

Disease-Associated *Streptococcus pneumoniae* Genetic Variation

Shimin Yang,¹ Jianyu Chen,¹ Jinjian Fu,¹ Jiayin Huang, Ting Li, Zhenjiang Yao, Xiaohua Ye

Streptococcus pneumoniae is an opportunistic pathogen that causes substantial illness and death among children worldwide. The genetic backgrounds of pneumococci that cause infection versus asymptomatic carriage vary substantially. To determine the evolutionary mechanisms of opportunistic pathogenicity, we conducted a genomic surveillance study in China. We collected 783 *S. pneumoniae* isolates from infected and asymptomatic children. By using a 2-stage genomewide association study process, we compared genomic differences between infection and carriage isolates to address genomic variation associated with pathogenicity. We identified 8 consensus k-mers associated with adherence, antimicrobial resistance, and immune modulation, which were unevenly distributed in the infection isolates. Classification accuracy of the best k-mer predictor for *S. pneumoniae* infection was good, giving a simple target for predicting pathogenic isolates. Our findings suggest that *S. pneumoniae* pathogenicity is complex and multifactorial, and we provide genetic evidence for precise targeted interventions.

Streptococcus pneumoniae is a pathogen that causes community-associated infections in young children <5 years of age (1,2). It can asymptotically colonize the nasopharynx and upper airway in healthy children (up to 60%) and can also invade sterile sites and lead to infections from mild to life-threatening, which can result in substantial illness and death worldwide (1,3,4). Despite the widespread use of pneumococcal vaccines to immunize children, *S. pneumoniae* remains the leading cause of life-threatening diseases. Worldwide, the increasing disease burden of *S. pneumoniae* is alarming; an estimated 1 million children <5 years of age die of pneumococcal disease every year (5). All pneumococcal diseases arise from bacterial colonization, and the adaptability of the virulence characteristics

enhances pneumococcal persistence in colonization of the host respiratory tract, suggesting that nasopharyngeal carriage of *S. pneumoniae* plays a key role in development and transmission of pneumococcal diseases (6). Pneumococcal disease is one of the most common infectious diseases caused by asymptomatic *S. pneumoniae* colonization in humans. Eliminating this opportunistic pathogenic bacterium requires knowledge of the pathogenicity-associated genetic elements that distinguish infection from carriage isolates. Previous studies have been limited to exploring virulence factors and molecular characterization of invasive *S. pneumoniae* isolates (7,8).

Whole-genome sequencing (WGS) has become a powerful tool for bacterial genotyping; costs have been decreasing as accessibility increases. The high-dimensional genomic data can provide unprecedented resolution for identifying subtle genomic variations. Genomewide association studies (GWAS) are increasingly used to detect novel genes and genetic elements associated with bacterial phenotypes, which may provide insight for future preventive strategies and control measures (9–12). In brief, traditional GWAS methods can be used to identify large numbers of common genetic variants, usually single-nucleotide polymorphisms (SNPs), to determine the genetic basis of bacterial phenotypes of interest. However, considering the high genomic plasticity of many species of bacteria, traditional GWAS methods can only partially identify the phenotype-associated genetic variants. To avoid the limitations of SNP-based GWAS, we used k-mers (DNA words of length k) as an alternative method, which can capture different types of variants (13,14).

To determine whether genetic variation is unevenly enriched in *S. pneumoniae* infection isolates, we used multiple GWAS analyses to compare genomic differences between infection and carriage isolates. Study protocols were approved by the Ethics Committee of Guangdong Pharmaceutical

Author affiliations: Guangdong Pharmaceutical University, Guangzhou, China (S. Yang, J. Chen, J. Huang, T. Li, Z. Yao, X. Ye); Liuzhou Maternity and Child Health Care Hospital, Liuzhou, China (J. Fu)

DOI: <https://doi.org/10.3201/eid3001.221927>

¹These authors contributed equally to this article.

University (2019–19) and the Ethics Committee of Liuzhou Maternity and Child Healthcare Hospital (2018–84). We obtained written informed consent from parents or legal guardians on behalf of the children.

Methods

Sampling

During 2015–2021, we collected clinical samples from infected children and nasal swab samples from healthy children in southern China (Guangxi and Guangdong Provinces). From hospitalized infected children, we collected 349 nonrepetitive pneumococcal isolates (e.g., blood, bronchoalveolar lavage fluid, sputum, middle ear fluid), of which 342 were noninvasive and 7 invasive. The eligibility criteria for infected children were having clinical infectious manifestations such as cough, respiratory secretions, abnormal lung sounds, dyspnea, or fever $>38^{\circ}\text{C}$, with or without infiltrates seen on chest radiographs; having *S. pneumoniae* infection diagnosed by clinical doctors on the basis of signs and symptoms; and having *S. pneumoniae* isolated from clinical infection sites. In terms of asymptomatic carriage isolates, we sampled 434 isolates from healthy children in kindergarten.

Whole-Genome Sequencing

We performed high-throughput genome sequencing on a HiSeq 2000 machine (Illumina, <https://www.illumina.com>) to obtain paired-end 150-bp reads. We assessed the quality of the raw sequenced reads by using FastQC version 0.11.5 (<https://github.com/s-andrews/FastQC>) and trimmed for low quality reads and adaptor regions by using Trimmomatic version 0.36 (<https://github.com/usadellab/Trimmomatic>). We then assembled trimmed reads by using SPAdes version 3.6.1 (<https://github.com/ablab/spades>). We used PathogenWatch (<https://pathogen.watch>) to predict global pneumococcal sequencing cluster (GPSC), multilocus sequence typing (MLST), and serotyping for all genomes.

Phylogenetic Analyses

To generate the variant sites with SNPs, we mapped assembled contigs to a standard reference genome *S. pneumoniae* R6 by using Snippy version 4.4.5 (<https://github.com/tseemann/snippy>). We used the generated core SNP alignment to construct a maximum-likelihood phylogenetic tree by using the generalized time reversible plus gamma model and 100 bootstrap replicates with FastTree version 2.1.10

(<http://www.microbesonline.org/fasttree>). We visualized and annotated the phylogenetic tree by using ChiPlot (<https://www.chiplot.online>).

Counting and Annotating k-mers

We scanned all k-mers that were 9- to 100-bp long from all assembled reads by using fsm-lite (<https://github.com/nvalimak/fsm-lite>) and filtered them to obtain 10,591,337 k-mers seen on 1%–99% of the total samples. To identify the relevant genes by using BWA-MEM (the Burrows-Wheeler Aligner with maximal exact matches alignment tool, <https://github.com/lh3/bwa>), we mapped all k-mers to 10 *S. pneumoniae* reference genomes (CGSP14, D39, Hungary^{19A}-6, R6, Taiwan^{19F}-14, TIGR4, Spain^{23F}-ST81, ATCC 49619, EF3030, and MDRSPN001) obtained from the Virulence Factor Database (<http://www.mgc.ac.cn/VFs>) and previous studies. We determined gene ontology annotations by using the UniProt (<https://beta.uniprot.org>).

Multiple GWAS Analyses of Disease-Associated k-mers

To explore the genomewide associations between genetic elements (k-mers) and *S. pneumoniae* disease status (infection or carriage), and thus to identify infection-associated k-mers, we used GWAS methods. Because of the high-dimensional genomic data structures, we used multiple GWAS methods: the linear mixed model (LMM; (<https://github.com/mgalardini/pyseer>), phylogenetic-based approach (Scoary; <https://github.com/AdmiralenOla/Scoary>), variable selection using random forests (VSURF; <https://github.com/robingenuer/VSURF>), and least absolute shrinkage and selection operator (LASSO; https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html) regression.

In brief, we used a 2-stage analysis process to detect the infection-associated k-mers by comprehensive GWAS analyses (Figure 1). First, we fitted a univariate LMM to initially screen infection-associated k-mers by using the Pyseer tool (version 1.3.10) (15). To correct for the population structure, we used the similarity pyseer command of Pyseer, which computes a similarity kinship matrix on the basis of the core genome SNPs. For covariates, the GWAS analysis used host age (years) and sex. Second, we used multiple methods (Scoary, LASSO, and VSURF) to minimize false-positive associations and identify consensus infection-associated k-mers by Venn diagram. In the GWAS analyses, we used the Bonferroni correction (α/N) to control for

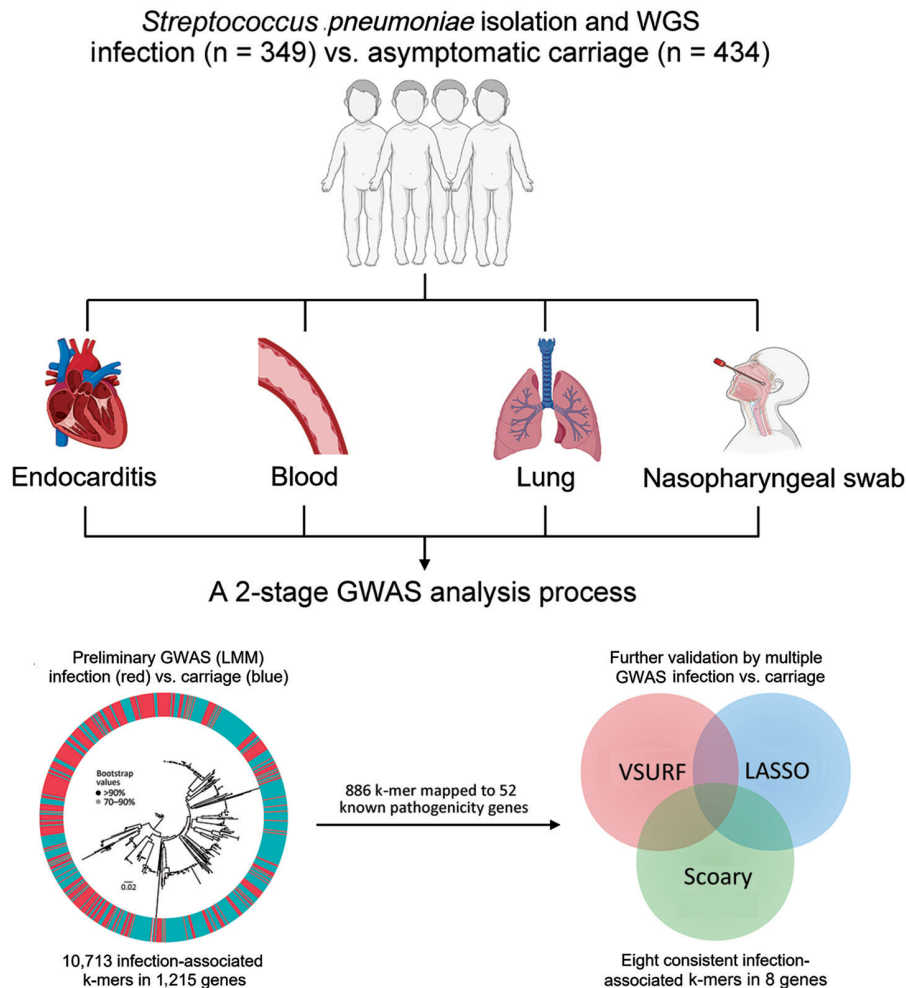


Figure 1. Two-stage GWAS analysis process used to detect infection-associated *Streptococcus pneumoniae* k-mers in study of disease-associated *Streptococcus pneumoniae* genetic variation. GWAS, genome-wide association studies; LASSO, least absolute shrinkage and selection operator; LMM, linear mixed model; VSURF, variable selection using random forests; WGS, whole-genome sequencing.

false-positive rates resulting from multiple comparisons of 1,418,815 k-mers (adjusted p value threshold 3.52×10^{-8}). Scoary is an ultrafast software tool for GWAS analyses that uses a phylogenetic-based method to adjust population structure. The LASSO regression is suitable for high-dimensional data structures, and the coefficients of nonrelevant variables can be compressed to zero to solve the problem of model overfitting (16). We used VSURF, based on random forest (RF), to perform a 2-step feature selection on the variables (17). Initially, VSURF ranks the variables according to the importance measure by using the RF permutation-based score of importance to obtain a subset of important variables, and then it uses a stepwise forward strategy for variable introduction based on the smallest out-of-bag error. More precisely, a variable is added only if the error decrease is larger than a threshold. We ranked the importance of k-mers by the mean decrease in impurity (mean decrease Gini), which is a measure of the predictor's contribution to the correct sample

classification. We compiled associated phenotype data for all 783 isolates (Appendix Table 1, <https://wwwnc.cdc.gov/EID/article/30/1/22-1927-App1.pdf>) and deposited sequences in the National Center for Biotechnology Information Sequence Read Archive database (<https://www.ncbi.nlm.nih.gov/sra>; projection no. PRJNA976286). The k-mer sequences and output results files from several GWAS analyses are publicly available (<https://doi.org/10.6084/m9.figshare.24466606.v3>).

Results

Characteristics of *S. pneumoniae* Isolates

Of the 349 children with *S. pneumoniae* infection, 342 (98.0%) had noninvasive disease (264 pneumonia, 49 bronchitis, 13 otitis media, 9 upper respiratory infection, 6 nasosinusitis, and 1 corneal ulcer), and 7 (2.0%) had invasive disease (6 bacteremia and 1 endocarditis). χ^2 test results indicated no differences between infection and carriage isolates with regard to host sex

($p = 0.359$) but significant differences with regard to age ($p < 0.001$) (Appendix Table 2).

Association between Genotypes and Disease Status

The most prevalent GPSCs for infection isolates were GPSC1 (45.9%), GPSC321 (9.2%), and GPSC852 (5.4%); the predominant GPSCs for carriage isolates were GPSC321 (16.1%), GPSC1 (15.4%), and GPSC23 (15.0%). In terms of sequence types (STs), the most common genotypes for infection isolates were ST271 (29.2%), ST320 (9.5%), and ST902 (7.2%); the predominant genotypes for carriage isolates were ST902 (15.9%), ST90 (13.8%), and ST271 (8.5%). The most prevalent serotypes for infection isolates were 19F (43.0%), 6B (15.2%), and 23F (8.3%); and the predominant serotypes for carriage isolates were 6B (32.7%), 19F (13.1%), and 15A (11.1%). The results indicated potential genotype differences between infection and carriage isolates. In addition, the phylogenetic tree based on core SNPs revealed that several genotypes (GPSCs/STs/serotypes) from infection and carriage isolates clustered in the same branches (Figure 2). Moreover, we found statistically significant differences in the proportion of specific GPSCs/STs/serotypes between infection and carriage isolates (Table 1), indicating that these isolates are associated with infection.

Preliminary Screening for Infection-Associated K-Mers by LMM

We identified 10,591,337 k-mers from the assemblies of 783 *S. pneumoniae* isolates and then filtered out low-frequency k-mers for a reduced matrix with 1,418,815 k-mers. Using those k-mers for GWAS, we performed a univariate LMM analysis to initially identify 22,790 infection-associated k-mers; 10,713 k-mers were successfully mapped to 1,215 unique genes (Figure 3, panel A; Appendix Figure 1). In the initial model with 10,713 k-mers, we used the RF model to assess the prediction effect. The classification balanced accuracy based on cross-validation was 93.60% (95% CI 91.48%–95.72%) (Table 2); the area under the curve (AUC), based on the out-of-bag risk scores of the classifier, was 0.98. In the LMM analysis, the QQ-plot indicated that population structure was well controlled at low p values ($p < 0.01$) (Appendix Figure 2). Because of the considerable redundancy among the genetic elements in risk prediction, studying all k-mer combinations had little benefit; therefore, we used a simpler model with 886 k-mers successfully mapped to 52 antibiotic resistance or virulence genes (Appendix Table 3). The classification balanced accuracy was 91.28% (95% CI 89.34%–93.22%) (Table 2); the AUC was 0.96, suggesting that the power of these 886 k-mers for

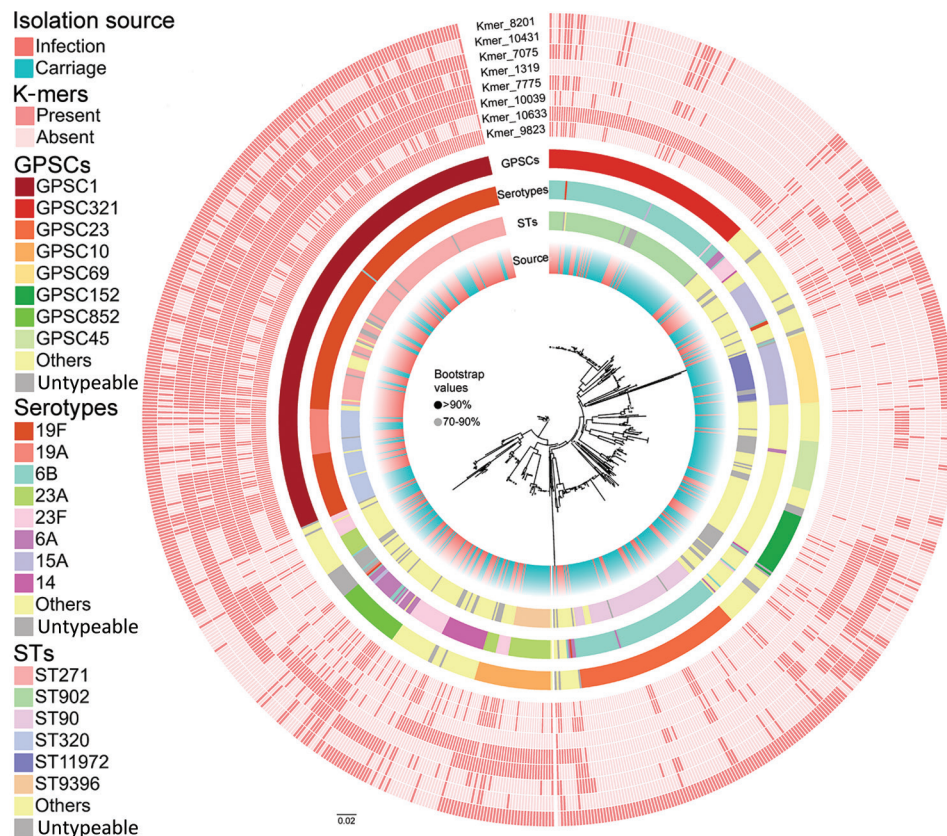


Figure 2. Whole-genome phylogenetic tree showing genetic similarity of 783 *Streptococcus pneumoniae* isolates in a study of disease-associated *Streptococcus pneumoniae* genetic variation. The colored strips at the tips of the tree (from inner to outer) represent isolate metadata (source, STs, serotypes, and GPSCs) and infection-associated k-mers found in the final model. GPSC, global pneumococcal sequencing cluster; ST, sequence type.

