5 *In vitro* transformation of *S. pneumoniae* ATCC 700669

5.1 Introduction

During the 1960s, independent studies were conducted in which mutations causing resistance to erythromycin (Iyer and Ravin, 1962), aminopterin (Ephrussi-Taylor *et al*., 1965), streptomycin (Chen and Ravin, 1966) or amethopterin (Sirotnak *et al*., 1969), or markers that overcame auxotrophic mutations preventing catabolism of maltose (Lacks, 1966) or biosynthesis of uracil (Morse and Lerman, 1969), were transferred between pneumococci. All of these studies concluded that independent mutations giving the same phenotype were transferred at reproducibly different rates, allowing markers to be categorized according to their transformation efficiency (Ephrussi-Taylor *et al*., 1965; Lacks, 1966; Sirotnak *et al*., 1969). The observation that markers induced by specific mutagens tended to transfer with similar efficiencies suggested that the disparity in transformation rates reflected the ease with which different types of mutation were transferred.

Following the advent of DNA sequencing, contemporaneous work studying the aminopterin resistance (*amiA*) and amylomaltase loci identified transversion mutations A•T↔C•G and C•G↔G•C as markers transferred with a high efficiency, the transversion A•T↔T•A having an intermediate efficiency, while transitions acted as low efficiency markers (Claverys *et al*., 1981; Lacks *et al*., 1982; Claverys *et al*., 1983), although some inconsistencies, ascribed to neighbouring sequence context, were identified. Deletions 3 bp or shorter also acted as low efficiency markers (Lacks *et al*., 1982; Gasc and Sicard, 1986; Gasc *et al*., 1987), while those 5 bp or longer were transferred as ‘very high efficiency markers’ (Claverys *et al*., 1981), although there is some evidence that this property is somewhat tempered as the deletion increases in length (Lacks, 1966; Claverys *et al*., 1980; Claverys *et al*., 1981; Lacks *et al*., 1982; Claverys *et al*., 1983; Gasc *et al*., 1987). Somewhat contradictory data indicating both insertions and deletions increase the rate with which flanking markers are transferred through transformation have also been reported (Lefevre *et al*., 1989; Pasta and Sicard, 1996).

Early in the study of transformation, a 'high efficiency integration' (*hex*) phenotype was recognized in recipient cells that no longer discriminated between low and high efficiency markers (Lacks, 1970), although deletions of the appropriate length remained very high efficiency markers in both backgrounds (Claverys *et al*., 1980). Strains exhibiting the *hex* phenotype were found to have an elevated spontaneous mutation rate, suggesting they had lost a DNA repair system (Tiraby and Fox, 1973). This function was found to be encoded by two genes, *hexA* (Balganesh and Lacks, 1985) and *hexB* (Prats *et al*., 1985), which were subsequently identified as being homologous with the mismatch repair (MMR) system in *Escherchia coli* and eukaryotes (Haber *et al*., 1988; Priebe *et al*., 1988; Mankovich *et al*., 1989; Prudhomme *et al*., 1989). Therefore the difference in transfer efficiency between base substitution markers reflects the rate at which the non-canonical base pairings formed by the invasion of genomic DNA by the acquired ssDNA are corrected by the mismatch repair system.

However, problems arise when trying to extrapolate from these relatively simple scenarios, the transfer of small numbers of selectable polymorphisms at specific loci, to understanding the frequent exchanges of sequence among the natural pneumococcal population. The interactions between small numbers of markers with different transformation efficiencies was found to be very complex (Gasc *et al*., 1989), and if correction of each mismatch within a recombination were as efficient as observed for individual polymorphisms many recombinations observed to occur *in vitro* and *in vivo* would be impossible (Lacks, 1966; Majewski *et al*., 2000). This appears to be a consequence of the saturation of the mismatch repair system by a relatively small number of polymorphisms in imported DNA (Humbert *et al*., 1995).

