Pneumo-Typer: integrated genomic surveillance tool for capsule genotype, serotype and sequence type in *Streptococcus pneumoniae* informs vaccine strategies

Xiangyang Li[2†], Huejie Zhang[1,3†], Yaoyao Zhu[4†], Zilin Yang[2], Xiangyu Wang[4], Guohui Zhang[2], Xuan Zhao[5], Yinyan Huang[6], Bingqing Li[1,5]* and Zhongrui Ma[1,4]*

[1]Department of Clinical Laboratory, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan 250021, China.

[2]College of Environment and Life Sciences, Bacterial Genome Data Mining & Bioinformatic Analysis Center, School of Sciences, Kaili University, Kaili 556011, China

[3]Department of Public Health and Health Management, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan 250117, China.

[4]Medical Science and Technology Innovation Center, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan 250117, China.

[5]Department of Pathogen Biology, School of Clinical and Basic Medical Sciences, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan 250117, China.

[6]Department of TB Control and Prevention, Hangzhou Center for Disease Control and Prevention (Hangzhou Health Supervision Institution), Hangzhou 310021, China.

[†]Xiangyang Li, Huajie Zhang and Yaoyao Zhu contributed equally to this work.

*Correspondence:

Bingqing Li

bingqingsdu@163.com

Zhongrui Ma

mazhongrui@sdfmu.edu.cn

Abstract

Background: *Streptococcus pneumoniae* persists as a leading global pathogen, with current capsule-based vaccines like PCV13 offering limited coverage against its extensive serotype diversity (>100 serotypes). This restricted coverage drives the emergence of non-vaccine serotypes. Moreover, vaccine efficacy is further compromised by immune evasion caused by capsule genotype (CapT) variations within vaccine-covered serotypes, while existing genome-based tools lack robust analytical capabilities for CapT.

Results: We developed Pneumo-Typer, a CapT visualization tool enabling rapid identification of CapT variations in *S. pneumoniae*. Application to 17,407 publicly available genomes from the NCBI database revealed widespread CapT variations across vaccine-covered serotypes, providing critical insights for optimizing vaccine formulations. In addition, Pneumo-Typer integrates high-accuracy serotype prediction (98.44% accuracy across 93 serotypes) and multilocus sequence typing (ST), facilitating comprehensive genomic surveillance. Besides, CapT and ST information can assist in serotype determination. Pneumo-Typer enables integrated analysis of serotype-CapT-ST relationships, demonstrating that while most serotypes associate with multiple STs and CapTs, ST and CapT lineages evolve independently. Strikingly, multi-serotype serogroups exhibited pronounced CapT diversity, suggesting evolutionary processes driving serogroup expansion through CapT plasticity.

Conclusions: In summary, Pneumo-Typer is an integrated platform for genomic surveillance of pneumococcal populations. By resolving previously underexplored serotype-CapT-ST linkages, the tool provides a framework for next-generation vaccine design, emphasizing the need for vaccine formulations that address both serotype replacement and CapT-driven immune evasion.

Keywords Pneumo-Typer, Capsule genotype, Serotype, Sequence type, Vaccine strategy, *Streptococcus pneumoniae*

Background

Streptococcus pneumoniae remains a leading global pathogen responsible for over 1.2 million annual deaths due to pneumonia, meningitis, and invasive pneumococcal disease (IPD), particularly among young children, the elderly, and immunocompromised individuals [1, 2]. The polysaccharide capsule, which defines over 100 distinct serotypes, is both a critical virulence factor and the primary target of pneumococcal vaccines [3, 4]. Current vaccines, such as the 13-valent pneumococcal conjugate vaccine (PCV13), have significantly reduced disease burden by targeting predominant serotypes [5].

However, pneumococcal vaccine efficacy is increasingly challenged by two phenomena: serotype replacement driven by the expansion of non-vaccine serotypes [6-8], and capsular genotype (CapT)-driven immune evasion within vaccine-targeted serotypes. While serotype replacement has been widely recognized and addressed through continuous updates to vaccine serotype composition [9], CapT variations remain an emerging concern. These genetic alterations in capsule biosynthesis can lead to antigenic shifts, allowing strains to evade vaccine-induced protection [10-13]. For instance, several strains classified as serotype 14 by molecular-based serotyping tools (SeroBA and SeroCall) exhibit significant alterations in their CapTs, allowing them to bypass vaccine protection [14]. Notably, these molecular methods may fail to fully capture CapT variations, further complicating surveillance efforts.

Traditional serotyping methods, such as the Quellung reaction, are labor-intensive and require specialized expertise [15-17]. Molecular-based serotyping tools, including PneumoCaT [18] and SeroCall [19], utilize sequence alignment approaches but depend on high-quality whole genome sequencing (WGS) data. Other tools, such as SeroBA [20], PneumoKITy [21], and PfaSTer [22], employ *k*-mer-based approaches for faster serotyping but may struggle with mosaic sequences, which are prevalent in pneumococcal populations [23]. Detecting and analyzing CapT variations remains particularly challenging due to the complexity of current bioinformatics workflows, which often require multiple platforms and substantial computational expertise [24, 25]. These challenges underscore the urgent need for high-resolution molecular surveillance tools capable of tracking both serotype dynamics and CapT variations.

While WGS has improved resolution for lineage tracking through sequence types (STs) [26] or global pneumococcal sequence clusters (GPSCs) [27], existing bioinformatics tools lack integrated analysis of capsular locus heterogeneity (CapT variations), serotype prediction, and ST assignment in a unified framework [28]. This gap hinders comprehensive surveillance of vaccine-driven selection pressures, as exemplified by the emergence of serotype 19A-ST320 [29], serotype 8-ST53 [30], and serotype 3-ST271 [31].

To address these challenges, we developed Pneumo-Typer, a high-throughput analytical pipeline that enables simultaneous identification of CapTs, serotypes, and STs (with an optional cgST) from assembled WGS data. The tool demonstrated high efficiency by processing 21,260 genomes within approximately one hour while achieving 98.44% serotyping accuracy across 93 serotypes. Application to 17,407 publicly available pneumococcal genomes from the NCBI database revealed extensive CapT variations, providing critical insights for optimizing vaccine formulations. Furthermore, systematic analysis elucidated complex CapT-serotype-ST relationships, revealing previously unrecognized patterns of pneumococcal evolution. This integrated tool bridges the gap between genomic surveillance and vaccine development, offering a scalable solution for monitoring pneumococcal evolution in the post-pneumococcal vaccine era.

Implementation

Pneumo-Typer workflow

Pneumo-Typer, developed in Perl, offers a distinct advantage over several available Python-based serotyping tools due to its smooth installation on UNIX platforms such as Linux and macOS. The tool requires specific Perl modules, including GD and GD::SVG, as well as BLAST+, BLAT, and Prodigal. For user convenience, Pneumo-Typer is also available as a Bioconda package with preinstalled dependencies. As shown in Fig. 1, the workflow of Pneumo-Typer is divided into three main steps:

Step 1: Input data processing

Genomes provided in GenBank or FASTA formats undergo systematic preparation. For each genome, gene nucleotide sequences (A) and TFT-formatted feature tables (B) - seven-column tab-delimited files detailing genomic features - are initially extracted [32].

Genomes lacking existing annotations are processed using Prodigal for automated gene prediction [33]. To ensure methodological consistency across datasets, Prodigal may optionally be applied to all input genomes regardless of their initial annotation status. For subsequent cgMLST (core genome multilocus sequence typing) analysis, a critical preprocessing step involves the seamless concatenation of all genomic contigs into a unified pseudo-chromosome (C), effectively simulating complete genome architecture for downstream comparative analyses.

Step 2: Sequence typing

MLST/cgMLST database retrieval. For MLST, seven allele genes (*aroE*, *gdh*, *gki*, *recP*, *spi*, *xpt*, *ddl*) along with their related allelic profiles specific to pneumococcus were sourced from the PubMLST website (https://pubmlst.org/) [34]. In parallel, for cgMLST, sequences from a collection of 1,222 allele genes and their matching allelic profiles were acquired for the determination of the cgST.

ST and cgST determination. Inputs A (gene sequences) or C (pseudo-chromosome) served as queries for allele identification. The ST determination was executed through a systematic approach: (i) The genes were compared with the MLST database using BLASTN. (ii) Stringent criteria (E-value = $1e^{-5}$, identity = 100%, coverage = 100%) were applied on the BLASTN results to identify multilocus sequence gene distribution. (iii) Identified genes were unified using dashes, resulting in a profile like aroE_13-gdh_13-gki_41-recP_5-spi_13-xpt_1-ddl_31. Missing allele genes in the genome were represented as “X”. (iv) STs were determined by matching profiles against known ST allelic profiles. Genomes with an “X” got an unknown ST label, and those with novel profiles were tagged as novel STs. The cgST determination followed a similar blueprint but with nuances: (i) BLAT [35] was harnessed for precise mapping of cgMLST loci against genomes. (ii) The "-fastMap" parameter sped up the process, fusing 1,222 cgMLST sequences into one, but for sequences like SPNE00213.fas that exceeded 5,000 bp, “-fastMap” was not used. (iii) Missing allele genes in the genome get a unique marker, like SPNE00001_N.

