Pneumo-Typer: a high-throughput capsule genotype visualization tool with integrated serotype and sequence type prediction for *Streptococcus pneumoniae*

Xiangyang Li[2]*, Yaoyao Zhu[1,4]*, Zilin Yang[2], Yanzi Li[4], Guohui Zhang[2], Xiangyu Wang[4], Xuan Zhao[5], Yinyan Huang[6], Huajie Zhang[3#], Bingqing Li[1,5#] & 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, Zhejiang, China.

[#] Corresponding authors.

E-mail addresses: mazhongrui@sdfmu.edu.cn (Z. Ma), bingqingsdu@163.com (B. Li), zhanghj@sdfmu.edu.cn (H. Zhang).

* Xiangyang Li and Yaoyao Zhu contributed equally to this work.

Abstract

Streptococcus pneumoniae remains a significant global pathogen, with capsule-based vaccines targeting only a subset of over 100 serotypes. Accurate serotyping is critical for evaluating vaccine impact and optimizing capsule formulations. While whole-genome sequencing (WGS)-based methods achieve high serotyping accuracy, they often struggle to identify serotype variants with altered capsule loci that may evade vaccine-induced immunity. Here, we present Pneumo-Typer, a high-throughput tool for capsule genotype (CapT) visualization, enabling rapid identification of isolates with capsular gene variations, including mutations and horizontal gene transfer. In addition, Pneumo-Typer integrates serotype and sequence type (ST) prediction, achieving 98.44% accuracy in predicting 93 serotypes across 17,683 genomes. Our analysis reveals substantial serotypes linked to multiple STs and CapTs, with STs and CapTs demonstrating relative evolutionary independence and limited CapT variation during ST expansion. Within multi-serotype serogroups, we observe striking CapT diversity, providing insights into the evolutionary dynamics of serogroup formation. These findings underscore the utility of Pneumo-Typer for epidemiological surveillance and vaccine design, contributing to efforts to mitigate the global burden of S. pneumoniae.

Introduction

Streptococcus pneumoniae (pneumococcus) is a major global pathogen responsible for a wide spectrum of diseases, including pneumonia, meningitis, sepsis, and otitis media, causing nearly 1.2 million deaths annually, particularly among young children, the elderly, and immunocompromised individuals[1, 2]. The introduction of pneumococcal conjugate vaccines (PCVs) has significantly reduced the burden of pneumococcal disease in many regions[3]. However, despite these successes, *S. pneumoniae* remains a considerable public health challenge due to its remarkable phenotypic and genetic plasticity, which enables rapid adaptation to selective pressures such as antibiotic use and vaccination[4]. This adaptability underscores the need for continuous surveillance of phenotypic and genetic characteristics, particularly serotypes, sequence types (STs), and capsular genotypes (CapTs), to better understand and manage the evolving epidemiology of pneumococcal disease[5, 6].

A critical virulence factor of *S. pneumoniae* is its polysaccharide capsule, which forms the basis for its classification into over 90 distinct serotypes[7, 8]. These serotypes vary in their invasiveness, antibiotic resistance, and responsiveness to vaccination[9, 10]. While PCVs target specific serotypes, vaccine-induced selective pressures have led to the emergence of non-vaccine serotypes in some regions, a phenomenon known as "serotype replacement"[11, 12]. This trend highlights the importance of accurate serotyping to monitor dynamic changes in circulating serotypes. Traditional serotyping methods, such as the Quellung reaction, are labor-intensive and require specialized expertise[13, 14, 15]. Molecular-based serotyping tools, such as PneumoCaT[16] and SeroCall[17], employ sequence alignment approaches but rely on high-quality whole genome sequencing (WGS) data. Other tools, like SeroBA[18], PneumoKITy[19], and PfaSTer[20], use *k*-mer-based approaches for faster typing but may struggle with mosaic sequences, which are common prevalent in pneumococcal populations[21].

Beyond serotype diversity, *S. pneumoniae* exhibits extensive genetic variability in STs, which are determined through multilocus sequence typing (MLST)[22]. STs provides insights into genetic relationships between pneumococcal strains, aiding in tracking clonal

expansions and understanding evolutionary trajectories[23, 24]. Recently, core genome sequence typing (cgST) has emerged as a higher-resolution alternative, clustering isolates based on 1,222 loci present in more than 95% of genomes[25, 26]. This method offers deeper insights into the genetic relatedness of isolates.

CapT, which focuses on the genetic determinants of capsule biosynthesis, can reveal cryptic or hybrid capsular types that may be missed by serotyping, uncovering antigenic shifts that enable strains to evade vaccine-induced immunity[27, 28, 29]. For instance, several strains identified as serotype 14 by SeroBA and SeroCall exhibit significant alterations in their capsule synthesis (*cps*) loci, allowing them to bypass vaccine protection[30]. Despite its importance in identifying vaccine escape variants, CapT remains underutilized due to the complexity of current bioinformatics workflows, which often require multiple platforms and considerable computational expertise[31, 32].

Despite the clear need for surveillance of pneumococcal serotypes, STs, and CapTs, existing tools for analyzing these factors are often fragmented and inefficient. Researchers frequently rely on multiple platforms or separate tools to analyze each feature, complicating workflows and hindering efforts to gain a comprehensive view of pneumococcal diversity. To address these challenges, we introduce Pneumo-Typer, a novel tool designed to simultaneously identify serotypes, STs (with an optional cgST), and CapTs in a single analysis pipeline from WGS data.

Here, we describe the development and validation of Pneumo-Typer using a comprehensive dataset of pneumococcal genomes from diverse geographical regions available in the NCBI database. Our results demonstrate that Pneumo-Typer accurately identifies serotypes, STs, and CapTs with high sensitivity and specificity. Furthermore, we analyze the relationships between serotype, ST, and CapT, uncovering or explaining certain phenomena and their potential applications. By providing a unified platform for pneumococcal typing, Pneumo-Typer has the potential to significantly enhance global efforts in epidemiological surveillance and vaccine design for *S. pneumoniae*.

Results

Overview of Pneumo-Typer

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.

The workflow of Pneumo-Typer is divided into three main steps: input data processing, sequence typing, and serotyping and visualization (Fig. 1). Key features include:

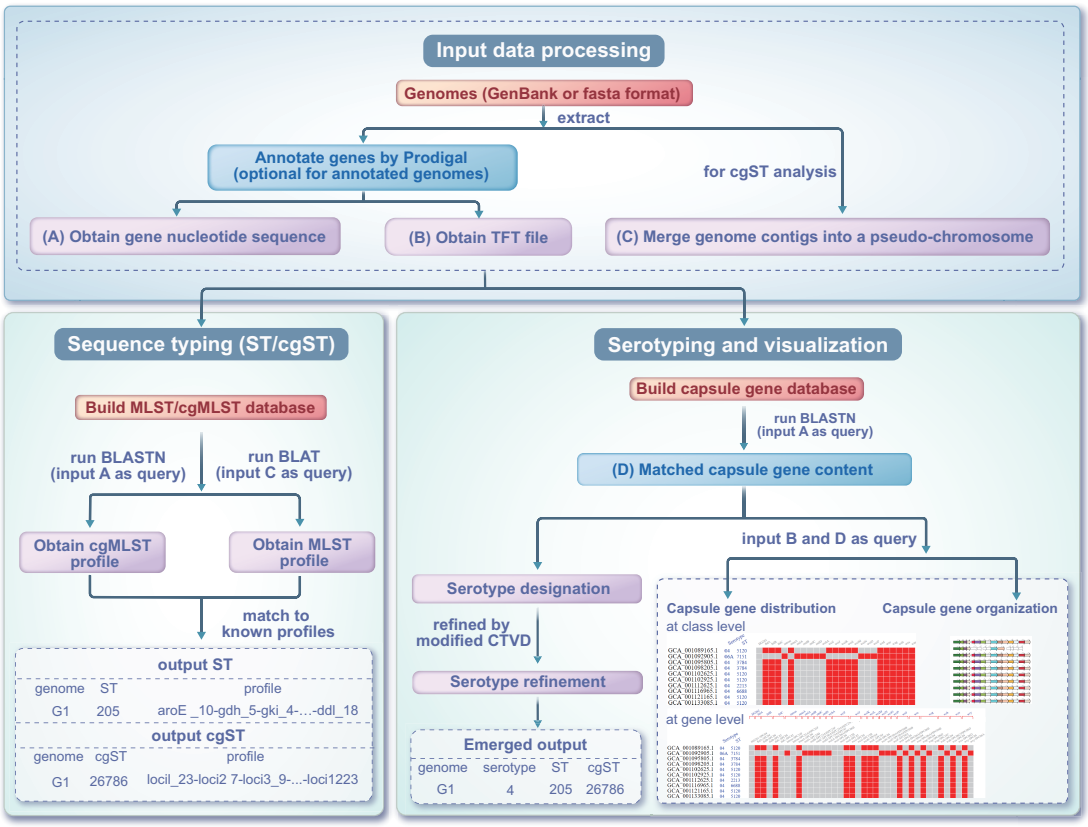
Simple Input: The only mandatory input for Pneumo-Typer is a directory containing genomes in either GenBank or fasta format, or a combination of both, ensuring straightforward acquisition of input data.

