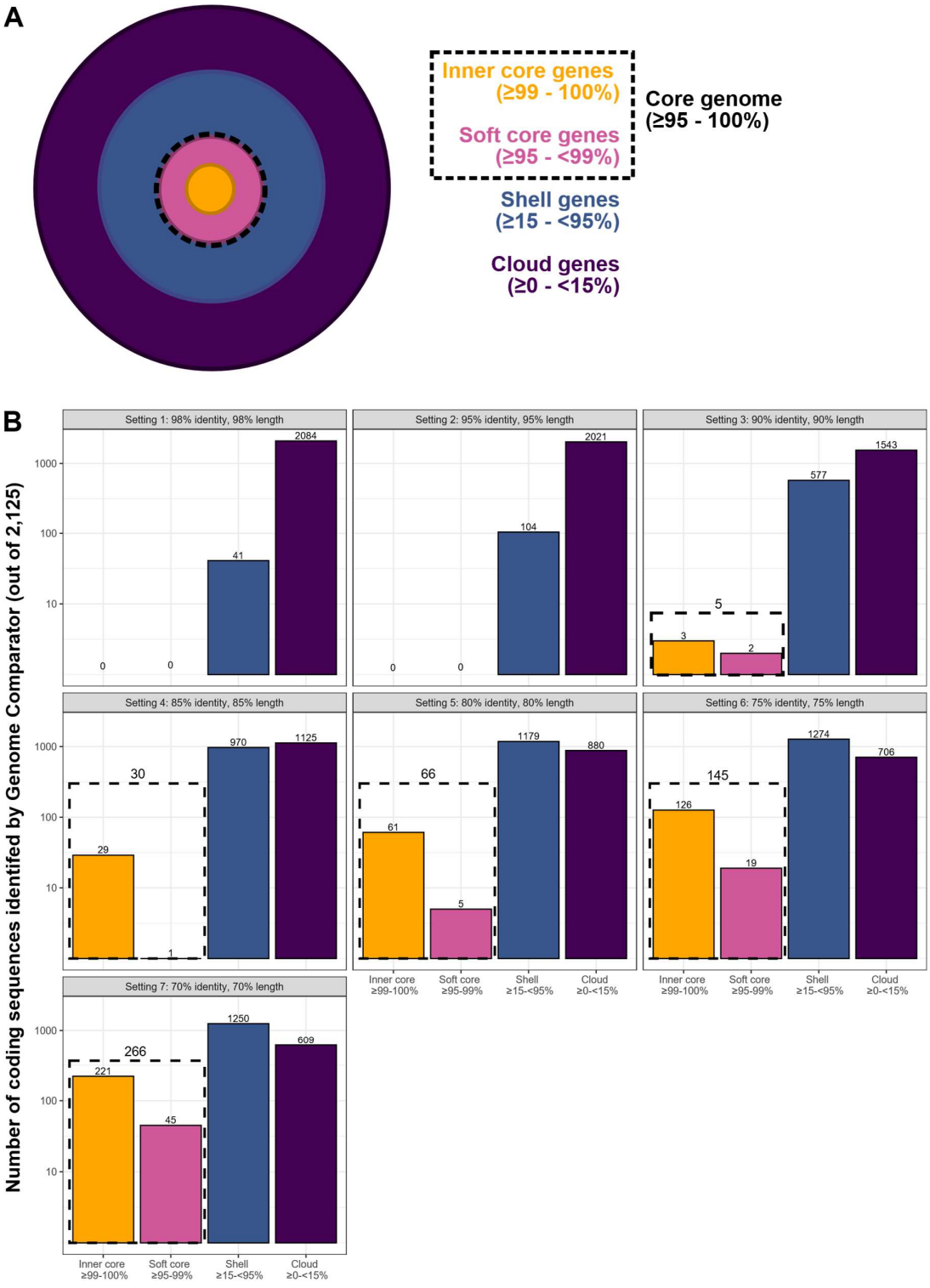


Figure 4.10 Different settings result in different streptococcal core genome sizes. (A-B) The dashed black line encompasses the core genome, which is comprised of the inner core and soft core genes. (A) Schematic overview of the definitions of the inner core, soft core, shell and cloud genes. (B) Seven different settings were applied to investigate the presence of the 2,125 pneumococcal coding sequences in the 1,919 *Streptococcus* genomes. Settings 1 and 2 result in no core genome.

In order to investigate the extent of variation within each set of genes, the distribution of mean *p*-distances was investigated for the inner core, soft core, shell and cloud genes defined using settings 3 to 7 (Figure 4.11). When looking at an alignment of two sequences, the *p*-distance is calculated as the proportion of observed nucleotide mismatches over the total alignment length [98]. Here, the *p*-distances represent the distance from a consensus sequence and mean *p*-distances were calculated per locus [73]. The largest locus had a sequence length of 14,331 bp and was part of the cloud genome for all settings. The mean *p*-distances of the core genome (combining the inner core and soft core genes) for each setting were: 0.047 for setting 3 (n=5 genes); 0.070 for setting 4 (n=30 genes); 0.094 for setting 5 (n=66 genes); 0.150 for setting 6 (n=145); 0.194 and for setting 7 (n=266). Overall, moving from setting 3 to 7, as the thresholds of nucleotide identity and alignment length became more lenient, there were more loci with a higher mean *p*-distance across the core genome, as expected (Figure 4.11).

The functional classification of the 2,125 coding genes of *Streptococcus pneumoniae* reference genome TIGR4 and of the genes within the core genomes determined using settings 3 to 7 were investigated (Figure 4.12). In TIGR4, 8.4% (178/2,125) of coding regions were involved in translation, ribosomal structure and biogenesis (category J), whilst in the core genome determined by setting 3 category J encompassed 100% (5/5), in setting 4 96.7% (29/30), in setting 5 78.8% (52/66), in setting 6 54.5% (79/145) and in setting 7 40.2% (107/266). In the core genomes determined by setting 5 to 7, the second largest functional class encompassed proteins involved in nucleotide transport and metabolism (category

F), which made up 7.6% (5/66) in setting 5, to 16.6% (24/145) in setting 6 and 16.2% (43/266) in setting 7, compared to 7.1% (151/2,125) in TIGR4.

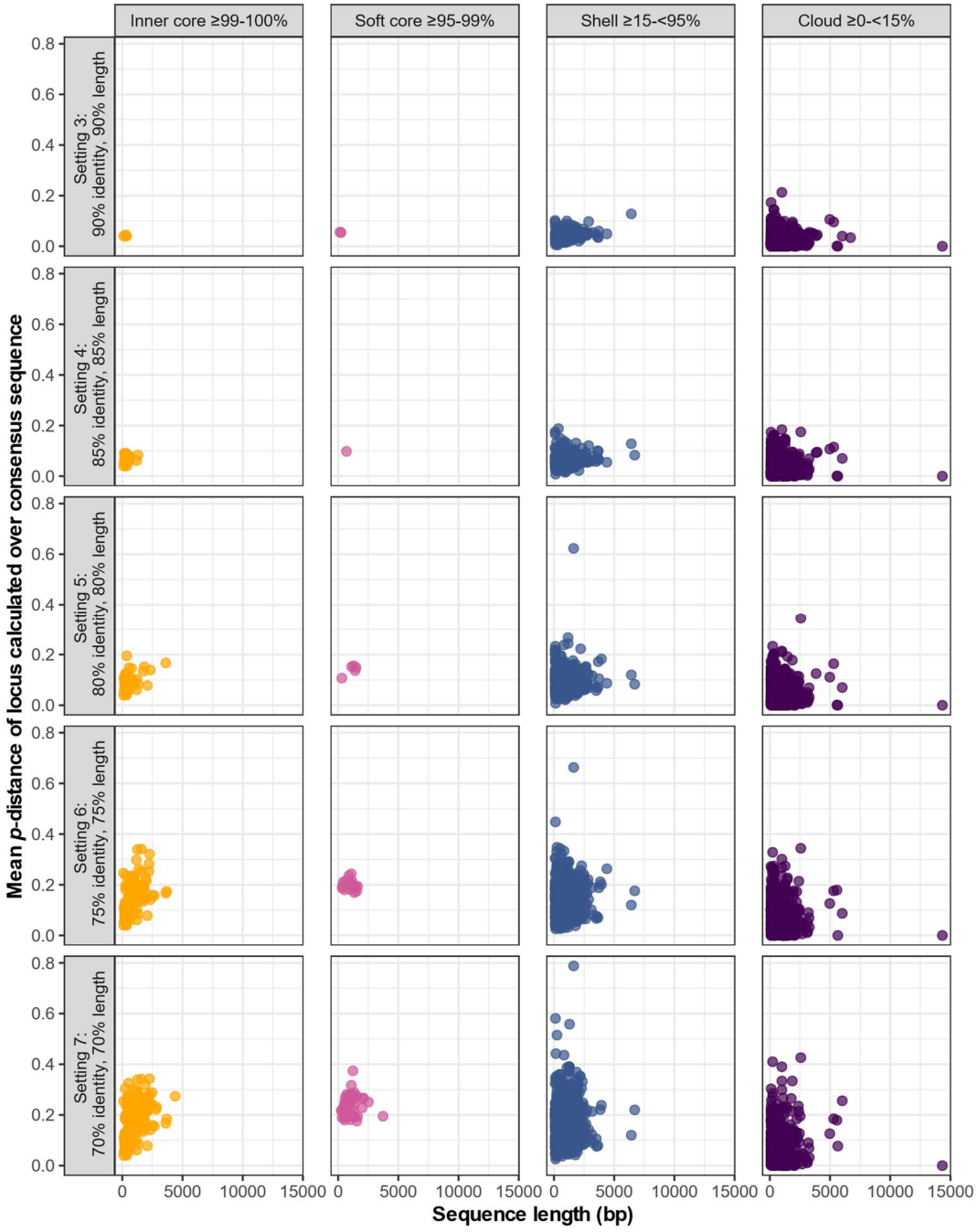


Figure 4.11 Mean *p*-distances of the inner core, soft core, shell and cloud genes. For each locus, the *p*-distance represents the dissimilarity in the multiple sequence alignments [98], calculated as the fraction of nucleotides differing from the consensus. The mean *p*-distance for each locus is shown. The number of inner core, soft core, shell and cloud genes varies depending on the different thresholds of nucleotide identity and alignment length applied by settings 3 to 7.

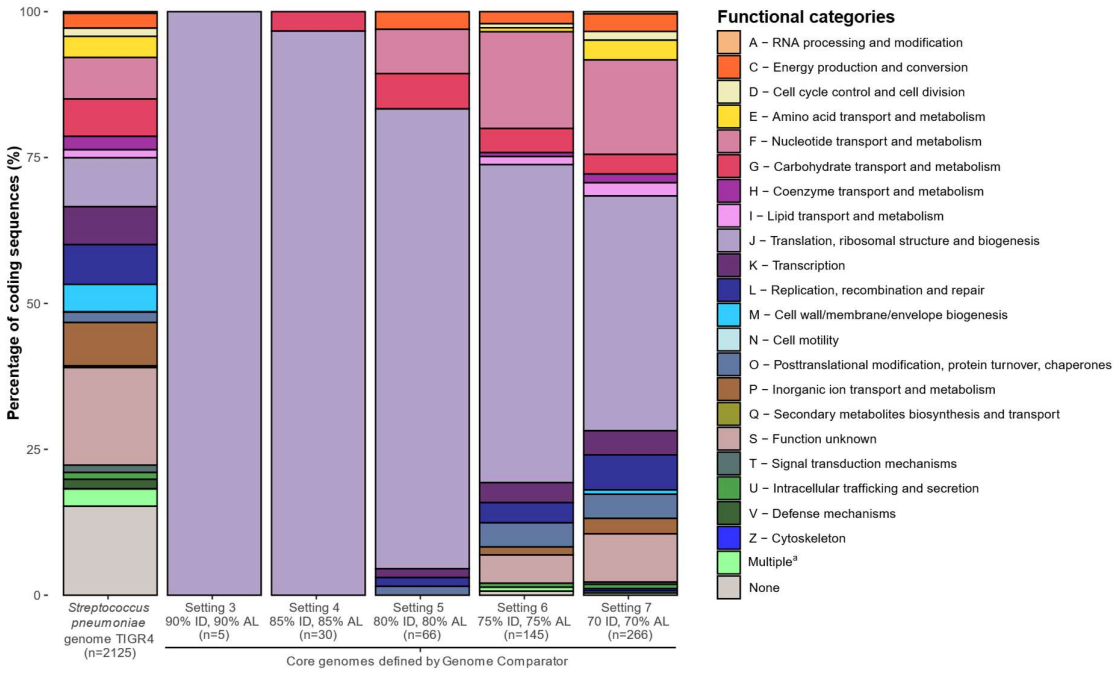


Figure 4.12 Functional classifications of the core genomes defined using different Genome Comparator settings compared to the coding sequences of *Streptococcus pneumoniae* reference genome TIGR4.
Functional classifications were determined by eggNOG-mapper using the Database of Clusters of Orthologous Genes (COGs) [269–272]. Percentages for each functional COG were calculated out of the total number of coding sequences of the core genome or reference genome. Settings 1 and 2 were excluded from this figure as they resulted in zero core genes.
Note: AL, alignment length; ID, identity.
[a]For 2.8% (63/2,215) of coding genes of *S. pneumoniae* reference genome TIGR4 multiple functional categories were found: CG (n=1), CO (n=4), CP (n=1), DJ (n=1), EG (n=4), EGP (n=3), EH (n=2), EJ (n=1), EM (n=2), EP (n=1), ET (n=4), FG (n=2), GK (n=7), GKT (n=1), GM (n=3), IQ (n=5), JKL (n=1), KL (n=2), KLT (n=3), KT (n=3), LM (n=1), LV (n=1), MU (n=1), NOT (n=1), NOU (n=1), NU (n=4), OU (n=1), PQ (n=1) and QT (n=1). For each of the core genomes determined using Setting 6 and Setting 7, one coding sequence was assigned two functional categories: OU (n=1).

Using setting 5 (minimum of 80% sequence identity and minimum of 80% alignment length), only 66 genes were present in ≥95% (≥1,824/1,919) of genomes of the 105 *Streptococcus* species. However, some species shared more genes with *S. pneumoniae* TIGR4 than others. In particular, there were nine *Streptococcus* species that shared more than 500/2,125 genes of *S. pneumoniae* TIGR4. Four of these species were represented in the dataset by a single genome each, and shared between 762 and 1,477 genes with *S. pneumoniae* TIGR4: *S. downii* (1,145 genes), *S. peroris* (763 genes), *S. timonensis* (762 genes) and *S. shenyangsis* (1,477 genes). Each genome of species *S. infantis, S. oralis, S. mitis, S. pseudopneumoniae* and *S. pneumoniae* contained between 795 and 1,964 genes of the 2,215 genes of *S. pneumoniae* TIGR4 (Figure 4.13). The elevated presence of *S. pneumoniae* TIGR4 genes in these nine species is consistent with their close phylogenetic clustering based on the 51 non-paralogous rMLST loci (Figure 4.3).

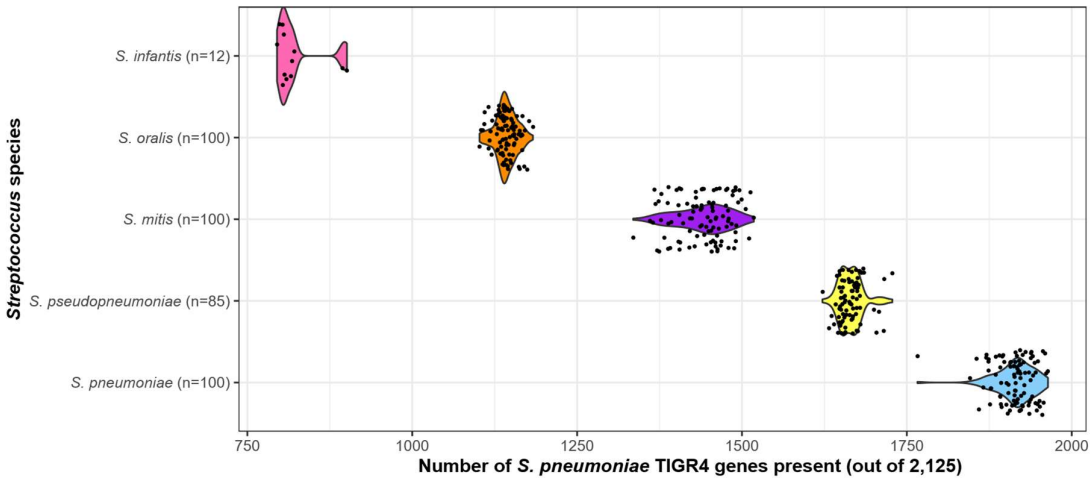


Figure 4.13 Closely related streptococcal genomes share more of the *S. pneumoniae* TIGR4 coding genes.
The distribution of the presence of the 2,125 coding sequences of *S. pneumoniae* reference genome TIGR4 assessed by minimum of 80% sequence identity and minimum of 80% alignment length (setting 5). Incomplete (end of contig) loci were counted as missing.

4.3.4.2 Panaroo

Panaroo was run using seven different settings to assess the number of streptococcal inner core, soft core, shell and cloud genes (Figure 4.14). Unlike the previous method, where the total number of genes was equal to 2,215 genes for all settings, the total number of clusters of orthologs identified by Panaroo ranged from 88,224 (setting 3) to 96,467 (setting 7). Without the constraint of requiring genes to be in the *S. pneumoniae* TIGR4 reference genome, and using this alternative approach that defines clusters of orthologous rather than performing a unidirectional BLASTN search, the identified core genome was larger. The core genome, defined as in Figure 4.10A (genes present in ≥95% of genomes in the *Streptococcus* dataset), ranged from 342 (setting 7) to 421 (setting 1). As settings 3 to 7 produced broadly similar results, further analyses were performed using the core genome defined by setting 5 (80% identity and 80% alignment length).

The functional classifications of the 412 core genes of the *Streptococcus* genus defined by Panaroo setting 5 were determined by eggNOG-mapper (Figure 4.15) [269–272]. As above, using the unidirectional BLASTN-based approach, the largest functional category, which encompassed 26.9% (111/412) of the core genes, was predicted to be involved in translation, ribosomal structure and biogenesis (category J). Overall, 12.6% (52/412) of the core genes had an unknown function (category S), 10.2% (42/412) of the core genes were predicted to be involved in nucleotide transport and metabolism (category F) and 7.8% (32/412) of the core genes in replication, recombination and repair (category L).

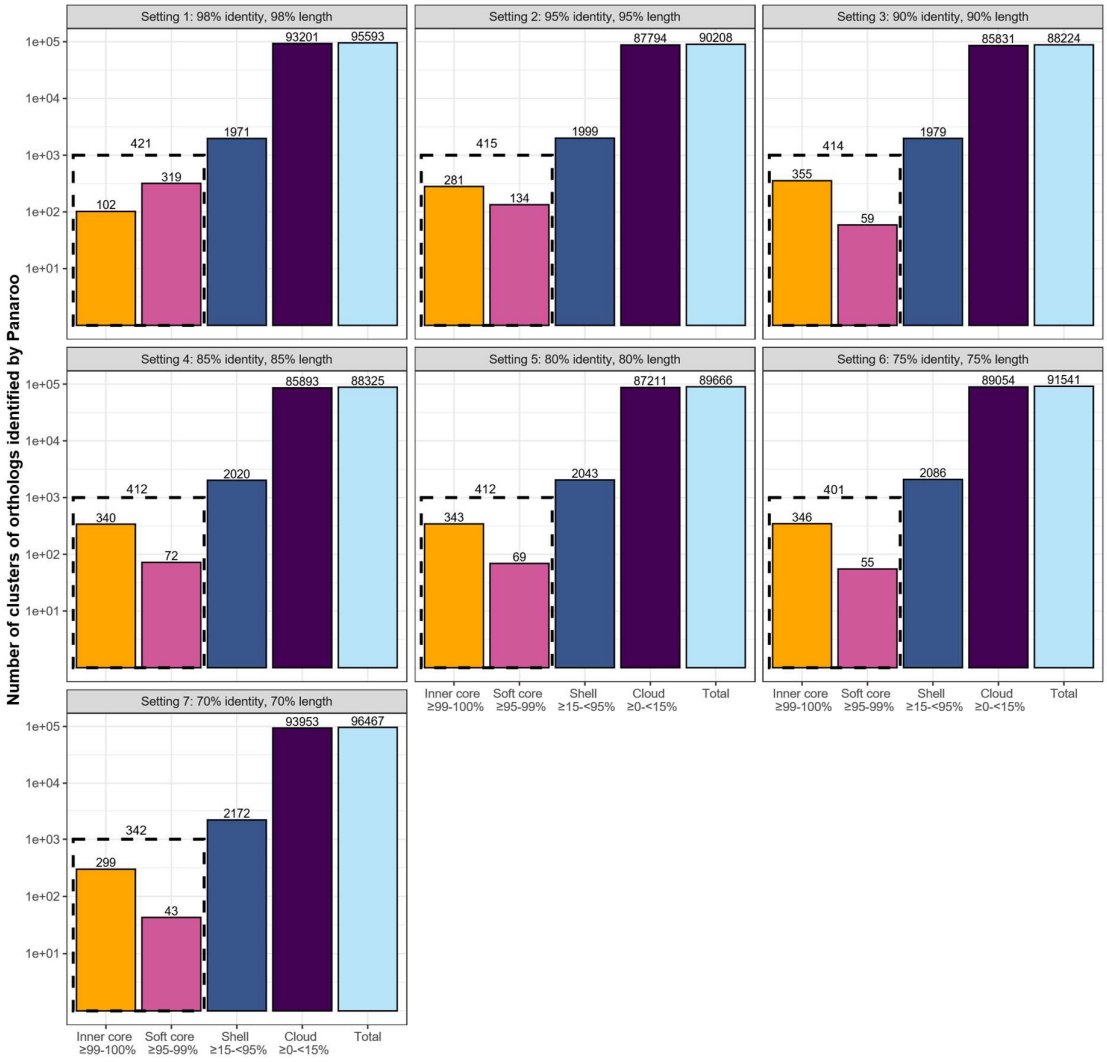


Figure 4.14 Different settings of Panaroo result in different numbers of inner core, soft core, shell and cloud genes.
Panaroo was run on 1,919 genomes of 105 different *Streptococcus* species.