Step 3: Serotyping and visualization

Capsule gene database construction. The capsule gene database was established by

collating GenBank files associated with capsule gene loci from reference strains, which covered a spectrum of 93 serotypes. The foundation of the database, comprising 90 serotypes, was extracted from the EMBL Nucleotide Sequence Database under accession numbers CR931632–CR931722 [36]. The dataset was enriched with two more serotypes 6C and 35D, sourced from NCBI under accession numbers JF911515 [37] and KY084476 [38], respectively. Serotype 6D's sequence was procured from the PneumoCaT database [18] and annotated leveraging Prokka v1.12 [39]. 6E and 23B1 were not included as independent serotypes in the capsule gene database since they emerged as novel CapT subtypes. For instance, the isolate carrying the “6E” capsule gene locus yielded the serotype 6B capsule [40, 41]. The nucleotide sequences of the 93 serotypes were extracted from the GenBank files by a custom Perl script. Subsequently, each capsule gene was tagged with its respective gene name and serotype, exemplified as *wzd*-SPC30. The genes underwent a two-step refinement: (i) exclusion of genes with incomplete sequences or encoding transposases. (ii) dereplication of genes sharing 100% nucleotide identity using a proprietary Perl script. Upon completion, the capsule gene database encompassed 1,290 genes spread across the 93 serotypes. The database included two principal files: a sequence file and a list file. The sequence file, formatted in fasta, delineated sequence IDs preceding their respective sequence data. Every sequence ID included a unique ID complemented by gene annotation, separated by a whitespace (e.g., SPC16F_0021 dTDP-D-glucose_4,6-dehydratase_RmlB). The list file, structured in a tab-delimited format, comprised two columns. The first column denoted the gene class and its abbreviation, distinctly separated by dual underscores. The second column hosted the unique ID for each gene, with a comma serving as a delimiter for multiple genes when required (e.g., wzy_wzy-SPC24B ['SPC24B_0019', 'SPC24B_0015']).

Serotype determination. Serotype determination was performed by querying Input A (gene sequences) against our established capsule gene database using BLASTN, with matches filtered by stringent thresholds (identity ≥70%, coverage ≥95%, E-value ≤1e−5). Genes aligning to multiple serotypes were counted across all relevant categories, and the predominant serotype was assigned based on the highest cumulative gene matches

(called Serotype Designation). Ambiguities were resolved using serotype-specific markers from a modified Capsular Typing Variant Database (CTVD) (called Serotype Refinement) (Table S1).

CapT nomenclature and visualization. We established a CapT nomenclature system. For example, the CapT of strain GCA_001457635.1 was designated as SPC01/15/93.33%, where “SPC01” indicates serotype 1, “15” represents the number of matched capsule genes, and “93.33%” signifies that 14 out of 15 matched capsule genes belong to serotype 1. For strains that could not be distinguished using these three criteria, numerical suffixes were added for further differentiation (e.g., SPC04/16/81.25%-1 and SPC04/16/81.25%-2), ensuring that each CapT corresponds to a unique capsule gene content. For CapT visualization, capsule gene distribution and operon organization were mapped using Inputs B (TFT files) and D (capsule gene content). Genes affected by frameshift mutations, nonsense mutations, or terminator codon mutations were not displayed in any CapT visualization (Table S2). A custom-modified version of Gcluster [32] generated two-tiered heatmaps (gene- and class-level resolution) from tab-delimited positional data, while TFT-derived annotations enabled operon structure visualization. CapT, Serotype and ST/cgST metadata were algorithmically superimposed onto these graphical outputs.

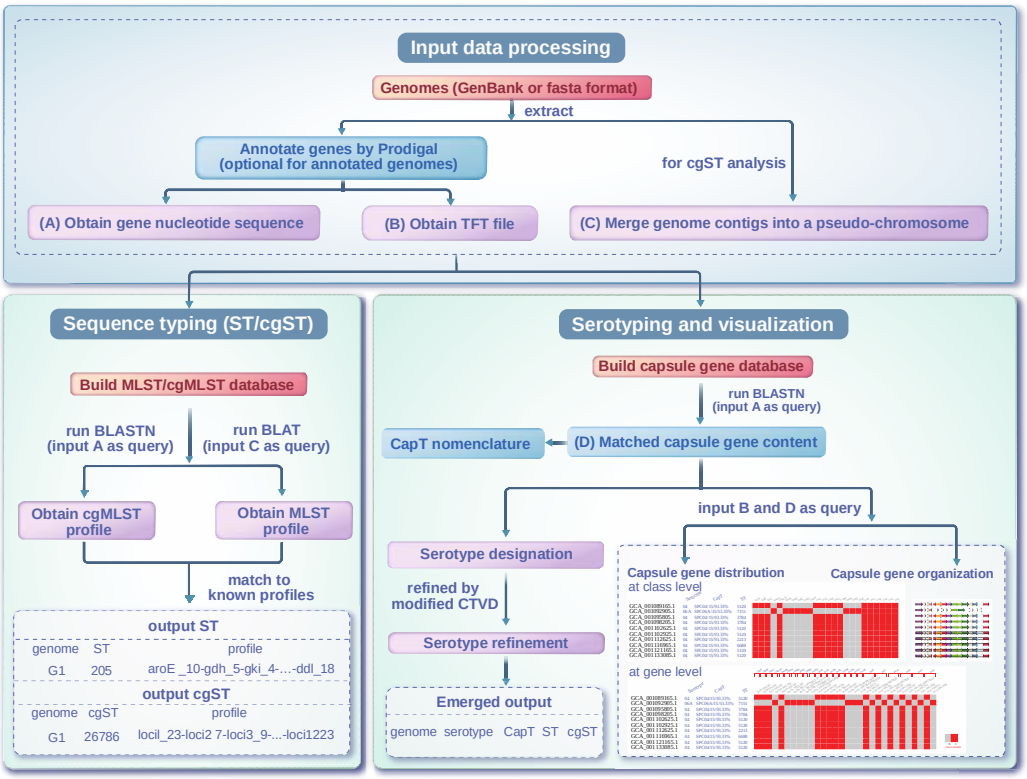

Fig. 1 The workflow of Pneumo-Typer.

Data acquisition, processing and analysis

Data acquisition and processing

To test the performance of Pneumo-Typer, all pneumococcal genomes from the NCBI database were downloaded as of Oct 5, 2022. These genomes, procured in GenBank format, were extracted using a custom Perl script aligned with a CSV file input. This CSV file was readily available from the NCBI genome database website (https://www.ncbi.nlm.nih.gov/genome/browse/) by searching with the term "*Streptococcus pneumoniae*". Furthermore, the serotype information was retrieved from the GenBank files through our proprietary Perl program. To ensure the quality and accuracy of the genomes, two diagnostic tools: CheckM [42] and QUAST [43] were employed. CheckM evaluated each genome's completeness and potential contamination based on lineage-specific sets of single-copy genes. Genomes boasting near-perfection (≥90% completeness) and negligible contamination (≤5% contamination) [44] were selected. QUAST offered metrics on assembly quality, and the genomes were sieved by stringent criteria, including contig size ≤ 500 bp and N50 ≥ 40 kb.

Data analysis

As shown in Fig. 2, after filtering process, 21,260 genomes were obtained, and used to assess the performance of Pneumo-Typer in ST and cgST determination. Among these, 17,667 genomes with recorded serotypes (covering 82 serotypes) were selected and combined with 16 genomes (covering 11 serotypes) downloaded from an earlier investigation [18] and assembled via MetaHIT [45], yielding a total of 17,683 genomes (covering 93 serotypes) for evaluating Pneumo-Typer's performance in serotype determination. Among these, 17,407 genomes that were accurately serotyped by Pneumo-Typer were used to analyze the relationships between serotypes, STs, and CapTs. Among these, 14,716 genomes within vaccine-covered serotypes were selected to evaluate Pneumo-Typer's performance in CapT analysis.

The analysis was performed on an Ubuntu 20.04 server equipped with four Intel Xeon Platinum 8260L 24-core processors and 512 GB RAM, harnessing 192 CPU threads. Parameter -p was set to “T”, prompting Prodigal to annotate all genomes under examination during Pneumo-Typer operation.

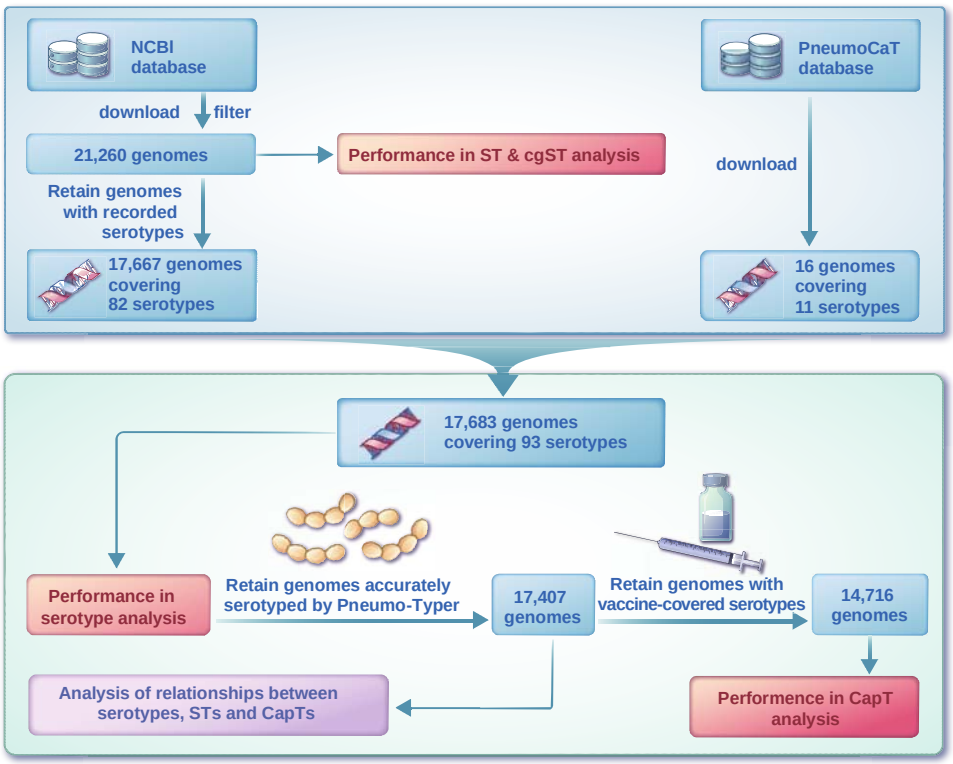

Fig. 2 Data acquisition, processing, and analysis workflow for this study.

