

# **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**

Melissa J. Jansen van Rensburg<sup>1\*</sup>, Duncan J. Berger<sup>1\*</sup>, Andy Fohrmann<sup>1</sup>, James E. Bray<sup>2</sup>, Keith A. Jolley<sup>2</sup>, Martin C.J. Maiden<sup>2</sup>, Angela B. Brueggemann<sup>1</sup>

\* Equal contributors

<sup>1</sup> Nuffield Department of Population Health, University of Oxford, UK

<sup>2</sup> Department of Biology, University of Oxford, Oxford

Corresponding author:

Prof Angela Brueggemann

[angela.brueggemann@ndph.ox.ac.uk](mailto:angela.brueggemann@ndph.ox.ac.uk)

## Abstract

Investigating the genomic epidemiology of major bacterial pathogens is integral to understanding transmission, evolution, colonisation, disease, antimicrobial resistance, and vaccine impact. Furthermore, the recent accumulation of large numbers of whole genome sequences for many bacterial species enhances the development of robust genome-wide typing schemes to define the overall bacterial population structure and lineages within it. Using previously published data, we developed the Pneumococcal Genome Library (PGL), a curated dataset of 30,976 genomes and contextual data for carriage and disease pneumococci recovered between 1916-2018 in 82 countries. We leveraged the size and diversity of the PGL to develop a core genome multilocus sequence typing (cgMLST) scheme comprised of 1,222 loci. Finally, using multilevel single-linkage clustering, we stratified pneumococci into hierarchical clusters based on allelic similarity thresholds, and defined these with a taxonomic life identification number (LIN) barcoding system. The PGL, cgMLST scheme, and LIN barcodes represent a high-quality genomic resource and fine-scale clustering approaches for the analysis of pneumococcal populations, which support the genomic epidemiology and surveillance of this leading global pathogen.

## Impact statement

Many thousands of pneumococcal genomes are available in the public domain, and this creates opportunities for the scientific community to re-use existing data; however, these data are most useful when the contextual data (provenance and phenotype) are also linked to the genomes. Therefore, we created a curated, open-access database in PubMLST that contained nearly 31,000 published pneumococcal genomes and the corresponding contextual data for each genome. This large and diverse pneumococcal database was used to create a novel cgMLST scheme and multilevel clustering method to define genetic lineages with high resolution and a standardised nomenclature. These are open-access resources for all to use and provide a unified framework for the characterisation of global pneumococcal populations.

## Introduction

In the first two decades of the 21<sup>st</sup> century, the capacity to sequence the whole genome of microbes transformed the fields of microbiology and infectious disease. Furthermore, genomic epidemiology and surveillance are playing increasingly important roles in public health, vaccine development, and the assessment of vaccine impact [1-4]. Thousands of bacterial genome sequences are publicly available, and the potential for re-use of these genomes is of major benefit to the scientific community.

*Streptococcus pneumoniae* (the pneumococcus) is a bacterium that primarily resides in the healthy human nasopharynx but is also a major cause of localised and invasive infections worldwide. In 2019, prior to the COVID-19 pandemic, pneumococci were estimated to cause over 650,000 (95% UI 553,000–777,000) deaths due to pneumonia and nearly 45,000 (95% UI 34,700–59,800) deaths from meningitis among people of all ages [5]. Nonpharmaceutical interventions implemented during the pandemic to contain the spread of SARS-CoV-2 led to significant and sustained reductions in invasive pneumococcal disease (IPD), but IPD is returning to pre-pandemic levels in many countries [6-7]. Pneumococcal conjugate vaccines are used in many countries worldwide and have successfully reduced the global burden of pneumococcal disease over the past twenty years; however, the pandemic disrupted vaccination programmes worldwide and restoring these programmes is a public health priority [8].

As of 2023, there were tens of thousands of pneumococcal genome sequences in public repositories including the European Nucleotide Archive and GenBank; however, these repositories typically included minimal corresponding contextual data (provenance and laboratory data), and often the genomes were only available as unassembled short-read data, which limited their use by many investigators. Therefore, the first objective of this study was to create a Pneumococcal Genome Library (PGL) that would contain assembled genomes and corresponding contextual data from the peer-reviewed published literature and make those data freely available to all users via the PubMLST platform. We also incorporated basic genome quality criteria into the PGL to enable users, including non-specialists, to query, filter, re-analyse and download published pneumococcal genomes.

Originally, the pneumococcal population structure was defined using seven-locus multilocus sequence typing (MLST) and clonal complexes (CCs; clusters of related isolates), but recently, whole genome sequences have been used to define genetic lineages more precisely [9-12]. A well-defined population structure is the foundation on which other criteria can be mapped and interpreted, such as the lineages causing disease versus those found among healthy

individuals, or antimicrobial-resistant lineages. Understanding the expected pneumococcal population structure is also necessary to detect and interpret any observed population-level changes to that structure. For example, pneumococci typically possess a polysaccharide capsule (serotype) and this is the primary vaccine antigen. There are over 100 antigenically different serotypes, and serotypes are typically associated with specific pneumococcal genotypes; therefore, new serotype/genotype combinations are a potential indication of capsular switching events that might have occurred in response to vaccination [13-15].

The MLST approach of defining alleles based upon sequence variation at a defined set of loci was transformative because sequence data are unambiguous and easily portable, and a common nomenclature was defined [16-17]. Assigning alleles and sequence types (STs) from large numbers of genomes is easily automated, whilst still relying on the expertise of a human data curator to ensure high quality data. Genomes can also be assessed using ribosomal MLST (rMLST), which characterises allelic diversity at the 53 bacterial ribosomal genes and is especially useful for species identification [18-19].

Core genome MLST (cgMLST) schemes have been implemented for several global pathogens to assess sequence variation at hundreds of core genes across the bacterial genome and increase the resolution of defined genetic lineages [20-25]. Therefore, the second objective of this study was to develop and implement a cgMLST scheme for pneumococci, which we extended to include a taxonomic life identification number (LIN) system to improve the resolution and clustering of genetic sublineages of pneumococci. The LIN approach was originally implemented for *Klebsiella pneumoniae* and has the main advantage of providing a definitive and stable 'barcode' for each genome that can be used to define and cluster groups of genetically related isolates, based on sequence variation within the core genes used in the cgMLST scheme [26-27].

Here, we describe the development of the PGL as a community resource of nearly 31,000 assembled pneumococcal genomes with their corresponding contextual data. We used the PGL to design a robust and reproducible pneumococcal cgMLST scheme of 1,222 core genes and developed a LIN barcoding system to cluster pneumococci across multiple levels, based upon sequence variation at those 1,222 genes. We implemented the PGL, cgMLST scheme, and LIN codes within PubMLST to provide open-access resources for the scientific community and a unified framework for the characterisation of global pneumococcal populations.

## Methods

### Literature search and article eligibility

We searched PubMed on 31 January 2020 using the terms “(‘*Streptococcus pneumoniae*’ OR ‘pneumococcus’) AND (‘genome sequencing’ OR ‘genome sequence’ OR ‘Genome, Bacterial’)”. An additional search was carried out on 15 July 2020 using the terms “(‘*Streptococcus pneumoniae*’ OR ‘pneumococcus’) AND ‘genome sequencing’”. Titles and abstracts of full-text, peer-reviewed articles published in English between 1 January 2000 and 15 July 2020 were manually screened for eligibility, and methods and results sections of an article were also reviewed if necessary. Eligible publications were original research articles that described genome data from at least one naturally occurring pneumococcal isolate.

### Genome data availability

The text and supplementary files of eligible published articles were searched for International Nucleotide Sequence Database Collaboration (INSDC) accession numbers, identifiers from other public databases, or genome data. For each article, we confirmed that the number of accessions, identifiers, or data files provided matched the total number of genomes described in the text, including any reference genomes and supplementary genome data. We also confirmed that any INSDC accession numbers were valid and corresponded to pneumococcal records.

### Data acquisition and genome assembly

Contextual data (country, year of isolation, source, diagnosis, sex, age, and serotype) were extracted from the published article and supplementary files, and data were manually cleaned to conform to allowed values accepted by the *S pneumoniae* PubMLST database. If available, assembled genomes were downloaded from the NCBI Nucleotide/Assembly databases. When assemblies were not available, short-read data was downloaded from ENA and genomes were assembled *de novo* using Velvet (v1.2.10) and VelvetOptimiser (v2.2.4) using a range of kmer sizes (19-191 bp) [28]. Assembled contigs shorter than 200 bp were removed. Genomes generated in earlier studies were assembled as previously described [9-10,29-40]. Illumina MiSeq data from two datasets assembled poorly with Velvet, so these genomes were assembled with SPAdes implemented in the INNUca pipeline v.4 [41-44]. Velvet assemblies were retained for the subset that did not assemble with the INNUca pipeline.

