Pherotype polymorphism in *Streptococcus pneumoniae* and its effects on population structure and recombination

Eric L. Miller[1,2], Benjamin A. Evans[1,3], Omar E. Cornejo[4], Ian S. Roberts[1§] and Daniel E. Rozen[1,2§]

Affiliations:

[1]Faculty of Biology, Medicine, and Health, University of Manchester, Manchester, United Kingdom

[2]Institute of Biology, Leiden University, Leiden, Netherlands

[3]Norwich Medical School, University of East Anglia, Norwich, United Kingdom

[4]School of Biological Sciences, Washington State University, Pullman, USA

[§] Corresponding authors: Ian S. Roberts

Oxford Road

Manchester M13 9PL

England, The United Kingdom

+44 161 275 7513

i.s.roberts@manchester.ac.uk

Daniel E. Rozen

Sylviusweg 72

2333 BE Leiden

The Netherlands

+31 (0) 71 527 7990

d.e.rozen@biology.leidenuniv.nl

Keywords: competence / pneumococcus / quorum sensing / balanced polymorphism / horizontal gene transfer

Abstract

Natural transformation in the Gram-positive pathogen *Streptococcus pneumoniae* occurs when cells become "competent", a state that is induced in response to high extracellular concentrations of a secreted peptide signal called CSP (Competence Stimulating Peptide) encoded by the *comC* locus. Two main CSP signal types (pherotypes) are known to dominate the pherotype diversity across strains. Using thousands of fully sequenced pneumococcal genomes, we confirm that pneumococcal populations are highly genetically structured and that there is significant variation among diverged populations in pherotype frequencies; most carry only a single pherotype. Moreover, we find that the relative frequencies of the two dominant pherotypes significantly vary within a small range across geographical sites. It has been variously proposed that pherotypes either promote genetic exchange among cells expressing the same pherotype, or conversely that they promote recombination between strains bearing different pherotypes. We distinguish these hypotheses using a bioinformatics approach by estimating recombination frequencies within and between pherotypes across 4,089 full genomes. Despite underlying population structure, we observe extensive recombination between populations; additionally, we found significantly higher rates of genetic exchange between strains expressing different pherotypes than among isolates carrying the same pherotype. Our results indicate that pherotypes do not restrict, and marginally facilitate, recombination between strains. Furthermore, our results suggest that the CSP balanced polymorphism does not causally underlie population differentiation. Therefore, when strains carrying different pherotypes encounter one another during co-colonization, genetic exchange can freely occur.

Introduction

The Gram-positive pathogen *Streptococcus pneumoniae* is responsible for up to 1 million deaths annually (O'Brien et al., 2009). *S. pneumoniae* is naturally transformable, and this ability to take up and recombine extracellular DNA across a broad size range, from 15 bp to 19 kb (Mostowy et al., 2014), is associated with the acquisition of antibiotic resistance genes and with capsular switching (Croucher et al., 2014a, 2014b, 2011). Transformation in *S. pneumoniae* occurs following the quorum-dependent induction of "competence", which is regulated by the secretion and detection of Competence Stimulating Peptide (CSP), encoded by *comC* (Håvarstein et al., 2006; Pestova et al., 1996). There are several alleles for *comC* and the gene encoding its cognate receptor, *comD*, although the vast majority of isolates carry either of two dominant mature signals encoded by *comC*, i.e. Csp-1 or Csp-2 (Evans and Rozen, 2013; Pozzi et al., 1996; Whatmore et al., 1999). Although these different allele combinations, referred to as pherotypes, are known to be mutually unresponsive, with isolates only responding to the CSP produced by isolates of their own pherotype (Iannelli et al., 2005), their role on the population structure of pneumococci remains unclear.

Two alternative hypotheses for the potential effects of multiple pherotypes on the population genetic structure of *S. pneumoniae* have been outlined. One hypothesis posits that because bacteria only bind and respond to their own CSP signal, pherotypes could ensure that bacteria only recombine with strains sharing the same pherotype, leading to a close association between pherotype and clonal structure (Håvarstein et al., 1997; Tortosa and Dubnau, 1999). In the second hypothesis, CSP-activated fratricide, whereby CSP-induced cells use secreted bacteriocins to kill uninduced cells, could alternatively facilitate recombination between strains with varying pherotypes. Importantly, this hypothesis assumes that strains of one pherotype preferentially kill strains of other pherotypes, after which the induced strains recombine with the DNA liberated from lysed cells (Claverys and Håvarstein, 2007; Claverys et al., 2006; Cornejo et al., 2010; Johnsborg et al., 2008). Although several recent studies have attempted to address the predictions of these models, results thus far are conflicting. While results from Carrolo et al. (Carrolo et al., 2014, 2009) support the idea that pherotypes facilitate within-pherotype recombination and therefore underlie population differentiation, results from Cornejo et al. (2010) are more consistent with the alternative. Importantly, the results of

these previous studies were limited to moderately small samples of strains, and analyses were performed on a limited number of markers (partial sequences of seven MLST loci), thereby reducing the ability to infer genomic-scale patterns of recombination. To overcome these limitations, our aim here is to address this question using a bioinformatics approach based upon analysis of recombination rates within and between pherotypes from 4,089 full pneumococcal genomes.

Materials and Methods

Genomic data

We obtained 4,089 *S. pneumoniae* genomes from five publicly available sets (Supplemental Table 1): 288 genomes from GenBank, which include 121 genomes of pathogenic strains from Atlanta, Georgia, The United States (Chancey et al., 2015); 3,017 genomes of carriage strains from Myanmar refugees in Maela, Thailand (Chewapreecha et al., 2014); 616 genomes of carriage strains from Massachusetts, the United States (Croucher et al., 2013a); 142 genomes of carriage strains from Rotterdam, the Netherlands (Bogaert et al., 2001; Miller et al., 2016); and 26 PMEN (Pneumococcal Molecular Epidemiology Network) genomes (McGee et al., 2001; Miller et al., 2016). These genomes were previously assembled and underwent quality control (Miller et al., 2016). We located *comC* and *comD* within these genomes using an iterative DNA reciprocal BLAST search as previously described (Miller et al., 2016). We excluded Rotterdam strain 724 and Maela strain 6983_6#45 from all pherotype analyses because the genomes of both strains carry two complete *comC* alleles encoding for both Csp-1 and Csp-2.

Phylogenetic analysis

To reconstruct the phylogenetic relationships of *comD*, we aligned full-length alleles using MUSCLE 3.8.425 (Edgar, 2004) and Geneious 7.1.5 (Kearse et al., 2012). After deleting sites with 5% or more gaps, we inferred the GTR+I+ γ substitution model using jModelTest 2.1.7 (Darriba et al., 2012). We used MrBayes 3.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) for phylogenetic reconstruction across four independent runs.

