Whole genome sequencing of *Streptococcus pneumoniae*: Development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline

Georgia Kapatai Corresp., 1 , **Carmen L Sheppard** 1 , **Ali Al-Shahib** 2 , **David J Litt** 1 , **Anthony P Underwood** 2 , **Timothy G Harrison** 1 , **Norman K Fry** 1

1 Respiratory and Vaccine Preventable Bacterial Reference Unit, Public Health England, London, United Kingdom

2 Infectious Disease Informatics, Public Health England, London, United Kingdom

Corresponding Author: Georgia Kapatai
Email address: Georgia.Kapatai@phe.gov.uk

Streptococcus pneumoniae typically express one of 92 serologically distinct capsule polysaccharide (cps) types (serotypes). Some of these serotypes are closely related to each other; using the commercially available typing antisera these are assigned to common serogroups containing types that show cross-reactivity. In this serotyping scheme, factor antisera are used to allocate serotypes within a serogroup, based on patterns of reactions. This serotyping method is technically demanding, requires considerable experience and the reading of the results can be subjective.

This study describes the analysis of the *S. pneumoniae* capsular operon genetic sequence to determine serotype distinguishing features and the development, evaluation and verification of an automated whole genome sequence (WGS)-based serotyping bioinformatics tool, PneumoCaT (**Pneumo**coccal **Ca**psule **T**yping).

Initially, WGS data from 871 *S. pneumoniae* isolates were mapped to reference cps locus sequences for the 92 serotypes. Thirty-two of 92 serotypes could be unambiguously identified based on sequence similarities within the cps operon. The remaining 60 were allocated to one of 20 'genogroups' that broadly correspond to the immunologically defined serogroups. By comparing the cps reference sequences for each genogroup, unique molecular differences were determined for serotypes within 18 of the 20 genogroups and verified using the set of 871 isolates. This information was used to design a decision-tree style algorithm within the PneumoCaT bioinformatics tool to predict to serotype level for 89/94 (92 + 2 molecular types/subtypes) from WGS data and to serogroup level for serogroups 24 and 32, which currently comprise 2.1% of UK referred, invasive isolates submitted to the National Reference Laboratory (NRL), Public Health England (June 2014 – July 2015). PneumoCaT was evaluated with an internal validation set of 2065 UK isolates covering 72/92 serotypes, including 19 non-typeable isolates and an external validation set of 2964 isolates from Thailand (n=2531), USA (n=181) and Iceland (n=252). PneumoCaT was able to predict serotype in 99.1% of the typeable UK isolates and in 99.0% of the non-UK isolates. Concordance was evaluated in UK isolates where further investigation was possible; in 91.5% of the cases the predicted capsular type was concordant with the serologically derived serotype. Following retesting, concordance increased to 99.3% and in most resolved cases (97.8%; 135/138) discordance was shown to be caused by errors in original serotyping. Replicate testing demonstrated that PneumoCaT gave 100% reproducibility of the predicted serotype result.

In summary, we have developed a WGS-based serotyping method that can predict capsular type to

serotype level for 89/94 serotypes and to serogroup level for the remaining four. This approach could be integrated into routine typing workflows in reference laboratories, reducing the need for phenotypic immunological testing.

1 Whole Genome Sequencing of *Streptococcus pneumoniae*:

2 Development, Evaluation and Verification of Targets for

3 Serogroup and Serotype Prediction using an Automated

4 Pipeline

5 Georgia Kapatai[1*], Carmen L. Sheppard[1], Ali Al-Shahib[2], David J. Litt[1], Anthony P. Underwood[2], Timothy G. Harrison[1]

6 and Norman K. Fry[1].

8 [1] Respiratory and Vaccine Preventable Bacterial Reference Unit, Microbiology Services Colindale, Public Health

9 England

10 [2] Infectious Disease Informatics, Microbiology Services Colindale, Public Health England

12 * Corresponding author

13 Email: georgia.kapatai@phe.gov.uk

Introduction

28 *Streptococcus pneumoniae* is a human respiratory tract pathogen that represents a leading cause

29 of invasive bacterial disease in children under 5 years of age and the elderly. Isolates of *S.*

30 *pneumoniae* are traditionally characterised in terms of the antigenicity of their capsular

31 polysaccharides (CPS) and currently there are 92 serologically distinct serotypes as defined by

32 Statens Serum Institut (SSI), Copenhagen, Denmark sera using the Neufeld reaction for capsular

33 swelling (referred to as Quellung)(Austrian; Statens Serum Institut, 2013). Additional 'serotypes'

34 (6E, 6F, 6G, 6H, 11E, 20A and 20B) have been identified by molecular methods and/or

35 monoclonal antibodies, but these are not distinguishable using the current commercial serotyping

36 sera and some may not be phenotypically distinct types, but simply genetic variants producing

37 the same polysaccharide structures (Calix & Nahm, 2010; Calix et al., 2012; Ko, Baek & Song,

38 2013a; Oliver et al., 2013a; Park et al., 2015; Burton et al., 2016). The polysaccharide capsule is

39 also the target of all currently licensed vaccines and therefore, due to selective pressure on

40 circulating strains, accurate identification of pneumococcal serotypes is essential for disease

41 surveillance, evaluation of the efficacy of the pneumococcal vaccines and to inform national

42 vaccine policy.

43 In the UK, the initial reduction of invasive pneumococcal disease (IPD) and carriage, due to the

44 introduction of PCV7 and PCV13 polysaccharide-conjugate vaccines, was followed by increases

45 in non-vaccine type IPD(Miller et al., 2011; Waight et al., 2015). Specifically, following

46 implementation of pneumococcal conjugate vaccine PCV13(Centers for Disease Control and

47 Prevention (CDC)., 2010) in 2010, changes in disease-causing serotypes included a significant

48 reduction in PCV13 serotypes 1, 6A, 7F and 19A, followed by a substantial increase in non-

49   vaccine serotypes 8, 12F, 15A, 15B/C, 22F, 23B and 24F observed in 2013/14 in children under

50   5 years of age, when compared with 2012/13 (Waight et al., 2015) (Figure 1). Surveillance

51   studies from other countries revealed a similar pattern of PCV13-serotype reduction and non-

52   vaccine serotype increase in IPD cases. These studies underline the need for continuous

53   surveillance to monitor the emergence of serotypes due to the clonal expansion of non-vaccine

54   serotypes (Richter et al., 2013; Regev-Yochay G., Paran Y., Bishara J., Oren I., Chowers M.,

55   Tziba Y., Istomin V., Weinberger M., Miron D., Temper V., 2015).

56   The Quellung reaction, in conjunction with the antisera from SSI, is currently considered the

57   gold standard for pneumococcal capsular typing and can identify all 92 serotypes (Austrian;

58   Habib, Porter & Satzke, 2014). Agglutination tests were developed later to reduce the time

59   required to serotype a batch of isolates by eliminating the need for microscopy. In these tests SSI

60   antisera reacts with the pneumococcal capsule and produces visible clumping (Kronvall G, 1973;

61   Porter, Ortika & Satzke, 2014). Based on the SSI serotype scheme there are 25 distinct serotypes

62   and 21 serogroups, each consisting of between 2 and 5 immunologically-related serotypes,

63   accounting for a further 67 serotypes (Statens Serum Institut, 2013). Despite the wide use of the

64   SSI sera, the serological methods can be difficult to interpret since, in certain cases, serotypes

65   within a serogroup can cross-react with some of the antisera making it impossible to resolve to

66   serotype level. In these cases, the serotypes are reported to serogroup level (Jauneikaite et al.,

67   2015).

68   In 2006, Bentley *et al.* (Bentley et al., 2006)*,* reported the sequences of the capsular biosynthetic

69   loci of 90 pneumococcal serotypes (serotypes 6C and 6D were not included) and linked the

70   capsular gene sequences to known structural and immunological patterns, thus paving the way

71   for molecular capsular typing. Since then, new molecular serotyping methods have been

72   described (microarray (Turner et al., 2011), multiplex PCR (Brito, Ramirez & Lencastre, 2003),

73   *in silico* inference from WGS data (Croucher et al., 2011; Everett et al., 2012; Metcalf et al.,

74   2016)), utilising the molecular differences between the capsular loci to determine serotype

75   (Jauneikaite et al., 2015). A review of methodologies for capsular typing of *S. pneumoniae*

76   including serological assays, semi-automated molecular tests, and one directly from WGS

77   data(Liyanapathirana et al., 2014) was published by Jauneikaite *et al* [14]. These molecular/WGS

78   assays have some advantages compared to the serological methods described above: easier

79   interpretation, multiplex capability and in some cases no requirement for culture. However, they

80   still have some limitations, especially if they were to be used for surveillance. Firstly, due to the

81   genetic similarities of certain serotypes, not all serogroups can be resolved. Secondly, most

82   assays described to date only include the most common serotypes so currently not all serotypes

83   can be detected (Magomani et al., 2014). Here we describe genetic capsular locus differences for

84   all but four serotypes and demonstrate the use of these differences in predicting capsular type

85   from WGS data using an automated bioinformatics tool that can be incorporated into routine

86   workflows.

87

88   **Materials and Methods**

89   **Isolate selection**

90   Reference strains (n=91) for all serotypes (excluding 6D) were acquired from Statens Serum

91   Institut. Reference strain for 6D was kindly provided by The National Institute for Health and

92   Welfare (THL), Finland. A total of 926 clinical isolates were selected from the archives of the

93   Public Health England (PHE) National Reference Lab as a test cohort; for serotypes found to

94   belong to a genogroup, at least 10 isolates were selected where available. Post genomic-sequence

95 data cleansing (to remove repeat isolates from the same patient, mixed cultures, other species and

96 MLST partial profiles or failures) resulted in 871 isolates (Development Set in Table 1). In

97 addition, 2079 prospective or research-related isolates were sequenced as part of the UK

98 validation cohort. This cohort covers 72 of the commonly circulating serotypes (including all

99 vaccine serotypes), and includes prospective isolates received by PHE during 2015, isolates

100 selected as part of research projects and epidemiological investigations (15A (n=196) and 19A

101 (n=249), respectively) and archived isolates for rarer serotypes. Post genomic-sequence data

102 cleansing of this dataset resulted in a total of 2065 isolates (Validation Set in Table 1). All

103 isolates were serotyped on receipt as part of the PHE enhanced surveillance programme using

104 slide agglutination with Statens Serum Institut typing sera.

105 Genomic data for non-UK isolates were obtained from *Streptococcus pneumoniae* isolate

106 database hosted in BIGSdb (http://pubmlst.org/software/database/bigsdb/)(Jolley & Maiden,

107 2010) and the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena). Specifically,

108 three collections were used; a set of 2531 isolates from Thailand initially described by

109 Chewapreecha et al(Chewapreecha et al., 2014), an Icelandic panel of 252 serogroup 6 isolates

110 described in van Tonder et al(van Tonder et al., 2015) and a USA panel of 181 invasive isolates

111 available in ENA as study SRP059723.

112 **DNA extraction and sequencing**

113 Isolates were grown overnight on horse blood agar (PHE Media Services) with 5% CO_2. DNA

114 was extracted from an entire plate of growth for each isolate using the QIAsymphony SP

115 automated instrument (Qiagen) and QIAsymphony DSP DNA Mini Kit, using the

116 manufacturer's recommended tissue extraction protocol for Gram negative bacteria (including a

117 1 hour pre-incubation with proteinase K in ATL buffer and RNAse A treatment). DNA

118   concentrations were measured using the Quant-iT dsDNA Broad-Range Assay Kit (Life

119   Technologies, Paisley, UK) and GloMax® 96 Microplate Luminometer (Promega, Southampton,

120   UK). DNA was sent for whole genome sequencing (WGS) by Illumina sequencing using the

121   PHE Genomic Services and Development Unit (Colindale, UK) (Dallman et al., 2014). Illumina

122   Nextera DNA libraries were constructed and sequenced using the Illumina HiSeq 2500.

123

124   Casava 1.8.2 (Illumina inc. San Diego, CA,USA) was used to deplex the samples and FASTQ

125   reads were processed with Trimmomatic (Bolger, Lohse & Usadel, 2014) to remove bases from

126   the trailing end that fall below a PHRED score of 30. K-mer identification software

127   (https://github.com/phe-bioinformatics/kmerid) was used to compare the sequence reads with a

128   panel of curated NCBI Refseq genomes to identify the species. A sample of k-mers (DNA

129   sequences of length k) in the sequence data are compared against the k-mers of 1769 reference

130   genomes representing 59 pathogenic genera obtained from RefSeq. The closest percentage match

131   is identified, and provides initial confirmation of the species. This step also identifies samples

132   containing more than one species of bacteria (i.e. mixed cultures) and any bacteria misidentified

133   as *Streptococcus pneumoniae* by the sending laboratory. Further analysis continued only if *S.*

134   *pneumoniae* was identified. FASTQ reads from all sequences in this study were submitted to

135   ENA using the ena_submission tool (https://github.com/phe-bioinformatics/ena_submission) and

136   can   be   found   at   the   PHE   Pathogens   BioProject   PRJEB14267   at   ENA

137   (http://www.ebi.ac.uk/ena/data/view/PRJEB14267; Supplementary tables S1-2).