### Creation of the PubMLST Pneumococcal Genome Library

In the *S pneumoniae* PubMLST database, isolate records were created that included the assembled genome, any corresponding provenance or laboratory data, and the PubMed

identifier of the original publication. A separate view of the *S pneumoniae* PubMLST database was created to enable users to access the PGL directly (https://pubmlst.org/organisms/streptococcus-pneumoniae/pgl). All known MLST and rMLST alleles were assigned by the BIGSdb autotagger tool in PubMLST [19]. New alleles and STs were manually curated and assigned by the database curators (ABB and JEB, respectively).

### rMLST species identification

The rMLST database (https://pubmlst.org/species-id) contains bacterial genomes compiled from the NCBI Assembly database and the assembly of short-read data [18-19]. The allelic variants of the rMLST genes of these genomes have been fully catalogued. For species identification purposes, the lowest common taxonomic node (LCTN) of each rMLST allele is calculated based upon the species annotations of the genomes that have that allele. For example, an allele observed in multiple *Streptococcus* species is assigned an LCTN of *Streptococcus* (a genus node), whereas an allele only observed in pneumococcal genomes is assigned an LCTN of *S pneumoniae*.

The rMLST species identification process includes three stages. Firstly, the query genome is scanned against the rMLST allele library using BLASTN (v. 2.12.0) and exact allelic matches are recorded [45]. Secondly, the LCTNs of the matched alleles are mapped onto the nodes of the bacterial taxonomic tree and the lowest observed non-overlapping taxonomic nodes are reported. Finally, the 'allele support' of each reported taxonomic node is calculated, which is the number of alleles observed for the reported node divided by the total number of alleles observed across all reported nodes (expressed as a percentage). A reported species node with an allele support above 90% indicates a high degree of confidence in that result.

### Genome quality control

The quality of each genome sequence was assessed based upon several criteria. Firstly, rMLST species identification was applied to each genome and interpreted as pass (only *S pneumoniae* detected, support  $\geq 90\%$ ), warning (only *S pneumoniae* detected, support  $< 90\%$ ), or fail (*S pneumoniae* not detected or *S pneumoniae* plus other organisms detected). Secondly, genomes were evaluated for evidence of mixed MLST and rMLST alleles: the presence of  $\geq 1$  allele at any MLST or rMLST locus (excluding putatively paralogous genes BACT000014 and BACT000062) flagged genomes that were potentially contaminated with non-pneumococcal DNA or consisted of multiple pneumococcal strains.

Thirdly, the interquartile deviation method was used to develop data-derived thresholds for genome size, GC content, number of contigs,  $N_{50}$ , number of Ns, and number of gaps.

Minimum and maximum thresholds (T) were set using the following equations, using  $c = 1.5$  for 'soft' thresholds and  $c = 2.2$  for 'hard' thresholds:

$$T_{min} = Q1 - (c * IQR)$$

$$T_{max} = Q3 + (c * IQR)$$

For each metric, genomes were categorised as 'pass' (between the soft thresholds), 'warning' (between the hard and soft thresholds), or 'fail' (outside the hard thresholds). The results of the rMLST species identification, MLST/rMLST allele curation, and genome quality thresholds were implemented in the *S. pneumoniae* PubMLST database to enable users to filter PGL data based on genome quality. Finally, genome completeness was assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs) (v.5.4.3) and the lactobacillales\_odb10 lineage dataset [46].

### Definition of a cgMLST genotyping scheme

Seventy-one complete pneumococcal genomes were available for download from NCBI in March 2019, of which 29 were excluded from analyses (non-RefSeq genomes,  $n = 8$ ; genomes with gaps,  $n = 5$ ; and genomes of STs represented more than once,  $n = 16$ , i.e. only one representative of each ST was selected). The chewBBACA software suite (v2.0.16) was run on the remaining 42 genomes using default parameters to create and validate a cgMLST scheme [47]. CreateSchema identified 3,139 complete, non-redundant coding sequences in each genome and alleles were defined for each of these genes using AlleleCall. Putatively paralogous genes ( $n = 42$ ) were removed using RemoveGenes and annotations for each gene were retrieved using UniprotFinder. Among the remaining 3,095 genes, 1,385 (44.8%) were detected in >95% of reference genomes, excluding genes with alleles that were 20% larger or smaller than the modal length of the distribution of matched genes.

The cgMLST scheme was assessed using 8,263 genomes from an early subset of the PGL, of which 86 were excluded from further analyses: likely contaminated based on presence of multiple alleles at rMLST loci ( $n = 74$ ); or overall size +/- 2 standard deviations from the mean pneumococcal genome size ( $n = 11$ ); or not a pneumococcus, based on rMLST species identification; or seven novel MLST alleles ( $n = 1$ ). The chewBBACA AlleleCall command was re-run on the 8,177 genomes using the 1,385 gene scheme. A further 629 genomes were removed due to high numbers of missing genes (>33) and or warnings (>26). One duplicated genome was removed. The remaining 7,547 genomes comprised the cgMLST scheme validation dataset. AlleleCall flagged 242 genes for manual review and 25 of these were removed from the scheme for these reasons: frequent absence ( $n = 15$ ); alleles 20% smaller



than the length mode of the distribution ( $n = 6$ ); transposases ( $n = 2$ ); paralogous hits ( $n = 1$ ); or frequently at the end of assembly contigs ( $n = 1$ ). Finally, a provisional 1,360 gene cgMLST scheme was implemented in PubMLST.

Initial seed alleles were identified using the validation set of 8,177 genomes plus an additional 796 recently assembled genomes. chewBBACA AlleleCall was used to identify alleles and only those at the mode length for each gene were retained. Automated curation within PubMLST (using the BIGSdb scannew.pl script, and thresholds of 98% sequence identity and 98% alignment length to existing alleles) was performed on the 7,547 genomes in the validation dataset. For genomes where alleles/genes were not identified, manual BLASTn searches were performed using lower thresholds of  $\geq 70\%$  sequence identity and at least 50% gene length. Truncated alleles, and gene sequences containing insertion sequences, mobile genetic elements or that were otherwise interrupted, were not assigned allele numbers. There were 138 core genes that contained a high proportion of disrupted or missing alleles and were thus excluded since they would be poorly reproducible in a typing scheme. Ultimately, 1,222 core genes were included in the final cgMLST scheme.

To identify the core gene alleles within each genome, three rounds of autocuration of the 1,222 genes were performed within PubMLST on the entire PGL, using the BIGSdb autotag.pl script, which used a subset of alleles as exemplars (defined using find\_exemplars.pl for each gene such that all known alleles were within 3% sequence identity to an exemplar allele of the same length). Initially, only hits with  $\geq 99\%$  sequence identity and 100% sequence alignment to an exemplar allele were allowed. The autocuration was then repeated, but the sequence identity threshold was first lowered to 97% and then to 96%; the results of these autocuration scans led to a final threshold of 97% sequence identity to exemplar alleles to define new alleles for any new genomes added to the PGL. Manual inspection and curation of 216 genes was also performed as described above to define any alleles not assigned automatically, using BLASTn with a threshold of 95% sequence identity to any known allele. EggNOG-mapper (v5.0.0) was used for functional annotation of cgMLST genes using one randomly selected representative of each coding sequence [48].

### Core genome alignment and phylogenetic analyses

A subset of the PGL was selected for phylogenetic analyses and comprised 1,870 pneumococcal genomes chosen at random, plus nine genomes of ST344 and 21 genomes of ST448. Fifty *S. pseudopneumoniae* genomes were included as an outgroup, chosen at random from a curated set of 77 *S. pseudopneumoniae* genomes from the rMLST database. The core gene sequences were retrieved by allele number from the sequence definitions database in



PubMLST for each of the 1,222 genes, aligned using MAFFT (v7.508; missing alleles were treated as gaps), and concatenated (total length 1.16 Mb) using a custom script [49-50]. A phylogenetic tree was created with FastTree (v2.1.11), using the generalised time-reversible (GTR) model of nucleotide evolution and a single rate for each site (the 'CAT' approximation) [51]. This tree was reconstructed to account for recombination using ClonalFrameML [52]. This process was repeated for all lineage-specific phylogenetic analyses. The resulting phylogenetic trees were visualised using ggtree (v3.4.2) and cophenetic distance was calculated using the cophenetic function in the R stats package (v4.2.2) [53].