Table 1. Association analysis between dominant genotypes and disease status from study of disease-associated *Streptococcus pneumoniae* genetic variation*

Genotype	Infection isolates, no. (%) (n = 349)	Carriage Isolates, no. (%) (n = 434)	χ^2	p value	OR (95% CI)
ST					
ST271	102 (29.2)	37 (8.5)	56.78	<0.001	4.43 (2.95–6.67)
ST902	25 (7.2)	69 (15.9)	13.97	<0.001	0.41 (0.25–0.66)
ST90	13 (3.7)	60 (13.8)	23.34	<0.001	0.24 (0.13–0.45)
ST320	33 (9.5)	24 (5.5)	4.42	0.036	1.78 (1.03–3.08)
ST11972	3 (0.9)	24 (5.5)	12.67	<0.001	0.15 (0.04–0.50)
ST9396	1 (0.3)	22 (5.1)	15.52	<0.001	0.05 (0.01–0.40)
Serotype					
19F	150 (43.0)	57 (13.1)	88.61	<0.001	4.99 (3.51–7.08)
6B	53 (15.2)	142 (32.7)	31.80	<0.001	0.37 (0.26–0.53)
15A	12 (3.4)	48 (11.1)	15.88	<0.001	0.29 (0.15–0.55)
23F	29 (8.3)	21 (4.8)	3.90	0.056	1.78 (0.96–3.35)
23A	10 (2.9)	32 (7.4)	7.74	0.005	0.37 (0.18–0.77)
6A	24 (6.9)	10 (2.3)	9.74	0.002	3.13 (1.48–6.64)
19A	15 (4.3)	11 (2.5)	1.87	0.171	1.73 (0.78–3.81)
14	15 (4.3)	11 (2.5)	1.87	0.171	1.73 (0.78–3.81)
GPSC					
GPSC1	160 (45.9)	67 (15.4)	88.88	<0.001	4.64 (3.32–6.48)
GPSC321	32 (9.2)	70 (16.1)	8.27	0.004	0.52 (0.34–0.82)
GPSC23	15 (4.3)	65 (15.0)	24.05	<0.001	0.25 (0.14–0.45)
GPSC10	9 (2.6)	28 (6.5)	6.44	0.011	0.38 (0.18–0.81)
GPSC69	3 (0.9)	31 (7.1)	18.39	<0.001	0.11 (0.04–0.35)
GPSC152	13 (3.7)	18 (4.2)	0.09	0.763	0.89 (0.44–1.83)
GPSC852	19 (5.4)	12 (2.8)	3.65	0.056	2.02 (0.98–4.17)

*Boldface indicates statistical significance. GPSC, global pneumococcal sequencing cluster; OR, odds ratio; ST, sequence type.

predicting disease status was close to that of the model with 10,713 k-mers.

In addition, we sorted the 886 disease-associated k-mers according to estimated importance (Figure 3, panel B). The k-mers were mainly associated with antimicrobial resistance (34%), adherence (20%), immune modulation (17%), and exoenzyme (1%). Moreover, the k-mers were divided into 3 functional gene ontology categories. Among those categories, proteolysis and cell wall organization were the largest subcategories in the biological process, membrane was the most enriched term in the cellular component, and serine-type D-Ala-D-Ala carboxypeptidase activity was the top term in the molecular function (Figure 3, panel C).

Further Validation of Infection-Associated k-mers by Multiple GWAS Analyses

To reduce the complexity of the model, we used 3 methods to identify consensus infection-associated k-mers (Figure 4). On the basis of the 886 k-mers screened above, we observed consensus on genomewide statistically significant associations for pathogenicity k-mers; 8 k-mers were identified by all 3 methods. When we used the simplest model with the 8 k-mers, the classification balanced accuracy was 90.89% (95% CI 89.48%–92.31%) (Table 2), and the AUC value was 0.93 (Figure 5, panel A), suggesting that the power of the 8 k-mers to predict disease status was comparable to that of the model with 886 k-mers. Of note, the k-mer predictors

still exhibited high classification balanced accuracy in the predominant GPSCs (95.34% for GPSC1 and 92.79% for GPSC321). The importance of the selected k-mers in the final model indicated that these predictors were mainly associated with adherence function (Figure 5, panel B). The highest ranked predictor (Kmer_9823 in sortase [*srtG1*]) achieved a classification accuracy of 79.57% on its own and also showed high classification accuracy in the predominant GPSCs (70.04% for GPSC1 and 85.29% for GPSC321). In addition, the best predictor (in *srtG1*) was associated with GPSC1 and GPSC321 (all $p < 0.05$). For the additional validation analysis that used the best RF classifier k-mer (in *srtG1*), 2 independent datasets of *S. pneumoniae* genomes with genotype distribution similar to that of our study were available on the National Center for Biotechnology Information Assembly database (<https://www.ncbi.nlm.nih.gov/assembly>) (data1: 60 noninvasive vs. 60 carriage isolates; data2: 60 invasive versus 60 carriage isolates; the prevalence of the predominant GPSCs [GPSC1 and GPSC321] was 58.3% for noninvasive, 55.0% for invasive and 30.0% for carriage isolates) (Appendix Table 4). Classification accuracy was 75.83% for data1 and 74.17% for data2, similar to that in the larger primary dataset in our study.

The proportion of k-mers differed significantly between infection and carriage isolates (all $p < 0.05$) (Figure 5, panel C), indicating that the proportion of k-mers was substantially higher in infection

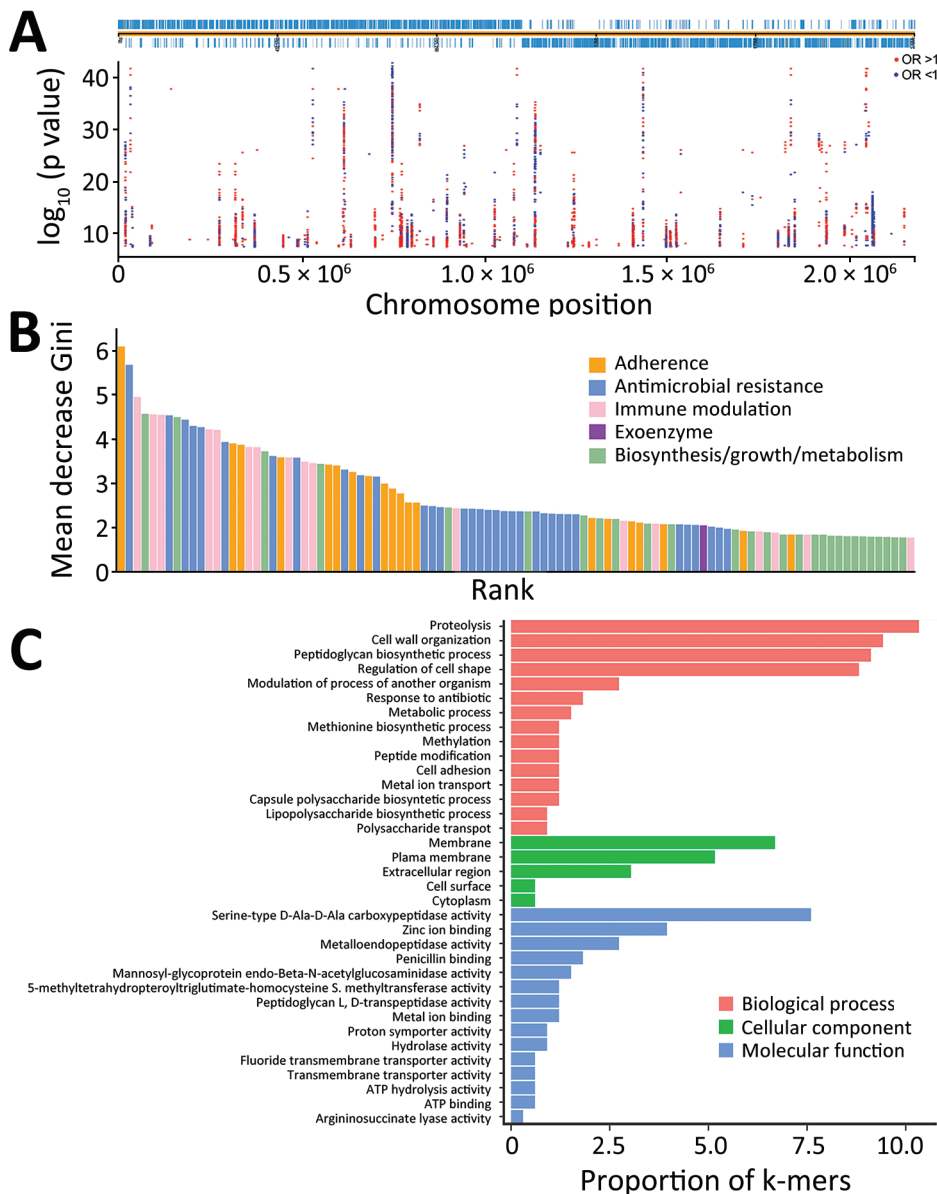


Figure 3. Preliminary screening for infection-associated k-mers by linear mixed model in study of disease-associated *Streptococcus pneumoniae* genetic variation. A) Manhattan plot showing statistical significance and chromosomal location of k-mers mapped to a complete reference genome (TIGR4; GenBank accession no. NC_003028.3). B) Importance of the top 100 k-mer predictors in a simpler model with 886 k-mers. C) Gene ontology annotations of the top 100 k-mer predictors. OR, odds ratio.

isolates than in carriage isolates. The effect of each k-mer on the estimated risk score (Figure 5, panel D), indicated by a point above the diagonal, indicates that the risk score is increased when the k-mer profile is present. The presence of k-mers associated with adherence genes markedly increased the risk for *S. pneumoniae* infection (odds ratio [OR] 1.88 for Kmer_9823, OR 1.65 for Kmer_10039, and OR 1.69 for Kmer_10431) (Table 3).

Discussion

To explore genomic differences between infection and carriage isolates, linking infection-associated genotypes with disease status is necessary. In our

study, the most common serotypes for infection isolates (19F, 6B, 23F) were consistent with the results from other regions of China (18–20) but differed from those from the United States and Japan (21,22). Moreover, we observed considerable ST diversity among infection isolates; the most prevalent genotypes were ST271, ST320, and ST902, a finding consistent with those of previous studies in China but different from those in developed and developing countries (23–25). The resolution of MLST and serotyping for inferring isolate relatedness is limited, so we also used GPSCs to characterize and compare different lineages (26). The most prevalent GPSCs among the infection isolates were GPSC1, GPSC321, and GPSC852, which

Table 2. Resubstitution estimate and cross-validation results based on random forest models used in study of disease-associated *Streptococcus pneumoniae* genetic variation*

Evaluation indicators	10,713 k-mer predictors		886 k-mer predictors		8 k-mer predictors	
	Resubstitution estimate	10-fold cross-validation estimate	Resubstitution estimate	10-fold cross-validation estimate	Resubstitution estimate	10-fold cross-validation estimate
Accuracy	98.60	93.23	96.42	90.81	90.93	90.04
Balanced accuracy	98.65	93.60	96.61	91.28	91.48	90.89
Sensitivity	99.13	94.48	97.91	92.87	94.27	93.72
Specificity	98.18	92.71	95.31	89.69	88.70	88.07
PPV	97.71	90.27	93.98	86.27	84.81	83.65
NPV	99.31	95.63	98.39	94.48	95.85	95.18
Kappa	0.97	0.86	0.93	0.81	0.81	0.80

*Values are percentages except for kappa, which is reported as a value ranging from -1 to 1. NPV, negative predictive value; PPV, positive predictive value.

differed from those in the United States and South Africa (27). Our findings suggest that discrepancy in genotypes on a global scale may be associated with different pathogenicity and evolutionary directions. In our study, associations between specific genotypes (such as 19F and GPSC1) and disease status differed significantly, which is consistent with findings from a study in India (28). Our findings indicate that the presence of specific pathogenic clones may promote infection. In a simple pathogenicity model, all pathogenic clones would belong to specific clusters of genetically related disease-causing isolates (i.e., pathogenic clone hypothesis; Figure 6, panel A), which has been observed for *Staphylococcus aureus* and *S. pneumoniae* isolates (29,30). That pathogenicity model is not suitable for all *S. pneumoniae* clones because many infection isolates clustered in the same branches of phylogenetic tree as carriage isolates. In addition, traditional genotypes provide little power for identifying small genetic variations at the genomic level (29), suggesting that those genotypes only partially explain the pathogenicity of *S. pneumoniae*.

Using high-throughput genome sequencing technologies and bacterial GWAS methods to further explore high-dimensional genetic variation between infection and carriage isolates is essential, thereby revealing the pathogenicity-associated genetic elements

of *S. pneumoniae*. According to the phylogenetic tree, we observed that infection isolates were markedly unevenly distributed across the phylogeny and also clustered with carriage isolates within several lineages, indicating that most lineages are capable of causing infection (i.e., opportunistic pathogenicity hypothesis; Figure 6, panel B). If this hypothesis is reasonable, then GWAS analyses would not detect numerous pathogenicity-associated k-mers. However, the LMM-based GWAS in our study detected 22,790 pathogenicity-associated k-mers. These findings suggest that the enrichment of genetic elements encoding pathogenicity traits may increase the pathogenicity of *S. pneumoniae* (i.e., pathogenic-determinant hypothesis; Figure 6, panel C), which is consistent with *Staphylococcus epidermidis* and avian pathogenic *Escherichia coli* (30,31). In this pathogenic-determinant model, horizontal gene transfer could spread genetic determinants in bacteria such as *S. pneumoniae* and *Klebsiella pneumoniae* (32–34), leading various clones to successfully cause disease.

High-throughput genomic data have brought substantial challenges to data analysis because of high-dimensional and highly correlated data structures. In our study, we identified infection-associated k-mers by using a 2-stage comprehensive GWAS analysis process, including LMM for initially screening

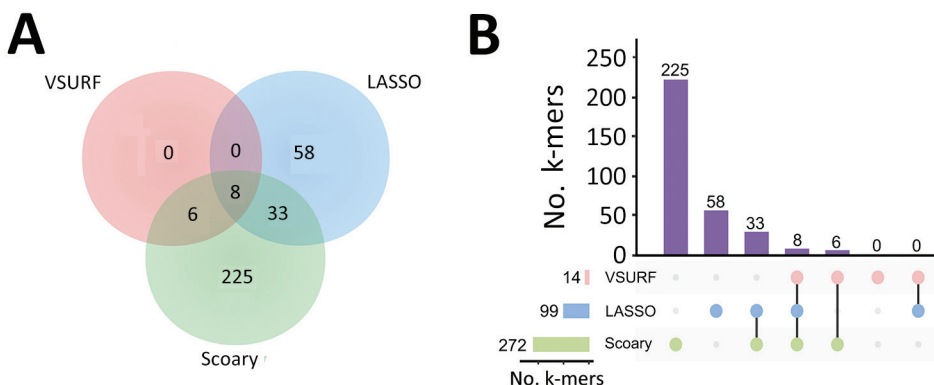


Figure 4. Further validation of infection-associated k-mers by multiple GWAS analyses in study of disease-associated *Streptococcus pneumoniae* genetic variation. A) Venn diagram visualization of the k-mers identified by 3 methods. B) UpSet plot visualization of the k-mers identified by 3 methods. LASSO, least absolute shrinkage and selection operator; VSURF, variable selection using random forests.