105 The short length of *comC* prevented reconstructing the phylogenetic history with confidence;
106 we instead used MUSCLE 3.8.425 (Edgar, 2004) and Geneious 7.1.5 (Kearse et al., 2012) to align
107 amino acid variants and produce a UPGMA phylogram of amino acid identity.
108 In order to characterize the overall genetic population structure in *S. pneumoniae*, we used
109 data from a previously reported full genome phylogenetic tree (Miller et al., 2016) based on alignment
110 to strain R6_uid57859 (Hoskins et al., 2001). In summary, we used genome sites found in at least
111 99.5% of all the above *S. pneumoniae* genomes (1,444,122 sites) and used the single best maximum
112 likelihood tree (ln(likelihood) of -3,2145,034.7) starting from fifteen unique random trees and from
113 fifteen unique parsimonious trees, as calculated by RAxML v8.2.4 (Stamatakis, 2006) and ExaML
114 v3.0 (Kozlov et al., 2015). This tree included 82 genomes from pneumococcus Complex 3 (Croucher
115 et al. 2013) and 240 PMEN-1 genomes (Croucher et al., 2011); we excluded these strains from all of
116 our analyses because they were originally sampled based on membership to clonal complexes and are
117 therefore biased with respect to population structure and pherotype. 43 *Streptococcus sp.* viridans
118 genomes were used as an outgroup for this tree, as previously reported (Miller et al., 2016).
119 For interspecific phylogenies, we identified *comC* and *comD* sequences from the available
120 NCBI genomes for *S. infantis* (5 genomes), *S. mitis* (45 genomes), *S. oralis* (18 genomes), and *S.*
121 *pseudopneumoniae* (40 genomes). The extensive sequence diversity in the mature CSP meant that we
122 lacked confidence in aligning *comC*. Therefore, we used BAli-Phy v.2.3.6 (Suchard and Redelings,
123 2006) for phylogenetic analysis because it estimates both the alignment and the tree simultaneously.
124 We used the HKY+ γ model of substitution and constrained the leader sequence of *comC* to align
125 together; BAli-Phy then estimated the full *comC* alignment and tree across eight independent runs
126 using the S07 model of indel mutations.
127 Using MUSCLE 3.8.425 (Edgar, 2004) and Geneious 7.1.5 (Kearse et al., 2012), we aligned
128 the full-length *comD* sequences and deleted sites with more than 5% gaps. For further analysis, we
129 used the filtered polymorphic sites from Gubbins (Croucher et al., 2015); any regions in individual
130 sequences with evidence of recombination from Gubbins were replaced with N's. We used a GTR + γ
131 model of nucleotide substitution as determined by jModelTest 2.1.7 (Darriba et al., 2012) to
132 reconstruct the phylogeny using MrBayes 3.2 across four independent runs.

133

134 **Shannon Diversity**

135 We located *comC* from 4 of 5 *S. infantis* genomes, 42 of 45 *S. mitis* genomes, all 18 *S. oralis*
136 genomes, 39 of 40 *S. pseudopneumoniae* genomes, and 4,076 of 4,089 *S. pneumoniae* genomes. *comD*
137 was also located in 4,040 *S. pneumoniae* genomes and from all other genomes except one *S. oralis*
138 genome. We calculated Shannon diversity within each species based on amino acid identity across the
139 entire gene. As each species had a different number of genomes, we calculated the Shannon diversity
140 based on sub-sampling of 18 genomes for *comC* and 17 genomes for *comD*, corresponding to the
141 minimum number of genomes in any species containing these genes, and we report the average of
142 1000 samples.

143

144 **Estimating population structure**

145 We used hierBAPS (Cheng et al., 2013) on the full genome (2,038,615 bp) alignment to strain
146 R6_uid57859 (Hoskins et al., 2001) in order to examine population structure. We used one run of 50,
147 60, 70, 80, 90, and 100 populations and four runs of 40 populations as upper bounds for the number of
148 populations with between 2 and 4 levels of division. We found a constant number of 29-31
149 populations at the first level of division, and we used the aggregate of all runs at this level to divide
150 genomes into populations for further analysis. We confirmed the inferred population structure using an
151 orthogonal metric; for this, we measured nucleotide diversity across the entire genome by dividing the
152 number of identical, non-gapped sites by the total number of non-gapped sites for pairwise
153 combinations of genomes using Python 2.7.8. We then compared average nucleotide diversity between
154 and within populations.

155

156 **Recombination events**

157 We calculated the expected frequency of within- and between-pherotype recombination by
158 assuming the null hypothesis that genetic exchange occurs randomly between pherotypes as a function
159 of their respective frequencies. Accordingly, if *i* represents the frequency of strains producing Csp-1,

160 the expected within Csp-1 recombination rate is i^2. It then follows that the expected fraction of all
161 recombination events that involve Csp-1 that occur strictly within-pherotype is

$$\frac{i^2}{i^2 + 2i(1-i)}$$

162 which reduces to:

$$\frac{i^2}{2i - i^2}$$

163 Similarly, the expected rate of between-pherotype recombination is *2ij*, where *i* and *j* represent the
164 frequencies of the focal pherotypes.

165 We used GeneConv 1.81a (Sawyer, 1999) to detect recombination events of at least 100 bp
166 between all pairwise comparisons of the 4,089 genomes; briefly, this program detects continuous
167 sections of DNA in pairwise alignments that have higher identity than the surrounding regions after
168 accounting for monomorphic sites. We analysed the genomes as 'circular', used Holm-Bonferroni
169 correction for multiple testing (Holm, 1979), and used a Gscale constant of 2, which scales to the
170 number of mismatches allowed while detecting recombination events.

171 We next calculated the fraction of observed recombination events relative to each pherotype.
172 By dividing the observed recombination frequencies by the expected recombination frequencies and
173 natural-log transforming this ratio, we derived an intuitive scaling: values greater than 0 indicate that
174 there is more recombination than expected by chance, while values less than 0 indicate the opposite.
175 We used the PropCIs package in R (R Core Team, 2013) to estimate 95% confidence intervals for this
176 ratio, and we tested if the observed recombination frequency differed significantly from the expected
177 recombination frequency using Pearson's χ^2 statistic using R (R Core Team, 2013). We additionally
178 calculated observed and expected recombination frequencies by classifying strains by populations in
179 place of pherotype.

180 To examine within and between pherotype recombination in more detail, we characterized the
181 unique recombination events occurring between strains carrying the two dominant pherotypes (Csp-1
182 and Csp-2), which together comprise 95.5% of strains. By highlighting unique recombination events,
183 the aim of this analysis was to remove the potentially biasing influence of vertical transmission, which

could cause more ancient recombination events to be overrepresented compared to newer events. In order to identify unique recombination events, we grouped events that shared an identical start and end position in the full genome alignment. Next, we calculated the average proportion of these unique recombination events taking place between strains with Csp-1 and Csp-2 for each group. We examined a range of cut-off points for the minimum proportion of strains that must be involved in each grouped recombination event, including: 0.05% of strains (≥2 strains, 1,663,312 events); 0.1% of strains (≥4 strains, 1,156,428 events); 0.5% of strains (≥20 strains, 381,084 events), 1.0% of strains (≥41 strains, 197,740 events), 2.5% of strains (≥102 strains, 59,980 events), and 5.0% of strains (≥204 strains, 19,428 events); this gradually focused the analysis on ancient recombination events. The expected proportion of between Csp-1 and Csp-2 recombination events was calculated as the expected encounter frequencies when only considering these two pherotypes, that is:

$$2 * \frac{i}{(i+j)} * \frac{j}{(i+j)}$$

where *i* and *j* are the frequencies of Csp-1 and Csp-2, respectively.

Results

comC and *comD* diversity across *S. pneumoniae* genomes and geographic sites

We estimated pherotype frequencies from 4,089 *S. pneumoniae* genomes that include extensive sampling in four geographic sites (Table 1). The global population is dominated by two pherotypes (the mature signals encoded by the *comC* gene): Csp-1 (72.0% of strains) and Csp-2 (23.5% of strains). We also identified a novel variant in 2.1% of strains designated Csp-1_Short, which is identical to Csp-1 but with the final 4 residues deleted. Three other related variants contain between 1 and 3 "NFF" amino acid repeats, which we designate Csp-4_R1, Csp-4_R2, and Csp-4_R3, respectively. Note that Csp-4_R2 has previously been labelled Csp-4 while Csp-4_R3 is also called Csp-3 (Whatmore et al., 1999). 0.27% of strains contained no *comC* sequence, which may result from incomplete genome sequencing; the low fraction of genomes without a *comC* sequence indicates strong selection against phenotypes that do not produce CSP. The overall frequency ratio of Csp-1::Csp-2 in previous studies was approximately 73.7%::26.3% after minority pherotypes were

removed (Carrolo et al 2009, Carrolo et al 2014, Cornejo et al 2010, Pozzi et al 1996, Vestrheim et al 2011); this is similar to the overall 75.4%::24.6% frequency ratio that we report after examining only Csp-1 and Csp-2 (Table 1). However, we found significant differences in Csp-1::Csp-2 frequency ratios between strains isolated from different geographic sites (Table 1), whose Csp-1 frequency ratios range from 77.6% to 65.9%. The Maela strains are significantly different from the Massachusetts and Rotterdam strains ($p = 2.14 \times 10^{-9}$ and $p = 1.46 \times 10^{-8}$ respectively, two-sample proportion test) but not significantly different from the Atlanta strains ($p = 0.527$, two-sample proportion test). The Atlanta strains are significantly different from the Massachusetts strains ($p = 0.0351$, two-sample proportion test), with both of these sets not significantly different from the Rotterdam strains ($p = 0.193$ and $p = 1.00$ respectively, two-sample proportion test).