### **Identification of population-wide variation in allelic mismatches**

Pairwise allelic mismatch dissimilarities and the Silhouette index ( $S_i$ ) were assessed using MSTclust (v0.21b), using a subset of 5000 PGL genomes (which provided detailed analyses in a reasonable processing time) and default parameters [27]. The Silhouette index ( $S_i$ ), a measure of cluster cohesiveness, was calculated for the full range of pairwise allelic mismatches among the core genes. The adjusted Rand Index  $R_i$  was calculated for each classification level and each pre-existing metric using the adjustedRandIndex function implemented in the mclust R package (v6.0.0) [54].

### **Comparison of clustering methods**

Clonal complexes based on the seven locus MLST scheme were assigned for all PGL isolates using Phyloviz and named after the predicted founder sequence type(s) [55]. Global Pneumococcal Sequence Clusters (GPSCs) were assigned using PopPunk (v2.6.0) and the GPS reference database (v6) [11, 56]. Mandrake was used to cluster a concatenated alignment of the 1,222 cgMLST genes from a randomly selected subset of 5000 PGL genomes, using default parameters [12].

## Results

### A diverse global dataset of 30,976 published pneumococcal genomes

A total of 211 articles published in 78 journals between 2000-2020 that contained pneumococcal whole genome sequence data were identified (Figure 1a; Supplementary Figure 1). Just over half ( $n=115$ , 54.5%) of the publications provided access to all analysed genome data, including reference genomes and contextual isolates. The remaining articles either did not provide the entire genome dataset reported in the publication or there were data integrity issues. This included six publications that contained suppressed genomes, with no published corrections or clarifications regarding the reason for suppression or potential impact on published analyses. Some data issues were resolved by further investigation and/or contacting the corresponding authors for more information (Supplementary Data 1).

Overall, the PGL was created with 33,303 genomes from 129 publications. Short-read data were downloaded and assembled for 24,967 genomes, and 8,336 assembled genomes were downloaded from public repositories. Overall, 30,976 (93.0%) genomes passed all quality control metrics (Supplementary Data 2). The genomes were generally highly contiguous (median  $N_{50}$  = 74.1 Kb; median  $L_{90}$  = 28 Kb; Figure 1b,c) and were of the expected size for a pneumococcal genome (median = 2.10 Mb, range = 1.95-2.26 Mb; Figure 1d). Genomes typically resulted in a full representation of the Lactobacillales BUSCOs (median complete and single-copy BUSCOs = 100%; Figure 1e; Supplementary Data 2; [46]). The remaining 2,165 genomes (6.5%) failed one or more quality checks, of which 67 showed evidence of contamination and 10 were not a pneumococcal genome (Supplementary Data 2). They were excluded from further analyses and filtered out of the PGL.

Pneumococci included in the PGL were recovered from 80 countries across six continents, and more than half of the collection was from South Africa ( $n = 4,887$ ), USA ( $n = 4,273$ ), The Netherlands ( $n = 3,511$ ), The Gambia ( $n = 2,859$ ), and Thailand ( $n = 2,305$ ; Figure 1f; Supplementary Data 3). Provenance data such as carriage/disease status and specimen source were available for 89.4% and 86.1% of pneumococci, respectively. Overall, 42.6% of pneumococci were nasopharyngeal carriage samples. Pneumococci were recovered between 1916-2018, and 58.0% were recovered between 2009-2018 (Figure 1g). Ninety-six serotypes were represented in the PGL, including 24 serotypes that are included in any licensed pneumococcal vaccine (Figure 1h).

There were 8,714 ribosomal sequence types (rSTs), and 4,401 sequence types (STs, seven-locus MLST scheme) that clustered into 1,482 CCs (Supplementary Data 4 and 5) represented

in the PGL. Variable-length-k-mer clustering identified 717 GPSCs (Supplementary Data 4).  
 333 Rarefaction analyses showed that the PGL effectively encompassed the known genetic  
 diversity of pneumococcal population clusters as defined by CCs and GPSCs (Figure 1i) but  
 had a more limited representation of the entire known genotyping diversity as defined by  
 336 pneumococcal STs and rSTs (>18,000 and >16,000, respectively; Supplementary Data 4).  
 There was a proportional representation of STs across CCs (Figure 1j).

### 339 **Defining a pneumococcal core genome multilocus sequence typing scheme**

Complete pneumococcal genomes ( $n = 42$ ) were used to define a provisional set of 1,385 non-  
 redundant core genes (coding sequences with start and stop codons) that were present in at  
 342 least 95% of the 42 complete genomes (Supplementary Data 6). A subset of 7,547 PGL  
 genomes was then used to assess the allele assignments for each of the 1,385 genes. Core  
 genes with high numbers of disrupted or missing alleles were excluded ( $n = 163$ ). This resulted  
 345 in a final cgMLST v1.0 scheme of 1,222 genes, which was consistent in size with previous  
 estimates of the pneumococcal core genome (range: 912-1,666 genes; [9-10, 59-60]).

348 The 1,222 core genes were evenly distributed across the pneumococcal genome and had a  
 diverse range of functions (Supplementary Figure 2; Supplementary Data 7). Per-locus  
 analysis using the pairwise homoplasy index found evidence of intragenic recombination in  
 351 36.2% of these core genes ( $n = 442$ ;  $p < 4.09 \times 10^{-5}$  after Bonferroni correction; Supplementary  
 Data 8) and greater allelic diversity as compared to non-recombining genes (Supplementary  
 Figure 3).

354 In total, 634,151 unique core gene alleles (range: 44-2,757 alleles per gene) were assigned  
 across the entire PGL (Supplementary Data 9). 96.3% of the PGL genomes were missing  
 357 fewer than 25 allele assignments across all 1,222 genes (Supplementary Data 10). The  
 number of core gene alleles assigned per genome did not vary substantially between the  
 majority of CCs, suggesting minimal phylogenetic bias resulting from missing data  
 360 (Supplementary Figures 4, 5). A core genome sequence type (cgST) was assigned to each  
 pneumococcus that had 25 or fewer missing core gene alleles ( $n = 29,893$  genomes), which  
 resulted in 27,531 unique cgSTs (Supplementary Data 4).

363 For comparison, 50 genomes of *S. pseudopneumoniae*, a closely related species, from the  
 rMLST database were screened for the presence of the 1,222 pneumococcal core genes.  
 366 Between 809 to 923 (median 884.5) of the 1,222 pneumococcal core genes were also  
 identified among the *S. pseudopneumoniae* genomes and alleles were assigned to those  
 genes (Supplementary Data 11).

### Defining the structure of pneumococcal populations using the cgMLST scheme

A set of 5000 PGL genomes was chosen at random and used to set population structure boundaries. Pairwise allelic mismatches among the 1,222 core genes formed a discontinuous distribution with three major peaks (Figure 2a; Supplementary Figure 6). The peak centred at 98.8% mismatches exclusively represented pairwise relationships between pneumococci and *S. pseudopneumoniae*. The peak centred at 93.0% mismatches predominantly represented comparisons between nontypable pneumococci of either CC344 or CC448 and other PGL pneumococci. CC344 and CC448 were previously implicated in conjunctivitis outbreaks and demonstrated a phylogenetic cluster distinct from other pneumococci [61-62]. The peak centred at 87.9% mismatches represented core gene differences between pneumococci of different CCs, sampled from different countries, in different sampling years. Therefore, a high-level species classification boundary of 96.6% allelic mismatches among 1,180 core genes was used to separate pneumococci from *S. pseudopneumoniae*.

There was no clear peak structure among pneumococci with fewer than 80% pairwise mismatches (Figure 2a; Supplementary Figure 6). To define lineage boundaries within the data, the distributions of allelic mismatches within pneumococcal groups that were defined by ribosomal MLST, seven-locus MLST, Mandrake clusters, GPSCs, and CCs, were compared (Figure 2b). In all cases the distributions were positively skewed and the majority of pairwise relationships were found below 50% mismatches.

Just over 98% of pneumococci within the same GPSC or Mandrake cluster, and 85% of pneumococci within the same CC, had 61.4% (n=750 genes) or fewer core gene allelic mismatches, thus a 'superlineage' boundary was defined at 61.4% mismatches. Core gene allelic mismatches among pneumococci with the same ST ranged from 0.0-38.2% but were predominantly low (median = 6.6% core gene allelic mismatches); therefore, a 'lineage' boundary (44.2% allelic mismatches or 540 genes) was defined to encompass all matching STs, which coincided with the highest  $S_i$  value.

Ribosomal STs (rSTs) catalogue sequence variation at ribosomal protein genes, which are conserved, robust to recombination effects, and typically used to differentiate bacterial species [18, 63]. The PGL data demonstrated that pneumococci with matching rSTs also shared the majority of core gene alleles (97.9% of pneumococci with matching rSTs had fewer than 13.1% (n=160 genes) core gene allelic mismatches). A similar observation was made among STs (69.3% of pneumococci of the same ST had fewer than 13.1% core gene allelic mismatches); therefore, we defined a 'sublineage' boundary at 13.1% core gene mismatches.

