

Supplementary Online Material: Polymorphism in the  
competence peptide and the limits to recombination in  
*Streptococcus pneumoniae*

Omar E. Cornejo\*, Lesley McGee, and Daniel E. Rozen

November 16, 2009

**Contents**

<b>1</b>	<b>Strain types of clinical isolates</b>	<b>2</b>
<b>2</b>	<b>Genetic diversity per geographic region</b>	<b>4</b>
<b>3</b>	<b>Evidence of admixture between CSP subpopulations</b>	<b>5</b>

## **1 Strain types of clinical isolates**

The sequences of the 6 loci employed in this work (*aroE*, *gki*, *gdh*, *recP*, *spi*, *xpt*) can be fetched from the [www.mlst.net](http://www.mlst.net) website by searching the corresponding strain type (ST) on the database. The sequences of each allele corresponding to the haplotype could be downloaded in fasta format. The alleles profile of strain B224(\*) has been already submitted to the [mlst.net](http://www.mlst.net), but no ST type has been assigned yet. Meanwhile, the corresponding ST type is available upon request to the authors.

Table S.1: List of strains and strain type (ST type) used in this work. The sequence for each locus can be retrieved from the [www.mlst.net](http://www.mlst.net) by looking at the corresponding ST type.

Isolate ID	CSP type	ST type	Isolate ID	CSP type	ST type
B040	2	315	B044	2	180
B049	2	247	B060	2	247
B283	2	460	B345	2	2283
B380	2	2284	B381	2	2285
B401	2	2287	B402	2	2288
B440	2	2291	B449	2	664
B485	2	2294	B491	2	3365
B496	2	2295	B694	2	180
B701	2	246	B707	2	260
B723	2	315	B761	2	1220
B789	2	247	B826	2	2304
B843	2	180	B026	1	643
B052	1	1624	B059	1	218
B069	1	228	B081	1	90
B082	1	615	B083	1	180
B100	1	306	B155	1	146
B156	1	458	B170	1	180
B198	1	557	B201	1	2278
B208	1	217	B218	1	217
B223	1	289	B224	1	NEW*
B243	1	2021	B299	1	180
B308	1	146	B311	1	113
B325	1	191	B344	1	2282
B364	1	218	B388	1	2286
B419	1	2289	B426	1	2290
B442	1	2421	B463	1	611
B470	1	2292	B473	1	2293
B486	1	458	B508	1	146
B551	1	180	B628	1	2275
B638	1	304	B641	1	289
B643	1	53	B679	1	228
B682	1	218	B683	1	2299
B686	1	58	B692	1	289
B695	1	304	B703	1	2300
B704	1	110	B713	1	3594
B714	1	156	B716	1	2301
B720	1	90	B725	1	2302
B755	1	306	B763	1	306
B766	1	306	B776 <sub>3</sub>	1	1073
B808	1	191	B810	1	615
B811	1	289	B816	1	2303
B817	1	615	B823	1	289
B839	1	224	B841	1	306
B845	1	53	B850	1	218

## 2 Genetic diversity per geographic region

The following table shows the average genetic diversity across MLST categorized by geographical region.

Table S.2: Genetic diversity per gene and overall of the clinical isolates, categorized by geographic origin

	<b>n<sup>a</sup></b>	<b>S<sup>b</sup></b>	<b>Hd<sup>c</sup> (s.d)<sup>d</sup></b>	<b><math>\pi</math><sup>e</sup> (s.d)<sup>e</sup></b>
By Region				
<b>South America</b>	12	64	0.95(0.07)	0.008(0.001)
<b>North America</b>	11	54	0.97(0.04)	0.008(0.001)
<b>Europe</b>	23	79	0.84(0.02)	0.009(0.001)
<b>South Africa</b>	27	118	0.99(0.01)	0.011 (0.001)
<b>New Zealand</b>	9	56	0.92(0.09)	0.007(0.001)
<b>Asia</b>	6	47	0.93(0.12)	0.009(0.001)

<sup>a</sup>**n** is the sample size

<sup>b</sup>**S** is the number of segregating sites

<sup>c</sup>**Hd** is the haplotypic diversity of the sample

<sup>d</sup>**s.d** is the standard deviation of the estimates

<sup>e</sup> $\pi$  is the pairwise genetic diversity estimated with a Jukes and Cantor correction

### 3 Evidence of admixture between CSP subpopulations

Genetic structures between the CSP populations was also assessed by Bayesian clustering using STRUCTURE v.3.2 [4]. Clustering of individuals was determined on the basis of their haplotypes at the 6 sequenced loci. An admixture model was used that includes linkage among polymorphisms for the ancestry model in which polymorphic sites in each locus were treated as a single linkage group and the independence of allele frequencies among populations was assumed [1]. The probability of assigning individuals into clusters was estimated using  $5.0 \times 10^5$  iterations, after  $1.0 \times 10^5$  iterations as a burn-in period. The number of clusters (K) was set to 2, and the run was replicated 20 times to test the stability of the results. Because we are interested in assessing the underlying structure as explained by the phenotypes, no subsequent clustering ( $K > 2$ ) was assessed. It is clearly observed that all individuals in either CSP1 or CSP2 subpopulations seem to present a mixed origin in their genetic background, consistent with frequent gene exchange between this subpopulations (see figure S1).