The polymorphism in CSP is mirrored in the histidine kinase receptor for CSP, *comD*. We found excellent concordance between the pherotype each strain carries and its corresponding *comD* sequence (Figure 1). 935 of 936 strains with Csp-2 and full-length *comD* alleles contained *comD* alleles within a well-supported clade (posterior probability (PP) = 1.00) distinct from the other common (> 0.2% occurrence) pherotypes. Similarly, all *comD* alleles found with Csp-4 variants cluster in a well-supported (PP = 1.00) clade, with Csp-1 strains then forming a paraphyletic group with their *comD* alleles. The Csp-1 paraphyletic group also contains *comD* alleles associated with Csp-1_Short, which suggests this CSP may bind to the ComD receptor similarly to Csp-1. This supports tight functional linkage within three groups of pherotypes and ComD histidine kinases: Csp-1 (which includes Csp-1_Short), Csp-2, and the Csp-4 derivatives.

Interspecific gene trees of the signal gene *comC* (Supplementary Figure 1) lacked support for most interspecific clades. *comC* variants do not strictly cluster by species and Csp-4 derivatives form a clade with other non-pneumoniae *Streptococcus*, which could be indicative of: i) horizontal transfer, which is not uncommon among related *Streptococcus* species (Balsalobre et al., 2003; Duesberg et al., 2008; King et al., 2005); or ii) the possibility that trans-specific polymorphism is maintained in this locus as a balanced polymorphism (Gao et al., 2015; Ségurel et al., 2013). This second explanation is unlikely for *S. pneumoniae* given that 115 of 118 *S. pneumoniae* alleles form a well-supported clade that excludes other species in the interspecific *comD* phylogenetic tree of the histidine kinase receptor

238 gene, in which we attempted to remove intragenic horizontal recombination (Supplementary Figure 2).
239 By contrast, *S. mitis* and *S. pseudopneumoniae* freely intermix in all other clades, which is a pattern
240 that supports trans-specific polymorphism between these two species. Shannon diversity of amino acid
241 variants is lowest in *S. pneumoniae* compared to *S. mitis*, *S. oralis*, and *S. pseudopneumoniae* for
242 ComC (0.75 compared to 1.89-2.70), the mature CSP (0.67 compared to 1.17-2.63), and ComD (1.51
243 compared to 2.66-2.78) (Supplementary Table 2). While diversity in *S. pneumoniae* is a product of the
244 Csp-1:Csp-2 polymorphism, other viridians-group species appear to maintain higher diversity through
245 a larger repertoire of CSP signal molecules (Supplementary Figure 1).
246
247 **Pherotype and population structure**
248 We inferred the population structure of these strains using 10 independent hierBAPS runs with
249 a full-genome alignment (Cheng et al. 2013; Figure 2); 23 of the resultant populations were invariant
250 across all runs. The remaining genomes were assigned to one of 17 populations that each contained 21
251 genomes or more (0.5% of the total number of genomes) that co-occurred in all ten runs. 118 genomes
252 (2.7%) could not be consistently classified into populations. Overall, this resulted in 40 estimated
253 populations alongside the unclassified genomes. 25% of populations were not monophyletic on the
254 full-genome phylogenetic tree, which is surprising but consistent with previous analysis of
255 pneumococcal populations (Figure 2; Chewapreecha et al., 2014; Croucher et al., 2013b). Consistent
256 with the hierBAPS analysis, we estimated less genome nucleotide diversity within a population than
257 between populations for 39 out of 40 populations (mean of per-within population nucleotide diversity
258 = 0.00404; mean of per-between population nucleotide diversity = 0.0111; $p < 1.2 \times 10^{-57}$ except for
259 population 1, p = 0.059; corrected Wilcox test). Pherotypes are not equally distributed across
260 populations (Figure 3A), with 20 populations fixed for a single pherotype. Similarly, geography has an
261 unsurprising effect on population structure, as 16 of the 40 populations consist solely of strains
262 collected in Maela, Thailand (Supplementary Table 3).
263
264 **Recombination Events**

To examine within versus between pherotype recombination frequencies, we compared the observed fraction of recombination events to the expected fraction of recombination events that assumes strains recombine randomly. Overall, we found no evidence that recombination is limited between populations. While populations 9 and 21 had significantly more within-population recombination events than expected (Figure 3B), 37 other populations had significantly less within-population recombination events than expected ($p \leq 1.09 \times 10^{-6}$ except for population 1, $p = 0.395$; Newcombe proportion test with Holm-Bonferroni correction). The pherotypic diversity of populations, as estimated by Shannon diversity, created an interesting pattern with the proportion of observed between-population recombination events (Figure 3C), although a linear relationship is not statistically significant (Spearman's rank correlation $\rho = 0.254$, $p = 0.280$ after removing populations with Shannon diversity = 0).

Figure 4 shows estimates of recombination for all common (> 0.2% of strains) pherotype combinations, with all observed recombination frequencies differing significantly from those expected ($p < 4.9 \times 10^{-6}$; Newcombe proportion test with Holm-Bonferroni correction; Figure 4A). Of the 30 between-pherotype comparisons, 15 estimates of recombination frequencies were significantly higher than expected, and 15 estimates were observed significant less frequently than expected. Importantly, we found a higher observed recombination frequency than expected between strains expressing the two dominant pherotypes Csp-1 and Csp-2 ($p < 10^{-99}$; Newcombe proportion test with Holm-Bonferroni correction). In total, recombination between these two pherotypes comprises 34.7% of all recorded recombination events; thus an excess of recombination for these pairs implies that the general role of pherotypes is to marginally increase between-pherotype recombination rates. Five of the six within-pherotype observed recombination frequencies are lower than expected, with Csp-4_R2 as the only exception (Figure 4A); three of these frequencies are the lowest of all pherotype combinations within their respective pherotype.

A potential caveat of these results is that we consider every pairwise recombination event as independent. However, recombination events could have occurred at any time during the ancestry of these strains and then persisted via vertical descent, which may lead to biased estimates of recombination frequencies if a single, historic recombination event is counted in each of the multiple

293 strains that descended from their common ancestor. As the true number of recombination events
294 through history is unknown, we considered the opposite extreme, in which pairwise recombination
295 events with identical start and end positions in the full-genome alignment originated from a single
296 recombination event. We focused only on Csp-1 and Csp-2 pherotypes, and we considered unique
297 recombination events found in a range of at least 0.05% of genomes (≥2 genomes) to at least 5.0% of
298 genomes (≥204 genomes) in order to focus on more ancient recombination events (Figure 4B). The
299 mean proportion of between Csp-1::Csp-2 events of five out of six distributions are significantly
300 different than a null expectation based on pherotype frequencies in the global population ($p < 0.0253$;
301 Wilcox test for difference to $\mu = 0.371$); however, no values differ by more than 4.0% from the null
302 expectation of 0.371, with three averages significantly above and two averages significantly below the
303 null value. This result indicates that any potential bias introduced by considering all recombination
304 events independently is, at most, marginal.
305
306 **Discussion**
307 Several studies, including this one, have found that the two dominant pherotypes of the
308 quorum-dependent regulator of competence in *S. pneumoniae* (i.e. Csp-1 and Csp-2) are maintained at
309 relative frequencies of roughly 70:30 (Carrolo et al 2014; Vestrheim et al 2011; Cornejo et al 2010;
310 Carrolo et al 2009; Pozzi et al 1996). Our results also indicate that there is subtle geographic variation
311 in the ratio of these dominant pherotypes, ranging from 66:34 to 78:22 (Table 1). These results
312 naturally lead to questions of how this pherotype ratio is maintained (to the near exclusion of other
313 pherotypes), especially at similar, yet not identical, levels in disparate geographic sites. The null
314 explanation is that pherotypes are neutral with respect to bacterial fitness, although three lines of
315 evidence counter this explanation. First, other studies suggest that *comC* is evolving under positive
316 selection (Cornejo et al 2010; Carrolo et al 2009). Second, other *Streptococcus* viridans species have a
317 higher diversity in CSP, ComC, and ComD; this diversity is caused by a large number of pherotypes
318 within each species as opposed to the two dominant pherotypes in *S. pneumoniae*. Third, each
319 geographic site is comprised of strains from anywhere between 14 and 36 different populations

