Chapter 4: Maela pneumococcal evolution and population-wide sequence exchange

---

Declaration of work contributions:

Analysis using BratNextGen discussed in 4.2.1.2 was performed by Dr Pekka Marttinen. Unless stated here, I was responsible for the analyses.

4. Maela pneumococcal evolution and population-wide sequence exchange

4.1 Introduction and aims

Understanding how pneumococcal populations evolve in carriage is essential for prevention and control of pneumococcal diseases. Prior to this study, evolutionary parameters, including rates of nucleotide substitution and homologous recombination, have been captured in a multidrug resistant lineage called PMEN-1 (Croucher, Harris *et al.* 2011). Diversity brought into the PMEN-1 genome by horizontal gene transfer, including homologous recombination, is frequently associated with genes conferring antibiotic resistance and antigenic variation. This introduction of genetic variation was proposed as a key contributor to the success of this lineage in evading clinical interventions and spreading globally. However, one can see from last chapter that the globally spread clones only make up a small proportion of the whole pneumococcal population at each locality, suggesting that observations from a single clone are unlikely to represent the whole pneumococcal community. Less was known about how other lineages in the population evolve and interact with each other.

This chapter explores the evolution and genetic interactions in a species-wide sampling of a pneumococcal population, from the densely sampled Maela collection.

This chapter aimed at:

i) Estimating lineage-specific evolutionary parameters including nucleotide substitution and recombination.

ii) Identifying the sources and sinks of recent recombination events.

4.2 Methods

4.2.1 Estimating lineage-specific evolutionary parameters

4.2.1.1 Criteria for selection of dominant clusters

Major lineages representing the whole population were selected for estimating and comparing lineage-specific evolutionary parameters. As currently available tools for detecting recombination are designed for a single lineage population rather multiple lineages (Croucher, Harris *et al.* 2011), the major clusters were chosen in such a way that the genetic diversity within the lineage must not exceed the limitations of recombination detection tool and the sample size of each cluster are is large enough to enable provide robust statistical power. Limitations of the recombination detection method was fully discussed in Croucher *et al.* 2013 (Croucher, Finkelstein *et al.* 2013). Out of eleven large primary clusters that comprised more than 100 isolates, seven primary clusters appear to have members that either share the same serotype/serogroup or differ by a few MLST locus variants, suggesting that isolates within these clusters are not too distant to exceed the tool limitations. Also, a phylogeny represents members of these main clusters on monophyletic branches. As a result, BC1-19F (n = 365 isolates), BC2-23F (n = 213 isolates), BC3-NT (n = 202 isolates), BC4-6B (n = 126 isolates), BC5-23A/F (n = 106 isolates), BC6-15B/C (n = 102 isolates) and BC7-14 (n = 102 isolates) (See **Appendix A**) were selected.

4.2.1.2 Recombination

Recombination fragments were identified using a phylogenetic-based algorithm as explained in (Croucher, Harris *et al.* 2011) and BratNextGen software (Marttinen, Hanage *et al.* 2012). Brief summaries of both algorithms are given below.

4.2.1.2.1 Croucher *et al.* method

Recombination fragments were identified using phylogenetic-based algorithms as explained in Croucher *et al.* 2011 (Croucher, Harris *et al.* 2011). In brief, the method

reconstructed a phylogeny based on the original alignments using RAxML v7.0.4 (Stamatakis 2006) followed by a reconstruction of nucleotide polymorphic sites on the phylogeny using PAML (Yang 2007). Regions on each phylogenetic branch where SNPs occur in clusters of significantly higher density than expected by chance (recombination events) were called and iteratively removed from the alignment. Remaining variants were then used to reconstruct the phylogeny, and recombination fragment detection repeated, iteratively, until there were no further changes in the tree topology. Recombination fragments were predicted from the SNPs dense regions in the whole genome alignment and were used to estimate lineage-specific rates of recombination.

4.2.1.2.2 BratNextGen method

In addition to above method, BratNextGen (**B**ayesian **R**ecombin**a**tion **T**racker) software (Marttinen, Hanage *et al.* 2012) was employed to identify recombination fragments. BratNextGen analysis was performed by Dr Pekka Marttinen, generating the predictions used in combination with results from Croucher *et al.* method described above. In brief, BratNextGen is a derivative of BAPS; the software exploits the data from different ancestral sources in each isolate predicted by BAPS (regarded as admixture events in the population) and expands the admixture model further by probabilistically characterising the origin of any particular site. This allows sites of recombination fragments with distinct origins to be identified, differentiating them from the rest of the genome. The software was applied on the whole genome alignment with the same default setting as in (Marttinen, Hanage *et al.* 2012). Significance of each putative recombinant segment (p-value <0.05) was determined through a bootstrap test with 100 replicates. As shown in (Marttinen, Hanage *et al.* 2012), the approach yielded highly similar recombination with the analysis of PMEN1 data (Croucher, Harris *et al.* 2011). Results generated from BratNextGen were thus used for cross-validation against results produced by Croucher *et al.* algorithm. Comparison between results generated by BratNextGen and Croucher *et al*. method will be discussed in this chapter.

4.2.1.2.3 Comparing level of recombination between different clusters

Following an identification of recombination, recombination per nucleotide substitution ratio (r/m) was calculated given the number of polymorphic sites produced by Croucher *et al.* algorithm, excluding any signals localised in the mobile genetic elements. The ratio r/m was calculated using two different approaches.

The first method modelled the relationship between recombination events and mutations as a linear regression under non-parametric distribution. Ranked recombination events and ranked number of SNPs were used as the outcome (y-axis) and the predictor variable (x-axis) respectively, with the slope representing r/m. ANCOVA tests were used to determine the significant difference in recombination rates between different clusters when statistical assumptions were met.

The second method used the arithmetic mean of r/m for each cluster, representing an average r/m from each branch within the cluster. The Kruskal-Wallis test was used to test for significant differences in r/m calculated by arithmetic mean. A consistency between two methods will give greater confidence over the results.

4.2.1.3 Mutation

4.2.1.3.1 BEAST

Mutation SNPs were separated from recombination SNPs using method described previously above (Croucher, Harris *et al.* 2011). However, there was difficulty in correlating the overall accumulation of SNPs through time from the whole cluster owing to a narrow sampling time frame. Therefore, correlations were performed within subclades of dominant clusters instead of using the whole cluster to capture the signals. Temporal signals and clock-like-ness in each subclade of the phylogenies were screened with Path-O-Gen v1.3 (Path-O-Gen 2010). Where positive correlations were observed (p-value <0.05), the mutation rates were then calculated with BEAST (**B**ayesian **E**volutionary **A**nalysis by **S**ampling **T**rees) (Drummond and Rambaut 2007) using the skyline population size prior and a relaxed lognormal clock model.

4.2.1.3.2 Comparing the rates between different clusters

The Kruskal-Wallis test was used to test for significant differences in mutation rates estimated from different clusters.

4.2.2 Tracing genetic exchanges through homologous recombination

Potential donor isolates for homologous recombination were identified from recombinant fragments detected in the recipients. Recombination fragments co-predicted by the methods described above were used as query sequences for nucleotide blast searches for potential donor blocks from 3,085 draft assembled genomes. Using blastall v 2.2.15 (Altschul, Gish *et al.* 1990), hits that have an exact match to the query sequence were likely to be potential donors. Several criteria and filters were applied onto these analyses to reduce detection of false positives. The rationale for the criteria will be discussed in full in the chapter.

Probabilities of a single isolate, as well as of each BAPS cluster acting as a donor for a recipient were then calculated. For each recipient isolate, “n” potential donor isolates were identified, and each donor isolate was assigned a probability of “1/n” of having been the donor. Isolates showing no particular hit for a particular search were given a probability of 0. The total frequency of each isolate in being the donor was represented by the sum of the above probabilities from all potential donation events. Isolates were grouped into lineages based on their population cluster from BAPS. The relationship between cluster size, cluster diversity and probability of being a donor were estimated using Spearmann’s ranking correlation.

4.3 Results

4.3.1 Estimating evolutionary rates within the population

Two evolutionary parameters, nucleotide substitution and homologous recombination were investigated here. This section documents how the two evolutionary signals were separated, identifying comparable nucleotide substitution yet heterogeneous recombination rates within the pneumococcal population and the potential biological implications.

4.3.1.1 Separating recombination signals from single nucleotide substitutions

To estimate the evolutionary rates, a phylogenetic network would be required to calibrate observed genetic changes with time. However, a conventional phylogeny typically assumes a single history from the whole genome, which is not true when recombination events have occurred. Different regions in the alignment affected by recombination may have different underlying phylogenies, resulting in an abrupt change in tree topology (Posada and Crandall 2002, Ruths and Nakhleh 2005). Here, signals from recombination were distinguished from nucleotide substitutions using the method described in Croucher *et al.* (Croucher, Harris *et al.* 2011). Recombination is generally identified as SNP dense regions in the genome. The algorithm searches the whole genome for regions with high SNP density and iteratively removes these suspected recombined regions until there is no further change in tree topology.

