THE LANCET
Microbe

Supplementary appendix 1

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: D'Aeth JC, Bertran M, Abdullahi F, et al. Whole-genome sequencing, strain composition, and predicted antimicrobial resistance of *Streptococcus pneumoniae* causing invasive disease in England in 2017–20: a prospective national surveillance study. *Lancet Microbe* 2025. https://doi.org/10.1016/j.lanmic.2025.101102

Supplementary materials

Whole-genome sequencing, strain composition and predicted antimicrobial resistance of *Streptococcus pneumoniae* causing invasive disease in England: prospective national surveillance, 2017 - 2020.

Joshua C. D'Aeth[1], Marta Bertran[2], Fariyo Abdullahi[2], Seyi Eletu[1], Erjola Hani[2], Norman K. Fry [1,2], Shamez N. Ladhani [2,3] and David J. Litt [1]

1 - Respiratory and Vaccine Preventable Bacteria Reference Unit, UK Health Security Agency, London, UK
2 - Immunisation and Vaccine Preventable Diseases Division, UK Health Security Agency, London, UK
3 - Paediatric Infectious Diseases Research Group, St George's University of London, London, UK

Contents

Methods

Sequencing and bioinformatic pipeline

Sequencing of *Streptococcus pneumoniae* isolates was performed as follows. Genomic DNA from pure *S. pneumoniae* cultures was extracted using the QIAGEN QIAsymphony DSP DNA Mini kit. Extracted DNA was sequenced on Illumina HiSeq 2500 platforms by the UKHSA Colindale Sequencing Laboratory (CSL). The deplexed fastq files produced by the CSL were then trimmed using the Trimmomatic [1] tool. When run routinely, any isolate's sequence data which has a yield < 100 MBp after Trimmomatic processing is sent for re-sequencing.

These trimmed reads were then analysed by the UKHSA pneumococcal WGS bioinformatic pipeline (Figure S1). The first step in the pipeline is species identification, performed through the KmerID tool. KmerID splits a set of species reference complete genomes from RefSeq into 18-mers. The percentage of 18-mers that appear at least twice in a set of query reads from a species reference is then reported as the top hit. Mixed cultures are detected by comparing the similarity of the references reported as the top hits. For *S. pneumoniae* the two reference sequences are strains ATCC 700669 (RefSeq accession: GCF_000026665.1) and 5652-06 (RefSeq accession: GCF_000252025.1). Reads which have *S. pneumoniae* as their top hit, and are not reported to be mixed, then proceed to the next step in the pipeline, multi-locus sequence typing (MLST). This is performed by the MOST [2] tool with the publicly available *S. pneumoniae* MLST scheme from PubMLST [3, 4]. MOST also provides quality-control (QC) on the reads by reporting the maximum non-consensus base percentage for MLST loci. Isolates with high (> 15%) non-consensus base scores at any of these seven loci are sent for re-sequencing. The final step in the pipeline is serotyping. This is performed using the PneumoCaT [5] tool, with manual slide agglutination serotyping performed on certain serogroups, 24, 32, and 35B, which PneumoCaT does not adequately distinguish.

All the tools in the routine pipeline used above have been validated previously, with KmerID used in Ashton *et al* 2015 [6], and Kapatai *et al* 2016 [5] where PneumoCaT has also been validated. MOST has been validated for *S. pneumoniae* in Tewolde *et al* 2016 [2] .

Of 13,944 isolates sent for sequencing, those that came from a pure unmixed culture and had a sequencing yield of > 100 Mbp were selected for further analysis in this study, which left 13,812 isolates.

Assembly and further QC of genomes

The 13,812 reads were then assembled using shovill v0.9. Two isolates could not be assembled, leaving 13,810 in total. Contigs of < 500 bp in length were then trimmed from the assemblies. QUAST v5.2.0 [7] was used to summarise assembly statistics, while CheckM v1.2.2 [8] was used to check the completeness and contamination of

the assemblies. From this, two isolates fell below the threshold of an n50 > 5,000, leaving 13,808. A further 55 isolates were found to have a contamination score of > 5%. Removing these isolates left a collection of 13,753 assemblies for further analysis.

PopPUNK v2.6.0 [9] was then run on this collection of assemblies in order to cluster isolates into strains and produce a core-genome distance based phylogeny using RapidNJ. From this a further four isolates were dropped from the collection during PopPUNK QC steps, identified as length outliers, leaving a total collection of 13,749 isolates. In order to place the isolates in a global context, the collection was also assigned to Global Pneumococcal Sequencing Clusters (GPSCs). This was performed using PopPUNK and the publicly available GPSC v6 database (`https://www.pneumogen.net/gps/training_command_line.html`). Dominant-GPSCs were defined as those containing $\geq$100 isolates in the collection, following, but not recapitulating, Gladstone et al 2019 [10]. Reads for the 13,749 isolates that have passed further assembly QC were deposited in the Sequence Read Archive (SRA) with BioProject Accession: PRJNA1034002 and PRJNA1027675. Individual isolate metadata can be found in Appendix 2 p1.

Clonal complexes (CCs) were also defined by grouping together the single-locus variants (SLVs) of MLST sequence types (STs). These were named after the largest constitutive ST within the complex. This grouping was performed with an in-house R script.

The output from the analysis pipeline outlined above for this study was benchmarked against the pipeline currently used by the Global Pneumococcal Sequencing project to analyse sequence data (`https://github.com/sanger-bentley-group/gps-pipeline`). A total of 100 isolates were randomly selected from the 13,749 isolates in this study's collection. All of the isolates passed the QC metrics for the GPS pipeline, the GPS assignment was consistent across both. One isolate, previously labeled as novel in our dataset, was instead updated with a recent ST from the GPS pipeline. Seven isolates differed in serotype, however all were within the same serogroup with the reported serotype differing due to alternative approaches in PneumoCaT and the seroBA tool used in the GPS pipeline.

Antimicrobial resistance

The presence of antimicrobial resistance (AMR) genes were detected using NCBI-AMRFinderPlus v3.1.40 [11] with the *S. pneumoniae* organism-specific database. The database version was v2022-08-09.1, which contains 6,218 unique AMR proteins and 161 separate point mutation reference sequences. The default thresholds of minimum identity were used on the assemblies that passed QC. The presence of a gene was taken to indicate resistance.

For resistance to β-lactam antibiotics, the random forest method developed and validated in Li *et al 2016* [12] and Li *et al* 2017 [13] was used. The resistance cat-

egorisation was based on the breakpoints for meningitis, which corresponds to the Clinical and Laboratories Standards Institute document M100-23.

Resistance to sulfamethoxazole and trimethoprim was determined by searching assemblies for mutations in the *folP* and *folA* genes respectively, as described in D'Aeth *et al* 2021 [14]. For sulfamethoxazole, indels within the amino acid sequence of *folP* from S61 were taken as evidence of resistance mutation, while for trimethoprim the mutation I100L in *folA* was taken as evidence of resistance. Resistance to co-trimoxazole, a combination of both sulfamethoxazole and trimethoprim, was predicted if either of the mutations in *folP* or *folA* were present.

All AMR profiles in this report are *in silico* predictions.

Phylogenetic and pangenome analysis

The 3,027 serotype 8 GPSC3 CC-53 isolates were combined with 137 GPSC3 ST53 serotype 8 isolates from the publicly available GPS study (`https://data-viewer.monocle.sanger.ac.uk/project/gps`) accessed on 30/07/2024. This collection of 3,164 isolates was mapped to the reference sequence for GPSC3, AP200 (accession code: NC_014494.1) using ska v1.0 [15] with a split kmer size of 31. This alignment was then input to Gubbins v3.3.5, using RapidNJ [16] as an initial phylogeny builder, RAxML v8.2.12 [17] as the main phylogeny builder, and pyjar [18] as the ancestral reconstruction option.

For GPSC12, an isolate's clade was determined from the phylogeny created in Bertran *et al* 2024 [19]. For the genome-wide association study (GWAS) described below, a phylogeny was formed by mapping the 1399 GPSC12 isolates in this study to the reference OXC141 isolate (Accession code: FQ312027.1), using ska v1.0 [15] with a split kmer size of 31. Gubbins v3.3.5 was then used to form a phylogeny from this alignment, using RapidNJ as the initial phylogeny builder [16], RAxML v8.2.12 [17], and pyjar as the ancestral reconstruction option.

For the GWAS study described below, the 1,399 GPSC12 isolates in this study had their genomes delineated into core and accessory genes by panaroo v1.5 [20]. Sequences were first annotated using Prokka v1.13, then panaroo v1.5 was run in strict mode, with a core-gene threshold of $\geq$99%, and set to produce a core gene alignment only.

Virulence factors

The 3,803 GPSC3 isolates and the 41 non-GPSC3 serotype 8 isolates with an assigned GPSC were searched for virulence genes. The virulence-factor database (VFDB) [21] was accessed on 02/08/2024, and the protein sequences from the 141 *S. pneumoniae* identified alleles were subset. DIAMOND v2.1.9 was then used to create a database of the 141 *S. pneumoniae* alleles, and isolates were then searched using the blastx functionality of DIAMOND [22]. Results were filtered such that the alignment and the reference sequence had the same length, alignments had a per-

cent identity score $\geq$95%, and that there were zero gaps in the alignment.