Finally, to differentiate very closely-related pneumococci more precisely, additional boundaries at 2.1%, 1.2%, 0.7%, 0.3%, 0.16% and 0.1% (corresponding to 25, 15, 8, 4, 2, and 1 gene(s), respectively) were used. The 2.1% boundary defined 'clonal group' and the remaining boundaries defined 'clonal subgroups'. Finally, a zero-mismatch boundary grouped cgMLST profiles that differed only by missing data.

To validate these boundaries, cgMLST allelic variation between pneumococci recovered from the same individuals either sampled concurrently (blood and cerebrospinal fluid; [64]), or longitudinally (1-52 weeks from birth; [65]), were compared (Figure 2c). Pneumococci recovered from the same study subjects formed clear peaks at <1% core gene allelic mismatches (Figure 2c).

### Multilevel clustering of PGL genomes

There were 27,531 unique core genome sequence types (cgSTs) in the entire PGL. Multilevel single-linkage clustering was performed using one representative of each unique cgST, which iteratively clustered cgMLST profiles based on pairwise allelic mismatches, and cgST profiles below each mismatch threshold were assigned to the same cgST cluster. Clustering was performed sequentially, with the input order determined by the tip order of cgSTs in a minimum spanning tree that was constructed based on allelic profile similarity.

During cgST clustering, a cgMLST-based life identification number (cgLIN) code was also assigned to each cgST. This multi-positional barcode has 11 numbers that reflect the cgST cluster assignment at its respective threshold at the boundaries defined above, ie species, superlineage, lineage, sublineage, clonal group, clonal subgroups, and the zero threshold. When any two barcodes were compared, the leftmost point of numerical dissimilarity reflected the threshold at which genomes no longer clustered together (Figure 3a). Therefore, cgLIN barcodes reflected an approximation of the phylogenetic relationships between genomes, for example: superlineage 0\_10 was composed of four lineages; lineage 0\_10\_0 was divided into five sublineages; and sublineage 0\_10\_0\_24 subdivided into seven distinct clonal groups (Figure 3b).

In total, 27,524 unique cgLIN codes were defined for 29,895 PGL genomes (96.5% of all PGL genomes; Supplementary Data 4). cgLIN codes defined 407 superlineages, 726 lineages, 2,782 sublineages and 12,246 clonal groups within the PGL, and 71.8% of the lineages and 50.5% of the sublineages were represented by >1 genome (median: 2 genomes; range: 1-

1,298; Supplementary Data 12). Over half of the PGL was represented by 15 superlineages  
and 24 lineages.

A subset of 1,800 pneumococcal and 50 *S. pseudopneumoniae* genomes was chosen at  
random from the PGL and rMLST databases, respectively, to determine the phylogenetic  
congruence of the assigned lineages. A maximum-likelihood phylogeny was reconstructed  
using nucleotide sequence alignments of all 1,222 cgMLST genes (Figure 4a). The phylogeny  
identified 221 lineages, and the 20 most prevalent lineages (47.4% of PGL genomes) were  
predominantly monophyletic and formed discrete phylogroups (Supplementary Figure 7).  
Comparison of allelic dissimilarity and core genome divergence (represented by cophenetic  
distance of the phylogeny) demonstrated only limited overlap of classification levels (Figure  
4b).

Among lineages, 45.5% were comprised of pneumococci from more than one country, and  
30.0% were from more than one continent (Supplementary Data 12). Consistent with previous  
studies, individual lineages were normally associated with one or a small number of serotypes  
(mode number of serotypes per lineage, 1; range, 1-21 serotypes; Supplementary Figure 8,  
Supplementary Data 12,13). Rarefaction analysis suggested wide variability in substructure  
diversity within each lineage (Figure 4d). For example, lineage 0\_0\_0 (serotype 1, CC217)  
demonstrated low clonal subgroup diversity despite extensive sampling ( $n = 1,319$  genomes),  
as compared to other lineages such as 0\_15\_0 (serotype 23F, CC81 and CC88;  $n = 681$ ) and  
0\_75\_0 (serotype 11A, CC53 and CC62;  $n = 1178$ ). The concordance between cgLIN  
clustering and other approaches was calculated using the adjusted Rand index (ARI), a  
measure of similarity between clustering approaches. At the lineage level, cgLIN clustering  
was nearly identical to GPSCs (ARI = 0.97), and highly concordant with clonal complexes (ARI  
= 0.79) and Mandrake clusters (ARI = 0.74) (Figure 4e).



## Discussion

Data sharing is essential for reproducibility and transparency in science. The open data movement has gained momentum, and many publishers, funders, and organisations encourage or require authors to share data [66]. There are strong arguments for data sharing in the field of pathogen genomics, particularly in a public health context such as when managing disease outbreaks [3, 67-68]. The COVID pandemic highlighted the benefits of rapid data sharing since publicly available SARS-CoV-2 genome data informed the development of vaccines, infection control strategies, and diagnostic assays [69]. An open data culture allows datasets to be repurposed to advance our understanding of the epidemiology, evolution, and biology of important human pathogens.

The PGL is a comprehensive, open-access database of nearly 31,000 curated pneumococcal genomes from a broad range of countries, serotypes, genotypes, clinical manifestations, and sampling years. The genomes and contextual data (provenance and phenotype) are easily accessible through the web-based PubMLST platform, which provides an extensive suite of third-party software to facilitate downstream analyses. The PGL aggregates existing genomes and helps to highlight underrepresented geographical regions that should be the focus of future genomic surveillance efforts.

Although the PGL contained genome data from 129 publications, many publications had data integrity issues that could not be resolved by the time of publication. This is contrary to open data principles and policies and is a barrier to the reproduction of published analyses to assess the validity of research findings [70-74]. Importantly, it also precludes integration of these datasets into surveillance efforts, potentially leading to unnecessary duplication of efforts or distorting estimates of global coverage of pneumococcal genomic surveillance. More stringent checks of adherence to open data standards during the publication process are necessary and could be made easier by the standardisation of data input formats. It also seems unlikely that these issues are unique to pneumococcal genomes and an assessment of published genome datasets for other microbial species is warranted.

Compared to the original seven-gene MLST, cgMLST provides much higher resolution of pneumococcal population structure, comparable to that of single nucleotide variants or variable length k-mer comparisons. Moreover, by retaining the methodological approach and increasing the number of genes, cgMLST retains the advantages of MLST, namely: consistency; standardised nomenclature; rapid allele assignment; and the representation of alleles as numerical indices [16, 25, 75]. Furthermore, this cgMLST scheme is differentiated



from other typing schemes by extensive manual curation of genes that were identified in a large validation set of genomes, which led to a large set of stable, reliably sequenced core genes and a robust genotyping scheme. Additionally, by applying this cgMLST scheme to the large PGL dataset we have already created an extensive database of pneumococcal allelic variation, and this reduces the amount of curation required going forward as new genomes are added to the PGL.

Using a wide range of fixed clustering thresholds, the pneumococcal population was differentiated at varying phylogenetic and epidemiologically relevant scales, which provided greater resolution than typing schemes with single clustering levels. The added complexity of multilevel clustering was counterbalanced with an intuitive barcoding system and classification level, to provide a consistent description of each clustering level. These analyses revealed many pneumococcal lineages, but since there were few obvious discontinuities in the percentage of pairwise allelic mismatches across the PGL, existing clustering metrics were used to guide the definition of LIN boundaries. The observed flat structure in pairwise allelic mismatches might be explained by geographical, serological or genotype-specific population substructures that are obscured by the large size of the PGL, or might be due to the naturally high recombination rates of pneumococci that may have created a natural gradient of allelic similarity. The analyses of clonal groups and clonal subgroups aimed to differentiate very closely related pneumococci and those boundaries were validated with studies of multiple and longitudinal sampling.

In conclusion, we have created a high-quality, open-access genomic resource that is representative of pneumococcal global diversity, and a novel pneumococcal cgMLST scheme and barcoding system to define and evaluate genetic lineages, in a manner that reflects the complex structure of pneumococcal populations.

## **Data availability**

The PGL, cgMLST scheme, and LIN barcodes are available from PubMLST ([https://pubmlst.org/bigssdb?db=pubmlst\\_spneumoniae\\_isolates\\_pgl](https://pubmlst.org/bigssdb?db=pubmlst_spneumoniae_isolates_pgl)). Genome accession numbers are available within each isolate record in PubMLST and in Supplementary Data 2.