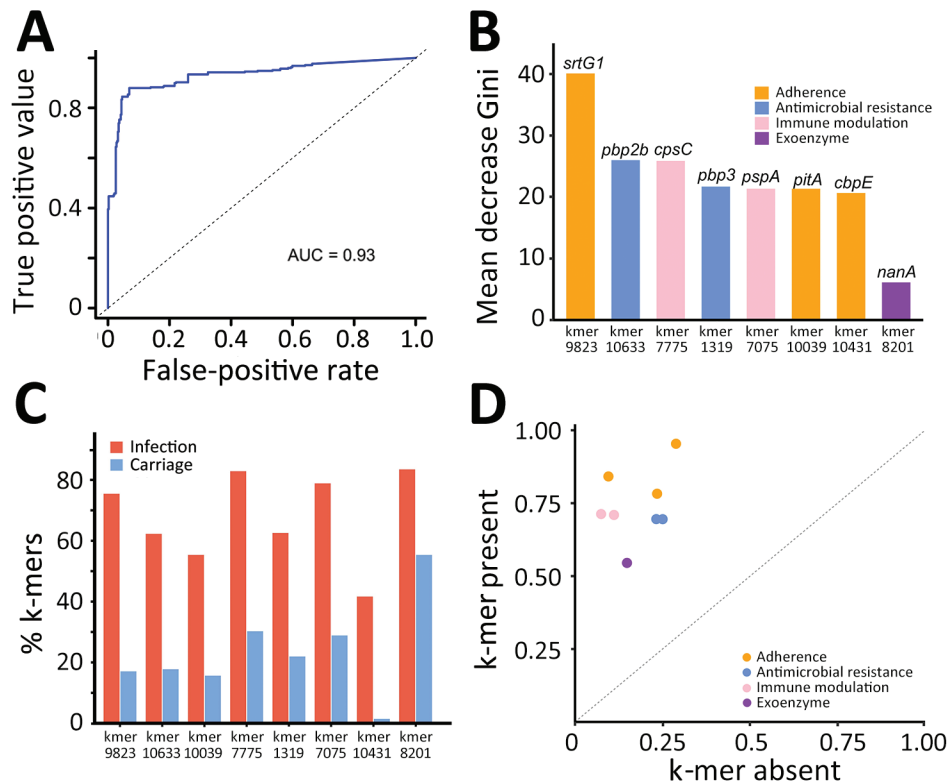


Figure 5. Prediction effect of the 8 k-mers identified in the final model used in study of disease-associated *Streptococcus pneumoniae* genetic variation. A) Receiver operating characteristic curve of the final model. B) Predictor importance of the 8 k-mers in the final model. C) Proportion of k-mer predictors between infection and carriage isolates. D) Change in risk score for a specific k-mer profile when the k-mer is present (y-axis) compared to absent (x-axis). AUC, area under the curve.

pathogenic k-mers and multiple GWAS methods for further validation. In the final prediction model, we identified 8 k-mer predictors, which mapped to genes associated with adherence, immune regulation, antibiotic resistance, and exoenzyme. Of the adherence-related genes, *srtG1* and the LPxTG-type surface-anchored protein (*pitA*) are important components of the pneumococcal pilus-2, which plays a crucial role in promoting adhesion, colonization, and cellular invasion (35,36). Classification accuracy of the most important k-mer in *srtG1* was high by itself, and that of the additional validation RF analysis based on open datasets was similar, suggesting that this predictor has great potential for predicting pathogenic isolates in a clinical setting. Phosphorylcholine esterase (*cbpE*) plays an important role in modulating both the

phosphorylcholine decoration of its surface and choline-bound surface adhesins, which may contribute to pneumococcal adherence and invasiveness (37). Capsular polysaccharide (CPS) is a major virulence factor in *S. pneumoniae*. Capsular polysaccharide protein C (CpsC) has been shown to affect the level of CPS expression and also regulate the assembly, export, and attachment of CPS to the cell wall (38). Pneumococcal surface protein A (PspA) plays role in preventing complement-mediated opsonization and is also capable of binding to lactoferrin, thereby preventing it from killing pneumococci (39). The infection-associated genes reported in our study (*cpsC* and *pspA*) are homologous to the genes associated with invasive pneumococci (*cpsA*, *cpsD*, and *pspC*) identified in previous studies (11,12), providing more evidence for

Table 3. Association analysis between k-mers and disease status used in study of disease-associated *Streptococcus pneumoniae* genetic variation*

k-mer	Genes	Infection isolates, no. (%), n = 349	Carriage isolates, no. (%), n = 434	p value	OR (95%CI)
Kmer_9823	<i>srtG1</i>	264 (75.6)	75 (17.3)	8.55×10^{-45}	1.88 (1.79–1.96)
Kmer_10633	<i>pbp2b</i>	218 (62.5)	78 (18.0)	2.68×10^{-37}	1.77 (1.71–1.84)
Kmer_10039	<i>pitA</i>	194 (55.6)	69 (15.9)	1.47×10^{-31}	1.65 (1.51–1.79)
Kmer_7775	<i>cpsC</i>	290 (83.1)	132 (30.4)	6.59×10^{-49}	1.75 (1.66–1.83)
Kmer_1319	<i>pbp3</i>	219 (62.8)	96 (22.1)	9.95×10^{-31}	1.79 (1.70–1.87)
Kmer_7075	<i>pspA</i>	276 (79.1)	126 (29.0)	4.31×10^{-44}	1.64 (1.55–1.72)
Kmer_10431	<i>cbpE</i>	146 (41.8)	7 (1.6)	3.38×10^{-45}	1.69 (1.62–1.76)
Kmer_8201	<i>nanA</i>	292 (83.7)	241 (55.5)	4.68×10^{-17}	1.66 (1.55–1.76)

*OR, odds ratio.

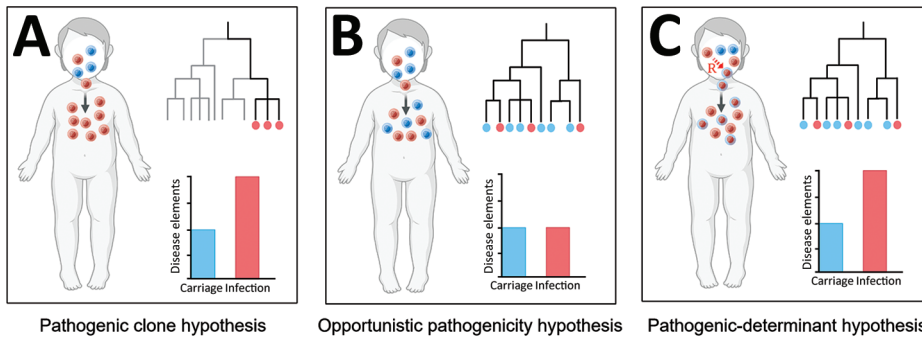


Figure 6. Pathogenicity models for genetically related disease-causing isolates used in study of disease-associated *Streptococcus pneumoniae* genetic variation. A) Pathogenic clone hypothesis; B) opportunistic pathogenicity hypothesis; C) pathogenic-determinant hypothesis.

S. pneumoniae pathogenicity. Neuraminidase A encoded by the *nanA* gene is an essential colonization factor for *S. pneumoniae* and promotes growth and survival of the bacteria in the upper respiratory tract (40). Antimicrobial drug use and abuse not only induce widespread multidrug-resistant pneumococci but also increase the susceptibility to invasive disease (41). For decades, penicillin has been the first choice for treatment of pneumococcal infection, and mutations in penicillin-binding proteins (PBPs) are essential for high-level penicillin resistance (42). Li et al. demonstrated that *pbp2b* and *pbp3* are associated with pneumococcal infection (42). One reason is that PBP2B and PBP3 are involved in the synthesis and growth of bacterial cell walls, which are crucial for the survival and virulence of pneumococci (43). In addition, a previous study revealed a potential association between penicillin resistance and GPSC1 (44), and our findings also showed that GPSC1 was associated with pneumococcal infection, suggesting that it cannot support a causal link between resistance and pneumococcal infection and may result from a lineage confounder. In summary, these infection-associated k-mers provide genetic evidence for revealing optimal risk factors for infection isolates, which may offer a theoretical basis for precise targeted interventions.

In this study, we attempted to use the comprehensive analysis strategy to identify pathogenic k-mers by well-characterized *S. pneumoniae* isolates from a single location so we could reduce redundancy of k-mer predictors, minimize false-positive associations, and avoid geographic variation. Our consensus findings of pathogenic k-mers from multiple GWAS methods may provide sufficient evidence for clarifying the complex multifactorial pathogenicity of *S. pneumoniae*. However, among the potential limitations, the first is that *S. pneumoniae* pathogenesis is a multifactorial and interacting process, but traditional GWAS methods identify the main effect of each genetic variation and ignore the complex gene-gene interactions (45). Therefore, future studies should use

the enrichment theory to determine the core functions or pathways for risk genes, which may provide new insights for understanding pathogenesis at functional levels (46,47). Second, although k-mers can reflect variation in bacterial genomes, we mapped infection-associated k-mers in our study to reference genomes to identify pathogenic genes, which cannot cover complete genomic variation in the entire species. To overcome those issues, we developed the extended k-mer-based GWAS methods to detect phenotype-specific k-mers without relying on prior annotations or reference genomes (48,49). Third, our study focused mainly on noninvasive rather than invasive isolates, and *S. pneumoniae* can transition from carriage to infection, suggesting potential similarity in carriage and noninvasive infection isolates. To improve the statistical power and comparability of exploring disease-associated markers, we included infection isolates from children with confirmed associated symptoms and carriage isolates from asymptomatic healthy children.

In conclusion, our 2-stage GWAS analyses identified a subset of 8 pathogenic k-mers associated with adherence, antimicrobial resistance, and immune modulation, indicating that the enrichment of genetic elements encoding pathogenicity traits may increase the pathogenicity of *S. pneumoniae*. The best predictor for *S. pneumoniae* infection achieved a high classification accuracy, giving a very simple target for predicting pathogenic isolates in a clinical setting. These findings suggest the complex multifactorial nature of *S. pneumoniae* pathogenicity and provide genetic evidence for the evolution of virulence and development of precise targeted interventions.

Acknowledgments

We sincerely thank all study children included in this study. We also thank the research staff and students at Guangdong Pharmaceutical University, China.

This work was supported by the National Natural Science Foundation of China (grant nos. 81973069 and 81602901),

the Guangdong Basic and Applied Basic Research Foundation (grant no. 2023A1515011583), and the Key Scientific Research Foundation of Guangdong Educational Committee (grant no. 2022ZDZX2033). The funders had no role in the study design, data collection, and analysis, or interpretation of the data.

About the Author

Ms. Yang is a graduate student at the Guangdong Pharmaceutical University, China. Her research interests include the pneumococcal disease, evolution of virulence, and genome-wide association study.

References

- Henriques-Normark B, Tuomanen EI. The pneumococcus: epidemiology, microbiology, and pathogenesis. *Cold Spring Harb Perspect Med*. 2013;3:a010215. <https://doi.org/10.1101/cshperspect.a010215>
- Mancuso G, Midiri A, Gerace E, Biondo C. Bacterial antibiotic resistance: the most critical pathogens. *Pathogens*. 2021;10:1310. <https://doi.org/10.3390/pathogens10101310>
- Subramanian K, Henriques-Normark B, Normark S. Emerging concepts in the pathogenesis of the *Streptococcus pneumoniae*: from nasopharyngeal colonizer to intracellular pathogen. *Cell Microbiol*. 2019;21:e13077. <https://doi.org/10.1111/cmi.13077>
- Bogaert D, De Groot R, Hermans PW. *Streptococcus pneumoniae* colonisation: the key to pneumococcal disease. *Lancet Infect Dis*. 2004;4:144–54. [https://doi.org/10.1016/S1473-3099\(04\)00938-7](https://doi.org/10.1016/S1473-3099(04)00938-7)
- Yao KH, Yang YH. *Streptococcus pneumoniae* diseases in Chinese children: past, present and future. *Vaccine*. 2008; 26:4425–33. <https://doi.org/10.1016/j.vaccine.2008.06.052>
- Kadioglu A, Weiser JN, Paton JC, Andrew PW. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat Rev Microbiol*. 2008;6:288–301. <https://doi.org/10.1038/nrmicro1871>
- Piet JR, Geldhoff M, van Schaik BD, Brouwer MC, Valls Seron M, Jakobs ME, et al. *Streptococcus pneumoniae* arginine synthesis genes promote growth and virulence in pneumococcal meningitis. *J Infect Dis*. 2014;209:1781–91. <https://doi.org/10.1093/infdis/jit818>
- Tunjungputri RN, Mobegi FM, Cremers AJ, van der Gaast-de Jongh CE, Ferwerda G, Meis JF, et al. Phage-derived protein induces increased platelet activation and is associated with mortality in patients with invasive pneumococcal disease. *MBio*. 2017;8:e01984–16. <https://doi.org/10.1128/mBio.01984-16>
- Chaguzo C, Ebruke C, Senghore M, Lo SW, Tientcheu PE, Gladstone RA, et al. Comparative genomics of disease and carriage serotype 1 pneumococci. *Genome Biol Evol*. 2022;14:evac052. <https://doi.org/10.1093/gbe/evac052>
- The CRYPTIC Consortium. Genome-wide association studies of global *Mycobacterium tuberculosis* resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. *PLoS Biol*. 2022;20:e3001755. <https://doi.org/10.1371/journal.pbio.3001755>
- Lees JA, Ferwerda B, Kremer PHC, Wheeler NE, Serón MV, Croucher NJ, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun*. 2019;10:2176. <https://doi.org/10.1038/s41467-019-09976-3>
- Obolski U, Gori A, Lourenço J, Thompson C, Thompson R, French N, et al. Identifying genes associated with invasive disease in *S. pneumoniae* by applying a machine learning approach to whole genome sequence typing data. *Sci Rep*. 2019;9:4049. <https://doi.org/10.1038/s41598-019-40346-7>
- Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat Commun*. 2016;7:12797. <https://doi.org/10.1038/ncomms12797>
- Gupta PK. Quantitative genetics: pan-genomes, SVs, and k-mers for GWAS. *Trends Genet*. 2021;37:868–71. <https://doi.org/10.1016/j.tig.2021.05.006>
- Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics*. 2018;34:4310–2. <https://doi.org/10.1093/bioinformatics/bty539>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30. <https://doi.org/10.48550/arXiv.1201.0490>
- Genuer R, Poggi J-M, Tuleau-Malot C. VSURF: an R package for variable selection using random forests. *R J*. 2015;7:19–33. <https://doi.org/10.32614/RJ-2015-018>
- Geng Q, Zhang T, Ding Y, Tao Y, Lin Y, Wang Y, et al. Molecular characterization and antimicrobial susceptibility of *Streptococcus pneumoniae* isolated from children hospitalized with respiratory infections in Suzhou, China. *PLoS One*. 2014;9:e93752. <https://doi.org/10.1371/journal.pone.0093752>
- Shi W, Li J, Dong F, Qian S, Liu G, Xu B, et al. Serotype distribution, antibiotic resistance pattern, and multilocus sequence types of invasive *Streptococcus pneumoniae* isolates in two tertiary pediatric hospitals in Beijing prior to PCV13 availability. *Expert Rev Vaccines*. 2019;18:89–94. <https://doi.org/10.1080/14760584.2019.1557523>
- Yu YY, Xie XH, Ren L, Deng Y, Gao Y, Zhang Y, et al. Epidemiological characteristics of nasopharyngeal *Streptococcus pneumoniae* strains among children with pneumonia in Chongqing, China. *Sci Rep*. 2019;9:3324. <https://doi.org/10.1038/s41598-019-40088-6>
- Suaya JA, Mendes RE, Singhs HL, Arguedas A, Reinert RR, Jodar L, et al. *Streptococcus pneumoniae* serotype distribution and antimicrobial nonsusceptibility trends among adults with pneumonia in the United States, 2009–2017. *J Infect*. 2020;81:557–66. <https://doi.org/10.1016/j.jinf.2020.07.035>
- Yanagihara K, Kosai K, Mikamo H, Mukae H, Takesue Y, Abe M, et al. Serotype distribution and antimicrobial susceptibility of *Streptococcus pneumoniae* associated with invasive pneumococcal disease among adults in Japan. *Int J Infect Dis*. 2021;102:260–8. <https://doi.org/10.1016/j.ijid.2020.10.017>
- Yan Z, Cui Y, Huang X, Lei S, Zhou W, Tong W, et al. Molecular characterization based on whole-genome sequencing of *Streptococcus pneumoniae* in children living in southwest China during 2017–2019. *Front Cell Infect Microbiol*. 2021;11:726740. <https://doi.org/10.3389/fcimb.2021.726740>
- Kellner JD, Ricketson LJ, Demczuk WHB, Martin I, Tyrrell GJ, Vanderkooi OG, et al. Whole-genome analysis of *Streptococcus pneumoniae* serotype 4 causing outbreak of invasive pneumococcal disease, Alberta, Canada. *Emerg Infect Dis*. 2021;27:1867–75. <https://doi.org/10.3201/eid2707.204403>
- Vorobieva S, Jensen V, Furberg AS, Slotved HC, Bazhukova T, Haldorsen B, Caugant DA, et al. Epidemiological and