The method has been successfully applied to a single lineage study of 240 PMEN-1 isolates (Croucher, Harris *et al.* 2011). However, it was not feasible to run the algorithm on the entire species-wide dataset of 3,085 genomes. A preliminary test on a smaller species-wide pneumococcal dataset of 127 isolates from Malawi (Everett, Cornick *et al.* 2012) showed that the method failed after reaching a high diversity threshold. A starting tree prior to the removal of recombination displayed a typical long-branch phylogeny, which is a result of the combined SNPs from both nucleotide substitutions and recombination. The removal of recombination in this diverse genetic background proved to be difficult. With each lineage having different evolutionary history, their recombination patterns are distinct in each isolate. This resulted in a

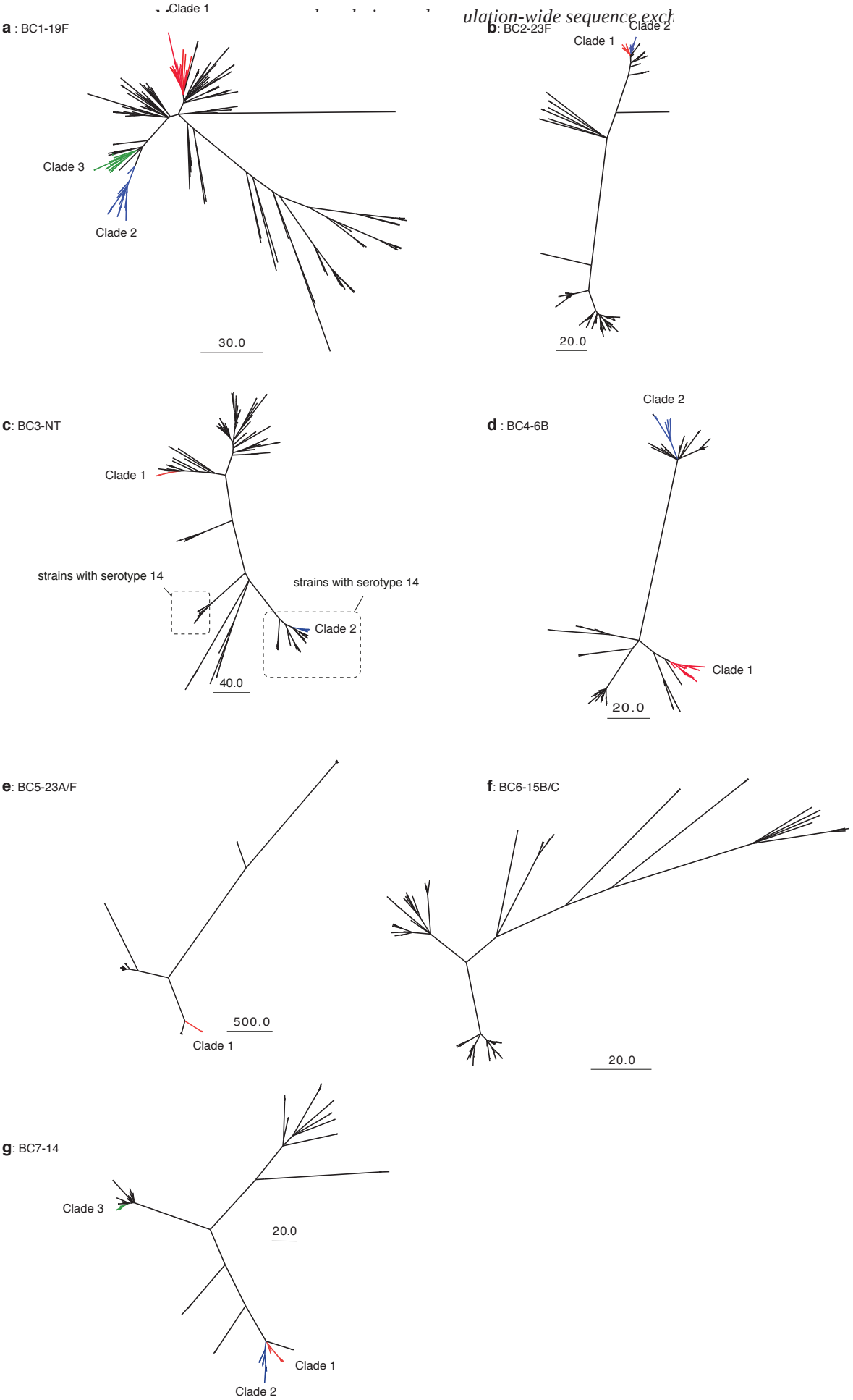
large number of putative recombination events in total. Removal of all recombination signatures from every single isolate removed nearly all the SNPs from the alignment after the first iteration. This left an inadequate number of remaining SNPs to construct a robust phylogenetic tree and consequently terminated all further iterations after the first round. In a single lineage scenario like PMEN-1, recombination from every single isolate accounted for 74% of the reference genome (Croucher, Harris *et al.* 2011), which left sufficient SNPs from nucleotide substitution to draw a vertical phylogeny. Therefore, the method was applied separately to each dominant lineage in Maela instead of the whole dataset.

The largest seven primary BAPS clusters, here denoted by BC1 to BC7 (**Figure 3.4 a**, **Appendix A**), were used as representatives for this analysis. Each cluster contains more than 100 samples and totalled 1,216 genomes accounting for 39.4% of the Maela pneumococcal population: BC1-19F (n=365 isolates), BC2-23F (n=213 isolates), BC3-NT (n=202 isolates), BC4-6B (n=126 isolates), BC5-23A/F (n=106 isolates), BC6-15B/C (n=102 isolates), and BC7-14 (n=102 isolates). It is possible that evolutionary rates from less prevalent lineages are significantly different from the dominant groups selected here and this should be investigated. However, they cannot be reliably assessed with their current sampling size nor can they be grouped together due to the complexity imposed on the computations. Therefore, minority groups were not considered here.

Analyses within focused clusters required genome alignment with higher resolution. To improve the resolution seen in previous chapter, sequence reads for each BAPS cluster were remapped against a closely related reference genome (see Methods) to allow greater sensitivity for detection of variants including single polymorphic changes and small insertions and deletions (indels). Applying the method described in (Croucher, Harris *et al.* 2011) on the improved alignment of each BAPS cluster separated recombination signals from nucleotide substitutions. This resulted in a final tree where only single nucleotide substitution accounts for its evolutionary history (**Figure 4.1**). Subsequently, predicted recombination events were allocated to each phylogenetic branch.

Figure 4.1 Nucleotide substitution based phylogeny and the clusters from which the nucleotide substitution rates were estimated.

Each panel represents a major Maela cluster: (a) BC1-19F, (b) BC2-23F, (c) BC3-NT, (d) BC4-6B, (e) BC5-23A/F, (f) BC6-15B/C and (g) BC-14. Subclades where substitution rates were estimated are highlighted in different colours (blue, green and red) and labelled accordingly. Please note that substitution rates cannot be confidently estimated from any clades in BC6-15B/C. The scale bar represents the number of SNPs. (**Figure is shown on the next page**)


a : BC1-19F
Clade 1
Clade 3
Clade 2
30.0
b: BC2-23F
Clade 1
Clade 2
20.0
c: BC3-NT
Clade 1
strains with serotype 14
strains with serotype 14
Clade 2
40.0
d : BC4-6B
Clade 2
Clade 1
20.0
e: BC5-23A/F
500.0
Clade 1
f: BC6-15B/C
20.0
g: BC7-14
Clade 3
20.0
Clade 1
Clade 2

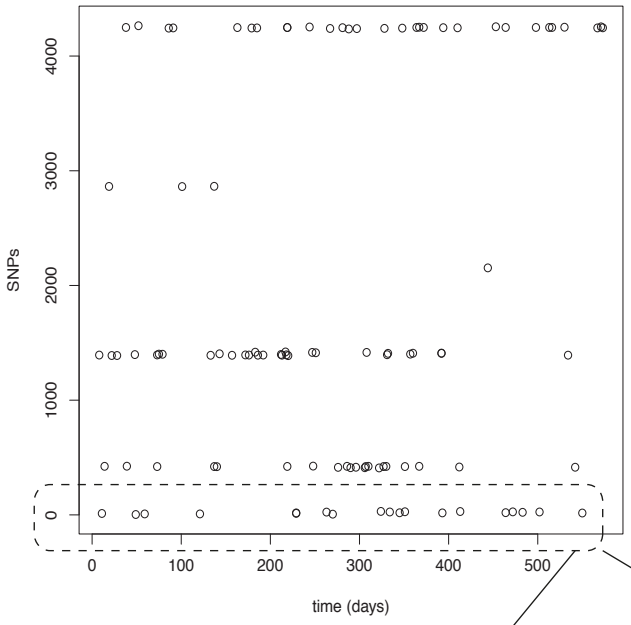
4.3.1.2 Rates of single nucleotide substitution

For each dominant lineage, the rate of single nucleotide substitution can be calculated from a rooted, time-measured phylogeny inferred from the nucleotide substitution tree discussed in previous section. However, there was difficulty in correlating the accumulation of nucleotide substitutions through time with the whole BAPS cluster. As each cluster consists of smaller subclades that co-evolve at the same time, the three-year time span of the isolate collection is not long enough to resolve the combined evolutionary signals from multiple subclades. To illustrate the difficulty in correlating nucleotide substitution and temporal signals from the whole cluster, **Figure 4.2** (top panel) illustrates the absence of nucleotide substitution-time signal when combined signals were considered using BC5-23A/F as an example. A linear regression cannot be determined from the plot as a whole. However, there are distinct clusters representing small subclades where there are good correlations between nucleotide substitution and time. When considering each subclade individually, the temporal signal and clock-likeness of each subclade phylogeny can be captured (**Figure 4.2** bottom panel). A correlation between substitution and time of each subclade was validated using Path-O-Gen (**Figure 4.3**). The subclades showing a positive correlation are highlighted in **Figure 4.1** and were used to estimate the substitution rates with Bayesian MCMC analysis (BEAST, see Methods).