138   **Polysaccharide capsule operon locus sequence analysis**

139   GenBank and FASTA files for the capsular locus sequences for the 90 serotypes published by

140   Bentley *et al.* (Bentley et al., 2006) were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/)

141 using accession numbers CR931632–CR931722(Bentley et al., 2006). For serotypes 6C and 6D

142 accession numbers JF911515.1 and HV580364.1 were used, respectively. Capsular locus

143 sequences for 6E and 23B1, a new 23B genetic subtype identified during the course of this

144 project, were derived in-house using the assembly-based approach described below. Isolates for

145 both types (6E n=12; 23B1 n=27) were assembled using SPAdes(Bankevich et al., 2012) and the

146 capsular locus sequences were extracted using the sequence for 6B and 23B, respectively. The

147 extracted sequences were aligned using ClustalW(Thompson, Higgins & Gibson, 1994) and the

148 alignment was visualized with MEGA6(Tamura et al., 2013). Based on the alignment, consensus

149 sequences were derived for each type and submitted to ENA as assemblies under PHE Pathogens

150 BioProject PRJEB14267 with accession numbers LT594598.1 and LT594599.1 for 6E and 23B1,

151 respectively. Using the GenBank annotation, gene sequences were retrieved from all capsular

152 types. BLAST+ (version 2.2.27)(Camacho et al., 2009) alignment was then used to detect gene-

153 level differences within each genogroup.

154 **Capsular typing tool implementation**

155 PneumoCaT   (Pneumococcal   Capsule   Typing;   https://github.com/phe-

156 bioinformatics/PneumoCaT), written in Python (version 2.7.6), utilises a two-step method to

157 assign capsular type (Figure 2). In the first step, reads from each readset are mapped to a multi-

158 fasta file containing 94 capsular locus sequences (92 serologically distinct serotypes + 2

159 molecular types/subtypes – 6E and 23B1) using bowtie2 (version 2.1.0; following options used: -

160 -fr --minins 300 --maxins 1100 -k 99999 -D 20 -R 3 -N 0 -L 20 -I S,1,0.50) (Langmead B. and

161 Salzberg S., 2012). Thresholds of coverage of >90% of the length of the sequence, minimum

162 depth of 5 reads per bp and a mean depth of > 20 reads over the entire length of the sequence are

163 implemented during this step. This step is considered successful if one or more capsular locus

164 sequences are returned.

165 If a match to a single capsular locus is returned and the capsular type predicted does not belong

166 to one of our defined genogroups (Table 2), then the software terminates here and reports this as

167 the predicted capsular type (Figure 3). If the match belongs to a genogroup or more than one

168 locus is matched then the software moves to the second step; a variant-based approach. If

169 multiple serotypes are matched in step 1 and they do not correspond to the same genogroup then

170 a 'Mixed sample' flag is called. Finally, if no serotypes are called due to low coverage (<90%)

171 then a 'Failed' flag is called.

172   The capsular types matched in step 1 (genogroup) are fed into step 2 and the relevant

173 information for this specific genogroup is retrieved, if genogroup information is available. This

174 information is organised in a multi-fasta reference file, with complete gene sequences for all

175 genes that display genetic diversity within the selected genogroup, and a database containing all

176 sequence variants pertaining to the same genogroup. Paired reads are then mapped to the loaded

177 reference file using bowtie2 and the tool tests for the variants loaded from the capsular type

178 variant (CTV) database. Specific functions were designed to detect the different types of

179 variants, such as presence/absence of genes, different alleles, single nucleotide polymorphisms

180 (SNPs) and loss-of-function mutations. Base calling and insertion/deletion detection were made

181 with SAMtools mpileup (version 0.1.19; following options used: -B -A)(Li H., Handsaker B.,

182 Wysoker A., Fennell T., Ruan J., Homer N., 2009), requiring a minimum-read depth of 5 and

183 using reads with minimum base mapping quality of 30. These base calls were only included in

184 the downstream analysis if they constituted > 80% of read depth (ie, a nonmixed base call). A

185 capsular type is assigned if a 100% match to all variants tested is achieved, else a 'Failed' flag is

186   returned. In the case of serogroups 24 and 32, where no distinguishing variants are available, the

187   software returns the serogroup.

Phylogenetic tree generation

189   Reads were mapped to selected reference sequence using bwa (version 0.7.12)(Li & Durbin,

190   2009). Variants were called using GATK 2.6.5(McKenna et al., 2010). Variants were then parsed

191   to retain high quality SNPs based on the following conditions: depth of coverage (DP) $\geq$5, AD

192   ratio (ratio between variant base and alternative bases) $\geq$0.8, Mapping Quality (MQ)$\geq$30, ratio of

193   reads with MQ0 to total number of reads $\leq$0.05. All positions that fulfilled the filtering criteria

194   in >0.9 of the samples were joined to produce a multiple fasta format file where the sequence for

195   each strain consists of the concatenated variants. This file was used as an input to generate a

196   maximum likelihood (ML) tree using RAxML(Stamatakis, 2014) with the following parameters

197   –m (substitutionModel) ASC_GTRCAT --asc_corr=lewis (Paul Lewis's model correction for

198   ascertainment bias when analysing variant-only data) –b (bootstrapRandomNumberSeed) 12345

199   -# (numberOfRuns) 1000.

200

Assembly-based sequence analysis

202   Genomic reads were assembled using SPAdes (version 2.5.1) *de novo* assembly

203   software(Bankevich et al., 2012) with the following parameters 'spades.py --careful -1

204   strain.1.fastq.gz -2 strain.2.fastq -t 4 -k 21,33,55,77,85,93'. The resulting contigs.fasta file was

205   converted into a BLAST database using blast+(Camacho et al., 2009) (version 2.2.27) and

206   queried using selected query sequence (i.e. gene or capsular operon sequences).

207 **Results**

208 **Genogroups**

209 Genetic clustering of *S. pneumoniae* strains based on the capsular locus sequence was previously

210 described by Mavroidi *et al.* (Mavroidi et al., 2007). Based on their analysis, 88 pneumococcal

211 serotypes were grouped in 8 major clusters and 20 subclusters. All but 6 serogroups fell within

212 the same genetic subcluster as previously described(Mavroidi et al., 2007) and 9 of the

213 subclusters contained serotypes from different serogroups. In this study we define genogroups as

214 groups of serotypes with high (>90%) sequence identity within the capsular locus. Based on

215 these criteria 21 genogroups were identified by comparing published capsular locus

216 sequences(Bentley et al., 2006). All genogroups contain serotypes grouped within the same

217 genetic subcluster as defined by Mavroidi *et al* (Mavroidi et al., 2007) (Table 2); 11 of the

218 genogroups correspond to serogroups as defined by the SSI sera, 6 of which also form genetic

219 subclusters described by Mavroidi *et al*. (Mavroidi et al., 2007). Of the remaining 10

220 genogroups, 6 are identical to previously described subclusters and 3 correspond to incomplete

221 serogroups. For example, serogroup 10 assigned previously to a single genetic subcluster is split

222 into two genogroups, 10A-10B and 10C-10F. Similarly, serogroup 19 was assigned to a single

223 subcluster, whereas in our analysis only serotypes 19B and 19C form a genogroup. The

224 remaining two serotypes are too genetically dissimilar to be included; 19A shares only 50% of

225 the capsular locus with 19F, its closest genetic neighbour within the serogroup, whereas 19F

226 shares 65% with 19B. The last genogroup is 33A, 33F and 37; serotypes 33A and 33F belong to

227 a genetic subcluster; serotype 37 was not included in the previous analysis since it does not use

228 the Wzx/Wzy-dependent pathway. However, previous analysis revealed that serotype 37 isolates

229   use the *tts* synthase gene which is located elsewhere in the genome to define serotype, but still

230   contain a defective 33F-like capsular locus (Llull D., Munoz R., Lopez R., 1999).

Serotype-defining sequence variants

232   Serotype-defining variants were detected following gene-based alignment of published capsular

233   locus sequences (Bentley et al., 2006). Sequences were extracted in GenBank format and for

234   each genogroup, gene profiling and sequence variant analysis was performed to determine

235   variant profiles that could predict to serotype level. Types of sequence variation investigated

236   include presence/absence of capsular biosynthesis-related genes, detection of inactivating

237   mutations such as early stop codons and frameshifts (insertions, deletions, truncations), detection

238   of differing alleles of genes that can be distinguished by coverage alone due to low sequence

239   similarity (<80%) and SNP differences that can be detected by variant calling. Reported variants

240   were then evaluated with our test cohort (n=871; Table 2) and variants with 100% incidence rate

241   (found in all isolates of serotype tested) were incorporated into capsular type variant (CTV)

242   database (Table 3). Many of the variants detected during this study correspond to structural and

243   genetic differences reported previously (Kolkman, van der Zeijst & Nuijten, 1998; Bentley et al.,

244   2006; Mavroidi et al., 2007; Song, Baek & Ko, 2011; Yang et al., 2011; Oliver et al., 2013b).

245   In two cases of previously reported frameshift mutations (7A and 15C), the mutations were

246   confirmed in our test cohort, but no consensus of the inactivated sequence was detected (20-80%

247   of reads carry frameshift mutation). In 7As, an insertion of a thymine (T) at residue 587 of *wcwD*

248   inactivates the glycosyltransferase (Table 3) leading to loss of the side branch D-Galp-(β1-2)-α-

249   D-Galp(Mavroidi et al., 2007).   During evaluation with a cohort of 34 isolates (7A, n=9; 7F,

250   n=25 – Table 2), all 7F isolates carried a wild type (wt) *wcwD* gene sequence. In comparison, 7/9

251   of the 7A isolates also carried a wt *wcwD* sequence, whereas 2/9 (including the SSI 7A type

252   strain) exhibited a mixed profile at residue 587, with only a percentage of the reads carrying the

253   insertion (60-70%). All 7A isolates were retested in the laboratory using the SSI sera and only

254   the two isolates with the mixed profile were re-typed as 7A whereas the 7/9 with a wt *wcwD*

255   were typed as 7F, suggesting that this mixed profile is indicative of 7As and can still lead to

256   phenotypic changes. In 15Cs, differences in the length of the TA tandem repeat region of the

257   *wciZ* gene lead to loss of function for the O-acetyltransferase in 15C and are responsible for the

258   differences in acetylation between 15B and 15C isolates(Bentley et al., 2006). However, unlike

259   the 7A/7F scenario where the two types can be easily distinguished, distinguishing 15B and 15C

260   isolates on the bench using SSI sera in the slide agglutination method is extremely challenging.

261   Reversible switching between the   15B and 15C serotypes was previously described in the

262   laboratory and in natural infection(Venkateswaran, Stanton & Austrian, 1983; Van Selm et al.,

263   2003). Thus, in practice it may not be possible to clearly distinguish between these two serotypes

264   in the laboratory using serological methods and for the reasons above such isolates appear as

265   15B/C in many publications(Laufer et al., 2010). Therefore, it was unsurprising that the capsular

266   typing tool had problems matching the serological results where isolates had previously been

267   reported as either 15B or 15C prior to a change to the reporting; the previous result mostly likely

268   due to the dominance of one type over the other in the mix. During evaluation with 30 isolates

269   (15B, n=16; 15C n=14), we encountered 15B and mixed 15B/C isolates but no pure 15C isolates.

270   A genetically pure 15C was found later during validation but a different frameshift mutation was

271   responsible for the inactivation of *wciZ*. Overall 11/16 15B isolates were correctly predicted as

272   15B and 5/16 were mixed 15B/C whereas 10/14 15C isolates were predicted as 15B and 4/14 as

273   15B/C. Since any 15B-15C-15B/C discordances cannot be resolved in the laboratory the 15B-

274   15B/C-15C isolates will always be reported as 15B/C.

Molecular type 6E

Serotype 6E is not a serologically distinct serotype, , but has been shown to be a distinct molecular subtype(Ko, Baek & Song, 2013b). Due to the low genetic similarity between the 6E capsular locus sequence and the other serogroup 6 capsular locus sequences, PneumoCaT was unable to determine a serotype for isolates with 6E capsular locus in the first step of the capsular typing tool due to coverage < 90% (n=12; 10 6B and 2 6A). In order to take account of this variation, the genomic sequences of these 12 isolates were used to determine a consensus 6E capsular locus sequence and this was introduced into the capsular locus typing tool. With this in place the tool can determine whether an isolate has a 6E molecular type and using the variant profiles described above can then determine the serological type (i.e. 6E(6B)).

Novel serogroup 7 subtype

Genogroup 7B, 7C and 40 can be distinguished using a panel of 9 SNPs (Table S3) all found within glycosyltransferase gene *wcwK*. Nucleotide differences at residues 46 and 385 are considered as more significant, as both result in distinct amino acids in the three serotypes. During evaluation using a cohort of 38 isolates (7B, n=15; 7C, n=21; 40, n=2 – Table 1 - Development Set), one isolate originally typed as serotype 7C in the laboratory, exhibited a novel codon at residue 385 which encoded for a distinct amino acid (CTT (Leu) compared to ACT (Thr) for serotype 40, TTT (Phe) for 7B and TGT (Cys) for 7C). The rest of the SNP profile matched the 7B profile. Further investigation revealed a distinct combination of factor sera reactions (Table 4) for this isolate suggesting that it might correspond to a novel serogroup 7, serotype.