- molecular characterization of *Streptococcus pneumoniae* carriage strains in pre-school children in Arkhangelsk, northern European Russia, prior to the introduction of conjugate pneumococcal vaccines. BMC Infect Dis. 2020;20:279. <https://doi.org/10.1186/s12879-020-04998-5>
26. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, et al.; Global Pneumococcal Sequencing Consortium. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine. 2019;43:338–46. <https://doi.org/10.1016/j.ebiom.2019.04.021>
 27. Lo SW, Gladstone RA, van Tonder AJ, Lees JA, du Plessis M, Benisty R, et al.; Global Pneumococcal Sequencing Consortium. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. Lancet Infect Dis. 2019;19:759–69. [https://doi.org/10.1016/S1473-3099\(19\)30297-X](https://doi.org/10.1016/S1473-3099(19)30297-X)
 28. Nagaraj G, Govindan V, Ganaie F, Venkatesha VT, Hawkins PA, Gladstone RA, et al. *Streptococcus pneumoniae* genomic datasets from an Indian population describing pre-vaccine evolutionary epidemiology using a whole genome sequencing approach. Microb Genom. 2021;7:000645. <https://doi.org/10.1099/mgen.0.000645>
 29. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327:469–74. <https://doi.org/10.1126/science.1182395>
 30. Méric G, Mageiros L, Pensar J, Laabei M, Yahara K, Pascoe B, et al. Disease-associated genotypes of the commensal skin bacterium *Staphylococcus epidermidis*. Nat Commun. 2018;9:5034. <https://doi.org/10.1038/s41467-018-07368-7>
 31. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, et al. Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. Nat Commun. 2021;12:765. <https://doi.org/10.1038/s41467-021-20988-w>
 32. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol. 2022;20:206–18. <https://doi.org/10.1038/s41579-021-00650-4>
 33. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. PLoS Genet. 2019;15:e1008114. <https://doi.org/10.1371/journal.pgen.1008114>
 34. Salvadori G, Junges R, Morrison DA, Petersen FC. Competence in *Streptococcus pneumoniae* and close commensal relatives: mechanisms and implications. Front Cell Infect Microbiol. 2019;9:94. <https://doi.org/10.3389/fcimb.2019.00094>
 35. Bagnoli F, Moschioni M, Donati C, Dimitrovska V, Ferlenghi I, Facciotti C, et al. A second pilus type in *Streptococcus pneumoniae* is prevalent in emerging serotypes and mediates adhesion to host cells. J Bacteriol. 2008;190:5480–92. <https://doi.org/10.1128/JB.00384-08>
 36. Dzaraly ND, Muthanna A, Mohd Desa MN, Taib NM, Masri SN, Rahman NIA, et al. Pilus islets and the clonal spread of piliated *Streptococcus pneumoniae*: a review. Int J Med Microbiol. 2020;310:151449. <https://doi.org/10.1016/j.ijmm.2020.151449>
 37. Hermoso JA, Lagartera L, González A, Stelter M, García P, Martínez-Ripoll M, et al. Insights into pneumococcal pathogenesis from the crystal structure of the modular teichoic acid phosphorylcholine esterase Pce. Nat Struct Mol Biol. 2005;12:533–8. <https://doi.org/10.1038/nsmb940>
 38. Kadioglu A, Weiser JN, Paton JC, Andrew PW. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. Nat Rev Microbiol. 2008;6:288–301. <https://doi.org/10.1038/nrmicro1871>
 39. Marquart ME. Pathogenicity and virulence of *Streptococcus pneumoniae*: cutting to the chase on proteases. Virulence. 2021;12:766–87. <https://doi.org/10.1080/21505594.2021.1889812>
 40. Brittan JL, Buckeridge TJ, Finn A, Kadioglu A, Jenkinson HF. Pneumococcal neuraminidase A: an essential upper airway colonization factor for *Streptococcus pneumoniae*. Mol Oral Microbiol. 2012;27:270–83. <https://doi.org/10.1111/j.2041-1014.2012.00658.x>
 41. Navarro-Torné A, Dias JG, Hrubá F, Lopalco PL, Pastore-Celentano L, Gauci AJ; Invasive Pneumococcal Disease Study Group. Risk factors for death from invasive pneumococcal disease, Europe, 2010. Emerg Infect Dis. 2015;21:417–25. <https://doi.org/10.3201/eid2103.140634>
 42. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE Jr, Walker H, et al. Penicillin-binding protein transpeptidase signatures for tracking and predicting β -lactam resistance levels in *Streptococcus pneumoniae*. MBio. 2016;7:e00756–16. <https://doi.org/10.1128/mBio.00756-16>
 43. Gibson PS, Veening JW. Gaps in the wall: understanding cell wall biology to tackle amoxicillin resistance in *Streptococcus pneumoniae*. Curr Opin Microbiol. 2023;72:102261. <https://doi.org/10.1016/j.mib.2022.102261>
 44. Egorova E, Kumar N, Gladstone RA, Urban Y, Voropaeva E, Chaplin AV, et al. Key features of pneumococcal isolates recovered in central and northwestern Russia in 2011–2018 determined through whole-genome sequencing. Microb Genom. 2022;8:mgen000851. <https://doi.org/10.1099/mgen.0.000851>
 45. Bai G, Vidal JE. Editorial: molecular pathogenesis of *Pneumococcus*. Front Cell Infect Microbiol. 2017;7:310. <https://doi.org/10.3389/fcimb.2017.00310>
 46. Chen L, Zhang YH, Wang S, Zhang Y, Huang T, Cai YD. Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways. PLoS One. 2017;12:e0184129. <https://doi.org/10.1371/journal.pone.0184129>
 47. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587–92. <https://doi.org/10.1093/nar/gkac963>
 48. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. PLoS Genet. 2018;14:e1007758. <https://doi.org/10.1371/journal.pgen.1007758>
 49. Aun E, Brauer A, Kisand V, Tenson T, Remm M. A k-mer-based method for the identification of phenotype-associated genomic biomarkers and predicting phenotypes of sequenced bacteria. PLoS Comput Biol. 2018;14:e1006434. <https://doi.org/10.1371/journal.pcbi.1006434>

Address for correspondence: Xiaohua Ye, Guangdong Pharmaceutical University, 283# Jianghai Dadao, Haizhu District, Guangzhou, China; email: smalltomato@163.com

EID cannot ensure accessibility for supplementary materials supplied by authors. Readers who have difficulty accessing supplementary content should contact the authors for assistance.

Disease-Associated *Streptococcus pneumoniae* Genetic Variation

Appendix

Appendix Table 1. Detailed information of 783 *Streptococcus pneumoniae* isolates in our study

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861694	Carriage	NP swab	M	3	GPSC4	14	ST876
PRJNA976286	SRR24861693	Carriage	NP swab	M	3	GPSC5	23A	ST338
PRJNA976286	SRR24861614	Carriage	NP swab	M	4	GPSC4	14	ST876
PRJNA976286	SRR24861040	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861025	Carriage	NP swab	F	3	GPSC4	14	ST876
PRJNA976286	SRR24861345	Carriage	NP swab	F	3	GPSC852	6A	-
PRJNA976286	SRR24861330	Carriage	NP swab	M	3	GPSC1	19F	-
PRJNA976286	SRR24861091	Carriage	NP swab	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861445	Carriage	NP swab	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861082	Carriage	NP swab	F	3	GPSC73	11A	ST99
PRJNA976286	SRR24861692	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861681	Carriage	NP swab	F	3	GPSC69	15A	ST11972
PRJNA976286	SRR24861670	Carriage	NP swab	M	5	GPSC10	23A	ST9396
PRJNA976286	SRR24861659	Carriage	NP swab	M	6	GPSC45	34	-
PRJNA976286	SRR24861648	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861637	Carriage	NP swab	M	6	GPSC1	19F	ST320
PRJNA976286	SRR24861274	Carriage	NP swab	F	5	GPSC4	14	ST876
PRJNA976286	SRR24861263	Carriage	NP swab	F	3	GPSC4	14	ST876
PRJNA976286	SRR24861252	Carriage	NP swab	M	3	GPSC45	34	ST9395
PRJNA976286	SRR24861625	Carriage	NP swab	M	3	GPSC1	19A	ST320
PRJNA976286	SRR24861613	Carriage	NP swab	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861602	Carriage	NP swab	F	4	GPSC45	34	ST9395
PRJNA976286	SRR24861239	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861228	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861217	Carriage	NP swab	M	4	GPSC1	19A	ST320

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861430	Carriage	NP swab	F	3	GPSC69	15A	ST6011
PRJNA976286	SRR24861419	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861408	Carriage	NP swab	F	3	GPSC14	23F	ST242
PRJNA976286	SRR24861062	Carriage	NP swab	M	5	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861051	Carriage	NP swab	M	5	GPSC1	19A	ST320
PRJNA976286	SRR24861598	Carriage	NP swab	M	4	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861587	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861576	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861213	Carriage	NP swab	F	4	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861202	Carriage	NP swab	M	5	GPSC1	19A	ST320
PRJNA976286	SRR24861191	Carriage	NP swab	F	5	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861404	Carriage	NP swab	F	5	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861393	Carriage	NP swab	F	4	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861382	Carriage	NP swab	F	5	GPSC152	15B	ST7768
PRJNA976286	SRR24861036	Carriage	NP swab	M	5	GPSC152	15B	ST3397
PRJNA976286	SRR24861024	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861796	Carriage	NP swab	F	5	GPSC1	19A	ST320
PRJNA976286	SRR24861561	Carriage	NP swab	M	6	GPSC45	34	ST11964
PRJNA976286	SRR24861550	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861539	Carriage	NP swab	M	6	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861176	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861165	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861154	Carriage	NP swab	F	5	GPSC1	19F	ST271
PRJNA976286	SRR24861367	Carriage	NP swab	F	6	GPSC45	34	-
PRJNA976286	SRR24861356	Carriage	NP swab	M	6	-	-	-
PRJNA976286	SRR24861344	Carriage	NP swab	M	5	-	-	ST10236
PRJNA976286	SRR24861781	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861770	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861759	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861524	Carriage	NP swab	M	5	GPSC1	19A	ST320

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861513	Carriage	NP swab	M	5	GPSC1	19A	ST320
PRJNA976286	SRR24861150	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861139	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861128	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861341	Carriage	NP swab	M	3	GPSC4	14	ST876
PRJNA976286	SRR24861329	Carriage	NP swab	M	2	GPSC45	34	ST9395
PRJNA976286	SRR24861318	Carriage	NP swab	M	4	GPSC45	34	ST9395
PRJNA976286	SRR24861755	Carriage	NP swab	F	3	GPSC23	6B	ST8526
PRJNA976286	SRR24861744	Carriage	NP swab	M	3	GPSC23	6B	ST8526
PRJNA976286	SRR24861733	Carriage	NP swab	F	4	GPSC45	34	ST9395
PRJNA976286	SRR24861498	Carriage	NP swab	M	4	GPSC45	34	ST9395
PRJNA976286	SRR24861487	Carriage	NP swab	M	4	GPSC45	34	-
PRJNA976286	SRR24861476	Carriage	NP swab	M	4	GPSC45	34	ST9395
PRJNA976286	SRR24861113	Carriage	NP swab	M	3	GPSC69	15A	ST6011
PRJNA976286	SRR24861102	Carriage	NP swab	M	5	-	-	ST7401
PRJNA976286	SRR24861090	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861303	Carriage	NP swab	M	6	-	16F	ST6542
PRJNA976286	SRR24861292	Carriage	NP swab	M	7	-	-	ST10236
PRJNA976286	SRR24861281	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861718	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861707	Carriage	NP swab	F	5	GPSC69	15A	-
PRJNA976286	SRR24861696	Carriage	NP swab	F	4	GPSC4	14	ST876
PRJNA976286	SRR24861461	Carriage	NP swab	F	5	GPSC69	15A	-
PRJNA976286	SRR24861450	Carriage	NP swab	M	4	GPSC4	14	ST876
PRJNA976286	SRR24861446	Carriage	NP swab	M	5	GPSC69	15A	-
PRJNA976286	SRR24861444	Carriage	NP swab	F	6	GPSC69	15A	ST11972
PRJNA976286	SRR24861443	Carriage	NP swab	F	6	GPSC158	16F	ST12671
PRJNA976286	SRR24861442	Carriage	NP swab	M	6	GPSC69	15A	ST11972
PRJNA976286	SRR24861441	Carriage	NP swab	M	6	GPSC230	35C	ST7752
PRJNA976286	SRR24861440	Carriage	NP swab	M	6	GPSC230	35C	ST7752
PRJNA976286	SRR24861439	Carriage	NP swab	M	6	GPSC230	35C	ST7752
PRJNA976286	SRR24861086	Carriage	NP swab	F	5	GPSC1	19F	ST271
PRJNA976286	SRR24861085	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861084	Carriage	NP swab	F	5	GPSC152	15C	ST6555
PRJNA976286	SRR24861083	Carriage	NP swab	F	6	GPSC230	13	ST2754
PRJNA976286	SRR24861081	Carriage	NP swab	F	6	GPSC230	13	ST2754
PRJNA976286	SRR24861080	Carriage	NP swab	M	5	-	6B	-

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861079	Carriage	NP swab	M	6	GPSC152	15C	ST6555
PRJNA976286	SRR24861078	Carriage	NP swab	M	6	GPSC23	6B	ST90
PRJNA976286	SRR24861077	Carriage	NP swab	M	5	GPSC165	34	-
PRJNA976286	SRR24861076	Carriage	NP swab	F	2	GPSC69	15A	ST11972
PRJNA976286	SRR24861075	Carriage	NP swab	M	4	GPSC244	6C	-
PRJNA976286	SRR24861074	Carriage	NP swab	F	3	GPSC244	6C	-
PRJNA976286	SRR24861073	Carriage	NP swab	M	4	GPSC45	34	-
PRJNA976286	SRR24861072	Carriage	NP swab	F	3	GPSC244	6C	-
PRJNA976286	SRR24861691	Carriage	NP swab	M	3	GPSC244	6C	-
PRJNA976286	SRR24861690	Carriage	NP swab	F	4	GPSC158	16F	-
PRJNA976286	SRR24861689	Carriage	NP swab	F	3	GPSC158	16F	-
PRJNA976286	SRR24861688	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861687	Carriage	NP swab	M	3	-	6B	-
PRJNA976286	SRR24861686	Carriage	NP swab	F	4	GPSC244	6C	-
PRJNA976286	SRR24861685	Carriage	NP swab	M	4	GPSC69	15A	-
PRJNA976286	SRR24861684	Carriage	NP swab	M	4	GPSC158	16F	-
PRJNA976286	SRR24861683	Carriage	NP swab	M	3	GPSC158	16F	-
PRJNA976286	SRR24861682	Carriage	NP swab	M	4	GPSC244	6C	-
PRJNA976286	SRR24861680	Carriage	NP swab	F	3	GPSC45	34	-
PRJNA976286	SRR24861679	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861678	Carriage	NP swab	M	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861677	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861676	Carriage	NP swab	M	4	GPSC45	6A	-
PRJNA976286	SRR24861675	Carriage	NP swab	F	4	-	14	-
PRJNA976286	SRR24861674	Carriage	NP swab	M	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861673	Carriage	NP swab	M	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861672	Carriage	NP swab	M	4	GPSC852	6B	ST3173
PRJNA976286	SRR24861671	Carriage	NP swab	M	3	GPSC165	34	-
PRJNA976286	SRR24861669	Carriage	NP swab	M	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861668	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861667	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861666	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861665	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861664	Carriage	NP swab	M	4	GPSC4	14	ST876
PRJNA976286	SRR24861663	Carriage	NP swab	M	4	GPSC186	35B	ST6327
PRJNA976286	SRR24861662	Carriage	NP swab	M	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861661	Carriage	NP swab	M	4	GPSC23	6B	ST90

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861660	Carriage	NP swab	M	4	GPSC45	6A	-
PRJNA976286	SRR24861658	Carriage	NP swab	M	3	GPSC186	35B	ST6327
PRJNA976286	SRR24861657	Carriage	NP swab	F	2	GPSC69	15A	ST11972
PRJNA976286	SRR24861656	Carriage	NP swab	F	6	GPSC212	35C	ST5972
PRJNA976286	SRR24861655	Carriage	NP swab	M	6	GPSC1	19A	ST320
PRJNA976286	SRR24861654	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861653	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861652	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861651	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861650	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861649	Carriage	NP swab	F	4	GPSC1	19F	ST236
PRJNA976286	SRR24861647	Carriage	NP swab	M	3	GPSC1	19F	ST236
PRJNA976286	SRR24861646	Carriage	NP swab	M	4	-	19A	ST10236
PRJNA976286	SRR24861645	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861644	Carriage	NP swab	M	3	GPSC152	15C	ST6555
PRJNA976286	SRR24861643	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861642	Carriage	NP swab	M	3	GPSC45	34	-
PRJNA976286	SRR24861641	Carriage	NP swab	M	4	GPSC152	15C	ST6555
PRJNA976286	SRR24861640	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861639	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861638	Carriage	NP swab	M	6	GPSC45	34	-
PRJNA976286	SRR24861636	Carriage	NP swab	M	5	GPSC45	34	-
PRJNA976286	SRR24861635	Carriage	NP swab	M	6	GPSC45	34	-
PRJNA976286	SRR24861634	Carriage	NP swab	F	5	GPSC23	6B	ST90
PRJNA976286	SRR24861633	Carriage	NP swab	F	6	-	6B	-
PRJNA976286	SRR24861632	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861631	Carriage	NP swab	M	5	GPSC177	35F	-
PRJNA976286	SRR24861278	Carriage	NP swab	M	4	-	-	ST10236
PRJNA976286	SRR24861277	Carriage	NP swab	M	5	GPSC152	15C	ST6555
PRJNA976286	SRR24861276	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861275	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861273	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861272	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861271	Carriage	NP swab	M	5	GPSC23	6B	-
PRJNA976286	SRR24861270	Carriage	NP swab	M	5	GPSC4	14	ST876
PRJNA976286	SRR24861269	Carriage	NP swab	M	5	GPSC152	15C	ST6555
PRJNA976286	SRR24861268	Carriage	NP swab	F	4	-	-	ST10236

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861267	Carriage	NP swab	F	3	GPSC177	35F	-
PRJNA976286	SRR24861266	Carriage	NP swab	F	4	-	6B	ST10236
PRJNA976286	SRR24861265	Carriage	NP swab	F	4	GPSC1	19F	ST236
PRJNA976286	SRR24861264	Carriage	NP swab	M	4	GPSC152	15B	ST6555
PRJNA976286	SRR24861262	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861261	Carriage	NP swab	M	4	GPSC165	34	ST7753
PRJNA976286	SRR24861260	Carriage	NP swab	F	2	GPSC10	23A	ST6227
PRJNA976286	SRR24861259	Carriage	NP swab	M	2	GPSC23	6B	ST90
PRJNA976286	SRR24861258	Carriage	NP swab	M	2	GPSC23	6B	ST90
PRJNA976286	SRR24861257	Carriage	NP swab	F	6	GPSC12	3	ST505
PRJNA976286	SRR24861256	Carriage	NP swab	M	4	GPSC47	6B	ST386
PRJNA976286	SRR24861255	Carriage	NP swab	M	5	-	-	-
PRJNA976286	SRR24861254	Carriage	NP swab	F	5	GPSC45	34	-
PRJNA976286	SRR24861253	Carriage	NP swab	M	5	GPSC152	15C	ST6555
PRJNA976286	SRR24861251	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861250	Carriage	NP swab	M	5	GPSC152	15C	ST6555
PRJNA976286	SRR24861249	Carriage	NP swab	F	4	GPSC152	15C	ST6555
PRJNA976286	SRR24861248	Carriage	NP swab	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861247	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861630	Carriage	NP swab	M	3	GPSC212	35A	ST7751
PRJNA976286	SRR24861629	Carriage	NP swab	M	3	GPSC45	34	-
PRJNA976286	SRR24861628	Carriage	NP swab	M	2	GPSC321	6B	ST902
PRJNA976286	SRR24861627	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861626	Carriage	NP swab	F	3	GPSC212	35A	ST7751
PRJNA976286	SRR24861624	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861623	Carriage	NP swab	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861622	Carriage	NP swab	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861621	Carriage	NP swab	F	3	GPSC12	3	ST180
PRJNA976286	SRR24861620	Carriage	NP swab	F	3	GPSC10	23A	ST9396
PRJNA976286	SRR24861619	Carriage	NP swab	F	4	GPSC212	35A	ST7751
PRJNA976286	SRR24861618	Carriage	NP swab	F	4	-	6B	ST63
PRJNA976286	SRR24861617	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861616	Carriage	NP swab	M	3	GPSC10	23A	ST9396
PRJNA976286	SRR24861615	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861612	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861611	Carriage	NP swab	M	3	GPSC10	23A	ST9396
PRJNA976286	SRR24861610	Carriage	NP swab	M	4	GPSC10	23A	ST9396