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 Capsular Typing Variant Database (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 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

133 represents the capsule operon organization (CPS_operon.svg), with distinct homologous
134 genes depicted in varied colors. To enable superior comparison, there is also an option to
135 map the capsule contexts onto a user-provided phylogenetic tree.



136

137 **Fig. 1 | The workflow of Pneumo-Typer.** During input data processing, genomes in GenBank
138 or fasta format are provided as input files. For each genome, gene nucleotide sequences (A)
139 and TFT (a seven-column tab-delimited feature table) files (B) are extracted. If a genome lacks
140 annotation, Prodigal is used for gene annotation. Optionally, Prodigal can annotate all input
141 genomes to maintain a uniform gene annotation process. For cgST analysis, all genome
142 contigs are seamlessly merged into a unified pseudo-chromosome (C). In the sequence typing
143 step, MLST and cgMLST databases are established. Inputs A or C are used as queries, and
144 BLASTN or BLAT is run against the established MLST and cgMLST databases to obtain MLST
145 or cgMLST profiles. These profiles are then matched against known MLST or cgMLST profiles
146 to output ST or cgST. In the serotyping and visualization step, a capsule gene database is
147 established. Input A is used as a query to run BLASTN against the established capsule gene
148 database, retrieving matched capsule gene content (D). Serotypes are assigned based on the

149 serotype(s) with the highest proportion of matched genes. Subsequently, serotype refinement
150 is performed using serotype-specific markers from a modified CTVD. The serotype, combined
151 with ST or cgST results, is then integrated and output. Additionally, inputs B and D are used to
152 generate and output heatmaps displaying the capsule gene distribution at the gene or class
153 level, as well as a map graphically representing the capsule operon organization.

154

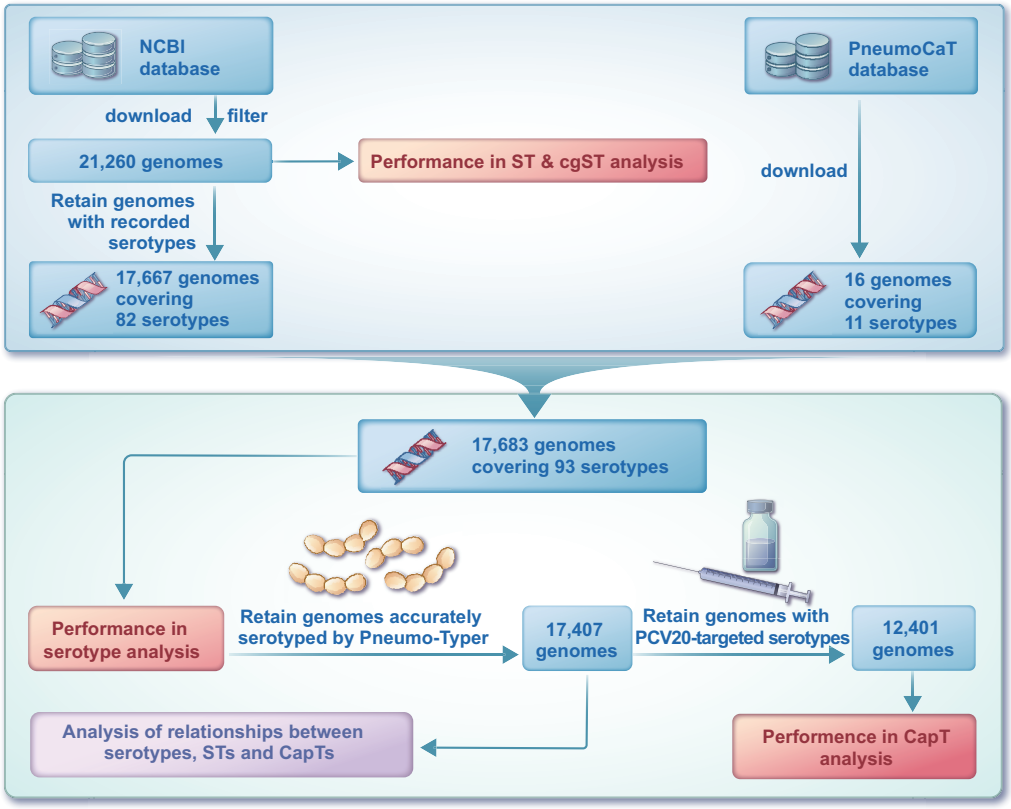
155 **Performance of Pneumo-Typer in serotyping**

156 The performance of Pneumo-Typer in serotyping was assessed using a dataset from the
157 NCBI Database, comprising a total of 21,260 pneumococcal genomes as of October 5,
158 2022. This dataset included 17,667 genomes with definitive serotypes, covering 82
159 serotypes. To evaluate Pneumo-Typer's performance across all 93 serotypes, 16
160 genomes representing the remaining 11 serotypes were obtained from an earlier
161 investigation[16] (Fig. 2).

162 At the serotype level, Pneumo-Typer achieved an accuracy of 84.98% (15,027 out of
163 17,683) in Serotype Designation (Step 1), which increased to 98.44% (17,407 out of
164 17,683) following Serotype Refinement (Step 2) (Table 1). For comparison, we tested
165 PneumoKITy and PfaSTer on the same dataset. PneumoKITy achieved an accuracy of
166 69.01% (12,203 out of 17,683), and PfaSTer achieved 90.29% (15,966 out of 17,683)
167 (Table S2). Pneumo-Typer achieved 100% prediction accuracy for 55 serotypes, 99%-100%
168 accuracy for 10 serotypes, and 98%-99% accuracy for 7 serotypes. However,
169 Pneumo-Typer had low prediction accuracy for some serotypes, such as 12A (0%), 25A
170 (0%), 32A (0%), 44 (0%), 9A (20%), and 35C (31.48%). At the serogroup level
171 Pneumo-Typer reached 99.84% accuracy (Table S3).

172 The visualization feature of Pneumo-Typer (Supplementary files 1, 2) allowed us to
173 specifically examine the CapTs of the mispredicted strains by Pneumo-Typer. Herein, a
174 strain recorded as serotype 1 but predicted as 16F had a capsule operon organization
175 distinctly resembling 16F (Fig. 3A). Similarly, two strains recorded as serotype 6D but
176 predicted as 6B showed a capsule operon organization more consistent with 6B, as 6B
177 had a gap between the *wciN* and *wciO* genes, which 6D lacked (Fig. 3B). Additionally, four

178 strains recorded as serotype 7B but predicted as 7C had a capsule operon organization
179 that matched both 7B and 7C (Fig. 3C), yet their capsule gene distribution at the gene
180 level aligned more closely with 7C (Fig. 3D). Excluding these misannotated serotypes
181 would further improve the serotype prediction accuracy of Pneumo-Typer. After exclusion,
182 the accuracy for serotype 1 would increase from 99.90% to 100%, for serotype 6D from
183 84.62% to 100%, and for serotype 7B from 86.67% to 100%. Therefore, Pneumo-Typer
184 can not only predict serotypes but also verify the correctness of annotated serotypes
185 through its CapT visualization features.