Serogroup 22

Biochemical or genetic differences between serotypes 22A and 22F have not been previously described. Initial genetic analysis revealed 99% identity over 100% of the capsular locus sequences of the two types (22A: CR931681; 22F: CR931682) suggesting a possible error in submission or serotyping. To investigate this, genomic data from the two SSI type strains (SSI-22A, SSI-22F) were assembled and using the reference sequence for 22A, the capsular locus sequences were extracted. The extracted sequences were compared to the reference sequences and the capsular locus sequence of SSI-22A was identical to the two reference sequences whereas the SSI-22F had 98% identity over 91% of the capsular locus of all three of the other sequences. Specifically, the region covering genes *wcwA* and *wcwC* shares no similarity between SSI-22F and 22A. The SSI-22F capsular locus sequence was submitted to ENA as assembly under PHE Pathogens BioProject PRJEB14267 (Accession No LT594600.1).

Novel molecular subtype 23B1

Serotypes 23A, 23B and 23F have distinct capsular locus sequences which allows for serotype prediction based on mapping coverage alone. A cohort of 72 isolates (23A, n=25; 23B, n=25; 23F, n=22) was used to test the coverage-based approach. All isolates typed as 23A or 23F and 21/25 isolates typed as 23B were correctly predicted; however 4 x 23B isolates gave a failed result in the capsular typing tool with highest coverage of approximately 70%. These isolates were re-tested in the laboratory using SSI sera and the previous results were confirmed. In order to investigate this further, more 23B isolates were sequenced and analysed using the capsular typing tool. In total 24/46 23B isolates failed the pipeline with 70% coverage. Following further investigation these isolates were identified as a novel molecular subtype (23B1) with ~70%

318   homology to the 23B capsular locus (unpublished data). The 23B1 capsular locus sequence was

319   introduced into PneumoCaT assuring 100% concordance with serological type.

Serogroup 24

321   Genetic differences between serotypes in serogroup 24 have previously been described(Mavroidi

322   et al., 2007) and were also confirmed following genetic analysis of the reference sequences

323   (Accession no. CR931686-8 for 24A, B and F respectively). Serotype 24A lacks *wzy,* a

324   polymerase gene and *rbsF,* a putative ribofuranose biosynthetic gene. In addition, the gene

325   sequence of the flippase gene *wzx* is unique to 24A as indicated by using a blast nucleotide query

326   against the NCBI database. This could only be confirmed by the SSI type strain (SSI-24A) since

327   no other 24A isolates were available in the PHE archive. According to Mavroidi *et al.* serotypes

328   24B and 24F are distinguished by loss of function of genes *abp1* and *rmlC* in 24B(Mavroidi et

329   al., 2007). This was confirmed when comparing the reference sequences. However, following

330   evaluation with a cohort of 25 isolates (6, 24B; 19, 24F) the frameshift in *rmlC* was not seen in

331   any of the 24B isolates, including the SSI-24B type strain whereas the frameshift in *abp1* was

332   only seen in SSI-24B strain, but no other 24B strains. This suggested that either all 24B isolates

333   are actually 24F or that the reference sequences are not representative of the circulating strains;

334   the serological typing was repeated for all isolates and original serotypes were confirmed,

335   therefore we surmised that the latter is more likely. Genomic SNP analysis was used to

336   investigate this further by mapping to either (a) the 24F capsular locus sequence (CR931688) or

337   (b) the whole genome of the non-capsular *S. pneumoniae* R6 strain (NC_003098). Subsequent

338   variant calling and filtering produced 145 and 19123 high quality SNPs, respectively.

339   Evolutionary history was inferred using the Maximum-likelihood method and phylogenetic trees

340   were drawn (Supplementary Figure 1). No serotype specific clustering was observed in either of

341   the trees suggesting that if there is a molecular basis for the observed serological differences, it

342   occurs using a mechanism that is not vertically inherited e.g rapid strand-slippage variation in a

343   microsatellite within a promoter region. Further investigation is underway but in the meantime,

344   PneumoCaT can predict isolates to serogroup level and serological analysis can distinguish

345   further to serotype level. Although genetic differences have been described above for

346   differentiating 24A from 24B/24F, they are not currently implemented in PneumoCaT due to

347   incomplete evaluation.

348   **Serogroup 32**

349   Structures for the capsular polysaccharide repeat unit of both serotype 32A and 32F are available

350   and differences based on the structure have previously been described(Bentley et al., 2006;

351   Mavroidi et al., 2007). However, when the two reference capsular locus sequences (32A:

352   CR931696; 32F: CR931697) were aligned, 99% identity was observed over 99% of the length of

353   the sequence. The only difference was a 5 bp gap at the intergenic region between *wcrN* and the

354   *HG272/3* pseudogene observed in the serotype 32 capsular locus. Gene-based analysis returned

355   no nucleotide differences within coding regions. To investigate whether there was a submission

356   error, as described for 22F, the capsular locus sequences (lacking the *tnp* regions at the beginning

357   of the capsular locus) of the 32A and 32F SSI type strains were extracted from the assembled

358   genomes of SSI-32A and SSI-32F strains and compared to the reference strains. Both type

359   sequences showed 100% identity with the respective reference sequences over the length of the

360   alignment. The serogroup 32 evaluation cohort only includes 5 isolates (32A, n=2; 32F, n=3 –

361   Table 1; Development Set) and following analysis with the capsular typing tool all isolates were

362   assigned to serogroup 32. Further investigation into the BAM files revealed that the 5 bp gap was

363   seen only in the SSI-32F type strain. More serogroup 32 isolates need to be analysed before any

364   conclusions can be reached, but no serogroup 32 isolates have been observed in the UK since

365   2007. At present the predicted serogroup 32 isolates can only be differentiated into serotypes

366   using serological analyses.

367   **Validation of the capsular typing tool**

368   Following development, the PneumoCaT tool was evaluated using two panels: (a) 2046 UK

369   isolates retrieved from PHE archive covering 72/92 serotypes, including all serotypes contained

370   in commercial vaccines (Table 5; excluding non-typeables) and (b) genomic data from 2964 non-

371   UK isolates from three distinct, publicly available datasets (Thailand; n=2531 - 65 serotypes,

372   USA; n=181 – serotype information not available- and Iceland; n=252 – serotypes 6A, 6B, 6C

373   and 6E).

374   Both panels were used to evaluate the typability of PneumoCaT based on the proportion of

375   isolates that were assigned a type. In the UK cohort, serotype was assigned initially to 98.2% of

376   the typeable isolates (2010/2046). Of the 36 failed isolates, 19 were called as mixed and

377   following retesting using a culture from a single colony, 18/19 were resolved, raising the

378   typability to 99.1%. In the non-UK cohort, a serotype was assigned to 99% of the isolates

379   (2934/2964) and 8/20 failed ones were called as mixed, suggesting that if retesting was possible

380   these could be resolved.

381   Concordance with currently used methodology (slide agglutination using SSI sera) was evaluated

382   using the UK cohort for which further laboratory testing was possible. Concordance was

383   estimated on the 2028 typeable isolates for which a serotype was called with both methods (2010

384   + 18 mixed resolved). Overall, 92.4% concordance was observed between the predicted

385   serotypes and the serologically-derived serotypes (n=1873). Discordance was observed in 155

386   isolates (Table 6) and in most cases (77%, n=119), the discordance involved serotypes using the

387   same factor sera set for serotype allocation (i.e. 12B/12F, 7A/7F discordances) or serotypes

388   using the same pool sera (i.e. serotype 31 and 7F both react with pool C serum)(Statens Serum

389   Institut, 2013) (n=22). Repeating the serological serotyping for these two sets resolved the initial

390   discordance in 88.7% of the cases (125/141). This suggests that discordance in these cases may

391   be attributed to difficulty in reading slide agglutination reactions during the serological analysis

392   with the SSI sera (i.e. auto-agglutination with all group 12 factor sera in the case of 12B/12F

393   typing). In one case a novel serogroup 9 serotype was predicted based on both the SNP-pattern

394   (Table S4) and sera reaction pattern that differed from the expected 9L pattern, showing a weak

395   reaction with factor sera 9g. In the remaining 14/155 cases, there are no shared sera between the

396   serotypes predicted by WGS and the serologically-derived serotypes. These isolates were

397   retested with both methods and all but one (unable to serotype due to autoagglutination) were

398   concordant suggesting a possible laboratory sample translocation during archiving and/or

399   retrieval of the isolate. Overall, following retesting, we observed a final concordance of 99.3%

400   (2013/2028), with only 18 discordant. In some of these cases serotyping could not resolve the

401   discordance because auto-agglutination leads to inconclusive results (n=10), whereas other cases

402   (n=8) exhibit a persistent discordance even after retesting with both methods (Table S9). In these

403   cases, the capsular operon sequence was investigated further using an assembly and blast

404   approach to query for both capsular operons. In all cases, the capsular operon sequence detected

405   did not match the serologically derived serotype. However, it must be noted that for each of these

406   serotypes, serotype was accurately predicted for the majority of cases and these discordant cases

407   are outliers (Table 5).

408   The serogroup 6 isolates from the Icelandic dataset (n=252) were originally used to demonstrate

409   the distribution of the new 6E type and molecular typing was used to characterized all isolates

410   (van Tonder et al., 2015). Although these isolates were not available for further investigation,

411   they were the only complete dataset with confirmed 6E isolates (molecular confirmation since

412   serological confirmation is not possible), so they were used to demonstrate the ability of

413   PneumoCaT to differentiate 6E from the other serogroup 6 types. Following analysis, 100%

414   typability and concordance was observed (n=252).

415   Reproducibility of the WGS method was investigated using a subset of 292 UK isolates covering

416   82/94 serotypes that were randomly selected for inclusion in the reproducibility study. For this

417   study each isolate was cultured twice and DNA extracted and sequenced at different times.

418   Concordance was seen in all cases.

419   **Non typeable isolates**

420   Virulence of *S. pneumoniae* is classically associated with the capsule, however some pathogenic

421   strains, often associated with non-invasive sites (i.e. respiratory isolates), are serologically non-

422   typeable (NT)(Park et al., 2012). Nineteen NT strains were selected from the PHE archive and

423   analysed using WGS and PneumoCaT to predict serotype. In 13/19 cases PneumoCaT failed to

424   predict serotype with a highest coverage <50% (Figure 4 and Table 5). This is consistent with

425   lack of a capsular operon and these isolates would be considered non-typeable. In addition, three

426   isolates that originally had been typed in the lab (19A and 2x38) but failed with <50% coverage

427   during PneumoCaT analysis were retested and found non-typeable in the lab, suggesting that this

428   cut off could be used an indicative of non-typeable isolates following further validation. In the

429   remaining 6/19 cases a serotype was predicted. In three of the cases retesting with SSI sera gave

430   the same serotype as predicted by PneumoCaT. In three other cases repeat serotyping by slide

431   agglutination consistently failed to give an acceptable result. In all of these cases the colonies

432   had an abnormal morphology (i.e. rough).

433   **Discussion**

434   Molecular capsular typing methods, such as multiplex PCR, have been readily used in reference

435   laboratories across the globe in place of the gold standard Quellung method("WHO | Laboratory

436   methods for the diagnosis of meningitis caused by Neisseria meningitidis, Streptococcus

437   pneumoniae, and Haemophilus influenzae"; European Centre for Diseases Prevention and

438   Control (ECDC), 2013). The lower cost and increased turnaround times associated with such

439   methods have made them the preferred serogrouping/typing method for a number of laboratories.

440   Many conventional and real-time PCR schemes cover the vaccine-related and most commonly

441   observed serotypes(Jauneikaite et al., 2015). However, these methods do not cover all serotypes

442   and many types cannot be distinguished from others with genetically closely related capsular

443   operons.   In this study, we present serotype-specific variant profiles that can be used to

444   distinguish 87/92 serologically-distinct serotypes and 2 molecular subtypes, and PneumoCaT, a

445   bioinformatics tool that uses the CTV (Capsular Typing Variant) database to predict capsular

446   type from WGS data.

447   A gene-based approach was used to identify variants that allow differentiation between closely-

448   related serotypes. During its development stage, a test cohort of 871 clinical isolates, including

449   the SSI-type strains, was used to evaluate the relevance of these variants in the population and

450   establish the CTV database. In many cases, these sequence variants were previously published

451   (e.g. serogroup 6)(Song, Baek & Ko, 2011), whereas in other cases structural differences have

452   been observed but no molecular information was provided (e.g. serogroup 9). One of the main

453   advantages of this approach is the ability to detect molecular subtypes (and potential new

454   phenotypes) even if no serological differences are observed using standard typing sera. For

455   example, isolates from serogroups 6 and 23 initially failed the development pipeline due to low

456   capsular locus coverage. Further investigation into the reason for the failures was able to identify

457   serotype 6E, a previously described molecular type(Ko, Baek & Song, 2013a) and a novel 23B

458   molecular subtype. Another example is serotype 19A, where two subgroups were identified

459   based on the mapping coverage of the 19A capsular locus. In this case, all isolates were still

460   predicted as 19A by PneumoCaT, since more than 90% coverage was observed for both

461   subgroups, but gene profile analysis revealed variability in the presence of cps19aO gene and

462   further functional and epidemiological analysis is currently underway to investigate whether this

463   has any phenotypic relevance. In terms of surveillance, the presence of these new molecular

464   subtypes could indicate introduction of new clones that may influence new pneumococcal

465   vaccine development. In some rare cases, a novel molecular type can also have corresponding

466   serological differences; for example, during the test phase of the variant profiles for genogroup

467   7B-7C-40, an isolate exhibited a distinct codon at position 385 and a distinct factor sera reaction

468   pattern (Table 4) upon retesting in our laboratory. Following further investigation by SSI this

469   isolate has now been confirmed as a novel serotype (authors' unpublished data). In addition,

470   during the validation phase of this study, SNP analysis revealed a novel serogroup 9 type with a

471   distinct sera reaction pattern, but this has not been confirmed.