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861609	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861608	Carriage	NP swab	M	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861607	Carriage	NP swab	M	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861606	Carriage	NP swab	F	4	GPSC158	16F	ST8250
PRJNA976286	SRR24861605	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861604	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861603	Carriage	NP swab	M	3	GPSC10	23A	ST9396
PRJNA976286	SRR24861601	Carriage	NP swab	F	3	GPSC10	23A	-
PRJNA976286	SRR24861600	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861599	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861246	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861245	Carriage	NP swab	F	4	GPSC10	23A	-
PRJNA976286	SRR24861244	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861243	Carriage	NP swab	F	4	GPSC10	23A	ST9396
PRJNA976286	SRR24861242	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861241	Carriage	NP swab	M	4	GPSC10	23A	-
PRJNA976286	SRR24861240	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861238	Carriage	NP swab	F	3	GPSC10	23A	ST9396
PRJNA976286	SRR24861237	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861236	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861235	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861234	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861233	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861232	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861231	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861230	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861229	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861227	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861226	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861225	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861224	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861223	Carriage	NP swab	F	5	GPSC10	23A	ST9396
PRJNA976286	SRR24861222	Carriage	NP swab	F	5	GPSC1	19F	ST271
PRJNA976286	SRR24861221	Carriage	NP swab	M	6	GPSC10	23A	ST9396
PRJNA976286	SRR24861220	Carriage	NP swab	F	5	GPSC10	23A	ST9396
PRJNA976286	SRR24861219	Carriage	NP swab	F	6	GPSC10	23A	ST9396
PRJNA976286	SRR24861218	Carriage	NP swab	M	6	GPSC23	6B	ST90

Bioproject	SRA accession	Disease	Age					
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861216	Carriage	NP swab	M	6	-	-	-
PRJNA976286	SRR24861215	Carriage	NP swab	M	5	GPSC14	23F	ST242
PRJNA976286	SRR24861438	Carriage	NP swab	M	5	GPSC14	23F	ST242
PRJNA976286	SRR24861437	Carriage	NP swab	F	5	GPSC1	19F	ST271
PRJNA976286	SRR24861436	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861435	Carriage	NP swab	F	5	-	-	-
PRJNA976286	SRR24861434	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861433	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861432	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861431	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861429	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861428	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861427	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861426	Carriage	NP swab	F	5	GPSC777	-	ST7502
PRJNA976286	SRR24861425	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861424	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861423	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861422	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861421	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861420	Carriage	NP swab	M	5	GPSC10	23A	-
PRJNA976286	SRR24861418	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861417	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861416	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861415	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861414	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861413	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861412	Carriage	NP swab	M	4	GPSC248	7C	ST2758
PRJNA976286	SRR24861411	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861410	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861409	Carriage	NP swab	F	5	GPSC14	23F	ST242
PRJNA976286	SRR24861407	NIPD	Bronchoalveolar lavage fluid	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861071	NIPD	Bronchoalveolar lavage fluid	F	1	GPSC1	19A	ST320
PRJNA976286	SRR24861070	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC24	23F	ST13646

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861069	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC1	19F	ST14665
PRJNA976286	SRR24861068	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC1	19F	ST14655
PRJNA976286	SRR24861067	NIPD	Bronchoalveolar lavage fluid	F	2	GPSC1	19F	ST271
PRJNA976286	SRR24861066	NIPD	Bronchoalveolar lavage fluid	M	4	GPSC4	14	ST876
PRJNA976286	SRR24861065	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC321	6B	-
PRJNA976286	SRR24861064	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861063	NIPD	Bronchoalveolar lavage fluid	F	0	GPSC43	2	ST4745
PRJNA976286	SRR24861061	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861060	NIPD	Bronchoalveolar lavage fluid	F	1	GPSC16	23F	ST81
PRJNA976286	SRR24861059	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC16	23F	ST81
PRJNA976286	SRR24861058	NIPD	Bronchoalveolar lavage fluid	F	0	GPSC1	19A	-
PRJNA976286	SRR24861057	NIPD	Bronchoalveolar lavage fluid	F	0	GPSC23	6B	-
PRJNA976286	SRR24861056	NIPD	Bronchoalveolar lavage fluid	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861055	NIPD	Bronchoalveolar lavage fluid	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861054	NIPD	Bronchoalveolar lavage fluid	F	2	GPSC1	19F	ST271
PRJNA976286	SRR24861053	NIPD	Bronchoalveolar lavage fluid	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861052	NIPD	Bronchoalveolar lavage fluid	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861050	NIPD	Bronchoalveolar lavage fluid	M	3	GPSC1	19F	ST271

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861049	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC248	7C	ST11967
PRJNA976286	SRR24861048	NIPD	Bronchoalveolar lavage fluid	F	0	GPSC248	7C	ST11967
PRJNA976286	SRR24861047	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC1	19F	-
PRJNA976286	SRR24861046	NIPD	Bronchoalveolar lavage fluid	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861045	NIPD	Sputum	M	2	GPSC1	19F	ST7178
PRJNA976286	SRR24861044	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861043	NIPD	Sputum	M	2	GPSC1	19F	ST320
PRJNA976286	SRR24861042	NIPD	Sputum	M	0	GPSC152	15C	ST3397
PRJNA976286	SRR24861041	NIPD	Bronchoalveolar lavage fluid	F	1	GPSC321	6B	-
PRJNA976286	SRR24861597	NIPD	Bronchoalveolar lavage fluid	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861596	NIPD	Sputum	F	3	GPSC5	23A	ST338
PRJNA976286	SRR24861595	NIPD	Sputum	F	3	GPSC1	19F	-
PRJNA976286	SRR24861594	NIPD	Bronchoalveolar lavage fluid	M	0	GPSC852	6A	ST3173
PRJNA976286	SRR24861593	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861592	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861591	NIPD	Sputum	F	2	GPSC852	6A	ST3173
PRJNA976286	SRR24861590	NIPD	Sputum	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861589	NIPD	Sputum	M	4	GPSC4	14	ST876
PRJNA976286	SRR24861588	NIPD	Sputum	M	1	GPSC23	6B	ST90
PRJNA976286	SRR24861586	NIPD	Sputum	M	1	GPSC1	19F	ST4768
PRJNA976286	SRR24861585	NIPD	Sputum	F	4	GPSC1	19F	ST4768
PRJNA976286	SRR24861584	NIPD	Sputum	M	3	GPSC853	6B	ST9789
PRJNA976286	SRR24861583	NIPD	Sputum	F	0	GPSC904	19F	ST2097
PRJNA976286	SRR24861582	NIPD	Sputum	M	1	GPSC23	6B	ST90
PRJNA976286	SRR24861581	NIPD	Sputum	F	4	GPSC1	19F	ST4768
PRJNA976286	SRR24861580	NIPD	Sputum	M	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861579	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861578	NIPD	Sputum	M	2	GPSC1	19F	ST236
PRJNA976286	SRR24861577	NIPD	Sputum	M	1	GPSC1	19F	ST236

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861575	NIPD	Sputum	M	3	GPSC16	23F	ST81*
PRJNA976286	SRR24861574	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC43	18C	ST3214
PRJNA976286	SRR24861573	NIPD	Sputum	M	3	GPSC321	6B	-
PRJNA976286	SRR24861572	NIPD	Sputum	M	1	GPSC852	6B	ST3173
PRJNA976286	SRR24861571	NIPD	Sputum	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861570	NIPD	Middle ear fluid	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861569	NIPD	Sputum	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861568	NIPD	Sputum	F	0	GPSC4	14	-
PRJNA976286	SRR24861567	NIPD	Sputum	F	0	GPSC4	14	ST876
PRJNA976286	SRR24861214	NIPD	Sputum	M	1	GPSC1	19F	ST236
PRJNA976286	SRR24861212	NIPD	Sputum	M	1	-	15A	-
PRJNA976286	SRR24861211	NIPD	Sputum	F	4	-	6C	-
PRJNA976286	SRR24861210	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC10	23F	ST230
PRJNA976286	SRR24861209	NIPD	Sputum	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861208	NIPD	Sputum	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861207	NIPD	Sputum	F	1	GPSC904	15A	-
PRJNA976286	SRR24861206	NIPD	Sputum	M	0	GPSC853	6A	ST9789
PRJNA976286	SRR24861205	NIPD	Sputum	F	1	GPSC1	19F	ST320
PRJNA976286	SRR24861204	NIPD	Sputum	F	2	GPSC16	23F	ST81
PRJNA976286	SRR24861203	NIPD	Sputum	F	0	GPSC321	6B	ST902
PRJNA976286	SRR24861201	NIPD	Sputum	M	0	GPSC1	19A	ST320
PRJNA976286	SRR24861200	NIPD	Middle ear fluid	F	1	GPSC12	3	ST505
PRJNA976286	SRR24861199	NIPD	Sputum	M	1	GPSC852	6A	ST3173
PRJNA976286	SRR24861198	NIPD	Sputum	F	1	GPSC230	6A	-
PRJNA976286	SRR24861197	NIPD	Sputum	M	1	GPSC10	23A	-
PRJNA976286	SRR24861196	NIPD	Sputum	F	3	GPSC12	3	ST505
PRJNA976286	SRR24861195	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861194	NIPD	Sputum	M	11	GPSC43	9V	ST11949
PRJNA976286	SRR24861193	NIPD	Sputum	M	3	GPSC382	6B	ST7397
PRJNA976286	SRR24861192	NIPD	Sputum	M	1	GPSC321	6B	ST902
PRJNA976286	SRR24861190	NIPD	Sputum	F	0	GPSC852	6A	ST3173
PRJNA976286	SRR24861189	NIPD	Sputum	F	0	GPSC13	6A	ST473
PRJNA976286	SRR24861188	NIPD	Sputum	M	1	GPSC152	15B	ST3397
PRJNA976286	SRR24861187	NIPD	Sputum	F	1	GPSC852	6A	ST6340

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861186	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC152	15B	ST9765
PRJNA976286	SRR24861185	NIPD	Sputum	M	0	GPSC852	6A	ST6340
PRJNA976286	SRR24861184	NIPD	Sputum	F	1	GPSC1	19F	-
PRJNA976286	SRR24861183	NIPD	Sputum	F	2	GPSC1	19F	ST271
PRJNA976286	SRR24861406	NIPD	Sputum	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861405	NIPD	Sputum	M	3	GPSC10	23F	ST230
PRJNA976286	SRR24861403	NIPD	Sputum	M	0	GPSC1	19F	ST320
PRJNA976286	SRR24861402	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861401	NIPD	Bronchoalveolar lavage fluid	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861400	NIPD	Sputum	M	2	-	-	-
PRJNA976286	SRR24861399	NIPD	Sputum	F	0	GPSC852	6A	ST6340
PRJNA976286	SRR24861398	NIPD	Sputum	F	4	GPSC904	19F	ST2097
PRJNA976286	SRR24861397	NIPD	Sputum	M	0	GPSC321	6B	ST902
PRJNA976286	SRR24861396	NIPD	Sputum	M	5	GPSC1	19F	ST320!
PRJNA976286	SRR24861395	NIPD	Sputum	F	1	GPSC321	6B	-
PRJNA976286	SRR24861394	NIPD	Sputum	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861392	NIPD	Sputum	M	0	GPSC321	6B	ST902
PRJNA976286	SRR24861391	NIPD	Sputum	F	3	GPSC1	19F	-
PRJNA976286	SRR24861390	NIPD	Middle ear fluid	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861389	NIPD	Sputum	F	0	GPSC1	19F	ST320
PRJNA976286	SRR24861388	NIPD	Sputum	F	2	GPSC43	9V	ST280
PRJNA976286	SRR24861387	NIPD	Sputum	M	2	GPSC1	19A	ST320
PRJNA976286	SRR24861386	NIPD	Sputum	F	1	GPSC1	19F	ST320
PRJNA976286	SRR24861385	NIPD	Bronchoalveolar lavage fluid	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861384	NIPD	Sputum	M	5	GPSC248	7C	ST11967
PRJNA976286	SRR24861383	NIPD	Sputum	M	0	-	15C	-
PRJNA976286	SRR24861381	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861380	NIPD	Sputum	F	3	GPSC10	23F	ST230
PRJNA976286	SRR24861379	NIPD	Puncture fluid	F	0	GPSC1	19A	ST320
PRJNA976286	SRR24861378	NIPD	Sputum	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861377	NIPD	Sputum	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861376	NIPD	Middle ear fluid	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861375	NIPD	Sputum	F	2	GPSC1	19F	ST271

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861039	NIPD	Middle ear fluid	F	9	GPSC852	6A	ST3173
PRJNA976286	SRR24861038	NIPD	Sputum	F	4	GPSC1	19A	ST320
PRJNA976286	SRR24861037	NIPD	Sputum	F	3	-	-	ST10236
PRJNA976286	SRR24861035	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861034	NIPD	Sputum	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861033	NIPD	Sputum	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861032	NIPD	Sputum	M	2	GPSC1	19F	ST320
PRJNA976286	SRR24861031	NIPD	Middle ear fluid	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861030	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861029	NIPD	Sputum	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861028	NIPD	Sputum	M	2	GPSC1	19F	ST320
PRJNA976286	SRR24861027	NIPD	Sputum	F	2	GPSC853	6A	ST9789
PRJNA976286	SRR24861026	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861023	NIPD	Sputum	F	0	GPSC152	15B	ST6555
PRJNA976286	SRR24861805	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861804	NIPD	Sputum	M	1	GPSC1	19F	ST320
PRJNA976286	SRR24861803	NIPD	Sputum	F	2	GPSC1	19F	ST271
PRJNA976286	SRR24861802	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861801	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861800	NIPD	Middle ear fluid	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861799	NIPD	Sputum	F	0	GPSC23	6B	ST6339
PRJNA976286	SRR24861798	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861797	NIPD	Sputum	M	1	GPSC1	19F	ST320!
PRJNA976286	SRR24861795	NIPD	Sputum	M	0	GPSC852	6A	ST3173
PRJNA976286	SRR24861794	NIPD	Sputum	M	1	GPSC5	23A	ST5242
PRJNA976286	SRR24861793	NIPD	Sputum	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861792	NIPD	Sputum	F	1	GPSC14	23F	ST242
PRJNA976286	SRR24861791	NIPD	Sputum	F	5	GPSC1	19F	ST271
PRJNA976286	SRR24861566	IPD	Blood	M	4	GPSC1	19A	ST320
PRJNA976286	SRR24861565	NIPD	Puncture fluid	M	7	GPSC152	15C	ST6555
PRJNA976286	SRR24861564	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861563	NIPD	Sputum	F	2	GPSC321	6B	ST902
PRJNA976286	SRR24861562	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861560	NIPD	Sputum	F	3	GPSC1	19A	ST320
PRJNA976286	SRR24861559	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861558	NIPD	Middle ear fluid	F	0	GPSC5	23A	ST5242
PRJNA976286	SRR24861557	NIPD	Middle ear fluid	M	0	GPSC1	19A	ST320

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861556	NIPD	Middle ear fluid	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861555	NIPD	Puncture fluid	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861554	NIPD	Puncture fluid	M	7	GPSC14	23F	ST242
PRJNA976286	SRR24861553	NIPD	Puncture fluid	M	12	GPSC1	19F	ST271
PRJNA976286	SRR24861552	NIPD	Puncture fluid	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861551	IPD	Blood	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861549	IPD	Blood	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861548	NIPD	Eye secretion	M	6	GPSC186	35B	-
PRJNA976286	SRR24861547	NIPD	Bronchoalveolar lavage fluid	M	3	GPSC4	14	ST876
PRJNA976286	SRR24861546	NIPD	Sputum	F	0	GPSC13	6A	ST473
PRJNA976286	SRR24861545	NIPD	Sputum	F	1	GPSC14	23F	ST2338
PRJNA976286	SRR24861544	NIPD	Sputum	F	0	GPSC383	28F	ST3398
PRJNA976286	SRR24861543	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861542	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861541	NIPD	Middle ear fluid	F	3	GPSC4	14	ST876
PRJNA976286	SRR24861540	NIPD	Sputum	M	0	GPSC47	6B	ST386
PRJNA976286	SRR24861538	NIPD	Sputum	M	4	-	15A	ST63
PRJNA976286	SRR24861537	NIPD	Sputum	M	0	GPSC1	19F	ST320
PRJNA976286	SRR24861536	NIPD	Sputum	M	1	GPSC852	6A	ST6918
PRJNA976286	SRR24861535	NIPD	Sputum	M	0	GPSC4	14	ST876
PRJNA976286	SRR24861182	NIPD	Sputum	M	1	GPSC212	15F	ST6202
PRJNA976286	SRR24861181	NIPD	Sputum	F	0	GPSC1	19A	ST320
PRJNA976286	SRR24861180	NIPD	Sputum	M	0	GPSC1	19F	ST320
PRJNA976286	SRR24861179	NIPD	Sputum	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861178	NIPD	Sputum	M	1	GPSC4	14	-
PRJNA976286	SRR24861177	NIPD	Sputum	M	1	GPSC4	14	ST876
PRJNA976286	SRR24861175	NIPD	Sputum	F	0	GPSC16	23F	ST81
PRJNA976286	SRR24861174	NIPD	Middle ear fluid	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861173	NIPD	Sputum	M	0	GPSC23	6B	ST90
PRJNA976286	SRR24861172	NIPD	Bronchoalveolar lavage fluid	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861171	NIPD	Sputum	M	0	GPSC152	15B	ST6555*
PRJNA976286	SRR24861170	NIPD	Sputum	F	0	GPSC23	6B	ST90
PRJNA976286	SRR24861169	NIPD	Sputum	F	0	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861168	NIPD	Sputum	F	4	GPSC1	19F	ST271