(Supplemental Table 3), yet each site maintains the approximate 70:30 ratio of major pherotypes. This inter-population pattern between geographic sites is unlikely to occur through neutral mechanisms.

Irrespective of the factors maintaining pherotype polymorphism, what are the consequences of their maintenance for pneumococcal populations? Two divergent hypotheses have been explored to answer this question. One holds that these variants reinforce population subdivision by restricting recombination to strains that carry the same pherotype (Carrolo et al 2009). The alternative hypothesis suggests that lysis of non-identical pherotypes by the process of fratricide leads to inter-pherotype transformation and therefore the elimination of pherotype-specific population structuring (Cornejo et al 2010). Our data, based on the analysis of recombination from more than 4,000 full genome sequences, are not clearly consistent with the extreme version of either hypothesis. We find a slight tendency towards increased between-pherotype recombination for the dominant pherotype classes (Csp-1 and Csp-2; Figure 4A), and these results clarify that pherotypes are not a barrier to recombination in this species. However, these data are not without limits, in particular the difficulty of inferring recombination among closely related strains, which could partly explain the lower frequencies of within-pherotype (Figure 4A) or within-population (Figure 3B) recombination. Because this should only have minimal influence on estimates of between-pherotype recombination frequencies, as strains carrying different pherotypes tend to come from different populations (Figure 3B), we do not view this is a bias that is likely to influence our conclusions.

S. pneumoniae live in surface-associated biofilm communities, where in the case of single-strain colonization, they will be surrounded by clone-mates. If some of these cells become competent and lyse the others via competence-induced fratricide (i.e. induced production of bacteriocins that target uninduced cells), there will be little signature of this event at the genomic level. However, it is increasingly clear that multiple-strain infections are more common than previously thought; co-colonization within the nasopharynx is observed in up to 50% of individuals (Wyllie et al 2014; Rodrigues et al 2013), and because rare variants are less likely to be detected during sampling, this may be an underestimate of the true occurrence of co-colonisation. In these cases, there is opportunity for genetic exchange between different genotypes. If the co-infecting genotypes express the same pherotype, both will respond to the same peptide signal and both genotypes could, in principle, release

348 DNA that will be available to the other. However, if the genotypes express different pherotypes, then
349 the genotype that first initiates competence may lyse the other non-responding genotype via
350 competence-induced fratricide, leading to more unidirectional uptake. While recombination during
351 single pherotype infections could reinforce any pre-existing association between pherotype and
352 population structure, recombination during mixed-pherotype infections would cause this association to
353 decline or disappear. Our results are most consistent with this latter scenario.

354 How often do mixed pherotype colonization events occur? If the likelihood of colonization is
355 random with respect to pherotype, then the probability that both pherotypes will be found to co-occur
356 is simply twice the product of these relative frequencies. In two recent studies where this has been
357 measured, mixed pherotype infections are common in cases where multiple strains are seen (47.5%
358 and 57.1% in strains from Portugal and Norway, respectively) and do not vary from the null
359 expectation of random colonization (Valente et al 2012, Vestrheim et al 2011). These results have
360 several important implications. First, they suggest that fratricide during the induction to competence
361 has minimal, if any, impact on the within-host competitive dynamics of co-occurring strains. Second,
362 they suggest that opportunities for recombination between strains expressing different pherotypes are
363 widespread. Accordingly, we would predict pherotypes to have little direct influence on population
364 structure, a prediction borne out by the results presented here and elsewhere (Figure 3A; Cornejo et al
365 2010). This, of course, does not preclude structure arising from other influences, e.g. serotype specific
366 immunity, differences in antibiotic resistance or other attributes leading to biases in colonization.
367 However, pherotypes neither appear to underlie this structure nor to eliminate it due to ubiquitous
368 inter-pherotype recombination (Figure 4A).

369 A final explanation for a fixed pherotype ratio is that diversity is maintained for reasons
370 wholly distinct from their effects on competence. In addition to competence, CSP induces more than
371 150 pneumococcal genes, and only a small fraction of these are required for DNA uptake and
372 recombination (Peterson et al 2004; Peterson et al 2000). Different concentrations of CSP are required
373 for competence activation between Csp-1 and Csp-2 with their respective receptors (Carrolo et al.,
374 2014; Iannelli et al., 2005), which determine the population density at which CSP induces these genes
375 and could drive large phenotypic differences between pherotype based only on *comC* and *comD*

variation. We were unsuccessful in finding genomic loci that co-associates with CSP via selection using a genome-wide association study with pherotype as a phenotype, which suggests that any ecological differentiation between phenotype may be caused only by differences in *comC* and *comD*. However, Carrolo et al. (2014) find striking differences in the ability for strains expressing Csp-1 and Csp-2 to form biofilms, with Csp-1 strains producing biofilms with greater biomass. Although the authors suggest that this difference may lead to differences in the colonization success and transmissibility of isolates, epidemiological data discussed above are not consistent with this possibility; co-colonization with multiple pherotypes occurs as often as expected by chance (Valente et al 2012, Vestrheim et al 2011). An alternative, which could lead to a form of balancing or frequency-dependent selection maintaining pherotypes at intermediate frequencies, is the possibility that differences in biofilm-associated biomass that influence stable colonization may trade-off with reductions in transmissibility. By this scenario, Csp-1-expressing strains may form more robust biofilms during colonization, while strains expressing Csp-2 are more proficient at dispersal. Although testing these possibilities is beyond the scope of this work, these ideas can potentially be examined empirically in both *in vivo* or *in vitro* models. Furthermore, they make the prediction that carriage duration should vary as a function of pherotype, a possibility that could potentially be retrospectively examined from epidemiological studies.

Although our results are unable to clarify the evolutionary factors that lead to the origin and maintenance of pherotypes in *S. pneumoniae*, our comprehensive analysis demonstrates the long-term effects of this polymorphism on recombination in this species. In summary, while pherotypes can apparently facilitate recombination between the major pherotype classes, this effect is weak and has no evident impact on population structure of this pathogen. Explanations for pneumococcal population structure therefore lie outside of pherotypes, and indeed, explanations for pherotypes may lie outside of their effects on recombination.

Acknowledgements

Bioinformatic work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation. This work made use of the facilities of N8 HPC Centre of Excellence, provided and

404 funded by the N8 consortium and the Engineering and Physical Sciences Research Council (grant
405 number EP/K000225/1). The Centre is co-ordinated by the Universities of Leeds and Manchester. This
406 work was supported by the Biotechnology and Biological Sciences Research Council (grant number
407 BB/J006009/1) to DER and ISR and by the Wellcome Trust (105610/Z/14/Z) to the University of
408 Manchester.