Statistical analyses

For the calculation of Simpson's diversity index values, the diversity command within the R package vegan v2.6.6.1 was used. Confidence intervals around the values were calculated using bootstrap sampling with 1,000 different replicates. Broad age-range categories of 0-4, 5-14, 15-44, 45-64 and ≥65 years were used in these calculations.

To assess the mean within-CC core-genome distances the R v4.4.1 ks.test function was used with default settings.

For the logistic regression analysis, variables were chosen to model both pathogen and host factors causing mortality. For pathogen factors, GPSC was chosen to represent the genetic diversity of an isolate, while AMR status was chosen to investigate the clinical effect of resistance. For patient factors, age was chosen to represent the diversity in host response to disease, clinical presentation to account for the severity of different pathologies, and year of isolation to account for any temporal effects. All predictor variables were coded as categorical variables. GPSC was coded with 24 levels matching the 24 dominant-GPSCs, those with greater than 100 isolates in the collection. GPSC19, which had the closest fatality rate to the aggregated rate within all dominant-GPSCs, was chosen as the reference group. Patient age was coded with patients split into discrete age groups: 0-4, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-64, 65-69, 70-74, 75-79, 80-84, and ≥85 years. The ≥85 age group was chosen as the reference group. For sampling year, the year with the highest number of isolates, 2019, was selected as the reference. For the clinical presentation, non-meningitis was the reference. Isolates were coded as AMR if they had at least one resistant gene predicted by the NCBI-AMRFinderPlus tool, or co-trimoxazole resistance, and/or were predicted to be resistant to penicillin at the meningitis breakpoints used by the random forest predictor [13]. Five isolates with incomplete penicillin breakpoints were removed from the analysis. The glm command within the R v4.4.1 stats package was used to fit the logistic regression model.

Sensitivity analyses for the regression were performed, incorporating the effect of vaccination with either a PCV vaccine (PCV7 or PCV13) or the PPV23 vaccine, extending the number of GPSCs to 26, and modelling the most frequent serotypes instead of GPSCs. Both PCV and PPV23 vaccine status was defined in one of five classes: Yes (at least one dose administered between 14 days and five years before date of case), Yes-Longer (at least one dose administered five years or more before date of case), Yes-Unsure (At least one dose administered, no data on date of vaccination), Unknown (No data on whether a case had been vaccination), and No (patient known not to have been vaccinated). PCV and PPV23 vaccine status were included as categorical variables, with the reference group for both being No. For the extended GPSC analysis, GPSC19 was once again chosen as the reference group, while for the serotype analysis, serotype 23A was chosen as the reference, having

the closest fatality rate of 17.5% to the dataset's overall rate of 17.4%.

To assess more fine-detailed associations of isolate's genetic characteristics with the case fatality rate, a GWAS analysis was conducted. This analysis was limited to the GPSC12 isolates taken from cases in patients in the $\geq$85 age group, a total of 302 isolates. The R package treeWAS v1.0 [23] was used. The analysis was conducted in two stages, with the filtered core gene alignment produced by panaroo, noted above, used as genetic data in stage one, and for stage two the accessory gene presence/absence matrix also produced by panaroo was used as the genetic data. Common input to both stages was the Gubbins GPSC12 phylogeny described above, subset to the 302 $\geq$85 years old isolates, and a csv file of the isolate's name and the outcome variable: whether the cases died within 30 days of laboratory case confirmation. treeWAS performed three tests to calculate terminal, simultaneous and subsequent association scores between input genetic data and the outcome variable, adapting significance thresholds based on corrections for multiple testing [23].

Results

GPSC3-CC53 phylogenetic analysis

The 3,027 GPSC3-CC53 isolates from the collection analysed in this study were combined with 137 isolates from the GPS database. Within the phylogeny of these 3,164 isolates, two clades were formed (appendix 1 p15). The smaller clade of 201 isolates contained 69 isolates (34·3%) from this study, along with 132 isolates (65·7%) from the GPS study, which were primarily from South Africa (104 isolates; 51·7%). The larger clade contained 2,963 isolates, of which 2,958 were from this study (99·8%), while four were from New Zealand and one from Slovenia.

Diversity of PBP alleles among resistant isolates

Of the 205 unique PBP profiles in the collection among the 1,149 isolates predicted to be resistant, 17 profiles were present in $\geq$10 isolates. Three profiles were present in $\geq$100 isolates: PBP1a-1, PBP2b-67, PBP2x-1 (n = 147); PBP1a-16, PBP2b-13, PBP2x-19 (n = 120); and PBP1a-23, PBP2b-27, PBP2x-77 (n = 108). There were three dominant-GPSCs which contained very high proportions of β-lactam resistance, $\geq$98%, GPSC5, GPSC9, and GPSC17. There were 24 different PBP profiles underlying the high-level of resistance in GPSC5, with all 147 PBP1a-7, PBP2b-67, PBP2x-1, mentioned above, within GPSC5, followed by PBP1a-0, PBP2b-1, PBP2x-1 (62 isolates; 20·3% of GPSC5 resistant isolates) and PBP1a-7, PBP2b-1, PBP2x-1 (52 isolates; 17% of resistant isolates) in terms of frequency. For GPSC9 there were 30 distinct profiles among the 173 resistant isolates, with three profiles found in $\geq$10 isolates: PBP1a-24, PBP2b-27, PBP2x-28 (n = 81; 46.8% of resistant GPSC9 isolates), PBP1a-67, PBP2b-27, PBP2x-35 (n = 26; 15% of resistant GPSC9 isolates), and PBP1a-24, PBP2b-27, PBP2x-179 (n = 19; 11% of GPSC9 resistant isolates). However, within GPSC17, all of whose 120 isolates was predicted to be resistant,

only one PBP profile was observed, PBP1a-16, PBP2b-13, PBP2x-19.

Diversity of Tn*916* among collection

Two dominant-GPSCs were observed to contain large numbers, $\geq$100, of putative Tn*916*-like elements, GPSC3 (n = 208) and GPSC9 (n = 170). Within GPSC9, where it was possible to reconstruct a Tn916-like insertion, 165 isolates contained a truncated 17kb Tn*6002* element, part of the Tn*916*-like family, with the *erm(B)* macrolide resistance gene and the *tet(M)* tetracycline resistance gene. This insertion had appeared to lose the 3.9kb region of the Tn*916* backbone immediately upstream of the *tet(M)* gene which contains the element's site-specific integrase. A further two isolates in GPSC9 contained a 23.5kb Tn*2009* insertion, which contained mef(A) and tet(M), while a further three isolates contained a 21kb element similar to the truncated Tn*6002* described above. Within GPSC3 193 isolates contained a complete 20.5kb Tn*6002* element, while a further 15 isolates contained Tn*916*-like elements. Most of these Tn*916*-like elements were found in CC-717 in GPSC3 (196 of 208; 94·2%), although there were seven isolates of the CC-53 lineage containing this.