186

187 **Fig. 2 | Data acquisition, processing, and analysis workflow for this study.** First, 21,260
188 high-quality *S. pneumoniae* genomes were downloaded and filtered from the NCBI database
189 (as of October 5, 2022) to assess the performance of Pneumo-Typer in ST and cgST analysis.
190 From these, 17,667 genomes with recorded serotypes (covering 82 serotypes) were selected
191 and combined with 16 genomes (covering 11 serotypes) downloaded from the PneumoCaT
192 database, yielding a total of 17,683 genomes (covering 93 serotypes) for evaluating
193 Pneumo-Typer's performance in serotype analysis. Among these, 17,407 genomes that were

accurately serotyped by Pneumo-Typer were used to analyze the relationships between serotypes, STs, and CapTs. Finally, 12,401 genomes with PCV20-targeted serotypes were selected to evaluate Pneumo-Typer's performance in CapT analysis.

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.

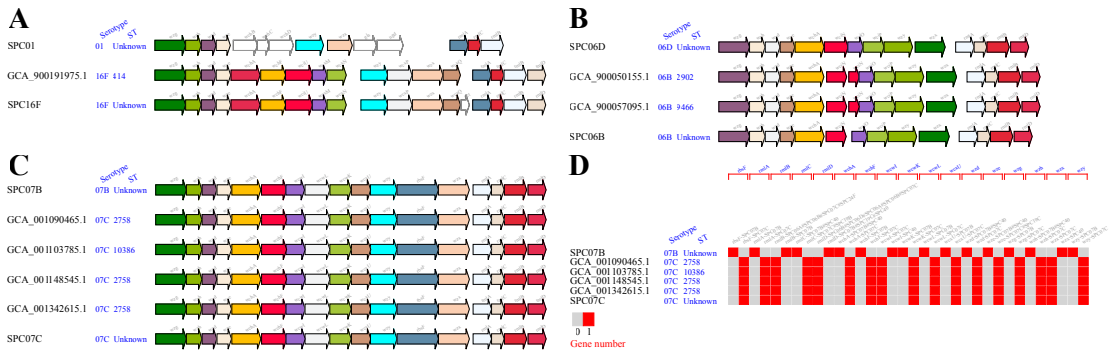


Fig. 3 | The role of CapT visualization in serotyping. From top to bottom: **A** CPS operon map of serotype 1 reference strain, GCA_900191975.1, and serotype 16F reference strain. **B** CPS operon map of serotype 6D reference strain, GCA_900050155.1, GCA_900057095.1, and serotype 6B reference strain.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

209 serotype 7C reference strain.

210

211 **Performance of Pneumo-Typer in sequence typing**

212 Pneumo-Typer demonstrated exceptional performance in the rapid assignment of STs,
213 processing 21,260 genomes in under 15 minutes. The assigned STs revealed the
214 presence of 3,250 known STs and the emergence of 18 novel STs (Table S4). However,
215 311 genomes remained unassigned to any ST, likely due to incomplete ST locus
216 identification in draft-status genomes. Among the identified STs, several exhibited high
217 isolate counts, notably ST217 (459 isolates), ST156 (390 isolates), ST180 (387 isolates),
218 ST199 (349 isolates), and ST191 (319 isolates). A substantial number of STs, 1,801, were
219 assigned to only one isolate each (Table S5).

220 Using Pneumo-Typer, 10,859 out of 21,260 genomes were categorized into 10,103
221 distinct cgSTs (Table S4). Additionally, the analysis revealed 10 novel cgST profiles. The
222 most frequently occurring cgST was cgST12235, accounting for 141 isolates, followed by
223 cgST1494, which accounted only for 7 isolates. A substantial number of cgSTs, 9,600,
224 were assigned to only one isolate each. However, a significant number of genomes,
225 10,391, could not be linked to established cgSTs, indicating that known cgSTs could only
226 be attributed to almost half of the genomes. This suggests a limitation in using cgST for
227 molecular typing, given the extensive number of gene loci required, which may not be fully
228 assembled in draft-status genomes.

229 **Bidirectional serotype-ST associations**

230 The performance of Pneumo-Typer in both serotyping and sequence typing enables a
231 systematic study elucidating the relationships between serotypes and STs. We selected
232 serotype and ST data from 17,407 genomes, which were accurately serotyped by
233 Pneumo-Typer (covering 90 serotypes), for analysis (Fig. 2). This selection excludes any
234 errors arising from potentially mispredicted or misannotated isolates. The relationships
235 between serotypes and STs (Table S6) can be analyzed in two ways: the STs associated
236 with each serotype and the serotypes associated with each ST.

237 First, we analyzed the STs associated with each serotype (Table S7) and found that the

238 vast majority—78 out of 90 serotypes—were associated with multiple STs (Fig. 4A). A
239 subset of serotypes, specifically 19F (≥284 STs), 6B (≥257 STs), 6A (≥238 STs), 19A
240 (≥216 STs), and 23F (≥182 STs), exhibited a particularly high number of coexisting STs. In
241 contrast, a small group of serotypes, including 7A, 10C, 11C, 11D, 11F, 16A, 19C, 24B, 42,
242 43, 47A, and 47F were observed to have only one unique ST association. However, the
243 ST richness of each serotype heavily depended on its sample size (Fig. 4B). To evaluate
244 the ST diversity of each serotype independently of sample size, we excluded serotypes
245 with fewer than 100 isolates to avoid potential bias and used Shannon's entropy as a
246 measure of diversity (higher Shannon's entropy indicates more STs with a more even
247 distribution). The results showed that serotypes 6A, 6B, 19F, 16F, and 19A had relatively
248 high ST diversity, while serotypes 5, 38, 7F, 8, and 12F had relatively low ST diversity (Fig.
249 4C), with ST diversity showing low correlation with sample size (Fig. 4D). These results
250 indicate significant differences in ST diversity among individual serotypes.
251 Conversely, each of the majority—2,618 out of 2,815 STs was linked to a unique
252 serotype (Table S8). For instance, all isolates of certain STs, such as the 458 isolates of
253 ST217 and the 110 isolates of ST3081, were exclusively associated with serotype 1.
254 Therefore, an isolate's ST can often help determine its serotype. We selected the STs of
255 the misannotated isolates (Fig. 3) from Table S8 and found that, except for ST2377 (not
256 found in Table S8), the serotypes associated with these STs did not match the recorded
257 serotypes, but they aligned with the predictions made by Pneumo-Typer (Table 2). It is
258 important to acknowledge that many STs were represented by only a limited number of
259 isolates, indicating the need for more comprehensive sampling to confirm these findings.
260 Besides, a small proportion of STs (197 out of 2,815) were associated with multiple
261 serotypes. Notably, ST156, ST162, ST199, ST172, and ST193 each were related to 10, 9,
262 8, 7, and 7 serotypes, respectively. Additionally, there were 167 STs associated with two
263 serotypes, 33 with three, 15 with four, and 3 each with five and six serotypes (Fig. 4E).
264 Given that STs represent highly conserved housekeeping genes with generally low
265 mutation rates, the phenomenon of a single ST corresponding to multiple serotypes is
266 likely due to capsular switch events[33].
267