472   Following the development stage, PneumoCaT was evaluated for typability, accuracy

473   (concordance with serological results) and reproducibility. Typability was evaluated as the

474   percentage of isolates for which a serotype was assigned by PneumoCaT, using two panels of

475   isolates (UK, n=2046; non-UK, n=2964), that combined, cover 82/94 capsular types, including

476   clones that cover 4 different countries (UK, USA, Iceland and Thailand). Using both panels, a

477   combined typability of 99% (99.1% for UK and 99% for non-UK) was achieved, demonstrating

478   that PneumoCaT can be used to analyse isolates from diverse geographical lineages. The panel

479   of 2028 typeable UK isolates (encompassing 72 serotypes) was then used to evaluate the

480   accuracy of the tool. When discordances between phenotypic serotyping results and PneumoCaT

481   serotype were investigated, the majority (119/155) were shown to be due to problems with

482   determining serotype within serogroup, when using the factor sera in the slide agglutination

483   method. Particular serogroups such as 12, 33 and 35 proved especially problematic (Table 5).

484   This could be due to the way the results are interpreted for the factor sera reactions when used in

485   slide agglutination. For example, in 24/69 Serogroup 12 cases a 12B/12F discordance was

486   observed; 18 serotype 12B isolates were assigned 12F capsular type by PneumoCaT whereas 6

487   12F isolates were assigned 12B. Based on the SSI typing scheme a reaction with all three

488   serogroup 12 factor sera (12b, 12c, 12e) is required to identify 12B, whereas a reaction with

489   factor serum 12b only is indicative of 12F isolates. Unfortunately, some factor sera agglutination

490   reactions can be weaker than others. Therefore, a weak reaction in all factor sera could be

491   mistaken for auto-agglutination or *vice versa*, leading to incorrect scoring. In most discordant

492   cases, repeating the phenotypic testing resolved the discordance. In some cases (n=10) a weak

493   reaction or auto-agglutination made interpretation of the serotyping result difficult, thus the

494   serotype could not be confidently called using the slide agglutination method with the standard

495   SSI sera. The standard factor sera obtained from Statens Serum Institut are optimised for use in

496   the Quellung reaction rather than slide agglutination and therefore, it is highly likely that if the

497   serotyping was performed using the gold-standard Quellung method by technicians with the

498   necessary experience, this could resolve the discrepancies in many cases and increase the

499   accuracy of the serotyping. The use of latex absorbed sera specifically designed for agglutination

500   reactions is also expected to improve the accuracy of the serotyping results for the more difficult

501   to distinguish serotypes. Late in this study a limited number of latex-absorbed subtyping sera

502   were kindly provided by SSI for use in agglutination reactions, and were used to re-serotype all

503   discordant serogroup 12 isolates. The latex sera was able to resolve 22/24 discordances and in

504   2/24 cases a serotype could not be confidently called due to autoagglutination.

505   In addition, some of the observed original discordances clearly came from simple record keeping

506   errors due to the nomenclature of the factor sera being very similar to the nomenclature for the

507   serotypes themselves, such as serotype 15A which is reacts with 15c factor sera, which

508   accounted for several discordant cases where 15A isolates had been originally recorded as 15C

509   due to 15c factor positive result. Similar discordances were also observed in serogroup 6 where

510   6A reacts with 6b factor sera, 6B with 6c and 6C with 6d factor sera resulting in mis-reporting.

511   This kind of error could be avoided by use of automated reporting mechanisms which take

512   electronically derived results (such as the PneumoCaT output) and enter them directly into

513   laboratory information systems, avoiding manual transcription.

514   Where occasional discordance persisted even after retesting, it involved pairs of serotypes within

515   the same serogroup or with known cross-reactions (Table S9).In these persistent cases, further

516   molecular investigation is underway to determine whether any regions outside of the capsular

517   operon might be involved in this discordance between phenotype and serotype. However, a

518   repeated serotyping error with these particular isolates cannot be excluded since other examples

519   of the same serotypes have concordance between WGS and serology. For example, one of the

520   isolates is serologically 29 where PneumoCaT predicted 35B. The two capsular operon

521   sequences share approximately 70% of the capsular locus sequence with 87% similarity and can

522   be distinguished serologically based on the reaction of a single factor sera (i.e. 35a gives positive

523   reaction if 35B and negative if 29, and there are known cross-reactions between 29 and some of

524   the factor sera for group 35). The presence of a complete serotype 35B capsular operon in this

525   discrepant isolate was confirmed by assembly as well as mapping and following analysis of the

526   capsular locus sequence no differences were found between this isolate and other 35Bs,

527   including two that were originally mistyped as 29. At this point, further analysis for external

528   factors is underway.

529   Three of the persistent discordances involved serotype 18C; 2 assigned 18B and one 18A

530   serotype by PneumoCaT. Following retesting by both WGS and sera; 2 persisted with 18C/18B

531   discordance, whereas the third isolate was re-serotyped as 18B and retyped as 18A by WGS. The

532   18C/B discordance is not unexpected as both serotypes react with 18e factor sera whereas 18c

533   sera reacts only with 18C serotypes, suggesting that a weak reaction or autoagglutination could

534   lead to false prediction. However, 18B/A discordance is unexpected as the serotypes react with

535   two different factor sera (18d for 18A and 18e for 18B). Interestingly, the molecular difference

536   distinguishing 18A and 18B is the presence of a non-functional *wciX* gene in 18B; this difference

537   should have no impact on phenotype, but has successfully differentiated 19 18A/B isolates (18A,

538   n=11 and 18B, n=8). We further examined the capsular sequence to determine additional

539   differences that could impact the phenotype. Specifically, we looked into *glf* gene, which based

540   on previous studies(Bentley et al., 2006) was present as a pseudogene in 18B, 18C, and 18F.

541   However, when these gene sequences were translated it was evident that a large part of the UDP-

542   galactopyranase mutase domain was still encoded; specifically within the translated *glf* sequence

543   in 18B and 18C a 167 amino acid ORF was present and a 204 amino acid ORF in 18F. Both

544   ORFs matched to published UDP-pyranase mutases when compared to the NCBI protein

545   database using BLAST. The genomic reads of the 19 18A/B and the discordant 18B/A isolate

546 were mapped to the 502 nucleotide coding region of 18B *glf* and as expected *glf* was present in

547 all 18Bs and none of the 18As. However, *glf* was also found in the 18B/A isolate which also

548 lacks *wciX*, a genetic marker for 18A. This finding can potentially explain the discordance

549 between phenotype and genotype and incorporation of the *glf* marker in CTV will prevent any

550 future 18A/B discordances and instead flag them for further investigation.

551 Another of the discordant isolates involved a rare serotype (41A/41F) and was difficult to resolve

552 due to the lack of representative isolates of serogroup 41. If implemented into the routine

553 laboratory, isolates of these rare serotypes would be subjected to traditional serotyping in

554 addition to WGS analysis until the pipeline could be further developed or confidence in the

555 results was assured.

556 The high typability, concordance and reproducibility rates (99.1, 99.2 and 100%, respectively)

557 observed with PneumoCaT suggest the method is highly robust and reliable, with less

558 subjectivity than traditional serotyping and full traceability of results. The sensitivity of the

559 method for serotype mixtures is greater than that of the standard method which means that

560 careful preparation and handling of the isolates for DNA extraction and WGS is necessary to

561 avoid cross-contamination, but also means that potentially genuinely mixed serotype cultures can

562 be recognised. For example, an isolate was initially typed as 19A and predicted as mixed 19A/3

563 with WGS when retrieved from the archives for testing. It was found to have colonies with two

564 different morphologies; one very mucoid and growing over a smaller non-mucoid colony form.

565 The mucoid colony was sampled and following retesting serotype 3 was assigned using both

566 traditional serotyping and PneumoCaT. Unfortunately, the smaller colonies could not be sampled

567 in a pure form but were presumably the original serotype 19A, given the PneumoCaT result.

568 A further advantage of the PneumoCaT method is the extra information obtained in terms of the

569 percentage mapping to the capsular reference and SNPs identified, which could lead to the

570 discovery of novel serotypes. Trends in the mapping percentage (6E and 23B1) and differences

571 in SNP patterns for those entering the CTV detection pathway could be recorded (novel

572 serogroup 7 and serogroup 9 subtypes). These novel types may be missed by standard serotyping

573 methods due to the way the typing sera are cross-absorbed to avoid reactions with known types,

574 this however does not avoid cross-reactions with novel types e.g. 6C which initially cross reacted

575 with 6A serotyping factor sera prior to production of specific factor sera for this type(Lambertsen

576 & Kerrn, 2010).

577 A small number of isolates (n=11) failed during the first stage (coverage-based) of the capsular

578 typing pipeline and, upon further examination, repeat serotyping was also unable to confidently

579 call a type in 5 cases suggesting that they might be true acapsular, non-typeable isolates. In 2/5

580 cases, % coverage was between 75-90% suggesting that a capsular operon was present but not

581 expressed. The remaining 6 isolates fall just short of the 90% coverage threshold (Figure 4), but

582 the top hit corresponded to the expected serotype/serogroup suggesting that a capsule could still

583 be expressed (Table S10).

584 Non-typeable isolates, lacking a functional capsular operon, also get a 'Failed' tag by

585 PneumoCaT but can easily be distinguished, based on coverage levels, from unusual isolates

586 where serotype prediction failed. Based on our analysis of 19 isolates, serologically non-typeable

587 isolates have a highest coverage value of less than 50% whereas for unusual isolates whereas

588 some reactions with serotyping sera are seen when the coverage falls within 60-90%. With

589 further data to confirm this observation, a 'Non-typeable' flag could be introduced to

590 PneumoCaT (Figure 4). During our analysis of these non-typeable isolates, in 32% of the cases a

591 serotype was predicted by PneumoCaT, indicating a functional operon was present. Upon

592 retesting, half of these gave the serotype predicted and half were still non-typeable, indicating

593 possible external regulatory elements that inhibit expression of the capsule in these organisms.

594 In summary, PneumoCaT is a robust, accurate, sensitive and expandable tool that could

595 revolutionise pneumococcal reference microbiology. This tool has the sensitivity to enable the

596 recognition of mixed serotypes or new subtypes that could be masked by the use of other

597 methods. The flexibility to introduce new types to the CTV database with relative ease means the

598 system can evolve to suit future challenges. The CTV database itself has the potential to inform

599 the further development of other molecular methods, for example PCR or microarray analysis for

600 determination of capsular type and it could be particularly useful for developing non-culture

601 typing schemes. PneumoCaT will enable more detailed surveillance of serotype drift and could

602 be used to target further phenotypic analysis of potential new serotypes. The automated nature of

603 the tool means that it can be incorporated into routine pipelines and results can be populated into

604 laboratory information systems using custom scripts, thus avoiding some of the potential errors

605 associated with manual result recording and entry and suiting a role in reference microbiology.

606 **Acknowledgements**

607 We thank McDonald Prest, Doris Omoigui, Tim Chambers and Maimuna Kimuli for retrieving

608 hundreds of isolates from archives and preparing DNA for sequencing. Cath Arnold and the team

609 in Genome Services and Development Unit, PHE Colindale for sequencing all isolates. Ella

610 Campion, Gurkiran Mankoo and John Duncan for performing routine *S. pneumoniae* serotyping

611 and repeat testing some isolates. Steve Platt and Tony McNiff for LIMS integration

612 developments. Mark Van der Linden, Anni Virolainen-Julkunen and Siira Lotta for sending the

613 6D pneumococcus strains. Pernille Landsbo Elverdal, and team (Statens Serum Institut) for

614 preparing and sending latex factor sera.

615 **References**

616 Austrian R. The quellung reaction, a neglected microbiologic technique. *The Mount Sinai*

617   *journal of medicine, New York* 43:699–709.

618 Bankevich A., Nurk S., Antipov D., Gurevich AA., Dvorkin M., Kulikov AS., Lesin VM.,

619   Nikolenko SI., Pham S., Prjibelski AD., Pyshkin A V., Sirotkin A V., Vyahhi N., Tesler G.,

620   Alekseyev MA., Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its

621   applications to single-cell sequencing. *Journal of computational biology : a journal of*

622   *computational molecular cell biology* 19:455–77. DOI: 10.1089/cmb.2012.0021.

623 Bentley SD., Aanensen DM., Mavroidi A., Saunders D., Rabbinowitsch E., Collins M., Donohoe

624   K., Harris D., Murphy L., Quail MA., Samuel G., Skovsted IC., Kaltoft MS., Barrell B.,

625   Reeves PR., Parkhill J., Spratt BG. 2006. Genetic analysis of the capsular biosynthetic locus

626   from all 90 pneumococcal serotypes. *PLoS Genetics*. DOI: 10.1371/journal.pgen.0020031.

627 Bolger AM., Lohse M., Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence

628   data. *Bioinformatics*. DOI: 10.1093/bioinformatics/btu170.

629 Brito DA., Ramirez M., Lencastre H De. 2003. Serotyping Streptococcus pneumoniae by

630   Multiplex PCR Serotyping Streptococcus pneumoniae by Multiplex PCR. DOI:

631   10.1128/JCM.41.6.2378.

