**Simulation of Recombinations**

The ancestral genotype was taken as that of *S. pneumoniae* ATCC 700669 [EMBL accession code: FM211187], the earliest known isolate of the PMEN1 lineage. One hundred nucleotide paired end Illumina reads, with an insert size of 250 nt, were simulated from the other complete publically-available pneumococcal genomes listed in Table 1. These were then mapped against the *S. pneumoniae* ATCC700669 genome using SMALT v0.7.8 and bases called using the criteria described in [transformation paper] with short insertions and deletions processed using the GATK indel realignment software [ref]. This was used to generate a whole genome alignment that served as the input for simulations of diversification through different recombination processes.

Diversification was simulated from *t* = 0 to *t* = *t*max­ in discrete steps. At each step, every extant sequence acquired a single base substitution. Furthermore, each sequence had a probability *p*C of being duplicated into two independently diversifying sequences, corresponding to a coalescent event in the phylogeny. Each sequence also had a probability *p*S of being sampled at each timestep, after which point it no longer diversified or served as the progenitor to any other sequences. The simulation was stopped at *t*max once the total number of sampled and extant sequences in the simulation reached (or exceeded) a pre-determined maximum, *n*max.

At each discrete time step, recombinations occurred with a fixed probability *p*R depending on the model. These involved exchanging the relevant region of the recipient sequence in the alignment for the corresponding homologous sequence from a randomly-selected donor. These events had a random start location, and extended for a length following a geometric distribution (defined by a λ parameter) also dependent on the model.

**Model A - Regular recombination model**: under this model, recombinations occurred with *p*R = 0.1. Their lengths followed a geometric distribution described by the parameter λR value’s of 0.00016 bp-1 (corresponding to a mean length of 6.3 kb). These parameter estimates are taken from the original analysis of the PMEN1 lineage.

**Model B - Heterogeneous recombination model**: under this model, two modes of recombination occur. Both ‘micro’ and ‘macro’ recombinations independently occurred with the same *p*R = 0.03, with the size distributions of both defined by λR. Each instance of ‘micro’ recombination introduced a single event whereas each instance of ‘macro’ recombination introduced *Q* segments of sequence, where *Q* was randomly drawn from a Poisson distribution with a mean of 2.3. These values were taken from the fit of the mixture model to the PMEN1 data described in this study.

**Model C - Correlated mixture model**: under this model, the ‘micro’ and ‘macro’ recombinations both again occurred independently with the same *p*R = 0.03. However, in this case the different processes were associated with different length distributions: ‘micro’ recombinations were defined using λΣ = 0.0021 bp-1 (corresponding to a mean length of 480 bp), whereas ‘macro’ recombinations were defined by λΩ = 0.00011 bp-1 (corresponding to a mean length of 8.8 kb). These values were taken from the fit of the mixture model to the PMEN1 data described in this study.

**Model D: Uncorrelated mixture model**: under this model, the ‘micro’ and ‘macro’ recombinations both again occurred independently with the same *p*R = 0.03. Each instance of ‘micro’ recombination introduced a single event whereas each instance of ‘macro’ recombination introduced *Q* segments of sequence, where *Q* was randomly drawn from a Poisson distribution with a mean of 2.3. However, both ‘micro’ and ‘macro’ recombinations were an equal mix of events drawn from the two size distributions defined by λΣ = 0.0021 bp-1 and λΩ = 0.00011 bp-1.

Three simulations were run for each of the models, all of which used *n*­max = 242 generated with the parameter values *p*C = 0.05 and *p*S = 0.025. The alignment of each set of sequences was then analysed as described for the original set of PMEN1 isolates. The different models of recombination were then fitted to these outputs, the results of which are displayed in Table 2.

**Table 1** Details of sequences used as sequence donors in simulations.

|  |  |
| --- | --- |
| **Genome** | **Accession code** |
| *Streptococcus pneumoniae* 670-6B | CP002176 |
| *Streptococcus pneumoniae* 70585 | CP000918 |
| *Streptococcus pneumoniae* AP200 | CP002121 |
| *Streptococcus pneumoniae* CGSP14 | CP001033 |
| *Streptococcus pneumoniae* D39 | CP000410 |
| *Streptococcus pneumoniae* G54 | CP001015 |
| *Streptococcus pneumoniae* gamPNI0373 | CP001845 |
| *Streptococcus pneumoniae* Hungary19A-6 | CP000936 |
| *Streptococcus pneumoniae* INV104 | FQ312030 |
| *Streptococcus pneumoniae* INV200 | FQ312029 |
| *Streptococcus pneumoniae* JJA | CP000919 |
| *Streptococcus pneumoniae* OXC141 | FQ312027 |
| *Streptococcus pneumoniae* P1031 | CP000920 |
| *Streptococcus pneumoniae* R6 | AE007317 |
| *Streptococcus pneumoniae* ST556 | CP003357 |
| *Streptococcus pneumoniae* Taiwan19F-14 | CP000921 |
| *Streptococcus pneumoniae* TCH8431/19A | CP001993 |
| *Streptococcus pneumoniae* TIGR4 | AE005672 |

**Table 2** Results of model fitting to simulated data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Run** | **AIC model 1** | **AIC model 2** | **AIC model 3** | **AIC model 4** | **Mean Micro** | **Mean Macro** |
| A | 1 |  |  |  |  |  |  |
| A | 2 |  |  |  |  |  |  |
| A | 3 |  |  |  |  |  |  |
| B | 1 |  |  |  |  |  |  |
| B | 2 |  |  |  |  |  |  |
| B | 3 |  |  |  |  |  |  |
| C | 1 |  |  |  |  |  |  |
| C | 2 |  |  |  |  |  |  |
| C | 3 |  |  |  |  |  |  |