Mean estimated substitution rates fell within the range of 1.45-4.81 x 10^{-6} substitutions per site per year with overlapping 95% credibility intervals (**Table 4.1**). These estimated rates are consistent with the previous report of the PMEN-1 lineage (1.57 x10^{-6} substitutions per site per year, 95% confidence interval 1.34 to 1.79 x 10^{-6}) (Croucher, Harris *et al.* 2011). Although the results appear to be consistent, the rates estimated here have broader confidence intervals due to the much shorter time span of the sampled collection.

Figure 4.2 Demonstration that clock-like signals can be detected from the subclades but not from the whole population.

Each dominant cluster is comprised of more than a single subclade that coevolve together. This plot used BC5-23A/F as an example. The clock signal cannot be detected using the whole cluster as there is confounding from the signals of the subclades.



When all members of the cluster was considered, there was no relationship between time and accumulation of nucleotide polymorphisms. However, distinct subclades can be observed from the plot.

A zoom into the subclade show that there is a positive correlation between time and SNPs, allowing mutation rate to be calculated.

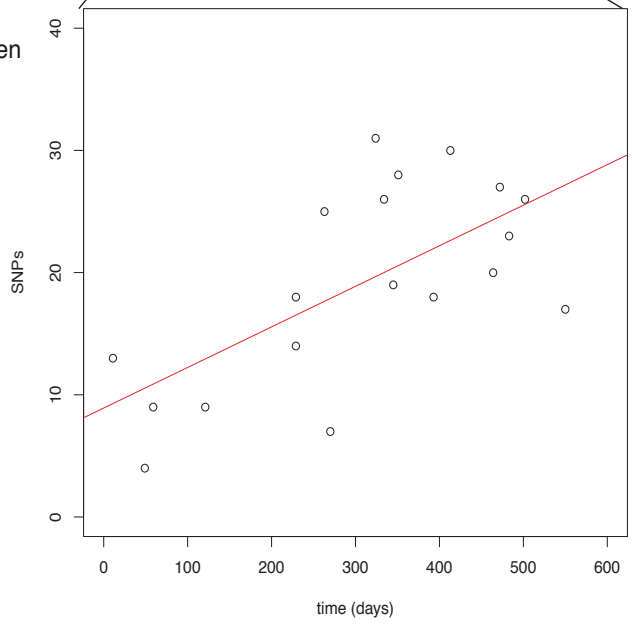


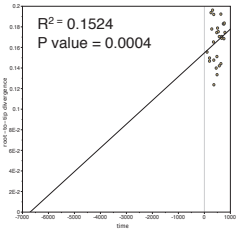
Figure 4.3 Clock-like signals from Path-O-Gen in the subclades where substitution rates were estimated

These subclades were highlighted in Figure 4.1. The y-axis reports root-to-tip divergence while the x-axis represents the time scale in days from the first date of collection, which was 12th November 2007. The first date (time = 0) is shown as a vertical dashed line.

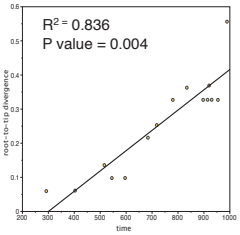
Please note that while a combination of high R^2 and low p-value is expected, regression with low R^2 and low p-value are observed in some cases. These do not necessarily conflict with each other as each represents different predictions. While R^2 measures how close the data are to the fitted regression line, the p-value tests the null hypothesis that the coefficient is equal to zero (no correlation).

(Figure is shown on the next page).

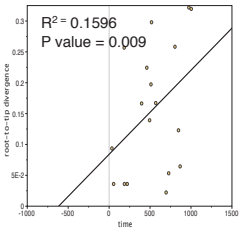
BC1-19F: clade 1



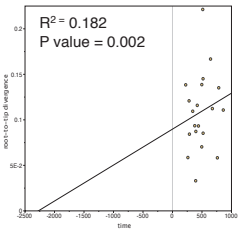
BC1-19F: clade 2



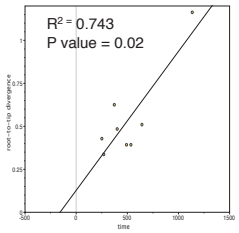
BC1-19F: clade 3



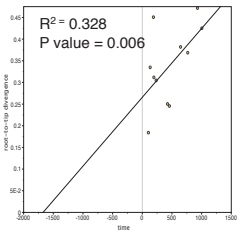
BC2-23F: clade 1



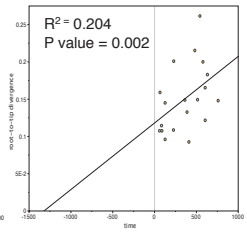
BC3-NT: clade 1



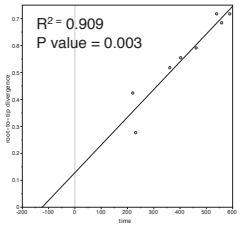
BC3-NT: clade 2



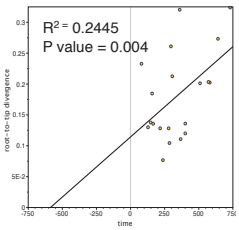
BC4-6B: clade 1



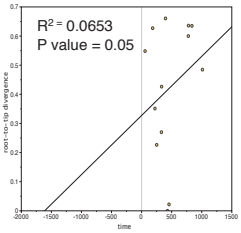
BC4-6B: clade 2



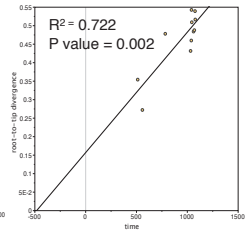
BC5-23A/F: clade 1



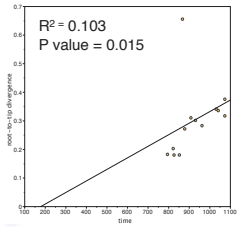
BC5-23A/F: clade 2



BC7-14: clade 1



BC7-14: clade 2



BC7-14: clade 3

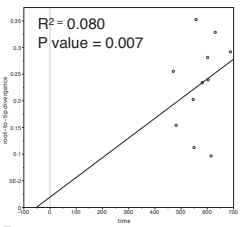


Table 4.1 Nucleotide substitution rates estimated by BEAST

Mean nucleotide substitution rate, the lower bound of the 95% highest posterior density (HPD) and the upper bound of the 95% HPD were tabulated respectively.

	Nucleotide substitution rate (substitutions per site per year)		
	mean	**lower bound**	**upper bound**
BC1-19F clade 1	2.60×10^{-6}	1.58×10^{-6}	3.71×10^{-6}
BC1-19F clade 2	2.86×10^{-6}	1.69×10^{-6}	4.07×10^{-6}
BC1-19F clade 3	2.35×10^{-6}	1.15×10^{-6}	3.68×10^{-6}
BC2-23F clade 1	1.83×10^{-6}	9.96×10^{-7}	2.70×10^{-6}
BC2-23F clade 2	1.45×10^{-6}	2.89×10^{-7}	2.66×10^{-6}
BC3-NT clade 1	4.49×10^{-6}	1.61×10^{-6}	9.78×10^{-6}
BC3-NT clade 2	2.39×10^{-6}	8.49×10^{-7}	4.38×10^{-6}
BC4-6B clade 1	3.08×10^{-6}	1.29×10^{-6}	5.13×10^{-6}
BC4-6B clade 2	3.07×10^{-6}	4.36×10^{-7}	9.83×10^{-6}
BC5-23A/F	3.26×10^{-6}	1.02×10^{-6}	6.60×10^{-6}
BC7-14 clade 1	2.79×10^{-6}	1.16×10^{-6}	4.67×10^{-6}
BC7-14 clade 2	4.81×10^{-6}	5.31×10^{-7}	9.77×10^{-6}
BC7-14 clade 3	4.39×10^{-6}	2.65×10^{-6}	8.14×10^{-6}

4.3.1.3 Rates of recombination

Recombination levels of the 7 dominant clusters were calculated, given numbers of recombination events and number of nucleotide substitutions computed for each branch of the phylogenetic tree as described earlier. Recombination signals can result from both site-specific recombination, potentially associated with mobile genetic elements, and homologous recombination. Signals localised in the regions of mobile genetic elements including prophages and integrative conjugative elements (ICEs) were removed so that only homologous recombination was considered.