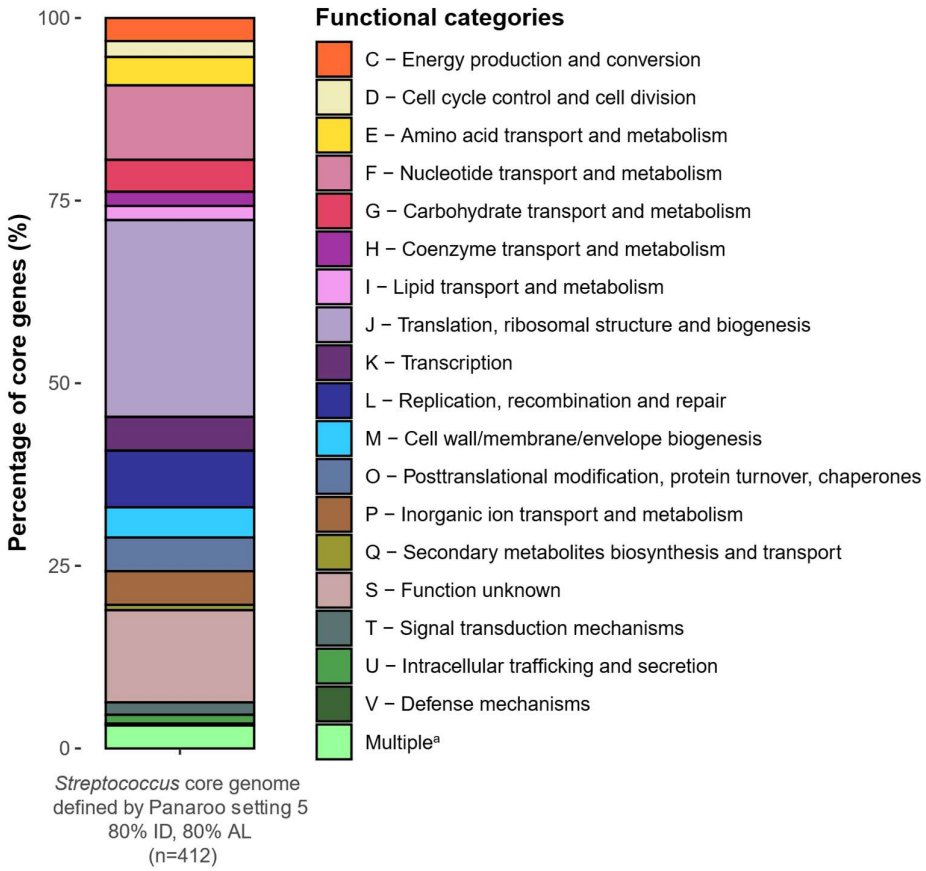


Figure 4.15 Functional classifications of the 412 core genes defined using Panaroo setting 5.
Functional classifications were determined by eggNOG-mapper using the Database of Clusters of Orthologous Genes (COGs) [269–272].
Note: AL, alignment length; ID, identity.
[a]For 3.2% (13/412) of the core genes multiple functional categories were found: EJ (n=1), FG (n=1), FJ (n=1), GK (n=1), IQ (n=3), JKL (n=2), KT (n=2), LU (n=1) and OU (n=1).

The 412 core genes were used to construct a genus-wide phylogenetic tree on which species were clearly demarcated (Figure 4.16). However, some species represented by a single genome clustered within another species (e.g. *S. shenyangsis* within the *S. mitis*) and eight *S. parasuis* isolates clustered within *S. suis*. Whilst the overall topology of the core genome tree was highly similar to the topology of the rMLST tree, there were some differences. Species of the mutans group – including *S. mutans, S. sobrinus, S. downei, S. ferus, S. macacae* and *S. ratti* [8] – were more dispersed on the core genome tree than on the rMLST-based

phylogeny. Whilst the mutans species formed a monophyletic group on the rMLST tree (Figure 4.3), the *S. sobrinus* and *S. downei* isolates were further removed from the other mutans group species on the core genome tree (Figure 4.16). In addition, the branch lengths in the core genome tree exceeded those in the rMLST phylogeny, e.g. the branch separating *S. pyogenes* and *S. dysgalactiae* is much longer in the core genome tree.

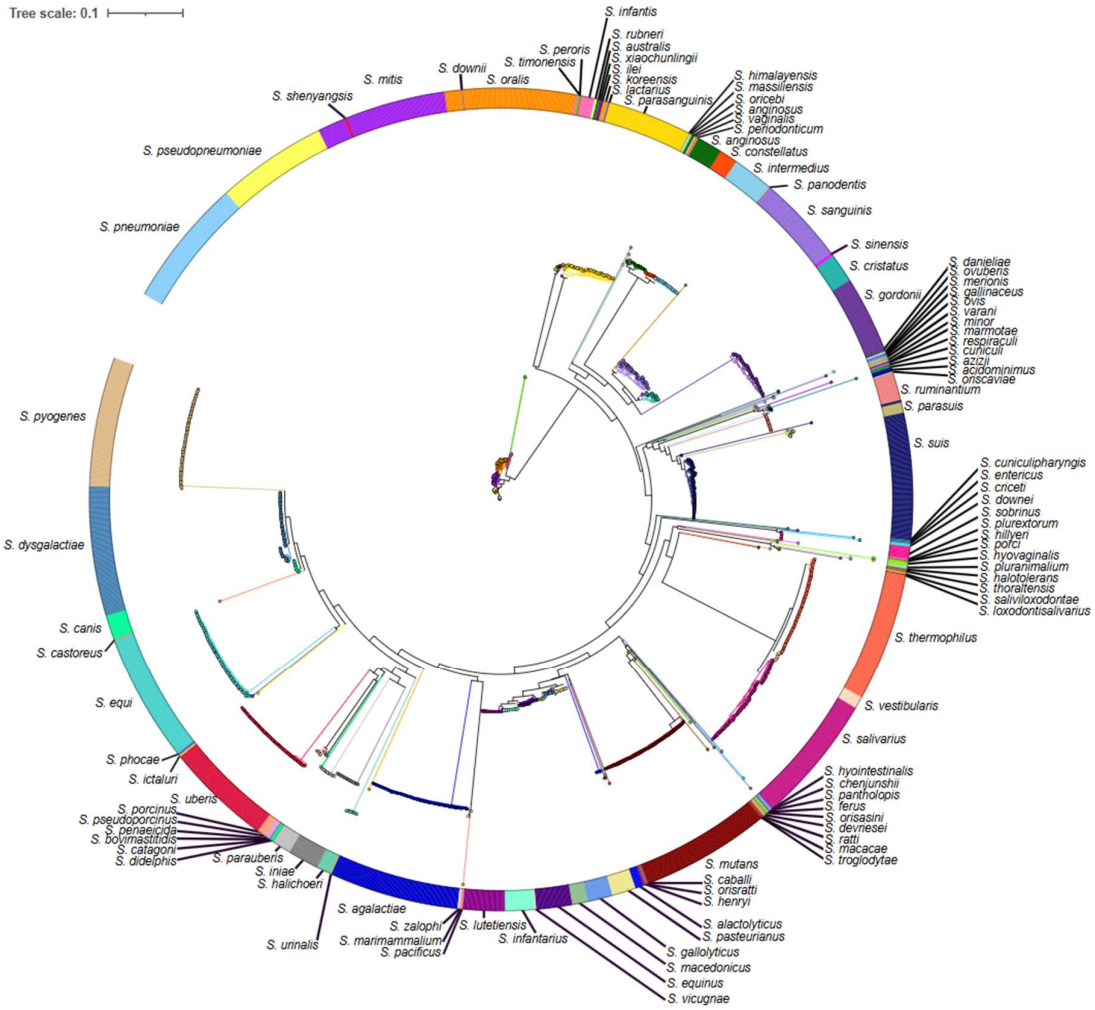


Figure 4.16 Maximum likelihood tree based on the nucleotide alignment of 412 core genes of 1,919 *Streptococcus* genomes.
The 105 *Streptococcus* species were highlighted with different colours and annotated, based on assignments made by those submitting the genome sequences to public repositories and confirmed by rMLST [73,89,157].
Note: *S., Streptococcus.*

4.3.5 Investigation of the presence of pneumococcal virulence factors

4.3.5.1 In silico amplicons of AMR genes

The *in silico* presence of amplicons of five qPCR assays targeting AMR genes was evaluated in the 1,919 *Streptococcus* genomes, comprising 100 *S. pneumoniae* genomes and 1,819 NPS genomes (Table 4.7). These published qPCR assays targeting AMR genes were designed for *S. pneumoniae*, where a positive PCR result for the *cat, ermB, mef* and *tetM* assays indicates resistance, whilst a positive *pbp2b* assay indicates susceptibility [147,280]. However, the absence *cat, ermB, mef* and *tetM* does not ensure susceptibility, since other resistance mechanisms may be present.

Among the five assays, the *tetM* assay yielded *in silico* positive results for the highest number of *Streptococcus* genomes: 21.2% (406/1,919). The vast majority of these, 404 genomes, had one *tetM* amplicon, whilst one genome had two *in silico* amplicons and one genome had three *in silico* amplicons (Table 4.7). Conversely, the *cat* assay tested *in silico* positive in the fewest *Streptococcus* genomes: 0.7% (13/1,919).

Table 4.7 *In silico* presence of qPCR amplicons for AMR genes in *Streptococcus* dataset (n=1,919)

Target [Study]	***Streptococcus* genomes with x amplicons (n)**				**Interpretation in *Streptococcus pneumoniae* if the assay is positive**
	x=0	**x=1**	**x=2**	**x=3**	
cat [280]	1,906	13	0	0	Chloramphenicol resistance
ermB [280]	1,686	224	9	0	Erythromycin and clindamycin resistance
mef [280]	1,659	260	0	0	Erythromycin resistance
pbp2b [280]	1,774	145	0	0	Penicillin susceptibility
tetM [280]	1,513	404	1	1	Tetracycline resistance

Note: NPS, non-pneumococcal *Streptococcus.*

The clustering of the *in silico* results of the *cat, ermB, mef* and *tetM* qPCR assays among genomes in the *Streptococcus* dataset was investigated further (Figure 4.17A). Of the genomes testing positive for *ermB* (n=233), the highest count of positive hits per species was found in *S. suis* (n=70), followed by *S. salivarius* (n=18). For genomes with an *in silico* positive *mef* amplicon (n=260), the highest count per species was found in *S. mitis* (n=49), followed by *S. salivarius* (n=44). Among genomes that tested *in silico* positive for *tetM* (n=406), the highest count per species was recorded for *S. agalactiae* (n=75), followed by *S. pseudopneumoniae* (n=42).

Despite close clustering on the core genome phylogeny, the presence of AMR determinants in *S. thermophilus* differed substantially from that in *S. salivarius*. None of the 100 *S. thermophilus* genomes tested *in silico* positive for the *cat, mef* or *tetM* assays, and only one genome was *in silico* positive for the *ermB* assay. In contrast, among the 100 *S. salivarius*, AMR determinants were frequently present: *ermB* (n=18), *mef* (n=44) and *tetM* (n=19).

Overall, 63.5% (1,219/1,919) of the isolates were *in silico* negative for all four of the *cat, ermB, mef* and *tetM* qPCR assays. The remaining 36.5% (700/1,919) of genomes tested positive for one to three of the AMR determinants, whilst none of the genomes were *in silico* positive for all four of the assays (Figure 4.17B). Most commonly, genomes tested *in silico* positive for only the *tetM* assay 11.5% (220/1,919) or for only the *mef* assay 9.7% (187/1,919). In total, only 0.9% (17/1,919) of genomes tested *in silico* positive for three of the assays

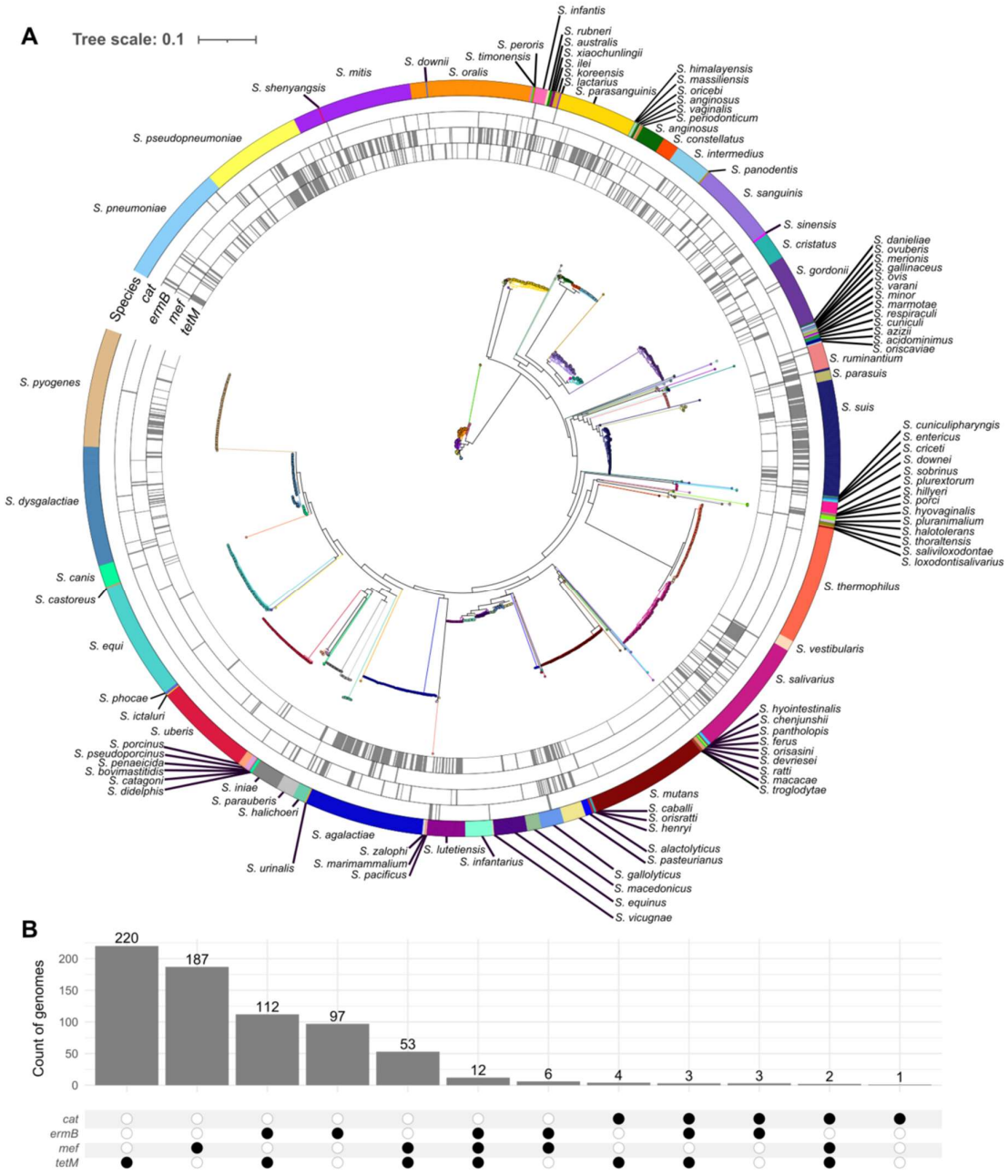


Figure 4.17 Clustering of AMR genes among *Streptococcus* genomes.
(A) Maximum likelihood phylogeny based on the 412 core genes of 1,919 *Streptococcus* genomes. From the inside to the outside ring, the annotations show the *in silico* qPCR results of *tetM, mef, ermB* and *cat* (marking presence in grey and absence in white), and the outside ring highlights the 105 *Streptococcus* species with different colours and annotations, as in Figure 4.16.
(B) Combinations of *in silico* positive hits for *cat, ermB, mef* and *tetM* among the *Streptococcus* genomes. Overall, 1,219/1,919 *Streptococcus* genomes tested *in silico* negative for all four of the qPCR assays and were excluded from the bar plot. Black circles indicate that an *in silico* positive qPCR amplicon was detected, whilst white circles indicate *in silico* negative qPCR result.
Note: AMR, anti-microbial resistance; NPS, non-pneumococcal *Streptococcus; S. Streptococcus*.

4.3.5.2 In silico amplicons of S. pneumoniae serotypes

The *in silico* presence of amplicons for the 41 standard PCR assays targeting pneumococcal serotypes/serogroups was determined across the 1,819 NPS genomes of the *Streptococcus* dataset. Out of these 41 assays, 11 assays produced *in silico* positive PCR amplicons in one to 13 NPS genomes each (Table 4.8). Nine of the 11 PCR assays were positive in a single *Streptococcus* species each (*S. mitis* or *S. oralis* or *S. infantis*), whilst the standard PCR assay for *S. pneumoniae* serotypes 10F/10C/33C was positive in genome(s) of two species: *S. gordonii* (n=1) and *S. oralis* (n=12) and the PCR assay for *S. pneumoniae* serotype 5 was positive in the genomes of three species: *S. mitis* (n=1) and *S. oralis* (n=1) and *S. pseudopneumoniae* (n=1).

The number of unique amplicons identified for each of these 11 assays ranged from one to six. For assays with multiple unique *in silico* amplicons, the variations were due to nucleotide changes rather than insertions or deletions, as the amplicon length remained constant for each assay. In standard PCR assays, *in vitro* product detection is typically achieved through gel electrophoresis, where the size of the *in vitro* amplicon is compared to a molecular size ladder run on the gel [254–259]. Therefore, potential differences in amplicon size between *S. pneumoniae* and NPS species would be detectable on a gel, facilitating differentiation [254–259]. However, for the 11 PCR assays for which *in silico* amplicons were found in the NPS genomes, all *in silico* amplicon lengths found in NPS genomes matched those published for *S. pneumoniae*, except for the PCR assay targeting serotype 2, where the amplicon length differed by only one nucleotide (289 bp found in *S. oralis* and 290 bp published for *S. pneumoniae*),

which would not be expected to appear materially different on a gel [147,254–259].

Table 4.8 Pneumococcal serotypes/serogroups PCR assays resulting in one or more positive *in silico* hits in NPS genomes (n=1,819).

***Streptococcus pneumoniae* serotype/ serogroups[a]**	**NPS genomes with x amplicons (n)**		**Unique amplicons (n)**	**NPS amplicon length (bp)[b]**	**Species**
	x=0	**x=1**			
1	1,818	1	1	280	*S. mitis* (n=1)
2	1,817	2	2	289	*S. oralis* (n=2)
5	1,816	3	2	362	*S. mitis* (n=1), *S. oralis* (n=1), *S. pseudopneumoniae* (n=1)
10A	1,816	3	2	628	*S. oralis* (n=3)
10F/10C/33C	1,806	13	6	248	*S. gordonii* (n=1), *S. oralis* (n=12)
13	1,818	1	1	655	*S. oralis* (n=1)
17F	1,818	1	1	693	*S. mitis* (n=1)
18C/18F/18B/18A	1,816	3	2	573	*S. mitis* (n=3)
20	1,818	1	1	514	*S. infantis* (n=1)
24F/24A/24B	1,817	2	1	99	*S. infantis* (n=1), *S. mitis* (n=1)
33F/33A/37	1,818	1	1	338	*S. oralis* (n=1)

[a]Published standard PCR assays to target *S. pneumoniae* serotypes/serogroups [254–259].
[b]The *in silico* amplicon length found in NPS species was the same as the *in vitro* amplicon length published for *S. pneumoniae* for all assays apart from the assay targeting serotype 2, which was 290 bp in *S. pneumoniae.*
Note: NPS, non-pneumococcal *Streptococcus; S. Streptococcus*.

The phylogenetic clustering of the 11 standard pneumococcal serotyping/serogrouping PCR assays that caused one or more *in silico* positive amplicons among NPS genomes was assessed (Figure 4.18). None of the NPS genomes were *in silico* positive for more than one serotyping/serogrouping PCR assay. All NPS species causing *in silico* positive were part of the α-haemolytic viridans streptococci, either the mitis group (*S. infantis, S. mitis, S. oralis, S. pseudopneumoniae*) or sanguinis group (*S. gordonii*) [8].

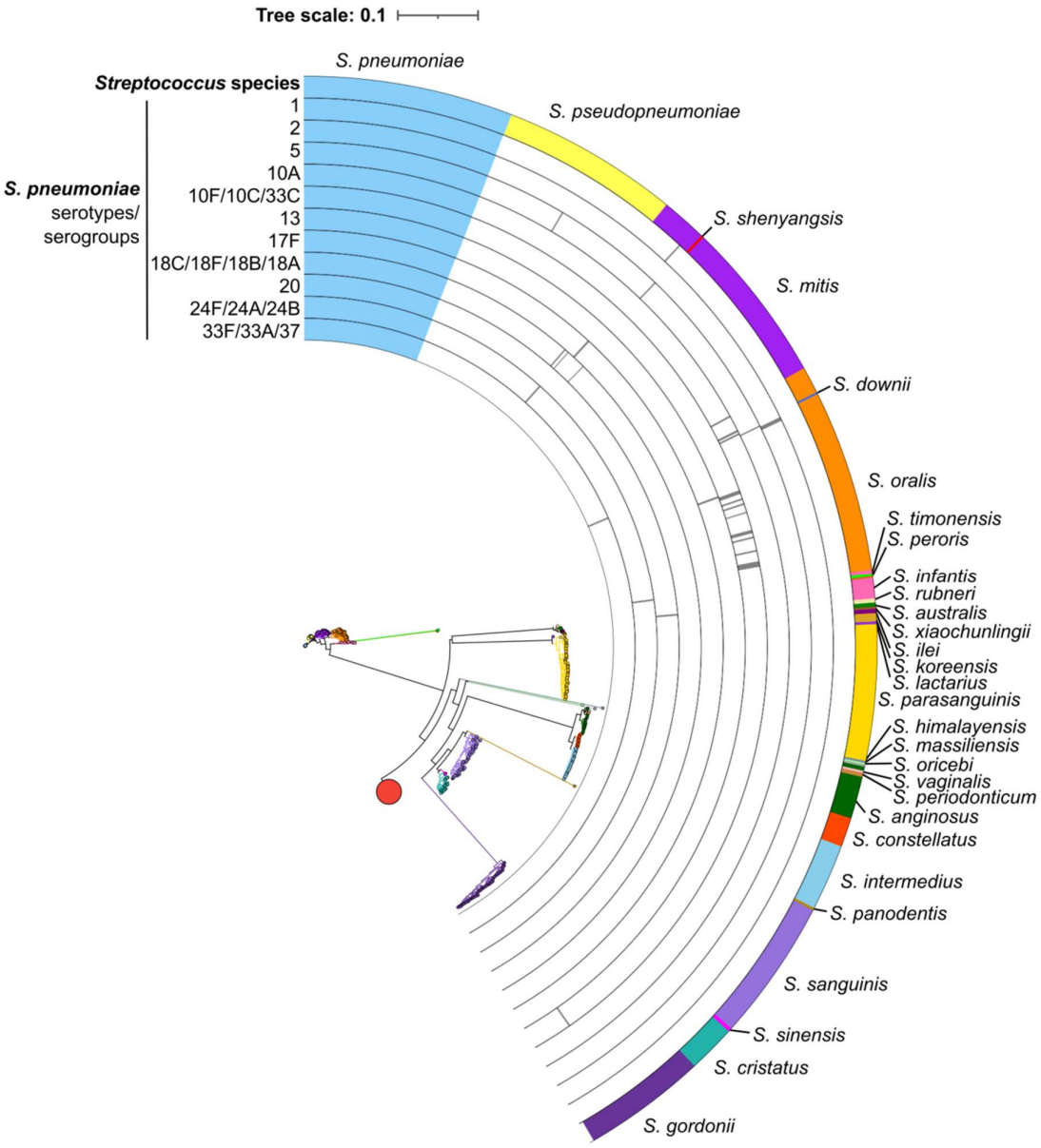


Figure 4.18 Clustering of *in silico* amplicon presence of pneumococcal serotypes/serogroups in NPS species on a core genome phylogeny. Maximum likelihood tree based on 412 *Streptococcus* core genes, with the outer ring highlighting the *Streptococcus* species using different colours and annotations, as in Figure 4.16. The *in silico* presence of amplicons of 11 pneumococcal serotype/serogroups PCR assays in NPS species is displayed in the 11 inner rings, with *in silico* positive hits marked grey and negative hits marked white. The *S. pneumoniae* genomes were excluded from the *in silico* analyses and are highlighted in blue. Branches of the tree without any *in silico* positive amplicons were collapsed and marked with a red circle.
Note: NPS, non-pneumococcal *Streptococcus; S., Streptococcus*.