Clonal Complexes analysis

The concept of MLST Clonal Complexes (CC) refers to STs that have differences in a single locus variant (SLV) based on the eBURST algorithm[46]. The analysis was performed using the PHYLOViZ 2.0 software[47].

ST or CapT diversity calculation

In this study, Shannon's entropy was employed to assess the diversity of STs or CapTs within individual serotypes. Shannon's entropy is equal to:

$$S(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

Where S(X) represents the Shannon's entropy of the variable X (either ST or CapT), $p(x_i)$ denotes the probability of X taking the *i*-th value x_i, and *n* is the total number of possible values for X.

Statistical analysis

Spearman's correlation was used to analyze the relationship between two sets of variables. A correlation coefficient $|R| \geq 0.8$ indicates a strong correlation, while $0.3 \leq |R| < 0.5$ suggests a weak correlation, and $|R| < 0.3$ denotes little to no correlation. To assess statistical differences between the two groups, an independent samples *t*-test was performed, with significance indicated by * $P < 0.05$.

Results

Key features of Pneumo-Typer

Pneumo-Typer has four key features:

Simple Input. The only mandatory input for Pneumo-Typer is a directory containing genomes in either GenBank or fasta format, or a combination of both, ensuring straightforward acquisition of input data.

Batch Processing. Pneumo-Typer efficiently handles the serotype, ST, cgST (optional), and CapT of a large array of pneumococcal genomes simultaneously using multiple threads. As a benchmark, it processed 21,260 genomes in approximately 53 minutes under default settings, with all genomes annotated using Prodigal and subjected to serotype and ST analysis, utilizing 190 threads on an Ubuntu 20.04 server powered by

four Intel Xeon Platinum 8260L processors.

Innovative Serotyping. Pneumo-Typer employs a novel two-step sequence alignment approach for serotyping. The first step, termed Serotype Designation, involves BLASTing genomes against a custom capsule gene database (covering 93 dominant serotypes). Serotype is assigned based on the serotype(s) with the highest proportion of matched genes. The second step, known as Serotype Refinement, further refines the serotype using serotype-specific markers from a modified CTVD (Table S1). If the serotypes predicted in Steps 1 and 2 do not belong to the same serogroup, the serotype from Step 1 is retained to avoid errors due to mosaic sequences.

Advanced CapT Visualization. Pneumo-Typer generates SVG-formatted heatmaps that display the capsule gene distribution at the gene (heatmap_gene.svg) or class (heatmap_class.svg) level. Additionally, it creates an SVG-formatted map that graphically represents the capsule operon organization (CPS_operon.svg), with distinct homologous genes depicted in varied colors. To enable superior comparison, there is also an option to map the capsule contexts onto a user-provided phylogenetic tree.

Performance of Pneumo-Typer in serotyping and sequence typing

At the serogroup level, Pneumo-Typer reached 99.84% accuracy (Table S3). At the serotype level, Pneumo-Typer achieved an accuracy of 84.98% (15,027 out of 17,683) after step Serotype Designation, which increased to 98.44% (17,407 out of 17,683) following step Serotype Refinement (Table 1). Herein, Pneumo-Typer achieved 100% prediction accuracy for 55 serotypes, 99%-100% accuracy for 10 serotypes, and 98%-99% accuracy for 7 serotypes. However, Pneumo-Typer had low prediction accuracy for some serotypes, such as 12A (0%), 25A (0%), 32A (0%), 44 (0%), 9A (20%), and 35C (31.48%). For comparison, we tested PneumoKITy and PfaSTer on the same dataset. PneumoKITy achieved an accuracy of 69.01% (12,203 out of 17,683), and PfaSTer achieved 90.29% (15,966 out of 17,683) (Table S4).

Table 1 The performance of Pneumo-Typer in serotyping

No.	Recorded serotype (*S*)	Serotype Designation (*D*)	Serotype Refinement (*R*)	Number	Percentage (%)	*D* = *S*	*R* = *S*

1	A	A	A	14646	82.83	yes	yes
2	A	B	A	2410	13.63	no	yes
3	A/B	A	A	272	1.54	yes	yes
4	A/B	A/B	A	79	0.45	yes	yes
7	A	B	B	211	1.19	no	no
5	A	A	B	30	0.17	yes	no
6	A	A/B	A/B	24	0.14	no	no
8	A	A/B	B	8	0.05	no	no
9	A	B/C	B/C	3	0.02	no	no
Total				**17683**	**100.00**	**Accuracy: 84.98%**	**Accuracy: 98.44%**

Notes: A, B, A/B, and B/C indicate possible serotype scenarios. Recorded serotypes refer to the serotypes documented in the NCBI Database. The definitions of Serotype Designation and Serotype Refinement are detailed in the "Overview of Pneumo-Typer" section.

The CapT visualization features of Pneumo-Typer allowed us to specifically examine the CapTs of the mispredicted strains by Pneumo-Typer. Herein, a strain recorded as serotype 1 but predicted as 16F had a capsule operon organization distinctly resembling 16F (Fig. 3A). Similarly, two strains recorded as serotype 6D but predicted as 6B showed a capsule operon organization more consistent with 6B, as 6B had a gap between the *wciN* and *wciO* genes, which 6D lacked (Fig. 3B). Additionally, four strains recorded as serotype 7B but predicted as 7C had a capsule operon organization that matched both 7B and 7C (Fig. 3C), yet their capsule gene distribution at the gene level aligned more closely with 7C (Fig. 3D). Excluding these misannotated serotypes would further improve the serotype prediction accuracy of Pneumo-Typer. After exclusion, the accuracy for serotype 1 would increase from 99.90% to 100%, for serotype 6D from 84.62% to 100%, and for serotype 7B from 86.67% to 100%. Therefore, Pneumo-Typer can not only predict serotypes but also verify the correctness of annotated serotypes through its CapT visualization features.

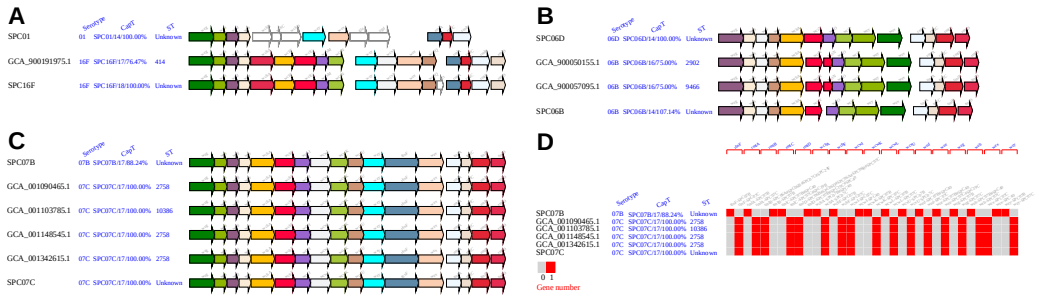

Fig. 3 The role of CapT visualization in serotyping. **A** CPS operon map of serotype 1 reference strain, GCA_900191975.1, and serotype 16F reference strain. **B** CPS operon map of serotype 6D reference strain, GCA_900050155.1, GCA_900057095.1, and serotype 6B reference strain. **C** CPS operon and **D** heatmap of serotype 7B reference strain, GCA_001090465.1, GCA_001103785.1, GCA_001148545.1, GCA_001342615.1, and serotype 7C reference strain.

Pneumo-Typer demonstrated exceptional performance in the rapid assignment of STs, processing 21,260 genomes in under 15 minutes. The assigned STs revealed the presence of 3,250 known STs and the emergence of 18 novel STs (Table S5). However, 311 genomes remained unassigned to any ST, likely due to incomplete ST locus identification in draft-status genomes. Among the identified STs, several exhibited high isolate counts, notably ST217 (459 isolates), ST156 (390 isolates), ST180 (387 isolates), ST199 (349 isolates), and ST191 (319 isolates). A substantial number of STs, 1,801, were assigned to only one isolate each (Table S6).

Besides, using Pneumo-Typer, 10,859 out of 21,260 genomes were categorized into 10,103 distinct cgSTs (Table S5). Additionally, the analysis revealed 10 novel cgST profiles. The most frequently occurring cgST was cgST12235, accounting for 141 isolates, followed by cgST1494, which accounted only for 7 isolates. A substantial number of cgSTs, 9,600, were assigned to only one isolate each. However, a significant number of genomes, 10,391, could not be linked to established cgSTs, indicating that known cgSTs could only be attributed to almost half of the genomes. This suggests a limitation in using cgST for molecular typing, given the extensive number of gene loci required, which may not be fully assembled in draft-status genomes.