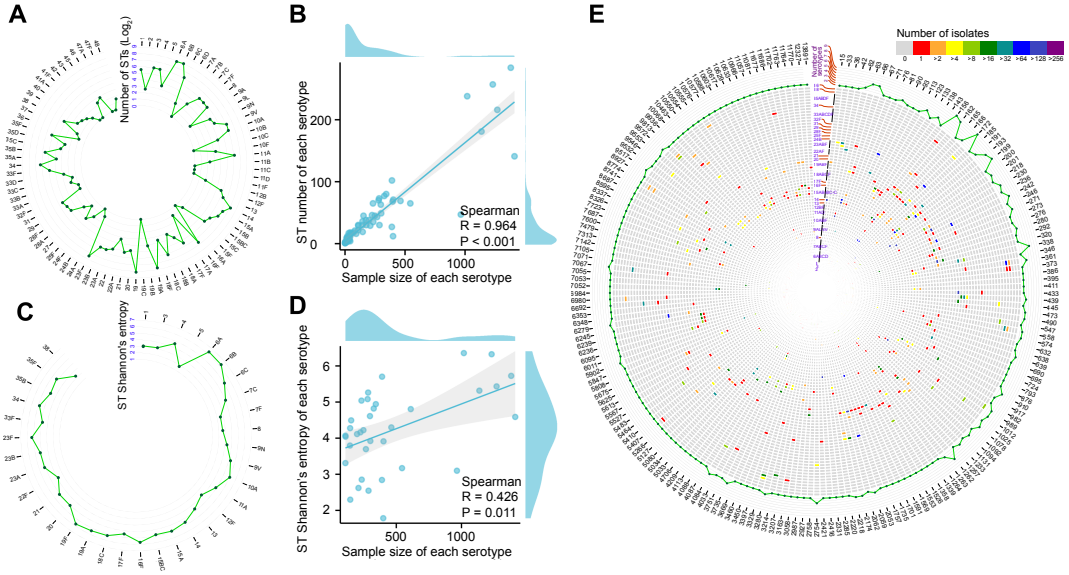


Fig. 4 | Serotype and ST associations. A Number of STs associated with each serotype. **B** Correlation between sample size and ST number of each serotype. **C** Shannon's entropy of STs associated with each serotype. **D** Correlation between sample size and ST Shannon's entropy of each serotype. **E** 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 2 | The role of STs in serotype determination.

Strain	Recorded serotype	Serotype by Pneumo-Typer	ST	ST: serotype|Number in Table S8
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

Serotypes represent phenotypes, while CapTs denote genotypes, with the latter determining the former. The CapTs of serotypes are diverse due to gene mutations and

282 horizontal gene transfer events. Thanks to Pneumo-Typer's capability to analyze CapTs,
283 we have, for the first time, observed the CapT diversity of each serotype (Table S9).
284 Based on capsule gene content, we established a naming convention for CapTs. For
285 example, the CapT of strain GCA_001457635.1 was named Serotype SPC01/15/93.33%.
286 Here, SPC01 indicates the serotype designation as serotype 1, 15 indicates 15 matched
287 capsule genes, and 93.33% indicates that 14 out of 15 matched capsule genes belong to
288 serotype 1. For strains indistinguishable by these three criteria, numerical codes were
289 added for further differentiation, e.g., Serotype SPC04/16/81.25% was subdivided into
290 Serotype SPC04/16/81.25%-1 and Serotype SPC04/16/81.25%-2, ensuring that each
291 CapT corresponds to a unique capsule gene content.

292 We selected 17,407 genomes with confirmed serotypes by Pneumo-Typer for analysis
293 (Fig. 2). The number of CapTs varied among serotypes, with 6A, 23F, 6B, 19F, and 19A
294 having the highest numbers (Fig. 5A). As expected, the number of CapTs was highly
295 correlated with sample size (Fig. 5B). We excluded serotypes with fewer than 100
296 samples and used Shannon's entropy to assess CapT diversity. The results showed that
297 serotypes 6A, 7F, 23F, 6C, and 23B had the highest Shannon's entropy values (Fig. 5C),
298 and CapT diversity was not correlated with sample size (Fig. 5D). These findings indicate
299 significant differences in CapT diversity among serotypes, suggesting that different
300 serotypes undergo capsule genetic changes at different frequencies. Furthermore, CapT
301 diversity and ST diversity reflect changes in different genomic regions, showing only a
302 weak correlation (Fig. 5E). This suggests that the evolution of CapTs and STs is not
303 synchronized.

304 Additionally, we observed that serotypes belonging to serogroups containing multiple
305 serotypes (e.g., 6A, 7F, and 23F) tended to exhibit higher CapT diversity compared to
306 those in serogroups that only contained a single serotype (e.g., 5, 1, and 3) (Fig. 5F). This
307 may be attributed to certain serotypes undergoing more frequent genetic changes in their
308 capsules, some of which lead to antigenic shifts. These shifts result in the emergence of
309 serotype variants that cluster under the same serogroup due to genetic or antigenic
310 similarities. However, serotypes 20 (ranked 6th) and 13 (ranked 7th) also displayed

relatively high Shannon's entropy values. This prompted us to investigate whether these serotypes might have serotype variants as well. An in-depth analysis of the heatmap (at the class level) for serotypes 20 and 13 (Supplementary file 1) revealed that among the 111 strains of serotype 20, 91 exhibited mutations of the *whaF* gene, which synthesizes a putative glycosyltransferase. Previous studies reported that serotyping methods using antibodies could not distinguish serotype 20-*whaF* mutants, designating them as serotype 20B[34]. Additionally, four strains of serotype 20 had mutations in the *wciG* gene, which synthesizes an O-acetyltransferase, crucial for distinguishing serotypes 35C and 42[35]. Inactivation of the *wciG* gene may also lead to capsule antigenic shifts in serotype 20. Regarding serotype 13, we observed that in 227 out of 267 strains, the *whaG* gene responsible for a glycosyltransferase was functional, contrary to the standard database that records it as a pseudogene. This phenomenon has not been previously reported.

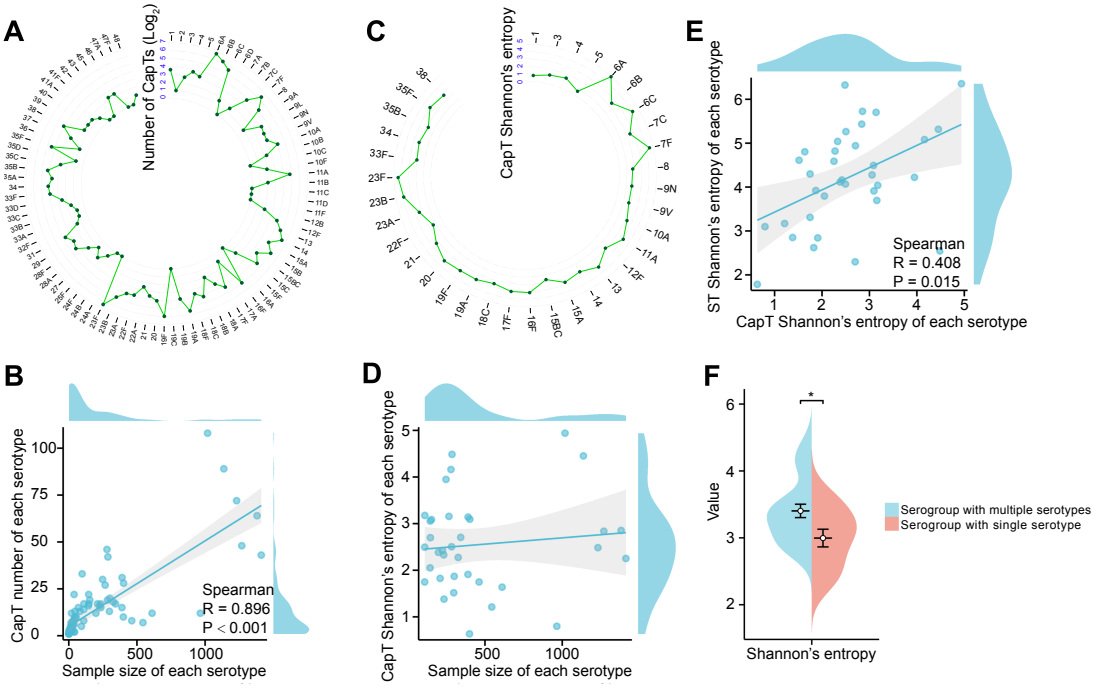


Fig. 5 | CapT diversity across serotypes. A Number of CapTs associated with each serotype. **B** Correlation between sample size and CapT number of each serotype. **C** Shannon's entropy of CapTs associated with each serotype. **D** Correlation between sample size and CapT Shannon's entropy of each serotype. **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.