4.4 Discussion

The *Streptococcus* genus encompasses a large group of species, some of which are challenging to differentiate due to their overlapping phenotypic and genotypic traits [8]. This chapter leveraged publicly available *Streptococcus* genomes to investigate phylogenetic clustering using various *in silico* typing methods. A robust core genome phylogeny was constructed, which enabled precise differentiation of many *Streptococcus* species and onto which virulence factors were mapped, providing valuable insights for both routine and research microbiology.

A comprehensive genomic dataset was compiled, comprising 1,919 high-quality genomes of 105 different *Streptococcus* species. The rMLST-based phylogenetic tree displayed clustering of mitis group species *S. cristatus* within sanguinis group species. This suggested that the mitis and sanguinis groups should be merged into one, as has been done by others such as Jensen et al. (2016), who included the sanguinis group within the mitis group streptococci [129].

Using rMLST typing as a benchmark for streptococcal speciation, the species-specific differences in GC-content and genome length were investigated (Figure 4.5). The distribution of the streptococcal GC-content largely aligned with the established theory that there is minimal variability in GC-content within the same species [290]. In addition, whilst the overall GC-content across prokaryotes is known to vary from 25-75%, the GC-content of the 1,919 streptococcal genomes studied here only ranged from 33.18-47.56% [290]. Interspecies differences in GC-content might originate from environmental adaptations [291]. Bacteria throughout the Firmicutes, which include other oral commensals such as species in the *Staphylococcus* genus, are known to have a low GC-content [6]. Whilst previous studies have suggested that *S. pneumoniae* has an overall larger

genome length compared to its closely related mitis group commensals that rarely cause disease following the “genome reduction hypothesis” [20], this was not observed in the large genomic dataset studied here (Figure 4.4).

Protein-encoding single locus targets were unable to accurately differentiate *Streptococcus* species, with species often intertwining, especially those closely related to *S. pneumoniae*. This chapter did not investigate the clustering of rRNA-encoding single locus targets like the 16S rRNA gene. Previous research has shown that the 16S rRNA sequences of *S. mitis, S. oralis, S. pneumoniae* and *S. pseudopneumoniae* share >99% identity, making this locus unsuitable for phylogenetic stratification [222]. However, Scholz et al. (2012) found that position 203 of the 16S rRNA gene can be used to distinguish *S. pneumoniae,* which has a cytosine residue at that position, from other mitis group streptococci, which have an adenosine at position 203 [292]. This differentiation relies on sequence comparison or restriction endonuclease digestion [292]. Large-scale *in silico* investigations of 16S rRNA genes using short-read WGS assemblies are complicated by the fragmentation of 16S rRNA sequences. This fragmentation arises due to assembly errors caused by the presence of multiple copies of the gene in *Streptococcus* genomes [20].

Some studies have combined several (usually four to seven) single locus typing targets to create a genus-wide multilocus typing scheme [20]. For example, Kilian et al. (2008) concatenated the partial sequences of *ddl, gdh, rpoB* and *sodA* to create a phylogenetic tree [20]. Creation of this type of scheme was not explored in this chapter; instead, a genus-wide streptococcal core genome was investigated.

In order to investigate the *Streptococcus* core genome, two contrasting methods and a range of different settings within each method were employed (Figure 4.10 and Figure 4.14). The first method used a unidirectional BLASTN approach of all coding sequences of a pneumococcal reference genome [73]. This resulted in a variable amount of streptococcal core genes (mean across seven settings: 73 genes, range: 0-226), with the number of core genes increasing as settings became more lenient. The second method used Panaroo, which provided a more stable estimate of the core genome size across settings (mean across seven settings: 402 genes, range: 324-421). The larger number of *Streptococcus* core genes identified by Panaroo compared to the unidirectional BLASTN approach can be attributed to differences in the nature of the clustering algorithms. Using the unidirectional BLASTN approach, the sum of the inner core, soft core, shell and cloud genes always equalled the number of genes in the pneumococcal reference genome (2,215 genes) across all settings. In contrast, Panaroo identified a total number of clusters of orthologs ranging from 88,224 to 96,467, depending on the settings used. The number of core genes identified by Panaroo is similar to previous findings that the *Streptococcus* genus has an estimated closed core genome of 369 genes [234]. Similarly, Barajas et al. (2019) identified a core genome of 404 genes across 108 isolates of 16 different streptococcal species [238].

Species that clustered more closely to *S. pneumoniae* on the rMLST and core genome phylogenies, shared a larger number of genes with *S. pneumoniae* TIGR4 than were included in the genus-wide *Streptococcus* core genome (Figure 4.13). This observation aligns with findings by Kilian and Tettelin (2019), who defined a core genome of 690 genes across *S. infantis, S. mitis, S. oralis, S.*

pneumoniae and *S. pseudopneumoniae* [293]. They also showed that the core genome increased to 894 core genes when *S. infantis* was removed, and to 995 genes when *S. oralis* was excluded [293]. Overall, the closer phylogenetic relationship to *S. pneumoniae* corresponded to a higher number of shared genes, as expected.

Prokaryotic core genomes have been associated with essential functions required for bacterial survival [294]. When investigating the core genome of 105 heterogeneous species of the *Streptococcus* genus, the core genes were expected to mainly contain genes performing vital functions. Indeed, the largest functional category across the core genes was involved with translation, ribosomal structure and biogenesis (Figure 4.12 and Figure 4.15). Similar results have been found for other genera; for example, the largest functional category of the *Campylobacter* core genome was also translation, ribosomal structure and biogenesis [295].

In both the rMLST and core genome phylogenies, there were several species represented by a single genome which clustered within another species, such as *S. shenyangsis* within *S. mitis*. *S. shenyangsis* was originally identified from the oropharynx of a healthy child in Shenyang (China) and was coined as a new species based on 16S rRNA sequencing, despite the *S. shenyangsis* 16S rRNA sequences showing 99.2% similarity with *S. oralis* and 98.9% similarity with *S. mitis* strains [296]. Overall, the definition of what constitutes a species is recursive upon the methods used and taxonomic changes have been suggested after increased availability of genomic data. For example, Jensen et al. (2016) suggested the reclassification of *S. oligofermentans* as a synonym of *S. cristatus* based on core genome phylogenetic clustering [129].

Limitations of this study include that the trees were unrooted, meaning evolutionary directionality could not be inferred. In addition, prevalence of genomes ranged from one to 100 per *Streptococcus* species. Whilst an upper limit of 100 genomes was imposed to prevent major bias, the higher frequency species might have skewed the core genome calculations, leading to overrepresentation of genes present in species present in high numbers. Furthermore, genomes failing the quality control metrics, or those that could not be differentiated to a species level by rMLST typing were excluded from the dataset [89]. Whilst this ensured high-quality genomes, future work could investigate the clustering of isolates defined only as "*Streptococcus sp."* based on rMLST typing, or the clustering of isolates for which rMLST speciation did not result in the same species as was originally assigned by those who submitted the genome to databases. Potentially, high-resolution core genome clustering could elucidate the taxonomy of these genomes, or discover new species. Indeed, Kumar et al. (2023) used PEPPAN to construct a core genome phylogenetic tree to investigate the clustering of streptococcal isolates that were previously unclassified [244].

This chapter looked for the *in silico* presence of PCR-based amplicons of AMR targets [280]. However, phenotypic methods remain the gold-standard for investigating the presence of AMR, which include disc diffusion to look for resistance or susceptibility, or broth dilution to determine minimum inhibitory concentrations (MICs) [245]. Furthermore, genotypic methods can accurately identify AMR in *S. pneumoniae* [297], and more recently machine learning methods have been employed to predict AMR phenotypes [298,299]. Future work could study AMR and *cps* genes across the *Streptococcus* phylogeny using both bioinformatics approaches and biochemical analyses. Future work could also look

at the *pbp* genes and explore the clustering of *pbp2x* across NPS species, building on existing literature that showed the presence of *pbp2x* in various viridans streptococci [251,252].

The strengths of this study include the diversity, size and quality of the genomic dataset: to our best knowledge, this dataset was larger and/or included a more diverse range of *Streptococcus* species than any other published study. Furthermore, genomes underwent comprehensive quality control, and species assignments were confirmed using rMLST. Among the 1,919 genomes of 105 *Streptococcus* species, a robust set of 412 core genes was identified after trialling two contrasting core genome identification methods and several settings. This set of core genes can be used in various settings, including the typing of streptococcal species from metagenomics samples [240]. Combining metagenomics and core genome investigations would allow studying the presence of streptococci in diverse environmental niches [240].

Overall, owing to the advancements in sequencing technology, there has been a substantial increase in the number of publicly available bacterial genomes [90]. This study leveraged publicly available WGS data to investigate streptococcal taxonomy and classification using core genome analyses. The high-resolution core genome phylogeny provides a framework for studying streptococcal evolution and epidemiology.

Chapter 5

General Discussion

5.1 Summary of findings

This thesis used genomic epidemiology to improve PCR- and LAMP-based diagnostic assays for *S. pneumoniae* and to enhance the understanding of the population structure across the *Streptococcus* genus.

5.1.1 A new multiplex qPCR assay for pneumococcal identification

In Chapter 2, I aimed to improve PCR-based diagnostics for the identification of *S. pneumoniae* by using large and diverse bacterial genome datasets. Although several targets have been suggested for the PCR-based identification of *S. pneumoniae*, many of these targets have historically only been validated on small *in vitro* panels [128,146,148,150,152]. Furthermore, *lytA* (the target recommended for pneumococcal identification by the CDC) has homologues in prophages, which are present throughout the *Streptococcus* genus [146,147,63,62]. In this chapter, I investigated the *in silico* performance of existing PCR-based assays *lytA* [146], *piaB* [148], *ply* [146], *psaA* [146], Spn9802 [150], SP2020 [128] and Xisco [152], using an extensive pneumococcal genome dataset (n=7,547) and a diverse NPS genome dataset (n=1,825) [74]. Whilst the existing assays generally performed well, *in silico* positive hits in *S. pseudopneumoniae* and/or *S. mitis* were found for all assays apart from Xisco. Overall, the *in silico* Youden's J statistics of Xisco and SP2020 were the highest (>0.99 each). Therefore, I designed a new multiplex qPCR assay targeting Xisco and SP2020, using large multiple sequence alignments of the genes extracted from diverse pneumococcal genomes to find conserved regions to target with the primers and probes.

Experimentally, the new multiplex qPCR assays could detect as few as nine copies of pneumococcal DNA per reaction and consistent qPCR C_q-value intervals between ten-fold dilutions indicated that the assays could reliably quantify

pneumococcal DNA. I compiled laboratory specificity (n=277) and sensitivity (n=246) panels and found 100% [$CI_{95\%}$ 98.7-100%] specificity and 100% [$CI_{95\%}$ 98.5-100%] sensitivity of the new SP2020 and Xisco assays *in vitro*. I assessed the *in silico* amplicon sequences of the qPCR assays across 6,090 pneumococcal genomes to devise an *in vitro* panel (n=248) that covered the amplicon diversity found across the pneumococcal population. Testing this *in vitro* amplicon diversity panel demonstrated that the new multiplex qPCR assays performed well across a range of different amplicon sequences and saved costly and laborious experiments by testing a small *in vitro* panel that covered the diversity found in a genome dataset nearly 25 times its size.

Multiplexing the qPCR primers and probes of Xisco and SP2020 allowed simultaneous detection of two pneumococcal genes, which could facilitate a dual-target approach that has been recommended for *S. pneumoniae* [300]. Within the experiments conducted for this thesis, multiplexing the pneumococcal targets saved reagents and the further incorporation of the published 16S rRNA target provided a useful internal control [168]. Overall, this chapter used an *in silico* approach, complemented by *in vitro* validation, to design and develop PCR-based diagnostics for *S. pneumoniae* in a rapid and low-cost manner [101].

5.1.2 A new real-time LAMP assay for pneumococcal identification

In Chapter 3, I used genomic data to develop a new real-time LAMP assay for the identification of *S. pneumoniae*. LAMP assays offer advantages over PCR assays, particularly in resource-limited settings, since they can produce faster results at lower costs without the need for a thermocycler machine [191]. Building on the results of Chapter 2, which demonstrated that Xisco and SP2020 were the most accurate targets for pneumococcal identification, Chapter 3 aimed to design LAMP

assays targeting these genes. However, the large amount of sequence diversity within the Xisco gene made the design of a LAMP assay unfeasible. Therefore, LAMP assays targeting SP2020 were created [207]. Among the six LAMP primer combinations targeting SP2020 that were tested, SP2020_3+, the only assay which included loop primers, had the fastest reaction time (12 minutes). The *in silico* performance of the new SP2020_3+ LAMP assay was compared to four published pneumococcal LAMP assays targeting *lytA* [201]*, ply* [204]*, psaA* [201] and SPNA45_01710 [205]. Overall, the *in silico* performance of the new SP2020_3+ LAMP assay was the highest (Youden's J statistic >0.99).

The SP2020_3+ assay had an *in vitro* limit of detection of 2.6 pneumococcal genome copies per µL and an *in vitro* sensitivity of 100% [$CI_{95\%}$ 98.5-100%]) and specificity of 100% [$CI_{95\%}$ 98.7-100%]. Using the amplicon diversity panel, 71 unique SP2020 alleles and 15 unique SP2020_3+ amplicon sequences were tested, which all gave rapid *in vitro* positive results. Overall, this new, fast, highly sensitive, specific, real-time LAMP assay could be particularly useful for application in low-resource settings. This assay is in alignment with the WHO recommendations for diagnostic assays for infectious disease surveillance in LMICs [209].

5.1.3 A streptococcal core genome phylogeny to map virulence factors

In Chapter 4, I evaluated the ability of various *in silico* typing methods to differentiate the genomes of 105 streptococcal species. I studied the size, function and clustering of the streptococcal core genome to construct a core genome phylogeny onto which the presence of AMR and virulence genes *in silico* amplicons was mapped. To achieve this, I first compiled a high-quality genome dataset of 105 *Streptococcus* species, selecting 1,919 genomes from a pool of

over 75,000 genomes and used rMLST as the benchmark for speciation [73,89]. Seven single locus typing genes that have been proposed for speciating streptococci were identified from literature: *groEL* [219]*, gyrB* [219]*, recA* [220]*, rpoB* [219]*, rpsB* [221]*, sodA* [219]*,* and *tuf* [220]. These genes were used to construct single locus maximum likelihood phylogenies [278]. These protein-encoding single locus typing genes were inadequate for speciation, as distinct species clusters were not formed; in particular, genomes of species closely related to *S. pneumoniae* often intermixed.

The streptococcal core genome size and function were investigated using two contrasting approaches: using a unidirectional BLAST search of the coding genes of a pneumococcal reference genome (using the Genome Comparator module part of $BIGS_{DB}$) and by defining clusters of orthologous genes (using Panaroo) [69,73,226]. A *Streptococcus* genus-wide core genome of 412 genes was defined, with the largest functional category of nearly 27% of core genes involved in translation, ribosomal structure and biogenesis. Phylogenetic clustering based on the core genome displayed clear demarcations. Mapping the presence of *in silico* amplicons of PCR assays targeting virulence factors onto this phylogenetic tree elucidated the widespread presence of AMR determinants (including *ermB, mef* and *tetM*) across the *Streptococcus* genus [147,280]. This was in contrast to the positive hits of *in silico* amplicons of PCR assays targeting the pneumococcal *cps* genes in NPS species, which were found only in viridans species, and nearly all in the subset of streptococci most closely related to *S. pneumoniae*, the mitis group [256,257,255,254,258,259,147].

5.2 Strengths

The overarching strength of this thesis was the repurposing of existing, high-quality, diverse bacterial genomes for *in silico* analyses. In Chapters 2 and 3, *in silico* approaches were exploited as a rapid and cost-effective method in three main ways. First, two large and diverse genomic datasets were compiled and used to assess the *in silico* performance of published diagnostic assays, by investigating the *in silico* presence of primer and probe binding sites [90,182]. This analysis relied on the ipcress tool, which is built into $BIGS_{DB}$, to search for *in silico* primer binding regions and to extract the amplicon sequences from genome assemblies [2,73]. Second, large pneumococcal multiple sequence alignments of target genes were used to identify the most conserved regions to target with the primers and probes of the new multiplex qPCR and real-time LAMP assays. Computational tools were key to the design and optimisation of successful molecular diagnostics by predicting the biophysical metrics such as the hairpin T_m and self-dimer T_m of proposed oligonucleotides [117,207]. Third, *in silico* analyses were utilised to design an amplicon diversity panel for use in laboratory validations: the amplicon sequences of interest of 6,090 diverse pneumococcal genomes were analysed *in silico* and used to carefully select an amplicon diversity panel. Testing a small *in vitro* panel that covered the diversity found in a global pneumococcal *in silico* dataset saved time and resources. Overall, this work exemplified the value of *in silico* analyses to aid the development of new *in vitro* diagnostics [101].

In Chapter 4, a comprehensive genomic dataset of the *Streptococcus* genus was compiled, encompassing 1,919 high-quality genomes of 105 *Streptococcus* species [73,89]. Having such a large and diverse dataset is crucial for determining

the genus-wide core genome; for example, the core genome would have been artificially large if only a small number of similar genomes had been included [301]. Genomes were selected from the rMLST database (https://pubmlst.org/species-id), demonstrating the value of open-access genome data archives such as the PubMLST databases [73,89]. The rMLST scheme was used as the benchmark for speciation and as a quality control metric [73,89]. Further strengths of this chapter include the thorough investigation of the *Streptococcus* core genome using two contrasting approaches and a range of settings [73,226]. A robust streptococcal core genome of 412 genes was defined using open-source bioinformatics tools PROKKA and Panaroo [226,274], and functional classifications were assigned to core genes using eggNOG-mapper [267,268]. Overall, this chapter was strengthened by the usage of diverse genomes and the application of advanced open-source bioinformatics tools; these analyses were made possible by the countless researchers and microbiology laboratories who made the genomic data they produced publicly available, and by those who developed advanced bioinformatics tools to analyse these genomic data [73,74,301].

5.3 Limitations

There were several limitations common to the analyses in this thesis. All analyses relied on the integrity of the underlying datasets, but some bias was inevitably present in both the pneumococcal and NPS datasets. In Chapters 2 and 3, whilst a large and diverse pneumococcal dataset was intentionally used, this represented only a subset of all of the global pneumococcal diversity and larger datasets of *S. pneumoniae* genomes have since been compiled [74,110]. In particular, the pneumococcal genome dataset used in Chapters 2 and 3 was biased towards carriage isolates collected in Europe, Asia and North America [73,74]. For the NPS species, there were large differences in the number of genomes that were available for different species in sequence read archives, with culturable, clinically relevant and industrially useful streptococci being over-represented [302,303,73,304]. Whilst upper limits of the number of genomes per species were imposed, there was still the potential for bias arising from the frequency of species in the datasets: in the NPS dataset used in Chapters 2 and 3 the frequency ranged from one to 180 genomes per species and in the *Streptococcus* dataset used in Chapter 4 the frequency ranged from one to 100 genomes per species.

A further limitation included end-of-contig hits in genomic regions of interest in short-read assemblies, despite the caution that had been taken to only include genomes that passed several assembly quality control metrics. This was of particular concern for genes which were present multiple times in the genomes and hence challenging to correctly assemble. Such genes included *lytA* which has homologs in prophages, which are often present in multiple copies in streptococcal genomes [62,63]. Furthermore, the copy numbers of the 16S rRNA gene vary

throughout prokaryotes and the 16S rRNA gene is known to be present multiple times in streptococci [305,306]. In short-read assemblies, extraction of a full 16S rRNA gene is therefore challenged by assembly errors and contig breaks [307]. A solution to this challenge would be to use long-read sequencing, particularly when using a hybrid assembly method that combines long-read and short-read data [186].

5.4 Future work

There are several avenues of further study which build on the research presented in this thesis. These include applications of the new PCR- and LAMP-based pneumococcal diagnostics, leveraging the *in silico* approaches to improve other diagnostic assays, and using sequencing data to address unanswered questions in the field.

5.4.1 Validation of diagnostic assays to directly test clinical samples

Future work includes the validation of the new qPCR and LAMP oligonucleotides for direct testing of clinical samples. Direct testing excludes a prior culture step: this is referred to as direct PCR for PCR-based analyses and direct LAMP for LAMP-based analyses [308,309]. In this thesis, the PCR- and LAMP- based assays were validated on DNA extracted from pure bacterial cultures. Further *in vitro* validation of these assays on clinical specimens such as blood, CSF, nasopharyngeal swabs, respiratory fluid, or saliva is crucial because the more complex molecular environment of these samples has the potential to influence both sensitivity and specificity [310]. Clinical samples, particularly blood, often contain compounds that inhibit molecular diagnostics and some clinical specimens, particularly saliva, are highly polymicrobial [310,311]. Whilst LAMP has been shown to be less affected by inhibitors commonly present in clinical samples than PCR [310,312], validation of the performance of the newly developed oligonucleotides on clinical samples is key to assuring clinical effectiveness. In CSF samples, LAMP assays can identify the causative agent of bacterial meningitis and hold promise for the development of rapid point-of-care assays [200,313]. Molecular assays have been used to detect pneumococcal DNA

from saliva samples [314], which are faster, easier and less invasive to collect than nasopharyngeal swabs [315].

5.4.2 Diagnostic surveillance in the African Meningitis Belt

The incidence of bacterial meningitis is greatest in the African Meningitis Belt, with a major proportion due to invasive meningococcal disease [316]. However, pneumococcal meningitis also causes a substantial number of infections [316]; the pneumococcal PCR- and particularly the LAMP-based diagnostics offer fast, cheap and easy-to-perform assays, which could contribute towards the strengthening of diagnostic capacity in Africa [317]. This is key to measuring the burden of disease, informing public policy and strategically planning the implementation of vaccines [139]. These methods have advantages over WGS-based diagnostics, at least in the short to medium term. Challenges around resources, technology and infrastructure will need to be overcome prior to routine implementation of WGS methods to track the genomic epidemiology of pathogens in the African Meningitis Belt [318].

The *in silico* approaches used in Chapters 2 and 3 could also be applied to any other pathogen with sufficient WGS data available. Therefore, this method could inform the development of a multiplex assay targeting several causes of meningitis [139], similar to the commercially available multiplex PCR-based BioFire FilmArray Meningitis/Encephalitis (ME) assay, which targets 14 pathogens [319–321]. A highly sensitive, specific, rapid and open-access multiplex LAMP assay targeting the major pathogens responsible for meningitis could be beneficial for surveillance efforts in the African Meningitis Belt [139,209].