Nevertheless, a linear relationship between the mean level of sequence divergence and the logarithm of the frequency of recombination events is observed, implying polymorphisms do constitute a significant barrier to the exchange of sequence between bacteria (Roberts and Cohan, 1993; Vulic *et al*., 1997; Majewski *et al*., 2000). This has been suggested to be the consequence of the requirement for a minimum threshold length of perfect sequence identity (a 'Minimal Efficiently Processed Segment', or MEPS) at each end of a recombination to allow a strand

exchange to occur (Majewski and Cohan, 1998). Based on the changing frequency of transfer with donor sequences of different levels of divergence from the recipient, the minimum summed length of the two MEPS flanking *S. pneumoniae* recombinations was estimated to be 27 bp (Majewski *et al*., 2000). In order to test how recombination events could occur with such constraints, yet allow the diversification observed in the PMEN1 isolates, an *in vitro* transformation of *S. pneumoniae* ATCC 700669 was performed.

5.2 Analysis of *in vitro* transformants

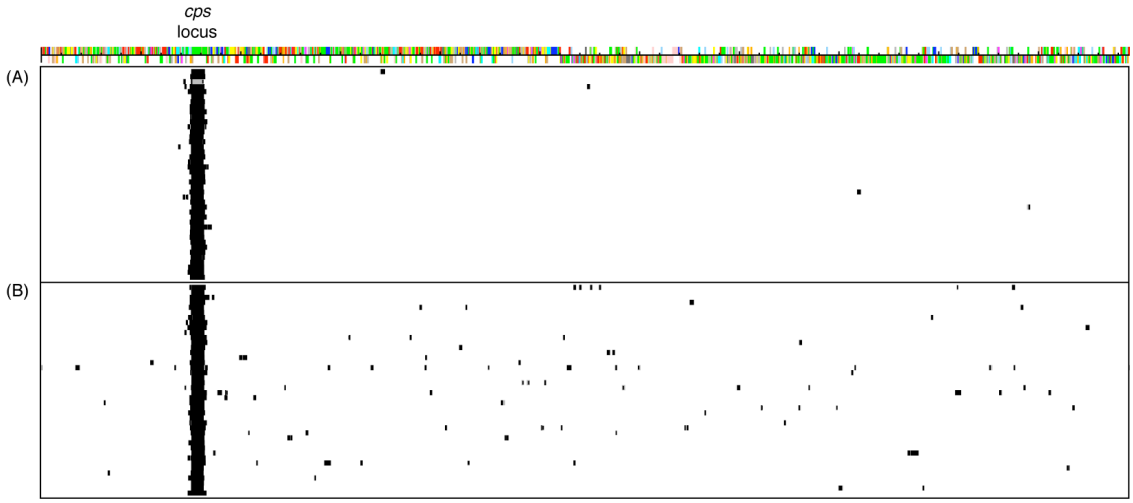

Figure 5.1 Distribution of recombination events across the genome. The annotation of the 2,182,008 bp *S. pneumoniae* ATCC 700669^{lab} genome is displayed across the top, coloured as described in Chapter 3. The *cps* locus, spanned by the selected recombination in transformants, is labelled. Underneath, each row represents an individual sequenced transformant. RSSs are indicated on the rows by black blocks, representing the donor sequence, flanked by grey FRs, which show the ambiguous boundaries between RSSs and the flanking polymorphisms with recipient alleles. The transformants are grouped as (A) *S. pneumoniae* ATCC 700669^{lab} transformed at 2 ng mL^{-1} donor DNA and (B) *S. pneumoniae* ATCC 700669^{lab} transformed at 200 ng mL^{-1} donor DNA.