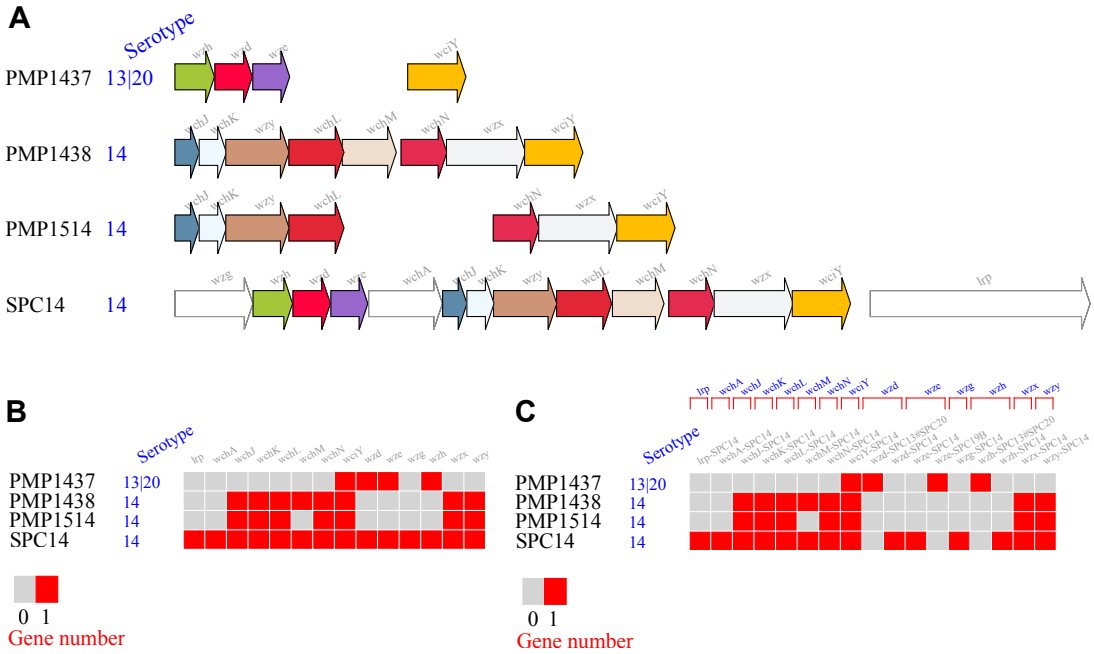
330

331 **Key capsular gene variations in PCV20-targeted serotypes**

332 To further test Pneumo-Typer's performance in detecting key capsular gene variations
333 causing potential capsule antigenic shifts, we visualized three serotype 14 variants
334 reported in a previous study[30]. Compared to the standard CapT of serotype 14, PMP1437,
335 PMP1438 and PMP1514 all exhibited multiple gene deletions, and PMP1514 had an
336 inserted unknown sequence (not matched to any genes in the capsule gene database).
337 PMP1437 appeared to be a hybrid, containing genes from serotype 13|20 and serotype
338 14 (Fig. 6). Therefore, Pneumo-Typer's CapT visualization feature can clearly identify
339 various capsular gene variations.

340 In the context of vaccine design, it is crucial to focus on potential capsule antigenic
341 shifts in vaccine-covered serotypes, as these directly impact vaccine efficacy. PCV20, the
342 latest pneumococcal conjugate vaccine, was approved for use in adults aged 18 years
343 and older on June 8, 2021, and subsequently recommended for routine use in infants and
344 children in June 2023. In this study, we analyzed the serotypes included in PCV20—1, 3,
345 4, 5, 6A, 6B, 7F, 8, 9V, 10A, 11A, 12F, 14, 15B, 18C, 19A, 19F, 22F, 23F, and 33F—to
346 identify key capsular gene variations, including gene mutations and horizontal gene
347 transfer (Table 3), based on heatmaps generated by Pneumo-Typer (Supplementary file
348 1). For serotype 3, 42 out of 544 strains (7.7%) exhibited mutations in the *wzd* gene
349 (responsible for capsular polysaccharide biosynthesis protein). In serotype 11A, mutations
350 were observed in 63 out of 402 strains (15.7%) in the *wcwC* gene (responsible for
351 O-acetyltransferase), and in 24 out of 402 strains (6.0%) in the *wcjE* gene (responsible for
352 putative acetyltransferase). For serotype 14, the *wciY* gene (responsible for putative
353 glycerol phosphotransferase) showed mutations in 1,232 out of 1,412 strains (87.1%). In
354 serotype 19A, the *rmlD* gene (responsible for dTDP-4-dehydrorhamnose reductase) was
355 mutated in 648 out of 1,269 strains (51.1%). Serotype 1 exhibited an insertion of the *rmlD*
356 gene from other serotypes in 8 out of 965 strains (0.8%). Serotype 33F showed an
357 insertion of the *wcyO* gene (responsible for putative glycosyltransferase) from serotype
358 33C in 23 out of 145 strains (15.9%). These results suggest that vaccines based on

359 standard strains may lose efficacy or become ineffective in responding to capsular gene
360 variations. Incorporating these CapT variants into vaccine formulations may be a potential
361 solution.



362

363 **Fig. 6 | Visualization of CapTs of three serotype 14 variants by Pneumo-Typer. A**
364 Capsule operon organization. **B** Capsule gene distribution at the class level. **C** Capsule gene
365 distribution at the gene level.

366

367 **Table 3 | Key capsular gene variations in PCV20-targeted serotypes.**

Serotype	Number	Gene mutation/number (≥3)	Horizontal gene transfer/number
1	965	*wzx*/4	*rmlD*/8; *wchO*/1
3	544	*wzd*/42	*wzh*/4
4	298	*wciJ*/5; *wciI*/3	N/A
5	400	N/A	N/A
6A	1018	*wzy*/12; *rmlD*/6	N/A
6B	1231	*wzy*/12; *rmlD*/5; *wzx*/4	N/A
7F	287	N/A	*wchQ*/2
8	210	*wchA*/8	N/A
9V	462	*wcjD*/6, *wzg*/4; *wcjA*/3; *wzx*/3	N/A
10A	233	*wcrG*/3	N/A
11A	402	*wcwC*/63; *wcjE*/24; *wzx*/9	N/A
12F	390	*wzx*/3	N/A
14	1412	*wciY*/1232; *wchA*/11; *wchN*/7; *wchM*/5; *wzx*/5;	N/A

		wzy/4; *wzg*/4; *wchK*/3	
15B	98	N/A	N/A
18C	284	*wchA*/6; *wzg*/5	*wcyO-glf*/1
19A	1269	*rmlD*/648; *wchQ*/4; *wzy*/3; *wchO*/3	*wcwD*/1
19F	1381	*wzg*/15; *rmlD*/3; *wze*/3; *wchO*/3	N/A
22F	235	N/A	*wchP*/1
23F	1137	*wzy*/5; *wzx*/4; *wzd*/4	*wcyO-glf*/1
33F	145	N/A	*wcyO*/23

368

369 **ST-CapT relationships and implications for vaccine design**

370 As mentioned earlier, STs and CapTs evolve relatively independently (Fig. 5E).
371 Nevertheless, we sought to identify potential patterns within the relationships between STs
372 and CapTs. When analyzing the relationships between STs and CapTs in PCV20-targeted
373 serotypes, we observed various patterns, including one-to-one, one-to-many, and
374 many-to-one correspondences (Supplementary file 3). However, these correspondences
375 were not random. To better illustrate these patterns, we further examined serotype 11A
376 and found that STs within CC200 and CC580 corresponded to serotype SPC11F/XX/XX,
377 STs within CC33, CC503, CC1079, and NA (unknown) corresponded to serotype
378 SPC11D/XX/XX, while STs in other CCs corresponded to serotype SPC11A/XX/XX (Fig.
379 7A). This suggests that STs with closely related gene sequences within the same clonal
380 complex (CC) tend to correspond to CapTs with similar capsule gene contents, indicating
381 the presence of vertical inheritance constraints within specific STs or CCs.