5.4.3 Applications of NAATs to monitor AMR

The *in silico* approaches used in Chapters 2 and 3 can be applied to improve other types of NAATs and to improve assays targeting other pathogens, AMR determinants or virulence genes. Traditionally, NAATs targeting AMR determinants are used to determine the presence of a gene and cannot assess the presence of point mutations [322]. However, this limitation can be overcome by performing melting curve analyses, where mismatches result in a shift in the T_m and allow interpretation of mutations causing AMR, as was demonstrated for *Neisseria gonorrhoeae* [323]. Furthermore, a direct qPCR assay for the identification of *N. gonorrhoeae* was used to measure its growth (through measuring C_q-values) in medium containing antimicrobials and compared to growth without antimicrobials, which allowed determination of the AMR phenotype of the pathogen through a dose-response model [324]. When integrated into point-of-care assays, NAATs could provide clinicians with information to prescribe targeted antimicrobial therapy and avoid unnecessary broad-spectrum antibiotics, which is one of the pillars in combatting the spread of AMR globally [325].

5.4.4 Differentiation of viridans group streptococci

Viridans streptococci are often classed as one large group when investigating disease burden, for example when DeSimone et al. (2015) looked at viridans group streptococci causing infective endocarditis [212]. Nevertheless, accurate differentiation of viridans streptococci to a species-level is crucial to better understand the disease burden and pathogenic potential of individual species. Despite frequently being used to type streptococci from clinical isolates in hospital settings, MALDI-TOF MS has low resolution for differentiating viridans streptococci to a species level [326]. In contrast, as demonstrated in Chapter 4, core genome

typing using WGS data allowed species-specific clustering of viridans streptococci. Future studies could use core genome differentiation of viridans streptococci to a species-level, paired with clinical data, for example from electronic health records, to investigate the potential pathogenic differences across species.

Metagenomic sequencing could be applied to better understand the oral microbiome, including viridans streptococci. Metagenomic sequencing provides a culture-independent, microbe-agnostic detection method [327]. Using this technique could help overcome one of the limitations of this thesis regarding bias in dataset composition, by providing a more comprehensive representation of the oral microbiome. Furthermore, metagenomics has the potential to identify new species [328]. Metagenomic sequencing can also be used to study ancient oral microbiome samples to investigate historical *Streptococcus* species, as was done by Achtman and Zhou (2020) [329] and Belman et al. (2022) [21].

5.4.5 Application of WGS in pneumococcal vaccine design and beyond

Beyond using bacterial WGS data to improve diagnostic assays, pathogen genomics also facilitates the study of transmission, evolution and epidemiology [92]. For *S. pneumoniae* in particular, genomic epidemiology has provided an invaluable method to monitor pre- and post-vaccine population structure and to inform which serotypes should be included in new pneumococcal vaccines [55]. Existing pneumococcal vaccines target serotypes and pneumococcal capsular switching facilitates vaccine evasion [55]. WGS investigations could potentially help identify a conserved protein expressed on the surface of pneumococci across all serotypes; this would be an ideal pneumococcal vaccine target, as the vaccine effectiveness would not be serotype-dependent [330]. Similarly, conserved surface proteins may be useful potential targets for vaccine development in other species.

5.5 Overall conclusions

Overall, this thesis leveraged genomic data to improve the differentiation of *S. pneumoniae* from other *Streptococcus* species, both through direct application in diagnostic assays and through phylogenetic analyses. The new multiplex qPCR and real-time LAMP assays presented in this thesis hold promise to improve the identification of *S. pneumoniae* in both clinical and research microbiology. The synergetic *in silico* plus *in vitro* approach can act as a template for the development of assays targeting other pathogens [101]. The core genome genus-wide phylogeny presented high-resolution insight into the population structure of the *Streptococcus* genus and was used as a framework onto which the presence of AMR and virulence factors was mapped. Genomic epidemiology offers avenues to improve the diagnostic assays and differentiation tools used in microbiology, facilitating enhanced detection and understanding of major human pathogens, such as *S. pneumoniae*.

References

1. Butler MEB. Genomic and experimental investigations into pneumococcal bacteriocins and their role in competition. Imperial College, London; 2022.

2. Ipcress Manual | EMBL's European Bioinformatics Institute [Internet]. [cited 2022 Sep 5]. Available from: https://www.ebi.ac.uk/about/vertebrate-genomics/software/ipcress-manual

3. Ferretti J, Köhler W. History of streptococcal research. In: Ferretti JJ, Stevens DL, Fischetti VA, editors. *Streptococcus pyogenes*: basic biology to clinical manifestations. Oklahoma City (OK): University of Oklahoma Health Sciences Center; 2016.

4. Etymologia: *Streptococcus*. Emerg Infect Dis. 2016 Nov;22(11):1977.

5. Parte AC. LPSN—list of prokaryotic names with standing in nomenclature. Nucl Acids Res. 2014 Jan;42(D1):D613–6.

6. Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB. Bergey's manual of systematic bacteriology: volume 3: the Firmicutes. Springer Science & Business Media; 2011. p. 655-711.

7. Patterson MJ. *Streptococcus*. In: Baron S, editor. Medical Microbiology. 4th ed. Galveston (TX): University of Texas Medical Branch at Galveston; 1996.

8. Facklam R. What happened to the streptococci: overview of taxonomic and nomenclature changes. Clin Microbiol Rev. 2002 Oct;15(4):613–30.

9. Doern CD, Burnham CAD. It's not easy being green: the viridans group streptococci, with a focus on pediatric clinical manifestations. J Clin Microbiol. 2010 Nov;48(11):3829–35.

10. Parks T, Barrett L, Jones N. Invasive streptococcal disease: a review for clinicians. British Medical Bulletin. 2015 Sep 1;115(1):77–89.

11. Lancefield RC. A serological differentiation of human and other groups of hemolytic streptococci. J Exp Med. 1933 Mar 31;57(4):571–95.

12. Köhler W. The present state of species within the genera *Streptococcus* and *Enterococcus*. Int J Med Microbiol. 2007 Jun;297(3):133–50.

13. Efstratiou A, Lamagni T. Epidemiology of *Streptococcus pyogenes*. In: Ferretti JJ, Stevens DL, Fischetti VA, editors. *Streptococcus pyogenes*: basic biology to clinical manifestations. 2nd ed. Oklahoma City (OK): University of Oklahoma Health Sciences Center; 2022.

14. Raabe VN, Shane AL. Group B *Streptococcus* (*Streptococcus agalactiae*). Microbiology Spectrum. 2019 Mar 22;7(2):10.1128/microbiolspec.gpp3-0007–2018.

15. Timoney JF. *Streptococcus*. In: Pathogenesis of bacterial infections in animals. John Wiley & Sons, Ltd; 2022. p. 565–87.

16. Coykendall AL. Classification and identification of the viridans streptococci. Clin Microbiol Rev. 1989 Jul;2(3):315–28.

17. Feng Y, Zhang H, Wu Z, Wang S, Cao M, Hu D, Wang C. *Streptococcus suis* infection. Virulence. 2014 May 15;5(4):477–97.

18. Leigh JA. *Streptococcus uberis*: a permanent barrier to the control of bovine mastitis? The Veterinary Journal. 1999 May 1;157(3):225–38.

19. Krzyściak W, Pluskwa KK, Jurczak A, Kościelniak D. The pathogenicity of the *Streptococcus* genus. Eur J Clin Microbiol Infect Dis. 2013 Nov 1;32(11):1361–76.

20. Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, Sørensen UBS. Evolution of *Streptococcus pneumoniae* and its close commensal relatives. Plos One. 2008 Jul 16;3(7):e2683.

21. Belman S, Chaguza C, Kumar N, Lo S, Bentley SD. A new perspective on ancient mitis group streptococcal genetics. Microbial Genomics. 2022;8(2):000753.

22. Watson DA, Musher DM, Jacobson JW, Verhoef J. A brief history of the pneumococcus in biomedical research: a panoply of scientific discovery. Clin Infect Dis. 1993 Nov;17(5):913–24.

23. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. J Exp Med. 1944 Feb 1;79(2):137–58.

24. Weiser JN, Ferreira DM, Paton JC. *Streptococcus pneumoniae*: transmission, colonization and invasion. Nat Rev Microbiol. 2018 Jun;16(6):355–67.

25. Henriques-Normark B, Tuomanen EI. The pneumococcus: epidemiology, microbiology, and pathogenesis. Cold Spring Harb Perspect Med. 2013 Jul;3(7):a010215.

26. Bogaert D, Groot R de, Hermans PWM. *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. The Lancet Infectious Diseases. 2004 Mar 1;4(3):144–54.

27. Neal EFG, Chan J, Nguyen CD, Russell FM. Factors associated with pneumococcal nasopharyngeal carriage: a systematic review. PLOS Glob Public Health. 2022 Apr 11;2(4):e0000327.

28. Usuf E, Bottomley C, Adegbola RA, Hall A. Pneumococcal carriage in Sub-Saharan Africa—a systematic review. Plos One. 2014 Jan 20;9(1):e85001.

29. Regev-Yochay G, Raz M, Dagan R, Porat N, Shainberg B, Pinco E, Keller N, Rubinstein E. Nasopharyngeal carriage of *Streptococcus pneumoniae* by adults and children in community and family settings. Clin Infect Dis. 2004 Mar 1;38(5):632–9.

30. Mehr S, Wood N. *Streptococcus pneumoniae*--a review of carriage, infection, serotype replacement and vaccination. Paediatr Respir Rev. 2012 Dec;13(4):258–64.

31. Mackenzie GA, Leach AJ, Carapetis JR, Fisher J, Morris PS. Epidemiology of nasopharyngeal carriage of respiratory bacterial pathogens in children and adults: cross-sectional surveys in a population with high rates of pneumococcal disease. BMC Infect Dis. 2010 Oct 23;10(1):304.

32. Wyllie AL, Rümke LW, Arp K, Bosch AATM, Bruin JP, Rots NY, Wijmenga-Monsuur AJ, Sanders EAM, Trzciński K. Molecular surveillance on *Streptococcus pneumoniae* carriage in non-elderly adults; little evidence for pneumococcal circulation independent from the reservoir in children. Sci Rep. 2016 Oct 7;6:34888.

33. Numminen E, Chewapreecha C, Turner C, Goldblatt D, Nosten F, Bentley SD, Turner P, Corander J. Climate induces seasonality in pneumococcal transmission. Sci Rep. 2015 Jun 12;5:11344.

34. Numminen E, Chewapreecha C, Turner C, Goldblatt D, Nosten F, Bentley SD, Turner P, Corander J. Corrigendum: Climate induces seasonality in pneumococcal transmission. Sci Rep. 2016 Mar 29;6:23307.

35. Ortqvist A, Hedlund J, Kalin M. *Streptococcus pneumoniae*: epidemiology, risk factors, and clinical features. Semin Respir Crit Care Med. 2005 Dec;26(6):563–74.

36. Marangu D, Zar HJ. Childhood pneumonia in low-and-middle-income countries: an update. Paediatr Respir Rev. 2019 Nov;32:3–9.

37. Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y, Lukšić I, Nair H, McAllister DA, Campbell H, Rudan I, Black R, Knoll MD. Burden of *Streptococcus pneumoniae* and *Haemophilus influenzae* type b disease in children in the era of conjugate vaccines: global, regional, and

national estimates for 2000-15. Lancet Glob Health. 2018 Jul;6(7):e744–57.

38. GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016 Oct 8;388(10053):1459–544.

39. Troeger C, Blacker B, Khalil IA, Rao PC, Cao J, Zimsen SRM, Albertson SB, Deshpande A, Farag T, Abebe Z, Adetifa IMO, Adhikari TB, Akibu M, Al Lami FH, Al-Eyadhy A, Alvis-Guzman N, Amare AT, Amoako YA, Antonio CAT, Aremu O, Asfaw ET, Asgedom SW, Atey TM, Attia EF, Avokpaho EFGA, Ayele HT, Ayuk TB, Balakrishnan K, Barac A, Bassat Q, Behzadifar M, Behzadifar M, Bhaumik S, Bhutta ZA, Bijani A, Brauer M, Brown A, Camargos PAM, Castañeda-Orjuela CA, Colombara D, Conti S, Dadi AF, Dandona L, Dandona R, Do HP, Dubljanin E, Edessa D, Elkout H, Endries AY, Fijabi DO, Foreman KJ, Forouzanfar MH, Fullman N, Garcia-Basteiro AL, Gessner BD, Gething PW, Gupta R, Gupta T, Hailu GB, Hassen HY, Hedayati MT, Heidari M, Hibstu DT, Horita N, Ilesanmi OS, Jakovljevic MB, Jamal AA, Kahsay A, Kasaeian A, Kassa DH, Khader YS, Khan EA, Khan MN, Khang YH, Kim YJ, Kissoon N, Knibbs LD, Kochhar S, Koul PA, Kumar GA, Lodha R, Magdy Abd El Razek H, Malta DC, Mathew JL, Mengistu DT, Mezgebe HB, Mohammad KA, Mohammed MA, Momeniha F, Murthy S, Nguyen CT, Nielsen KR, Ningrum DNA, Nirayo YL, Oren E, Ortiz JR, Pa M, Postma MJ, Qorbani M, Quansah R, Rai RK, Rana SM, Ranabhat CL, Ray SE, Rezai MS, Ruhago GM, Safiri S, Salomon JA, Sartorius B, Savic M, Sawhney M, She J, Sheikh A, Shiferaw MS, Shigematsu M, Singh JA, Somayaji R, Stanaway JD, Sufiyan MB, Taffere GR, Temsah MH, Thompson MJ, Tobe-Gai R, Topor-Madry R, Tran BX, Tran TT, Tuem KB, Ukwaja KN, Vollset SE, Walson JL, Weldegebreal F, Werdecker A, West TE, Yonemoto N, Zaki MES, Zhou L, Zodpey S, Vos T, Naghavi M, Lim SS, Mokdad AH, Murray CJL, Hay SI, Reiner RC. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet Infectious Diseases. 2018 Nov 1;18(11):1191–210.

40. Kwambana-Adams B. Global burden of meningitis and implications for strategy. Lancet Neurol. 2023 Aug;22(8):646–8.

41. Brueggemann AB, Rensburg MJJ van, Shaw D, McCarthy ND, Jolley KA, Maiden MCJ, Linden MPG van der, Amin-Chowdhury Z, Bennett DE, Borrow R, Brandileone MCC, Broughton K, Campbell R, Cao B, Casanova C, Choi EH, Chu YW, Clark SA, Claus H, Coelho J, Corcoran M, Cottrell S, Cunney RJ, Dalby T, Davies H, Gouveia L de, Deghmane AE, Demczuk W, Desmet S, Drew RJ, Plessis M du, Erlendsdottir H, Fry NK, Fuursted K, Gray SJ, Henriques-Normark B, Hale T, Hilty M, Hoffmann S, Humphreys H, Ip M, Jacobsson S, Johnston J, Kozakova J,

Kristinsson KG, Krizova P, Kuch A, Ladhani SN, Lâm TT, Lebedova V, Lindholm L, Litt DJ, Martin I, Martiny D, Mattheus W, McElligott M, Meehan M, Meiring S, Mölling P, Morfeldt E, Morgan J, Mulhall RM, Muñoz-Almagro C, Murdoch DR, Murphy J, Musilek M, Mzabi A, Perez-Argüello A, Perrin M, Perry M, Redin A, Roberts R, Roberts M, Rokney A, Ron M, Scott KJ, Sheppard CL, Siira L, Skoczyńska A, Sloan M, Slotved HC, Smith AJ, Song JY, Taha MK, Toropainen M, Tsang D, Vainio A, Sorge NM van, Varon E, Vlach J, Vogel U, Vohrnova S, Gottberg A von, Zanella RC, Zhou F. Changes in the incidence of invasive disease due to *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Neisseria meningitidis* during the COVID-19 pandemic in 26 countries and territories in the Invasive Respiratory Infection Surveillance Initiative: a prospective analysis of surveillance data. The Lancet Digital Health. 2021 Jun 1;3(6):e360–70.

42. Shaw D, Abad R, Amin-Chowdhury Z, Bautista A, Bennett D, Broughton K, Cao B, Casanova C, Choi EH, Chu YW, Claus H, Coelho J, Corcoran M, Cottrell S, Cunney R, Cuypers L, Dalby T, Davies H, Gouveia L de, Deghmane AE, Demczuk W, Desmet S, Domenech M, Drew R, Plessis M du, Duarte C, Erlendsdóttir H, Fry NK, Fuursted K, Hale T, Henares D, Henriques-Normark B, Hilty M, Hoffmann S, Humphreys H, Ip M, Jacobsson S, Johnson C, Johnston J, Jolley KA, Kawabata A, Kozakova J, Kristinsson KG, Krizova P, Kuch A, Ladhani S, Lâm TT, León ME, Lindholm L, Litt D, Maiden MCJ, Martin I, Martiny D, Mattheus W, McCarthy ND, Meehan M, Meiring S, Mölling P, Morfeldt E, Morgan J, Mulhall R, Muñoz-Almagro C, Murdoch D, Murphy J, Musilek M, Mzabi A, Novakova L, Oftadeh S, Perez-Argüello A, Pérez-Vázquez M, Perrin M, Perry M, Prevost B, Roberts M, Rokney A, Ron M, Sanabria OM, Scott KJ, Sheppard C, Siira L, Sintchenko V, Skoczyńska A, Sloan M, Slotved HC, Smith AJ, Steens A, Taha MK, Toropainen M, Tzanakaki G, Vainio A, Linden MPG van der, Sorge NM van, Varon E, Vohrnova S, Gottberg A von, Yuste J, Zanella R, Zhou F, Brueggemann AB. Trends in invasive bacterial diseases during the first 2 years of the COVID-19 pandemic: analyses of prospective surveillance data from 30 countries and territories in the IRIS Consortium. The Lancet Digital Health. 2023 Sep 1;5(9):e582–93.

43. Griffith F. The significance of pneumococcal types. J Hyg (Lond). 1928 Jan;27(2):113–59.

44. Andre GO, Converso TR, Politano WR, Ferraz LFC, Ribeiro ML, Leite LCC, Darrieux M. Role of *Streptococcus pneumoniae* proteins in evasion of complement-mediated immunity. Front Microbiol. 2017 Feb 20;8:224.

45. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen HB, Nahm MH. Pneumococcal capsules and their types: past, present, and future. Clin Microbiol Rev. 2015 Jul;28(3):871–99.

46. Geno KA, Gilbert GL, Song JY, Skovsted IC, Klugman KP, Jones C, Konradsen HB, Nahm MH. Erratum: Pneumococcal capsules and their

types: past, present, and future. Clin Microbiol Rev. 2021 Mar 17;34(2):e00320-20.

47. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW, Gladstone RA, Turner P, Keenan JD, Breiman RF, Nahm MH. A new pneumococcal capsule type, 10D, is the 100th serotype and has a large *cps* fragment from an oral *Streptococcus*. mBio. 2020 May 19;11(3):10.1128/mbio.00937-20.

48. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. J Infect Dis. 2003 May 1;187(9):1424–32.

49. Sheppard CL, Manna S, Groves N, Litt DJ, Amin-Chowdhury Z, Bertran M, Ladhani S, Satzke C, Fry NK. PneumoKITy: A fast, flexible, specific, and sensitive tool for *Streptococcus pneumoniae* serotype screening and mixed serotype detection from genome sequence data. Microbial Genomics. 2022;8(12):000904.

50. Epping L, van Tonder AJ, Gladstone RA, Bentley SD, Page AJ, Keane JA. SeroBA: rapid high-throughput serotyping of *Streptococcus pneumoniae* from whole genome sequence data. Microb Genom. 2018 Jun 15;4(7):e000186.

51. Grabenstein JD, Klugman KP. A century of pneumococcal vaccination research in humans. Clin Microbiol Infect. 2012 Oct;18 Suppl 5:15–24.

52. Kim NH, Lee J, Lee SJ, Lee H, Kim KH, Park SE, Lee HJ. Immunogenicity and safety of pneumococcal 7-valent conjugate vaccine (diphtheria CRM_{197} protein conjugate; Prevenar) in Korean infants: differences that are found in Asian children. Vaccine. 2007 Nov 7;25(45):7858–65.

53. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, Liñares J, Jacobs MR, Kristinsson KG, Beall BW, Klugman KP, Parkhill J, Hakenbeck R, Bentley SD, Brueggemann AB. Pneumococcal capsular switching: a historical perspective. J Infect Dis. 2013 Feb 1;207(3):439–49.

54. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease following pneumococcal vaccination: a discussion of the evidence. Lancet. 2011 Dec 3;378(9807):1962–73.

55. Bentley SD, Lo SW. Global genomic pathogen surveillance to inform vaccine strategies: a decade-long expedition in pneumococcal genomics. Genome Med. 2021 May 17;13:84.

56. Brito IL. Examining horizontal gene transfer in microbial communities. Nat Rev Microbiol. 2021 Jul;19(7):442–53.

57. Chaguza C, Cornick JE, Everett DB. Mechanisms and impact of genetic recombination in the evolution of *Streptococcus pneumoniae*. Comput Struct Biotechnol J. 2015 Apr 8;13:241–7.

58. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022 Apr;20(4):206–18.

59. Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD. A high-resolution view of genome-wide pneumococcal transformation. PLOS Pathogens. 2012 Jun 14;8(6):e1002745.

60. Chewapreecha C, Harris SR, Croucher NJ, Turner C, Marttinen P, Cheng L, Pessia A, Aanensen DM, Mather AE, Page AJ, Salter SJ, Harris D, Nosten F, Goldblatt D, Corander J, Parkhill J, Turner P, Bentley SD. Dense genomic sampling identifies highways of pneumococcal recombination. Nat Genet. 2014 Mar;46(3):305–9.

61. 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 Aug;187(15):5376–86.

62. Rezaei Javan R, Ramos-Sevillano E, Akter A, Brown J, Brueggemann AB. Prophages and satellite prophages are widespread in *Streptococcus* and may play a role in pneumococcal pathogenesis. Nat Commun. 2019 Oct 24;10(1):4852.

63. Brueggemann AB, Harrold CL, Rezaei Javan R, van Tonder AJ, McDonnell AJ, Edwards BA. Pneumococcal prophages are diverse, but not without structure or history. Sci Rep. 2017 Feb 20;7(1):42976.

64. Song JY, Eun BW, Nahm MH. Diagnosis of pneumococcal pneumonia: current pitfalls and the way forward. Infect Chemother. 2013 Dec;45(4):351–66.

65. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977 Dec;74(12):5463–7.

66. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016 Jan;107(1):1–8.

67. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequence analysis. Nature. 1986 Jun 12;321(6071):674–9.

68. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh, W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L lng, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom

E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995 Jul 28;269(5223):496–512.

69. *Streptococcus pneumoniae* TIGR4, complete genome [Internet]. 2014 [cited 2024 Feb 22]. Available from: http://www.ncbi.nlm.nih.gov/nuccore/AE005672.3

70. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, Durkin AS, Gwinn M, Kolonay JF, Nelson WC, Peterson JD, Umayam LA, White O, Salzberg SL, Lewis MR, Radune D, Holtzapple E, Khouri H, Wolf AM, Utterback TR, Hansen CL, McDonald LA, Feldblyum TV, Angiuoli S, Dickinson T, Hickey EK, Holt IE, Loftus BJ, Yang F, Smith HO, Venter JC, Dougherty BA, Morrison DA, Hollingshead SK, Fraser CM. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science. 2001 Jul 20;293(5529):498–506.