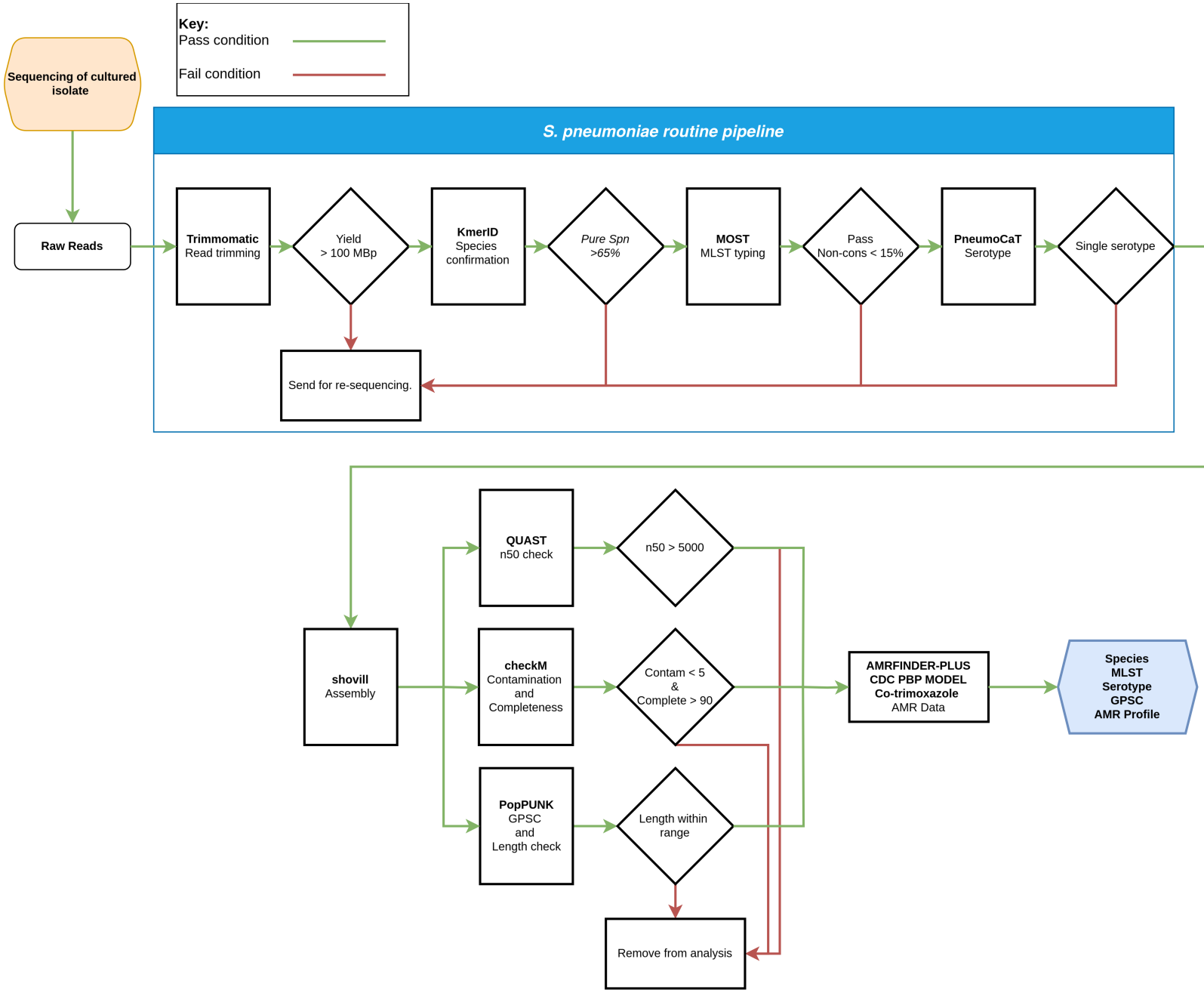

Figure S1: Analysis pipeline for pneumococcal sequence data. Flow chart detailing the analyses performed on *S. pneumoniae* isolates used in this report, highlighting the routine pipeline used for surveillance by UKHSA, and the added analysis steps used for this study.

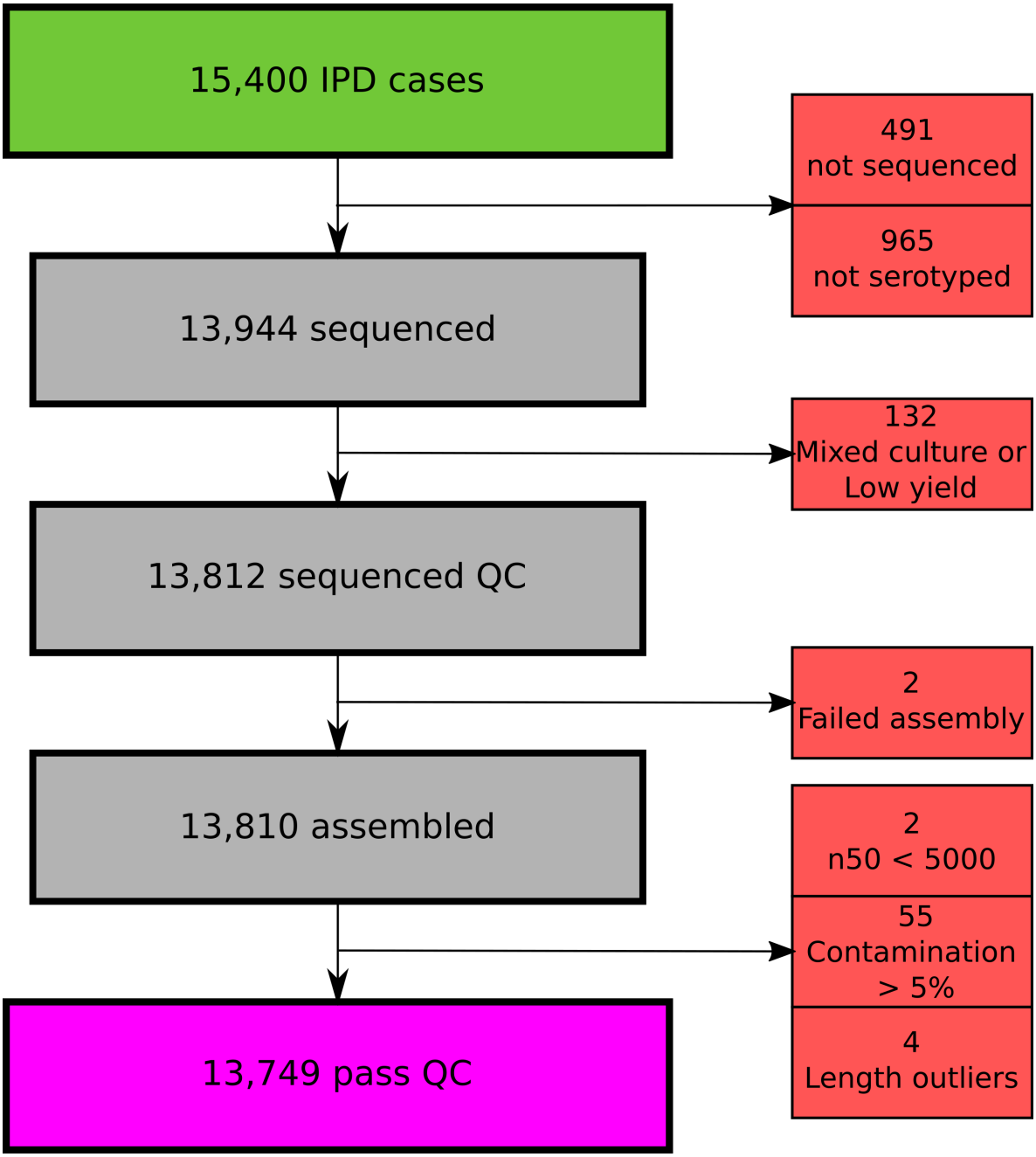

Figure S2: Isolate inclusion flowchart. Flowchart showing the flow of data from the initial 15,400 reported IPD cases between 1st of July 2017 and February 29th 2020, to the 13,749 isolates with WGS data analysed in this study. Exclusion reasons and numbers are listed in the red boxes to the right of the flowchart.

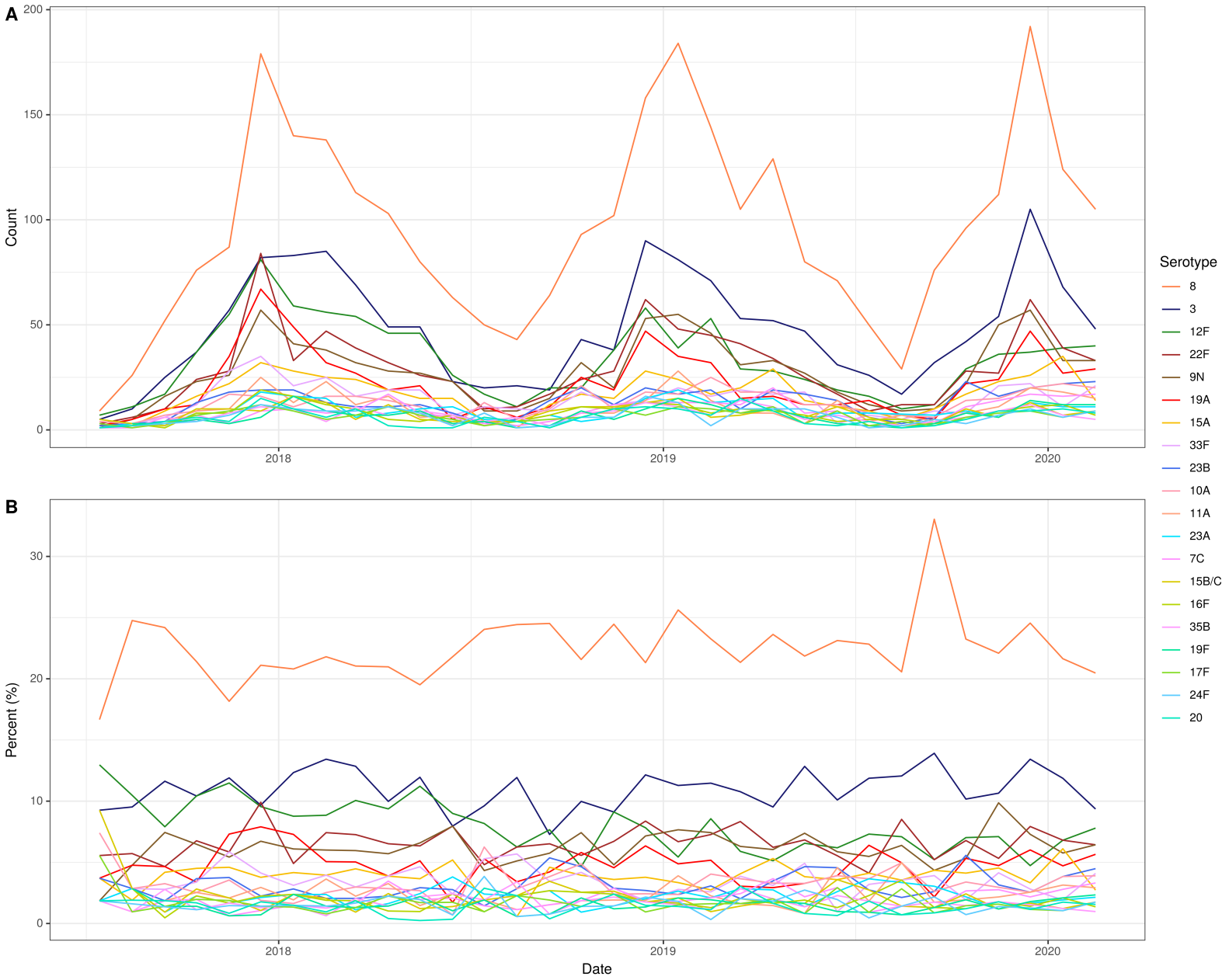

Figure S3: Serotype composition over time of the collection. (A) The overall counts of serotypes within the 13,749 isolate collection by month of the study period from July 1st 2017 to February 29th 2020. Lines are coloured by serotype in question, while the legend is arranged from the most frequent serotypes overall to the least frequent. Only the 20 most frequently observed serotypes are plotted. **(B)** The percentage of isolates expressing a certain serotype during a particular month of the study period. As for section A, only the 20 most frequently observed serotypes are expressed.