382 Due to the vertical inheritance constraints, it is possible to predict certain CapTs based
383 on specific STs. To explore this further, we selected isolates with key variations in capsular
384 genes from PCV20-targeted serotypes listed in Table 4 and mapped the relationships
385 between their STs and CapTs (Fig. 7B). Notably, ST303 (serotype 1-*rmlD*), ST700
386 (serotype 3-*wzd*), ST673 (serotype 33F-*wcyO*), ST99 (serotype 11A-*wcwC*), and ST62
387 (serotype 11A-*wcjE*) were stood out. When these STs are prevalent, it becomes essential
388 to pay close attention to CapT variations when designing vaccines targeting these STs.
389 For instance, serotype 11A-ST99 was prevalent in South Korea from 2004 to 2013,
390 accounting for 27.8% of total serotype 11A isolates[36]. 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[37]. In our study, all ST673 isolates had an insertion of the *wcyO* gene, while ST717 and ST100 did not. Therefore, to effectively target infections caused by ST99 or ST673, it is crucial to use serotype 11A strains with *wcwC* gene mutations or serotype 33F strains with *wcyO* gene insertions in vaccine design.

Fig. 7 | Sankey diagram illustrates the relationships between STs and CapTs for **A** serotype 11A and **B** several serotypes with capsular gene variations.

Discussion

For a long time, epidemiological surveillance of *S. pneumoniae* has primarily focused on analyzing clinical outcomes, age distribution, geographical location, colonization or infection sites, and vaccination status. Further studies have often involved antibiotic

405 resistance and the distribution of serotypes and sequence types (STs) among isolates. In
406 recent years, the identification of increasing cases of vaccine escape driven by capsular
407 gene variations has heightened the demand for CapT detection[30].

408 Pneumo-Typer was developed in response to this need, enabling the simultaneous
409 identification of serotype, ST (with optional cgST), and CapT in *S. pneumoniae*.
410 Importantly, Pneumo-Typer not only integrates the analysis of serotypes, STs, and CapT
411 but also introduces an innovative approach to serotyping. Furthermore, it establishes a
412 novel visualization framework and nomenclature system for CapT analysis. With features
413 such as simple input, batch processing, innovative serotyping, and advanced visualization,
414 Pneumo-Typer marks a significant advancement in pneumococcal research.

415 In terms of serotyping, Pneumo-Typer employs a novel two-step sequence alignment
416 approach. Unlike PneumoCaT[16], which relies on full-genome alignment and struggles to
417 determine serotypes for genomes with less than 90% capsular operon sequence
418 coverage, Pneumo-Typer requires only a match to one gene from the capsule gene
419 database to assign a serotype. This approach makes Pneumo-Typer more inclusive,
420 accommodating genomes with incomplete *cps* loci or varying degrees of capsule gene
421 variation. Despite this inclusivity, its accuracy remains uncompromised, achieving 98.44%,
422 outperforming other serotyping tools like PneumoKITy (69.01%) and PfaSTer (90.29%).

423 Moreover, we applied CapT and ST data to *S. pneumoniae* serotype determination for
424 the first time. Mispredictions caused by annotation discrepancies can be corrected using
425 Pneumo-Typer's CapT visualization capabilities. Additionally, STs are a valuable aid in
426 serotype determination, as each ST is predominantly associated with a specific serotype.

427 CapT represents the genetic foundation of serotype, while the serotype reflects these
428 genetic characteristics at the antigenic level. However, due to the diversity of CapT within
429 individual serotypes, it is no longer feasible to deduce the CapT directly from the serotype.
430 CapT plays a critical role in vaccine design, particularly for glycoconjugate vaccines, as it
431 provides essential insights into capsule antigens. Before the advent of Pneumo-Typer,
432 CapT analysis required a complex, multi-step bioinformatics workflow involving several
433 software tools. Pneumo-Typer simplifies this process by enabling the visualization of
434 capsule gene distribution and operon organization, allowing for a more intuitive

435 understanding of capsular gene variations. Additionally, Pneumo-Typer's operon
436 organization map can reveal potential insertions of unknown sequences (represented as
437 gaps, as shown in PMP1514 in Fig. 6), prompting further investigation into the origins of
438 these sequences.

439 In the case illustrated in Fig. 6, three isolates were identified as serotype 14 by SeroBA
440 and SeroCall. Although Pneumo-Typer also classified two of these isolates as serotype 14
441 (with another classified as serotype 13/20), its CapT visualization feature clearly reveals
442 significant differences in the *cps* loci of these isolates compared to the standard serotype
443 14 strain. The inclusion of CapT data helps overcome the limitations of molecular methods
444 in predicting serotypes. By leveraging Pneumo-Typer's CapT analysis capabilities, we
445 further identified numerous capsular gene variations within serotypes targeted by the
446 PCV20 vaccine, providing valuable insights for future vaccine design.

447 The strong performance of Pneumo-Typer in integrating serotypes, STs, and CapT has
448 allowed us to efficiently analyze the relationships between these factors, a task that was
449 previously challenging or impossible. Through this analysis, several interesting patterns
450 have emerged. First, a single serotype can correspond to multiple STs, and the diversity of
451 STs varies across different serotypes. This finding supports the notion that a single
452 serotype often encompasses significant genetic diversity within its isolates[26]. Conversely,
453 a single ST is typically associated with only one serotype. This observation is consistent
454 with the findings of Enright and Spratt, who reported that capsular genes spread
455 horizontally within the pneumococcal population over the long term, while clonal
456 expansion with rare serotype changes occurs in the short term[38].

457 Second, individual serotypes generally exhibit varying degrees of CapT diversity.
458 Serotypes belonging to multi-serotype serogroups tend to show higher CapT diversity
459 compared to single-serotype groups, consistent with Croucher *et al.*'s observation that
460 multi-serotype serogroups experience a higher frequency of within-group serotype
461 switching events[39].

462 Finally, our analysis highlights the relative evolutionary independence between STs and
463 CapTs, with limited CapT changes occurring during ST expansion. This can be explained
464 by the distinct evolutionary pressures acting on core genomes (represented by STs) and

465 non-core genomes (represented by CapTs). While STs evolve conservatively under
466 purifying selection, CapTs are subject to immune pressure and change only when
467 necessary. Horizontal gene transfer allows for rare but significant shifts in capsule
468 structure, while vertical inheritance preserves CapT stability during ST expansion. Thus,
469 capsule changes are limited but adaptive, occurring primarily in response to strong
470 external pressures such as vaccination or immune evasion[40, 41, 42, 43].

471 In conclusion, Pneumo-Typer facilitates rapid, simultaneous, and accurate analysis of
472 serotypes, STs, and CapTs, contributing significantly to the epidemiological surveillance
473 and vaccine design of *S. pneumoniae*. Its user-friendly design makes it an excellent tool
474 for researchers. Looking forward, we plan to develop an online version of Pneumo-Typer
475 to further enhance its accessibility and ease of use for non-bioinformatics professionals.

476 **Methods**

477 **Capsule gene database construction**

478 The capsule gene database was established by collating GenBank files associated with
479 capsule gene loci from reference strains, which covered a spectrum of 93 serotypes. The
480 foundation of the database, comprising 90 serotypes, was extracted from the EMBL
481 Nucleotide Sequence Database under accession numbers CR931632–CR931722[7]. The
482 dataset was enriched with two more serotypes 6C and 35D, sourced from NCBI under
483 accession numbers JF911515[44] and KY084476[45], respectively. Serotype 6D's sequence
484 was procured from the PneumoCaT database[16] and annotated leveraging Prokka v1.12[46].
485 6E and 23B1 were not included as independent serotypes in the capsule gene database
486 since they emerged as novel CapT subtypes. For instance, the isolate carrying the "6E"
487 capsule gene locus yielded the serotype 6B capsule[47, 48].