Performance of Pneumo-Typer in identifying CapT variations

Serotypes represent phenotypic characteristics, whereas CapT denotes the genotype, with the latter determining the former. However, due to the limitations of molecular-based serotyping tools, the serotype results obtained may not always accurately reflect CapT variations, which can significantly impact the identification of potential capsule antigenic shifts. To address this issue, we employed CapT visualization features generated by Pneumo-Typer and evaluated its performance using three serotype 14 variants reported in a previous study firstly [14]. WGS data from these three serotype 14 variants (PMP1437, PMP1438, and PMP1514), along with the serotype 14 reference strain (SPC14), were analyzed using Pneumo-Typer, resulting in distinct types of CapT profiles (Fig. 4). Although Pneumo-Typer misclassified all three isolates as serotype 14 or 13|20, its CapT visualization features revealed critical genetic differences. Specifically, PMP1437, PMP1438, and PMP1514 all exhibited multiple gene deletions (Fig. 4B). Additionally, PMP1514 contained an inserted sequence that did not match any known genes in our capsule gene database (Fig. 4A). Moreover, PMP1437 appeared to be a hybrid, incorporating genes from both serotype 13|20 and serotype 14 (Fig. 4C). These findings demonstrate that Pneumo-Typer's CapT visualization features effectively identify CapT variations, providing valuable insights into genetic changes that may not be captured by conventional serotyping methods.

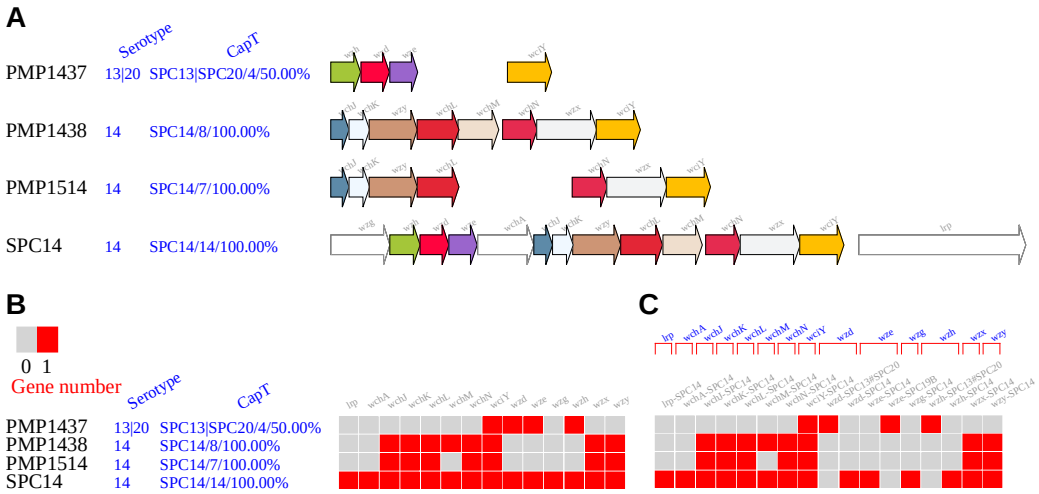

Fig. 4 CapT visualization of three serotype 14 variants by Pneumo-Typer. **A** Capsule operon organization. **B** Capsule gene distribution at class level. **C** Capsule gene distribution at the

gene level. Serotype 14 variants: PMP1437, PMP1438, and PMP1514; Serotype 14 reference strain: SPC14. Any genomic region that does not match the capsule gene database appears as a gap in the Capsule operon organization diagram.

In the context of vaccine design, monitoring capsule antigenic shifts in vaccine-covered serotypes is crucial, as these shifts directly impact vaccine efficacy. To address this issue, we analyzed the CapT profiles of serotypes included in PPSV23, PCV13, PCV15, PCV20, PCV21, and PCV24 [48] to identify CapT variations, including gene mutations and horizontal gene transfer events (Table 2). Notably, several CapT variations with high prevalence warrant further attention (Fig. 5). For serotype 14, mutations were identified in the *wciY* gene, which encodes a putative glycerol phosphotransferase, in 1,232 out of 1,412 strains (87.1%). However, previous studies have shown that the disruption of *wciY* does not impact capsular production [49]. The *rmlD* gene, which encodes dTDP-4-dehydrorhamnose reductase, was mutated in 648 out of 1,269 serotype 19A strains (51.1%) and 75 out of 248 serotype 23B strains (30.2%). Given that *rmlD* is essential for L-rhamnose biosynthesis in the capsule, its disruption is expected to influence capsular structure. For serotype 11A, mutations were detected in the *wcwC* gene, encoding an O-acetyltransferase, in 63 out of 402 strains (15.7%). In serotype 20, mutations in the *whaF* gene were identified in 91 out of 111 strains (82.0%). Notably, the *whaF* mutant has been designated as serotype 20A, whereas mutations in *wciG* (detected in 4 out of 111 strains) have led to the classification of serotype 20C. These variants are distinct from serotype 20B (which corresponds to serotype 20 in this study), with 20A and 20B included in PCV21 and PCV24, respectively [50, 51]. Additionally, in serotype 33F, an insertion of the *wcyO* gene, encoding a putative acetyltransferase from serotype 33C, was observed in 23 out of 145 strains (15.9%). Similarly, serotype 17F exhibited an insertion of the *glf* gene, encoding UDP-galactopyranose mutase, likely acquired from serotype 33B, in 49 out of 154 strains (31.8%). Given the role of *glf* in the biosynthesis of UDP-galactofuranose (Gal*f*) [52], the impact of this insertion on the capsule structure of serotype 17F remains unclear. These findings highlight the potential risk that vaccines based on standard serotype strains may lose efficacy or fail to protect against emerging

CapT variants. Integrating these CapT variants into future vaccine formulations may serve as a viable strategy to maintain or enhance vaccine coverage. Moreover, Pneumo-Typer proved to be an effective tool for serotype subtyping, as demonstrated in this study by its ability to differentiate between the serotype 20 subtypes 20A, 20B, and 20C.

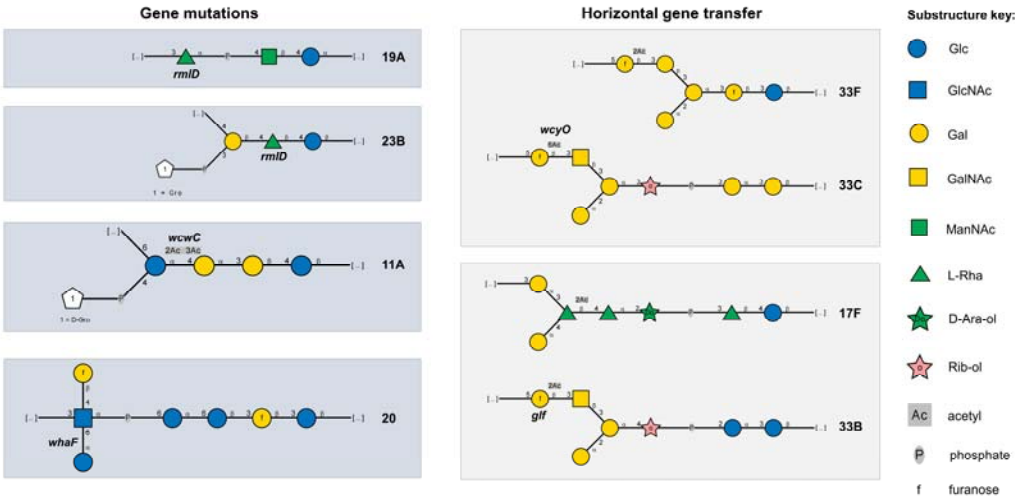

Fig. 5 Capsule repeating unit of several serotypes with CapT variations in this study.

Table 2 Key CapT variations in vaccine-targeted serotypes

Serotype	Number	PPSV23	PCV13	PCV15	PCV20	PCV21	PCV24	Gene mutation/number (≥3)	Horizontal gene transfer/number
14	1412	▪	▪	▪	▪		▪	*wciY*/1232; *wchA*/11; *wchN*/7; *wchM*/5; *wzx*/5; *wzy*/4; *wzg*/4; *wchK*/3	N/A
19F	1381	▪	▪	▪	▪		▪	*wzg*/15; *rmlD*/3; *wze*/3; *wchO*/3	N/A
19A	1269	▪	▪	▪	▪	▪	▪	*rmlD*/648; *wchQ*/4; *wzy*/3; *wchO*/3	*wcwD*/1
6B	1231	▪	▪	▪	▪		▪	*wzy*/12; *rmlD*/5; *wzx*/4	N/A
23F	1137	▪	▪	▪	▪		▪	*wzy*/5; *wzx*/4; *wzd*/4	*wcyO-glf*/1
6A	1018		▪	▪	▪	▪	▪	*wzy*/12; *rmlD*/6	N/A
1	965	▪	▪	▪	▪		▪	*wzx*/4	*rmlD*/8; *wchO*/1
3	544	▪	▪	▪	▪	▪	▪	*wzd*/42	*wzh*/4
9V	462	▪	▪	▪	▪		▪	*wcjD*/6, *wzg*/4; *wcjA*/3; *wzx*/3	N/A
11A	402	▪			▪	▪	▪	*wcwC*/63; *wcjE*/24; *wzx*/9	N/A

5	400	▪	▪	▪	▪		▪	N/A	N/A
15B	397	▪			▪	▪	▪	N/A	N/A
16F	391					▪		N/A	N/A
12F	390	▪			▪	▪	▪	*wzx*/3	N/A
15A	343					▪		N/A	N/A
35B	301					▪		*wzg*/3	N/A
4	298	▪	▪	▪	▪		▪	*wciJ*/5; *wciI*/3	N/A
7F	287	▪	▪	▪	▪	▪	▪	N/A	*wchQ*/2
18C	284	▪	▪	▪	▪		▪	*wchA*/6; *wzg*/5	*wcyO-glf*/1
23B	248					▪		*rmlD*/75; *wchW*/3	N/A
22F	235	▪		▪	▪	▪	▪	N/A	*wchP*/1
10A	233	▪			▪	▪	▪	*wcrG*/3	N/A
23A	225					▪		N/A	N/A
8	210	▪			▪	▪	▪	*wchA*/8	N/A
17F	154	▪				▪	▪	*wchA*/3	*glf*/49; *wcjE*/1
33F	145	▪		▪	▪	▪	▪	N/A	*wcyO*/23
20(B)	111	▪				▪(20A)	▪(20B)	*whaF*/91 (20A); *wciG*/4 (20C)	N/A
9N	110	▪				▪	▪	N/A	N/A
31	82					▪		*wzy*/4	N/A
2	38	▪					▪	N/A	N/A
24F	13					▪		N/A	N/A
Coverage against total 17,407 (%)		**69.48**	**61.40**	**63.58**	**72.96**	**40.83**	**75.33**		

Bidirectional serotype-ST associations

The performance of Pneumo-Typer in both serotyping and sequence typing enables a systematic study elucidating the relationships between serotypes and STs. The relationships between serotypes and STs (Table S7) can be analyzed in two ways: the STs associated with each serotype and the serotypes associated with each ST.