Here, the level of recombination was estimated using the relative scale of **r**ecombination events per number of nucleotide substitutions or **m**utations (r/m) observed in each branch. For the numerator (r), the number of recombination events was used instead of numbers of polymorphic sites introduced by recombination as originally given in (Feil, Maiden *et al.* 1999). Depending on the genetic distance between the donors and recipients of a recombination event, numbers of polymorphic sites incorporated in a single recombination event can be variable. Closely related DNA donors may have lower sequence variation compared to distant donors, and thus potentially bring a bias when comparing level of recombination across different clusters. With the use of recombination events instead of polymorphic sites introduced by recombination, the r/m calculated here is expected to be lower than reported earlier (Croucher, Harris *et al.* 2011, Croucher, Finkelstein *et al.* 2013).

The ratio r/m was calculated and compared by two different approaches.

i) By modelling the relationship between recombination events and mutations as a linear regression: recombination events ~ nucleotide substitutions (**Figure 4.4**), using the ranked recombination events as the outcome, and ranked number of nucleotide substitutions as the predictor variable. The slope of each plot represents r/m. Where the assumptions of linearity were met, r/m was calculated and reported in **Table 4.2**.

ii) By using the arithmetic mean of r/m of a cluster. For each cluster, the r/m was calculated separately for each branch and then averaged. Mean and distribution of the r/m of each cluster are tabulated in **Table 4.2**.

With the exception of BC7-14, the r/m calculated from two different approaches shows a good overlap within each cluster, generating a result in which one can have more confidence. The ratio was found to be less than 1 in all studied clusters, indicating that recombination events occur less frequently than nucleotide substitutions.

Figure 4.4 Recombinations per mutation (r/m) of each cluster calculated by linear regression.

Due to the large sample size available in our studies, we alternatively calculated the ratio of recombination events (y axis) over single nucleotide substitutions (x axis) observed on each branch of the slope (r/m) of the linear regression. The number is tabulated in **Table 4.2**. For comparison of r/m by linear regression, all the data are ranked to accommodate the non-parametric ANCOVA analysis. (**Figure is shown on the next page**)

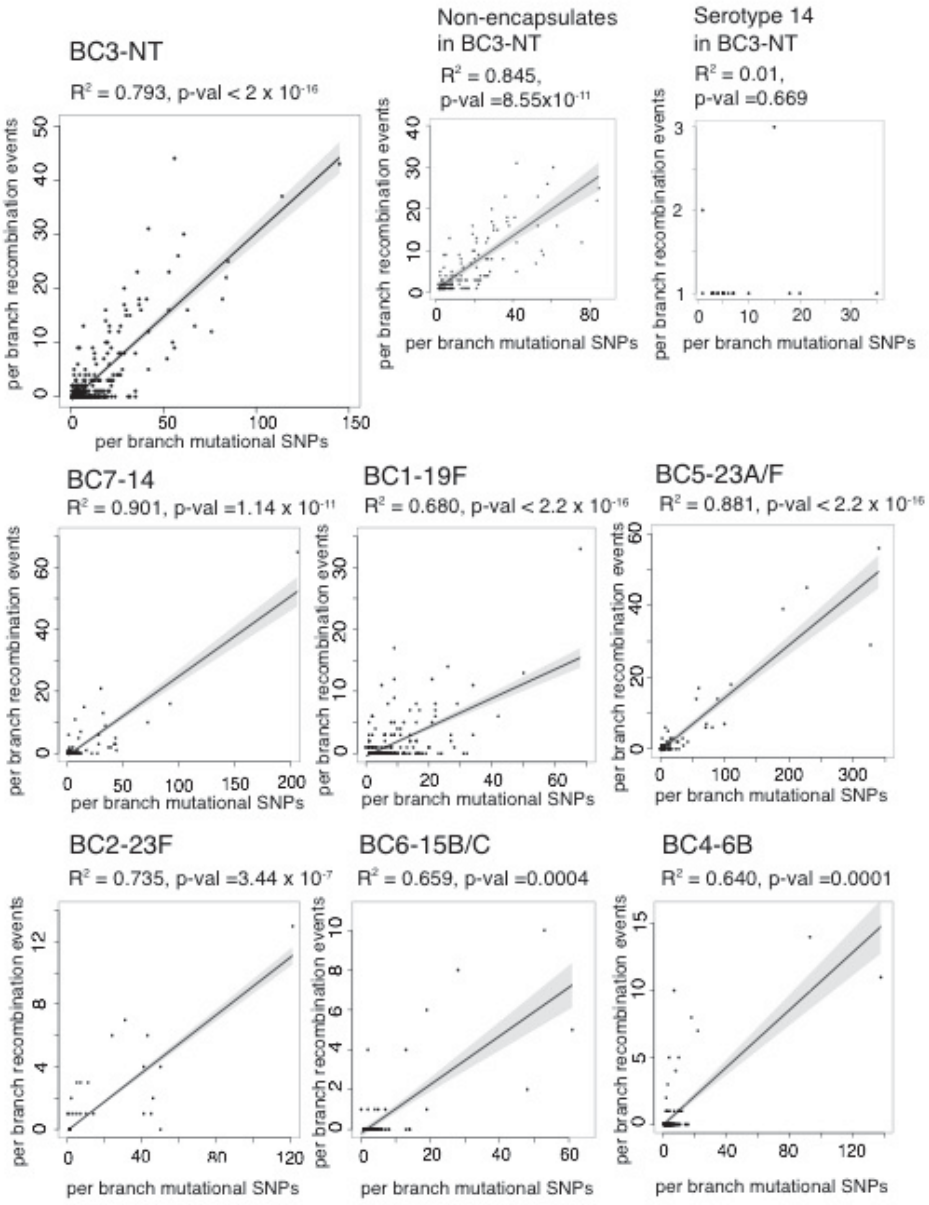

BC3-NT
R^2 = 0.793, p-val < 2 x 10^{-16}
Non-encapsulates in BC3-NT
R^2 = 0.845, p-val =8.55x10^{-11}
Serotype 14 in BC3-NT
R^2 = 0.01, p-val =0.669
BC7-14
R^2 = 0.901, p-val =1.14 x 10^{-11}
BC1-19F
R^2 = 0.680, p-val < 2.2 x 10^{-16}
BC5-23A/F
R^2 = 0.881, p-val < 2.2 x 10^{-16}
BC2-23F
R^2 = 0.735, p-val =3.44 x 10^{-7}
BC6-15B/C
R^2 = 0.659, p-val =0.0004
BC4-6B
R^2 = 0.640, p-val =0.0001
per branch recombination events
per branch mutational SNPs

Table 4.2 Recombination per mutation (r/m) calculated from linear regression and arithmetic mean.

Testing clusters	r/m		Hypothesis	Test and p-value	
	Estimated by linear regression (95% confident interval)	Estimated by arithmetic mean (95% confident interval)		ANCOVA (difference in slope calculated by linear regression)	Kruskal-Wallis (difference in arithmetic mean)
BC1-19F	0.233 (0.210-0.256)	0.299 (0.195-0.263)	r/m of BC3-NT > other clusters	$1.10x10^{-3}$	$1.76x10^{-5}$
BC2-23F	0.092 (0.068-0.117)	0.140 (0.068-0.212)			
BC3-NT	0.310 (0.284-0.336)	0.320 (0.289-0.351)			
BC4-6B	0.107 (0.005-0.209)	0.132 (0.037-0.227)			
BC5-23A/F	0.147 (0.132-0.162)	0.146 (0.100-0.192)			
BC6-15B/C	0.122 (0.070-0.174)	0.200 (0.115-0.285)			
BC7-14	0.257 (0.211-0.395)	0.148 (0.086-0.210)			
NT within BC3-NT	0.341 (0.288-0.395)	0.343 (0.307-0.379)	Within BC3-NT, r/m of NT > serotype 14	NA	$2.44x10^{-3}$
Serotype 14 within BC3-NT	Assumptions of the linear regression models were not met	0.203 (0.164-0.242)			

4.3.1.4 Comparison of evolutionary rates within the population

Next, evolutionary parameters estimated in dominant clusters were compared (**Figure 4.5** top panel). There was no significant difference in rates of nucleotide substitutions between major clusters (Kruskal-Wallis test p value = 0.98). However, the levels of recombination, estimated by the r/m ratio, were significantly different between clusters (Kruskal-Wallis test p value = 1.24×10^{-8}) (**Figure 4.5** bottom panel). The difference in levels of recombination is consistent with previous genome-based (Croucher, Finkelstein *et al.* 2013) and *vitro* studies (Ravin 1959, Yother, McDaniel *et al.* 1986, Hsieh, Wang *et al.* 2006). This might suggest a potential difference in speed of adaptation to changes in environment within the population.