488 The nucleotide sequences of the 93 serotypes were extracted from the GenBank files
489 by a custom Perl script. Subsequently, each capsule gene was tagged with its respective
490 gene name and serotype, exemplified as *wzd*-SPC30. The genes underwent a two-step
491 refinement: (i) exclusion of genes with incomplete sequences or encoding transposases.
492 (ii) dereplication of genes sharing 100% nucleotide identity using a proprietary Perl script.
493 Upon completion, the capsule gene database encompassed 1,290 genes spread across
494 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']).

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/)[49]. 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.

Execution of Pneumo-Typer

Input data processing. It extracted gene nucleotide sequences and TFT format files (a seven-column tab-delimited feature table) for each genome[50]. If a genome was without an annotation, Prodigal[51] was used for gene annotation. Optionally, Prodigal could also annotate all input genomes to maintain a uniform gene annotation process. For cgMLST analysis, it seamlessly merged all genome contigs into a unified pseudo-chromosome.

Sequence typing (ST/cgST). The ST analysis 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 analysis followed a similar blueprint but with nuances: (i) BLAT[52] was harnessed for precise mapping of cgMLST loci against

525 genomes. (ii) The "-fastMap" parameter sped up the process, fusing 1,222 cgMLST
526 sequences into one, but for sequences like SPNE00213.fas that exceeded 5,000 bp,
527 “-fastMap” was not used. (iii) Missing allele genes in the genome get a unique marker, like
528 SPNE00001_N.

529 *Serotyping and visualization.* The serotyping strategy was as follows: (i) Genes were
530 aligned against the capsule gene database using BLASTN. (ii) Results underwent
531 rigorous filtering (identity ≥ 70%, coverage ≥ 95%, E-value = $1e^{-5}$) and then aligned with
532 the list file in the capsule gene database for a detailed gene distribution. (iii) The capsule
533 genes involved in every serotype were computed. If a gene aligned with multiple
534 serotypes, it was tallied for each. (iv) Serotype designation: the predominant serotype for
535 a genome was the serotype with the most capsule genes. (v) Serotype refinement:
536 serotype-specific markers from a modified CTVD based on PneumoCaT refined
537 serotyping results for certain serotypes. (vi) The serotyping results and ST/cgST results
538 were merged. The visualization was accomplished with the assistance of the software
539 Gcluster[50], which underwent extensive modifications specifically for this study. The
540 visualization strategy was as follows: (i) A tab-delimited file was produced, detailing the
541 potential capsule gene distribution, categorized by capsule gene locus in each genome. (ii)
542 This data was converted into heatmaps, presenting the capsule gene distribution at the
543 class or gene level. (iii) This data, in conjunction with TFT format files, was visualized as
544 capsule gene organization. (iv) The predicted serotype and ST/cgST information was
545 integrated for a comprehensive visual insight.

546 **Data acquisition and processing**

547 To test the performance of Pneumo-Typer, all pneumococcal genomes from the NCBI
548 database were downloaded as of Oct 5, 2022. These genomes, procured in GenBank
549 format, were extracted using a custom Perl script aligned with a CSV file input. This CSV
550 file was readily available from the NCBI genome database website
551 (https://www.ncbi.nlm.nih.gov/genome/browse/) by searching with the term
552 "*Streptococcus pneumoniae*". Furthermore, the serotype information was retrieved from
553 the GenBank files through our proprietary Perl program. To ensure the quality and
554 accuracy of the genomes, two diagnostic tools: CheckM[53] and QUAST[54] were employed.

555 CheckM evaluated each genome's completeness and potential contamination based on
556 lineage-specific sets of single-copy genes. Genomes boasting near-perfection (≥90%
557 completeness) and negligible contamination (≤5% contamination)[55] were selected.
558 QUAST offered metrics on assembly quality, and the genomes were sieved by stringent
559 criteria, including contig size ≤ 500 bp and N50 ≥ 40 kb. After filtering process, 21,260
560 genomes were obtained, and then input into Pneumo-Typer. The analysis was performed
561 on an Ubuntu 20.04 server equipped with four Intel Xeon Platinum 8260L 24-core
562 processors and 512 MB RAM, harnessing 192 CPU threads. Parameter -p was set to "T",
563 prompting Prodigal to annotate all genomes under examination during Pneumo-Typer
564 operation. Additionally, 16 metagenomes were procured in fastq format from an earlier
565 investigation[16]. After assembly via MetaHIT[56], these genomes underwent serotype
566 determination using Pneumo-Typer.

567 **Clonal Complexes analysis**

568 The concept of MLST Clonal Complexes (CC) refers to STs that have differences in a
569 single locus variant (SLV) based on the eBURST algorithm[57]. The analysis was performed
570 using the PHYLOViZ 2.0 software[58].

571 **ST or CapT diversity calculation**

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

574 $$\mathrm{S(X)} = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

575 Where S(X) represents the Shannon's entropy of the variable X (either ST or CapT),
576 $p(x_i)$ denotes the probability of X taking the *i*-th value x_i, and *n* is the total number of
577 possible values for X.

578 **Statistical analysis**

579 Spearman's correlation was used to analyze the relationship between two sets of
580 variables. A correlation coefficient |R| ≥ 0.8 indicates a strong correlation, while 0.3 ≤ |R| <
581 0.5 suggests a weak correlation, and |R| < 0.3 denotes little to no correlation. To assess
582 statistical differences between the two groups, an independent samples *t*-test was
583 performed, with significance indicated by * $P < 0.05$.

Data availability

All data that support the findings of this paper are presented within the paper and the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Code availability

The source code related to 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".

Acknowledgements

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).

Competing interests

The authors declare no competing interests.

References

1. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. *Nature reviews Microbiology* **16**, 355-367 (2018).
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. *The Lancet Infectious diseases* **18**, 1191-1210 (2018).
3. Wahl B*, et al.* Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national

616 estimates for 2000-15. *The Lancet Global health* **6**, e744-e757 (2018).
617 4. Telford JL. Bacterial genome variability and its impact on vaccine design. *Cell host &*
618 *microbe* **3**, 408-416 (2008).
619 5. Varvio SL, Auranen K, Arjas E, Mäkelä PH. Evolution of the capsular regulatory genes
620 in Streptococcus pneumoniae. *The Journal of infectious diseases* **200**, 1144-1151
621 (2009).
622 6. Jefferies JM, Smith A, Clarke SC, Dowson C, Mitchell TJ. Genetic analysis of diverse
623 disease-causing pneumococci indicates high levels of diversity within serotypes and
624 capsule switching. *Journal of clinical microbiology* **42**, 5681-5688 (2004).
625 7. Bentley SD*, et al.* Genetic analysis of the capsular biosynthetic locus from all 90
626 pneumococcal serotypes. *Plos Genet* **2**, e31 (2006).
627 8. Geno KA*, et al.* Pneumococcal Capsules and Their Types: Past, Present, and Future.
628 *Clinical microbiology reviews* **28**, 871-899 (2015).
629 9. Song JH, Dagan R, Klugman KP, Fritzell B. The relationship between pneumococcal
630 serotypes and antibiotic resistance. *Vaccine* **30**, 2728-2737 (2012).
631 10. Liñares J, Ardanuy C, Pallares R, Fenoll A. Changes in antimicrobial resistance,
632 serotypes and genotypes in Streptococcus pneumoniae over a 30-year period. *Clinical*
633 *microbiology and infection : the official publication of the European Society of Clinical*
634 *Microbiology and Infectious Diseases* **16**, 402-410 (2010).
635 11. Weinberger DM, Malley R, Lipsitch M. Serotype replacement in disease after
636 pneumococcal vaccination. *Lancet (London, England)* **378**, 1962-1973 (2011).
637 12. Croucher NJ*, et al.* Population genomics of post-vaccine changes in pneumococcal
638 epidemiology. *Nat Genet* **45**, 656-663 (2013).
639 13. Mauffrey F*, et al.* Comparison of sequential multiplex PCR, sequetyping and whole
640 genome sequencing for serotyping of Streptococcus pneumoniae. *PloS one* **12**,
641 e0189163 (2017).
642 14. Abdul Rahman NA*, et al.* The Molecular Approaches and Challenges of Streptococcus
643 pneumoniae Serotyping for Epidemiological Surveillance in the Vaccine Era. *Polish*
644 *journal of microbiology* **72**, 103-115 (2023).
645 15. Jauneikaite E*, et al.* Current methods for capsular typing of Streptococcus
646 pneumoniae. *Journal of microbiological methods* **113**, 41-49 (2015).
647 16. Kapatai G*, et al.* Whole genome sequencing of Streptococcus pneumoniae:
648 development, evaluation and verification of targets for serogroup and serotype
649 prediction using an automated pipeline. *PeerJ* **4**, e2477 (2016).
650 17. Knight JR*, et al.* Determining the serotype composition of mixed samples of
651 pneumococcus using whole-genome sequencing. *Microb Genom* **7**, (2021).
652 18. Epping L*, et al.* SeroBA: rapid high-throughput serotyping of Streptococcus
653 pneumoniae from whole genome sequence data. *Microb Genom* **4**, (2018).
654 19. Sheppard CL*, et al.* PneumoKITy: A fast, flexible, specific, and sensitive tool for
655 Streptococcus pneumoniae serotype screening and mixed serotype detection from
656 genome sequence data. *Microb Genom* **8**, (2022).
657 20. Lee JT, Li X, Hyde C, Liberator PA, Hao L. PfaSTer: a machine learning-powered
658 serotype caller for Streptococcus pneumoniae genomes. *Microbial genomics* **9**,
659 (2023).