First, we analyzed the STs associated with each serotype (Table S8) and found that the vast majority—78 out of 90 serotypes—were associated with multiple STs (Fig. 6A). A subset of serotypes, specifically 19F (≥284 STs), 6B (≥257 STs), 6A (≥238 STs), 19A (≥216 STs), and 23F (≥182 STs), exhibited a particularly high number of coexisting STs. In contrast, a small group of serotypes, including 7A, 10C, 11C, 11D, 11F, 16A, 19C, 24B, 42, 43, 47A, and 47F were observed to have only one unique ST association. However, the

ST richness of each serotype heavily depended on its sample size (Fig. 6B). To evaluate the ST diversity of each serotype independently of sample size, we excluded serotypes with fewer than 100 isolates to avoid potential bias and used Shannon's entropy as a measure of diversity (higher Shannon's entropy indicates more STs with a more even distribution). The results showed that serotypes 6A, 6B, 19F, 16F, and 19A had relatively high ST diversity, while serotypes 5, 38, 7F, 8, and 12F had relatively low ST diversity (Fig. 6C), with ST diversity showing low correlation with sample size (Fig. 6D). These results indicate significant differences in ST diversity among individual serotypes.

Conversely, each of the majority—2,618 out of 2,815 STs was linked to a unique serotype. For instance, all isolates of certain STs, such as the 458 isolates of ST217 and the 110 isolates of ST3081, were exclusively associated with serotype 1. Therefore, an isolate's ST can often help determine its serotype. We selected the STs of the misannotated isolates (Fig. 3) from and found that, except for ST2377 (not found in Table S9), the serotypes associated with these STs (Table S9) did not match the recorded serotypes, but they aligned with the predictions made by Pneumo-Typer (Table 3). It is important to acknowledge that many STs were represented by only a limited number of isolates, indicating the need for more comprehensive sampling to confirm these findings. Besides, a small proportion of STs (197 out of 2,815) were associated with multiple serotypes. Notably, ST156, ST162, ST199, ST172, and ST193 each were related to 10, 9, 8, 7, and 7 serotypes, respectively. Additionally, there were 167 STs associated with two serotypes, 33 with three, 15 with four, and 3 each with five and six serotypes (Fig. 6E). Given that STs represent highly conserved housekeeping genes with generally low mutation rates, the phenomenon of a single ST corresponding to multiple serotypes is likely due to capsular switch events [53].

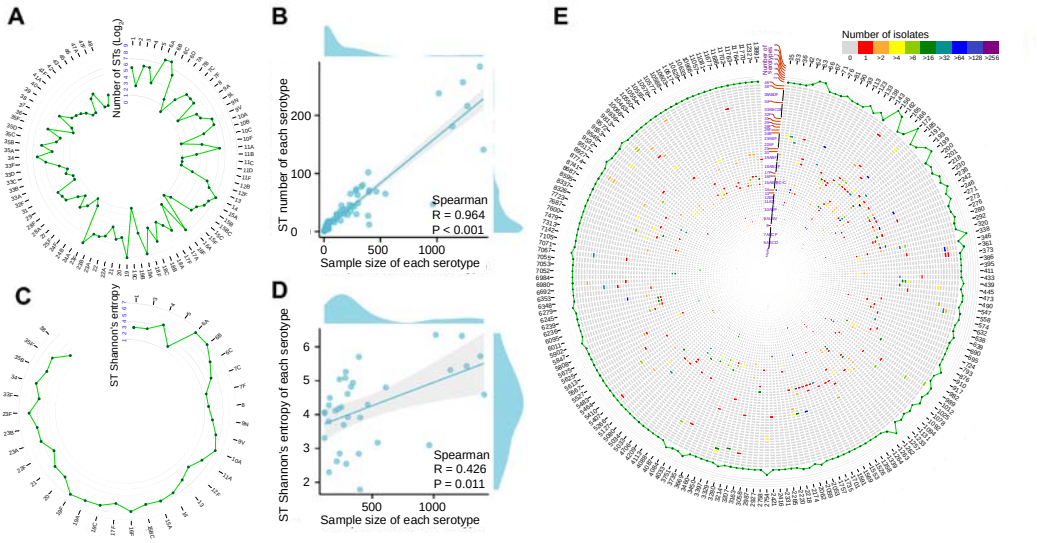

Fig. 6 Serotype and ST associations. **A** Number of ST types associated with each serotype. **B** Correlation between sample size and ST type number of each serotype. **C** Shannon's entropy of STs associated with each serotype. **D** Correlation between sample size and ST Shannon's entropy of each serotype. **E** Number of serotypes (two or more) associated with each ST. Progressing from the innermost to the outermost part, rings 1 through 70 present heatmaps depicting the count of isolates per serotype for each ST, while ring 71 illustrates the total number of serotypes corresponding to each ST.

Table 3 The role of STs in serotype determination

Strain	Recorded serotype	Serotype by Pneumo-Typer	ST	ST: Serotype|Number
GCA_900191975.1	1	16F	ST414	ST414: 16F|3
GCA_900050155.1	6D	6B	ST2902	ST2902: 6B|9
GCA_900057095.1	6D	6B	ST9466	ST9466: 6B|2
GCA_001090465.1; GCA_001148545.1; GCA_001342615.1	7B	7C	ST2758	ST2758: 7C|2; 15B|15C|1
GCA_001103785.1	7B	7C	ST2377	ST2377: N/A

Excessive CapT diversity across serotypes

We analyzed CapT diversity across serotypes (Table S10) and found that the number of CapT types varied among serotypes (Fig. 7A). However, as expected, the number of CapT types was highly correlated with sample size (Fig. 7B), indicating that raw counts alone do not accurately reflect true CapT diversity. To address this, we excluded

serotypes with fewer than 100 samples and applied Shannon's entropy to assess CapT diversity. The results showed that serotypes 6A, 7F, 23F, 6C, and 23B had relatively high CapT diversity (Fig. 7C), with no correlation to sample size (Fig. 7D). These findings highlight significant differences in CapT diversity among serotypes, suggesting that different serotypes undergo capsule genetic modifications at varying frequencies. Furthermore, CapT diversity and ST diversity reflect changes in distinct genomic regions, showing only a weak correlation (Fig. 7E). This suggests that CapT and ST evolution occur independently rather than in a synchronized manner.

Additionally, serotypes belonging to serogroups with multiple serotypes (e.g., 6A, 7F, and 23F) exhibited higher CapT diversity compared to those in serogroups that contain a single serotype (e.g., 5, 1, and 3) (Fig. 7F). This may be attributed to certain serotypes undergoing more frequent genetic changes in their capsules, with dominant CapT variants gradually emerging as new subtypes within the serotype or serogroup. Notably, serotypes 20 (ranked 6th) and 13 (ranked 7th) also displayed relatively high CapT diversity. As previously discussed, serotype 20 has been further classified into subtypes 20A, 20B, and 20C (Table 2). Regarding serotype 13, we found that in 227 out of 267 strains, the *whaG* gene, which encodes a glycosyltransferase, was functional despite being recorded as a pseudogene in standard databases. However, to date, serotype 13 has not been formally subdivided into distinct subtypes.

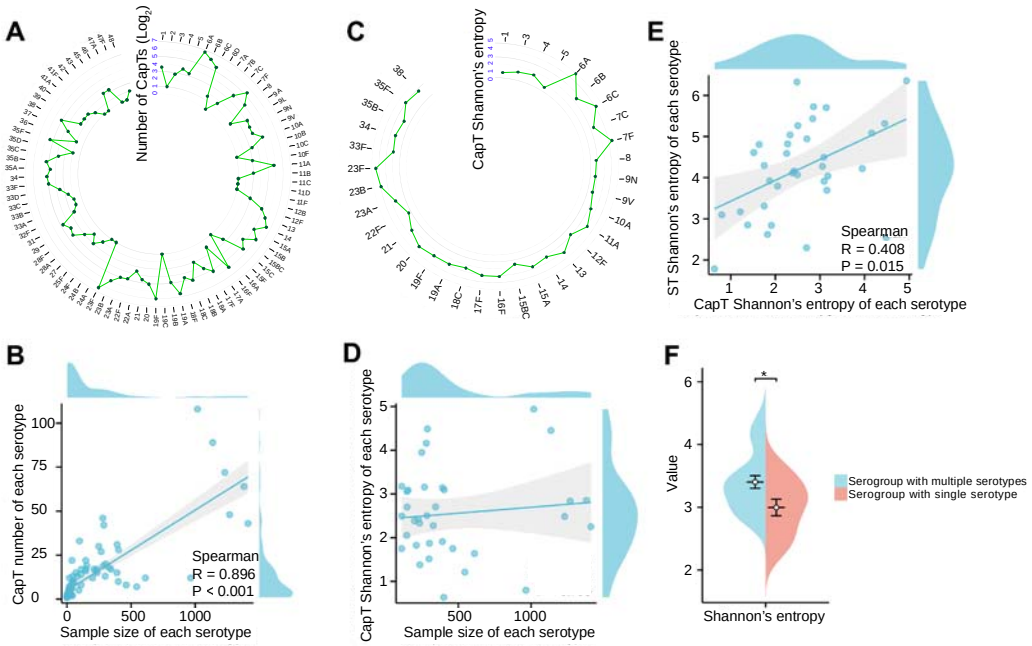

Fig. 7 CapT diversity across serotypes. **A** Number of CapT types associated with each serotype. **B** Correlation between sample size and CapT type number of each serotype. **C** Shannon's entropy of CapTs associated with each serotype. **D** Correlation between sample size and CapT Shannon's entropy of each serotype. **E** Correlation between CapT Shannon's entropy and ST Shannon's entropy of each serotype. **F** Difference in CapT Shannon's entropy between serogroups with multiple serotypes and serogroups with single serotype.