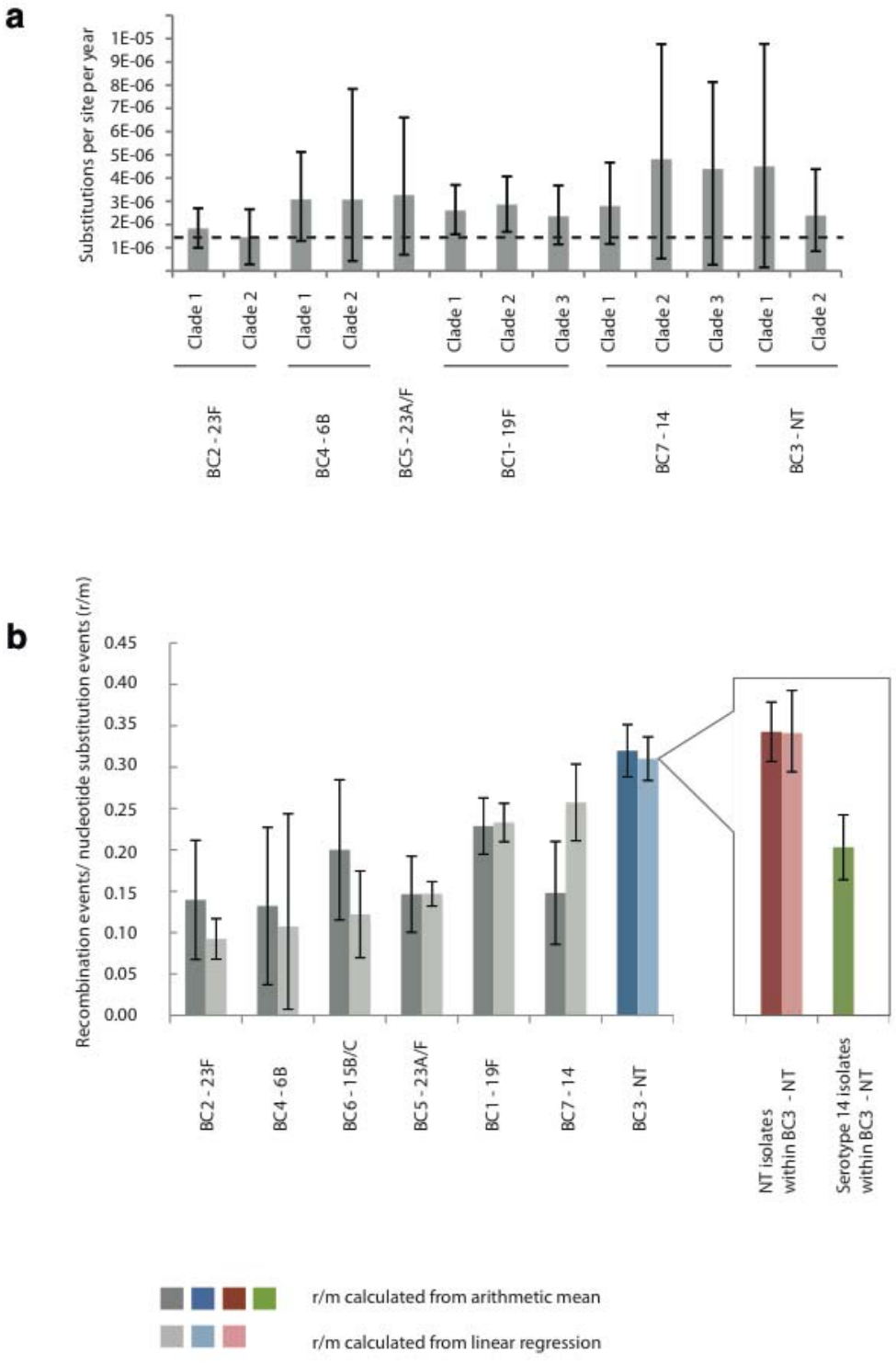
The highest recombination rate was observed in a group dominated by nontypable isolates (BC3-NT) (comparing NT against other groups: Mann-Whitney test p value = 1.76×10^{-5}, ANCOVA test p value = 0.0011, **Table 4.2**). This is consistent with the observation of increased transformation efficiency in capsule defective mutants (Pearce, Iannelli *et al.* 2002); a property widely exploited in laboratory experiments. This is consistent with the general concept that the capsule might act as a physical barrier to DNA uptake *in natura*. Cluster BC3-NT includes a sub-set of isolates able to express the serotype 14 capsule, thereby providing a route to test the idea of the capsule limiting DNA uptake within the same genetic background. When calculated separately, the r/m was significantly higher in the NT isolates compared to serotype 14 isolates (Man-Whitney test p value = 2.44×10^{-3}, **Table 4.2**), indicating that encapsulation reduces recombination efficiency.

Given similar substitution rates, the differences in rates of recombination are likely to be an important factor in the ability of clusters to acquire exogenous DNA with potential selective advantages. With higher level of DNA uptake observed in NTs, one may speculate that they might be fast adapters to environmental changes, and potentially promote the whole population in adjusting to new challenges by disseminating their selectively advantageous genes. However, such a role cannot be established without knowing how much NTs contribute their DNA to the population DNA pool. This question will be tackled in the next section.

Figure 4.5 Comparison of evolutionary parameters estimated in dominant clusters

(a) Comparison of single nucleotide substitution rates estimated using BEAST. Error bars correspond to 95% credibility intervals. The dashed line represents the mutation rate estimated in a previous pneumococcal study of 1.57×10^{-6} substitutions per site per year (95% confidence interval of $1.34\text{-}1.79 \times 10^{-6}$) (Croucher, Harris *et al.* 2011). (b) Comparison of recombination events per mutation (r/m) across the dominant clusters quantified by two separate methods: linear regression on each branch of the appropriate phylogeny and the arithmetic mean of r/m on each branch. Error bars represent 95% confidence intervals. BC3-NT (in blue) has the highest r/m ratio, with its subclusters NT and serotype 14 highlighted in red and green, respectively. Please note that the assumptions of the linear regression models were not met for serotype 14.

(**Figure is shown on the next page**)


a
Substitutions per site per year
1E-05
9E-06
8E-06
7E-06
6E-06
5E-06
4E-06
3E-06
2E-06
1E-06
Clade 1
Clade 2
BC2 - 23F
BC4 - 6B
BC5 - 23A/F
BC1- 19F
BC7 - 14
BC3 - NT
b
Recombination events/ nucleotide substitution events (r/m)
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
BC2 - 23F
BC4 - 6B
BC6 - 15B/C
BC5 - 23A/F
BC1 - 19F
BC7 - 14
BC3 - NT
NT isolates within BC3 - NT
Serotype 14 isolates within BC3 - NT
r/m calculated from arithmetic mean
r/m calculated from linear regression

4.3.2 Population-wide sequence exchange

With high sampling density, potential sources of recombination fragments (referred to as the "donor blocks") can be determined given the sequence's identity to the recombined fragments detected in the recipient strains (referred to as the "recipient block"). This allowed a unique opportunity to capture potential recombination donors, which are important players at the other end of the gene flow.

4.3.2.1 Searching for potential recombination donors and search criteria

The recipient blocks identified in BC1- BC7 were searched using BLAST against the rest of the pneumococcal genomes in the Maela data set for identical matches (donor blocks). As the Maela pneumococcal population has co-evolved in the same geographical area, the chance of different sampled pneumococcal lineages exchanging genetic content is high. Several criteria, which will be discussed in the next section, have been tested and applied in this analysis to reduce the false positives and maximise the search specificity. As a result, 443 out of 928 unique recipient blocks were found to have identical matches elsewhere in the data set (**Table 4.3**).

Table 4.3 Numbers of recombination events used for the search, and number of recipient blocks where potential donors were identified following strict search criteria

Total number of recipient blocks used for the search following different criteria			Number of recipient blocks where potential donor blocks were identified
A) All recombination events predicted by method described in Croucher *et al.* (generated in 4.1.1)	**B)** Recombination detected at the external nodes using method in Croucher *et al.*, which are likely to represent recent recombination (criteria discussed in 4.2.1.1)	**C)** Recombination blocks predicted by method described in Marttinen *et al.* that show overlap with regions in B) (criteria discussed in 4.2.1.2)	**D)** Recipient blocks in C) where potential donors can be found (applying criteria discussed in 4.2.1.3, 4.2.1.4, and 4.2.1.5)
2,209	620*	928*	443

* Please note that some recombination blocks predicted by the method of Croucher *et al.* contain multiple blocks predicted by that of Marttinen *et al.*

Criteria used for identifying potential recombination donors include: focusing the search on recent recombination events, using overlapping recombination blocks predicted by two independent algorithms, allowing only identical hits with no unknown mapping character “N”, and an overall check on the search specificity. Each point and its rationale are discussed as follows.

4.3.2.1.1 Only recent recombination events were considered

A focus on recent recombination occurring on the external branches alone reduces the chances of the donor detection being confounded by subsequent recombination events. Therefore only recent recipient blocks detected on the external branches (identified using the algorithm described in Croucher, Harris *et al.* 2011 were considered.

4.3.2.1.2 Using two independent algorithms for predicting recombination fragments in the recipients

Any methods, to a certain level, report false positives. Based on detecting the density of SNPs in a sliding window, the method described in Croucher, Harris *et al.* 2011 allows for high flexibility such that poor mapping regions flanked by high SNP densities can be counted as recombination fragments. This can be observed, for example, for surface protein encoding genes where high sequence diversity may affect mapping against the reference genome due to sequence mismatches. The algorithm is designed to merge such mismatched gaps flanked by SNPs dense region together as single recombination fragments. Results generated by this method thus represent all possible cases of recombination.