409

References

Balsalobre, L., Ferrandiz, M.J., Linares, J., Tubau, F., de la Campa, A.G., 2003. Viridans group streptococci are donors in horizontal transfer of topoisomerase IV genes to *Streptococcus pneumoniae*. Antimicrob. Agents Chemother. 47, 2072–2081. doi:10.1128/aac.47.7.2072-2081.2003

Bogaert, D., Engelen, M.N., Timmers-Reker, A.J.M., Elzenaar, K.P., Peerbooms, P.G.H., Coutinho, R.A., De Groot, R., Hermans, P.W.M., 2001. Pneumococcal carriage in children in the Netherlands: A molecular epidemiological study. J. Clin. Microbiol. 39, 3316–3320. doi:10.1128/JCM.39.9.3316-3320.2001

Carrolo, M., Pinto, F.R., Melo-Cristino, J., Ramirez, M., 2014. Pherotype influences biofilm growth and recombination in *Streptococcus pneumoniae*. PLoS One 9, e92138. doi:10.1371/journal.pone.0092138

Carrolo, M., Pinto, F.R., Melo-Cristino, J., Ramirez, M., 2009. Pherotypes are driving genetic differentiation within *Streptococcus pneumoniae*. BMC Microbiol. 9, 191. doi:10.1186/1471-2180-9-191

Chancey, S.T., Agrawal, S., Schroeder, M.R., Farley, M.M., Tettelin, H.H., Stephens, D.S., 2015. Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. Front. Microbiol. 6, 1–14. doi:10.3389/fmicb.2015.00026

Cheng, L., Connor, T.R., Sirén, J., Aanensen, D.M., Corander, J., 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol. Biol. Evol. 30, 1224–1228. doi:10.1093/molbev/mst028

Chewapreecha, C., Harris, S.R., Croucher, N.J., Turner, C., Marttinen, P., Cheng, L., Pessia, A., Aanensen, D.M., Mather, A.E., Page, A.J., Salter, S.J., Harris, D., Nosten, F., Goldblatt, D., Corander, J., Parkhill, J., Turner, P., Bentley, S.D., 2014. Dense genomic sampling identifies highways of pneumococcal recombination. Nat. Genet. 46, 305–9. doi:10.1038/ng.2895

Claverys, J.-P., Håvarstein, L.S., 2007. Cannibalism and fratricide: mechanisms and raisons d'être. Nat. Rev. Microbiol. 5, 219–229. doi:10.1038/nrmicro1613

Claverys, J.-P., Prudhomme, M., Martin, B., 2006. Induction of competence regulons as a general response to stress in Gram-positive bacteria. Annu. Rev. Microbiol. 60, 451–475. doi:10.1146/annurev.micro.60.080805.142139

Cornejo, O.E., McGee, L., Rozen, D.E., 2010. Polymorphic competence peptides do not restrict recombination in *Streptococcus pneumoniae*. Mol. Biol. Evol. 27, 694–702. doi:10.1093/molbev/msp287

Croucher, N.J., Chewapreecha, C., Hanage, W.P., Harris, S.R., McGee, L., van der Linden, M., Song, J.-H., Ko, K.S., de Lencastre, H., Turner, C., Yang, F., Sá-Leão, R., Beall, B., Klugman, K.P., Parkhill, J., Turner, P., Bentley, S.D., 2014a. Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. Genome Biol. Evol. 6, 1589–602. doi:10.1093/gbe/evu120

Croucher, N.J., Finkelstein, J. a, Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P., Lipsitch, M., 2013a. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat. Genet. 45, 656–63. doi:10.1038/ng.2625

Croucher, N.J., Finkelstein, J.A., Pelton, S.I., Mitchell, P.K., Lee, G.M., Parkhill, J., Bentley, S.D., Hanage, W.P., Lipsitch, M., 2013b. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat. Genet. 45, 656–63. doi:10.1038/ng.2625

Croucher, N.J., Hanage, W.P., Harris, S.R., McGee, L., van der Linden, M., de Lencastre, H., Sá-Leão, R., Song, J.-H., Ko, K.S., Beall, B., Others, 2014b. Variable recombination dynamics during the emergence, transmission and'disarming'of a multidrug-resistant pneumococcal clone. BMC Biol. 12, 49. doi:10.1186/1741-7007-12-49

Croucher, N.J., Harris, S.R., Fraser, C., Quail, M. a, Burton, J., van der Linden, M., McGee, L., von Gottberg, A., Song, J.H., Ko, K.S., Pichon, B., Baker, S., Parry, C.M., Lambertsen, L.M., Shahinas, D., Pillai, D.R., Mitchell, T.J., Dougan, G., Tomasz, A., Klugman, K.P., Parkhill, J., Hanage, W.P., Bentley, S.D., 2011. Rapid pneumococcal evolution in response to clinical interventions. Science 331, 430–434. doi:10.1126/science.1198545

Croucher, N.J., Mitchell, A.M., Gould, K. a., Inverarity, D., Barquist, L., Feltwell, T., Fookes, M.C., Harris, S.R., Dordel, J., Salter, S.J., Browall, S., Zemlickova, H., Parkhill, J., Normark, S.,

Henriques-Normark, B., Hinds, J., Mitchell, T.J., Bentley, S.D., 2013c. Dominant role of nucleotide substitution in the diversification of serotype 3 pneumococci over decades and during a single infection. PLoS Genet. 9. doi:10.1371/journal.pgen.1003868

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., Harris, S.R., 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 43, e15–e15. doi:10.1093/nar/gku1196

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9, 772–772. doi:10.1038/nmeth.2109

Duesberg, C.B., Malhotra-Kumar, S., Goossens, H., McGee, L., Klugman, K.P., Welte, T., Pletz, M.W.R., 2008. Interspecies recombination occurs frequently in quinolone resistance-determining regions of clinical isolates of Streptococcus pyogenes. Antimicrob. Agents Chemother. 52, 4191–3. doi:10.1128/AAC.00518-08

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340

Evans, B.A., Rozen, D.E., 2013. Significant variation in transformation frequency in *Streptococcus pneumoniae*. ISME J. 7, 791–9. doi:10.1038/ismej.2012.170

Gao, Z., Przeworski, M., Sella, G., 2015. Footprints of ancient-balanced polymorphisms in genetic variation data from closely related species. Evolution 69, 431–46. doi:10.1111/evo.12567

Håvarstein, L.S., Hakenbeck, R., Gaustad, P., 1997. Natural competence in the genus *Streptococcus*: Evidence that streptococci can change pherotype by interspecies recombinational exchanges. J. Bacteriol. 179, 6589–6594.

Håvarstein, L.S., Martin, B., Johnsborg, O., Granadel, C., Claverys, J.P., 2006. New insights into the pneumococcal fratricide: Relationship to clumping and identification of a novel immunity factor. Mol. Microbiol. 59, 1297–1307. doi:10.1111/j.1365-2958.2005.05021.x

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70. doi:10.2307/4615733

Hoskins, J., Alborn, W.E., Arnold, J., Blaszczak, L.C., Burgett, S., Estrem, S.T., Fritz, L., Fu, D., Fuller, W., Geringer, C., Gilmour, R., Glass, J.S., Khoja, H., Kraft, A.R., Lagace, R.E., Lee, L.N., Lefkowitz, E.J., Lu, J.I.N., Matsushima, P., Mundy, C.W., Nicas, T.I., Norris, F.H., Peery, R.B., Robertson, G.T., Rockey, P., Sun, P., Winkler, M.E., Yang, Y., Young-bellido, M., Zhao, G., Zook, C. a, Baltz, R.H., Jaskunas, R., Rosteck, P.R., Skatrud, P.L., Glass, J.I., States, U., 2001. Genome of the bacterium *Streptococcus pneumoniae* strain R6. J. Bacteriol. 183, 5709–5717. doi:10.1128/JB.183.19.5709

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755. doi:10.1093/bioinformatics/17.8.754

Iannelli, F., Oggioni, M.R., Pozzi, G., 2005. Sensor domain of histidine kinase ComD confers competence pherotype specificity in *Streptoccoccus pneumoniae*. FEMS Microbiol. Lett. 252, 321–326. doi:10.1016/j.femsle.2005.09.008