CapT-ST relationships and implications for vaccine design

As mentioned earlier, STs and CapTs evolve independently. Nevertheless, we aimed to identify potential patterns in the relationships between CapTs and STs (Table S11). To visualize these relationships, we used a Sankey diagram to illustrate the associations between CapTs and STs for PCV20-covered serotypes (see Supplementary files). Taking serotype 11A as an example, we found that STs within CC200 and CC580 corresponded to SPC11F/XX/XX, STs within CC33, CC503, CC1079, and NA (unknown) corresponded to SPC11D/XX/XX, while STs in other CCs corresponded to SPC11A/XX/XX (Fig. 8A). This suggests that STs with closely related gene sequences within the same clonal complex (CC) tend to be associated with similar CapTs, indicating the presence of vertical inheritance constraints within specific STs or CCs.

Due to these vertical inheritance constraints, when certain STs become prevalent, it is not only essential to monitor their corresponding serotypes but also to consider their associated CapTs, as this has implications for vaccine design. To explore this further, we selected several CapT variations from Table 2 and mapped the relationships between their CapTs and STs (Fig. 8B). Notably, ST303 (serotype 1-*rmlD*), ST700 (serotype 3-*wzd*), ST673 (serotype 33F-*wcyO*), ST99 (serotype 11A-*wcwC*), and ST62 (serotype 11A-*wcjE*) stood out as significant. When these STs are prevalent, it becomes crucial to consider CapT variations in vaccine development. For instance, serotype 3-ST700 exhibits a capsule variant phenotype and has demonstrated enhanced vaccine escape potential following PCV13 introduction [54]. In our dataset, all ST700 isolates lacked the *wzd* gene. Additionally, serotype 11A-ST99 was prevalent in South Korea between 2004 and 2013, accounting for 27.8% of all serotype 11A isolates [55]. In our study, we found that all ST99 isolates carried mutations in the *wcwC* gene. Furthermore, during 2014/15–2018/19 in England, the major circulating STs of serotype 33F were ST717, ST100, and ST673 [56]. In our study, all ST673 isolates contained an insertion of the *wcyO* gene. Therefore, to effectively target infections caused by ST700, ST99, or ST673, it is crucial to incorporate the corresponding CapT variants into vaccine strain selection during vaccine development.

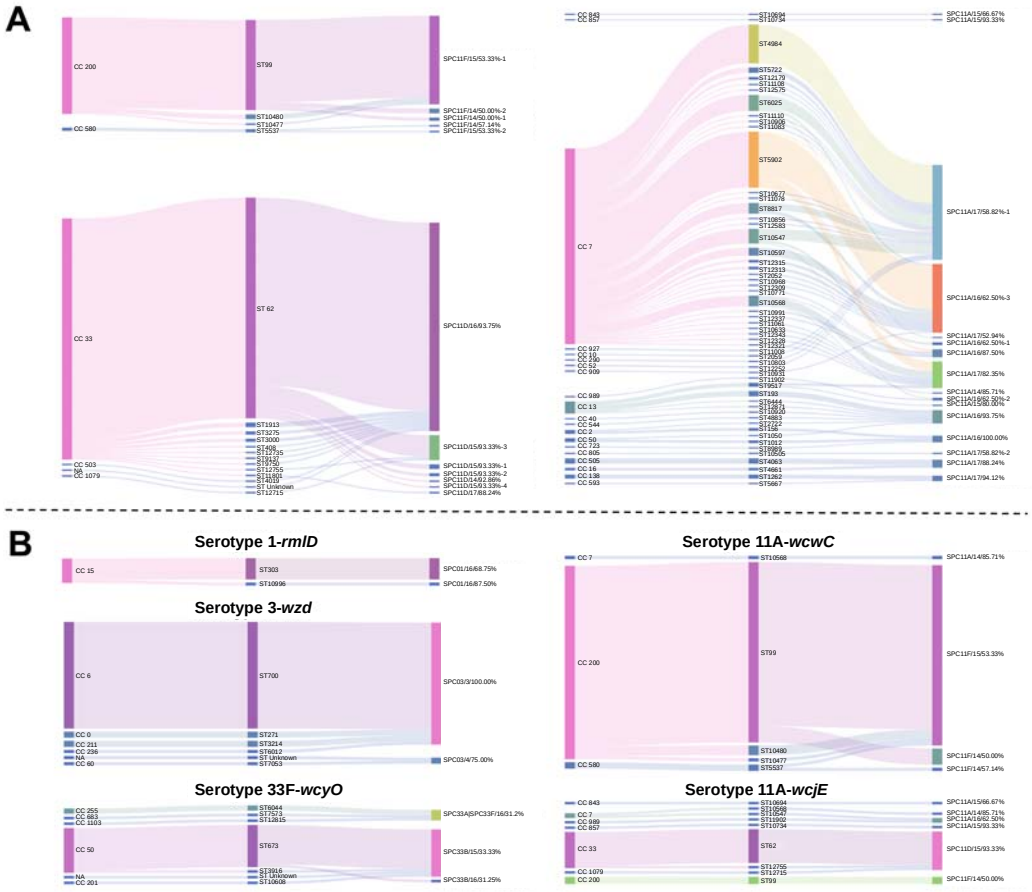

Fig. 8 Sankey diagram illustrates the relationship between CapTs and STs for **A** serotype 11A and **B** several serotypes with CapT variations.

Discussion

For a long time, epidemiological surveillance of *S. pneumoniae* has primarily focused on clinical outcomes, age distribution, geographical variations, colonization or infection sites, and vaccination status. Further studies have commonly involved analyzing antibiotic resistance patterns and the distribution of serotypes and STs (or cgST, GPSC) among isolates. In recent years, the increasing occurrence of vaccine escape driven by GapT variations has underscored the need for reliable CapT detection. In response, Pneumo-Typer was developed to enable the simultaneous identification of serotype, ST (with optional cgST), and CapT in *S. pneumoniae*.

CapT serves as the genetic basis of serotype determination, while serotypes represent these genetic characteristics at the antigenic level. However, due to significant CapT

diversity within individual serotypes, it is no longer feasible to directly infer CapT from serotype alone. CapT is particularly critical for vaccine design, especially in the development of glycoconjugate vaccines, as it provides key insights into capsular antigenic structures. Before the advent of Pneumo-Typer, CapT analysis required a labor-intensive multi-step bioinformatics workflow involving multiple software tools. Pneumo-Typer streamlines this process by integrating visualization features that illustrate capsule gene distribution and operon organization, offering an intuitive understanding of CapT variations.

In the case illustrated in Fig. 4, three isolates were identified as serotype 14 by both SeroBA and SeroCall. While Pneumo-Typer classified two of these isolates as serotype 14 (with another classified as serotype 13/20), its CapT visualization feature revealed substantial differences in the *cps* loci of these isolates compared to the standard serotype 14 strain. Additionally, Pneumo-Typer's operon organization map identified potential insertions of unknown sequences (depicted as gaps, as seen in PMP1514), prompting further investigation into the origins of these sequences. The inclusion of CapT data thus addresses the limitations of conventional molecular methods for serotype prediction. By leveraging Pneumo-Typer's CapT analysis capabilities, we identified multiple CapT variations within vaccine-covered serotypes, providing valuable insights for future vaccine design.

Beyond its strengths in CapT detection, Pneumo-Typer also demonstrates robust performance in traditional serotyping and sequence typing. For serotyping, it employs a novel two-step sequence alignment strategy. Unlike PneumoCaT, which relies on full-genome alignment and struggles to determine serotypes for genomes with less than 90% capsular operon sequence coverage, Pneumo-Typer requires only a match to a single capsule gene from its database to assign a serotype. This makes Pneumo-Typer more inclusive, accommodating genomes with incomplete *cps* loci or varying degrees of CapT variation. Despite its inclusivity, Pneumo-Typer maintains high accuracy (98.44%), outperforming other serotyping tools such as PneumoKITy (69.01%) and PfaSTer (90.29%). It is important to note that since SeroBA and SeroCall only support raw sequencing reads, whereas Pneumo-Typer, PneumoKITy, and PfaSTer operate on

assembled WGS data, a direct comparison with SeroBA and SeroCall was not conducted in this study. However, we plan to incorporate a raw-read analysis mode in future versions of Pneumo-Typer.

Furthermore, this study represents the first application of CapT and ST data in *S. pneumoniae* serotype determination. Annotation discrepancies that lead to serotype mispredictions can be corrected using Pneumo-Typer's CapT visualization features. Additionally, ST data provide valuable support for serotype determination, as each ST is predominantly associated with a specific serotype.