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861167	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861166	NIPD	Sputum	M	3	GPSC321	6B	-
PRJNA976286	SRR24861164	NIPD	Sputum	F	3	GPSC904	15A	ST63
						:9		
PRJNA976286	SRR24861163	NIPD	Sputum	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861162	NIPD	Sputum	F	3	GPSC1	6B	-
PRJNA976286	SRR24861161	NIPD	Sputum	M	4	GPSC852	3	ST3173
PRJNA976286	SRR24861160	NIPD	Sputum	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861159	NIPD	Middle ear fluid	F	0	GPSC858	3	ST10085
PRJNA976286	SRR24861158	NIPD	Sputum	M	1	GPSC904	15A	ST63
						:9		
PRJNA976286	SRR24861157	NIPD	Sputum	F	2	GPSC852	6A	ST3173
PRJNA976286	SRR24861156	NIPD	Sputum	F	4	GPSC158	16F	ST8250
PRJNA976286	SRR24861155	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861153	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861152	NIPD	Sputum	F	7	GPSC230	13	-
PRJNA976286	SRR24861151	NIPD	Sputum	F	6	GPSC321	6B	ST902
PRJNA976286	SRR24861374	NIPD	Sputum	M	2	GPSC43	9N	-
PRJNA976286	SRR24861373	NIPD	Sputum	F	5	GPSC1	19F	ST271
PRJNA976286	SRR24861372	NIPD	Sputum	M	4	GPSC904	15A	ST63
						:9		
PRJNA976286	SRR24861371	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861370	NIPD	Sputum	F	4	GPSC321	6B	ST902*
PRJNA976286	SRR24861369	NIPD	Sputum	M	0	GPSC853	6A	-
PRJNA976286	SRR24861368	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861366	NIPD	Sputum	F	1	GPSC69	15A	ST11972
PRJNA976286	SRR24861365	NIPD	Sputum	M	0	GPSC23	6B	ST90
PRJNA976286	SRR24861364	NIPD	Bronchoalveolar lavage fluid	F	1	GPSC43	9N	-
PRJNA976286	SRR24861363	NIPD	Sputum	F	2	GPSC1	19F	ST271
PRJNA976286	SRR24861362	NIPD	Sputum	M	3	GPSC1	19F	ST320
PRJNA976286	SRR24861361	NIPD	Sputum	F	2	GPSC16	23F	ST81
PRJNA976286	SRR24861360	NIPD	Sputum	M	3	GPSC244	6C	ST7767*
PRJNA976286	SRR24861359	NIPD	Sputum	M	0	-	23F	-
PRJNA976286	SRR24861358	NIPD	Sputum	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861357	NIPD	Sputum	F	0	GPSC10	23A	ST9396
PRJNA976286	SRR24861355	NIPD	Sputum	F	2	GPSC152	6B	-

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861354	NIPD	Sputum	F	6	GPSC69	15A	ST11972
PRJNA976286	SRR24861353	NIPD	Sputum	F	0	GPSC152	15B	ST7768
PRJNA976286	SRR24861352	NIPD	Sputum	M	10	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861351	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861350	NIPD	Sputum	M	0	GPSC23	6B	ST90
PRJNA976286	SRR24861349	NIPD	Bronchoalveolar lavage fluid	M	10	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861348	NIPD	Sputum	F	4	GPSC10	23F	ST230
PRJNA976286	SRR24861347	NIPD	Sputum	M	1	GPSC321	6B	ST902
PRJNA976286	SRR24861346	NIPD	Sputum	M	0	GPSC10	23F	ST230
PRJNA976286	SRR24861343	NIPD	Sputum	M	0	GPSC1	19F	-
PRJNA976286	SRR24861790	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861789	NIPD	Sputum	M	1	GPSC321	6B	ST902
PRJNA976286	SRR24861788	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861787	NIPD	Sputum	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861786	NIPD	Bronchoalveolar lavage fluid	M	1	GPSC152	15C	-
PRJNA976286	SRR24861785	NIPD	Sputum	F	1	GPSC13	6B	ST1876
PRJNA976286	SRR24861784	NIPD	Sputum	M	8	GPSC1	19F	ST271
PRJNA976286	SRR24861783	NIPD	Bronchoalveolar lavage fluid	F	6	-	14	-
PRJNA976286	SRR24861782	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861780	NIPD	Sputum	F	2	GPSC10	23F	ST230
PRJNA976286	SRR24861779	NIPD	Sputum	M	1	GPSC852	6A	ST3173
PRJNA976286	SRR24861778	NIPD	Sputum	M	1	GPSC158	16F	ST8250
PRJNA976286	SRR24861777	NIPD	Sputum	F	0	GPSC1	19F	ST320
PRJNA976286	SRR24861776	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861775	NIPD	Sputum	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861774	NIPD	Sputum	F	3	GPSC5	23A	ST338
PRJNA976286	SRR24861773	NIPD	Sputum	F	4	GPSC1	19F	-
PRJNA976286	SRR24861772	NIPD	Sputum	M	1	GPSC1	19F	-
PRJNA976286	SRR24861771	NIPD	Sputum	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861769	NIPD	Sputum	M	1	GPSC321	6B	ST902
PRJNA976286	SRR24861768	NIPD	Sputum	F	10	GPSC1	19A	ST320
PRJNA976286	SRR24861767	NIPD	Sputum	M	1	GPSC1	19F	ST320
PRJNA976286	SRR24861766	NIPD	Sputum	M	0	GPSC1	19F	ST271

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861765	NIPD	Sputum	M	0	GPSC1	19A	ST320
PRJNA976286	SRR24861764	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861763	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861762	NIPD	Sputum	M	0	GPSC321	6B	ST902
PRJNA976286	SRR24861761	NIPD	Sputum	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861760	NIPD	Sputum	M	0	GPSC43	17F	ST1263
PRJNA976286	SRR24861534	NIPD	Sputum	M	3	GPSC853	6A	ST9789
PRJNA976286	SRR24861533	NIPD	Sputum	F	0	GPSC321	6B	ST902
PRJNA976286	SRR24861532	NIPD	Sputum	F	0	GPSC321	6B	ST902
PRJNA976286	SRR24861531	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861530	NIPD	Sputum	M	2	GPSC321	6B	ST902
PRJNA976286	SRR24861529	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861528	NIPD	Sputum	M	2	GPSC23	6B	ST90
PRJNA976286	SRR24861527	NIPD	Sputum	F	2	GPSC23	6B	ST90
PRJNA976286	SRR24861526	NIPD	Sputum	F	2	-	14	-
PRJNA976286	SRR24861525	NIPD	Sputum	M	3	GPSC852	6A	ST6918
PRJNA976286	SRR24861523	NIPD	Sputum	M	3	GPSC73	11A	ST99
PRJNA976286	SRR24861522	NIPD	Sputum	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861521	NIPD	Sputum	M	0	GPSC4	14	-
PRJNA976286	SRR24861520	NIPD	Sputum	M	1	GPSC321	19F	-
PRJNA976286	SRR24861519	NIPD	Sputum	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861518	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861517	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861516	NIPD	Sputum	F	1	GPSC1	19F	ST236
PRJNA976286	SRR24861515	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861514	NIPD	Sputum	M	3	GPSC152	15C	ST3397
PRJNA976286	SRR24861512	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861511	NIPD	Sputum	F	4	GPSC852	6A	ST6340
PRJNA976286	SRR24861510	NIPD	Sputum	M	3	GPSC1	19F	ST320
PRJNA976286	SRR24861509	NIPD	Sputum	M	0	GPSC852	6B	ST3173
PRJNA976286	SRR24861508	NIPD	Sputum	M	1	GPSC1	19A	ST320
PRJNA976286	SRR24861507	NIPD	Sputum	M	6	GPSC850	23F	ST1437
PRJNA976286	SRR24861506	NIPD	Sputum	F	2	GPSC1	19F	ST320
PRJNA976286	SRR24861505	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861504	NIPD	Sputum	M	3	-	19F	-
PRJNA976286	SRR24861503	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861149	NIPD	Sputum	M	1	GPSC4	14	ST876

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861148	NIPD	Sputum	M	0	GPSC1	19F	ST320
PRJNA976286	SRR24861147	NIPD	Sputum	M	0	GPSC14	23F	ST880!
PRJNA976286	SRR24861146	NIPD	Sputum	F	1	GPSC321	6B	ST902
PRJNA976286	SRR24861145	NIPD	Sputum	F	3	GPSC152	15B	-
PRJNA976286	SRR24861144	NIPD	Sputum	F	1	GPSC1	19F	ST1464
PRJNA976286	SRR24861143	NIPD	Sputum	F	7	GPSC852	6A	ST3173
PRJNA976286	SRR24861142	NIPD	Sputum	M	0	GPSC5	23A	ST5242
PRJNA976286	SRR24861141	NIPD	Sputum	F	0	GPSC73	11A	ST99
PRJNA976286	SRR24861140	NIPD	Sputum	F	3	GPSC1	19F	-
PRJNA976286	SRR24861138	NIPD	Sputum	F	0	GPSC5	23A	ST5242
PRJNA976286	SRR24861137	NIPD	Sputum	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861136	NIPD	Sputum	M	0	-	19F	-
PRJNA976286	SRR24861135	NIPD	Sputum	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861134	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861133	NIPD	Sputum	M	0	GPSC5	23A	ST338
PRJNA976286	SRR24861132	NIPD	Sputum	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861131	NIPD	Sputum	M	0	GPSC687	39	ST6318
PRJNA976286	SRR24861130	NIPD	Sputum	F	1	GPSC10	23A	ST6227
PRJNA976286	SRR24861129	NIPD	Sputum	M	1	GPSC165	33C	ST6578
PRJNA976286	SRR24861127	NIPD	Sputum	F	3	GPSC1	19F	ST320
PRJNA976286	SRR24861126	NIPD	Sputum	M	2	GPSC1	19F	ST320
PRJNA976286	SRR24861125	NIPD	Sputum	F	2	GPSC321	6B	ST902
PRJNA976286	SRR24861124	NIPD	Sputum	F	1	GPSC1	19F	ST271
PRJNA976286	SRR24861123	NIPD	Sputum	M	0	GPSC321	6B	ST902
PRJNA976286	SRR24861122	NIPD	Sputum	F	3	-	23F	-
PRJNA976286	SRR24861121	NIPD	Sputum	M	1	GPSC1	19F	ST320
PRJNA976286	SRR24861120	NIPD	Sputum	M	2	GPSC1	19F	ST271
PRJNA976286	SRR24861119	NIPD	Sputum	M	3	GPSC1	19F	ST271
PRJNA976286	SRR24861342	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861340	NIPD	Sputum	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861339	NIPD	Pus	M	2	GPSC1	19A	ST320
PRJNA976286	SRR24861338	NIPD	Sputum	M	1	GPSC1	19F	ST271
PRJNA976286	SRR24861337	NIPD	Sputum	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861336	NIPD	Sputum	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861335	NIPD	Sputum	F	1	GPSC23	6A	ST90
PRJNA976286	SRR24861334	NIPD	Sputum	M	2	GPSC852	6A	ST3173
PRJNA976286	SRR24861333	IPD	Blood	F	4	-	14	-

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861332	NIPD	Sputum	M	6	GPSC850	23F	ST1437
PRJNA976286	SRR24861331	NIPD	Sputum	M	1	GPSC23	6B	ST90
PRJNA976286	SRR24861328	NIPD	Sputum	F	0	-	-	-
PRJNA976286	SRR24861327	IPD	Blood	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861326	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861325	NIPD	Sputum	M	0	GPSC16	23F	ST81
PRJNA976286	SRR24861324	NIPD	Sputum	F	0	GPSC16	23F	ST81
PRJNA976286	SRR24861323	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861322	NIPD	Sputum	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861321	NIPD	Sputum	M	0	GPSC1	19F	-
PRJNA976286	SRR24861320	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861319	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861317	NIPD	Sputum	F	0	GPSC152	14	-
PRJNA976286	SRR24861316	IPD	Blood	M	2	GPSC1	19A	ST320
PRJNA976286	SRR24861315	NIPD	Sputum	M	0	GPSC1	19F	ST271
PRJNA976286	SRR24861314	NIPD	Sputum	M	0	GPSC1	19F	ST1968
PRJNA976286	SRR24861313	NIPD	Sputum	F	2	GPSC321	6B	ST902
PRJNA976286	SRR24861312	NIPD	Sputum	M	0	GPSC13	19F	ST5501
PRJNA976286	SRR24861311	NIPD	Sputum	F	3	GPSC152	15C	ST3397
PRJNA976286	SRR24861758	NIPD	Sputum	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861757	IPD	Blood	F	0	GPSC1	19F	ST271
PRJNA976286	SRR24861756	Carriage	NP swab	F	6	GPSC852	6B	ST3173
PRJNA976286	SRR24861754	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861753	Carriage	NP swab	M	5	GPSC16	23F	ST81
PRJNA976286	SRR24861752	Carriage	NP swab	F	4	GPSC1	19A	ST320
PRJNA976286	SRR24861751	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861750	Carriage	NP swab	F	4	GPSC23	6B	ST96
PRJNA976286	SRR24861749	Carriage	NP swab	F	4	GPSC16	23F	ST81
PRJNA976286	SRR24861748	Carriage	NP swab	M	4	GPSC23	6B	ST96
PRJNA976286	SRR24861747	Carriage	NP swab	F	4	GPSC5	23A	ST5242
PRJNA976286	SRR24861746	Carriage	NP swab	F	4	GPSC1	19F	-
PRJNA976286	SRR24861745	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861743	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861742	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861741	Carriage	NP swab	M	5	GPSC16	23F	ST81
PRJNA976286	SRR24861740	Carriage	NP swab	F	6	GPSC23	6B	ST90
PRJNA976286	SRR24861739	Carriage	NP swab	F	6	GPSC73	11A	-

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861738	Carriage	NP swab	F	6	-	16F	ST6542
PRJNA976286	SRR24861737	Carriage	NP swab	M	5	-	16F	ST6542
PRJNA976286	SRR24861736	Carriage	NP swab	F	6	-	16F	ST6542
PRJNA976286	SRR24861735	Carriage	NP swab	M	5	GPSC852	6A	-
PRJNA976286	SRR24861734	Carriage	NP swab	M	4	GPSC852	6A	ST6340
PRJNA976286	SRR24861732	Carriage	NP swab	M	4	GPSC230	6B	ST3263
PRJNA976286	SRR24861731	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861730	Carriage	NP swab	M	4	GPSC852	6A	ST6340
PRJNA976286	SRR24861729	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861728	Carriage	NP swab	F	4	GPSC904	15A	ST374
						;9		
PRJNA976286	SRR24861727	Carriage	NP swab	M	3	GPSC13	6A	ST473
PRJNA976286	SRR24861502	Carriage	NP swab	M	3	GPSC321	6B	ST902
PRJNA976286	SRR24861501	Carriage	NP swab	M	6	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861500	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861499	Carriage	NP swab	M	6	GPSC1	19F	ST320
PRJNA976286	SRR24861497	Carriage	NP swab	F	6	GPSC904	15A	ST63
						;9		
PRJNA976286	SRR24861496	Carriage	NP swab	M	6	GPSC904	15A	ST374
						;9		
PRJNA976286	SRR24861495	Carriage	NP swab	F	6	GPSC321	6B	ST902
PRJNA976286	SRR24861494	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861493	Carriage	NP swab	M	6	GPSC850	23F	ST7497
PRJNA976286	SRR24861492	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861491	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861490	Carriage	NP swab	F	4	GPSC1	19F	ST320
PRJNA976286	SRR24861489	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861488	Carriage	NP swab	M	5	-	-	ST10236
PRJNA976286	SRR24861486	Carriage	NP swab	M	4	GPSC73	11A	ST99
PRJNA976286	SRR24861485	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861484	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861483	Carriage	NP swab	F	5	GPSC850	23F	ST1437
PRJNA976286	SRR24861482	Carriage	NP swab	M	4	GPSC321	6B	ST902
PRJNA976286	SRR24861481	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861480	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861479	Carriage	NP swab	F	6	GPSC16	23F	-

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861478	Carriage	NP swab	F	3	GPSC45	34	ST11945
PRJNA976286	SRR24861477	Carriage	NP swab	F	4	GPSC45	34	ST11945
PRJNA976286	SRR24861475	Carriage	NP swab	M	3	GPSC73	11A	ST99
PRJNA976286	SRR24861474	Carriage	NP swab	F	3	GPSC45	34	ST11945
PRJNA976286	SRR24861473	Carriage	NP swab	M	3	GPSC1	19F	ST320
PRJNA976286	SRR24861472	Carriage	NP swab	F	3	GPSC1	19F	ST320
PRJNA976286	SRR24861471	Carriage	NP swab	M	3	GPSC1	19F	ST320
PRJNA976286	SRR24861118	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861117	Carriage	NP swab	F	3	GPSC1	19F	ST320
PRJNA976286	SRR24861116	Carriage	NP swab	M	5	GPSC850	23F	ST7497
PRJNA976286	SRR24861115	Carriage	NP swab	F	7	GPSC850	23F	ST7497
PRJNA976286	SRR24861114	Carriage	NP swab	M	6	GPSC904	15A	ST63
							;9	
PRJNA976286	SRR24861112	Carriage	NP swab	M	5	GPSC152	15C	ST10098
PRJNA976286	SRR24861111	Carriage	NP swab	F	5	GPSC1	19F	ST320
PRJNA976286	SRR24861110	Carriage	NP swab	M	5	GPSC152	15B	ST10098
PRJNA976286	SRR24861109	Carriage	NP swab	F	6	GPSC152	15C	ST10098
PRJNA976286	SRR24861108	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861107	Carriage	NP swab	M	5	GPSC73	11A	ST6739
PRJNA976286	SRR24861106	Carriage	NP swab	F	5	GPSC321	6B	ST902
PRJNA976286	SRR24861105	Carriage	NP swab	F	4	GPSC850	23F	ST7497
PRJNA976286	SRR24861104	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861103	Carriage	NP swab	M	4	GPSC1	19F	ST320
PRJNA976286	SRR24861101	Carriage	NP swab	M	4	GPSC1	19A	ST320
PRJNA976286	SRR24861100	Carriage	NP swab	M	5	GPSC10	23A	ST9396
PRJNA976286	SRR24861099	Carriage	NP swab	M	5	GPSC1	19F	ST320
PRJNA976286	SRR24861098	Carriage	NP swab	F	5	GPSC1	19F	ST320
PRJNA976286	SRR24861097	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861096	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861095	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861094	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861093	Carriage	NP swab	M	3	GPSC73	11A	ST99
PRJNA976286	SRR24861092	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861089	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861088	Carriage	NP swab	F	3	GPSC1	19F	ST271
PRJNA976286	SRR24861087	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861310	Carriage	NP swab	M	6	GPSC321	6B	ST902