71. Santoro F, Iannelli F, Pozzi G. Genomics and genetics of *Streptococcus pneumoniae*. Microbiol Spectr. 2019 May;7:10.1128/microbiolspec.gpp3-0025-2018.

72. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, Thomson NR, Iqbal Z. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLOS Biology. 2021 Nov 9;19(11):e3001421.

73. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.

74. Jansen Van Rensburg MJ, Berger DJ, Fohrmann A, Bray JE, Jolley KA, Maiden MCJ, Brueggemann AB. Development of the Pneumococcal Genome Library, a core genome multilocus sequence typing scheme, and a taxonomic life identification number barcoding system to investigate and define pneumococcal population structure [Internet]. preprint; 2023 Dec. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.12.19.571883

75. Forde BM, O'Toole PW. Next-generation sequencing technologies and their impact on microbial genomics. Briefings in Functional Genomics. 2013 Sep 1;12(5):440–53.

76. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. 2010 Jan;11(1):31–46.

77. Nyrén P. The history of pyrosequencing. Methods Mol Biol. 2007;373:1–14.

78. Donkor ES. Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics. Genes (Basel). 2013 Oct 14;4(4):556–72.

79. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, deWinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2;323(5910):133–8.

80. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nat Rev Microbiol. 2015 Dec;13(12):787–94.

81. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends Genet. 2018 Sep;34(9):666–81.

82. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Feb 7;21(1):30.

83. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15(2):141–61.

84. Croucher NJ, Harris SR, Grad YH, Hanage WP. Bacterial genomes in epidemiology—present and future. Philos Trans R Soc Lond B Biol Sci. 2013 Mar 19;368(1614):20120202.

85. Didelot X, Falush D. Inference of bacterial microevolution using multilocus sequence data. Genetics. 2007 Mar 1;175(3):1251–66.

86. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015 Feb 12;11(2):e1004041.

87. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research. 2015 Feb 18;43(3):e15–e15.

88. Thrash A, Hoffmann F, Perkins A. Toward a more holistic method of genome assembly assessment. BMC Bioinformatics. 2020 Jul 6;21(4):249.

89. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology (Reading). 2012 Apr;158(Pt 4):1005–15.

90. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012 Sep;13(9):601–12.

91. GenBank and WGS Statistics [Internet]. [cited 2024 May 5]. Available from: https://www.ncbi.nlm.nih.gov/genbank/statistics/

92. Eyre DW. Infection prevention and control insights from a decade of pathogen whole-genome sequencing. Journal of Hospital Infection. 2022 Apr;122:180–6.

93. Uelze L, Grützke J, Borowiak M, Hammerl JA, Juraschek K, Deneke C, Tausch SH, Malorny B. Typing methods based on whole genome sequencing data. One Health Outlook. 2020 Feb 18;2(1):3.

94. Haeckel E. Generelle Morphologie der Organismen Vols I and II. Georg Reimer; 1866.

95. Kutschera U. Haeckel's 1866 tree of life and the origin of eukaryotes. Nat Microbiol. 2016 Jul 26;1(8):1–1.

96. Gregory TR. Understanding Evolutionary Trees. Evo Edu Outreach. 2008 Apr;1(2):121–37.

97. Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London, UK: Murray; 1859.

98. Zvelebil M, Baum JO. Understanding bioinformatics. Garland Science; 2007. p. 223–313

99. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet. 2018;19(1):9–20.

100. He Y, Ong E, Huffman A. Chapter 26 - Bacterial whole-genome determination and applications. In: Tang YW, Hindiyeh MY, Liu D, Sails A, Spearman P, Zhang JR, editors. Molecular Medical Microbiology (Third Edition) Academic Press; 2024 p. 511–25.

101. Jansen van Rensburg MJ, Swift C, Cody AJ, Jenkins C, Maiden MCJ. Exploiting bacterial whole-genome sequencing data for evaluation of diagnostic assays: *Campylobacter* species identification as a case study. Journal of Clinical Microbiology. 2016 Dec;54(12):2882–90.

102. Seth-Smith HMB, Thomson NR. Whole-genome sequencing of bacterial sexually transmitted infections: implications for clinicians. Current Opinion in Infectious Diseases. 2013 Feb;26(1):90.

103. Ripa T, Nilsson P. A variant of *Chlamydia trachomatis* with deletion in cryptic plasmid: implications for use of PCR diagnostic tests. Weekly releases (1997–2007). 2006 Nov 9;11(45):3076.

104. Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broeker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R. Identification of vaccine candidates against serogroup B *meningococcus* by whole-genome sequencing. Science. 2000 Mar 10;287(5459):1816–20.

105. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, Barash SC, Rosen CA, Masure HR, Tuomanen E, Gayle A, Brewah YA, Walsh W, Barren P, Lathigra R, Hanson M, Langermann S, Johnson S, Koenig S. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. Infection and Immunity. 2001 Mar;69(3):1593–8.

106. Simar SR, Hanson BM, Arias CA. Techniques in bacterial strain typing: past, present, and future. Curr Opin Infect Dis. 2021 Aug 1;34(4):339–45.

107. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TEA, Walker AS. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med. 2013 Sep 26;369(13):1195–205.

108. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 2015 Sep 1;31(17):2877–8.

109. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res. 2019 Feb;29(2):304–16.

110. Gladstone RA, Lo SW, Lees JA, Croucher NJ, Tonder AJ van, Corander J, Page AJ, Marttinen P, Bentley LJ, Ochoa TJ, Ho PL, Plessis M du, Cornick JE, Kwambana-Adams B, Benisty R, Nzenze SA, Madhi SA, Hawkins PA, Everett DB, Antonio M, Dagan R, Klugman KP, Gottberg A von, McGee L, Breiman RF, Bentley SD. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. eBioMedicine. 2019 May 1;43:338–46.

111. Gladstone RA, Lo SW, Goater R, Yeats C, Taylor B, Hadfield J, Lees JA, Croucher NJ, van Tonder AJ, Bentley LJ, Quah FX, Blaschke AJ, Pershing NL, Byington CL, Balaji V, Hryniewicz W, Sigauque B, Ravikumar KL, Almeida SCG, Ochoa TJ, Ho PL, du Plessis M, Ndlangisa KM, Cornick JE, Kwambana-Adams B, Benisty R, Nzenze SA, Madhi SA, Hawkins PA, Pollard AJ, Everett DB, Antonio M, Dagan R, Klugman KP, von Gottberg A, Metcalf BJ, Li Y, Beall BW, McGee L, Breiman RF, Aanensen DM, Bentley SD, The Global Pneumococcal Sequencing Consortium. Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates. Microb Genom. 2020 May;6(5):e000357.

112. Maiden MCJ, van Rensburg MJJ, Bray JE, Earle SG, Ford SA, Jolley KA, McCarthy ND. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728–36.

113. Eyre DW, Peto TEA, Crook DW, Walker AS, Wilcox MH. Hash-based core genome multilocus sequence typing for *Clostridium difficile*. J Clin Microbiol. 2019 Dec 23;58(1):e01037-19.

114. Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology (Reading). 1998 Nov;144 (Pt 11):3049–60.

115. King AC, Kumar N, Mellor KC, Hawkins PA, McGee L, Croucher NJ, Bentley SD, Lees JA, Lo SW. Comparison of gene-by-gene and genome-wide short nucleotide sequence based approaches to define the global population structure of *Streptococcus pneumoniae* [Internet]. 2024 [cited 2024 Jul 6]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.05.29.596230

116. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998 Mar 17;95(6):3140–5.

117. Geneious Prime version 2022.1.1 [Internet]. Available from: https://www.geneious.com

118. *Streptococcus pneumoniae* [Internet]. PubMLST. [cited 2024 Mar 20]. Available from: https://pubmlst.org/organisms/streptococcus-pneumoniae

119. Enright MC, Spratt BG. Extensive variation in the *ddl* gene of penicillin-resistant *Streptococcus pneumoniae* results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. Mol Biol Evol. 1999 Dec;16(12):1687–95.

120. Feil EJ, Enright MC, Spratt BG. Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between *Neisseria meningitidis* and *Streptococcus pneumoniae*. Research in Microbiology. 2000 Jul 8;151(6):465–9.

121. Ej F, Bc L, Dm A, Wp H, Bg S. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. Journal of bacteriology. 2004 Mar;186(5).

122. Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009 May 18;10(1):152.

123. 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. 2017 Jan 1;33(1):128–9.

124. 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 May 8;13(1):87.

125. Jansen van Rensburg MJ. The genomic epidemiology of *Campylobacter* from the Republic of South Africa. University of Oxford; 2015.

126. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020 Jul;21(7):428–44.

127. Challa S, Neelapu NRR. Phylogenetic trees: applications, construction, and assessment. In: Hakeem KR, Shaik NA, Banaganapalli B, Elango R, editors. Essentials of bioinformatics, volume III: In Silico Life Sciences: Agriculture. Cham: Springer International Publishing; 2019. p. 167–92.

128. Tavares DA, Handem S, Carvalho RJ, Paulo AC, de Lencastre H, Hinds J, Sá-Leão R. Identification of *Streptococcus pneumoniae* by a real-time PCR assay targeting SP2020. Sci Rep. 2019 Dec;9(1):3285.

129. Jensen A, Scholz CFP, Kilian M 2016. Re-evaluation of the taxonomy of the mitis group of the genus *Streptococcus* based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. International Journal of Systematic and Evolutionary Microbiology. 2016;66(11):4803–20.

130. Satzke C, Turner P, Virolainen-Julkunen A, Adrian PV, Antonio M, Hare KM, Henao-Restrepo AM, Leach AJ, Klugman KP, Porter BD, Sá-Leão R, Scott JA, Nohynek H, O'Brien KL. Standard method for detecting upper

respiratory carriage of *Streptococcus pneumoniae*: updated recommendations from the World Health Organization pneumococcal carriage working group. Vaccine. 2013 Dec 17;32(1):165–79.

131. Richter SS, Heilmann KP, Dohrn CL, Riahi F, Beekmann SE, Doern GV. Accuracy of phenotypic methods for identification *of Streptococcus pneumoniae* isolates included in surveillance programs. J Clin Microbiol. 2008 Jul;46(7):2184–8.

132. Arbique JC, Poyart C, Trieu-Cuot P, Quesne G, Carvalho M da GS, Steigerwalt AG, Morey RE, Jackson D, Davidson RJ, Facklam RR. Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. Journal of Clinical Microbiology. 2004 Oct;42(10):4686–96.

133. Sadowy E, Hryniewicz W. Identification of *Streptococcus pneumoniae* and other mitis streptococci: importance of molecular methods. Eur J Clin Microbiol Infect Dis. 2020 Dec;39(12):2247–56.

134. Ikryannikova LN, Lapin KN, Malakhova MV, Filimonova AV, Ilina EN, Dubovickaya VA, Sidorenko SV, Govorun VM. Misidentification of alpha-hemolytic streptococci by routine tests in clinical practice. Infection, Genetics and Evolution. 2011 Oct 1;11(7):1709–15.

135. Werno AM, Christner M, Anderson TP, Murdoch DR. Differentiation of *Streptococcus pneumoniae* from nonpneumococcal streptococci of the *Streptococcus mitis* group by matrix-assisted laser desorption ionization–time of flight mass spectrometry. Journal of Clinical Microbiology. 2012;50(9):2863–7.

136. De Bel A, Wybo I, Piérard D, Lauwers S. Correct implementation of matrix-assisted laser desorption ionization-time of flight mass spectrometry in routine clinical microbiology. Journal of Clinical Microbiology. 2010 May;48(5):1991–2.

137. Roberts T, Chandna A, Watthanaworawit W, Thaiprakong A, Soeng S, Simmalavong M, Phoumin P, Saengchun W, Khatta N, Hinfonthong P, Kaewpundoem N, Lee SJ, Perrone C, Amos B, Turner P, Ashley EA, Ling CL. Impact of delayed processing of positive blood cultures on organism detection: a prospective multi-centre study. BMC Infect Dis. 2022 Jun 1;22(1):517.

138. Pizzutti K, Perez VP, Barbiero C, d'Azevedo PA, Fischer GB, Dias C. Identifying pneumococci in parapneumonic pleural effusion: is there a role for culture-independent methods? Pediatric Pulmonology. 2020;55(2):484–9.

139. Diallo K, Feteh VF, Ibe L, Antonio M, Caugant DA, du Plessis M, Deghmane AE, Feavers IM, Fernandez K, Fox LM, Rodrigues CMC,

Ronveaux O, Taha MK, Wang X, Brueggemann AB, Maiden MCJ, Harrison OB. Molecular diagnostic assays for the detection of common bacterial meningitis pathogens: a narrative review. EBioMedicine. 2021 Mar;65:103274.

140. Garibyan L, Avashia N. Polymerase chain reaction. Journal of Investigative Dermatology. 2013 Mar 1;133(3):1–4.

141. Artika IM, Dewi YP, Nainggolan IM, Siregar JE, Antonjaya U. Real-time polymerase chain reaction: current techniques, applications, and role in COVID-19 diagnosis. Genes (Basel). 2022 Dec 16;13(12):2387.

142. Buchan BW, Ledeboer NA. Emerging technologies for the clinical microbiology laboratory. Clinical Microbiology Reviews. 2014 Oct;27(4):783–822.

143. Shen CH. Chapter 9 - Amplification of nucleic acids. In: Shen CH, editor. Diagnostic molecular biology. Academic Press; 2019. p. 215–47.

144. Ruiz-Villalba A, Ruijter JM, van den Hoff MJB. Use and misuse of C_q in qPCR data analysis and reporting. Life (Basel). 2021 May 29;11(6):496.

145. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER, Masignani V. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biology. 2010 Oct 29;11(10):R107.

146. Carvalho M da GS, Tondella ML, McCaustland K, Weidlich L, McGee L, Mayer LW, Steigerwalt A, Whaley M, Facklam RR, Fields B, Carlone G, Ades EW, Dagan R, Sampson JS. Evaluation and improvement of real-time PCR assays targeting *lytA*, *ply*, and *psaA* genes for detection of pneumococcal DNA. J Clin Microbiol. 2007 Aug;45(8):2460–6.

147. *Streptococcus* Lab Resources and Protocols | CDC [Internet]. 2023 [cited 2024 Jun 12]. Available from: https://www.cdc.gov/streplab/pneumococcus/resources.html

148. Trzciński K, Bogaert D, Wyllie A, Chu MLJN, Ende A van der, Bruin JP, Dobbelsteen G van den, Veenhoven RH, Sanders EAM. Superiority of trans-oral over trans-nasal sampling in detecting *Streptococcus pneumoniae* colonization in adults. Plos One. 2013 Mar 28;8(3):e60520.

149. Suzuki N, Seki M, Nakano Y, Kiyoura Y, Maeno M, Yamashita Y. Discrimination of *Streptococcus pneumoniae* from viridans group streptococci by genomic subtractive hybridization. J Clin Microbiol. 2005 Sep;43(9):4528–34.

150. Ganaie FA, Govindan V, Ravi Kumar KL. Standardisation and evaluation of a quantitative multiplex real-time PCR assay for the rapid identification of *Streptococcus pneumoniae*. Pneumonia (Nathan). 2015 Dec 1;6:57–66.

151. Strålin K, Herrmann B, Abdeldaim G, Olcén P, Holmberg H, Mölling P. Comparison of sputum and nasopharyngeal aspirate samples and of the PCR gene targets *lytA* and Spn9802 for quantitative PCR for rapid detection of pneumococcal pneumonia. J Clin Microbiol. 2014 Jan;52(1):83–9.

152. Salvà-Serra F, Connolly G, Moore ERB, Gonzales-Siles L. Detection of "Xisco" gene for identification of *Streptococcus pneumoniae* isolates. Diagnostic Microbiology and Infectious Disease. 2018 Apr 1;90(4):248–50.

153. Croxen MA, Lee TD, Azana R, Hoang LMY 2018. Use of genomics to design a diagnostic assay to discriminate between *Streptococcus pneumoniae* and *Streptococcus pseudopneumoniae*. Microbial Genomics. 2018;4(7):e000175.

154. Belman S, Pesonen H, Croucher NJ, Bentley SD, Corander J. Estimating between country migration in pneumococcal populations. G3 Genes|Genomes|Genetics. 2024 Mar 20;jkae058.

155. Zhang D, Luo Y, Zeng X, Yu Y, Wu Y. Developing a multiplex PCR-based assay kit for bloodstream infection by analyzing genomic big data. J Clin Lab Anal. 2022 Oct;36(10):e24686.

156. Lemmon GH, Gardner SN. Predicting the sensitivity and specificity of published real-time PCR assays. Ann Clin Microbiol Antimicrob. 2008 Sep 25;7:18.

157. Ribosomal MLST [Internet]. PubMLST. [cited 2022 Aug 19]. Available from: https://pubmlst.org/species-id

158. Alhakami H, Mirebrahim H, Lonardi S. A comparative evaluation of genome assembly reconciliation tools. Genome Biol. 2017 May 18;18:93.

159. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. 2021 Jul 2;49(W1):W293–6.

160. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004 Mar 1;32(5):1792–7.

161. Youden WJ. Index for rating diagnostic tests. Cancer. 1950 Jan;3(1):32–5.

162. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

163. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455–77.

164. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, Ellington M, Swift C, Green J, Underwood A. MOST: a modified MLST typing tool based on short read sequencing. PeerJ. 2016;4:e2308.

165. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG, Fry NK. Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. PeerJ. 2016;4:e2477.

166. Sheppard CL, Groves N, Andrews N, Litt DJ, Fry NK, Southern J, Miller E. The genomics of *Streptococcus pneumoniae* carriage isolates from UK children and their household contacts, pre-PCV7 to post-PCV13. Genes (Basel). 2019 Sep 6;10(9):E687.

167. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

168. Badell E, Guillot S, Tulliez M, Pascal M, Panunzi LG, Rose S, Litt D, Fry NK, Brisse S. Improved quadruplex real-time PCR assay for the diagnosis of diphtheria. J Med Microbiol. 2019 Oct;68(10):1455–65.

169. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry. 2009 Apr 1;55(4):611–22.

170. R Core Team (2023). R: a language and environment for statistical computing. [Internet]. R Foundation for Statistical Computing, Vienna, Austria.; Available from: https://www.R-project.org/

171. Dorai-Raj S. binom: binomial confidence intervals for several parameterizations [Internet]. 2022 [cited 2024 Aug 2]. Available from: https://cran.r-project.org/web/packages/binom/index.html

172. Krassowski M, Arts M, Lagger C, Max. krassowski/complex-upset: v1.3.5 [Internet]. Zenodo; 2022 [cited 2024 Feb 25]. Available from: https://zenodo.org/records/7314197

173. Wilke CO. cowplot: streamlined plot theme and plot annotations for 'ggplot2' [Internet]. 2024 [cited 2024 Feb 28]. Available from: https://cran.r-project.org/web/packages/cowplot/index.html

174. Wickham H, François R, Henry L, Müller K, Vaughan D, Software P, PBC. dplyr: a grammar of data manipulation [Internet]. 2023 [cited 2024 Feb 28]. Available from: https://cran.r-project.org/web/packages/dplyr/index.html

175. Wickham H. Data analysis. In: Wickham H, editor. ggplot2: Elegant Graphics for data analysis [Internet]. Cham: Springer International Publishing; 2016 [cited 2024 Feb 25]. p. 189–201. Available from: https://doi.org/10.1007/978-3-319-24277-4_9

176. Wickham H, Hester J, Francois R, Bryan J, Bearrows S, Software P, PBC, Jylänki J, Jørgensen M. [Internet]. 2024 [cited 2024 Feb 28]. Available from: https://cran.r-project.org/web/packages/readr/index.html

177. Müller K, Wickham H, Francois R, Bryan J, RStudio. tibble: simple data frames [Internet]. 2023 [cited 2024 Feb 28]. Available from: https://cran.r-project.org/web/packages/tibble/index.html

178. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the tidyverse. JOSS. 2019 Nov 21;4(43):1686.

179. Scientific Image and Illustration Software | BioRender [Internet]. [cited 2023 May 7]. Available from: https://www.biorender.com/

180. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999 Apr 1;8(2):135–60.

181. Daniels CC, Rogers PD, Shelton CM. A review of pneumococcal vaccines: current polysaccharide vaccine recommendations and future protein antigens. J Pediatr Pharmacol Ther. 2016;21(1):27–35.

182. Fournier PE, Dubourg G, Raoult D. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Medicine. 2014 Nov 29;6(11):114.

183. Garriss G, Nannapaneni P, Simões AS, Browall S, Subramanian K, Sá-Leão R, Goossens H, de Lencastre H, Henriques-Normark B. Genomic characterization of the emerging pathogen *Streptococcus pseudopneumoniae*. mBio. 2019 Jun 25;10(3):10.1128/mbio.01286-19.

184. Stadhouders R, Pas SD, Anber J, Voermans J, Mes THM, Schutten M. The effect of primer-template mismatches on the detection and

quantification of nucleic acids using the 5′ nuclease assay. J Mol Diagn. 2010 Jan;12(1):109–17.

185. Wick RR, Judd LM, Holt KE. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. PLOS Computational Biology. 2023 Mar 2;19(3):e1010905.

186. De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M, Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Walker AS, Stoesser N. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. Microb Genom. 2019 Sep 4;5(9):e000294.

187. Miellet WR, van Veldhuizen J, Litt D, Mariman R, Wijmenga-Monsuur AJ, Badoux P, Nieuwenhuijsen T, Thombre R, Mayet S, Eletu S, Sheppard C, van Houten MA, Rots NY, Miller E, Fry NK, Sanders EAM, Trzciński K. It takes two to tango: combining conventional culture with molecular diagnostics enhances accuracy of *Streptococcus pneumoniae* detection and pneumococcal serogroup/serotype determination in carriage. Frontiers in Microbiology 2022;13.

188. Fabri LV, Azzopardi KI, Osowicki J, Frost HR, Smeesters PR, Steer AC. An *emm*-type specific qPCR to track bacterial load during experimental human *Streptococcus pyogenes* pharyngitis. BMC Infect Dis. 2021 May 21;21:463.

189. Habets MN, Cremers AJH, Bos MP, Savelkoul P, Eleveld MJ, Meis JF, Hermans PWM, Melchers WJ, de Jonge MI, Diavatopoulos DAY 2016. A novel quantitative PCR assay for the detection of *Streptococcus pneumoniae* using the competence regulator gene target *comX*. Journal of Medical Microbiology. 65(2):129–36.

190. Defeating meningitis by 2030: a global road map [Internet]. [cited 2022 Dec 4]. Available from: https://www.who.int/publications-detail-redirect/9789240026407

191. Hassan MM, Grist LF, Poirier AC, La Ragione RMY 2022. JMM profile: loop-mediated isothermal amplification (LAMP): for the rapid detection of nucleic acid targets in resource-limited settings. Journal of Medical Microbiology. 2022;71(5):001522.

192. Mori Y, Notomi T. Loop-mediated isothermal amplification (LAMP): a rapid, accurate, and cost-effective diagnostic method for infectious diseases. J Infect Chemother. 2009 Apr;15(2):62–9.

193. Notomi T, Okayama H, Masubuchi H, Yonekawa T, Watanabe K, Amino N, Hase T. Loop-mediated isothermal amplification of DNA. Nucleic Acids Res. 2000 Jun 15;28(12):E63.