However, a search for donor blocks using recipient blocks as the sequence query required a good sequence quality as well as a confident prediction of recombination fragments. To reduce false positives generated in one method, as well as eliminate regions with poor mapping quality due to naturally high sequence diversity allowed in previous algorithm, another algorithm was used to co-detect recombination regions.

BratNextGen was developed to identify foreign DNA fragments that were introduced into the genome by recombination and this has been successfully applied to

pneumococcal genomes (Marttinen, Hanage *et al.* 2012). Unlike the previous algorithm, which takes poor quality regions with flanking recombination signals into account, BratNextGen handles missing data in a different manner and leads to better sequence quality in predicted recombination fragments.

Here, recombination fragments predicted by two independent algorithms were compared. A large majority of predicted regions show overlaps. This can be demonstrated in **Figure 4.6 a** using the phylogeny and overlapping recombined regions detected in BC7-14, one of the smallest dominant clusters, as an example. The length and quality of the predicted fragments from both algorithms were investigated. **Figure 4.6 b-c** and **Table 4.4** summarise comparisons of the length and percent of unknown characters "N" found in the fragments from both algorithms respectively. The results show that while the method described in Croucher, Harris *et al.* 2011 generally predicts larger recombination blocks, BratNextGen (Marttinen, Hanage *et al.* 2012) gave predicted sequences with higher quality. To optimise the output, only sequences co-predicted from both algorithms were used in the next part of the analysis.

Figure 4.6 Comparison of two recombination detection methods

(a) Genome view of recombination fragments predicted by both algorithms. Recombination regions are aligned with taxa on the phylogenetic tree (left). Genome coordinates are labelled on top. Recombination regions exclusively predicted by methods described in Croucher, Harris *et al.* 2011 and Marttinen, Hanage *et al.* 2012 are highlighted in red and blue, respectively. Overlapping regions predicted by both algorithms are highlighted in dark grey. (b) A histogram showing the length of recombination fragments (bp) predicted by two algorithms. (c) Sequence quality of recombination fragments predicted by two algorithms reported as percent "N". For (b) and (c), fragments predicted by tools described in Croucher, Harris *et al.* 2011 and Marttinen, Hanage *et al.* 2012 are shaded in red and blue respectively.

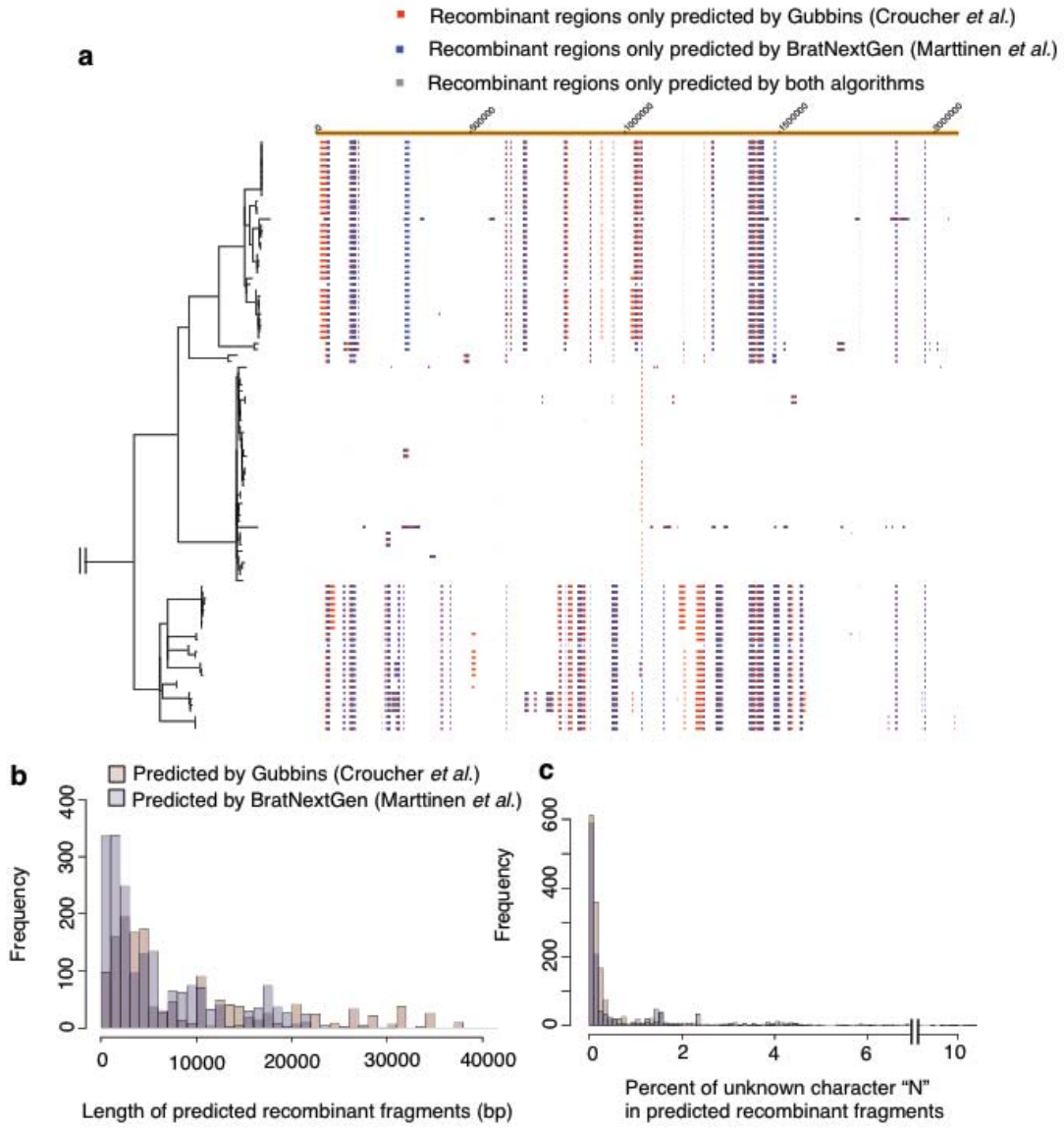


Table 4.4 Comparison of two recombination detection methods as given by Figure 4.6

	methods	1st Quantile	Median	Mean	3rd Quantile
Length of recipient blocks in **Figure 4.6 b**	Croucher *et al.*	2403	4558	10150	14350
	Marttinen *et al.*	1189	3427	5777	8500
	Co-predicted	1177	3057	5418	21263
Percent of "N" in recipient blocks in **Figure 4.6 c**	Croucher *et al.*	0.084	0.200	6.05	2.36
	Marttinen *et al.*	0.066	0.153	5.59	2.00

4.3.2.1.3 No unknown mapping character "N" was allowed

As "N" s, unknown nucleotides, generate non-specific matches in BLAST searches, recipient and donor blocks were checked for sequence quality such that no sequences with "N"s were used.

4.3.2.1.4 Hits must be identical matches

Recombination recipient blocks were used as query sequences for nucleotide blast searches. They were blasted against themselves as positive controls. Any hits that have an exact match to the score given by the positive control are likely to be potential donors. Although recombination may lead to insertions or deletions over recombining regions (Claverys, Lefevre *et al.* 1980, Lefevre, Mostachfi *et al.* 1989, Pasta and Sicard 1996), indels were not considered here as the searches were performed in closely related strains, which require high specificity for the recipient-donor relationship to be drawn.

4.3.2.1.5 A relationship between query length and search specificity

Length of recipient blocks (blast queries) and the number of potential donors detected (blast hits) were checked for search specificity. Queries with successful hits ranged

between 10 - 6,846 bp, with a mean length of 1,162 bp (**Table 4.5**). **Figure 4.7 a** shows similar distribution in length of recipient blocks shown by all queries and queries where hits were detected. This suggests that the blast search did not impose any preference over query length. A decrease in query length usually correlates with increasing broadness or generality of the hits (Phan 2006). The distribution of hits can be modelled with a negative exponential function (**Figure 4.7 b**) in this result. The decay constant of 2.56×10^{-4} bp^{-1} observed here means that on average, the numbers of hit clusters are reduced by half for every 3,903 bp extension of sequence query. These behaviours gave some confidence to the search results. Therefore, all sequence queries that passed criteria described earlier were used for blast searches.