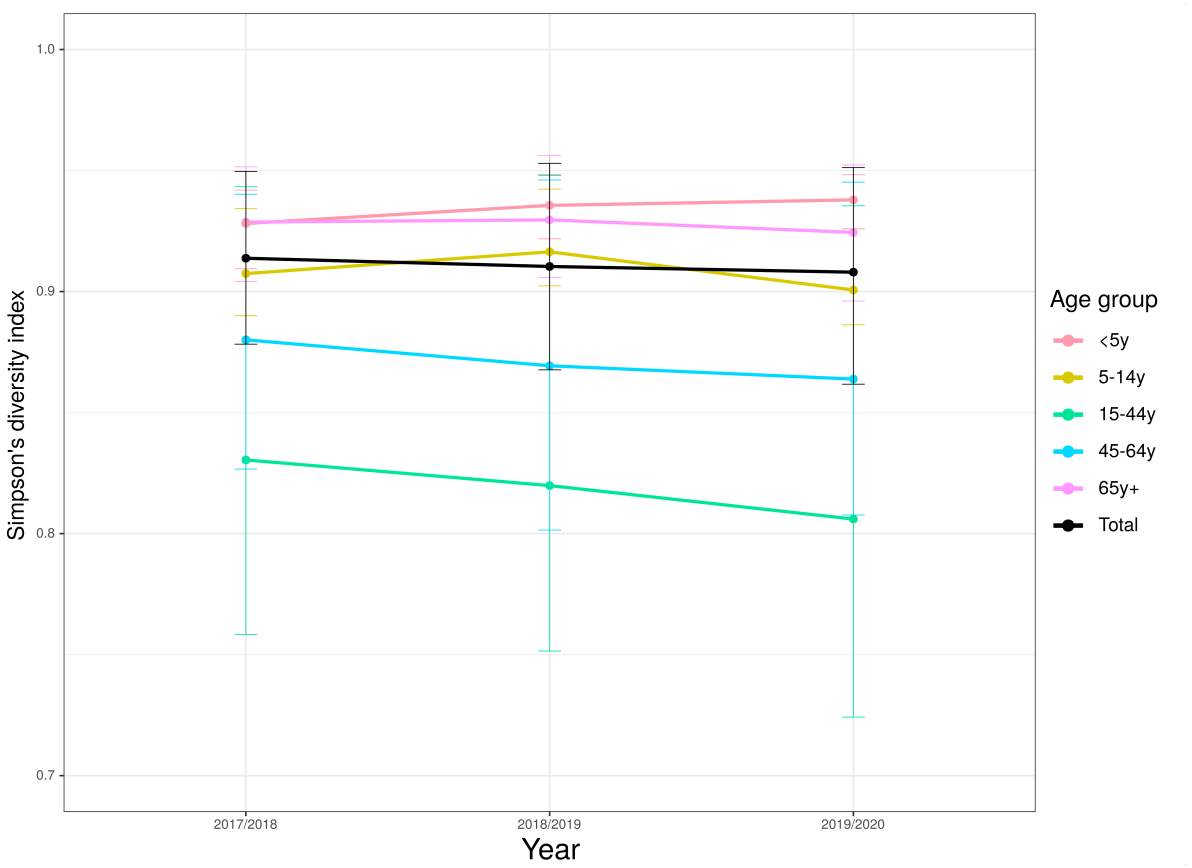

Figure S4: Serotype diversity over time. The Simpson's diversity index of isolate serotype within patient age and epidemiological years. Lines are coloured by the patient age group, total represents all age groups. Error bars represent the 95% confidence interval for values calculated from bootstrap sampling.

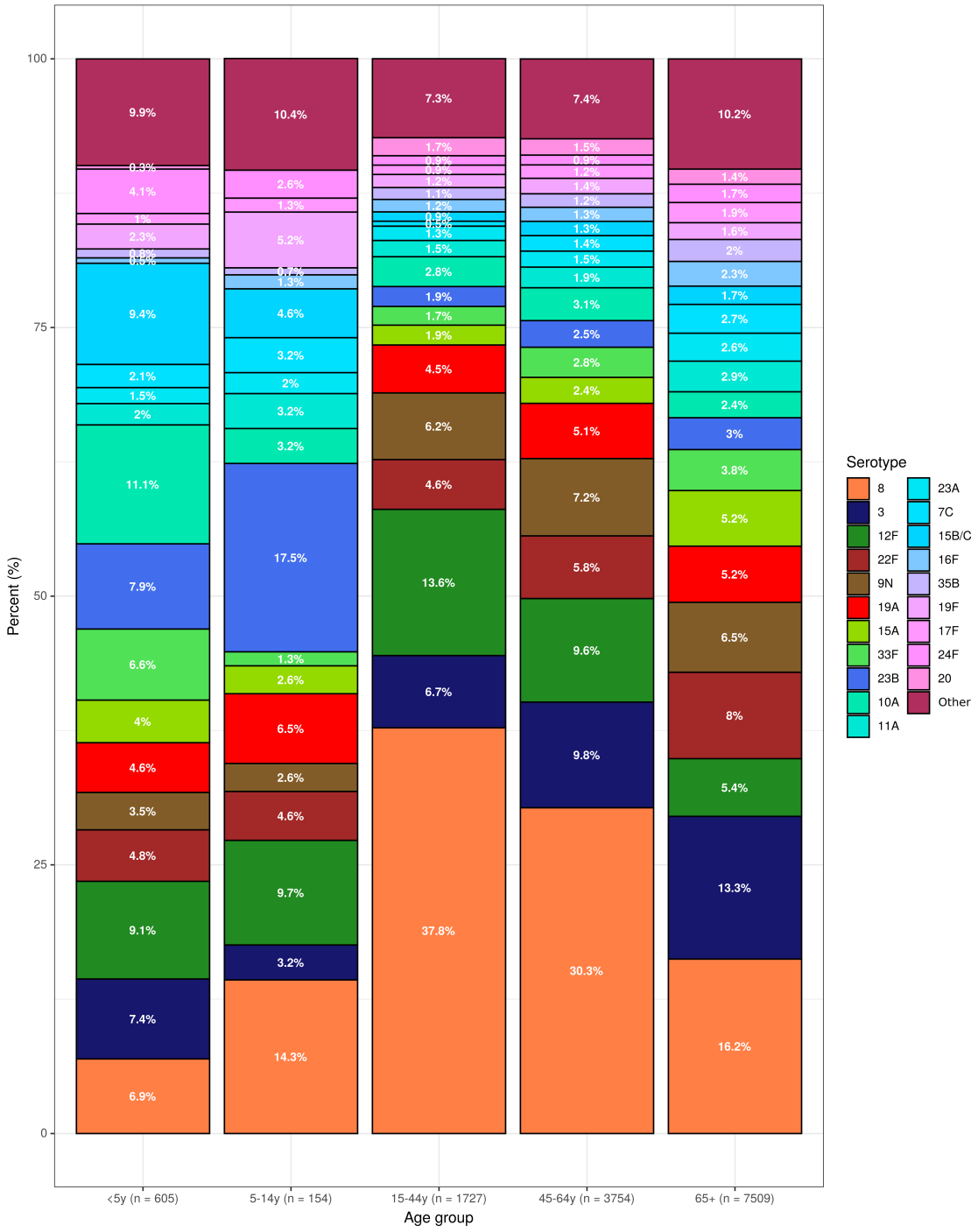

Figure S5: Serotype composition by age group. The percentage of the top 20 serotypes in each patient age group in the collection. Bars are coloured by the serotype, with only the top 20 serotypes coloured, other serotypes are grouped in the other category. The key is ordered from the most common serotype, serotype 8, to the least common of the top 20, serotype 20.