194. Nagamine K, Hase T, Notomi T. Accelerated reaction by loop-mediated isothermal amplification using loop primers. Molecular and Cellular Probes. 2002 Jun 1;16(3):223–9.

195. Wong YP, Othman S, Lau YL, Radu S, Chee HY. Loop-mediated isothermal amplification (LAMP): a versatile technique for detection of micro-organisms. Journal of Applied Microbiology. 2018;124(3):626–43.

196. Mori Y, Nagamine K, Tomita N, Notomi T. Detection of loop-mediated isothermal amplification reaction by turbidity derived from magnesium pyrophosphate formation. Biochem Biophys Res Commun. 2001 Nov 23;289(1):150–4.

197. Zhou D, Guo J, Xu L, Gao S, Lin Q, Wu Q, Wu L, Que Y. Establishment and application of a loop-mediated isothermal amplification (LAMP) system for detection of cry1Ac transgenic sugarcane. Sci Rep. 2014 May 9;4:4912.

198. Seki M, Yamashita Y, Torigoe H, Tsuda H, Sato S, Maeno M. Loop-mediated isothermal amplification method targeting the *lytA* gene for detection of *Streptococcus pneumoniae*. J Clin Microbiol. 2005 Apr;43(4):1581–6.

199. Kim DW, Kilgore PE, Kim EJ, Kim SA, Anh DD, Dong BQ, Kim JS, Seki M. The enhanced pneumococcal LAMP assay: a clinical tool for the diagnosis of meningitis due to *Streptococcus pneumoniae*. Plos One. 2012;7(8):e42954.

200. Seki M, Kilgore PE, Kim EJ, Ohnishi M, Hayakawa S, Kim DW. Loop-mediated isothermal amplification methods for diagnosis of bacterial meningitis. Frontiers in Pediatrics 2018;6.

201. Du S, Yan C, Du B, Zhao H, Xue G, Zheng P, Feng Y, Cui J, Gan L, Feng J, Fan Z, Xu Z, Fu T, Fu H, Zhang Q, Li N, Zhang R, Liu S, Li X, Cui X, Zhou Y, Zhang Q, Chen Y, Yuan J. Development of loop-mediated isothermal amplification assay targeting *lytA* and *psaA* genes for rapid and visual diagnosis of *Streptococcus pneumoniae* pneumonia in children. Front Microbiol. 2021;12:816997.

202. Xia Y, Guo XG, Zhou S. Rapid detection of *Streptococcus pneumoniae* by real-time fluorescence loop-mediated isothermal amplification. J Thorac Dis. 2014 Sep;6(9):1193–9.

203. Cima-Cabal MD, Vázquez-Espinosa E, Vazquez F, García-Suárez M del M. Detection of *Streptococcus pneumoniae* in urine by loop-mediated isothermal amplification. J Pediatr Infect Dis. 2021 Jan;16(1):18–25.

204. Wang Y, Wang Y, Li D, Xu J, Ye C. Detection of nucleic acids and elimination of carryover contamination by using loop-mediated isothermal amplification and Antarctic thermal sensitive uracil-DNA-glycosylase in a

lateral flow biosensor: application to the detection of *Streptococcus pneumoniae*. Microchim Acta. 2018 Mar 7;185(4):212.

205. Higgins O, Clancy E, Cormican M, Boo TW, Cunney R, Smith TJ. Evaluation of an internally controlled multiplex *Tth* endonuclease cleavage loop-mediated isothermal amplification (TEC-LAMP) assay for the detection of bacterial meningitis pathogens. Int J Mol Sci. 2018 Feb 9;19(2):524.

206. Wang H, Ma Z, Qin J, Shen Z, Liu Q, Chen X, Wang H, An Z, Liu W, Li M. A versatile loop-mediated isothermal amplification microchip platform for *Streptococcus pneumoniae* and *Mycoplasma pneumoniae* testing at the point of care. Biosens Bioelectron. 2019 Feb 1;126:373–80.

207. LAMP primer designing software [PrimerExplorer] [Internet]. [cited 2022 Sep 5]. Available from: https://primerexplorer.jp/e/v5_manual/index.html

208. WarmStart® LAMP Kit (DNA & RNA) | NEB [Internet]. [cited 2022 Sep 5]. Available from: https://international.neb.com/products/e1700-warmstart-lamp-kit-dna-rna#Product%20Information

209. Mabey D, Peeling RW, Ustianowski A, Perkins MD. Diagnostics for the developing world. Nat Rev Microbiol. 2004 Mar;2(3):231–40.

210. Kaneko H, Kawana T, Fukushima E, Suzutani T. Tolerance of loop-mediated isothermal amplification to a culture medium and biological substances. Journal of Biochemical and Biophysical Methods. 2007 Apr 10;70(3):499–501.

211. Parte AC, Sardà Carbasse J, Meier-Kolthoff JP, Reimer LC, Göker M. List of Prokaryotic names with Standing in Nomenclature (LPSN) moves to the DSMZ. International Journal of Systematic and Evolutionary Microbiology. 2020;70(11):5607–12.

212. DeSimone DC, Tleyjeh IM, Correa de Sa DD, Anavekar NS, Lahr BD, Sohail MR, Steckelberg JM, Wilson WR, Baddour LM, Mayo Cardiovascular Infections Study Group. Incidence of infective endocarditis due to viridans group streptococci before and after the 2007 American heart association's prevention guidelines: an extended evaluation of the Olmsted county, Minnesota, population and nationwide inpatient sample. Mayo Clin Proc. 2015 Jul;90(7):874–81.

213. Abranches J, Zeng L, Kajfasz J, Palmer S, Chakraborty B, Wen Z, Richards V, Brady L, Lemos J. Biology of oral streptococci. Microbiol Spectr. 2018 Oct;6(5):10.1128/microbiolspec.GPP3-0042–2018.

214. To KN, Cornwell E, Daniel R, Goonesekera S, Jauneikaite E, Chalker V, Le Doare K. Evaluation of matrix-assisted laser desorption ionisation time-of-flight mass spectrometry (MALDI-TOF MS) for the identification of Group B *Streptococcus*. BMC Res Notes. 2019 Feb 14;12(1):85.

215. Moore WEC, Stackebrandt E, Kandler O, Colwell RR, Krichevsky MI, Truper HG, Murray RGE, Wayne LG, Grimont PAD, Brenner DJ, Starr MP, Moore LH. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic and Evolutionary Microbiology. 1987 Oct 1;37(4):463–4.

216. Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, Da Costa MS, Rooney AP, Yi H, Xu XW, De Meyer S, Trujillo ME. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. International Journal of Systematic and Evolutionary Microbiology. 2018 Jan 1;68(1):461–6.

217. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology. 2007;57(1):81–91.

218. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci USA. 2005 Feb 15;102(7):2567–72.

219. Glazunova OO, Raoult D, Roux V. Partial sequence comparison of the *rpoB*, *sodA*, *groEL* and *gyrB* genes within the genus *Streptococcus*. International Journal Of Systematic And Evolutionary Microbiology. 2009 Sep 1;59(9):2317–22.

220. Kosecka-Strojek M, Wolska M, Żabicka D, Sadowy E, Międzobrodzki J. Identification of clinically relevant *Streptococcus* and *Enterococcus* species based on biochemical methods and 16S rRNA, *sodA*, *tuf*, *rpoB*, and *recA* gene sequencing. Pathogens. 2020 Nov;9(11):939.

221. Wyllie AL, Pannekoek Y, Bovenkerk S, van Engelsdorp Gastelaars J, Ferwerda B, van de Beek D, Sanders EAM, Trzciński K, van der Ende A. Sequencing of the variable region of *rpsB* to discriminate between *Streptococcus pneumoniae* and other streptococcal species. Open Biol. 2017 Sep;7(9):170074.

222. Kawamura Y, Hou Xg, Sultana F, Miura H, Ezaki T. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. International Journal of Systematic and Evolutionary Microbiology. 1995;45(2):406–8.

223. Innings A, Krabbe M, Ullberg M, Herrmann B. Identification of 43 *Streptococcus* species by pyrosequencing analysis of the *rnpB* gene. J Clin Microbiol. 2005 Dec;43(12):5983–91.

224. Zeigler DR. Gene sequences useful for predicting relatedness of whole genomes in bacteria. International Journal of Systematic and Evolutionary Microbiology. 2003;53(6):1893–900.

225. Bentley RW, Leigh JA, Collins MD. Intrageneric structure of *Streptococcus* based on comparative analysis of small-subunit rRNA sequences. International Journal of Systematic and Evolutionary Microbiology. 1991;41(4):487–94.

226. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, Frost SDW, Corander J, Bentley SD, Parkhill J. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biology. 2020 Jul 22;21(1):180.

227. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. Microb Genom. 2018 Mar 15;4(3):e000166.

228. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003 Sep;13(9):2178–89.

229. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res. 2018 Jan 9;46(1):e5.

230. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: a fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. GigaScience. 2019 Oct 1;8(10):giz119.

231. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, Perrin A, Médigue C, Calteau A, Cruveiller S, Matias C, Ambroise C, Rocha EPC, Vallenet D. PPanGGOLiN: depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020 Mar 19;16(3):e1007732.

232. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. Genome Res. 2020 Nov;30(11):1667–79.

233. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.

234. Gao XY, Zhi XY, Li HW, Klenk HP, Li WJ. Comparative genomics of the bacterial genus *Streptococcus* illuminates evolutionary implications of species groups. Plos One. 2014 Jun 30;9(6):e101229.

235. Richards VP, Palmer SR, Pavinski Bitar PD, Qin X, Weinstock GM, Highlander SK, Town CD, Burne RA, Stanhope MJ. Phylogenomics and

the dynamic genome evolution of the genus *Streptococcus*. Genome Biol Evol. 2014 Apr;6(4):741–53.

236. Patel S, Gupta RS. Robust demarcation of fourteen different species groups within the genus *Streptococcus* based on genome-based phylogenies and molecular signatures. Infection, Genetics and Evolution. 2018 Dec 1;66:130–51.

237. Rasmussen LH, Dargis R, Højholt K, Christensen JJ, Skovgaard O, Justesen US, Rosenvinge FS, Moser C, Lukjancenko O, Rasmussen S, Nielsen XC. Whole genome sequencing as a tool for phylogenetic analysis of clinical strains of mitis group streptococci. Eur J Clin Microbiol Infect Dis. 2016 Oct;35(10):1615–25.

238. Barajas HR, Romero MF, Martínez-Sánchez S, Alcaraz LD. Global genomic similarity and core genome sequence diversity of the *Streptococcus* genus as a toolkit to identify closely related bacterial species in complex environments. PeerJ. 2019 Jan 14;6:e6233.

239. Jensen CS, Iversen KH, Dargis R, Shewmaker P, Rasmussen S, Christensen JJ, Nielsen XC. *Streptococcus pseudopneumoniae*: use of whole-genome sequences to validate species identification methods. Journal of Clinical Microbiology. 2021 Jan 21;59(2):10.1128/jcm.02503-20.

240. Chalita M, Ha S min, Kim YO, Oh HS, Yoon SH, Chun J. Improved metagenomic taxonomic profiling using a curated core gene-based bacterial database reveals unrecognized species in the genus *Streptococcus*. Pathogens. 2020 Mar;9(3):204.

241. Velsko IM, Perez MS, Richards VP. Resolving phylogenetic relationships for *Streptococcus mitis* and *Streptococcus oralis* through core- and pan-genome analyses. Genome Biology and Evolution. 2019 Apr 1;11(4):1077–87.

242. Lefébure T, Stanhope MJ. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biology. 2007 May 2;8(5):R71.

243. Achtman M, Zhou Z, Charlesworth J, Baxter L. EnteroBase: hierarchical clustering of 100 000s of bacterial genomes into species/subspecies and populations. Philos Trans R Soc Lond B Biol Sci. 2022;377(1861):20210240.

244. Kumar S, Bansal K, Sethi SK. Comparative genomics reconciliations of genus *Streptococcus* resolves its taxonomy and elucidates biotechnological importance of their constituent species. Ecological Genetics and Genomics. 2023 Dec 1;29:100205.

245. Haenni M, Lupo A, Madec JY. Antimicrobial resistance in *Streptococcus* spp. Microbiology Spectrum. 2018 Mar 29;6(2):10.1128/microbiolspec.arba-0008–2017.

246. Watkins ER, Kalizang'Oma A, Gori A, Gupta S, Heyderman RS. Factors affecting antimicrobial resistance in *Streptococcus pneumoniae* following vaccination introduction. Trends in Microbiology. 2022 Dec 1;30(12):1135–45.

247. Andrejko K, Ratnasiri B, Hausdorff WP, Laxminarayan R, Lewnard JA. Antimicrobial resistance in paediatric *Streptococcus pneumoniae* isolates amid global implementation of pneumococcal conjugate vaccines: a systematic review and meta-regression analysis. Lancet Microbe. 2021 Sep;2(9):e450–60.

248. Rofael S, Leboreiro Babe C, Davrandi M, Kondratiuk AL, Cleaver L, Ahmed N, Atkinson C, McHugh T, Lowe DM. Antibiotic resistance, bacterial transmission and improved prediction of bacterial infection in patients with antibody deficiency. JAC Antimicrob Resist. 2023 Dec;5(6):dlad135.

249. D'Aeth JC, van der Linden MP, McGee L, de Lencastre H, Turner P, Song JH, Lo SW, Gladstone RA, Sá-Leão R, Ko KS, Hanage WP, Breiman RF, Beall B, Bentley SD, Croucher NJ, The GPS Consortium. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. Rainey PB, Soldati-Favre D, Falush D, Kilian M, editors. eLife. 2021 Jul 14;10:e67113.

250. Jensen A, Valdórsson O, Frimodt-Møller N, Hollingshead S, Kilian M. Commensal streptococci serve as a reservoir for β-lactam resistance genes in *Streptococcus pneumoniae*. Antimicrob Agents Chemother. 2015;59(6):3529–40.

251. Hakenbeck R, Kaminski K, König A, van der Linden M, Paik J, Reichmann P, Zähner D. Penicillin-binding proteins in beta-lactam-resistant *Streptococcus pneumoniae*. Microb Drug Resist. 1999;5(2):91–9.

252. van der Linden M, Otten J, Bergmann C, Latorre C, Liñares J, Hakenbeck R. Insight into the diversity of penicillin-binding protein 2x alleles and mutations in viridans streptococci. Antimicrobial Agents and Chemotherapy. 2017 Apr 24;61(5):10.1128/aac.02646-16.

253. Chaffanel F, Charron-Bourgoin F, Libante V, Leblond-Bourget N, Payot S. Resistance genes and genetic elements associated with antibiotic resistance in clinical and commensal isolates of *Streptococcus salivarius*. Applied and Environmental Microbiology. 2015 Jun 15;81(12):4155–63.

254. Dias CA, Teixeira LM, Carvalho M da G, Beall B. Sequential multiplex PCR for determining capsular serotypes of pneumococci recovered from Brazilian children. Journal of Medical Microbiology. 2007;56(9):1185–8.

255. Carvalho M da G, Pimenta FC, Gertz RE, Joshi HH, Trujillo AA, Keys LE, Findley J, Moura IS, Park IH, Hollingshead SK, Pilishvili T, Whitney CG, Nahm MH, Beall BW, Active Bacterial Core Surveillance Team. PCR-based quantitation and clonal diversity of the current prevalent invasive serogroup 6 pneumococcal serotype, 6C, in the United States in 1999 and 2006 to 2007. J Clin Microbiol. 2009 Mar;47(3):554–9.

256. Pai R, Gertz RE, Beall B. Sequential multiplex PCR approach for determining capsular serotypes of *Streptococcus pneumoniae* isolates. J Clin Microbiol. 2006 Jan;44(1):124–31.

257. Carvalho M da G, Pimenta FC, Jackson D, Roundtree A, Ahmad Y, Millar EV, O'Brien KL, Whitney CG, Cohen AL, Beall BW. Revisiting pneumococcal carriage by use of broth enrichment and PCR techniques for enhanced detection of carriage and serotypes. J Clin Microbiol. 2010 May;48(5):1611–8.

258. Pimenta FC, Robert E. Gertz J, Roundtree A, Yu J, Nahm MH, McDonald RR, Carvalho M da G, Beall BW. Rarely occurring 19A-like *cps* locus from a serotype 19F pneumococcal isolate indicates continued need of serology-based quality control for PCR-based serotype determinations. Journal of Clinical Microbiology. 2009 Jul;47(7):2353.

259. Menezes AP de O, Reis JN, Ternes YM, Andrade AL, Pimenta FC, Carvalho M da G, Beall B. Update of pneumococcal PCR serotyping assay for detection of a commonly occurring type 19F *wzy* variant in Brazil. J Clin Microbiol. 2013 Jul;51(7):2470–1.

260. Carvalho M da G, Bigogo GM, Junghae M, Pimenta FC, Moura I, Roundtree A, Li Z, Conklin L, Feikin DR, Breiman RF, Whitney CG, Beall B. Potential nonpneumococcal confounding of PCR-based determination of serotype in carriage. J Clin Microbiol. 2012 Sep;50(9):3146–7.

261. Carvalho M da G, Pimenta FC, Moura I, Roundtree A, Gertz RE, Li Z, Jagero G, Bigogo G, Junghae M, Conklin L, Feikin DR, Breiman RF, Whitney CG, Beall BW. Non-pneumococcal mitis-group streptococci confound detection of pneumococcal capsular serotype-specific loci in upper respiratory tract. PeerJ. 2013;1:e97.

262. Lessa FC, Milucky J, Rouphael NG, Bennett NM, Talbot HK, Harrison LH, Farley MM, Walston J, Pimenta F, Gertz RE, Rajam G, Carvalho M da G, Beall B, Whitney CG. *Streptococcus mitis* expressing pneumococcal serotype 1 capsule. Sci Rep. 2018 Dec 19;8(1):17959.

263. Pimenta F, Gertz RE, Park SH, Kim E, Moura I, Milucky J, Rouphael N, Farley MM, Harrison LH, Bennett NM, Bigogo G, Feikin DR, Breiman R, Lessa FC, Whitney CG, Rajam G, Schiffer J, da Gloria Carvalho M, Beall B. *Streptococcus infantis*, *Streptococcus mitis*, and *Streptococcus oralis* strains with highly similar *cps*5 loci and antigenic relatedness to serotype 5 pneumococci. Front Microbiol. 2018;9:3199.

264. Beall B. Potential epidemiologic and historical implications of capsular serotypes shared by pneumococci and their nonpneumococcal relatives. J Infect Dis. 2020 Jul 6;222(3):343–6.

265. Gertz RE, Pimenta FC, Chochua S, Larson S, Venero AK, Bigogo G, Milucky J, Carvalho M da G, Beall B. Nonpneumococcal strains recently recovered from carriage specimens and expressing capsular serotypes highly related or identical to pneumococcal serotypes 2, 4, 9A, 13, and 23A. mBio. 2021 May 18;12(3):e01037-21.

266. Nucleotide BLAST: Search nucleotide databases using a nucleotide query [Internet]. [cited 2022 Sep 6]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

267. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Tamura K, editor. Molecular Biology and Evolution. 2021 Dec 9;38(12):5825–9.

268. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research. 2019 Jan 8;47(D1):D309–14.

269. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research. 2000 Jan 1;28(1):33–6.

270. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021 Jan 8;49(D1):D274–81.

271. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015 Jan;43(Database issue):D261-269.

272. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct 24;278(5338):631–7.

273. Seemann T. tseemann/prokka [Internet]. 2024 [cited 2024 Mar 21]. Available from: https://github.com/tseemann/prokka

274. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068–9.

275. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar 8;11:119.

276. Tonkin-Hill G. gtonkinhill/panaroo [Internet]. 2024 [cited 2024 Apr 9]. Available from: https://github.com/gtonkinhill/panaroo

277. Katoh K, Misawa K, Kuma K ichi, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059–66.

278. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximum-likelihood trees for large alignments. Plos One. 2010 Mar 10;5(3):e9490.

279. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution. 2009 Jul 1;26(7):1641–50.

280. Velusamy S, Tran T, Mongkolrattanothai T, Walker H, McGee L, Beall B. Expanded sequential quadriplex real-time polymerase chain reaction (PCR) for identifying pneumococcal serotypes, penicillin susceptibility, and resistance markers. Diagn Microbiol Infect Dis. 2020 Jun;97(2):115037.

281. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Research. 2024 Apr 13;gkae268.

282. Wickham H, RStudio. forcats: tools for working with categorical variables (factors) [Internet]. 2023 [cited 2024 Mar 11]. Available from: https://cran.r-project.org/web/packages/forcats/index.html

283. Xu S, Chen M, Yu G. Use ggbreak to effectively utilize plotting space to deal with large datasets and outliers. Front Genet. 2021 Nov 2;12:774846.

284. Schauberger P, Walker A, Braglia L, Sturm J, Garbuszus JM, Barbone JM. openxlsx: read, write and edit xlsx files [Internet]. 2023 [cited 2024 Mar 11]. Available from: https://cran.r-project.org/web/packages/openxlsx/index.html

285. Wickham H, Vaughan D, Girlich M, Ushey K, Software P, PBC. tidyr: tidy messy data [Internet]. 2024 [cited 2024 Mar 11]. Available from: https://cran.r-project.org/web/packages/tidyr/index.html

286. Visually distinct colors generator [Internet]. [cited 2024 May 18]. Available from: https://mokole.com/palette.html

287. Developers IW. Inkscape [Internet]. [cited 2023 May 18]. Available from: https://inkscape.org/

288. Pilarczyk-Zurek M, Sitkiewicz I, Koziel J. The clinical view on *Streptococcus anginosus* group – opportunistic pathogens coming out of hiding. Front Microbiol. 2022 Jul 8;13:956677.

289. Lemos JA, Palmer SR, Zeng L, Wen ZT, Kajfasz JK, Freires IA, Abranches J, Brady LJ. The biology of *Streptococcus mutans*. Microbiol Spectr. 2019 Jan;7(1).

290. Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. Proceedings of the National Academy of Sciences. 1962 Apr;48(4):582–92.

291. Mann S, Chen YPP. Bacterial genomic G+C composition-eliciting environmental adaptation. Genomics. 2010 Jan 1;95(1):7–15.

292. Scholz CFP, Poulsen K, Kilian M. Novel molecular method for identification of *Streptococcus pneumoniae* applicable to clinical microbiology and 16S rRNA sequence-based microbiome studies. J Clin Microbiol. 2012 Jun;50(6):1968–73.

293. Kilian M, Tettelin H. Identification of virulence-associated properties by comparative genome analysis of *Streptococcus pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, three *S. oralis* subspecies, and *S. infantis*. mBio. 2019 Sep 3;10(5):e01985-19.

294. Costa SS, Guimarães LC, Silva A, Soares SC, Baraúna RA. First steps in the analysis of prokaryotic pan-genomes. Bioinform Biol Insights. 2020 Aug 7;14:1177932220938064.

295. Zhong C, Qu B, Hu G, Ning K. Pan-genome analysis of *Campylobacter*: insights on the genomic diversity and virulence profile. Microbiology Spectrum. 2022 Sep 7;10(5):e01029-22.