Figure 4.7 Query length and search specificity

(a) A histogram showing distribution of length of recipient blocks from recombination events detected at the tip of the phylogenies. Shaded in grey are all recipient blocks where identical hits were detected from the rest of the population. (b) A plot showing association between the length of sequence queries (recipient blocks) and the diversity of detected hits (potential donor blocks classified by secondary BAPS clusters). The data was modelled as an exponential decay with the line of best fit (red line) and the 95% confidence interval (dashed red lines)

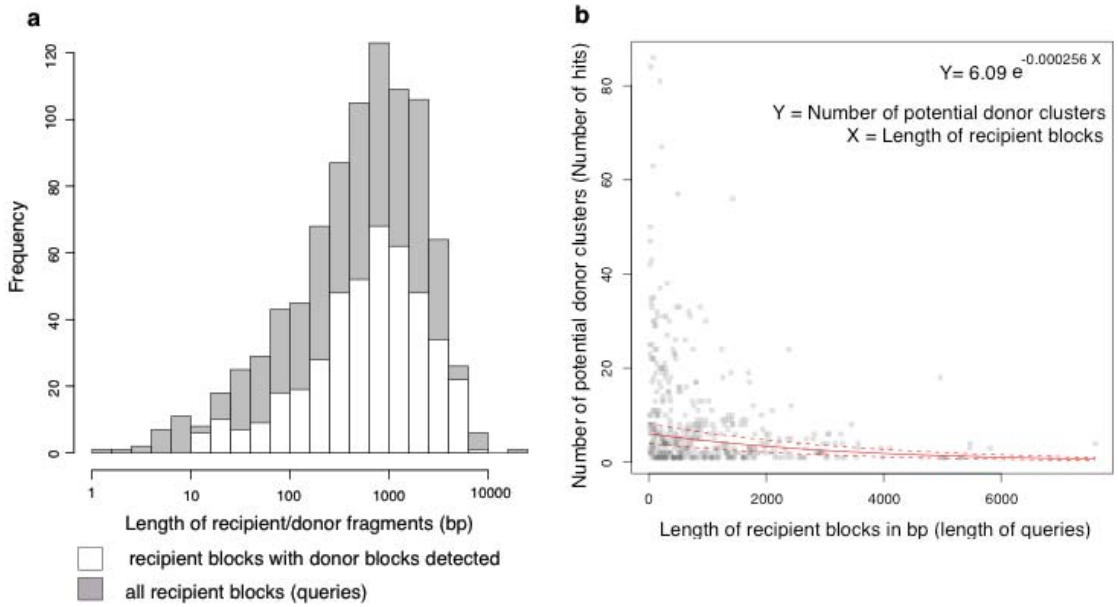


Table 4.5 Distribution of length of recipient blocks described in Figure 4.7 a

Length of fragments (bp)	**Min**	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Max**
All recipient blocks (queries)	3	197	604	1,072	1,412	22,970
Recipient blocks with donor blocks detected (identical hits)	10	275	741	1,162	1,513	6,846

4.3.2.2 Nature of sequence exchange

Following identification of potential recombination donors in 4.2.1, donor strains were classified based on their BAPS clusters (**Appendix A**). This highlighted some interesting biologically features of the genetic flow between the recipient and donor strains.

4.3.2.2.1 A single recipient strain can have multiple donors

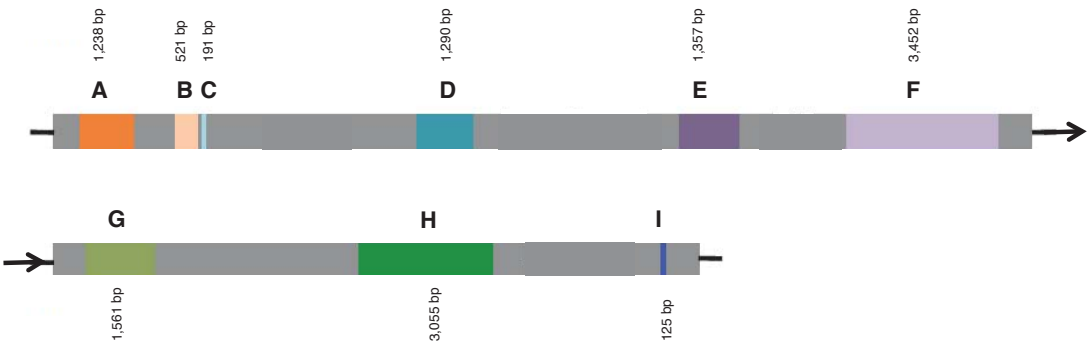
When considering recombination events found in an individual recipient isolate, donors for a single recipient strain could come from a single or multiple genetic backgrounds. For example, isolate SMRU1452 had nine recipient blocks with identical hits detected in eight different clusters, each with a different serotype (**Figure 4.8, Table 4.6**). Eight of the nine recipient blocks (block A, B, C, D, F, G, H, I) were repeatedly detected in one particular cluster, sBC145-sBC152 (serotype 6B), while the remaining block (block E) was only detected in sBC140-sBC143 (serotype NT). Taken together, these observations suggest that the recent ancestor of isolate SMRU1452 had recombined with members of sBC145-152 (serotype 6B) and sBC140-143 (serotype NT), resulting in the import of eight and one DNA region of diversity, respectively.

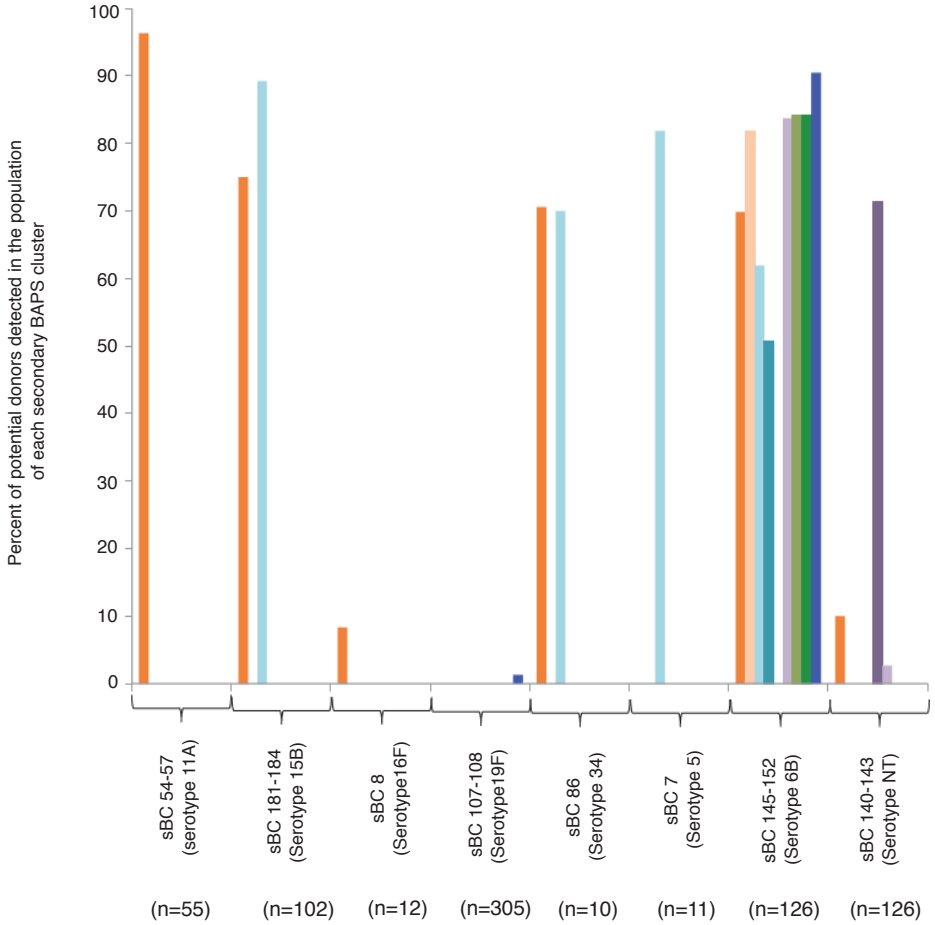
In addition, the shared ancestry of multiple blocks A, B, C, D, E, F, G, H and I possibly suggests that a single transformation event can result in the replacement of DNA at multiple non-adjacent loci of the recipient strain. This observation is consistent with (Hiller, Ahmed *et al.* 2010) where simultaneous replacement of multiple loci following a single transformation event was captured from a clinical strain. Together, these highlight the magnitude of genetic changes potentially introduced by a single transformation event.

Figure 4.8 Multiple potential donors for a single recipient

Top panel: Nine predicted recombination fragments in SMRU1452 (fragments A-I) are highlighted in different colours and are ordered according to their locations on the genome with their size labelled. Bottom panel: the bar chart presents the possible sources of each recombination fragment based on the above colour scheme. The y-axis gives the proportion of hits detected per population of particular lineage. For example, recombination fragment A of 1,236 bp in length was found to have identical matches in 96.29%, 75%, 70.59%, 69.84%, 10% and 8.33 % of the population of secondary BAPS clusters of serotype 11A, 15B, 34, NT, and 16F, respectively. The number of isolates in each secondary cluster was given in the parentheses. (**Figure is shown on the next page**)