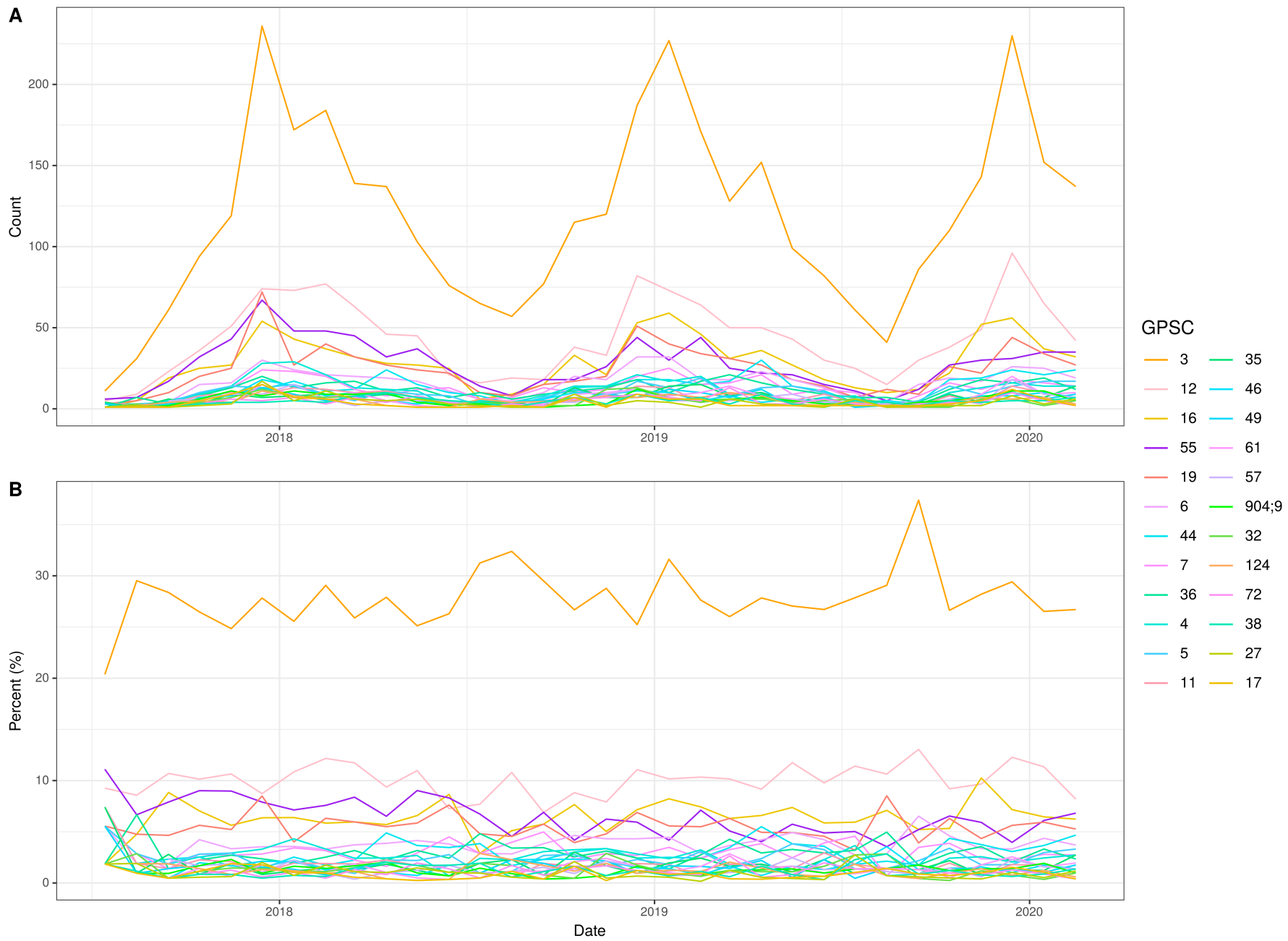

Figure S6: Global pneumococcal sequencing cluster composition over time. **(A)** The overall counts of the 24 dominant-GPSCs by month over the time period of the study. Lines are coloured by GPSC in question, with the legend ordered by the frequency of GPSCs in the collection. **(B)** The percentage of isolates expressing a certain GPSC over the study time period.

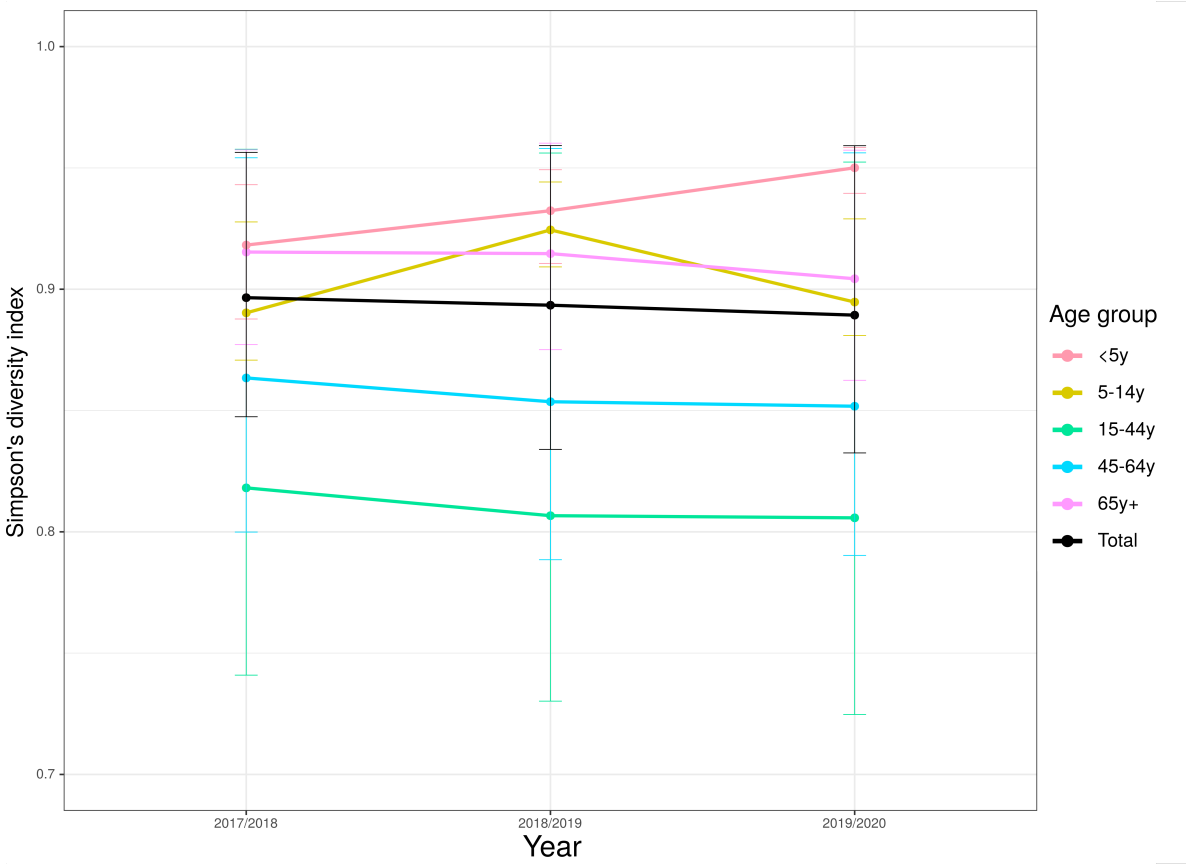

Figure S7: GPSC diversity over time. The Simpson's diversity index of isolate GPSC within patient age and epidemiological years. Lines are coloured by the patient age group, total represents all age groups. Error bars represent the 95% confidence interval for Simpson's diversity index values calculated from bootstrap sampling.