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861309	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861308	Carriage	NP swab	F	4	GPSC1	19F	ST320
PRJNA976286	SRR24861307	Carriage	NP swab	F	4	GPSC152	15B	ST3397
PRJNA976286	SRR24861306	Carriage	NP swab	F	4	GPSC152	15C	ST3397
PRJNA976286	SRR24861305	Carriage	NP swab	F	3	GPSC1	19F	ST320
PRJNA976286	SRR24861304	Carriage	NP swab	F	5	GPSC321	23F	ST902
PRJNA976286	SRR24861302	Carriage	NP swab	M	5	GPSC904	15A	ST63
							;9	
PRJNA976286	SRR24861301	Carriage	NP swab	M	5	GPSC152	15C	ST6555
PRJNA976286	SRR24861300	Carriage	NP swab	M	6	GPSC321	6B	ST902
PRJNA976286	SRR24861299	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861298	Carriage	NP swab	F	6	GPSC850	23F	ST7497
PRJNA976286	SRR24861297	Carriage	NP swab	F	6	GPSC321	6B	ST902
PRJNA976286	SRR24861296	Carriage	NP swab	M	5	GPSC1	19F	ST320
PRJNA976286	SRR24861295	Carriage	NP swab	M	6	GPSC852	23F	ST3173
PRJNA976286	SRR24861294	Carriage	NP swab	F	3	GPSC321	6B	ST902
PRJNA976286	SRR24861293	Carriage	NP swab	M	5	GPSC10	23F	ST230
PRJNA976286	SRR24861291	Carriage	NP swab	M	6	GPSC904	15A	ST374
							;9	
PRJNA976286	SRR24861290	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861289	Carriage	NP swab	F	6	GPSC69	15A	ST11972
PRJNA976286	SRR24861288	Carriage	NP swab	F	5	GPSC1	19F	ST8781
PRJNA976286	SRR24861287	Carriage	NP swab	M	6	GPSC69	15A	ST11972
PRJNA976286	SRR24861286	Carriage	NP swab	M	4	GPSC321	15A	ST902
PRJNA976286	SRR24861285	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861284	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861283	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861282	Carriage	NP swab	F	3	GPSC23	6B	ST90
PRJNA976286	SRR24861280	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861279	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861726	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861725	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861724	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861723	Carriage	NP swab	M	3	GPSC23	6B	ST90
PRJNA976286	SRR24861722	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861721	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861720	Carriage	NP swab	F	4	GPSC73	11A	ST99

Bioproject	SRA accession	Disease		Age				
accession	number	phenotypes	Isolation source	Sex	(years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861719	Carriage	NP swab	F	4	GPSC16	23F	ST81
PRJNA976286	SRR24861717	Carriage	NP swab	F	3	GPSC16	23F	ST81
PRJNA976286	SRR24861716	Carriage	NP swab	M	3	GPSC158	16F	ST8250
PRJNA976286	SRR24861715	Carriage	NP swab	F	6	GPSC5	23A	ST338
PRJNA976286	SRR24861714	Carriage	NP swab	M	5	GPSC23	6B	ST90
PRJNA976286	SRR24861713	Carriage	NP swab	F	4	GPSC1	19F	ST271
PRJNA976286	SRR24861712	Carriage	NP swab	F	5	GPSC852	15A	-
PRJNA976286	SRR24861711	Carriage	NP swab	F	5	GPSC852	15A	ST3173
PRJNA976286	SRR24861710	Carriage	NP swab	M	5	-	28A	-
PRJNA976286	SRR24861709	Carriage	NP swab	F	4	GPSC852	6A	ST3173
PRJNA976286	SRR24861708	Carriage	NP swab	F	4	GPSC5	23A	ST5242
PRJNA976286	SRR24861706	Carriage	NP swab	M	5	GPSC69	15A	ST6011
PRJNA976286	SRR24861705	Carriage	NP swab	F	4	GPSC852	6A	ST3173
PRJNA976286	SRR24861704	Carriage	NP swab	F	4	GPSC852	6A	ST3173
PRJNA976286	SRR24861703	Carriage	NP swab	M	2	GPSC321	6B	-
PRJNA976286	SRR24861702	Carriage	NP swab	M	4	GPSC850	23F	ST7497
PRJNA976286	SRR24861701	Carriage	NP swab	M	4	-	11A	-
PRJNA976286	SRR24861700	Carriage	NP swab	M	3	GPSC73	11A	ST99
PRJNA976286	SRR24861699	Carriage	NP swab	M	3	GPSC73	11A	ST99
PRJNA976286	SRR24861698	Carriage	NP swab	F	4	GPSC73	11A	ST99
PRJNA976286	SRR24861697	Carriage	NP swab	F	4	GPSC850	23F	ST7497
PRJNA976286	SRR24861695	Carriage	NP swab	F	3	GPSC73	11A	ST99
PRJNA976286	SRR24861470	Carriage	NP swab	F	3	GPSC73	11A	ST99
PRJNA976286	SRR24861469	Carriage	NP swab	F	4	GPSC321	6B	ST902
PRJNA976286	SRR24861468	Carriage	NP swab	M	4	GPSC23	6B	ST90
PRJNA976286	SRR24861467	Carriage	NP swab	M	5	GPSC1	19F	ST271
PRJNA976286	SRR24861466	Carriage	NP swab	M	5	GPSC321	6B	ST902
PRJNA976286	SRR24861465	Carriage	NP swab	M	6	GPSC1	19F	ST271
PRJNA976286	SRR24861464	Carriage	NP swab	M	6	GPSC1	19F	ST271
PRJNA976286	SRR24861463	Carriage	NP swab	F	5	GPSC5	23A	ST5242
PRJNA976286	SRR24861462	Carriage	NP swab	F	4	GPSC23	6B	ST90
PRJNA976286	SRR24861460	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861459	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861458	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861457	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861456	Carriage	NP swab	F	3	GPSC69	15A	ST11972
PRJNA976286	SRR24861455	Carriage	NP swab	F	4	GPSC69	15A	ST11972

Bioproject accession	SRA accession number	Disease phenotypes	Isolation source	Sex	Age (years)	GPSCs	Serotypes	MLST
PRJNA976286	SRR24861454	Carriage	NP swab	M	4	GPSC1	19F	ST271
PRJNA976286	SRR24861453	Carriage	NP swab	F	3	GPSC69	15A	ST11972
PRJNA976286	SRR24861452	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861451	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861449	Carriage	NP swab	M	3	GPSC69	15A	ST11972
PRJNA976286	SRR24861448	Carriage	NP swab	F	4	GPSC69	15A	ST11972
PRJNA976286	SRR24861447	Carriage	NP swab	M	5	GPSC12	3	ST180

Appendix Table 2. Demographics of participants contributing *S. pneumoniae* isolates*

Feature	Infection isolates (n=349)	Carriage Isolates (n=434)	χ^2	<i>P</i>
Gender				
Male	202(57.9)	237(54.6)	0.84	0.359
Female	147(42.1)	197(45.4)		
Age (years)				
≤5	332(95.1)	371(85.5)	19.62	<0.001
>5	17(4.9)	63(14.5)		

*Data are presented as no. (%) or as otherwise indicated. Bold text indicates statistical significance.

Appendix Table 3. Function or pathogenic role of the 886 disease-associated k-mers (genes)

Gene	k-mer		Average		Function and/or Pathogenic	
	hits	$-\log(p)$	OR	Protein Name	Role	GO annotations
<i>pcpA</i>	618	18.07	1.38	Pneumococcal Choline-Binding Protein A	Protection Against Lung Infection and Sepsis	GO:0033925, GO:0008152, GO:0035821, GO:0005576
<i>pitA</i>	216	32.55	1.87	Pilus-2 Subunit, Ancillary Protein	Adherence	GO:0016020
<i>pavB</i>	210	21.02	1.28	Pneumococcal Adherence and Virulence Protein B	Adherence and Colonization	GO:0016020
<i>pbp1A</i>	111	58.68	2.03	penicillin-binding protein PBP1A	Antibiotic Resistance	GO:0008955, GO:0009002, GO:0008658, GO:0006508, GO:0016020
<i>cps4D</i>	79	9.43	1.44	Capsular Polysaccharide Biosynthesis Protein	Immune Modulation	GO:0004715, GO:0005524, GO:0045227, GO:0016310, GO:0005737

Gene	k-mer		Average		Function and/or Pathogenic	
	hits	-log(p)	OR	Protein Name	Role	GO annotations
<i>phtD</i>	68	8.02	1.29	Pneumococcal Histidine Triad D	Adherence and Immune Evasion	-
<i>lga</i>	23	52.22	1.53	IgA1 Protease	Immune Modulation	GO:0004222, GO:0008270, GO:0006508, GO:0005576, GO:0016020
<i>zmpB</i>	10	31.94	1.52	Zinc Metalloprotease B	Immune Evasion and Colonization	GO:0004222, GO:0008270, GO:0006508, GO:0005576, GO:0009986, GO:0016020
<i>cpsC</i>	9	34.27	1.56	Capsular Polysaccharide Biosynthesis Protein	Immune Modulation	GO:0005351, GO:0045227, GO:0009103, GO:0015774, GO:0005886
<i>pbp3</i>	9	26.22	1.47	Penicillin-Binding Protein 3	Antibiotic Resistance	GO:0009002, GO:0008360, GO:0071555, GO:0009252, GO:0006508
<i>pspA</i>	8	46.76	1.54	Pneumococcal Surface Protein A	Immune Modulation	GO:0046872, GO:0007155, GO:0030001, GO:0005886
<i>lanM</i>	7	48.02	1.56	Type 2 Lantipeptide Synthetase LanM	Toxin Production and Resistance	GO:0031179
<i>metE</i>	7	19.26	1.54	5-methyltetrahydropteroyltri glutamate-- homocysteine S-methyltransferase	Amino Acid Biosynthesis	GO:0003871, GO:0008270, GO:0009086, GO:0032259
<i>cbpA</i>	7	31.55	1.41	Choline Binding Protein A	Adherence, Immune Evasion, Colonization and Invasion	GO:0033925, GO:0008152, GO:0035821
<i>cbpE</i>	7	42.54	1.46	Choline Binding Protein E	Adherence	GO:0033925, GO:0008152, GO:0035821
<i>zmpA</i>	6	39.38	1.57	Zinc Metalloprotease A, IgA1	IgA1 Protease Enzyme and Colonization	GO:0004222, GO:0008236, GO:0008270, GO:0006508, GO:0005576, GO:0016020
<i>zmpD</i>	6	24.81	1.49	Zinc Metalloprotease D, IgA1 Paralog Protease	Immune Evasion and Colonization	GO:0004222, GO:0008270, GO:0006508, GO:0005576, GO:0016020
<i>folP</i>	6	16.25	1.42	Dihydropteroate Synthase	Biosynthesis of Cofactors, Prosthetic groups, and Carriers	GO:0004156, GO:0046872, GO:0046656, GO:0046654

Gene	k-mer		Average		Function and/or Pathogenic	
	hits	-log(p)	OR	Protein Name	Role	GO annotations
<i>pbp2B</i>	6	59.17	1.56	Penicillin-Binding Protein 2B	Antibiotic Resistance	GO:0071972, GO:0008658, GO:0008360, GO:0071555, GO:0009252, GO:0046677, GO:0005886
<i>secA2</i>	5	16.73	1.38	Accessory Sec System Translocase SecA2	Protein and Peptide Secretion and Trafficking	GO:0008564, GO:0005524, GO:0065002, GO:0006605, GO:0008564, GO:0017038, GO:0005886, GO:0005737
<i>crcB</i>	4	19.60	1.50	Fluoride Efflux Transporter CrcB	Unknown Function	GO:1903425, GO:0005886
<i>rrgA</i>	4	12.41	1.44	Pilus-1 Tip Protein (Adhesin)	Adherence	GO:0016020
<i>argH</i>	4	47.63	1.39	Argininosuccinate Lyase	Amino Acid Biosynthesis	GO:0004056, GO:0042450, GO:0005737
<i>srtG1</i>	3	44.07	1.76	PI-2 pilus system class B sortase SrtG1	Adherence	GO:0016787, GO:0016020
<i>cbpJ</i>	3	14.53	1.54	Choline-Binding Protein J	Virulence	GO:0035821
<i>psrP</i>	3	23.18	1.49	Pneumococcal Serine-Rich Repeats Protein	Adherence	GO:0003677, GO:0007155, GO:0052031, GO:0044010, GO:0005576, GO:0009275, GO:0009986
<i>srtC-1</i>	3	31.01	1.46	Mediates host cell adhesion	Adherence	GO:0016787, GO:0016020
<i>aph(3')-IIIa</i>	3	27.45	1.45	aminoglycoside O-phosphotransferase APH(3')-IIIa	Amino acid biosynthesis	GO:0008910, GO:0005524, GO:0046872, GO:0016310, GO:0046677
<i>nanA</i>	3	18.77	1.44	Neuraminidase A	Hydrolytic Enzyme, Adherence and Colonization	GO:0052794, GO:0052795, GO:0052796, GO:0004308, GO:0009313, GO:0006689, GO:0005576, GO:0043231, GO:0005737, GO:0016020
<i>galR</i>	3	17.74	1.44	Galactose Operon Repressor	Regulatory Functions and DNA Interactions	GO:0000976, GO:0003700, GO:0006355
<i>phtA</i>	3	21.84	1.43	Pneumococcal Histidine Triad A	Adherence and Immune Evasion	GO:0008965
<i>catA</i>	3	17.95	1.35	Type A Chloramphenicol O-acetyltransferase	Antibiotic Resistance	GO:0008811, GO:0046677

Gene	k-mer		Average		Function and/or Pathogenic	
	hits	-log(p)	OR	Protein Name	Role	GO annotations
<i>pbp1B</i>	3	36.59	1.54	Penicillin-Binding Protein 1B	Antibiotic Resistance	GO:0008955, GO:0009002, GO:0008658, GO:0006508, GO:0016020
<i>pbp2X</i>	3	42.31	1.44	Penicillin-Binding Protein 2X	Antibiotic Resistance	GO:0008658, GO:0008360, GO:0071555, GO:0009252, GO:0007049, GO:0051301, GO:0046677, GO:0005886
<i>pspC</i>	3	46.49	1.55	Pneumococcal Surface Protein C	Adherence, Immune Evasion, Colonization and Invasion	GO:0033925, GO:0008152, GO:0035821
<i>mef(A)</i>	2	30.46	1.68	Macrolide Efflux MFS Transporter Mef(A)	Antibiotic Resistance	GO:0022857, GO:0005886
<i>rrgC</i>	2	20.72	1.62	Pilus-1 Anchore Protein	Adherence	GO:0016020
<i>clpX</i>	2	45.23	1.59	ATP-dependent Clp protease ATP-binding subunit ClpX	Degradation of Proteins, Peptides, and Glycopeptides	GO:0016887, GO:0140662, GO:0016887, GO:0051603, GO:0051301, GO:0009376
<i>pitB</i>	2	11.59	1.58	Pilus-2 Subunit, Backbone Protein	Adherence	GO:0016020
<i>cbpG</i>	2	14.74	1.37	Choline-Binding Protein G	Adherence and Colonization	GO:0008236, GO:0035821, GO:0006508
<i>gtfA</i>	2	9.36	1.30	Accessory Sec System Glycosyltransferase GtfA	Protein Modification and Repair	GO:0016757, GO:0000166, GO:0018242, GO:0005886, GO:0005737, GO:0017122
<i>lytA</i>	2	7.71	1.24	Autolysin (N-Acetyl-Muramoyl-L-Alanine Amidase)	Autolytic Enzyme, Cell Wall Digestion and Autolysis	GO:0008745, GO:0030435, GO:0071555, GO:0009253, GO:0030420, GO:0035821, GO:0005576
<i>rrgB</i>	1	26.27	1.72	Pilus-1 Backbone Protein	Adherence	GO:0016020
<i>fusA</i>	1	24.15	1.63	Fructooligosaccharide ABC transporter substrate-binding protein FusA	Protein Synthesis	GO:0046872, GO:0015774, GO:0005886
<i>msr(D)</i>	1	22.59	1.60	ABC-F type ribosomal protection protein Msr(D)	Transport and Binding proteins, Toxin Production and Resistance	GO:0005524
<i>srtC-2</i>	1	28.46	1.54	PI-1 pilus system sortase SrtC-2	Adherence	GO:0016787, GO:0016020

Gene	k-mer		Average		Function and/or Pathogenic	
	hits	-log(p)	OR	Protein Name	Role	GO annotations
<i>galU</i>	1	23.80	1.54	UTP--glucose-1-phosphate uridylyltransferase galU	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	GO:0003983, GO:0006011, GO:0009058
<i>mltG</i>	1	16.52	1.47	Endolytic Transglycosylase MltG	Part of the elongasome which synthesizes peripheral peptidoglycan.	GO:0008932, GO:0071555, GO:0009252GO:0005886
<i>gpsB</i>	1	11.91	1.47	Cell Division Regulator GpsB	Mediates Protein Phosphorylation and Penicillin-Binding Protein Interactions	GO:0008360, GO:0007049, GO:0051301, GO:0005737
<i>tet(M)</i>	1	21.23	1.44	Tetracycline Resistance Ribosomal Protection Protein Tet(M)	Antibiotic Resistance	GO:0003924, GO:0005525, GO:0046677, GO:0006412
<i>pavA</i>	1	7.51	1.42	Pneumococcal Adherence and Virulence Protein A	Adherence, Immune Evasion, Colonization and Translocation	GO:0000049, GO:0043023, GO:0072344, GO:0005576, GO:0042603, GO:0009986, GO:0005737
<i>ilvC</i>	1	8.37	1.37	Ketol-acid reductoisomerase	Amino Acid Biosynthesis	GO:0004455, GO:0000287, GO:0050661, GO:0009097, GO:0009099