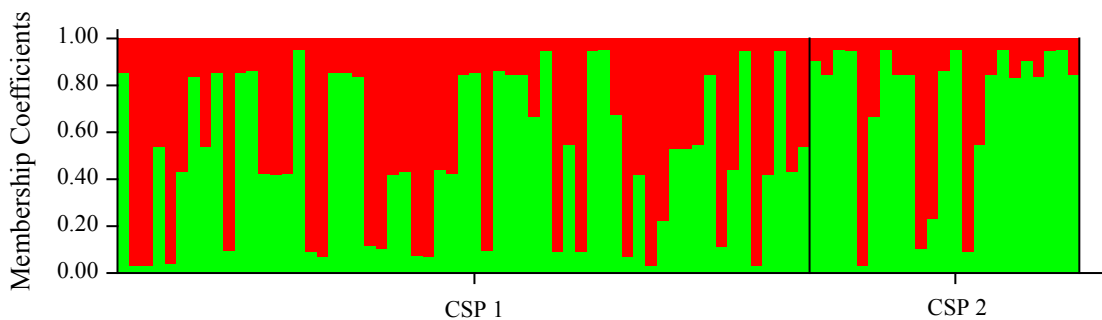


Figure S.1: BarPlot showing the membership coefficient to the subpopulations (for  $K=2$ ) as estimated under a model that allows for admixture.

Also consistent with the results shown in figure 1 on the main body of the manuscript, analyses performed with ClonalFrame [2] show that there is no clear clustering between isolates with different CSP types (see figure S2). The isolates presenting CSP1 or CSP2 phenotype does not seem to belong (all of them) to a particular clonal complex. Instead, numerous clusters present mixed phenotype composition. Multiple runs under different initial conditions (5 different runs) and similar run conditions were performed and

showed consistent results. The conditions for the runs were 100,000 generations burnin, 2,000,000 generations for the MCMC run after burnin, 10 branch-swapping events per iteration (2 additional runs were performed with 5 events per iteration). Convergence of the runs was checked by examination of the traces and also by comparing different runs via Gelman and Rubin statistics as suggested by the authors. The statistic was around 1.001 for  $\theta$ , 1.01 for  $R$ ,  $\nu$ , and  $\delta$  and 1.03 for TMRCA.

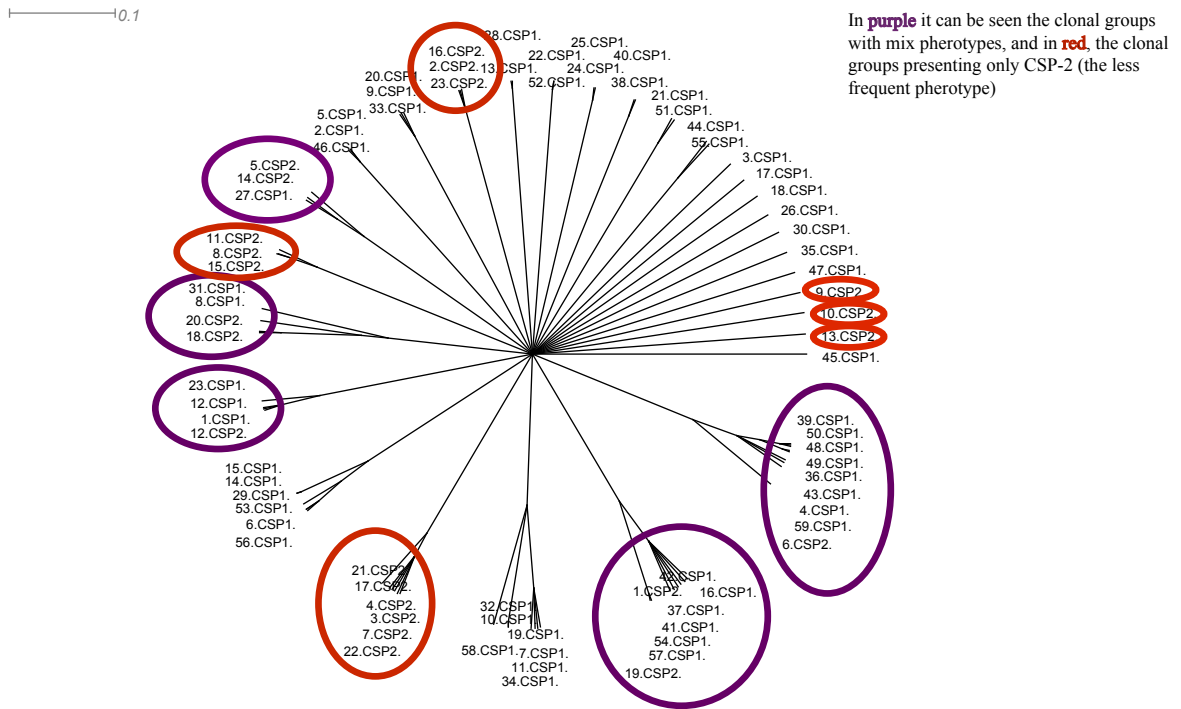


Figure S.2: Consensus tree obtained, as inferred with ClonalFrame. Drawing was done with Split-Tree4 [3]. The purple ovals show clusters of closely related isolates that have different pherotypes, while the red ovals show clusters that present pherotype 2 or CSP-2 (the less frequent pherotype)

## References

- [1] D. Falush, M. S. and J. Pritchard. 2003. Inference of population structure using multi-locus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**:1567–1587.
- [2] Didelot, X. and D. Falush. 2007. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**:1251–1266. URL <http://dx.doi.org/10.1534/genetics.106.063305>.
- [3] Huson, D. and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**:254–267. URL <http://mbe.oxfordjournals.org/cgi/content/full/23/2/254?ijkey=GvcBFw4QPIORzUE&keytype=ref>.
- [4] J. K. Pritchard, M. S. and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**:945–959.