296. Liu D, Xiao C, Li X, Sun Y, Qi H, Zou Y. *Streptococcus shenyangsis* sp. nov., a new species isolated from the oropharynx of a healthy child from Shenyang China. Curr Microbiol. 2021 Jul 1;78(7):2821–7.

297. Metcalf BJ, Gertz RE, Gladstone RA, Walker H, Sherwood LK, Jackson D, Li Z, Law C, Hawkins PA, Chochua S, Sheth M, Rayamajhi N, Bentley SD, Kim L, Whitney CG, McGee L, Beall B. Strain features and distributions in pneumococci from children with invasive disease before and after 13-valent conjugate vaccine implementation in the USA. Clin Microbiol Infect. 2016 Jan;22(1):60.e9-60.e29.

298. Kaya DE, Ülgen E, Kocagöz AS, Sezerman OU. A comparison of various feature extraction and machine learning methods for antimicrobial resistance prediction in *Streptococcus pneumoniae*. Front Antibiot, 2023 Mar 24;2.

299. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, Hawkins PA, Tran T, McGee L, Beall BW, on behalf of the Active Bacterial Core surveillance team. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (*pbp*) sequences. BMC Genomics. 2017 Aug 15;18(1):621.

300. Miellet WR, Almeida ST, Trzciński K, Sá-Leão R. *Streptococcus pneumoniae* carriage studies in adults: importance, challenges, and key issues to consider when using quantitative PCR-based approaches. Front Microbiol. 2023 Feb 16;14:1122276.

301. Tonkin-Hill G, Corander J, Parkhill J. Challenges in prokaryote pangenomics. Microb Genom. 2023 May 25;9(5):mgen001021.

302. Overmann J, Abt B, Sikorski J. Present and future of culturing bacteria. Annual Review of Microbiology. 2017 Sep 8;71(Volume 71, 2017):711–30.

303. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016 Jan 4;44(D1):D733–45.

304. Albright S, Louca S. Trait biases in microbial reference genomes. Sci Data. 2023 Feb 9;10(1):84.

305. Bentley RW, Leigh JA. Determination of 16S ribosomal RNA gene copy number in *Streptococcus uberis, S. agalactiae, S. dysgalactiae* and *S. parauberis*. FEMS Immunology and Medical Microbiology. 1995 Sep 1;12(1):1–7.

306. Winand R, Bogaerts B, Hoffman S, Lefevre L, Delvoye M, Van Braekel J, Fu Q, Roosens NH, De Keersmaecker SC, Vanneste K. Targeting the 16S rRNA gene for bacterial identification in complex mixed samples: comparative evaluation of second (Illumina) and third (Oxford Nanopore technologies) generation sequencing technologies. Int J Mol Sci. 2019 Dec 31;21(1):298.

307. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, Gruca A, Grynberg M, Kajava AV, Promponas VJ, Anisimova M, Jakobsen KS, Linke D. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019 Dec 2;47(21):10994–1006.

308. Cascella R, Strafella C, Ragazzo M, Zampatti S, Borgiani P, Gambardella S, Pirazzoli A, Novelli G, Giardina E. Direct PCR: a new pharmacogenetic approach for the inexpensive testing of HLA-B*57:01. Pharmacogenomics J. 2015 Apr;15(2):196–200.

309. Tran DH, Tran HT, Le UP, Vu XD, Trinh TBN, Do HDK, Than VT, Bui LM, Vu VV, Nguyen TL, Phung HTT, Le VP. Direct colorimetric LAMP assay for rapid detection of African swine fever virus: a validation study during an outbreak in Vietnam. Transboundary and Emerging Diseases. 2021;68(4):2595–602.

310. Sidstedt M, Rådström P, Hedman J. PCR inhibition in qPCR, dPCR and MPS—mechanisms and solutions. Anal Bioanal Chem. 2020 Apr 1;412(9):2009–23.

311. Wyllie AL, Mbodj S, Thammavongsa DA, Hislop MS, Yolda-Carr D, Waghela P, Nakahata M, Stahlfeld AE, Vega NJ, York A, Allicock OM, Wilkins G, Ouyang A, Siqueiros L, Strong Y, Anastasio K, Alexander-Parrish R, Arguedas A, Gessner BD, Weinberger DM. Persistence of pneumococcal carriage among older adults in the community despite COVID-19 mitigation measures. Microbiol Spectr. 11(3):e04879-22.

312. Nwe MK, Jangpromma N, Taemaitree L. Evaluation of molecular inhibitors of loop-mediated isothermal amplification (LAMP). Sci Rep. 2024 Mar 11;14(1):5916.

313. Dou M, Sanjay ST, Dominguez DC, Liu P, Xu F, Li X. Multiplexed instrument-free meningitis diagnosis on a polymer/paper hybrid microfluidic biochip. Biosens Bioelectron. 2017 Jan 15;87:865–73.

314. Wyllie AL, Rots NY, Wijmenga-Monsuur AJ, van Houten MA, Sanders EAM, Trzciński K. Saliva as an alternative sample type for detection of pneumococcal carriage in young children. Microbiology (Reading). 2023 Oct 11;169(10):001394.

315. Laxton CS, Peno C, Hahn AM, Allicock OM, Perniciaro S, Wyllie AL. The potential of saliva as an accessible and sensitive sample type for the detection of respiratory pathogens and host immunity. The Lancet Microbe. 2023 Oct 1;4(10):e837–50.

316. Pinilla-Monsalve GD, Llanos-Leyton N, González MC, Manrique-Hernández EF, Rey-Serrano JJ, Quiñones-Bautista JA. Socioepidemiological macro-determinants associated with the cumulative incidence of bacterial meningitis: a focus on the African Meningitis Belt. Front Neurol. 2023 Feb 14;14:1088182.

317. Mfuh KO, Abanda NN, Titanji BK. Strengthening diagnostic capacity in Africa as a key pillar of public health and pandemic preparedness. PLOS Glob Public Health. 2023 Jun 13;3(6):e0001998.

318. Retchless AC, Fox LM, Maiden MCJ, Smith V, Harrison LH, Glennie L, Harrison OB, Wang X. Toward a global genomic epidemiology of meningococcal disease. The Journal of Infectious Diseases. 2019 Oct 31;220(Supplement_4):S266–73.

319. Obaro S, Hassan-Hanga F, Medugu N, Olaosebikan R, Olanipekun G, Jibir B, Gambo S, Ajose T, Duru C, Ebruke B, Davies HD. Comparison of bacterial culture with BioFire® FilmArray® multiplex PCR screening of archived cerebrospinal fluid specimens from children with suspected bacterial meningitis in Nigeria. BMC Infect Dis. 2023 Oct 2;23(1):641.

320. Cailleaux M, Pilmis B, Mizrahi A, Lourtet-Hascoet J, Nguyen Van JC, Alix L, Couzigou C, Vidal B, Tattevin P, Le Monnier A. Impact of a multiplex PCR assay (FilmArray®) on the management of patients with suspected central nervous system infections. Eur J Clin Microbiol Infect Dis. 2020 Feb;39(2):293–7.

321. Liesman RM, Strasburg AP, Heitman AK, Theel ES, Patel R, Binnicker MJ. Evaluation of a commercial multiplex molecular panel for diagnosis of infectious meningitis and encephalitis. Journal of Clinical Microbiology. 2018 Mar 26;56(4):10.1128/jcm.01927-17.

322. Anjum MF, Zankari E, Hasman H. Molecular methods for detection of antimicrobial resistance. Microbiology Spectrum. 2017 Dec 7;5(6):10.1128/microbiolspec.arba-0011–2017.

323. Donà V, Kasraian S, Lupo A, Guilarte YN, Hauser C, Furrer H, Unemo M, Low N, Endimiani A. Multiplex real-time PCR assay with high-resolution melting analysis for characterization of antimicrobial resistance in *Neisseria gonorrhoeae*. Journal of Clinical Microbiology. 2016 Jul 25;54(8):2074–81.

324. Chen L, Shin DJ, Zheng S, Melendez JH, Gaydos CA, Wang TH. Direct-qPCR Assay for coupled identification and antimicrobial susceptibility testing of *Neisseria gonorrhoeae*. ACS Infect Dis. 2018 Sep 14;4(9):1377–84.

325. Johnson AP, Ashiru-Oredope D, Beech E. Antibiotic stewardship initiatives as part of the UK 5-year antimicrobial resistance strategy. Antibiotics. 2015 Dec;4(4):467–79.

326. Angeletti S, Dicuonzo G, Avola A, Crea F, Dedej E, Vailati F, Farina C, De Florio L. Viridans group streptococci clinical isolates: MALDI-TOF mass spectrometry versus gene sequence-based identification. Plos One. 2015 Mar 17;10(3):e0120502.

327. Govender KN, Street TL, Sanderson ND, Eyre DW. Metagenomic sequencing as a pathogen-agnostic clinical diagnostic tool for infectious diseases: a systematic review and meta-analysis of diagnostic test accuracy studies. J Clin Microbiol. 2021 Aug 18;59(9):e0291620.

328. Baker JL, Morton JT, Dinis M, Alvarez R, Tran NC, Knight R, Edlund A. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. Genome Res. 2021 Jan;31(1):64–74.

329. Achtman M, Zhou Z. Metagenomics of the modern and historical human oral microbiome with phylogenetic studies on *Streptococcus mutans* and *Streptococcus sobrinus*. Philosophical Transactions of the Royal Society B: Biological Sciences. 2020 Oct 5;375(1812):20190573.

330. Silva PH, Vázquez Y, Campusano C, Retamal-Díaz A, Lay MK, Muñoz CA, González PA, Kalergis AM, Bueno SM. Non-capsular based immunization approaches to prevent *Streptococcus pneumoniae* infection. Front Cell Infect Microbiol. 2022 Sep 26;12:949469.

331. Sloas MM, Barrett FF, Chesney PJ, English BK, Hill BC, Tenover FC, Leggiadro RJ. Cephalosporin treatment failure in penicillin- and cephalosporin-resistant *Streptococcus pneumoniae* meningitis. Pediatr Infect Dis J. 1992 Aug;11(8):662–6.

332. Coffey TJ, Berrón S, Daniels M, Garcia-Leoni ME, Cercenado E, Bouza E, Fenoll A, Spratt BG. Multiply antibiotic-resistant *Streptococcus pneumoniae* recovered from Spanish hospitals (1988-1994): novel major clones of serotypes 14, 19F and 15F. Microbiology (Reading). 1996 Oct;142 (Pt 10):2747–57.

333. Hall LM, Whiley RA, Duke B, George RC, Efstratiou A. Genetic relatedness within and between serotypes of *Streptococcus pneumoniae* from the United Kingdom: analysis of multilocus enzyme electrophoresis, pulsed-field gel electrophoresis, and antimicrobial resistance patterns. J Clin Microbiol. 1996 Apr;34(4):853–9.

334. Jabes D, Nachman S, Tomasz A. Penicillin-binding protein families: evidence for the clonal nature of penicillin resistance in clinical isolates of pneumococci. J Infect Dis. 1989 Jan;159(1):16–25.

335. Smith AM, Klugman KP. Three predominant clones identified within penicillin-resistant South African isolates of *Streptococcus pneumoniae*. Microb Drug Resist. 1997;3(4):385–9.

336. Gherardi G, Whitney CG, Facklam RR, Beall B. Major related sets of antibiotic-resistant pneumococci in the United States as determined by pulsed-field gel electrophoresis and *pbp1a-pbp2b-pbp2x-dhf* restriction profiles. J Infect Dis. 2000 Jan;181(1):216–29.

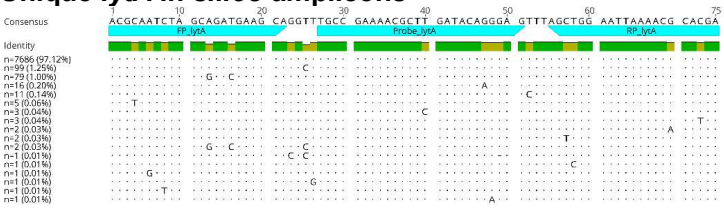
337. Syrogiannopoulos GA, Grivea IN, Beratis NG, Spiliopoulou AE, Fasola EL, Bajaksouzian S, Appelbaum PC, Jacobs MR. Resistance patterns of *Streptococcus pneumoniae* from carriers attending day-care centers in Southwestern Greece. Clin Infect Dis. 1997 Aug;25(2):188–94.

Appendices

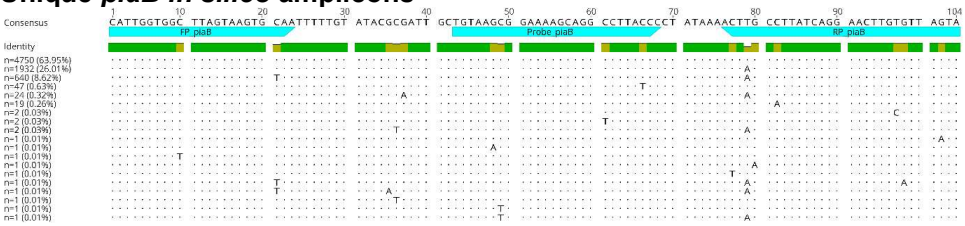
7.1 Appendices to Chapter 2

7.1.1 Nucleotide sequence alignments of *in silico* PCR amplicons

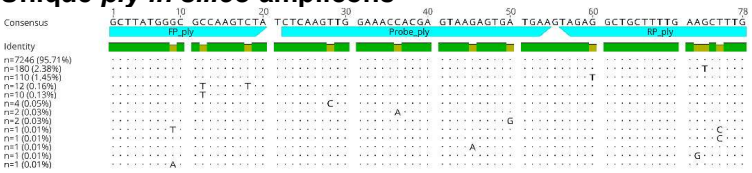
Unique *lytA in silico* amplicons



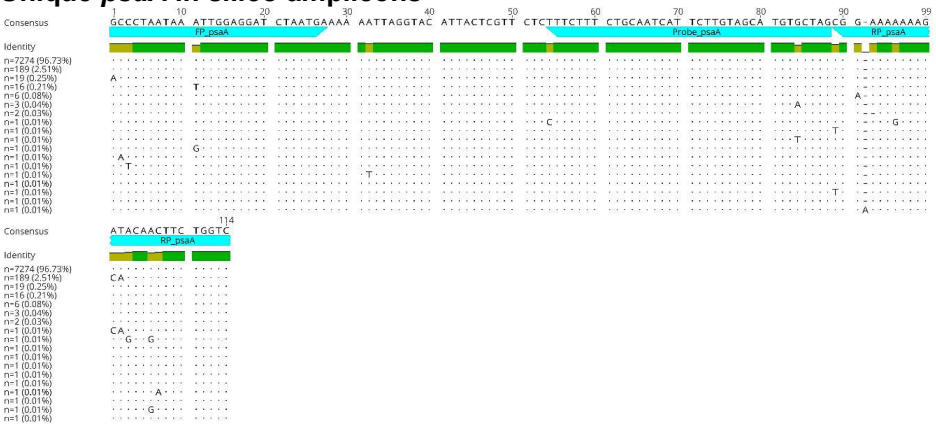
Unique *piaB in silico* amplicons



Unique *ply in silico* amplicons

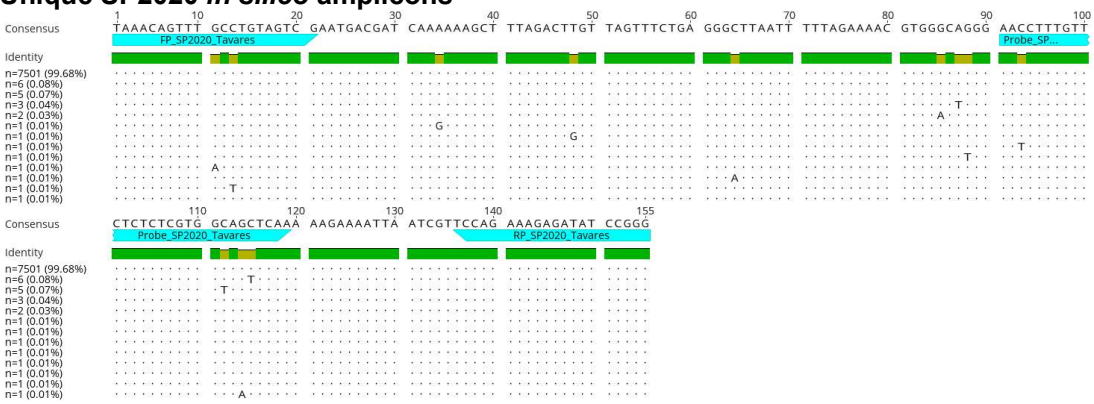


Unique *psaA in silico* amplicons

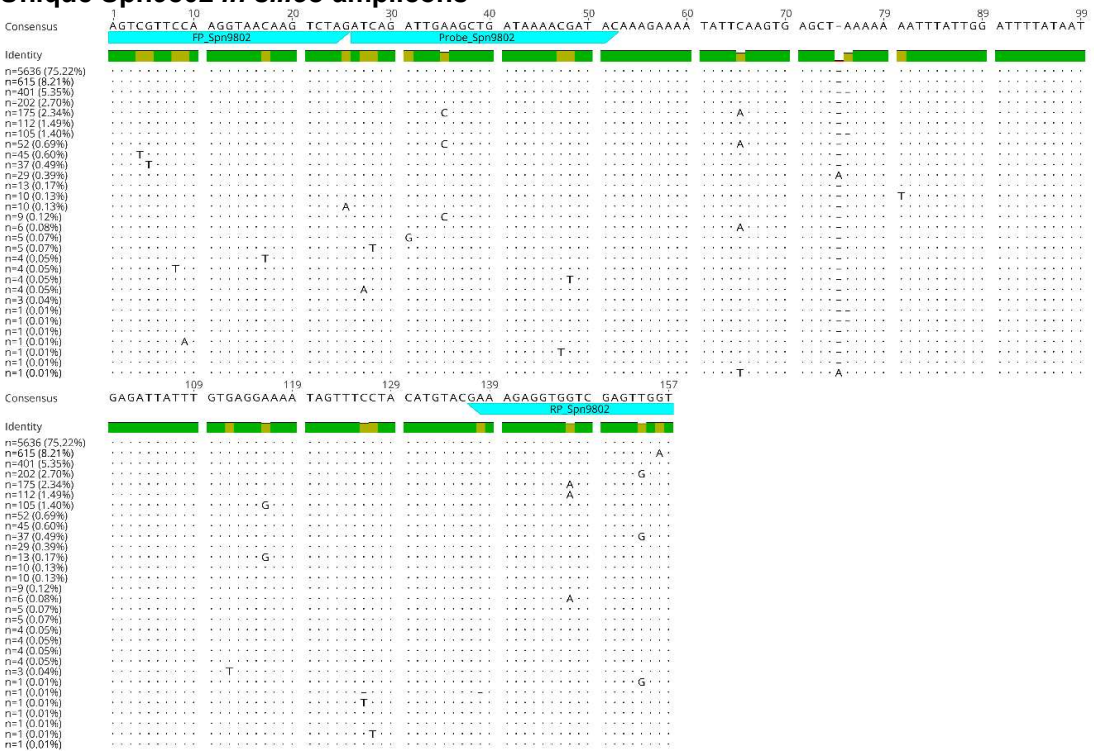


Appendix Figure 7.1 Alignments of nucleotide sequences of qPCR *in silico* amplicons for the four previously published assays *lytA, piaB, ply* and *psaA*. The consensus sequence shows the nucleotides representing at least 50% of nucleotides in that position across the amplicons. Dots indicate that the nucleotide is identical to the consensus; dashes indicate a sequence gap. Primer and probe binding regions are annotated (FP, forward primer; RP, reverse primer). The 'identity' row shows the nucleotide sequence pairwise identity: 100%; brown, 30-100%. In the first column, n (%) indicates the frequency of that amplicon sequence out of the total number of predicted amplicons in the *in silico* pneumococcal study dataset for that assay.

Unique SP2020 *in silico* amplicons

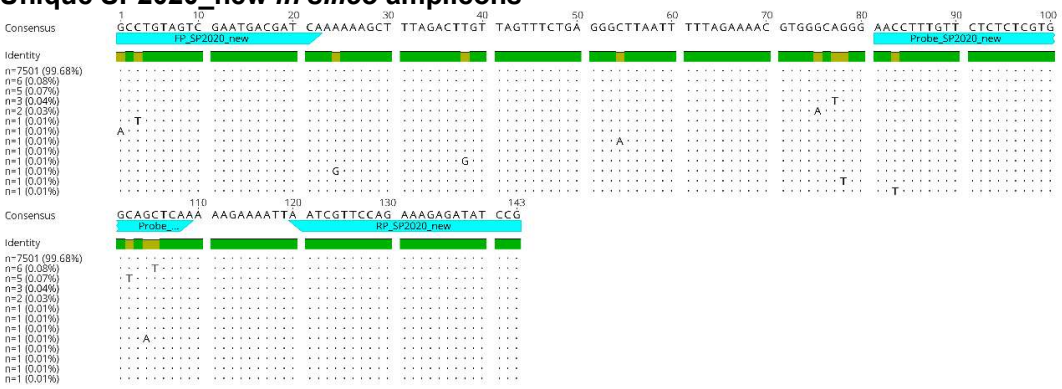


Unique Spn9802 *in silico* amplicons

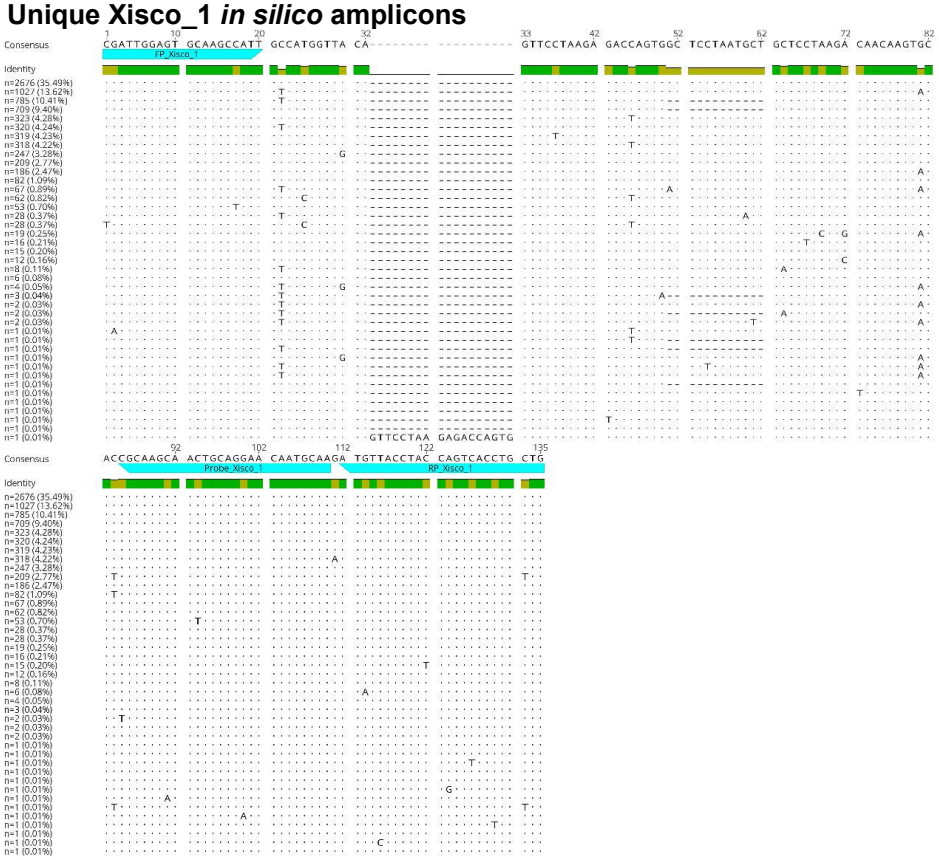


Appendix Figure 7.2 Alignments of nucleotide sequences of qPCR *in silico* amplicons for the two previously published assays SP2020 and Spn9802. The consensus sequence shows the nucleotides representing at least 50% of nucleotides in that position across the amplicons. Dots indicate that the nucleotide is identical to the consensus; dashes indicate a sequence gap. Primer and probe binding regions are annotated (FP, forward primer; RP, reverse primer). The 'identity' row shows the nucleotide sequence pairwise identity: 100%; brown, 30-100%. In the first column, n (%) indicates the frequency of that amplicon sequence out of the total number of predicted amplicons in the *in silico* pneumococcal study dataset for that assay.