Appendix Table 4. List of genomes available on NCBI including those used for validation of risk prediction analyses

Disease phenotypes	Isolation source	GPSCs	Serotypes	MLST	NCBI genome accession number
IPD	Blood	GPSC544	6A	ST11918	GCA_901289825.1
IPD	Blood	GPSC141	6B	ST874	GCA_901293355.1
IPD	CSF	GPSC1	19F	ST236	GCA_001103965.1
IPD	Clinical specimen	GPSC1	19F	ST320	GCA_901286745.1
IPD	Clinical specimen	GPSC1	19A	ST320	GCA_901289905.1
IPD	Clinical specimen	GPSC1	19A	ST320	GCA_901289945.1
IPD	Clinical specimen	GPSC1	19F	ST271	GCA_901290385.1
IPD	Clinical specimen	GPSC1	19F	ST271	GCA_901290515.1
IPD	CSF	GPSC1	19F	ST236	GCA_901290595.1
IPD	Blood	GPSC1	19F	ST320	GCA_001162305.1
IPD	CSF	GPSC141	6B	ST874	GCA_901292895.1

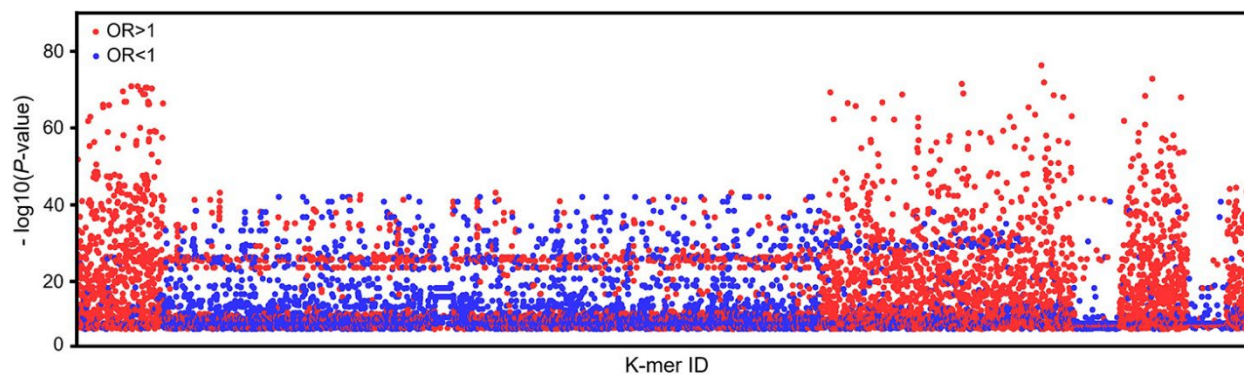
Disease phenotypes	Isolation source	GPSCs	Serotypes	MLST	NCBI genome accession number
IPD	Blood	GPSC141	6B	ST874	GCA_901293085.1
IPD	Blood	GPSC1	19F	ST236	GCA_901293215.1
IPD	Blood	GPSC9	14	ST5285	GCA_901293445.1
IPD	Blood	GPSC1	19A	ST320	GCA_001118025.1
IPD	Blood	GPSC1	19A	ST271	GCA_001129725.1
IPD	Blood	GPSC9	14	ST63	GCA_901295225.1
IPD	Blood	GPSC213	6A	ST1988	GCA_901305565.1
IPD	Blood	GPSC185	6B	ST2016	GCA_901329145.1
IPD	Blood	GPSC1	19F	ST236	GCA_901330015.1
IPD	Blood	GPSC1	19F	ST1396	GCA_901331045.1
IPD	Blood	GPSC23	6B	ST11237	GCA_901332835.1
IPD	Clinical specimen	GPSC23	23F	ST311	GCA_901333965.1
IPD	Clinical specimen	GPSC16	23F	ST81	GCA_901334305.1
IPD	Clinical specimen	GPSC14	23F	ST242	GCA_901337455.1
IPD	Blood	GPSC1	19F	ST1421	GCA_901276655.1
IPD	Blood	GPSC6	14	ST156	GCA_901330195.1
IPD	Peritoneal fluid	GPSC1	19F	ST1421	GCA_901338195.1
IPD	Blood	GPSC1	19A	ST320	GCA_901294185.1
IPD	CSF	GPSC105	6B	ST5625	GCA_901217015.1
IPD	Pleural fluid	GPSC1	19A	ST320	GCA_901330725.1
IPD	CSF	GPSC1	19F	ST1421	GCA_901218465.1
IPD	CSF	GPSC16	23F	ST81	GCA_901252795.1
IPD	CSF	GPSC1	19F	ST1421	GCA_901273885.1
IPD	CSF	GPSC29	6A	ST11310	GCA_901293175.1
IPD	CSF	GPSC5	23F	ST338	GCA_901294975.1
IPD	CSF	GPSC1	19F	ST236	GCA_901329605.1
IPD	CSF	GPSC47	6B	ST315	GCA_901330955.1
IPD	Blood	GPSC24	6A	ST4598	GCA_901332755.1
IPD	CSF	GPSC6	23F	ST156	GCA_901338475.1
IPD	Blood	GPSC1	19F	ST271	GCA_901251715.1
IPD	Blood	GPSC1	19F	ST1421	GCA_901218775.1
IPD	Blood	GPSC1	19F	ST1421	GCA_901213735.1
IPD	Blood	GPSC1	19F	ST651	GCA_901334775.1
IPD	Pleural fluid	GPSC23	6B	ST90	GCA_901215735.1
IPD	Pleural fluid	GPSC23	6B	ST1121	GCA_901215755.1
IPD	Blood	GPSC1	19F	ST236	GCA_901315645.1
IPD	Blood	GPSC189	6B	ST5619	GCA_901252455.1
IPD	Blood	GPSC1	19F	ST651	GCA_901309215.1

Disease phenotypes	Isolation source	GPSCs	Serotypes	MLST	NCBI genome accession number
IPD	Blood	GPSC6	14	ST156	GCA_901293545.1
IPD	Blood	GPSC1	19A	ST320	GCA_901294425.1
IPD	pleural fluid	GPSC1	19A	ST320	GCA_901295055.1
IPD	Blood	GPSC1	19F	ST5459	GCA_901215195.1
IPD	Blood	GPSC1	19F	ST925	GCA_901246325.1
IPD	Blood	GPSC321	6B	ST2757	PATH2282
IPD	-	GPSC321	6B	ST902	GCA_901287305.1
IPD	-	GPSC321	6B	ST902	GCA_901305435.1
IPD	Blood	GPSC21	19F	ST347	GCA_901217095.1
IPD	Blood	GPSC21	19F	ST9930	GCA_901216415.1
IPD	Blood	GPSC21	19F	ST2715	GCA_901216645.1
NIPD	Sputum	GPSC23	6B	ST90	GCA_001101845.1
NIPD	Sputum	GPSC1	19F	ST-	GCA_001146585.1
NIPD	Sputum	GPSC1	19F	ST236	GCA_001330955.1
NIPD	Middle ear fluid	GPSC1	19F	ST320	GCA_901332285.1
NIPD	Middle ear fluid	GPSC1	19F	ST320	GCA_901332745.1
NIPD	Sputum	GPSC1	19F	ST236	GCA_001168545.1
NIPD	Middle ear fluid	GPSC1	19F	ST320	GCA_901301755.1
NIPD	Eye discharge	GPSC1	19A	ST320	GCA_901340745.1
NIPD	Middle ear fluid	GPSC1	19F	ST320	GCA_901302685.1
NIPD	Sputum	GPSC1	19F	ST236	GCA_001090545.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_001990165.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_001902455.1
NIPD	Middle ear fluid	GPSC1	19A	ST320	GCA_901332215.1
NIPD	Auditory canal	GPSC1	19A	ST320	GCA_001896085.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_002081405.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_002081315.1
NIPD	Sputum	GPSC1	19A	ST320	GCA_019399065.1
NIPD	Sputum	GPSC1	19A	ST320	GCA_019398885.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_019399165.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_019399085.1
NIPD	Sputum	GPSC1	19A	ST320	GCA_019397725.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_019398005.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_019397965.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_019397925.1
NIPD	Ear canal secretions	GPSC1	19A	ST320	GCA_019397365.1
NIPD	Sputum	GPSC1	19F	ST320	GCA_019397765.1
NIPD	Sputum	GPSC1	19F	ST271	GCA_019397645.1

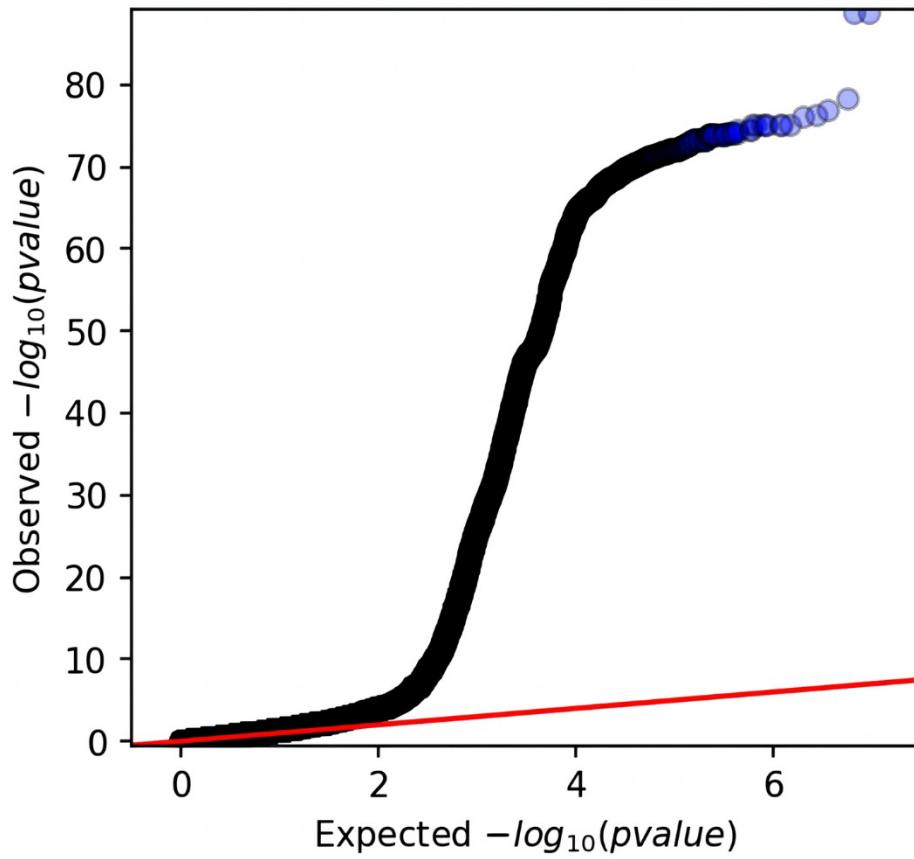
Disease phenotypes	Isolation source	GPSCs	Serotypes	MLST	NCBI genome accession number
NIPD	Sputum	GPSC1	19F	ST271	GCA_019397505.1
NIPD	Middle ear fluid	GPSC1	19F	ST236	GCA_001168485.1
NIPD	Sputum	GPSC1	19F	ST236	GCA_001330735.1
NIPD	Sputum	GPSC23	6B	ST90	GCA_001138625.1
NIPD	Sputum	GPSC23	6B	ST90	GCA_001132385.1
NIPD	Sputum	GPSC23	6B	ST90	GCA_001329835.1
NIPD	Middle ear fluid	GPSC23	6B	ST90	GCA_001329855.1
NIPD	Clinical specimen	GPSC23	6B	ST90	GCA_901286825.1
NIPD	Clinical specimen	GPSC23	6B	ST90	GCA_901286845.1
NIPD	Middle ear fluid	GPSC7	23F	ST9751	GCA_901246435.1
NIPD	Middle ear fluid	GPSC7	23F	ST439	GCA_901247245.1
NIPD	Middle ear fluid	GPSC7	23F	ST629	GCA_901247555.1
NIPD	Middle ear fluid	GPSC7	23F	ST629	GCA_901247755.1
NIPD	Middle ear fluid	GPSC7	23F	ST36	GCA_901332705.1
NIPD	Middle ear fluid	GPSC16	23F	ST81	GCA_901302635.1
NIPD	Sputum	GPSC230	6A	ST16328	GCA_019398925.1
NIPD	Sputum	GPSC13	6A	ST473	GCA_019398795.1
NIPD	Sputum	GPSC852	6A	ST11968	GCA_019397945.1
NIPD	Alveolar lavage fluid	GPSC852	6A	ST3173	GCA_019397425.1
NIPD	Sputum	Not assigned	19F	ST4662	GCA_002081035.1
NIPD	Middle ear fluid	GPSC1	19A	ST320	GCA_901301585.1
NIPD	Middle ear fluid	GPSC1	19A	ST320	GCA_901302985.1
NIPD	Middle ear fluid	GPSC1	19A	ST320	GCA_901301655.1
NIPD	Sputum	GPSC4	14	ST876	GCA_002081325.1
NIPD	Sputum	GPSC4	14	ST876	GCA_002081195.1
NIPD	Sputum	GPSC4	14	ST876	GCA_002081175.1
NIPD	Sputum	GPSC4	14	ST876	GCA_019397605.1
NIPD	Sputum	GPSC321	6B	ST902	GCA_019398745.1
NIPD	Sputum	GPSC321	6B	ST902	GCA_019397845.1
NIPD	Endotracheal tube tip	GPSC321	6B	ST902	GCA_019397625.1
NIPD	Middle ear fluid	GPSC32	7F	ST11901	GCA_901302105.1
NIPD	Sputum	GPSC23	6B	ST90	GCA_001173005.1
NIPD	Sputum	GPSC23	6B	ST90	GCA_900005375.1
Carriage	NP swab	GPSC141	6B	ST5293	GCA_901247035.1
Carriage	NP swab	GPSC141	6B	ST5293	GCA_901219865.1
Carriage	NP swab	GPSC116	23A	ST7707	GCA_901219685.1
Carriage	NP swab	GPSC69	15A	ST4965	GCA_901265045.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286105.1

Disease phenotypes	Isolation source	GPSCs	Serotypes	MLST	NCBI genome accession number
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286145.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286195.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286265.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286295.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286315.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286345.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286355.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286365.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286415.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901286455.1
Carriage	NP swab	GPSC165	34	ST4640	GCA_901290455.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901309365.1
Carriage	NP swab	GPSC1	19F	ST236	GCA_001110065.1
Carriage	NP swab	GPSC45	34	ST12204	GCA_901309895.1
Carriage	NP swab	GPSC1	19F	ST271	GCA_901311365.1
Carriage	NP swab	GPSC1	19F	ST1464	GCA_901311855.1
Carriage	NP swab	GPSC69	15A	ST6011	GCA_901312265.1
Carriage	NP swab	GPSC5	23A	ST5242	GCA_901312355.1
Carriage	NP swab	GPSC9	15A	ST63	GCA_901312435.1
Carriage	NP swab	GPSC16	23F	ST81	GCA_901312475.1
Carriage	NP swab	GPSC9	15A	ST63	GCA_901312655.1
Carriage	NP swab	GPSC9	15A	ST63	GCA_901313005.1
Carriage	NP swab	GPSC1	19F	ST271	GCA_901313485.1
Carriage	NP swab	GPSC1	19F	ST271	GCA_901313715.1
Carriage	NP swab	GPSC509	6B	ST9597	GCA_901220635.1
Carriage	NP swab	GPSC7	23A	ST439	GCA_901214025.1
Carriage	NP swab	GPSC16	23F	ST81	GCA_901214365.1
Carriage	NP swab	GPSC69	15A	ST5448	GCA_901214405.1
Carriage	NP swab	GPSC69	15A	ST5448	GCA_901215885.1
Carriage	NP swab	GPSC23	6B	ST90	GCA_901216465.1
Carriage	NP swab	GPSC23	6B	ST5628	GCA_901217565.1
Carriage	NP swab	GPSC23	6B	ST5628	GCA_901217935.1
Carriage	NP swab	GPSC76	6B	ST1092	GCA_901218755.1
Carriage	NP swab	GPSC1	19F	ST9906	GCA_901220035.1
Carriage	NP swab	GPSC5	23A	ST338	GCA_901253265.1
Carriage	NP swab	GPSC5	23A	ST338	GCA_901254525.1
Carriage	NP swab	GPSC23	6B	ST5628	GCA_901257675.1
Carriage	NP swab	GPSC1	19F	ST236	GCA_901260265.1

Disease phenotypes	Isolation source	GPSCs	Serotypes	MLST	NCBI genome accession number
Carriage	NP swab	GPSC1	19F	ST236	GCA_001099865.1
Carriage	NP swab	GPSC23	6B	ST7416	GCA_901261585.1
Carriage	NP swab	GPSC1	19F	ST1421	GCA_901291835.1
Carriage	NP swab	GPSC16	19F	ST81	GCA_901328025.1
Carriage	NP swab	GPSC322	34	ST7441	GCA_901329445.1
Carriage	NP swab	GPSC69	15A	ST5453	GCA_901330255.1
Carriage	NP swab	GPSC69	15A	ST5448	GCA_901337705.1
Carriage	NP swab	GPSC237	34	ST12190	GCA_901340485.1
Carriage	NP swab	GPSC1	19F	ST236	GCA_900170135.1
Carriage	NP swab	GPSC1	19F	ST236	GCA_001163425.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001158765.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001128505.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001329995.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001133245.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001171445.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001155585.1
Carriage	NP swab	GPSC1	19F	ST4414	GCA_001330135.1



Appendix Figure 1. Manhattan plot showing statistical significance of all k-mers.



Appendix Figure 2. QQ-plot showing the expected and observed P-values for the GWAS analysis using LMM.