632 Burton RL., Geno KA., Saad JS., Nahm MH. 2016. Pneumococcus with the "6E" cps Locus

633   Produces Serotype 6B Capsular Polysaccharide. *Journal of clinical microbiology* 54:967–

634   71. DOI: 10.1128/JCM.03194-15.

635   Calix JJ., Porambo RJ., Brady AM., Larson TR., Yother J., Abeygunwardana C., Nahm MH.

636   2012. Biochemical, genetic, and serological characterization of two capsule subtypes among

637   Streptococcus pneumoniae serotype 20 strains: Discovery of a new pneumococcal serotype.

638   *Journal of Biological Chemistry*. DOI: 10.1074/jbc.M112.380451.

639   Calix JJ., Nahm MH. 2010. A new pneumococcal serotype, 11E, has a variably inactivated wcjE

640   gene. *The Journal of infectious diseases*. DOI: 10.1086/653123.

641   Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009.

642   BLAST+: architecture and applications. *BMC bioinformatics*. DOI: 10.1186/1471-2105-10-

643   421.

644   Centers for Disease Control and Prevention (CDC). 2010. Licensure of a 13-valent

645   pneumococcal conjugate vaccine (PCV13) and recommendations for use among children -

646   Advisory Committee on Immunization Practices (ACIP), 2010. *MMWR. Morbidity and*

647   *mortality weekly report* 59:258–61.

648   Chewapreecha C., Harris SR., Croucher NJ., Turner C., Marttinen P., Cheng L., Pessia A.,

649   Aanensen DM., Mather AE., Page AJ., Salter SJ., Harris D., Nosten F., Goldblatt D.,

650   Corander J., Parkhill J., Turner P., Bentley SD. 2014. Dense genomic sampling identifies

651   highways of pneumococcal recombination. *Nature genetics* 46:305–9. DOI:

652   10.1038/ng.2895.

653   Croucher NJ., Harris SR., Fraser C., Quail MA., Burton J., van der Linden M., McGee L., von

654   Gottberg A., Song JH., Ko KS., Pichon B., Baker S., Parry CM., Lambertsen LM., Shahinas

655   D., Pillai DR., Mitchell TJ., Dougan G., Tomasz A., Klugman KP., Parkhill J., Hanage WP.,

656   Bentley SD. 2011. Rapid pneumococcal evolution in response to clinical interventions.

657   *Science (New York, N.Y.)*. DOI: 10.1126/science.1198545.

658   Dallman TJ., Chattaway MA., Cowley LA., Doumith M., Tewolde R., Wooldridge DJ.,

659   Underwood A., Ready D., Wain J., Foster K., Grant KA., Jenkins C. 2014. An investigation

660   of the diversity of strains of enteroaggregative escherichia coli isolated from cases

661   associated with a large multi-pathogen foodborne outbreak in the UK. *PLoS ONE*. DOI:

662   10.1371/journal.pone.0098103.

663   European Centre for Diseases Prevention and Control (ECDC). 2013. *European Centre for*

664   *Disease Prevention and Control. Surveillance of invasive bacterial diseases in Europe,*

665   *2011*. DOI: 10.2900/1510.

666   Everett DB., Cornick J., Denis B., Chewapreecha C., Croucher N., Harris S., Parkhill J., Gordon

667   S., Carrol ED., French N., Heyderman RS., Bentley SD. 2012. Genetic Characterisation of

668   Malawian Pneumococci Prior to the Roll-Out of the PCV13 Vaccine Using a High-

669   Throughput Whole Genome Sequencing Approach. *PLoS ONE*. DOI:

670   10.1371/journal.pone.0044250.

671   Habib M., Porter BD., Satzke C. 2014. Capsular Serotyping of Streptococcus pneumoniae Using

672   the Quellung Reaction Video Link. *J. Vis. Exp* 84:512083791–51208. DOI: 10.3791/51208.

673   Jauneikaite E., Tocheva AS., Jefferies JM., Gladstone RA., Faust SN., Christodoulides M.,

674   Hibberd ML., Clarke SC. 2015. Current methods for capsular typing of Streptococcus

675   pneumoniae. *Journal of Microbiological Methods*:41–9. DOI:

676   10.1016/j.mimet.2015.03.006.

677   Jolley KA., Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the

678   population level. *BMC bioinformatics* 11:595. DOI: 10.1186/1471-2105-11-595.

679   Ko KS., Baek JY., Song JH. 2013a. Capsular gene sequences and genotypes of "serotype 6E"

680   streptococcus pneumoniae isolates. *Journal of Clinical Microbiology*. DOI:

681   10.1128/JCM.01645-13.

682   Ko KS., Baek JY., Song JH. 2013b. Capsular gene sequences and genotypes of "serotype 6E"

683   streptococcus pneumoniae isolates. *Journal of Clinical Microbiology*. DOI:

684   10.1128/JCM.01645-13.

685   Kolkman MA., van der Zeijst BA., Nuijten PJ. 1998. Diversity of capsular polysaccharide

686   synthesis gene clusters in Streptococcus pneumoniae. *Journal of biochemistry* 123:937–45.

687   Kronvall G. 1973. A RAPID SLIDE-AGGLUTINATION METHOD FOR TYP ING

688   PNEUMOCOCCI BY MEANS OF SPECIFIC ANTIBODY ADSORBED TO PROTEIN

689   A-CONTAINING STAPHYLOCOCCI. *Journal of Medical Microbiology*:187–190.

690   Lambertsen L., Kerrn MB. 2010. Test of a novel Streptococcus pneumoniae serotype 6C type

691   specific polyclonal antiserum (factor antiserum 6d) and characterisation of serotype 6C

692   isolates in Denmark. *BMC infectious diseases* 10:282. DOI: 10.1186/1471-2334-10-282.

693   Langmead B. and Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*

694   9:357–9.

695   Laufer AS., Thomas JC., Figueira M., Gent JF., Pelton SI., Pettigrew MM. 2010. Capacity of

696   serotype 19A and 15B/C Streptococcus pneumoniae isolates for experimental otitis media:

697   Implications for the conjugate vaccine. *Vaccine* 28:2450–7. DOI:

698   10.1016/j.vaccine.2009.12.078.

699   Li H., Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.

700   *Bioinformatics (Oxford, England)* 25:1754–60. DOI: 10.1093/bioinformatics/btp324.

701   Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N. et al. 2009. The sequence

702   alignment/map format and SAMtools. Bioinformatics. *Bioinformatics* 25:2078–9.

703   Liyanapathirana V., Ang I., Tsang D., Fung K., Ng TK., Zhou H., Ip M. 2014. Application of a

704   target enrichment-based next-generation sequencing protocol for identification and

705   sequence-based prediction of pneumococcal serotypes. *BMC microbiology* 14:60. DOI:

706   10.1186/1471-2180-14-60.

707   Llull D., Munoz R., Lopez R. GE. 1999. A Single Gene (tts) located outside the cap Locus

708   directs the formation of Streptococcus pneumoniae type 37 capsular polysaccharide. .

709   *Journal of Experimental Medicine* 190:241–251.

710   Magomani V., Wolter N., Tempia S., Du Plessis M., De Gouveia L., Von Gottberg A. 2014.

711   Challenges of Using Molecular Serotyping for Surveillance of Pneumococcal Disease.

712   *Journal of clinical microbiology* 52:3271–3276. DOI: 10.1128/JCM.01061-14.

713   Mavroidi A., Aanensen DM., Godoy D., Skovsted IC., Kaltoft MS., Reeves PR., Bentley SD.,

714   Spratt BG. 2007. Genetic relatedness of the Streptococcus pneumoniae capsular

715   biosynthetic loci. *Journal of Bacteriology*. DOI: 10.1128/JB.00836-07.

716   McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K.,

717   Altshuler D., Gabriel S., Daly M., DePristo MA. 2010. The Genome Analysis Toolkit: a

718   MapReduce framework for analyzing next-generation DNA sequencing data. *Genome*

719   *research* 20:1297–303. DOI: 10.1101/gr.107524.110.

720   Metcalf BJ., Gertz RE., Gladstone RA., Walker H., Sherwood LK., Jackson D., Li Z., Law C.,

721   Hawkins PA., Chochua S., Sheth M., Rayamajhi N., Bentley SD., Kim L., Whitney CG.,

722   McGee L., Beall B., Active Bacterial Core surveillance team. 2016. Strain features and

723   distributions in pneumococci from children with invasive disease before and after 13-valent

724 conjugate vaccine implementation in the USA. *Clinical microbiology and infection : the*

725 *official publication of the European Society of Clinical Microbiology and Infectious*

726 *Diseases* 22:60.e9–60.e29. DOI: 10.1016/j.cmi.2015.08.027.

727 Miller E., Andrews NJ., Waight PA., Slack MP., George RC. 2011. Herd immunity and serotype

728 replacement 4 years after seven-valent pneumococcal conjugate vaccination in England and

729 Wales: an observational cohort study. *The Lancet. Infectious diseases* 11:760–8. DOI:

730 10.1016/S1473-3099(11)70090-1.

731 Oliver MB., Van Der Linden MPG., Küntzel SA., Saad JS., Nahm MH. 2013a. Discovery of

732 streptococcus pneumoniae serotype 6 variants with glycosyltransferases synthesizing two

733 differing repeating units. *Journal of Biological Chemistry*. DOI: 10.1074/jbc.M113.480152.

734 Oliver MB., Jones C., Larson TR., Calix JJ., Zartler ER., Yother J., Nahm MH. 2013b.

735 Streptococcus pneumoniae serotype 11D has a bispecific glycosyltransferase and expresses

736 two different capsular polysaccharide repeating units. *The Journal of biological chemistry*

737 288:21945–54. DOI: 10.1074/jbc.M113.488528.

738 Park IH., Kim K-H., Andrade AL., Briles DE., McDaniel LS., Nahm MH. 2012. Nontypeable

739 pneumococci can be divided into multiple cps types, including one type expressing the

740 novel gene pspK. *mBio* 3. DOI: 10.1128/mBio.00035-12.

741 Park IH., Geno KA., Yu J., Oliver MB., Kim K-H., Nahm MH. 2015. Genetic, Biochemical, and

742 Serological Characterization of a New Pneumococcal Serotype, 6H, and Generation of a

743 Pneumococcal Strain Producing Three Different Capsular Repeat Units. *Clinical and*

744 *Vaccine Immunology*. DOI: 10.1128/CVI.00647-14.

745 Porter BD., Ortika BD., Satzke C. 2014. Capsular Serotyping of Streptococcus pneumoniae by

746   Latex Agglutination Video Link. *J. Vis. Exp*. DOI: 10.3791/51747.

747 Regev-Yochay G., Paran Y., Bishara J., Oren I., Chowers M., Tziba Y., Istomin V., Weinberger

748   M., Miron D., Temper V. RG and DR. 2015. Early impact of PCV7/PCV13 sequential

749   introduction to the national pediatric immunization plan,on adult invasive pneumococcal

750   disease:A nationwide surveillance study. *Vaccine*:1135–42.

751 Richter SS., Heilmann KP., Dohrn CL., Riahi F., Diekema DJ., Doern G V. 2013. Evaluation of

752   pneumococcal serotyping by multiplex PCR and quellung reactions. *Journal of Clinical*

753   *Microbiology*. DOI: 10.1128/JCM.01876-13.

754 Van Selm S., Van Cann LM., Kolkman MAB., Van der Zeijst BAM., Van Putten JPM. 2003.

755   Genetic Basis for the Structural Difference between Streptococcus pneumoniae Serotype

756   15B and 15C Capsular Polysaccharides. *Infection and Immunity*. DOI:

757   10.1128/IAI.71.11.6192-6198.2003.

758 Song JH., Baek JY., Ko KS. 2011. Comparison of capsular genes of Streptococcus pneumoniae

759   Serotype 6A, 6B, 6C, and 6D isolates. *Journal of Clinical Microbiology*. DOI:

760   10.1128/JCM.02628-10.

761 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of

762   large phylogenies. *Bioinformatics (Oxford, England)* 30:1312–3. DOI:

763   10.1093/bioinformatics/btu033.

764 Statens Serum Institut. 2013. *Streptococcus pneumoniae: Textbook in Diagnosis, Serotyping,*

765   *Virulence Factors and Enzyme-linked Immunosorbent Assay (ELISA) for Measuring*

766   *Pneumococcal Antibodies.*

767 Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. 2013. MEGA6: Molecular

768   evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*. DOI:

769   10.1093/molbev/mst197.

770   Thompson JD., Higgins DG., Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of

771   progressive multiple sequence alignment through sequence weighting, position-specific gap

772   penalties and weight matrix choice. *Nucleic acids research* 22:4673–80.

773   van Tonder AJ., Bray JE., Roalfe L., White R., Zancolli M., Quirk SJ., Haraldsson G., Jolley

774   KA., Maiden MCJ., Bentley SD., Haraldsson Á., Erlendsdóttir H., Kristinsson KG.,

775   Goldblatt D., Brueggemann AB. 2015. Genomics Reveals the Worldwide Distribution of

776   Multidrug-Resistant Serotype 6E Pneumococci. *Journal of clinical microbiology* 53:2271–

777   85. DOI: 10.1128/JCM.00744-15.