Johnsborg, O., Eldholm, V., Bjørnstad, M.L., Håvarstein, L.S., 2008. A predatory mechanism dramatically increases the efficiency of lateral gene transfer in *Streptococcus pneumoniae* and related commensal species. Mol. Microbiol. 69, 245–253. doi:10.1111/j.1365-2958.2008.06288.x

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. doi:10.1093/bioinformatics/bts199

King, S.J., Whatmore, A.M., Dowson, C.G., 2005. NanA, a neuraminidase from *Streptococcus pneumoniae*, shows high levels of sequence diversity, at least in part through recombination with *Streptococcus oralis*. J. Bacteriol. 187, 5376–5386. doi:10.1128/JB.187.15.5376

Kozlov, A.M., Aberer, A.J., Stamatakis, A., 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics 31, 2577–2579. doi:10.1093/bioinformatics/btv184

McGee, L., McDougal, L., Zhou, J., Spratt, B.G., Tenover, F.C., George, R., Hakenbeck, R., Hryniewicz, W., Lefévre, J.C., Tomasz, A., Klugman, K.P., 2001. Nomenclature of major antimicrobial-resistant clones of *Streptococcus pneumoniae* defined by the pneumococcal molecular epidemiology network. J. Clin. Microbiol. 39, 2565–2571.

520 doi:10.1128/JCM.39.7.2565-2571.2001
521 Miller, E.L., Abrudan, M.I., Roberts, I.S., Rozen, D.E., 2016. Diverse ecological strategies are
522 encoded by *Streptococcus pneumoniae* bacteriocin-like peptides. Genome Biol. Evol. 8, 1072–
523 1090. doi:10.1093/gbe/evw055
524 Mostowy, R., Croucher, N.J., Hanage, W.P., Harris, S.R., Bentley, S., Fraser, C., 2014. Heterogeneity
525 in the frequency and characteristics of homologous recombination in pneumococcal evolution.
526 PLoS Genet. 10, 1–15. doi:10.1371/journal.pgen.1004300
527 O'Brien, K.L., Wolfson, L.J., Watt, J.P., Henkle, E., Deloria-Knoll, M., McCall, N., Lee, E.,
528 Mulholland, K., Levine, O.S., Cherian, T., 2009. Burden of disease caused by *Streptococcus*
529 *pneumoniae* in children younger than 5 years: global estimates. Lancet 374, 893–902. doi:S0140-
530 6736(09)61204-6 [pii]\r10.1016/S0140-6736(09)61204-6 [doi]
531 Pestova, E. V, Håvarstein, L.S., Morrison, D. a, 1996. Regulation of competence for genetic
532 transformation in *Streptococcus pneumoniae* by an auto-induced peptide pheromone and a two-
533 component regulatory system. Mol. Microbiol. 21, 853–862. doi:10.1046/j.1365-
534 2958.1996.501417.x
535 Pozzi, G., Masala, L., Iannelli, F., Manganelli, R., Håvarstein, L.S., Piccoli, L., Simon, D., Morrison,
536 D.A., 1996. Competence for genetic transformation in encapsulated strains of *Streptococcus*
537 *pneumoniae*: two allelic variants of the peptide pheromone. J. Bacteriol. 178, 6087–6090.
538 R Core Team, 2013. R: A Language and Environment for Statistical Computing.
539 Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed
540 models. Bioinformatics 19, 1572–1574. doi:10.1093/bioinformatics/btg180
541 Sawyer, S. A. 1999. GENECONV: A computer package for the statistical detection of gene
542 conversion. Distributed by the author, Department of Mathematics, Washington University in St.
543 Louis, available at http://www.math.wustl.edu/~sawyer.
544 Ségurel, L., Thompson, E.E., Flutre, T., Lovstad, J., Venkat, A., Susan, W., Moyse, J., Ross, S.,
545 Gamble, K., Sella, G., 2013. Correction for Segurel et al., The ABO blood group is a trans-
546 species polymorphism in primates. Proc. Natl. Acad. Sci. 110, 6607–6607.
547 doi:10.1073/pnas.1304029110
548 Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
549 thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.
550 doi:10.1093/bioinformatics/btl446
551 Suchard, M.A., Redelings, B.D., 2006. BAli-Phy: simultaneous Bayesian inference of alignment and
552 phylogeny. Bioinformatics 22, 2047–2048. doi:10.1093/bioinformatics/btl175
553 Tortosa, P., Dubnau, D., 1999. Competence for transformation: A matter of taste. Curr. Opin.
554 Microbiol. 2, 588–592. doi:10.1016/S1369-5274(99)00026-0
555 Whatmore, A.M., Barcus, V. a, Christopher, G., Dowson, C.G., 1999. Genetic diversity of the
556 Streptococcal competence (*com*) gene locus. J. Bacteriol. 181, 3144–3154.
557

558 **Table 1.** Frequencies of pherotypes within geographic sites.

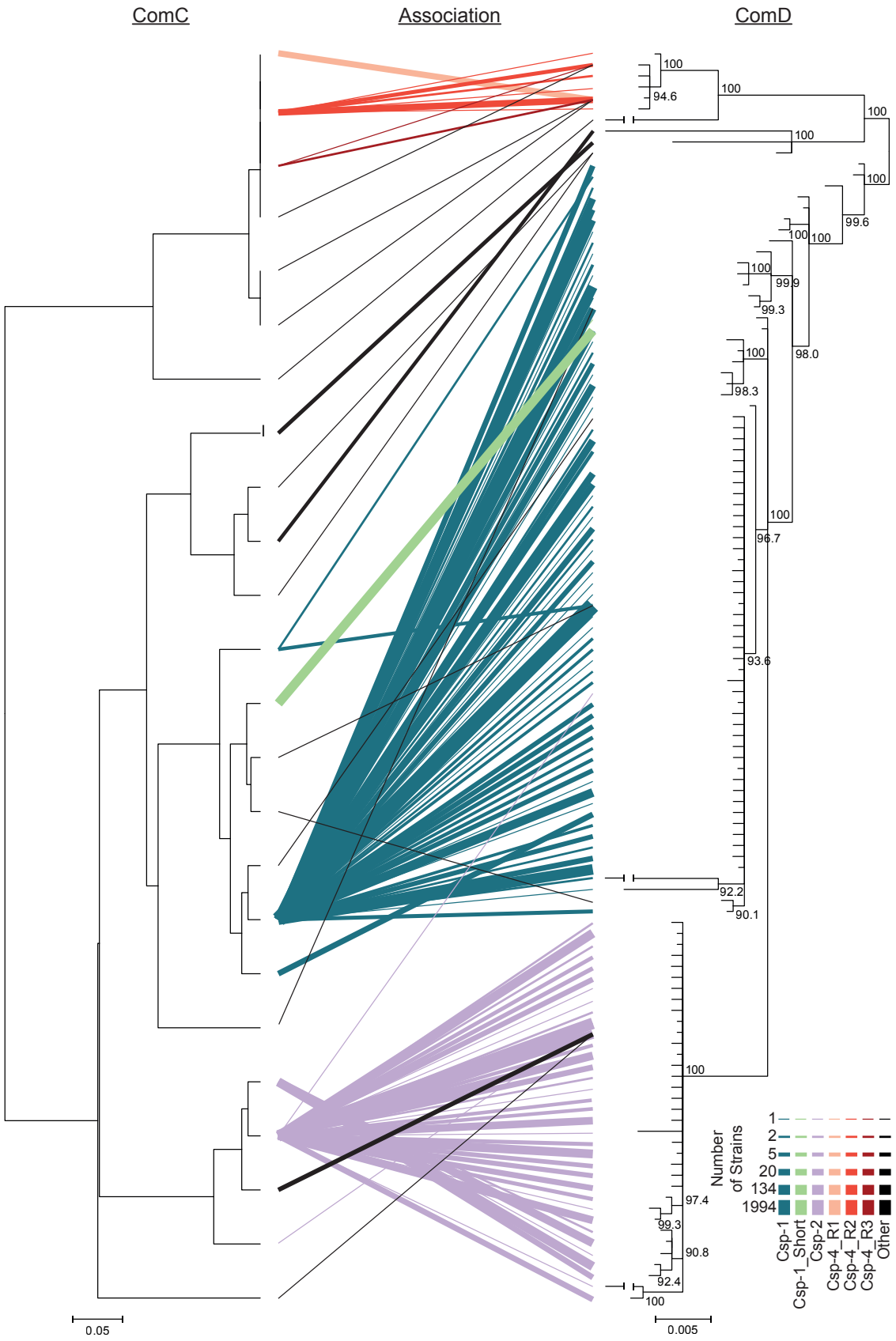
Pherotype	Secreted Peptide	Maela, Thailand	Atlanta, USA	Mass., USA	Rotterdam, The Netherlands	All 4,089 Genomes
Csp-1	EMRLSKFFRDFILQRKK	0.736	0.702	0.643	0.634	0.720
Csp-1_Short	EMRLSKFFRDFIL	0.029	0	0	0	0.021
Csp-2	EMRISRIILDFLFLRKK	0.213	0.24	0.333	0.324	0.235
Csp-4_1R	EMRKMNEKSFNIFNFF------RRR	0.005	0	0.003	0	0.004
Csp-4_2R	EMRKMNEKSFNIFNFFNFF---RRR	0.004	0.008	0.018	0.035	0.007
Csp-4_3R	EMRKMNEKSFNIFNFFNFFNFFRRR	0.005	0	0.002	0	0.004
Other[a]	–	0.007	0.017	0.002	0.007	0.007
None	–	0.002	0.033	0	0	0.003