## **Code availability**

The code used for data analyses is available at: <https://github.com/brueggemann-lab>

## **Authors and contributors**

Conceptualisation: MJJvR, ABB. Literature search and data acquisition: MJJvR, AF. Data curation: MJJvR, DJB, AF, JEB, KAJ, ABB. Genome assembly: JEB. Data analyses: MJJvR, DJB, JEB, KAJ, ABB. PubMLST platform and software development: JEB, KAJ, MCJM. Data visualisation: MJJvR, DJB. PubMLST funding and infrastructure: KAJ, MCJM, ABB. Writing of first draft: MJJvR, DJB, ABB. All authors contributed to the final version of the manuscript.

## **Competing interests**

The authors declare no competing interests.

## **Funding information**

This study was funded by a Wellcome Trust Investigator Award to ABB (grant number 206394/Z/17/Z), a Wellcome Trust Biomedical Resource Grant to MJCM, ABB, and KAJ (grant number 218205/Z/19/Z), and a contribution to the PGL was made by the Meningitis Research Foundation to MJCM and ABB.

## **Acknowledgements**

We gratefully acknowledge all authors who shared their genomic data on publication, thereby making the PGL possible. Mario Ramirez and João Carriço were helpful in the early development of this work, in particular with discussions about chewBBACA, and the provision of SPAdes genome assemblies. Madeleine Butler and Femke Ahlers contributed to data curation in the early development of the cgMLST scheme as part of their bioinformatics training.

## Figure legends

### Figure 1: Summary of the creation and contents of the Pneumococcal Genome Library

(PGL). a) Schematic overview of the creation of the PGL, containing 30,976 genome assemblies. b-d) Assembly statistics for all genomes. e) Number of complete and single-copy Benchmarking Universal Single-Copy Orthologs for all genomes. f) Worldwide sampling density of the PGL, coloured by the number of genomes from each country. g) Years of collection of the PGL pneumococci. h) The 24 most prevalent serotypes in the PGL, marked to indicate serotypes included in any licensed pneumococcal vaccine. i) Rarefaction curves depicting the number of unique clusters observed for each of the four different taxonomic groups, plotted from random subsets of each sample size in triplicate. Ribosomal sequence type (rST), sequence type (ST; 7-locus multilocus sequence type), clonal complex (CC), and Global Pneumococcal Sequence Cluster (GPSC). j) Number of unique sequence types within each clonal complex.

### Figure 2: Distribution of pairwise core genome MLST (cgMLST) allelic differences across globally dispersed pneumococcal populations.

a) Pairwise cgMLST allelic differences between 5000 randomly selected PGL genomes. X-axis values are plotted as a percentage of all 1,222 cgMLST loci, excluding pairwise relationships where one or both loci had unassigned alleles. The species boundary separates pneumococcal and *S. pseudopneumoniae* genomes. b) Distribution of pairwise cgMLST allelic differences between genomes belonging to the same taxonomic group: ribosomal sequence type (rST), sequence type (ST; 7-locus multilocus sequence type), Mandrake cluster (MC), Global Pneumococcal Sequence Cluster (GPSC), clonal complex (CC), and bacterial species. c). Distribution of pairwise cgMLST allelic differences between pneumococci isolated from the same individual either concurrently (blood and cerebrospinal fluid samples) or longitudinally (1-52 weeks post-birth), as published by Lees et al and Chaguza et al, respectively [64-65].

### Figure 3: Schematic representation of the cgMLST life identification number (LIN) barcoding approach.

a) Overview of LIN barcode construction for each taxonomic level. A simplified phylogeny (left) depicts the allelic mismatch values used to define each single-linkage clustering threshold. The leftmost point of numerical dissimilarity in the barcode indicates the threshold at which genomes no longer cluster together. b) A demonstration of phylogenetic relationships using superlineage 0\_10 as an example and superlineage 0\_130 as an outgroup. The relationships among lineages (left), sublineages (middle) and clonal groups (right) and the corresponding LIN barcodes are shown.

**Figure 4: Global pneumococcal population structure and taxonomic classification.**

a) Maximum-likelihood phylogeny of 1,800 pneumococcal and 50 *S pseudopneumoniae* genomes, based on a nucleotide alignment of 1,222 cgMLST loci and rooted with *S pseudopneumoniae*. The 20 most prevalent pneumococcal lineages are highlighted and annotated with CC, last three differentiating digits of the LIN barcode, and predominant serotype(s). b) Comparison of similarity among genomes represented by the phylogeny in panel a. Pairwise relationships were calculated for all 1,850 genomes and randomly subset down to 50,000 pairs. Points are coloured by the closest inferred taxonomic relationship. c) Geographical diversity of pneumococcal lineages. The ten most prevalent lineages are highlighted. d) Rarefaction curves depicting the number of unique clonal groups observed for each of the ten most prevalent lineages, plotted from random subsets of each sample size in triplicate. e) Similarity between cgMLST cluster classification levels defined in this publication and other measures of population structure. Concordance between clustering metrics was measured using the adjusted Rand Index.

## References

1. Bambini S, Rappuoli R. The use of genomics in microbial vaccine development. *Drug Discov Today*. 2009 Mar;14(5-6):252-60. doi: 10.1016/j.drudis.2008.12.007. Epub 2009 Jan 15. PMID: 19150507; PMCID: PMC7108364.
2. Hill DM, Lucidarme J, Gray SJ, Newbold LS, Ure R, et al. Genomic epidemiology of age-associated meningococcal lineages in national surveillance: an observational cohort study. *Lancet Infect Dis*. 2015 Dec;15(12):1420-8. doi: 10.1016/S1473-3099(15)00267-4. Epub 2015 Oct 27. Erratum in: *Lancet Infect Dis*. 2016 Jan;16(1):16. PMID: 26515523; PMCID: PMC4655307.
3. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat Rev Genet*. 2018 Jan;19(1):9-20. doi: 10.1038/nrg.2017.88. Epub 2017 Nov 13. PMID: 29129921; PMCID: PMC7097748.
4. Baker KS. Microbe hunting in the modern era: reflecting on a decade of microbial genomic epidemiology. *Curr Biol*. 2020 Oct 5;30(19):R1124-R1130. doi: 10.1016/j.cub.2020.06.097. PMID: 33022254; PMCID: PMC7534602.
5. GBD 2019 Antimicrobial Resistance Collaborators. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2022 Dec 17;400(10369):2221-2248. doi: 10.1016/S0140-6736(22)02185-7. Epub 2022 Nov 21. PMID: 36423648; PMCID: PMC9763654.
6. Brueggemann AB, Jansen van Rensburg MJ, Shaw D, McCarthy ND, Jolley KA, et al. 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. *Lancet Digit Health*. 2021 Jun;3(6):e360-e370. doi: 10.1016/S2589-7500(21)00077-7. Erratum in: *Lancet Digit Health*. 2021 May 26; PMID: 34045002; PMCID: PMC8166576.
7. Shaw D, Abad R, Amin-Chowdhury Z, Bautista A, Bennett D, et al. 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. *Lancet Digit Health*. 2023 Sep;5(9):e582-e593. doi: 10.1016/S2589-7500(23)00108-5. Epub 2023 Jul 27. PMID: 37516557.
8. <https://www.who.int/news/item/15-07-2022-covid-19-pandemic-fuels-largest-continued-backslide-in-vaccinations-in-three-decades>
9. van Tonder AJ, Mistry S, Bray JE, Hill DM, Cody AJ, et al. Defining the estimated core genome of bacterial populations using a Bayesian decision model. *PLoS Comput Biol*.