778   Turner P., Hinds J., Turner C., Jankhot A., Gould K., Bentley SD., Nosten F., Goldblatt D. 2011.

779   Improved detection of nasopharyngeal cocolonization by multiple pneumococcal serotypes

780   by use of latex agglutination or molecular serotyping by microarray. *Journal of Clinical*

781   *Microbiology*. DOI: 10.1128/JCM.00157-11.

782   Venkateswaran PS., Stanton N., Austrian R. 1983. Type Variation of Strains of Streptococcus

783   pneumoniae in Capsular Serogroup 15. *Journal of Infectious Diseases* 147:1041–1054.

784   DOI: 10.1093/infdis/147.6.1041.

785   Waight PA., Andrews NJ., Ladhani SN., Sheppard CL., Slack MPE., Miller E. 2015. Eff ect of

786   the 13-valent pneumococcal conjugate vaccine on invasive pneumococcal disease in

787   England and Wales 4 years after its introduction: an observational cohort study. *Articles*

788   *Lancet Infect Dis* 15:535–43.

789   WHO | Laboratory methods for the diagnosis of meningitis caused by Neisseria meningitidis,

790   Streptococcus pneumoniae, and Haemophilus influenzae

791   Yang J., Nahm MH., Bush CA., Cisar JO. 2011. Comparative structural and molecular

792   characterization of Streptococcus pneumoniae capsular polysaccharide serogroup 10. *The*

793   *Journal of biological chemistry* 286:35813–22. DOI: 10.1074/jbc.M111.255422.

794

795

796

797

Figure 1

Serotype distribution of UK pneumococcal isolates

Number of IPD isolates referred to PHE per epidemiological year (July 2011 to June 2015).

Serotypes with less than 10 isolates in all periods were removed.

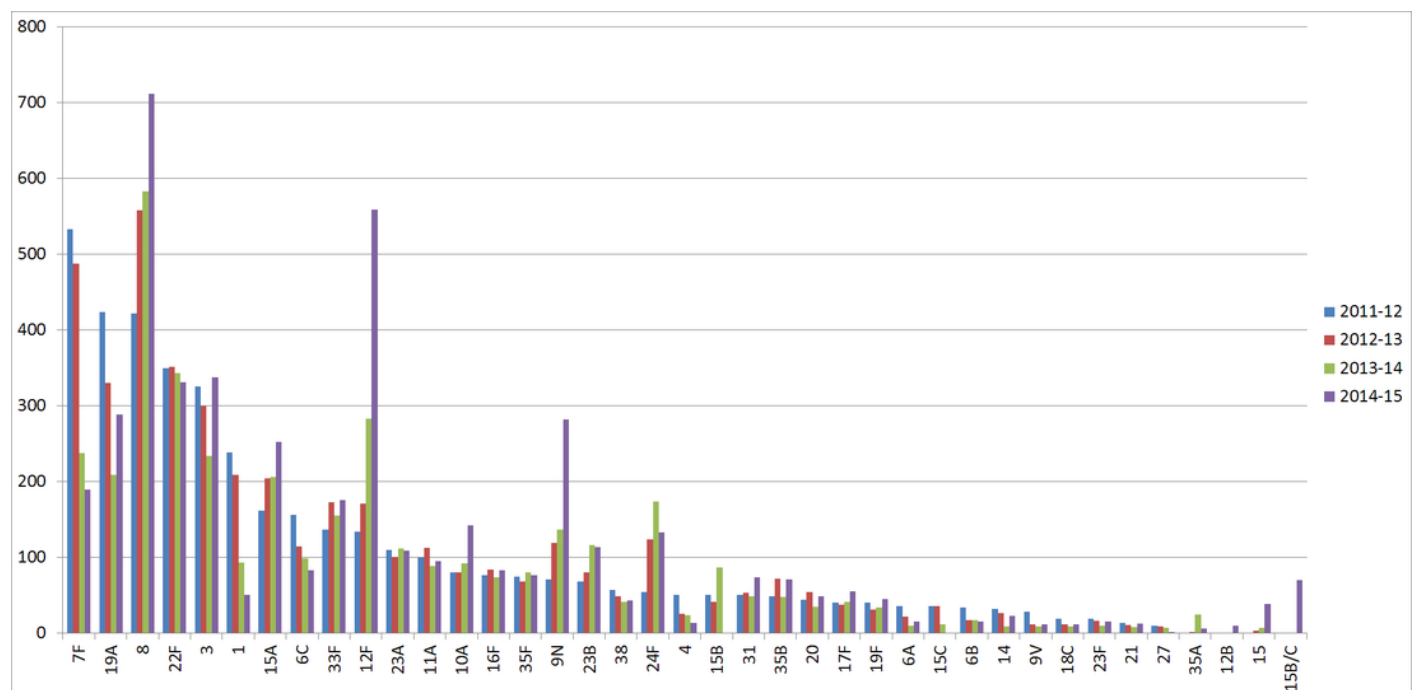



Figure 2

PneumoCaT process workflow

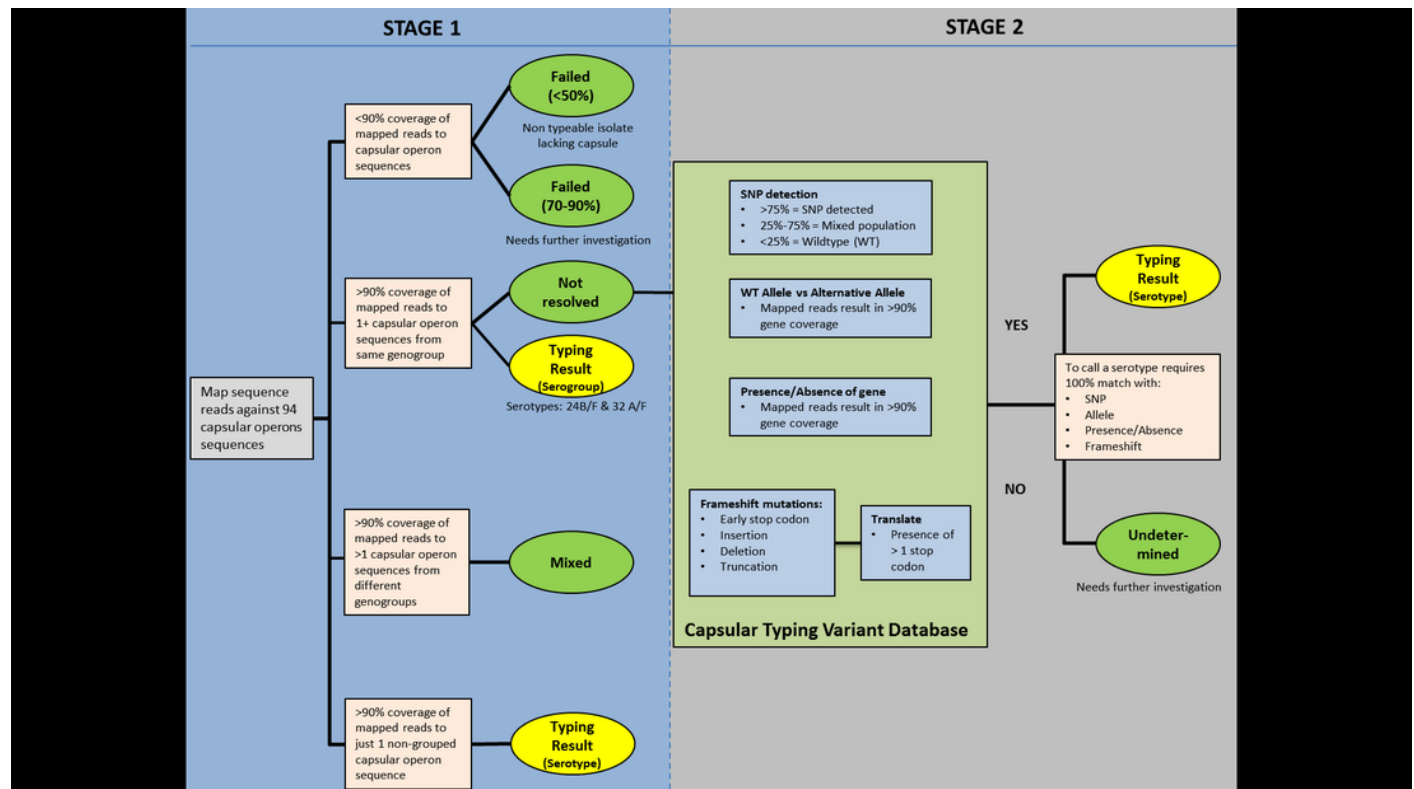



Figure 3

Average % coverage of the top and second hit during stage 1 of the PneumoCaT for serotypes determined by mapping coverage only.

Error bars correspond to standard deviation.

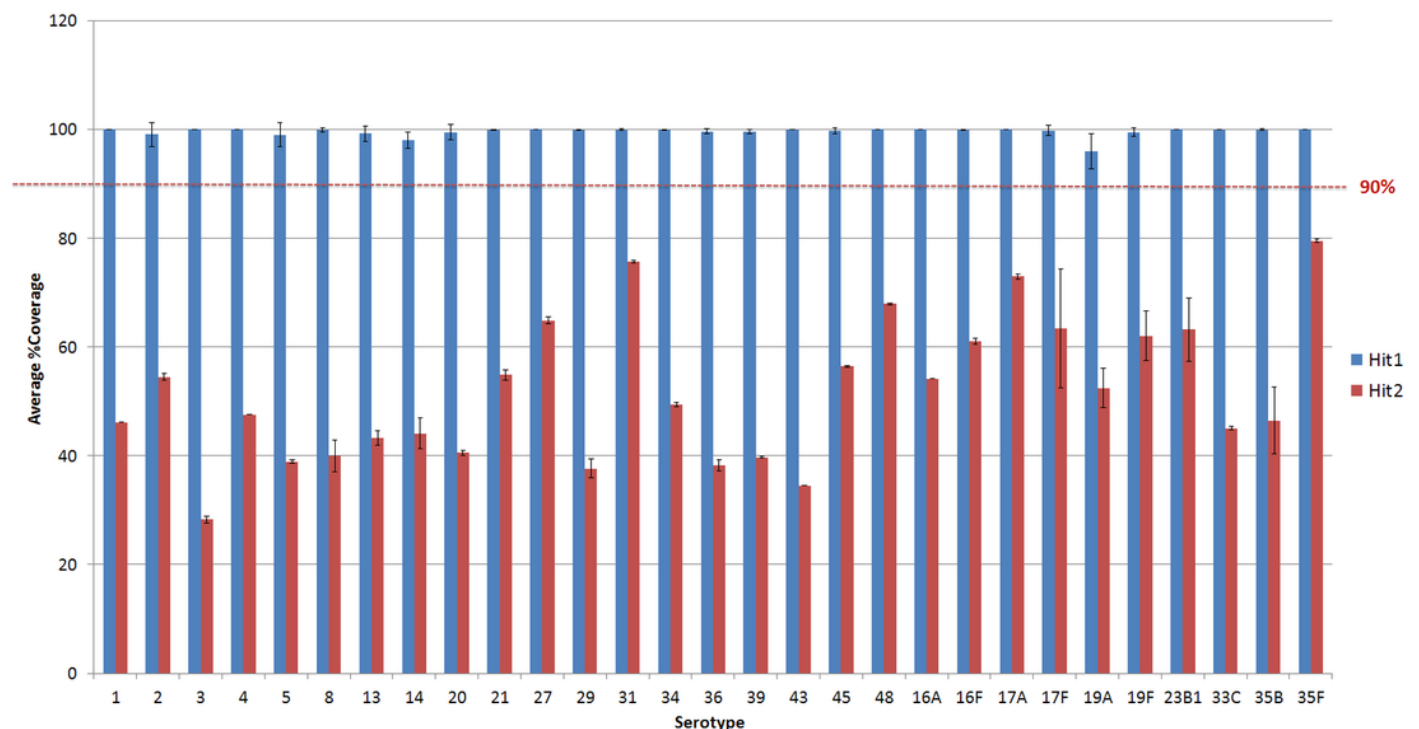
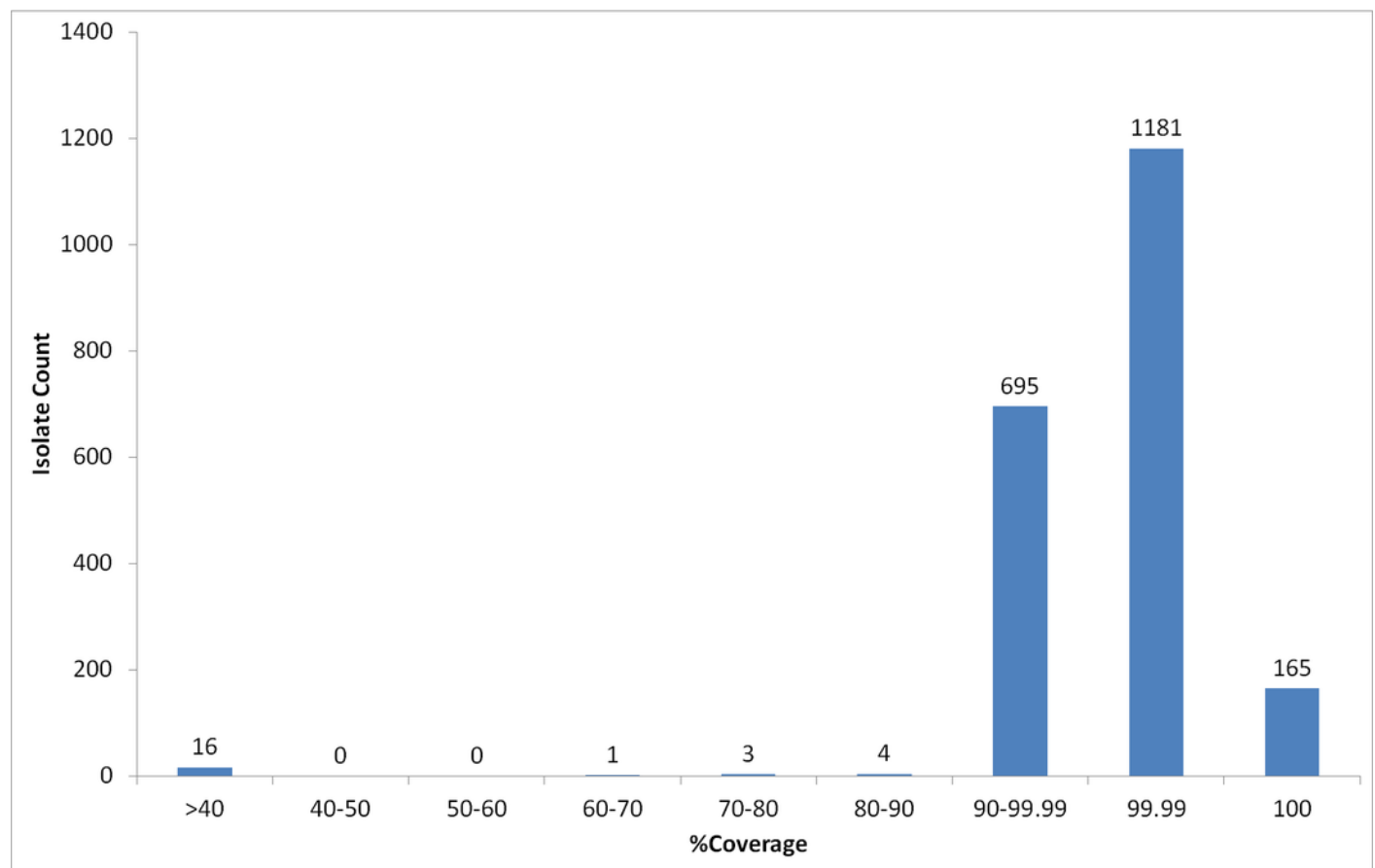