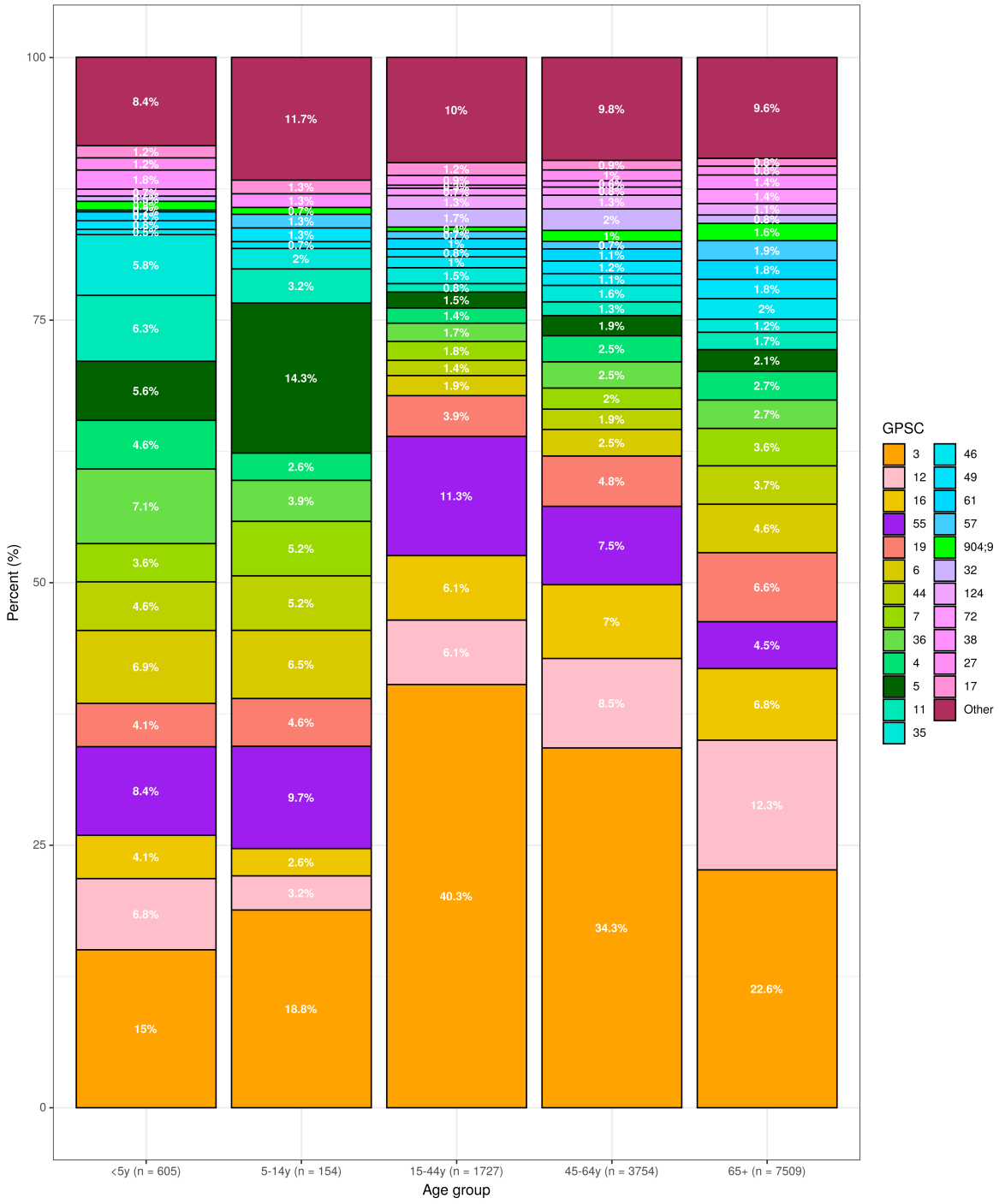

Figure S8: GPSC composition by age group. The percentage of the 24 dominant-GPSCs in each patient age group. Bars are coloured by the GPSC, with the key ordered by the frequency of the GPSCs.

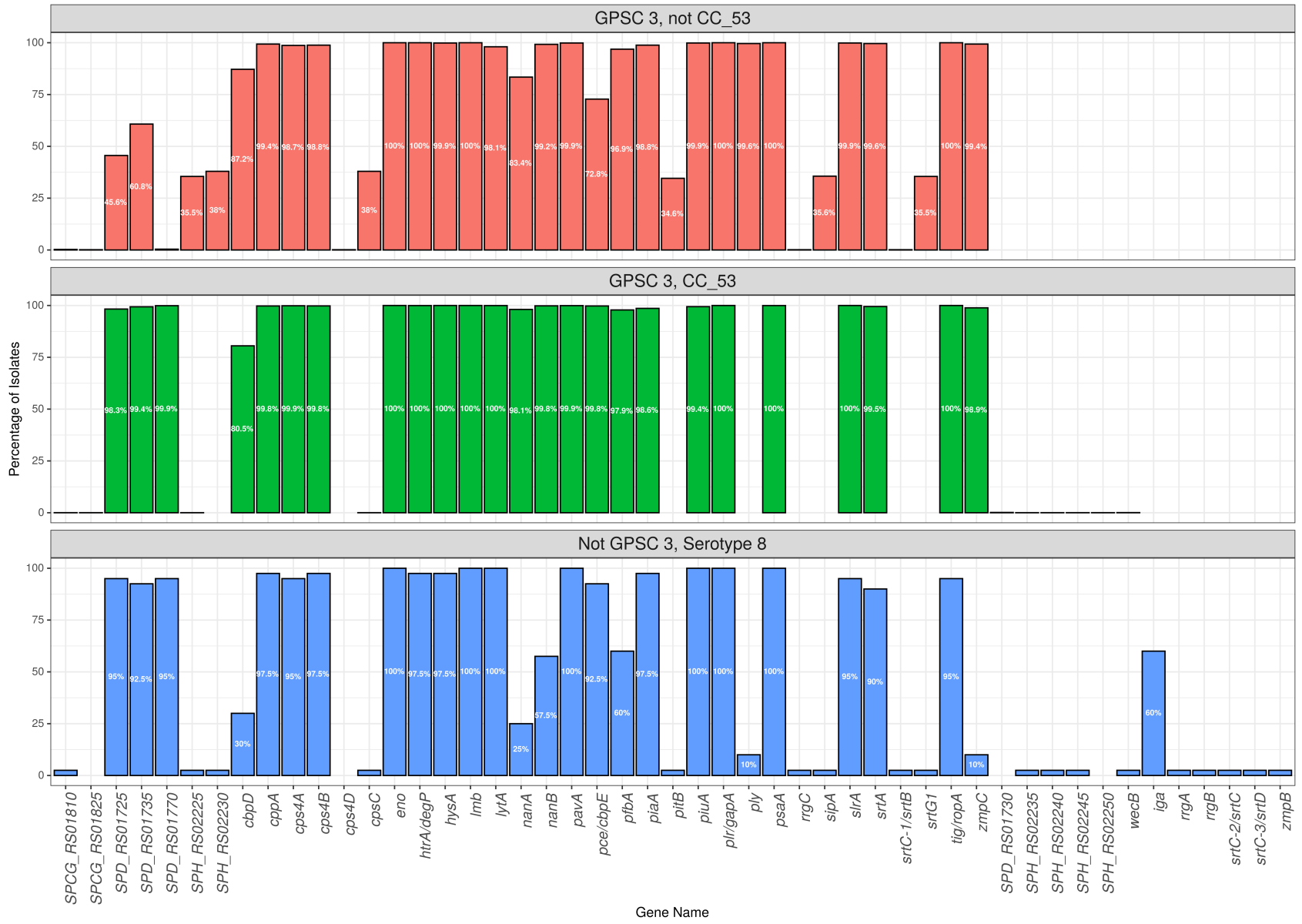

Figure S9: Distribution of virulence genes among comparison groups. The percentage of isolates in three comparison groups containing 49 separate virulence genes. Group GPSC3, not CC_53 N = 772; Group GPSC3, CC_53 N = 3027; Group Not GPSC3, Serotype 8 N = 40.

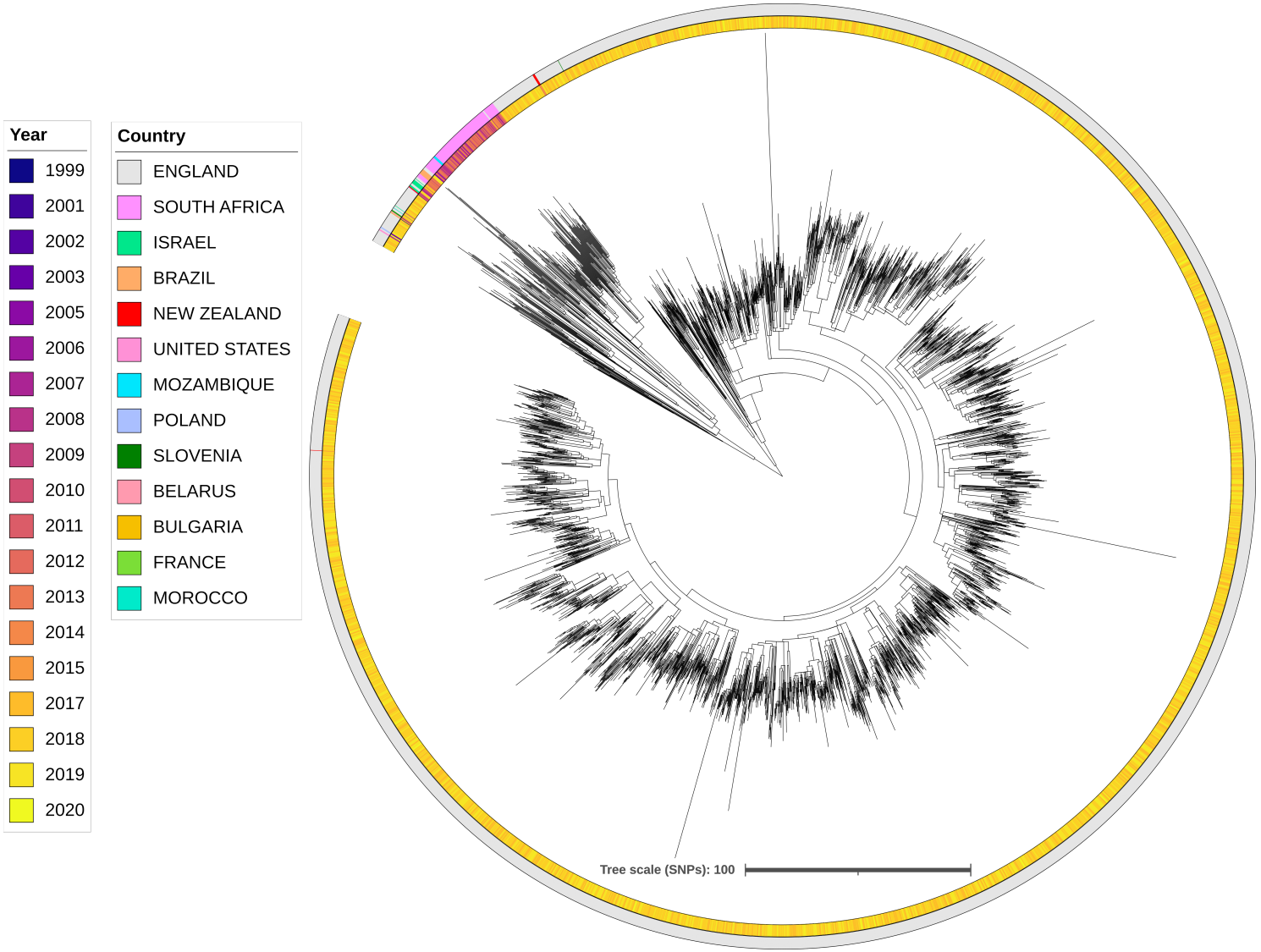

Figure S10: GPSC3 maximum likelihood phylogeny from global collection. Phylogeny formed from the non-recombinant regions of the alignment of 3,164 GPSC3 isolates. The inner annotation ring represents the year of isolation, while the outer ring represents the country of origin for an isolate.