- 2014 Aug 21;10(8):e1003788. doi: 10.1371/journal.pcbi.1003788. PMID: 25144616; PMCID: PMC4140633.
10. van Tonder AJ, Bray JE, Jolley KA, Jansen van Rensburg M, Quirk SJ, et al. Genomic analyses of >3,100 nasopharyngeal pneumococci revealed significant differences between pneumococci recovered in four different geographical regions. *Front Microbiol.* 2019 Feb 25;10:317. doi: 10.3389/fmicb.2019.00317. PMID: 30858837; PMCID: PMC6398412.
11. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, et al; Global Pneumococcal Sequencing Consortium. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine.* 2019 May;43:338-346. doi: 10.1016/j.ebiom.2019.04.021. Epub 2019 Apr 16. PMID: 31003929; PMCID: PMC6557916.
12. Lees JA, Tonkin-Hill G, Yang Z, Corander J. Mandrake: visualizing microbial population structure by embedding millions of genomes into a low-dimensional representation. *Philos Trans R Soc Lond B Biol Sci.* 2022 Oct 10;377(1861):20210237. doi: 10.1098/rstb.2021.0237. Epub 2022 Aug 22. PMID: 35989601; PMCID: PMC9393562.
13. Brueggemann AB, Pai R, Crook DW, Beall B. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathog.* 2007 Nov;3(11):e168. doi: 10.1371/journal.ppat.0030168. PMID: 18020702; PMCID: PMC2077903.
14. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, et al. 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):e00937-20. doi: 10.1128/mBio.00937-20. PMID: 32430472; PMCID: PMC7240158.
15. Manna S, Werren JP, Ortika BD, Bellich B, Pell CL, et al. *Streptococcus pneumoniae* serotype 33G: genetic, serological, and structural analysis of a new capsule type. *Microbiol Spectr.* 2023 Dec 7:e0357923. doi: 10.1128/spectrum.03579-23. Epub ahead of print. PMID: 38059623.
16. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A.* 1998 Mar 17;95(6):3140-5. doi: 10.1073/pnas.95.6.3140. PMID: 9501229; PMCID: PMC19708.
17. 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):3049-3060. doi: 10.1099/00221287-144-11-3049. PMID: 9846740.

18. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology (Reading). 2012 Apr;158(Pt 4):1005-1015. doi: 10.1099/mic.0.055459-0. Epub 2012 Jan 27. PMID: 22282518; PMCID: PMC3492749.
19. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018 Sep 24;3:124. doi: 10.12688/wellcomeopenres.14826.1. PMID: 30345391; PMCID: PMC6192448.
20. Harrison OB, Cehovin A, Skett J, Jolley KA, Massari P, et al. *Neisseria gonorrhoeae* population genomics: use of the gonococcal core genome to improve surveillance of antimicrobial resistance. J Infect Dis. 2020 Nov 9;222(11):1816-1825. doi: 10.1093/infdis/jiaa002. PMID: 32163580; PMCID: PMC7653085.
21. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ. Core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates. J Clin Microbiol. 2017 Jul;55(7):2086-2097. doi: 10.1128/JCM.00080-17. Epub 2017 Apr 26. PMID: 28446571; PMCID: PMC5483910.
22. Abdel-Glil MY, Thomas P, Linde J, Jolley KA, Harmsen D, et al. Establishment of a publicly available core genome multilocus sequence typing scheme for *Clostridium perfringens*. Microbiol Spectr. 2021 Oct 31;9(2):e0053321. doi: 10.1128/Spectrum.00533-21. Epub 2021 Oct 27. PMID: 34704797; PMCID: PMC8549748.
23. Gonzalez-Escalona N, Jolley KA, Reed E, Martinez-Urtaza J. Defining a core genome multilocus sequence typing scheme for the global epidemiology of *Vibrio parahaemolyticus*. J Clin Microbiol. 2017 Jun;55(6):1682-1697. doi: 10.1128/JCM.00227-17. Epub 2017 Mar 22. PMID: 28330888; PMCID: PMC5442524.
24. Abdel-Glil MY, Thomas P, Brandt C, Melzer F, Subbaiyan A, et al. Core genome multilocus sequence typing scheme for improved characterization and epidemiological surveillance of pathogenic *Brucella*. J Clin Microbiol. 2022 Aug 17;60(8):e0031122. doi: 10.1128/jcm.00311-22. Epub 2022 Jul 19. PMID: 35852343; PMCID: PMC9387271.
25. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 2013 Oct;11(10):728-36. doi: 10.1038/nrmicro3093. Epub 2013 Sep 2. PMID: 23979428; PMCID: PMC3980634.
26. Vinatzer BA, Tian L, Heath LS. A proposal for a portal to make earth's microbial diversity easily accessible and searchable. Antonie Van Leeuwenhoek. 2017



Oct;110(10):1271-1279. doi: 10.1007/s10482-017-0849-z. Epub 2017 Mar 9. PMID: 28281028.

27. Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, et al. A dual barcoding approach to bacterial strain nomenclature: genomic taxonomy of *Klebsiella pneumoniae* strains. Mol Biol Evol. 2022 Jul 2;39(7):msac135. doi: 10.1093/molbev/msac135. PMID: 35700230; PMCID: PMC9254007.
28. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Curr Protoc Bioinformatics. 2010 Sep;Chapter 11:Unit 11.5. doi: 10.1002/0471250953.bi1105s31. PMID: 20836074; PMCID: PMC2952100.
29. Bogaardt C, van Tonder AJ, Brueggemann AB. Genomic analyses of pneumococci reveal a wide diversity of bacteriocins - including pneumocyclicin, a novel circular bacteriocin. BMC Genomics. 2015 Jul 28;16(1):554. doi: 10.1186/s12864-015-1729-4. PMID: 26215050; PMCID: PMC4517551.
30. Brueggemann AB, Harrold CL, Rezaei Javan R, van Tonder AJ, McDonnell AJ, et al. Pneumococcal prophages are diverse, but not without structure or history. Sci Rep. 2017 Feb 20;7:42976. doi: 10.1038/srep42976. PMID: 28218261; PMCID: PMC5317160.
31. Kurioka A, van Wilgenburg B, Javan RR, Hoyle R, van Tonder AJ, et al. Diverse *Streptococcus pneumoniae* strains drive a mucosal-associated invariant T-Cell response through major histocompatibility complex class I-related molecule-dependent and cytokine-driven pathways. J Infect Dis. 2018 Mar 5;217(6):988-999. doi: 10.1093/infdis/jix647. PMID: 29267892; PMCID: PMC5854017.
32. Quirk SJ, Haraldsson G, Erlendsdóttir H, Hjálmarsdóttir MÁ, van Tonder AJ, et al. Effect of vaccination on pneumococci isolated from the nasopharynx of healthy children and the middle ear of children with otitis media in Iceland. J Clin Microbiol. 2018 Nov 27;56(12):e01046-18. doi: 10.1128/JCM.01046-18. PMID: 30257906; PMCID: PMC6258863.
33. Quirk SJ, Haraldsson G, Hjálmarsdóttir MÁ, van Tonder AJ, Hrafnkelsson B, et al. Vaccination of Icelandic children with the 10-valent pneumococcal vaccine leads to a significant herd effect among adults in Iceland. J Clin Microbiol. 2019 Mar 28;57(4):e01766-18. doi: 10.1128/JCM.01766-18. PMID: 30651396; PMCID: PMC6440763.
34. Rezaei Javan R, van Tonder AJ, King JP, Harrold CL, Brueggemann AB. Genome sequencing reveals a large and diverse repertoire of antimicrobial peptides. Front Microbiol. 2018 Aug 27;9:2012. doi: 10.3389/fmicb.2018.02012. PMID: 30210481; PMCID: PMC6120550.

35. 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. doi: 10.1038/s41467-019-12825-y. PMID: 31649284; PMCID: PMC6813308.
36. van Tonder AJ, Bray JE, Roalfe L, White R, Zancolli M, et al. Genomics reveals the worldwide distribution of multidrug-resistant serotype 6E pneumococci. *J Clin Microbiol.* 2015 Jul;53(7):2271-85. doi: 10.1128/JCM.00744-15. PMID: 25972423; PMCID: PMC4473186.
37. van Tonder AJ, Bray JE, Quirk SJ, Haraldsson G, Jolley KA, et al. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. *Microb Genom.* 2016 Oct 1;2(10):000090. doi: 10.1099/mgen.0.000090. PMID: 28133541; PMCID: PMC5266551.
38. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, et al. The multidrug-resistant PMEN1 pneumococcus is a paradigm for genetic success. *Genome Biol.* 2012 Nov 16;13(11):R103. doi: 10.1186/gb-2012-13-11-r103. PMID: 23158461; PMCID: PMC3580495.
39. Wyres KL, van Tonder A, Lambertsen LM, Hakenbeck R, Parkhill J, et al. Evidence of antimicrobial resistance-conferring genetic elements among pneumococci isolated prior to 1974. *BMC Genomics.* 2013 Jul 24;14:500. doi: 10.1186/1471-2164-14-500. PMID: 23879707; PMCID: PMC3726389.
40. Wyres KL, Lambertsen LM, Croucher NJ, McGee L, von Gottberg A, et al. Pneumococcal capsular switching: a historical perspective. *J Infect Dis.* 2013 Feb 1;207(3):439-49. doi: 10.1093/infdis/jis703. Epub 2012 Nov 21. PMID: 23175765; PMCID: PMC3537446.
41. Andam CP, Mitchell PK, Callendrello A, Chang Q, Corander J, et al. Genomic epidemiology of penicillin-nonsusceptible pneumococci with nonvaccine serotypes causing invasive disease in the United States. *J Clin Microbiol.* 2017 Apr;55(4):1104-1115. doi: 10.1128/JCM.02453-16. Epub 2017 Jan 18. PMID: 28100596; PMCID: PMC5377837.
42. Chang B, Morita M, Lee KI, Ohnishi M. Whole-Genome sequence analysis of *Streptococcus pneumoniae* strains that cause hospital-acquired pneumonia infections. *J Clin Microbiol.* 2018 Apr 25;56(5):e01822-17. doi: 10.1128/JCM.01822-17. PMID: 29444837; PMCID: PMC5925718.
43. <https://github.com/B-UMMI/INNUca>