21. 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. *Frontiers in microbiology* **11**, 1883 (2020).
22. Enright MC, Spratt BG. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology (Reading, England)* **144 (Pt 11)**, 3049-3060 (1998).
23. Candeias C, Félix S, Handem S, de Lencastre H, Sá-Leão R. Clonal Changes in the Pneumococcal Population Carried by Portuguese Children during Six Years of Private Use of the 13-Valent Pneumococcal Conjugate Vaccine: the Relative Contribution of Clonal Expansion, Clonal Emergence, and Capsular Switch Events. *Microbiology spectrum* **11**, e0290922 (2023).
24. Ruiz García Y, Nieto Guevara J, Izurieta P, Vojtek I, Ortega-Barría E, Guzman-Holst A. CIRCULATING CLONAL COMPLEXES AND SEQUENCE TYPES OF STREPTOCOCCUS PNEUMONIAE SEROTYPE 19A WORLDWIDE: THE IMPORTANCE OF MULTIDRUG RESISTANCE: A SYSTEMATIC LITERATURE REVIEW. *Expert review of vaccines* **20**, 45-57 (2021).
25. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* **3**, 124 (2018).
26. Brueggemann AB, Griffiths DT, Meats E, Peto T, Crook DW, Spratt BG. Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. *The Journal of infectious diseases* **187**, 1424-1432 (2003).
27. Ganaie F*, et al.* A New Pneumococcal Capsule Type, 10D, is the 100th Serotype and Has a Large cps Fragment from an Oral Streptococcus. *mBio* **11**, (2020).
28. Makarewicz O*, et al.* Whole Genome Sequencing of 39 Invasive Streptococcus pneumoniae Sequence Type 199 Isolates Revealed Switches from Serotype 19A to 15B. *PloS one* **12**, e0169370 (2017).
29. van Tonder AJ*, et al.* Putatively novel serotypes and the potential for reduced vaccine effectiveness: capsular locus diversity revealed among 5405 pneumococcal genomes. *Microbial genomics* **2**, 000090 (2016).
30. Manna S*, et al.* Variants of Streptococcus pneumoniae Serotype 14 from Papua New Guinea with the Potential to Be Mistyped and Escape Vaccine-Induced Protection. *Microbiology spectrum* **10**, e0152422 (2022).
31. DePristo MA*, et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498 (2011).
32. van Tonder AJ*, et al.* Putative novel cps loci in a large global collection of pneumococci. *Microbial genomics* **5**, (2019).
33. Crisafulli G*, 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* **13**, 151-161 (2013).
34. Calix JJ*, et al.* Biochemical, genetic, and serological characterization of two capsule subtypes among Streptococcus pneumoniae Serotype 20 strains: discovery of a new

704 pneumococcal serotype. *The Journal of biological chemistry* **287**, 27885-27894
705 (2012).
706 35. Geno KA, Bush CA, Wang M, Jin C, Nahm MH, Yang J. WciG O-Acetyltransferase
707 Functionality Differentiates Pneumococcal Serotypes 35C and 42. *Journal of clinical*
708 *microbiology* **55**, 2775-2784 (2017).
709 36. Manna S*, et al.* A novel genetic variant of Streptococcus pneumoniae serotype 11A
710 discovered in Fiji. *Clinical microbiology and infection : the official publication of the*
711 *European Society of Clinical Microbiology and Infectious Diseases* **24**,
712 428.e421-428.e427 (2018).
713 37. Amin-Chowdhury Z*, et al.* Invasive pneumococcal disease due to 22F and 33F in
714 England: A tail of two serotypes. *Vaccine* **39**, 1997-2004 (2021).
715 38. Enright MC, Spratt BG. A multilocus sequence typing scheme for Streptococcus
716 pneumoniae: identification of clones associated with serious invasive disease.
717 *Microbiology* **144**, 3049-3060 (1998).
718 39. Croucher NJ*, et al.* Selective and genetic constraints on pneumococcal serotype
719 switching. *PLoS genetics* **11**, e1005095 (2015).
720 40. Beddek AJ, Li MS, Kroll JS, Jordan TW, Martin DR. Evidence for capsule switching
721 between carried and disease-causing Neisseria meningitidis strains. *Infection and*
722 *immunity* **77**, 2989-2994 (2009).
723 41. Bohlin J, Brynildsrud OB, Sekse C, Snipen L. An evolutionary analysis of genome
724 expansion and pathogenicity in Escherichia coli. *BMC genomics* **15**, 882 (2014).
725 42. Croucher NJ*, et al.* Rapid pneumococcal evolution in response to clinical interventions.
726 *Science (New York, NY)* **331**, 430-434 (2011).
727 43. Bikard D, Hatoum-Aslan A, Mucida D, Marraffini LA. CRISPR interference can prevent
728 natural transformation and virulence acquisition during in vivo bacterial infection. *Cell*
729 *host & microbe* **12**, 177-186 (2012).
730 44. Elberse K*, et al.* Sequence diversity within the capsular genes of Streptococcus
731 pneumoniae serogroup 6 and 19. *PLoS One* **6**, e25018 (2011).
732 45. Geno KA, Saad JS, Nahm MH. Discovery of Novel Pneumococcal Serotype 35D, a
733 Natural WciG-Deficient Variant of Serotype 35B. *J Clin Microbiol* **55**, 1416-1425
734 (2017).
735 46. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford,*
736 *England)* **30**, 2068-2069 (2014).
737 47. Burton RL, Geno KA, Saad JS, Nahm MH. Pneumococcus with the "6E" cps Locus
738 Produces Serotype 6B Capsular Polysaccharide. *J Clin Microbiol* **54**, 967-971 (2016).
739 48. Kapatai G*, et al.* Pneumococcal 23B Molecular Subtype Identified Using Whole
740 Genome Sequencing. *Genome Biol Evol* **9**, 2122-2135 (2017).
741 49. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the
742 population level. *BMC Bioinformatics* **11**, 595 (2010).
743 50. Li X, Chen F, Chen Y. Gcluster: a simple-to-use tool for visualizing and comparing
744 genome contexts for numerous genomes. *Bioinformatics* **36**, 3871-3873 (2020).
745 51. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal:
746 prokaryotic gene recognition and translation initiation site identification. *BMC*
747 *Bioinformatics* **11**, 119 (2010).

52. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).
53. 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 research* **25**, 1043-1055 (2015).
54. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)* **29**, 1072-1075 (2013).
55. Chen MY*, et al.* Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. *ISME J* **15**, 211-227 (2021).
56. 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* **31**, 1674-1676 (2015).
57. Francisco AP, Bugalho M, Ramirez M, Carrico JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics* **10**, 152 (2009).
58. 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* **33**, 128-129 (2017).