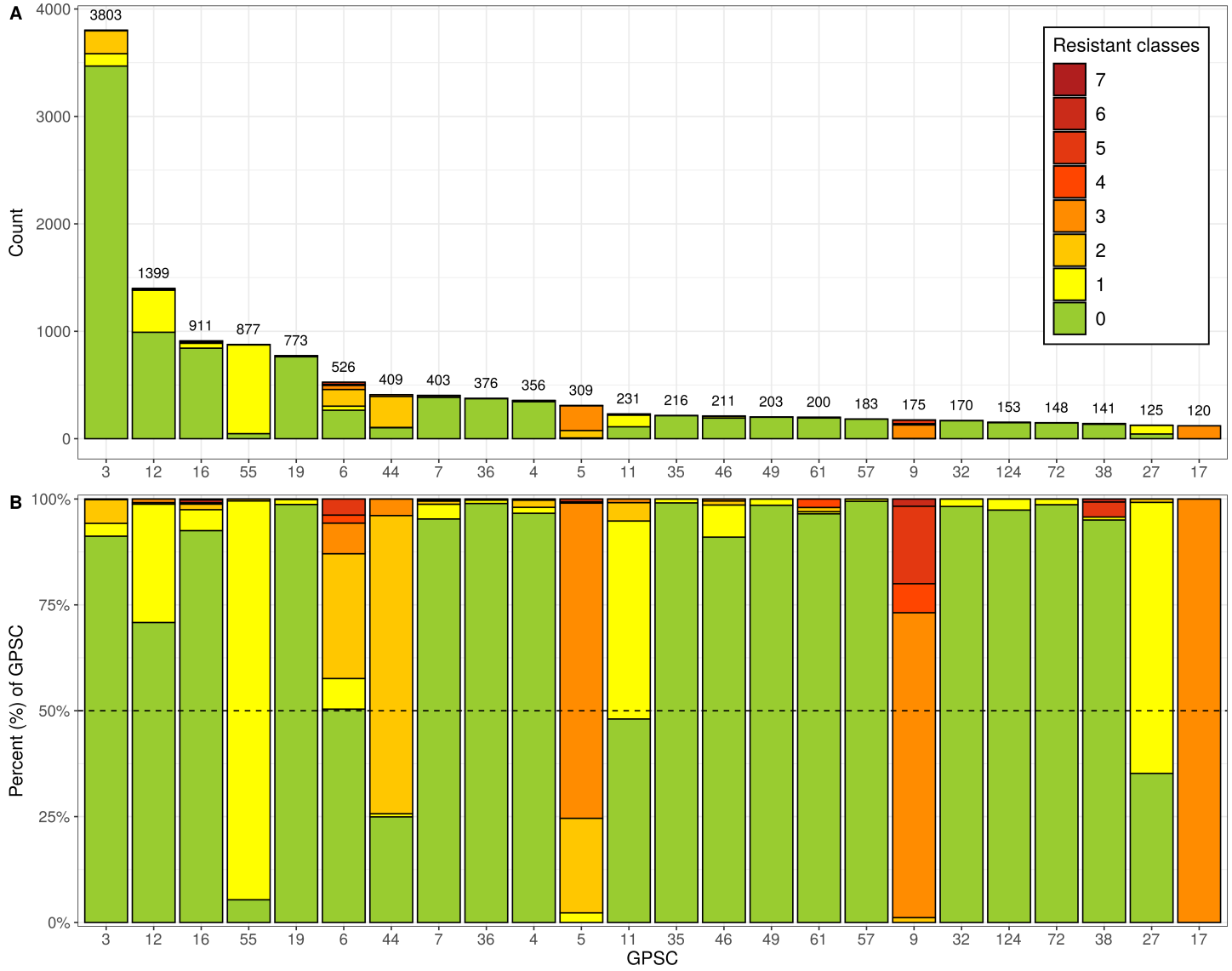

Figure S11: AMR isolates within dominant-GPSCs. A The count of isolates predicted to be resistant to a set number of antimicrobial subclasses within each of the 24 dominant-GPSCs. The bars are coloured by the number of different antimicrobial subclasses isolates are predicted to be resistant to. A value of 0 represents isolates predicted to be susceptible, 1 represents isolates predicted to be resistant to any single class of antimicrobial (i.e β-lactams). GPSCs are arranged on the x-axis in order of overall size in the collection, with the total number of isolates printed above the bar. **B** The percentage of each GPSC's isolates predicted to be resistant to the set number of different antimicrobial classes.

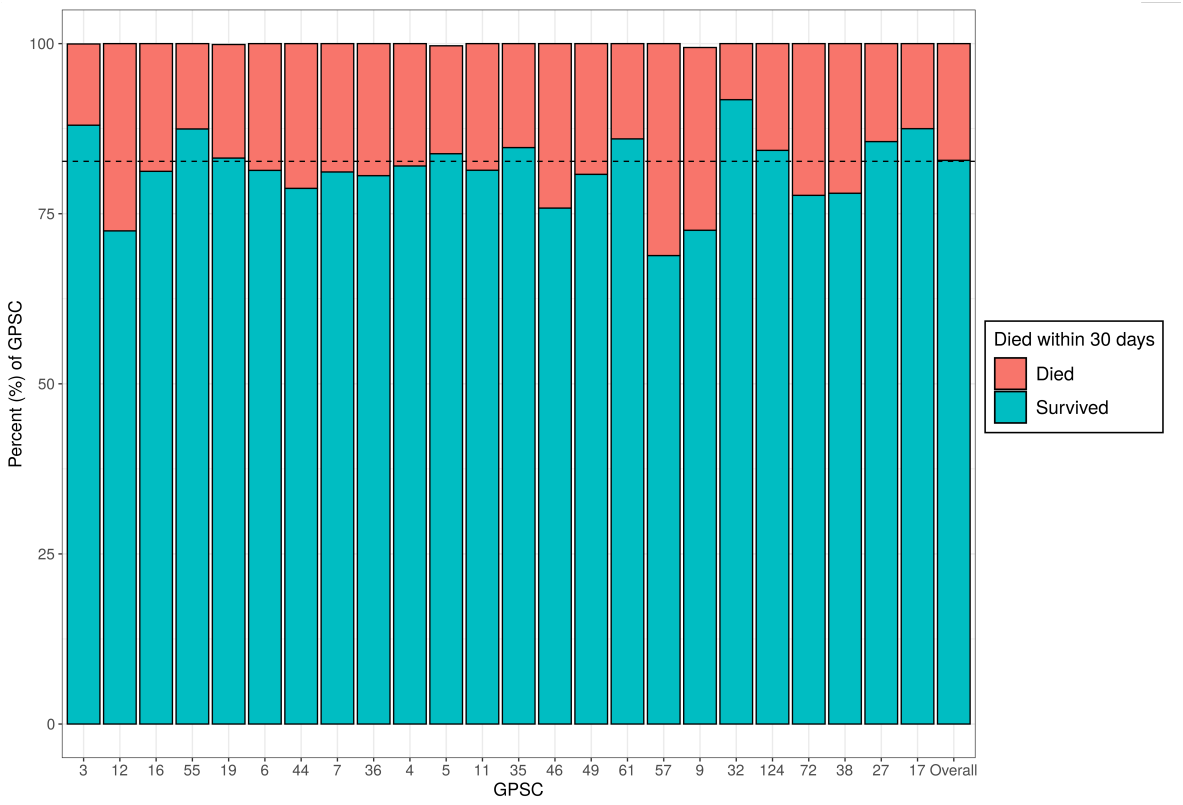

Figure S12: 30-day case fatality rate by GPSC. Bars represent the percent of each dominant-GPSC's isolates that either died or survived 30 days after laboratory confirmation of case. The final bar represents the overall split across the 24 dominant-GPSCs, with the dashed horizontal line, the split from this value. GPSCs are ordered on the x axis by their size in the collection, from the largest GPSC3, to the smallest GPSC17.