- 798 44. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes de  
novo assembler. *Curr Protoc Bioinformatics*. 2020 Jun;70(1):e102. doi:  
10.1002/cpbi.102. PMID: 32559359.
- 801 45. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. BLAST+:  
architecture and applications. *BMC Bioinformatics*. 2009 Dec 15;10:421. doi:  
10.1186/1471-2105-10-421. PMID: 20003500; PMCID: PMC2803857.
- 804 46. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: novel  
and streamlined workflows along with broader and deeper phylogenetic coverage for  
scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021 Sep  
807 27;38(10):4647-4654. doi: 10.1093/molbev/msab199. PMID: 34320186; PMCID:  
PMC8476166.
- 810 47. Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, et al. chewBBACA: A  
complete suite for gene-by-gene schema creation and strain identification. *Microb  
Genom*. 2018 Mar;4(3):e000166. doi: 10.1099/mgen.0.000166. Epub 2018 Mar 15.  
PMID: 29543149; PMCID: PMC5885018.
- 813 48. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, et al.  
eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology  
resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019 Jan  
816 8;47(D1):D309-D314. doi: 10.1093/nar/gky1085. PMID: 30418610; PMCID:  
PMC6324079.
49. Katoh K, Asimenos G, Toh H. Multiple alignment of DNA sequences with MAFFT.  
819 *Methods Mol Biol*. 2009;537:39-64. doi: 10.1007/978-1-59745-251-9\_3. PMID:  
19378139.
50. <https://github.com/brueggemann-lab>
- 822 51. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees  
for large alignments. *PLoS One*. 2010 Mar 10;5(3):e9490. doi:  
10.1371/journal.pone.0009490. PMID: 20224823; PMCID: PMC2835736.
- 825 52. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole  
bacterial genomes. *PLoS Comput Biol*. 2015 Feb 12;11(2):e1004041. doi:  
10.1371/journal.pcbi.1004041. PMID: 25675341; PMCID: PMC4326465.
- 828 53. Yu G, Lam TT, Zhu H, Guan Y. Two methods for mapping and visualizing associated  
data on phylogeny using ggtree. *Mol Biol Evol*. 2018 Dec 1;35(12):3041-3043. doi:  
10.1093/molbev/msy194. PMID: 30351396; PMCID: PMC6278858.
- 831 54. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and  
density estimation using Gaussian finite mixture models. *R J*. 2016 Aug;8(1):289-317.  
PMID: 27818791; PMCID: PMC5096736.

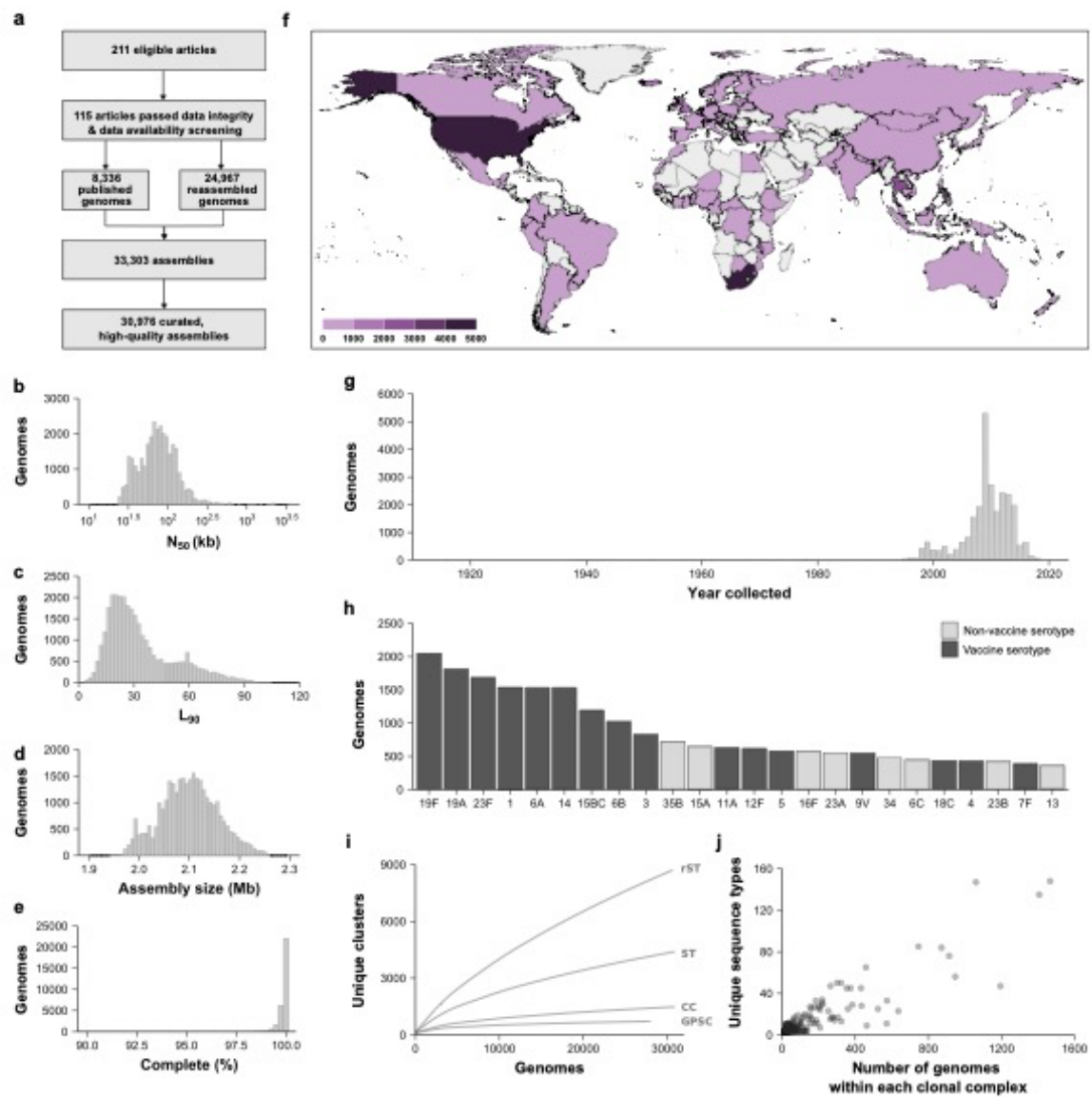
55. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics*. 2017 Jan 1;33(1):128-129. doi: 10.1093/bioinformatics/btw582. Epub 2016 Sep 6. PMID: 27605102.
56. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res*. 2019 Feb;29(2):304-316. doi: 10.1101/gr.241455.118. Epub 2019 Jan 24. PMID: 30679308; PMCID: PMC6360808.
57. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, et al; Global Pneumococcal Sequencing Consortium. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine*. 2019 May;43:338-346. doi: 10.1016/j.ebiom.2019.04.021. Epub 2019 Apr 16. PMID: 31003929; PMCID: PMC6557916.
58. Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, et al. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol*. 2010;11(10):R107. doi: 10.1186/gb-2010-11-10-r107. Epub 2010 Oct 29. PMID: 21034474; PMCID: PMC3218663.
59. Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, et al. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol*. 2007 Nov;189(22):8186-95. doi: 10.1128/JB.00690-07. Epub 2007 Aug 3. PMID: 17675389; PMCID: PMC2168654.
60. Rosconi F, Rudmann E, Li J, Surujon D, Anthony J, Frank M, et al. A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat Microbiol*. 2022 Oct;7(10):1580-1592. doi: 10.1038/s41564-022-01208-7. Epub 2022 Sep 12. PMID: 36097170; PMCID: PMC9519441.
61. Porat N, Greenberg D, Givon-Lavi N, Shuval DS, Trefler R, et al. The important role of nontypable *Streptococcus pneumoniae* international clones in acute conjunctivitis. *J Infect Dis*. 2006 Sep 1;194(5):689-96. doi: 10.1086/506453. Epub 2006 Jul 28. PMID: 16897669.
62. Marimon JM, Ercibengoa M, García-Arenzana JM, Alonso M, Pérez-Trallero E. *Streptococcus pneumoniae* ocular infections, prominent role of unencapsulated isolates in conjunctivitis. *Clin Microbiol Infect*. 2013 Jul;19(7):E298-305. doi: 10.1111/1469-0691.12196. Epub 2013 Mar 20. PMID: 23517475.
63. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MC, et al. The landscape of realized homologous recombination in pathogenic bacteria. *Mol Biol Evol*. 2016 Feb;33(2):456-71. doi: 10.1093/molbev/msv237. Epub 2015 Oct 29. PMID: 26516092; PMCID: PMC4866539.