Figure 4

Percent coverage distribution of top serotype hit in PneumoCaT stage 1 for all *S. pneumoniae* isolates (n=2065) in the study.

Number of isolates from each serotype included in this study.

Serotype	Development Set	Validation Set	Total
1	4	41	45
2	3	9	12
3	5	44	49
4	3	43	46
5	3	41	44
8	7	70	77
13	3	20	23
14	5	43	48
20	3	42	45
21	3	21	24
27	5	23	28
29	2	21	23
31	8	22	30
34	3	22	25
36	2	5	7
37	8	22	30
38	16	23	39
39	2	2	4
40	2		2
42	2		2
43	1		1
44	1		1
45	2		2
46	6		6
48	2	5	7
06A	22	41	63
06B	23	43	66
06C	25	22	47
06D	5	2	7
07A	2	5	7
07B	15	4	19
07C	21	24	45
07F	32	40	72
09A	8	6	14
09L	9	2	11
09N	17	44	61
09V	33	45	78
10A	13	44	57
10B	2	7	9
10C	1		1
10F	24	22	46
11A	38	44	82
11B	6	2	8

11C	2	4	6
11D	1		1
11F	1		1
12A	4	2	6
12B	5	24	29
12F	28	44	72
15A	18	196	214
15B	17	41	58
15B/C		8	8
15C	14	24	38
15F	2	2	4
16A	1		1
16F	15	21	36
17A	3		3
17F	15	41	56
18A	4	11	15
18B	4	9	13
18C	5	41	46
18F	2	5	7
19A	28	249	277
19B	1	1	2
19C	1	1	2
19F	18	41	59
22A	9	2	11
22F	31	43	74
23A	26	23	49
23B	45	79	124
23F	22	41	63
24A	1	2	3
24B	6		6
24F	19		19
Serogroup 24		31	31
25A	3	1	4
25F	10	1	11
28A	15	19	34
28F	4	1	5
32A	2		2
32F	3		3
33A	3	5	8
33B	4	1	5
33C	2	1	3
33D	1		1
33F	19	43	62
35A	9	24	33

35B	29	23	52
35C	2	4	6
35F	13	22	35
41A	3	1	4
41F	2		2
47A	1		1
47F	1		1
NT		19	19
Grand Total	**871**	**2067**	**2938**

Comparison of genogroups defined in this study and genetic subclusters defined by Mavroidi et al.

Genogroups	Genetic subclusters
6A, 6B, 6C, 6D, 6E	6A, 6B
7A, 7F	7A, 7F
7B, 7C, 40	7B, 7C, 40
9A, 9L, 9V, 9N	9A, 9L, 9V, 9N
10A, 10B	10A, 10B, 10C, 10F
10C, 10F	
11A, 11B, 11C, 11D, 11F	11A, 11B, 11C, 11D, 11F
12A, 12B, 12F, 44, 46	12A, 12B, 12F, 44, 46
15A, 15B, 15C, 15F	15A, 15B, 15C, 15F
18A, 18B, 18C, 18F	18A, 18B, 18C, 18F
19B, 19C	19A, 19B, 19C, 19F
22A, 22F	22A, 22F
23A, 23B, 23F	23A, 23B, 23F
24A, 24B, 24F	24A, 24B, 24F, 17F, 48
25A, 25F, 38	25A, 25F, 38
28A, 28F	28A, 28F, 16F
32A, 32F	32A, 32F, 27
33A, 33F, 37	33A, 33F
33B, 33D	33B, 33D
35A, 35C, 42	35A, 35C, 42
41A, 41F	41A, 41F, 17A, 31

Table 3: Genetic differentiation of serotypes within serogroups.

Description of variants present in the CTV database.

Serogroup	Serotype	Distinguishing genetic features	Functional Effect
6	6A/6B and 6C/6D	A>G 583 in *wciP*	amino acid substitution (Ser195Asn) which results to different rhamnose-ribitol linkages (1 →3 in 6A/6C and 1→in 6B/6D)(Mavroidi et al., 2007; Sheppard et al., 2010; Baek et al., 2014)
	6A/6C and 6B/6D and 6E	*wciNα* in 6A and 6B / *wciNβ* in 6C and 6B / *wciNγ* in 6E	Allele *wciNα* encodes for galactosyl-transferase whereas *wciNβ* is 200 bp shorter and encodes for a glycosyl-transferase - consistent with changes in structure(Park et al., 2007) . *WciNγ* is a chimeric form of *wciNα* (75%) and *wciNβ* (25%).
7 and serotype 40	7A/7F	Frameshift mutation insT 587 in 7A *wcw*D gene	Loss of function of glycosyltransferase leading to loss of side branch for 7A(Mavroidi et al., 2007). *"Mixed: ['07A','07F']" result corresponds to 7A phenotype (see Results).
	7B/7C/40	SNPs in *wcwK (*Table S1)	Amino acid changes - *wcwK* encodes for a GT but 7C and 40 structure not known
9	9A/9V	Frameshift mutation delG 722 in 9A *wcjE*	Loss of function of O-acetyltranferase leads to differences in acetylation
	9L/9N	SNPs in genes wchA, wcjA, wcjB and wzy (Table S2)	Amino acid changes - *wcjA* and *wcjB* encode for glycosyltranferases (GT) and changes in these are consistent with presence of glucose in 9N instead of galactose present in residue 3 of the polysaccharide repeat unit of the other three serotypes(Mavroidi et al., 2007)
	9A/9V/9L/9N	presence of an additional O-acetyltransferase encoded by *wcjD* in 9A-9V	Differences in acetylation
10	10A/10B/10C/10F	10A/10B carries gene *wcrG*, whereas 10C/10F carries genes *wcrH* and *wciG*	*wcrH* encodes for GT and is responsible for side branch linkage Galf(1-6)Galp present in 10F but not in 10A; w*crG* encodes for GT and it catalyzes the linkage of Galp(1-6) side branch in 10A(Aanensen et al., 2007)

	10A/10B/10C/10F	10A/10C have *wcrCα* whereas 10B/10F have *wcrCβ*	*wcrCβ* allele is described as *wcrF* and both genes encode for glucosyltransferases and are responsible for the differences observed in the linkage between galactose and ribitol-5-phosphate(Yang et al., 2011)
11	11A/11B/11C/11D/11F	genes *wcwC* and *wcjE* are present in 11A, 11D and 11F whereas gene *wcwR* is present in 11B and 11C(Mavroidi et al., 2007)	*wcwC, wcjE* and *wcwR* are acetyltransferase genes - differences in acetylation
	11A/11B/11C/11D/11F	Frameshift mutation delA 130 in *gct* in 11B and 11F	presence of Gro-1P correlates with an intact *gct* gene in types 11A and 11C; *gct* is frameshifted in types 11F and 11B, and Rib-ol is present in the CPS instead of Gro(Mavroidi et al., 2007)
	11A/11D/11F	*wcrL pos 334*: codon AAT (Asn) in 11A; codon ACT (Ser) in 11D(Oliver et al., 2013) and codon GCT (Ala) in 11F	*wcrL* encodes for a GT - donor sugar for WcrL is GlcpNAc in types 11F, 11B, and 11C but Glcp in type 11A(Mavroidi et al., 2007)
12 and serotypes 44 and 46	12A/12B/12F/44/46	SNPs in genes *wcxD* and *wcxF* (Table S3)	Both genes encode for GTs present only in this genogroup(Mavroidi et al., 2007)- effect on sugar chain unknown (no structure for 12B, 44 and 46)
15	15A/15B/15C/15F	15F has 4 additional genes; *glf, rmlB, rmlD* and *wcjE*(Bentley et al., 2006)	*glf, rmlB* and *rmlD* are involved in sugar biosynthesis; *wcjE* encodes for an acetyltraferase.

	15A/15B/15C	*15A wchL* has 81% identity in the first 300 bps compare to the allele found in 15B/15C, whereas 15A *wzd* has 69% identity in the last 300 bps when compared to the 15B/C allele	*wchL* encodes for a GT; *wzd* is involved in translocation of mature CPS to the cell surface and thus is responsible for determining the length of the capsule polysaccharide chain(Bentley et al., 2006)
	15B/15C	difference in TA tandem repeat region near position 413 of *wciZ,* leading to frameshift in 15C(Bentley et al., 2006)	*wciZ* encodes for an O-acetyltransferase - differences in acetylation. *15B, 15B/C and 15C results can be assigned (see Results section)
16	16A, 16F	mapping only	
17	17A, 17F	mapping only	
18	18A/18B/18C/18F	18F has an extra acetyltransferase gene (*wcxM)* and type 18A lacks the acetyltransferase gene *wciX*(Mavroidi et al., 2007)	Differences in acetylation
	18B/18C	G>T 168 in *wciX* leads to early stop codon in 18B(Mavroidi et al., 2007)	*wciX* encodes for an acetylotranferase - difference in acetylation
19	19A, 19F	mapping only	
	19B/19C	19B lacks genes *wchU, HG264* and *glf*	*wchU* encodes for a putative GT and could be responsible for the additional glucose in the capsular polysaccharide repeat unit of 19C; *glf* encodes for a UDP-galactopyranose mutase whereas *HG264* has no functional product

22	22A/22F	*wcwA* and *wcwC* share no similarity between 22A and 22F.	*wcwA,* encoding for a putative glycosyl-transferase and *wcwC,* encoding for a putative O-acetyltranferase - structure for 22A unknown
23	23A/23B/23F	distinct *wzy* sequence in all serotypes	wzy encodes for a polymerase and differences in sequence should account for the different polymerization linkages(Mavroidi et al., 2007) - structures for 23A and 23B unknown
	23A/23B/23F	*wchA* is identical in 23B and 23F but distinct in 23A.	*wchA* encodes for a glycosyl-1-phosphatase transferase(Aanensen et al., 2007) - structures for 23A and 23B unknown
25 and serotype 38	25A/25F/38	*wcyV* missing in 38(Mavroidi et al., 2007)	*wcyV,* wcyD *and* wcyC encode for GTs(Aanensen et al., 2007) - no structures available for 25A, 25F or 38
	25A/25F/38	*wcyDα* in serogroup 25 and *wcyDβ* in serotype 38	
	25A/25F/38	SNPs in *wcyC* (Table S4)	
28	28A/28F	SNPs in *wciU* (Table S5)	*wciU* encodes for a GT - no structures available
33 and serotype 37	33A/33F/37	37 carries *tts* - a transferase gene	*tts* is responsible for the polysaccharide capsule synthesis in 37(Waite et al., 2003)
	33A/33F	Frameshift mutation insT 433 in 33F *wcjE* gene	Loss of function of O-acetyltranferase leads to differences in acetylation(Mavroidi et al., 2007)
	33B/33D	*wciNα* in 33B / *wciNβ*in 33C	wciNα encodes for a putative glycosyltranferase whereas wciNα encodes for a putative galactosyltransferase - consistent with differences in structure
	33C	mapping only	
35 and serotype 42	35B, 35F	mapping only	
	35A/35C/42	SNPs in genes *mnp1, wcrL* and *wzh* (Table S6)	*mnp1* encodes for a putative nucleotidyltranferase (NDP-mannitol pathway), *wcrL*, a GT and *wzh,* a protein-tyrosine phosphatase - consistent with differences in structure
	35A/35C/42	Frameshift mutation insA 248 in 35A *wcrK*(Mavroidi et al., 2007)	wcrK encodes for a GT - consistent with differences in structure

41	41A/41F	Frameshift mutation delG 23 in 41A *wcrX*(Mavroidi et al., 2007)	*wcrX* encodes for a acetyltranferase - differences in acetylation
47	47A, 47F	mapping only	

2 **References:**

3 Aanensen DM., Mavroidi A., Bentley SD., Reeves PR., Spratt BG. 2007. Predicted functions and linkage specificities of the products of the

4   Streptococcus pneumoniae capsular biosynthetic loci. *Journal of Bacteriology*. DOI: 10.1128/JB.00837-07.

5 Baek JY., Park IH., Song JH., Ko KS. 2014. Prevalence of isolates of Streptococcus pneumoniae putative serotype 6E in South Korea. *Journal of*

6   *Clinical Microbiology*. DOI: 10.1128/JCM.00228-14.