[a]Other pherotypes each found in less than 0.2% of genomes, as well as two genomes of both Csp-1 and Csp-2 pherotype.

559

560

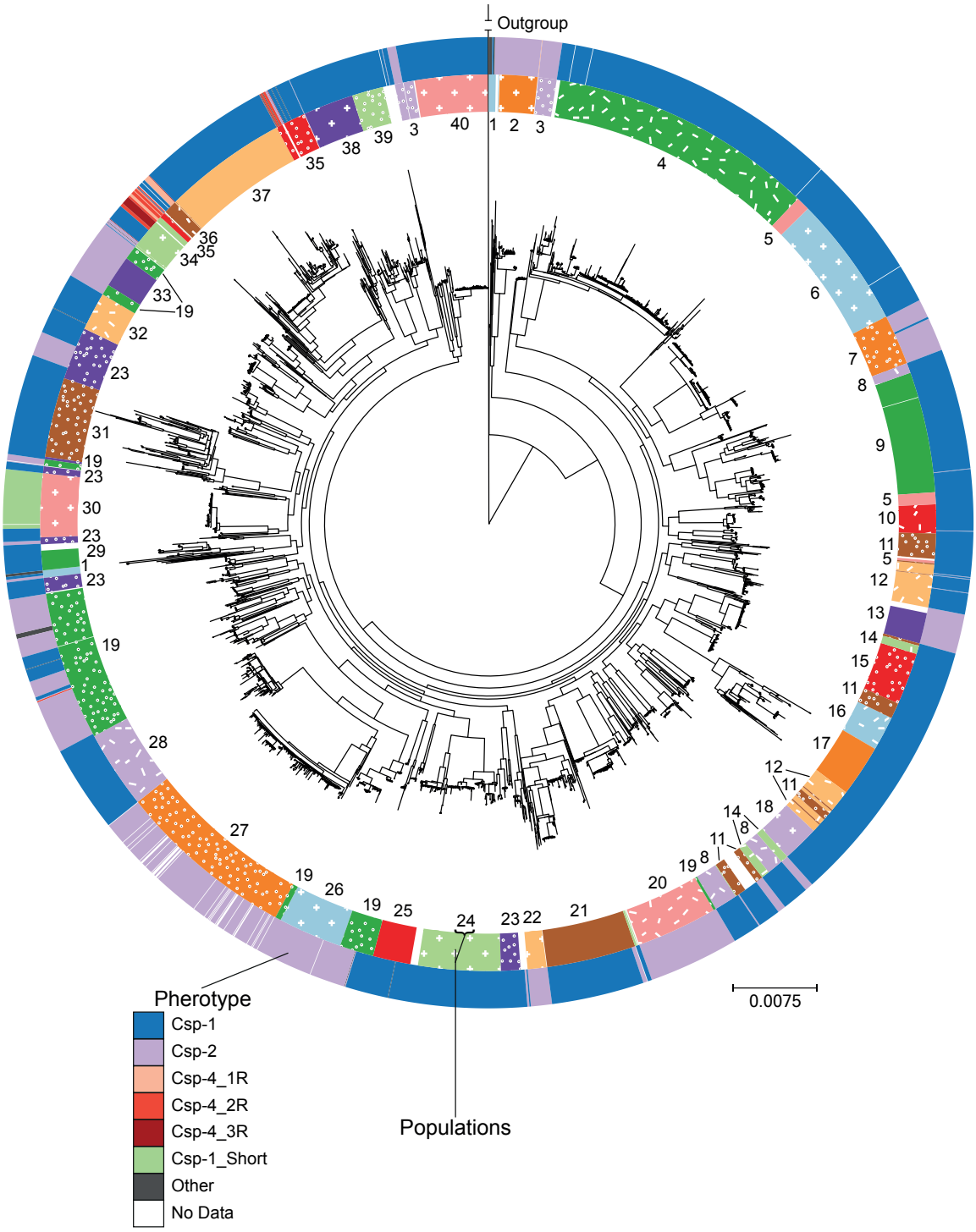
561 **Figure 1.** Associations between ComC and *comD*.



562

563

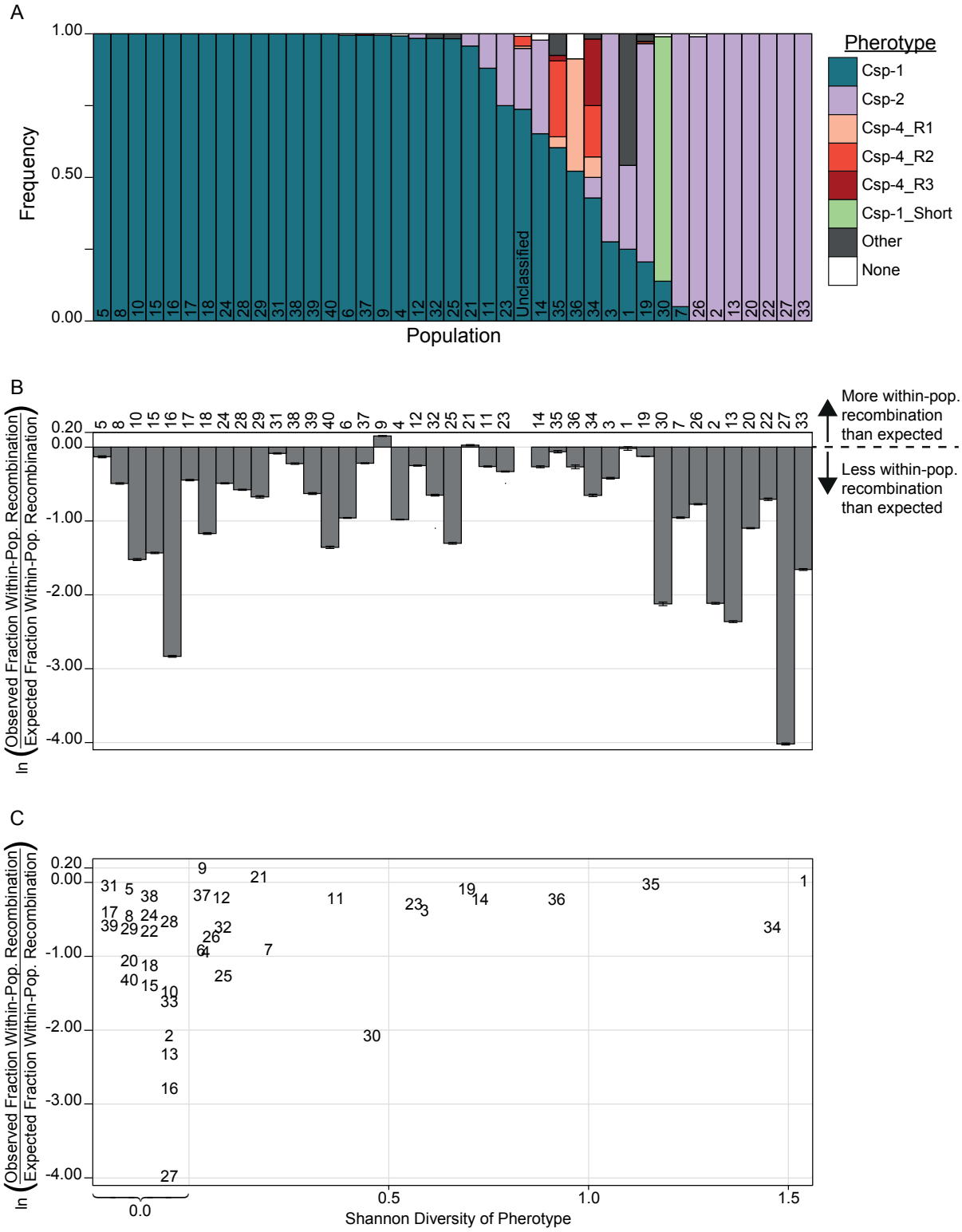
564 **Figure 2.** Phylogenetic relationship between *S. pneumoniae* genomes.