Size and relative position of each recombinant blcoks





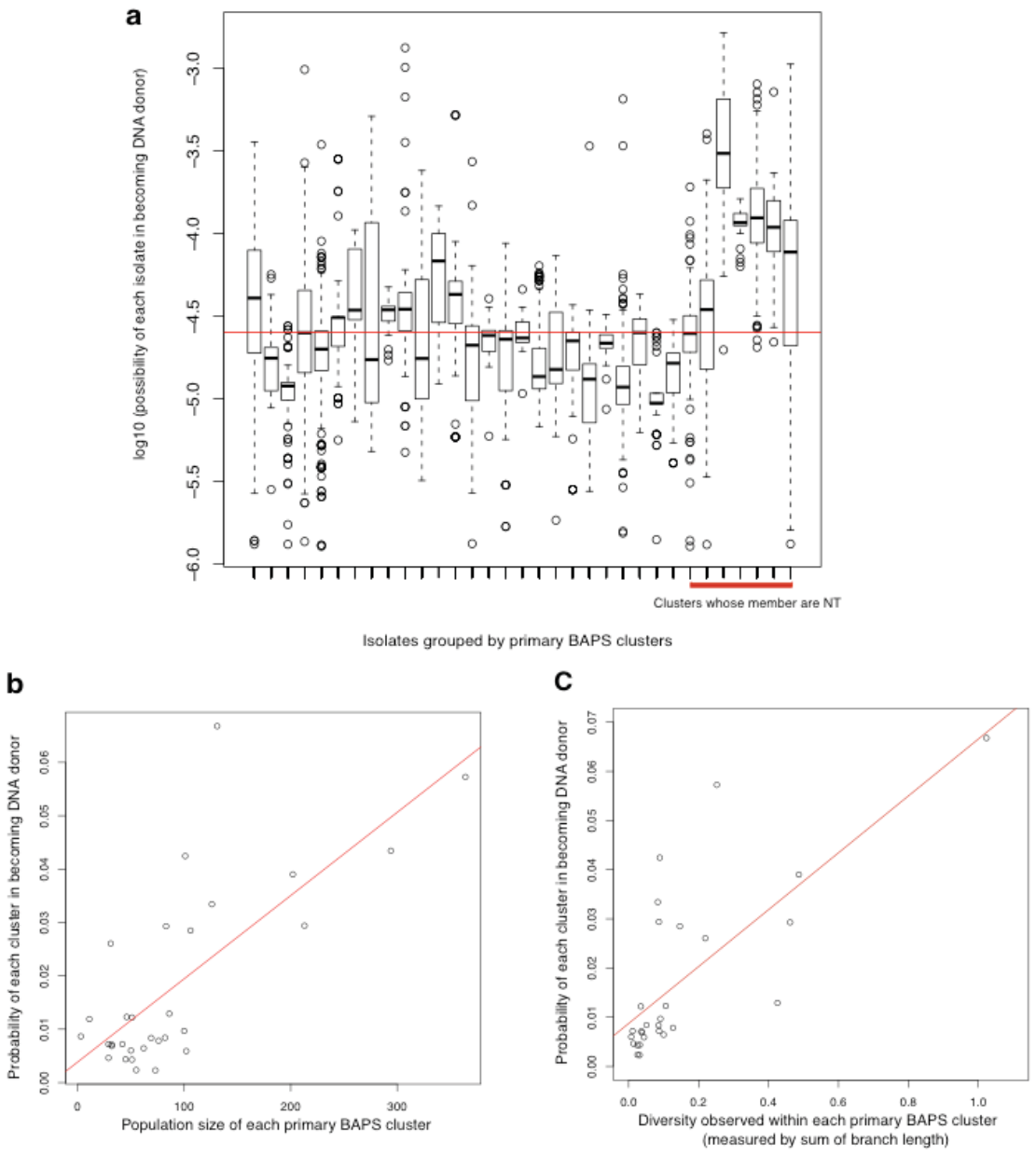
4.3.2.2.2 Probability of a single isolate acting as a donor

The probability of each isolate being a donor was calculated. For each recipient isolate where "n" potential donor isolates were identified, a probability of "1/n" of having been the donor was assigned to each potential donor isolate. Isolates showing no hit for a particular search were assigned a probability of 0. The total likelihood of each isolate for being a donor was represented by the sum of the above probabilities from all donation events. This gave a mean frequency of 2.53×10^{-5} donation events per isolate i.e. each isolate has a probability of 1/39,537 to donate its DNA in recombination events detected here.

The probabilities from individual isolates were then grouped into lineages based on their primary BAPS clusters. The boxplots (**Figure 4.9 a**) showed the distribution of probabilities of isolates within that cluster acting as a donor. Heterogeneity in the donation frequency between each cluster within the population was observed. There was a higher probability of NT isolates acting as the donor than the rest of the population (Mann-Whitney U test between NT isolates and other clusters, p value $< 2.2 \times 10^{-16}$). However, identical matches could come from shared recipient of recombinant fragments as well as true donors. As the two events cannot be distinguished based on the sequenced isolate alone, this finding should be interpreted with caution. As shown earlier, NT isolates are efficient recipients with higher acquisition of recombinant DNA. Therefore the results presented here may be confounded by NT isolates being co-recipients as well as recombination donors.

Figure 4.9 Trends in genetic exchange

(a) Boxplots represent distribution of donation probability of isolates within each cluster. A red bar represents a mean frequency of donation event of any isolates ($2.53x10^{-5}$), i.e. each isolate has a probability of 1/39,537 to donate DNA in a recombination event. (b) and (c) respectively show positive correlations between potential donor clusters (based on primary BAPS clusters) and outer population size, and separately cluster diversity.



4.3.2.2.3 Probability of a cluster acting as a donor and its relationship with cluster size and diversity

To reduce the number of detections from non-donor isolates, the probability of being the donor was considered as a cluster. Cluster filters were applied to remove random hits from non-related clusters, particularly the co-recipients that potentially confounded the result. Either i) clusters most commonly detected as sources for each recipient isolate or ii) clusters detected as the sole source of recombinant DNA in each recipient isolate, were allowed. To demonstrate, a case where multiple potential donor clusters were identified from nine recipient blocks of a single recipient isolate SMRU1452 (discussed in 4.2.2.1) was used here. Tabulated in **Table 4.6**, a hit to each cluster was scored as "1" while no hit was scored "0". The most common sources came from sBC145-152 and this was selected based on the first criterion. Also, sole-source donors were detected in sBC145-152 (sole sources for recipient blocks B, D, G, and H) and sBC140-143 (sole source for recipient block E). These clusters were chosen according to the second criterion. This effectively reduced the number of potential donor clusters from eight to two confident clusters.

For a search matching "n" potential donor clusters for each recipient block, each identified cluster was assigned a probability of "1/n" of being the donor. Clusters that did not contain any potential donors were given a probability of 0. The total probability of each cluster acting as donor is presented as the sum of above probabilities.

The probability of being recombination donors was associated with two other characters detected in each cluster: the cluster population size, and the genetic diversity within each cluster. The population size represented the number of isolates detected in a particular cluster. Plotting the probability of a cluster being a donor against the cluster size gave a positive correlation ($\rho = 0.592$, p-value = 2.69×10^{-4}, **Figure 4.9 b**). Another feature, the diversity within a cluster was calculated as the sum of total branch length, which is proportional to the number of polymorphic sites observed in a particular cluster. A positive correlation was also observed between the probability of a cluster being a donor and the cluster diversity ($\rho = 0.773$, p-value =

1.45 x 10^{-6}, **Figure 4.9 c**). This is consistent with the concept that higher diversity gives a greater chance to discover minority instances (Wang, Yao *et al.* 2009), some of which were captured as the donor for recombinant fragments.

This result suggests that the likelihood for each cluster of being a recombination donor increases with the cluster population size and the cluster diversity. Both features were observed in the clusters of NT isolates. Therefore, NTs could potentially be major donors, contributing their DNA more frequently to the population gene pool.

Table 4.6 Potential donors for each recombinant fragment detected in isolate SMRU1452.

		Recipient blocks								
		A	B	C	D	E	F	G	H	I
Donor clusters	sBC 54-57	1	0	0	0	0	0	0	0	0
	sBC 181-184	1	0	1	0	0	0	0	0	0
	sBC8	1	0	0	0	0	0	0	0	0
	sBC107-108	0	0	0	0	0	0	0	0	1
	sBC 86	1	0	1	0	0	0	0	0	0
	sBC 7	0	0	1	0	0	0	0	0	0
	sBC 145-152	1	1	1	1	0	1	1	1	1
	sBC 140-143	1	0	0	0	1	1	0	0	0

4.4 Conclusion

This chapter summarises the evolution and genetic exchange observed in a densely sampled pneumococcal carriage population. A high sampling density of 3,085 isolates collected over a 3-year period in a refugee camp allowed comparisons of evolutionary rates within the population as well as identification of the source and sink of sequence exchange. Heterogeneity in rates of recombination for both donation and receipt of DNA suggests a structure to genetic flux within the population.

The high rate of receipt of recombination in NT pneumococci is consistent with that observed in NT lineages from some other species (Connor, Corander *et al.* 2012). This is consistent with the general concept that capsule might act as a physical barrier for DNA uptake. The higher rates of both receipt and donation of recombinant fragments observed here in NTs suggest that these clusters might function as hubs of gene flow in the Maela pneumococcal population. Though an increased recombination rate could bring transient benefit, there are potential long-term disadvantages due to increasing genomic instability (Giraud, Matic *et al.* 2001). So it is notable that sporadic switches between the NT and encapsulated states (discussed in chapter 3) may serve as a mechanism to modulate the trade-off between benefit and cost of increased recombination rates.

This chapter introduced the key players of genetic exchange in the Maela pneumococcal population. The next chapter will explore the genes that have been exchanged in relation to selection pressure, particularly the high consumption of antibiotics observed in this refugee camp.