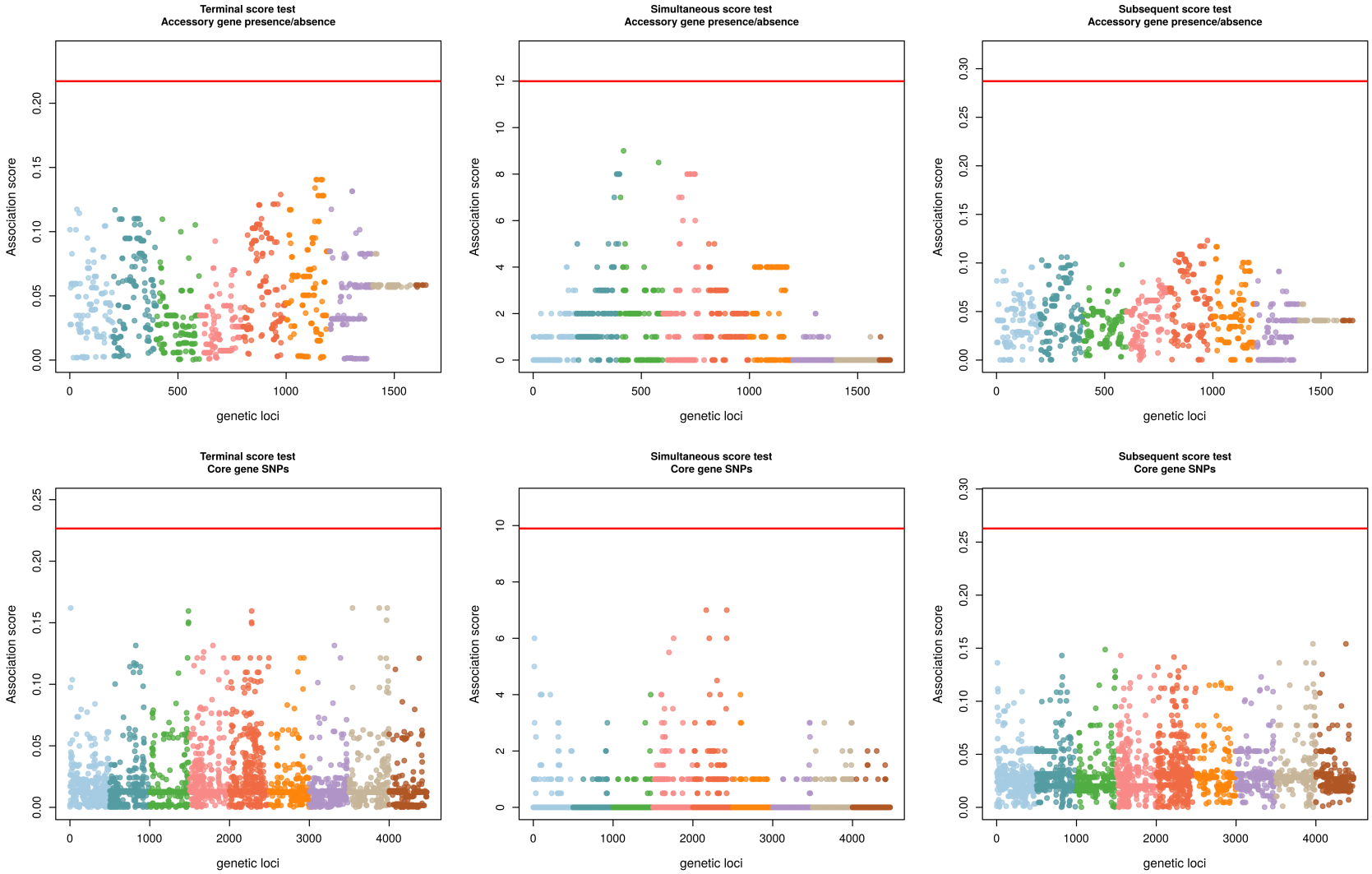

Figure S13: Manhattan plots from GWAS analysis of GPSC12 isolates. The first row of plots represents the GWAS association scores calculated by treeWAS for the accessory gene presence/absence table for the 302 GPSC12 isolates in cases where patients where aged 85+. The three plots correspond to one of the three tests which treeWAS performs, terminal score, simultaneous score, or subsequent scoring. The red horizontal line represents the corrected association score for multiple testing. The second row of plots represents the same scoring metric as described above, except for the SNPs present in the core genome alignment created for the 302 GPSC12 isolates in cases aged 85+.

References

[1] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[2] Rediat Tewolde, Timothy Dallman, Ulf Schaefer, Carmen L Sheppard, Philip Ashton, Bruno Pichon, Matthew Ellington, Craig Swift, Jonathan Green, and Anthony Underwood. Most: a modified mlst typing tool based on short read sequencing. *PeerJ*, 4:e2308, 2016.

[3] Keith A Jolley, James E Bray, and Martin CJ Maiden. Open-access bacterial population genomics: Bigsdb software, the pubmlst. org website and their applications. *Wellcome open research*, 3, 2018.

[4] Mark C Enright and Brian G Spratt. A multilocus sequence typing scheme for streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology*, 144(11):3049–3060, 1998.

[5] Georgia Kapatai, Carmen L Sheppard, Ali Al-Shahib, David J Litt, Anthony P Underwood, Timothy G Harrison, and Norman K Fry. Whole genome sequencing of streptococcus pneumoniae: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ*, 4:e2477, 2016.

[6] Philip M. Ashton, Satheesh Nair, Tansy M. Peters, Janet A. Bale, David G. Powell, Anaïs Painset, Rediat Tewolde, Ulf Schaefer, Claire Jenkins, Timothy J. Dallman, Elizabeth M. De Pinna, and Kathie A. Grant. Identification of salmonella for public health surveillance using whole genome sequencing. *PeerJ*, 2016, 2016.

[7] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.

[8] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[9] John A Lees, Simon R Harris, Gerry Tonkin-Hill, Rebecca A Gladstone, Stephanie W Lo, Jeffrey N Weiser, Jukka Corander, Stephen D Bentley, and Nicholas J Croucher. Fast and flexible bacterial genomic epidemiology with poppunk. *Genome research*, 29(2):304–316, 2019.

[10] Rebecca A Gladstone, Stephanie W Lo, John A Lees, Nicholas J Croucher, Andries J Van Tonder, Jukka Corander, Andrew J Page, Pekka Marttinen, Leon J Bentley, Theresa J Ochoa, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*, 43:338–346, 2019.

[11] Michael Feldgarden, Vyacheslav Brover, Narjol Gonzalez-Escalona, Jonathan G Frye, Julie Haendiges, Daniel H Haft, Maria Hoffmann, James B Pettengill, Arjun B Prasad, Glenn E Tillman, et al. Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific reports*, 11(1):12728, 2021.

[12] Yuan Li, Benjamin J. Metcalf, Sopio Chochua, Zhongya Li, Robert E. Gertz, Hollis Walker, Paulina A. Hawkins, Theresa Tran, Cynthia G. Whitney, Lesley McGee, and Bernard W. Beall. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in streptococcus pneumoniae. *mBio*, 7, 2016.

[13] Yuan Li, Benjamin J. Metcalf, Sopio Chochua, Zhongya Li, Robert E. Gertz, Hollis Walker, Paulina A. Hawkins, Theresa Tran, Lesley McGee, and Bernard W. Beall. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (pbp) sequences. *BMC Genomics*, 18, 8 2017.

[14] Joshua C D'Aeth, Mark PG van der Linden, Lesley McGee, Herminia de Lencastre, Paul Turner, Jae-Hoon Song, Stephanie W Lo, Rebecca A Gladstone, Raquel Sá-Leão, Kwan Soo Ko, William P Hanage, Robert F Breiman, Bernard Beall, Stephen D Bentley, Nicholas J Croucher, and The GPS Consortium. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. *eLife*, 10:e67113, jul 2021.

[15] S. R. Harris. Ska: Split kmer analysis toolkit for bacterial genomic epidemiology. *bioRxiv*, 2018.

[16] Martin Simonsen, Thomas Mailund, and Christian N. S. Pedersen. Rapid neighbour-joining. In Keith A. Crandall and Jens Lagergren, editors, *Algorithms in Bioinformatics*, pages 113–122, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[17] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 01 2014.

[18] Tal Pupko, Itsik Pe, Ron Shamir, and Dan Graur. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, 17(6):890–896, 06 2000.

[19] Marta Bertran, Joshua C D'Aeth, Fariyo Abdullahi, Seyi Eletu, Nick J Andrews, Mary E Ramsay, David J Litt, and Shamez N Ladhani. Invasive pneumococcal disease 3 years after introduction of a reduced 1+1 infant 13-valent pneumococcal conjugate vaccine immunisation schedule in england: a prospective national observational surveillance study. *The Lancet Infectious Diseases*, 2 2024. doi: 10.1016/S1473-3099(23)00706-5.

[20] Gerry Tonkin-Hill, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A. Lees, Rebecca A. Gladstone, Stephanie Lo, Christopher Beaudoin, R. Andres Floto, Simon D.W. Frost, Jukka Corander, Stephen D. Bentley, and Julian Parkhill. Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome Biology*, 21, 7 2020.

[21] Bo Liu, Dandan Zheng, Siyu Zhou, Lihong Chen, and Jian Yang. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research*, 50(D1):D912–D917, 11 2021.

[22] Benjamin Buchfink, Klaus Reuter, and Hajk Georg Drost. Sensitive protein alignments at tree-of-life scale using diamond. *Nature Methods*, 18:366–368, 4 2021.

[23] Caitlin Collins and Xavier Didelot. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Computational Biology*, 14, 2 2018.