565

566

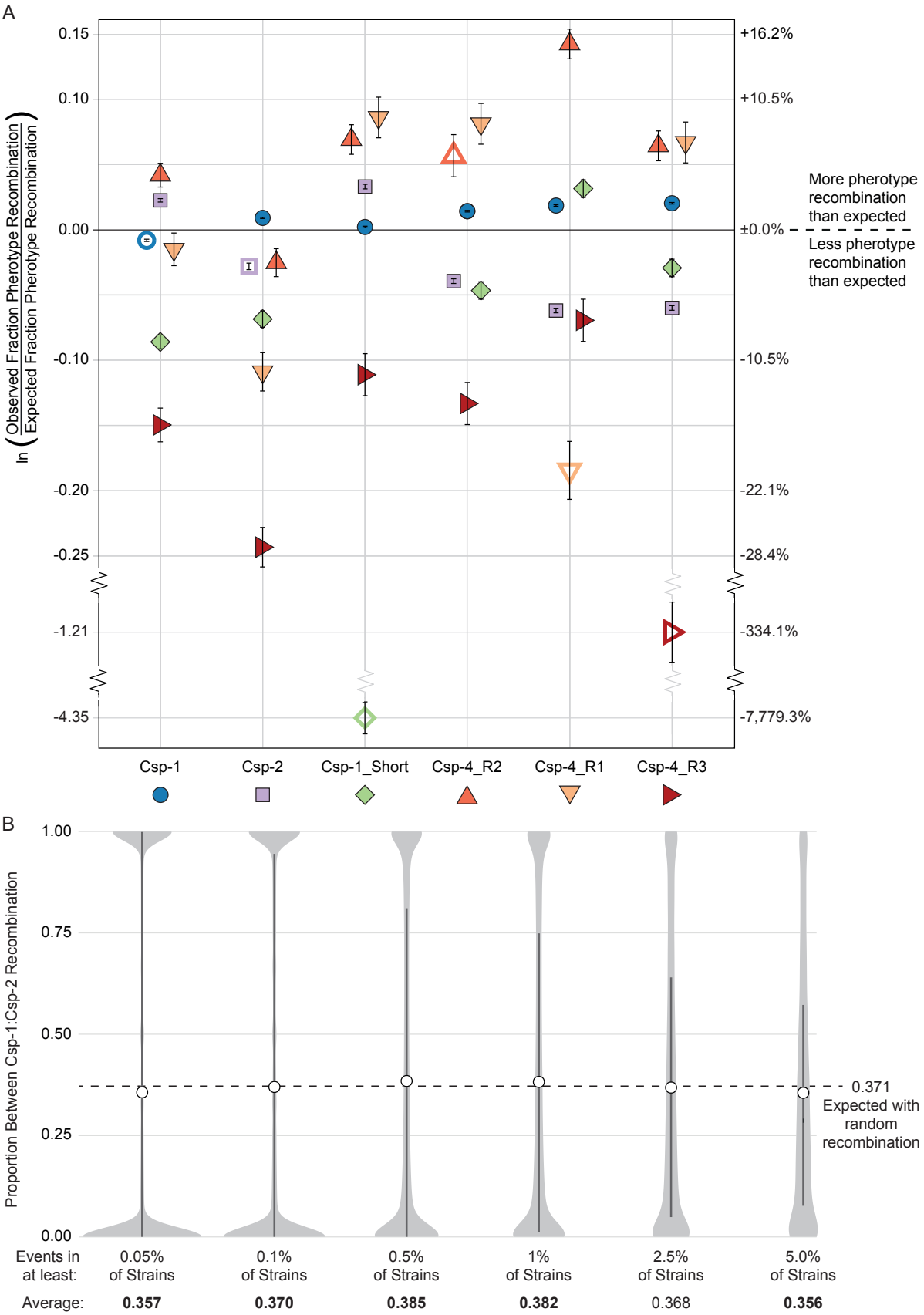
567 **Figure 3.** Relationship between pherotypes and populations.



568

569

570 **Figure 4.** Recombination between and within pherotypes.



571
572

573 **Figure Legends**

574 **Fig. 1.** Associations between ComC and *comD*. The UPGMA clustering of ComC genes is shown next
575 to the inferred phylogenetic relationship between *comD* alleles. ComC amino acid variants and *comD*
576 nucleotide alleles found within the same genome are show as lines connecting the two phylograms,
577 with thicker lines showing associations found in more strains. Line thickness is on a log scale.
578 Classification of *comD* alleles is based on co-occurring CSPs within genomes, with a 99.8%
579 correlation for the Csp-1 *comD* group, 99.2% correlation for the Csp-2 *comD* group, and 95.3%
580 correlation for the Csp-4 *comD* group.

581

582 **Fig. 2.** Phylogenetic relationship between *S. pneumoniae* genomes. The inner coloured ring shows the
583 population grouping of strains as determined by hierBAPS, shown as a number and a colour/pattern.
584 Genomes not classified into a population are white in the inner ring. The outer ring denotes pherotype.

585

586 **Fig. 3**. Relationship between pherotypes and populations. A) Distribution of pherotypes within each
587 population. B) Observed / expected fraction of within-population recombination for each population,
588 with values below zero indicating less observed within-population recombination than expected. Error
589 bars show 95% confidence intervals. C) The same recombination ratio as part B with populations'
590 Shannon diversity of pherotypes, with populations shown as numbers.

591

592 **Fig. 4.** Recombination between and within pherotypes. A) Observed / expected fraction of
593 recombination within and between pherotypes. Empty shapes are within-pherotype comparisons,
594 while filled shapes show between-pherotype comparisons. Colours correspond to pherotypes as in
595 Figure 1. Error bars show 95% confidence intervals. B) Distribution of the proportion of
596 recombination events between Csp-1 and Csp-2, in which all recombination events with identical
597 breaks points in the full-genome alignment are grouped as a single event. Light grey shapes show
598 density estimate; dark bars incorporate the 25th to 75th percentile; white circles indicate the average
599 proportion. Averages in bold are significantly different from the null expectation.