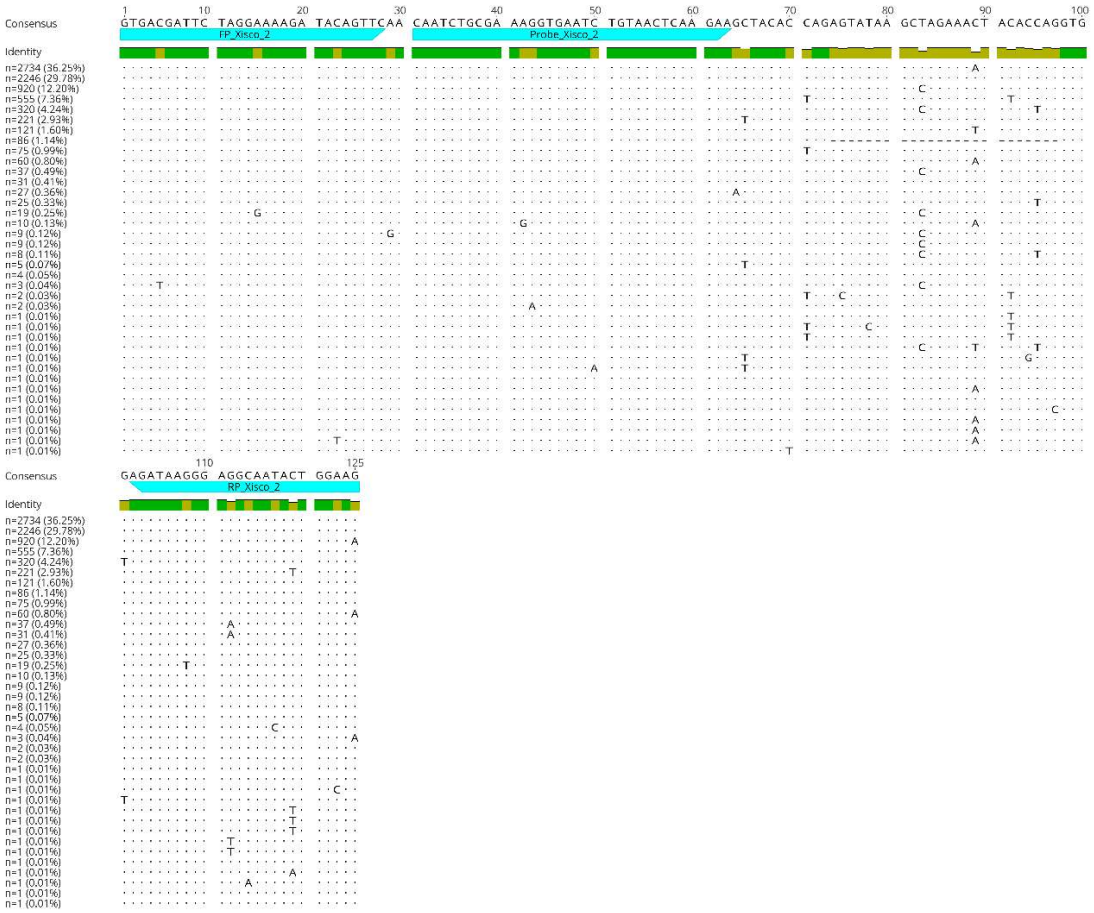
Unique SP2020_new *in silico* amplicons



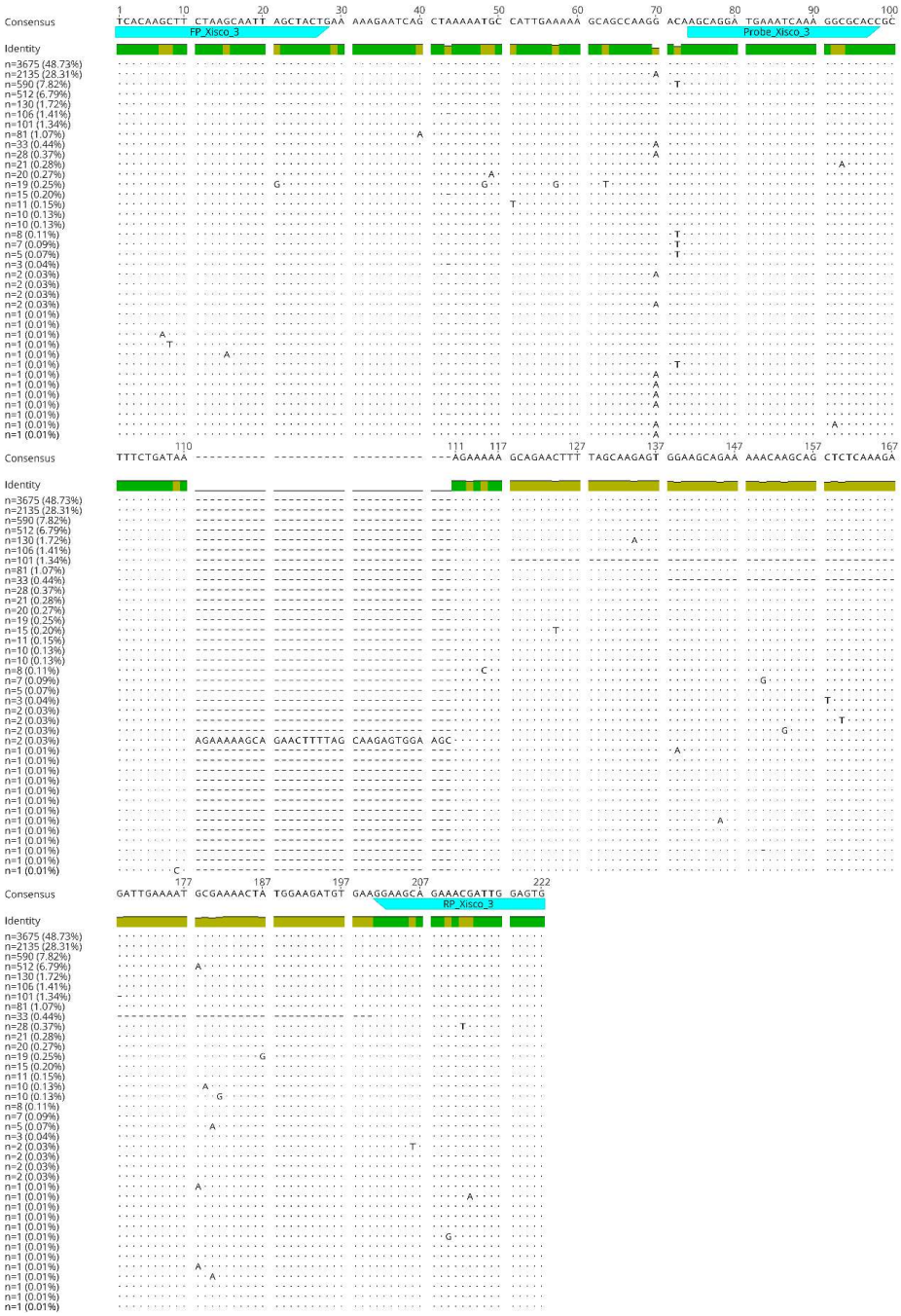
Unique Xisco_1 *in silico* amplicons



Appendix Figure 7.3 Alignments of nucleotide sequences of qPCR *in silico* amplicons for two newly designed assays SP2020_new and Xisco_1.
The consensus sequence shows the nucleotides representing at least 50% of nucleotides in that position across the amplicons. Dots indicate that the nucleotide is identical to the consensus; dashes indicate a sequence gap. Primer and probe binding regions are annotated (FP, forward primer; RP, reverse primer). The 'identity' row shows the nucleotide sequence pairwise identity: 100%; brown, 30-100%. In the first column, n (%) indicates the frequency of that amplicon sequence out of the total number of predicted amplicons in the *in silico* pneumococcal study dataset for that assay.



Appendix Figure 7.4 Alignments of nucleotide sequences of qPCR *in silico* amplicons for the newly designed assay Xisco_2.
The consensus sequence shows the nucleotides representing at least 50% of nucleotides in that position across the amplicons. Dots indicate that the nucleotide is identical to the consensus; dashes indicate a sequence gap. Primer and probe binding regions are annotated (FP, forward primer; RP, reverse primer). The 'identity' row shows the nucleotide sequence pairwise identity: 100%; brown, 30-100%. In the first column, n (%) indicates the frequency of that amplicon sequence out of the total number of predicted amplicons in the *in silico* pneumococcal study dataset for that assay.



Appendix Figure 7.5 Alignments of nucleotide sequences of qPCR *in silico* amplicons for the newly designed assay Xisco_3.
The consensus sequence shows the nucleotides representing at least 50% of nucleotides in that position across the amplicons. Dots indicate that the nucleotide is identical to the consensus; dashes indicate a sequence gap. Primer and probe binding regions are annotated (FP, forward primer; RP, reverse primer). The 'identity' row shows the nucleotide sequence pairwise identity: 100%; brown, 30-100%. In the first column, n (%) indicates the frequency of that amplicon sequence out of the total number of predicted amplicons in the *in silico* pneumococcal study dataset for that assay.

7.2 Appendices to Chapter 3

7.2.1 Biophysical metrics of pneumococcal LAMP primers

Appendix Table 7.1 Biophysical metrics of LAMP primers calculated automatically by Primer Explorer (v5) [207].

Assay name	Primer name	Length (bp)	T_m (°C)
SP2020_1	SP2020_1_F3	25	57.65
	SP2020_1_B3	22	56.46
	SP2020_1_F2	19	56.02
	SP2020_1_F1c	24	61.01
	SP2020_1_B2	19	56.63
	SP2020_1_B1c	23	60.81
SP2020_2	SP2020_2_F3	22	55.08
	SP2020_2_B3	25	56.15
	SP2020_2_F2	19	57.67
	SP2020_2_F1c	24	60.10
	SP2020_2_B2	21	55.05
	SP2020_2_B1c	25	60.85
SP2020_3	SP2020_3_F3	19	56.01
	SP2020_3_B3	18	55.33
	SP2020_3_F2	25	57.53
	SP2020_3_F1c	25	60.73
	SP2020_3_B2	25	57.65
	SP2020_3_B1c	20	61.11
	SP2020_3_LB	21	60.12
	SP2020_3_LF	20	60.04
SP2020_4	SP2020_4_F3	21	55.28
	SP2020_4_B3	18	55.33
	SP2020_4_F2	19	57.96
	SP2020_4_F1c	23	60.34
	SP2020_4_B2	25	57.65
	SP2020_4_B1c	20	61.11
SP2020_5	SP2020_5_F3	24	55.73
	SP2020_5_B3	20	55.35
	SP2020_5_F2	18	55.25
	SP2020_5_F1c	25	60.43
	SP2020_5_B2	20	57.23
	SP2020_5_B1c	20	61.49

Note: bp, base pairs; T_m, melting temperature; °C, degrees Celsius.

7.2.2 Pneumococcal genomes used for phylogenetic analysis of *in silico* positive LAMP assays in NPS genomes

Appendix Table 7.2 Published *Streptococcus pneumoniae* genomes.

Strain name	ATCC number	Serotype	Year	Reference
Tennessee23F-4	ATCC 51916	23F	1991	[331]
Spain14-5	ATCC 700902	14	1990	[332]
England14-9	ATCC 700676	14	1994	[333]
CSR14-10	ATCC 700677	14	1987	[334]
S.Africa19A-13	ATCC 700904	19A	1988	[335]
Maryland6B-17	ATCC BAA-342	6B	1997	[336]
Tennessee14-18	ATCC BAA-340	14	1997	[336]
Greece6B-22	ATCC BAA-658	6Bii (6E)	1996	[337]
Denmark12F-34	ATCC BAA-1660	12F	1995	[114]
Netherlands14-35	ATCC BAA-1661	14	1980	[114]
Netherlands18C-36	ATCC BAA-1662	18C	1980	[114]
Netherlands7F-39	ATCC BAA-1665	7F	1984	[114]

Note: ATCC, American Type Culture Collection.

7.3 International conference abstracts

7.3.1 Abstract poster presentation ISPPD 2022

International Symposium on Pneumococci and Pneumococcal Diseases (ISPPD)

Toronto, Canada 19-23 June 2022

First author poster presentation in section “Diagnosis and management”

In silico evaluation of >9300 genomes to assess the sensitivity and specificity of PCR assays currently recommended for the molecular detection of pneumococci

F. M. Ahlers[1], M. J. Jansen van Rensburg[1], K. K. Malhi[1], M. E. B. Butler[2], J. E. Bray[3], K. A. Jolley[3], O. B. Harrison[3], M. C. J.* Maiden[3], A. B. Brueggemann[1]

[1]Nuffield Department of Population Health, Big Data Institute, University of Oxford, Oxford, United Kingdom

[2]Department of Infectious Disease, Imperial College, London, London, United Kingdom

[3]Department of Zoology, University of Oxford, Oxford, United Kingdom

Background: PCR-based diagnostics provide rapid methods to identify pneumococci, but potentially misidentify closely-related streptococci and bacteriophage *lytA* homologue. This study aimed to investigate the *in silico* sensitivity and specificity of published PCR/qPCR assays recommended for pneumococcal diagnostics.

Methods: Target loci *lytA* (6 assays) and *piaB, ply, psaA,* SP2020, Spn9802 and Xisco (1 assay each) were evaluated using two curated datasets: 7,547 genomes from the PubMLST Pneumococcal Genome Library

(https://pubmlst.org/organisms/streptococcus-pneumoniae/pgl); and 1,825 genomes from 55 non-pneumococcal *Streptococcus* species. Pneumococcal gene sequences were identified and stored within the PubMLST database, unique amplicon sequences were aligned, and nucleotide mismatches in the primer/probe regions were counted. Putative PCR amplifications among non-pneumococcal streptococci were assessed using the *in silico* PCR tool (https://bigsdb.readthedocs.io/en/latest/data_analysis/in_silico_pcr.html). Sensitivity and specificity of each assay was calculated.

Results: For sensitivity, sequences were assessed only if the full-length PCR amplicon could be extracted, thus the denominator varied per assay (data not shown). Assay sensitivity, allowing for 0-2 nucleotide mismatches, was: Xisco (96.50-99.93%); SP2020 (99.43-99.63%); *piaB* (63.00-98.47%); *lytA* (88.00-99.82%); *psaA* (96.43-99.62%); *ply* (95.68-99.99%); and Spn9802 (82.11-99.47%). Assay specificity, allowing for 0-2 nucleotide mismatches, was: Xisco (100%); SP2020 (99.95%); *piaB* (99.84-100%); *lytA* (99.67*-99.95%); *psaA* (99.29-99.51%); *ply* (97.26-97.53%); and Spn9802 (96.00-99.89%). All *lytA* assays detected both pneumococcal and bacteriophage genes. False positives were viridans streptococci.

Conclusions: *In silico* analyses demonstrated high sensitivity and specificity for all assays, but false positives and amplification of bacteriophage genes remain a concern. This is a cost-effective method to investigate the suitability of PCR targets before and after *in vitro* assay development.

*An incorrectly reported initial of one of the authors' names and a number in the abstract have been corrected in this appendix.

7.3.2 Abstract oral presentation ECCMID 2023

33rd European Congress of Clinical Microbiology & Infectious Diseases (ECCMID)

Copenhagen, Denmark 15-18 April 2023

First author oral presentation in section “Diagnostic microbiology”

Utilising large and diverse bacterial genome datasets to improve the identification of *Streptococcus pneumoniae* via PCR-based diagnostics

F. M. Ahlers[1], D. Litt[2], M. J. Jansen van Rensburg[1], J. E. Bray[3], K. A. Jolley[3], S. Eletu[2], J. Coelho[5], B. Pichon[4], O. B. Harrison[3], M. C. J. Maiden[3], D. W. Eyre[1], N. K. Fry[2,6], A. B. Brueggemann[1]

[1]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[2]Respiratory and Vaccine Preventable Bacteria Reference Unit, UK Health Security Agency, London, United Kingdom

[3]Department of Biology, University of Oxford, Oxford, United Kingdom

[4]Healthcare Associated Infections, Fungal, AMR, AMU & Sepsis Division, UK Health Security Agency, London, United Kingdom

[5]Antimicrobial Resistance and Healthcare Associated Infection Unit, UK Health Security Agency, London, United Kingdom

[6]Immunisation and Vaccine Preventable Diseases, UK Health Security Agency, London, United Kingdom

Background: Accurate identification of *Streptococcus pneumoniae* (pneumococcus) is crucial for clinical microbiology, surveillance and epidemiology,

but can be complicated by the misidentification of non-pneumococcal *Streptococcus* (NPS) species. The aims of this study were to use genome datasets to assess the *in silico* accuracy of current qPCR/PCR targets, and then develop and validate *in vitro* assays for the most promising target: the 'Xisco' gene (TIGR4 genome locus SP_1992).

Methods: Three Xisco qPCR primer/probe sets were designed and compared to six published qPCR primer/probe sets for *lytA, piaB, ply, psaA,* Spn9802 and SP2020*,* and one published PCR primer set for Xisco, using the *in silico* PCR tool in PubMLST (https://bigsdb.readthedocs.io/en/latest/data_analysis/in_silico_pcr.html). Overall, 7,547 pneumococcal genomes, and 1,825 genomes that represented 55 NPS species, were analysed *in silico*. Putative qPCR/PCR amplicon sequences were aligned and up to four nucleotide mismatches per primer or probe were allowed to be considered a positive result. The three Xisco qPCR primer/probe sets and one published primer/probe set each for *lytA* and *piaB* were tested *in vitro* using 246 pneumococci from the 2022 UK surveillance collection and 277 isolates of 21 different NPS species.

Results: Among all 7,547 pneumococcal genomes, the Xisco primer/probe sets correctly detected virtually all true positives *in silico* (99.9% of true positives, 95% CI 99.8-100%) and Xisco was never detected among 1,825 NPS genomes. In contrast, all other non-Xisco qPCR/PCR targets detected 96.8-100% of pneumococci *in silico,* and produced *in silico* false positives for between 1 and 82 NPS genomes (*Streptococcus pseudopneumoniae* or *Streptococcus mitis*). Therefore, *in vitro* qPCR assays were developed for Xisco, with a resulting overall sensitivity of 100% (246/246, 95% CI 98.5-100%) among pneumococci and overall

specificity of 100% (277/277, 95% CI 98.7-100%) among NPS isolates, which was comparable or superior to the tested *lytA* and *piaB in vitro* assays.

Conclusions: Large-scale *in silico* analyses of diverse genome datasets are a cost-effective method to investigate the putative accuracy of qPCR primers/probes prior to the more expensive development of *in vitro* laboratory assays.

7.3.3 Abstract poster presentation ECCMID 2023

33rd European Congress of Clinical Microbiology & Infectious Diseases (ECCMID)

Copenhagen, Denmark 15-18 April 2023

First author poster presentation in section “Diagnostic microbiology”

Development of a real-time loop-mediated isothermal amplification (LAMP) assay targeting SP2020 for the identification of *Streptococcus pneumoniae*

F. M. Ahlers[1], D. Litt[2], M. J. Jansen van Rensburg[1], J. E. Bray[3], K. A. Jolley[3], J. Coelho[4], B. Pichon[5], O. B. Harrison[3], M. C. J. Maiden[3], D. W. Eyre[1], N. K. Fry[2,6], A. B. Brueggemann[1]

[1]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

[2]Respiratory and Vaccine Preventable Bacteria Reference Unit, UK Health Security Agency, London, United Kingdom

[3]Department of Biology, University of Oxford, Oxford, United Kingdom

[4]Antimicrobial Resistance and Healthcare Associated Infection Unit, UK Health Security Agency, London, United Kingdom

[5]Healthcare Associated Infections, Fungal, AMR, AMU & Sepsis Division, UK Health Security Agency, London, United Kingdom

[6]Immunisation and Vaccine Preventable Diseases, UK Health Security Agency, London, United Kingdom

Background: *Streptococcus pneumoniae* (pneumococcus) is a major global cause of morbidity and mortality. Rapid, affordable molecular diagnostics have the potential to improve the detection of pneumococci from clinical specimens. The aim of this study was to develop a real-time, loop-mediated isothermal amplification (LAMP) assay targeting the SP2020 locus, a putative GntR-family transcriptional regulator.

Methods: LAMP primers targeting SP2020 were designed using Prime Explorer software. The presence or absence of SP2020, and putative accuracy of SP2020 primers, were assessed in two well-curated datasets: 7,547 pneumococcal genomes and 1,825 genomes of 55 non-pneumococcal *Streptococcus* (NPS) species. The accuracy of the primers was investigated using an *in silico* PCR tool (https://bigsdb.readthedocs.io/en/latest/data_analysis/in_silico_pcr.html). A maximum of four nucleotide mismatches per primer binding region was allowed for a positive predicted *in silico* reaction. One set of SP2020 primers was tested *in vitro* on three sets of isolates: 246 pneumococci that caused invasive disease in the UK in 2022; 277 isolates of 21 different NPS species; and 71 pneumococci selected to represent unique SP2020 alleles. Real-time detection was facilitated with SYBR green and three independent dilution series were used to determine the limit of detection.

Results: Overall, 99.6% (7518/7547, 95% CI 99.4-99.7%) of pneumococcal genomes were positive *in silico* using the SP2020 LAMP primers, whereas only 0.1% (1/1825, 95% CI 0.0-0.3%) of NPS genomes were positive (*Streptococcus pseudopneumoniae*). The reaction time of the SP2020 LAMP *in vitro* assay was 12min. *In vitro* sensitivity was 100% (246/246, 95% CI 98.5-100%) and specificity was 100% (277/277, 95% CI 98.7-100%). All pneumococci selected to represent the 71 different SP2020 alleles tested positive *in vitro* using this LAMP assay. The SP2020 limit of detection was 65 copies per 25 µL reaction.

Conclusions: We utilised large genomic datasets to develop *in silico* primers that detected the allelic diversity of SP2020 in the pneumococcal population, and validated these primers *in vitro* in a highly sensitive, specific and rapid LAMP assay, which could be especially useful in low-resource settings.