The integration of serotype, ST, and CapT data in Pneumo-Typer has enabled a comprehensive analysis of their relationships, revealing several key insights. First, a single serotype can correspond to multiple STs, and the degree of ST diversity varies across serotypes. This observation aligns with prior findings that individual serotypes often exhibit considerable genetic diversity among isolates [57]. Conversely, a single ST is typically associated with only one serotype, reflecting the tendency for capsular genes to spread horizontally over long evolutionary timescales, while clonal expansion with limited serotype switching occurs in the short term [58]. Second, different serotypes display varying degrees of CapT diversity, with multi-serotype serogroups showing higher CapT diversity compared to single-serotype serogroups. This finding is consistent with previous research indicating that multi-serotype serogroups undergo frequent within-group serotype switching events [59]. Lastly, our analysis highlights the relative evolutionary independence between STs and CapTs. This can be explained by the distinct selective pressures acting on core genomes (represented by STs) and non-core genomes (represented by CapTs): STs evolve conservatively under stabilizing selection, whereas CapT variations occur adaptively in response to external pressures such as vaccination or immune evasion [60-63].

Conclusions

In summary, Pneumo-Typer provides a rapid, accurate, and integrated approach for analyzing serotypes, STs, and CapTs, significantly contributing to epidemiological surveillance and vaccine design for *S. pneumoniae*. Its user-friendly design makes it an excellent tool for researchers. Moving forward, we plan to develop an online version of

Pneumo-Typer to further enhance accessibility and usability, particularly for researchers without bioinformatics expertise.

Availability and requirements

Project name: Pneumo-Typer

Project home page: https://www.microbialgenomic.cn/Pneumo-Typer.html or https://github.com/Xiangyang1984/Pneumo-Typer

Operating system(s): Platform independent

Programming language: Perl

Other requirements: None

License: GNU General Public License v3.0 or any later version (GPL-3.0-or-later)

Any restrictions to use by non-academics: None

Abbreviations

CapT: Capsule gentoype; cgST: core genome sequence type; CTVD: Capsular Typing Variant Database; GPSCs: Pneumococcal sequence clusters; IPD: Invasive pneumococcal disease; PCV: Pneumococcal conjugate vaccine; ST: Sequence type; WGS: Whole genome sequencing

Declarations

Ethics approval and consent to participate

This study only utilizes data that has been previously published or publicly stored in NCBI database.

Consent for publication

Not applicable

Availability of Data and materials

Pneumo-Typer is freely available at https://github.com/Xiangyang1984/Pneumo-Typer and https://www.microbialgenomic.cn/Pneumo-Typer.html. Pneumo-Typer can also be installed through conda: "conda install -c bioconda pneumo-typer". A detailed README is included. The entire sets of CapT visualization figures of 17,683 pneumococcal genomes and 93 standard serotypes, together with Sankey diagram for PCV20-covered serotypes, are available in Supplementary files. Additional data related to this paper may be requested from the authors.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Natural Science Foundation of Shandong Province (ZR2024QC359), the National Natural Science Foundation of China (32000647, 21966015, and 21507012), Science and Technology Innovation Development Project of Tai'an City (2023NS432), Medical and Health Science and Technology Project of Shandong Province (202402070770), the "1000 Level" Talent Training Project of Guizhou Province (QQCRC [2022]201703), the Basic Research Program of Qiandongnan Miao and Dong Autonomous Prefecture (QDNKHJC[2024]0007), the Quality Improvement Project of Kaili University (No. 24), the Specialized Fund for the Doctoral Development of Kaili University (BSFZ202210), the Foundation Research Project of Kaili University (YTH-XM2024022 and YTH-PT202404), and the Hangzhou Medical Health Science and Technology Project (A20220614).

Authors' contributions

ZM and BL conceived and supervised research. ZM and XL devised the data analysis strategies and directed the data analysis. XL implemented the software program while HZ and YZ performed the data analysis and prepared the figures and tables. ZM, ZY, XW, GZ, XZ and YH contributed to the writing of the manuscript. All of the authors have read and approved the final manuscript.

Acknowledgements

The authors thank Dr. Yingying Hao from the Shandong Provincial Hospital Affiliated to Shandong First Medical University for her needs and feedback regarding the use of the Pneumo-Typer software.

References

1. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. Nat Rev Microbiol. 2018;16(6):355-67; doi:10.1038/s41579-018-0001-8.
2. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Infect Dis. 2018;18(11):1191-210; doi:10.1016/s1473-3099(18)30310-4.
3. Kwun MJ, Ion AV, Cheng HC, D'Aeth JC, Dougan S, Oggioni MR et al. Post-vaccine

epidemiology of serotype 3 pneumococci identifies transformation inhibition through prophage-driven alteration of a non-coding RNA. Genome Med. 2022;14(1):144; doi:10.1186/s13073-022-01147-2.

4. Ganaie FA, Beall BW, Yu J, van der Linden M, McGee L, Satzke C et al. Update on the evolving landscape of pneumococcal capsule types: new discoveries and way forward. Clin Microbiol Rev. 2025;38(1):e0017524; doi:10.1128/cmr.00175-24.
5. Wahl B, O'Brien KL, Greenbaum A, Majumder A, Liu L, Chu Y et al. Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000-15. Lancet Glob Health. 2018;6(7):e744-e57; doi:10.1016/s2214-109x(18)30247-x.
6. Lo SW, Mellor K, Cohen R, Alonso AR, Belman S, Kumar N et al. Emergence of a multidrug-resistant and virulent Streptococcus pneumoniae lineage mediates serotype replacement after PCV13: an international whole-genome sequencing study. Lancet Microbe. 2022;3(10):e735-e43; doi:10.1016/s2666-5247(22)00158-6.
7. Lewnard JA, Hanage WP. Making sense of differences in pneumococcal serotype replacement. Lancet Infect Dis. 2019;19(6):e213-e20; doi:10.1016/s1473-3099(18)30660-1.
8. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after pneumococcal vaccination. Lancet. 2011;378(9807):1962-73; doi:10.1016/s0140-6736(10)62225-8.
9. Ramos B, Vadlamudi NK, Han C, Sadarangani M. Future immunisation strategies to prevent Streptococcus pneumoniae infections in children and adults. Lancet Infect Dis. 2025; doi:10.1016/s1473-3099(24)00740-0.
10. Ganaie F, Saad JS, McGee L, van Tonder AJ, Bentley SD, Lo SW et al. A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large cps Fragment from an Oral Streptococcus. mBio. 2020;11(3); doi:10.1128/mBio.00937-20.
11. Makarewicz O, Lucas M, Brandt C, Herrmann L, Albersmeier A, Rückert C et al. Whole Genome Sequencing of 39 Invasive Streptococcus pneumoniae Sequence Type 199 Isolates Revealed Switches from Serotype 19A to 15B. PLoS One. 2017;12(1):e0169370; doi:10.1371/journal.pone.0169370.
12. van Tonder AJ, Bray JE, Quirk SJ, Haraldsson G, Jolley KA, Maiden MCJ et al. Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. Microb Genom. 2016;2(10):000090; doi:10.1099/mgen.0.000090.
13. Ganaie FA, Saad JS, Lo SW, McGee L, Bentley SD, van Tonder AJ et al. Discovery and Characterization of Pneumococcal Serogroup 36 Capsule Subtypes, Serotypes 36A and 36B. J Clin Microbiol. 2023;61(4):e0002423; doi:10.1128/jcm.00024-23.
14. Manna S, Spry L, Wee-Hee A, Ortika BD, Boelsen LK, Batinovic S et al. Variants of Streptococcus pneumoniae Serotype 14 from Papua New Guinea with the Potential to Be Mistyped and Escape Vaccine-Induced Protection. Microbiol Spectr. 2022;10(4):e0152422; doi:10.1128/spectrum.01524-22.
15. Mauffrey F, Fournier É, Demczuk W, Martin I, Mulvey M, Martineau C et al. Comparison of sequential multiplex PCR, sequetyping and whole genome sequencing for serotyping of Streptococcus pneumoniae. PLoS One. 2017;12(12):e0189163; doi:10.1371/journal.pone.0189163.
16. Abdul Rahman NA, Mohd Desa MN, Masri SN, Taib NM, Sulaiman N, Hazman H et al. The

Molecular Approaches and Challenges of Streptococcus pneumoniae Serotyping for Epidemiological Surveillance in the Vaccine Era. Pol J Microbiol. 2023;72(2):103-15; doi:10.33073/pjm-2023-023.