7 Bentley SD., Aanensen DM., Mavroidi A., Saunders D., Rabbinowitsch E., Collins M., Donohoe K., Harris D., Murphy L., Quail MA., Samuel G.,

8   Skovsted IC., Kaltoft MS., Barrell B., Reeves PR., Parkhill J., Spratt BG. 2006. Genetic analysis of the capsular biosynthetic locus from all 90

9   pneumococcal serotypes. *PLoS Genetics*. DOI: 10.1371/journal.pgen.0020031.

10 Mavroidi A., Aanensen DM., Godoy D., Skovsted IC., Kaltoft MS., Reeves PR., Bentley SD., Spratt BG. 2007. Genetic relatedness of the

11   Streptococcus pneumoniae capsular biosynthetic loci. *Journal of Bacteriology*. DOI: 10.1128/JB.00836-07.

12 Oliver MB., Jones C., Larson TR., Calix JJ., Zartler ER., Yother J., Nahm MH. 2013. Streptococcus pneumoniae serotype 11D has a bispecific

13   glycosyltransferase and expresses two different capsular polysaccharide repeating units. *The Journal of biological chemistry* 288:21945–54.

14   DOI: 10.1074/jbc.M113.488528.

15   Park IH., Pritchard DG., Cartee R., Brandao A., Brandileone MCC., Nahm MH. 2007. Discovery of a new capsular serotype (6C) within serogroup 6

16   of Streptococcus pneumoniae. *Journal of clinical microbiology* 45:1225–33. DOI: 10.1128/JCM.02199-06.

17   Sheppard CL., Pichon B., George RC., Hall LMC. 2010. Streptococcus pneumoniae isolates expressing a capsule with epitopes of both serotypes

18   6A and 6B. *Clinical and Vaccine Immunology*. DOI: 10.1128/CVI.00335-10.

19   Waite RD., Penfold DW., Struthers JK., Dowson CG. 2003. Spontaneous sequence duplications within capsule genes cap8E and tts control phase

20   variation in Streptococcus pneumoniae serotypes 8 and 37. *Microbiology (Reading, England)* 149:497–504. DOI: 10.1099/mic.0.26011-0.

21   Yang J., Nahm MH., Bush CA., Cisar JO. 2011. Comparative structural and molecular characterization of Streptococcus pneumoniae capsular

22   polysaccharide serogroup 10. *The Journal of biological chemistry* 286:35813–22. DOI: 10.1074/jbc.M111.255422.

Factor serum reactions for novel serogroup 7 serotype compared to other types within the group.

	7b	7c	7e	7f
7F	+	-	-	-
7A	(+)	+	-	-
7B	-	-	+	-
7C	-	-	-	+
novel	-	-	+	+

Detail breakdown of the concordance analysis for the UK validation panel (n=2065; 2046 typeable and 19 non-typeables).

* - if not possible to resolve discordance using serology a 15B/C serotype is reported

¥ - Novel 9 serotype by WGS and SSI sera

Serotype	Total	INITIAL			AFTER RETESTING		
		Concordant	Discordant	Failed WGS	Concordant	Discordant	Non-typeable
1	41	40	1		40		
2	9	8	1		8		
3	44	44			46		
4	43	43			44		
5	41	40	1		40		
8	70	70			72		
13	20	20			20		
14	43	40	3		40	1	
20	42	39	3		40		
21	21	21			21		
27	23	23			24		
29	21	9	11	1	9	1	1
31	22	22			23		
34	22	22			24		
36	5	3	1	1	3		
37	22	21	1		22		
38	23	19	1	3	19		1
39	2	2			2		
40					1		
42							
43							
44							
45							
46							
48	5	5			5		
06A	41	41			42		
06B	43	37	6		37		
06C	22	21	1		24		
06D	2	1	1		2		
07A	5	1	4		1		
07B	4	3	1		3		
07C	24	22	2		22		
07F	40	38	2		43		
09A	6	1	5		1		1
09L	2	1	1		1		
09N	44	42	2		43		
09V	45	45			49		
10A	44	41	3		42	1	
10B	7	6	1		6		
10C							
10F	22	19	1	2	20		
11A	44	40	2	2	46		

11B	2	2			2		
11C	4		4				
11D							
11F							
12A	2		1	1			
12B	23	5	18		9		
12F	44	38	6		58		2
15A	196	191	4	1	200		
15B*	41	40		1	42		
15B/C*	8	8			26		
15C*	24	16	6	2	1		
15F	2		2				
16A							
16F	21	21			25		
17A							
17F	41	41			42		
18A	11	11			11		
18B	9	8	1		8		
18C	41	38	3		39	3	
18F	5	3	2		3		
19A	249	245	3	1	246		
19B	1	1			1		
19C	1		1				
19F	41	41			42		
22A	2	2			2		
22F	43	43			47		
23A	23	22	1		25		
23B	79	78	1		90		
23F	40	27	11	2	27		
24A	2		2				
Serogroup 24	31	30	1		34		
25A	1		1				1
25F	1	1			1		
28A	19	19			19		
28F	1		1				
Serogroup 32							
33A	5		5				2
33B	1	1			1		
33C	1			1			
33D							
33F	43	43			48		
35A	24	4	20		4		1
35B	23	23			52		
35C	4		4			1	1
35F	22	21	1		22		

41A 41F 47A 47F	1		1			1	
NOVEL 9¥					1		
Total	**2046**	**1873**	**155**	**18**	**2013**	**8**	**10**
NT	19	13	6		13	3	
Grand Total	**2065**	**1886**	**161**	**18**	**2026**	**11**	**10**

Investigation of discordant isolates in the Validation Set

Isolate	Slide Agglutination		WGS	
	Initial	Repeat	Initial	Repeat
PHESPV0001	18F	3	3	
PHESPV0002	09L	Novel 9 pattern	Novel 9	Novel 9
PHESPV0012	09A	09V	09V	
PHESPV0013	09A	09V	09V	
PHESPV0016	12B	12F	12F	12F
PHESPV0024	35C	35B	35B	
PHESPV0025	07A	07F	07F	
PHESPV0027	09A	09V	09V	
PHESPV0029	18B	18C	18C	
PHESPV0070	06B	06D	06D	
PHESPV0071	NT	27	27	
PHESPV0101	09N	09N	23B	09N
PHESPV0112	06B	17F	17F	
PHESPV0121	09N	8	8	
PHESPV0128	1	15A	15A	
PHESPV0174	09A	09V	09V	
PHESPV0189	2	11A	11A	
PHESPV0194	29	22F	22F	
PHESPV0197	29	35B	35B	
PHESPV0200	29	34	34	
PHESPV0204	36	22F	22F	
PHESPV0209	23F	23B	23B	
PHESPV0211	29	35B	35B	
PHESPV0235	NT	35B	35B	
PHESPV0303	07A	07F	07F	
PHESPV0322	15A	15B/C	15B	
PHESPV0348	24A	24F	Serogroup 24	
PHESPV0349	35A	35B	35B	
PHESPV0380	24A	24F	Serogroup 24	
PHESPV0438	15A	15B/C	15B	
PHESPV0448	15F	15B	15B	
PHESPV0458	23F	23B	23B	
PHESPV0489	35A	35B	35B	
PHESPV0518	35A	35B	35B	
PHESPV0525	35A	35B	35B	
PHESPV0526	35A	35B	35B	
PHESPV0537	15A	15B/C	15B/C	
PHESPV0555	06B	06C	06C	
PHESPV0556	23F	23B	23B	
PHESPV0557	35A	35B	35B	

PHESPV0558	**11C**	11A	11A	
PHESPV0562	**15F**	15B	15B	
PHESPV0563	**33A**	33F	33F	
PHESPV0566	**23F**	23B	23B1	
PHESPV0582	**19A**	19F	19F	
PHESPV0586	**06B**	06C	06C	
PHESPV0591	**35A**	35B	35B	
PHESPV0605	**35A**	35B	35B	
PHESPV0608	20	11A	11A	
PHESPV0613	**35A**	35B	35B	
PHESPV0616	06B	33F	33F	
PHESPV0624	**35A**	35B	35B	
PHESPV0654	**35A**	35B	35B	
PHESPV0663	**35A**	35B	35B	
PHESPV0667	**12B**	12F	12F	
PHESPV0669	**35A**	35B	35B	
PHESPV0678	**23F**	23B	23B	
PHESPV0681	**12B**	12F	12F	
PHESPV0691	NT	37	37	
PHESPV0698	24F	07F	07F	
PHESPV0700	07F	20	20	
PHESPV0744	10A	16F	16F	
PHESPV0753	**35A**	35B	35B	
PHESPV0761	**12B**	12F	12F	12F
PHESPV0773	**35A**	35B	35B	
PHESPV0779	**12B**	12F	12F	
PHESPV0780	**12B**	12F	12F	
PHESPV0794	**12B**	12F	12F	
PHESPV0796	**12B**	12F	12F	12F
PHESPV0797	**12B**	12F	12F	
PHESPV0804	**35A**	35B	35B	
PHESPV0805	**33A**	33F	33F	
PHESPV0812	**10B**	10F	10F	
PHESPV0820	**12A**	12F	12F	
PHESPV0836	**23F**	23B	23B	
PHESPV0839	**11C**	11A	11A	
PHESPV0840	**33A**	33F	33F	
PHESPV0843	**35A**	35B	35B	
PHESPV0845	**23F**	23A	23A	
PHESPV0854	**06B**	06A	06A	
PHESPV0861	**35A**	35B	35B	
PHESPV0862	18F	23B	23B1	
PHESPV0888	**11C**	11A	11A	
PHESPV0891	**11C**	11A	11A	
PHESPV0894	37	16F	16F	

PHESPV0898	19A	3	3	
PHESPV0912	14	15B	15B	
PHESPV1001	29	35B	35B	
PHESPV1005	14	12F	12F	12F
PHESPV1014	35A	35B	35B	
PHESPV1017	23F	35F	35F	
PHESPV1018	14	23A	23A	
PHESPV1020	23F	23B	23B1	
PHESPV1084	20	22F	22F	
PHESPV1140	12F	12B	12B	12B
PHESPV1168	15A	15A	19F	15A
PHESPV1178	06C	23B	23B1	
PHESPV1200	07B	24F	Serogroup 24	
PHESPV1208	12F	12B	12B	12B
PHESPV1244	23A	15A	15A	
PHESPV1283	12F	12B	12B	
PHESPV1390	07F	31	31	
PHESPV1401	10A	22F	22F	
PHESPV1406	12F	12B	12B	
PHESPV1413	19A	8	8	
PHESPV1418	06D	06C	06C	
PHESPV1648	23F	23B	23B	
PHESPV1650	23F	23B	23B1	
PHESPV1652	07A	07F	07F	
PHESPV1654	35F	34	34	
PHESPV1668	11A	16F	16F	
PHESPV1716	07A	07F	07F	
PHESPV1734	11A	16F	16F	
PHESPV1790	20	24F	Serogroup 24	
PHESPV1809	28F	23B	23B	
PHESPV1845	15C	15A	15A	
PHESPV1847	15C	15A	15A	
PHESPV1848	15C	15A	15A	
PHESPV1864	5	4	4	
PHESPV1868	15C	15A	15A	
PHESPV1869	15C	15A	15A	
PHESPV1874	15C	15A	15A	
PHESPV1883	29	35B	35B	
PHESPV1919	29	35B	35B	
PHESPV1935	29	35B	35B	
PHESPV1941	10F	10A	10A	
PHESPV1951	35C	35B	35B	
PHESPV2005	12B	12F	12F	

PHESPV2007	**12B**	12F	12F
PHESPV2017	**23B**	23A	23A
PHESPV2018	**19C**	19A	19A
PHESPV2028	**12B**	12F	12F
PHESPV2032	**29**	35B	35B
PHESPV2035	**12B**	12F	12F
PHESPV2039	**12B**	12F	12F
PHESPV2042	**12B**	12F	12F
PHESPV2043	07C	33F	33F
PHESPV2049	**12B**	12F	12F
PHESPV2052	**12B**	12F	12F
PHESPV2062	**12B**	12F	12F
PHESPV2063	**07C**	40	40