64. Lees JA, Kremer PHC, Manso AS, Croucher NJ, Ferwerda B, et al. Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom.* 2017 Jan 31;3(1):e000103. doi: 10.1099/mgen.0.000103. PMID: 28348877; PMCID: PMC5361624.
65. Chaguza C, Senghore M, Bojang E, Gladstone RA, Lo SW, et al. Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation. *Nat Commun.* 2020 Jul 10;11(1):3442. doi: 10.1038/s41467-020-17327-w. PMID: 32651390; PMCID: PMC7351774.
66. Tenopir C, Rice NM, Allard S, Baird L, Borycz J, et al. Data sharing, management, use, and reuse: practices and perceptions of scientists worldwide. *PLoS One.* 2020 Mar 11;15(3):e0229003. doi: 10.1371/journal.pone.0229003. PMID: 32160189; PMCID: PMC7065823.
67. Orata FD, Keim PS, Boucher Y. The 2010 cholera outbreak in Haiti: how science solved a controversy. *PLoS Pathog.* 2014 Apr 3;10(4):e1003967. doi: 10.1371/journal.ppat.1003967. PMID: 24699938; PMCID: PMC3974815.
68. Pettengill JB, Markell A, Conrad A, Carleton HA, Beal J, et al. A multinational listeriosis outbreak and the importance of sharing genomic data. *Lancet Microbe.* 2020 Oct;1(6):e233-e234. doi: 10.1016/S2666-5247(20)30122-1. Epub 2020 Oct 7. PMID: 35544215.
69. Rito T, Fernandes P, Duarte R, Soares P. Evaluating data sharing of SARS-CoV-2 genomes for molecular epidemiology across the COVID-19 pandemic. *Viruses.* 2023 Feb 17;15(2):560. doi: 10.3390/v15020560. PMID: 36851774; PMCID: PMC9959893.
70. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. PMID: 26978244; PMCID: PMC4792175.
71. <https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy>
72. <https://www.ukri.org/publications/ukri-open-access-policy/uk-research-and-innovation-open-access-policy/>
73. <https://openaccess.gatesfoundation.org/how-to-comply/data-sharing-requirements/>
74. <https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards>
75. Blanc DS, Magalhães B, Koenig I, Senn L, Grandbastien B. Comparison of whole genome (wg-) and core genome (cg-) MLST (BioNumerics™) versus SNP variant calling for epidemiological investigation of *Pseudomonas aeruginosa*. *Front Microbiol.*

2020 Jul 22;11:1729. doi: 10.3389/fmicb.2020.01729. PMID: 32793169; PMCID: PMC7387498.

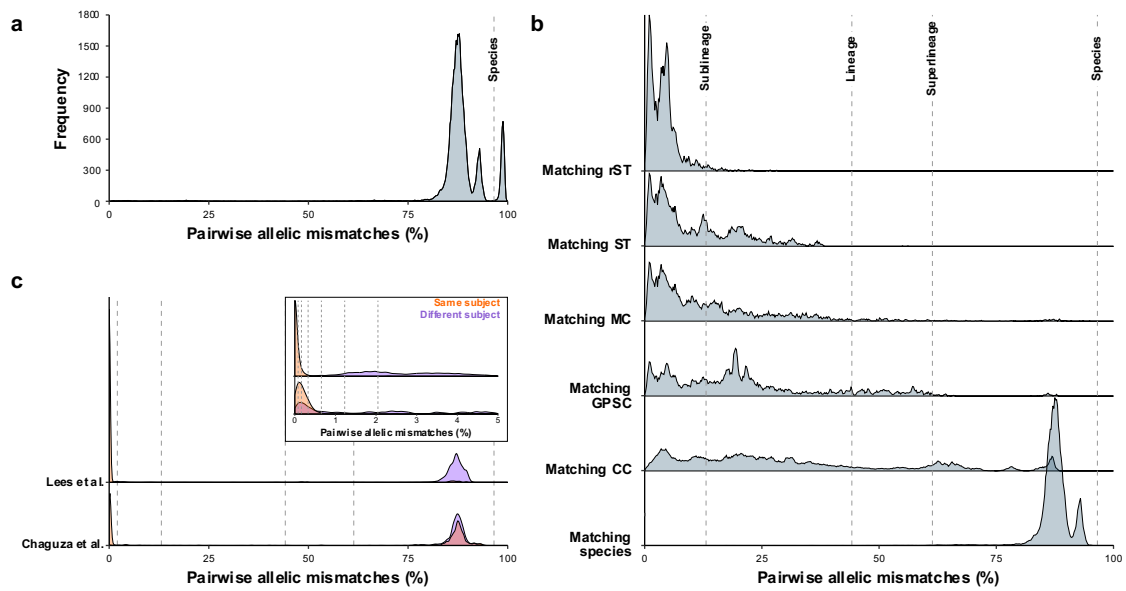
909

**Figure 1**

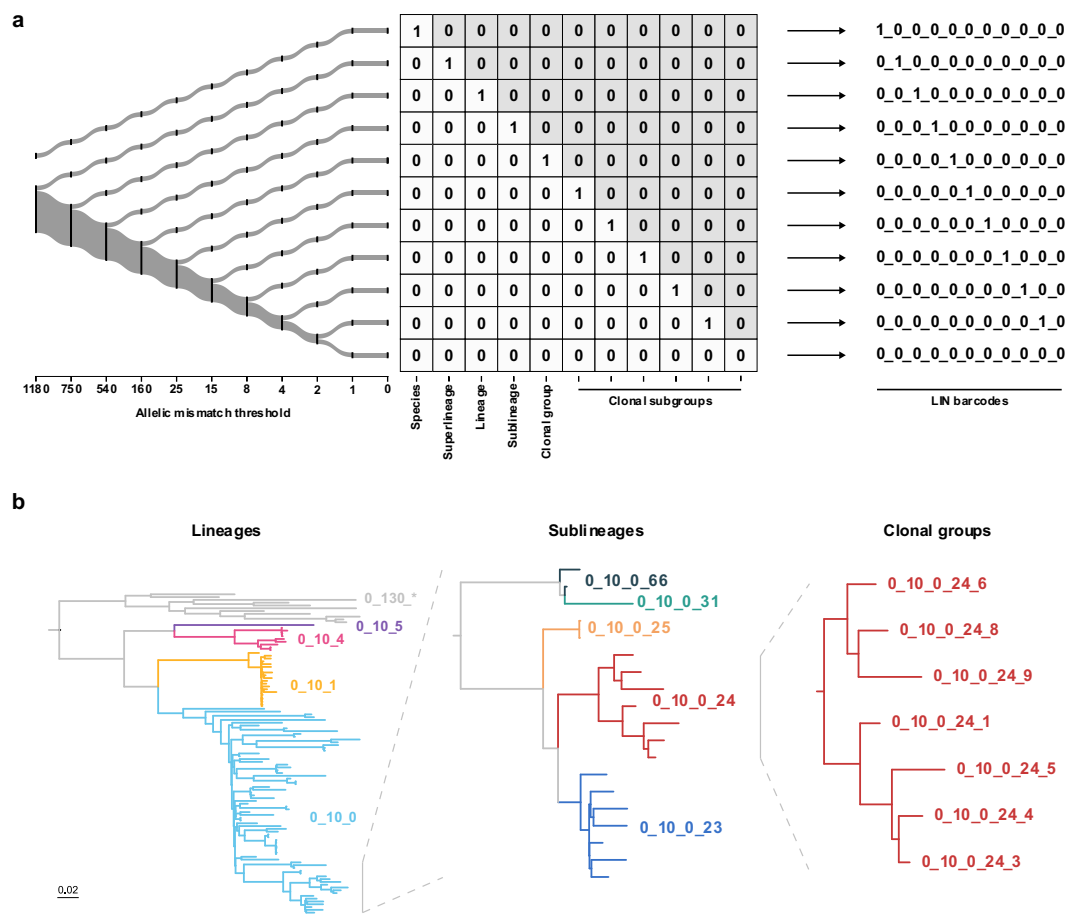




**Figure 2**



**Figure 3**



**Figure 4**