17. Jauneikaite E, Tocheva AS, Jefferies JM, Gladstone RA, Faust SN, Christodoulides M et al. Current methods for capsular typing of Streptococcus pneumoniae. J Microbiol Methods. 2015;113:41-9; doi:10.1016/j.mimet.2015.03.006.
18. Kapatai G, Sheppard CL, Al-Shahib A, Litt DJ, Underwood AP, Harrison TG et al. Whole genome sequencing of Streptococcus pneumoniae: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. PeerJ. 2016;4:e2477; doi:10.7717/peerj.2477.
19. Knight JR, Dunne EM, Mulholland EK, Saha S, Satzke C, Tothpal A et al. Determining the serotype composition of mixed samples of pneumococcus using whole-genome sequencing. Microb Genom. 2021;7(1); doi:10.1099/mgen.0.000494.
20. Epping L, van Tonder AJ, Gladstone RA, The Global Pneumococcal Sequencing C, Bentley SD, Page AJ et al. SeroBA: rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. Microb Genom. 2018;4(7); doi:10.1099/mgen.0.000186.
21. Sheppard CL, Manna S, Groves N, Litt DJ, Amin-Chowdhury Z, Bertran M et al. PneumoKITy: A fast, flexible, specific, and sensitive tool for Streptococcus pneumoniae serotype screening and mixed serotype detection from genome sequence data. Microb Genom. 2022;8(12); doi:10.1099/mgen.0.000904.
22. Lee JT, Li X, Hyde C, Liberator PA, Hao L. PfaSTer: a machine learning-powered serotype caller for Streptococcus pneumoniae genomes. Microb Genom. 2023;9(6); doi:10.1099/mgen.0.001033.
23. Lepuschitz S, Weinmaier T, Mrazek K, Beisken S, Weinberger J, Posch AE. Analytical Performance Validation of Next-Generation Sequencing Based Clinical Microbiology Assays Using a K-mer Analysis Workflow. Front Microbiol. 2020;11:1883; doi:10.3389/fmicb.2020.01883.
24. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43(5):491-8; doi:10.1038/ng.806.
25. van Tonder AJ, Gladstone RA, Lo SW, Nahm MH, du Plessis M, Cornick J et al. Putative novel cps loci in a large global collection of pneumococci. Microb Genom. 2019;5(7); doi:10.1099/mgen.0.000274.
26. Enright MC, Spratt BG. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. Microbiology (Reading). 1998;144 (Pt 11):3049-60; doi:10.1099/00221287-144-11-3049.
27. Gladstone RA, Lo SW, Goater R, Yeats C, Taylor B, Hadfield J et al. Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates. Microb Genom. 2020;6(5); doi:10.1099/mgen.0.000357.
28. Eldholm V, Osnes MN, Bjørnstad ML, Straume D, Gladstone RA. A genome-based survey of invasive pneumococci in Norway over four decades reveals lineage-specific responses to vaccination. Genome Med. 2024;16(1):123; doi:10.1186/s13073-024-01396-3.
29. Nakano S, Fujisawa T, Chang B, Ito Y, Akeda H, Fujita J et al. Whole-Genome Analysis-Based

Phylogeographic Investigation of Streptococcus pneumoniae Serotype 19A Sequence Type 320 Isolates in Japan. Antimicrob Agents Chemother. 2022;66(2):e0139521; doi:10.1128/aac.01395-21.

30. Calvo-Silveria S, González-Díaz A, Marimón JM, Cercenado E, Quesada MD, Casabella A et al. Resilience and emergence of pneumococcal serotypes and lineages in adults post-PCV13 in Spain: A multicentre study. J Infect Public Health. 2025;18(1):102619; doi:10.1016/j.jiph.2024.102619.
31. Scherer EM, Beall B, Metcalf B. Serotype-Switch Variant of Multidrug-Resistant Streptococcus pneumoniae Sequence Type 271. Emerg Infect Dis. 2021;27(6):1689-92; doi:10.3201/eid2706.203629.
32. Li X, Chen F, Chen Y. Gcluster: a simple-to-use tool for visualizing and comparing genome contexts for numerous genomes. Bioinformatics. 2020;36(12):3871-3; doi:10.1093/bioinformatics/btaa212.
33. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119; doi:10.1186/1471-2105-11-119.
34. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010;11:595; doi:10.1186/1471-2105-11-595.
35. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64; doi:10.1101/gr.229202.
36. Bentley SD, Aanensen DM, Mavroidi A, Saunders D, Rabbinowitsch E, Collins M et al. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. 2006;2(3):e31; doi:10.1371/journal.pgen.0020031.
37. Elberse K, Witteveen S, van der Heide H, van de Pol I, Schot C, van der Ende A et al. Sequence diversity within the capsular genes of Streptococcus pneumoniae serogroup 6 and 19. PLoS One. 2011;6(9):e25018; doi:10.1371/journal.pone.0025018.
38. Geno KA, Saad JS, Nahm MH. Discovery of Novel Pneumococcal Serotype 35D, a Natural WciG-Deficient Variant of Serotype 35B. J Clin Microbiol. 2017;55(5):1416-25; doi:10.1128/JCM.00054-17.
39. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9; doi:10.1093/bioinformatics/btu153.
40. Burton RL, Geno KA, Saad JS, Nahm MH. Pneumococcus with the "6E" cps Locus Produces Serotype 6B Capsular Polysaccharide. J Clin Microbiol. 2016;54(4):967-71; doi:10.1128/JCM.03194-15.
41. Kapatai G, Sheppard CL, Troxler LJ, Litt DJ, Furrer J, Hilty M et al. Pneumococcal 23B Molecular Subtype Identified Using Whole Genome Sequencing. Genome Biol Evol. 2017;9(8):2122-35; doi:10.1093/gbe/evx092.
42. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25(7):1043-55; doi:10.1101/gr.186072.114.
43. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5; doi:10.1093/bioinformatics/btt086.
44. Chen MY, Teng WK, Zhao L, Hu CX, Zhou YK, Han BP et al. Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. ISME J. 2021;15(1):211-27;

doi:10.1038/s41396-020-00775-z.

45. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-6; doi:10.1093/bioinformatics/btv033.
46. Francisco AP, Bugalho M, Ramirez M, Carrico JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009;10:152; doi:10.1186/1471-2105-10-152.
47. Nascimento M, Sousa A, Ramirez M, Francisco AP, Carrico JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2017;33(1):128-9; doi:10.1093/bioinformatics/btw582.
48. Garcia Quesada M, Peterson ME, Bennett JC, Hayford K, Zeger SL, Yang Y et al. Serotype distribution of remaining invasive pneumococcal disease after extensive use of ten-valent and 13-valent pneumococcal conjugate vaccines (the PSERENADE project): a global surveillance analysis. Lancet Infect Dis. 2025;25(4):445-56; doi:10.1016/s1473-3099(24)00588-7.
49. Kolkman MA, Wakarchuk W, Nuijten PJ, van der Zeijst BA. Capsular polysaccharide synthesis in Streptococcus pneumoniae serotype 14: molecular analysis of the complete cps locus and identification of genes encoding glycosyltransferases required for the biosynthesis of the tetrasaccharide subunit. Mol Microbiol. 1997;26(1):197-208; doi:10.1046/j.1365-2958.1997.5791940.x.
50. Yu J, Ravenscroft N, Davey P, Liyanage R, Lorenz O, Kuttel MM et al. New pneumococcal serotype 20C is a WciG O-acetyltransferase deficient variant of canonical serotype 20B. Microbiol Spectr. 2025;13(1):e0244324; doi:10.1128/spectrum.02443-24.
51. Calix JJ, Porambo RJ, Brady AM, Larson TR, Yother J, Abeygunwardana C et al. Biochemical, genetic, and serological characterization of two capsule subtypes among Streptococcus pneumoniae Serotype 20 strains: discovery of a new pneumococcal serotype. J Biol Chem. 2012;287(33):27885-94; doi:10.1074/jbc.M112.380451.
52. Tanner JJ, Boechi L, Andrew McCammon J, Sobrado P. Structure, mechanism, and dynamics of UDP-galactopyranose mutase. Arch Biochem Biophys. 2014;544:128-41; doi:10.1016/j.abb.2013.09.017.
53. Crisafulli G, Guidotti S, Muzzi A, Torricelli G, Moschioni M, Masignani V et al. An extended multi-locus molecular typing schema for Streptococcus pneumoniae demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. Infect Genet Evol. 2013;13:151-61; doi:10.1016/j.meegid.2012.09.008.
54. Kalizang'oma A, Swarthout TD, Mwalukomo TS, Kamng'ona A, Brown C, Msefula J et al. Clonal Expansion of a Streptococcus pneumoniae Serotype 3 Capsule Variant Sequence Type 700 With Enhanced Vaccine Escape Potential After 13-Valent Pneumococcal Conjugate Vaccine Introduction. J Infect Dis. 2024;230(1):e189-e98; doi:10.1093/infdis/jiae040.
55. Manna S, Ortika BD, Dunne EM, Holt KE, Kama M, Russell FM et al. A novel genetic variant of Streptococcus pneumoniae serotype 11A discovered in Fiji. Clin Microbiol Infect. 2018;24(4):428.e1-.e7; doi:10.1016/j.cmi.2017.06.031.
56. Amin-Chowdhury Z, Groves N, Sheppard CL, Litt D, Fry NK, Andrews N et al. Invasive pneumococcal disease due to 22F and 33F in England: A tail of two serotypes. Vaccine.

2021;39(14):1997-2004; doi:10.1016/j.vaccine.2021.02.026.

57. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. J Infect Dis. 2003;187(9):1424-32; doi:10.1086/374624.

58. Enright MC, Spratt BG. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. Microbiology. 1998;144(11):3049-60; doi:https://doi.org/10.1099/00221287-144-11-3049.

59. Croucher NJ, Kagedan L, Thompson CM, Parkhill J, Bentley SD, Finkelstein JA et al. Selective and genetic constraints on pneumococcal serotype switching. PLoS Genet. 2015;11(3):e1005095; doi:10.1371/journal.pgen.1005095.

60. Beddek AJ, Li MS, Kroll JS, Jordan TW, Martin DR. Evidence for capsule switching between carried and disease-causing Neisseria meningitidis strains. Infect Immun. 2009;77(7):2989-94; doi:10.1128/iai.00181-09.

61. Bohlin J, Brynildsrud OB, Sekse C, Snipen L. An evolutionary analysis of genome expansion and pathogenicity in Escherichia coli. BMC Genomics. 2014;15(1):882; doi:10.1186/1471-2164-15-882.

62. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M et al. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331(6016):430-4; doi:10.1126/science.1198545.

63. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. Cell Host Microbe. 2012;12(2):177-86; doi:10.1016/j.chom.2012.06.003.