5.2.1 Genome-wide exchange of sequence between pneumococci

The precise isolate of *S. pneumoniae* ATCC 700669 used in the experiment is hereafter designated *S. pneumoniae* ATCC 700669^{lab} (see Material and Methods). This was transformed with genomic DNA from a rough derivative of *S. pneumoniae* TIGR4 that carries a kanamycin resistance marker at the capsule biosynthesis (*cps*) locus (Pearce *et al*., 2002). Multiple transformations were performed using a concentration of either 2 ng mL^{-1} or 200 ng mL^{-1} of donor genomic DNA. Recombinations affecting the *cps* locus were detected either through selection with

kanamycin alone, or kanamycin supplemented with penicillin. This latter condition was used to test the hypothesis that the transfer of *cps* loci from penicillin-sensitive pneumococci, such as TIGR4Δ*cps*, to penicillin resistant strains, such as ATCC 700669, may be inhibited by selection against any co-transfer of the flanking antibiotic-sensitive penicillin-binding protein gene alleles.

With selection on just kanamycin, the transformation with the lower concentration of DNA produced 38 fold fewer transformants (Wilcoxon rank sum test, $p = 1.5\text{x}10^{-4}$), suggesting that the availability of the marker was limiting the rate of transformation. Dual selection with penicillin as well caused a small, but non-significant, decrease in transformation rates: with 2 ng mL^{-1} DNA, a 12.6% decrease was observed (Wilcoxon rank sum test, $p = 0.53$), while with 200 ng mL^{-1}, there was a fall of 9.6% (Wilcoxon rank sum test, $p = 0.21$). Therefore, selection for β-lactam resistance does not appear to significantly inhibit exchange with penicillin sensitive lineages at the *cps* locus, instead suggesting a strong limitation on the size of recombination events reducing the impact of linkage between genes. To test this hypothesis, 21 isolates from each of the four examined conditions (low and high DNA concentration, and with and without penicillin selection) were sequenced using the Illumina platform.

Alignment of the complete genome sequences of the donor and recipient strains identified 21,512 base substitutions, 476 insertions in TIGR4Δ*cps* relative to ATCC 700669^{lab} (1 bp – 14,153 bp in size) and 578 insertions in ATCC 700669^{lab} (1 bp – 76,827 bp in size). By mapping Illumina reads simulated from the donor sequence to that of the recipient, it was possible to identify 16,067 base substitutions, all but 15 of which were also found through whole genome alignment. Sequence data from the 84 transformants identified 2,347 polymorphic sites, of which just 67 did not correspond to polymorphisms transferred from the donor. Eleven of these sites represent difficulties with mapping; a further nine appear to have arisen through intragenomic recombinations affecting an IS element and the repetitive surface protein gene *pclA*. The remaining 47 sites appear to be spontaneous mutations occurring *in vitro*; 25 of these are C•G→T•A substitutions likely to represent the consequences of cytosine oxidation and deamination (Kreutzer and Essigmann, 1998), which may be caused by

the high levels of hydrogen peroxide produced by *S. pneumoniae* during aerobic growth (Pericone *et al*., 2000).

Recombinant sequence segments (RSSs) were detectable in transformant sequence data as loci containing donor alleles at polymorphic sites, defining the minimum size of the recombination, bounded by recipient alleles at the flanking polymorphic sites, demarcating the maximum size of the exchange (Figure 5.1). The actual length may be estimated as being the median (L50) between these two limits, positioning the boundary half way through each flanking region (FR), expressed as a distance relative to the donor ($L50_D$) or recipient ($L50_R$) genome. The selected recombination at the *cps* locus was detected in all transformants, with at least one of the 84 isolates having recombinant sequence between the genomic loci 294,349 bp and 340,522 bp; this region of the chromosome is henceforth referred to as the 'primary locus', as these recombinations have been driven by selection. Furthermore, 112 unselected, 'secondary' recombinations were observed outside the *cps* locus, with one strain having a total of 18 RSSs. The mean proportion of the recipient genome found to have undergone recombination was 1.4%, ranging up to a maximum of 2.3%. Secondary recombinations were significantly more common in the strains transformed at a high concentration of DNA (mean of 2.48 secondary events per strain) than at a low concentration (mean of 0.26 secondary events per strain; Wilcoxon rank sum test, $p = 2.0 \times 10^{-8}$). Hence the effective concentration of DNA available for recombination inside the cell can vary. This implies that recombination events involving separate DNA strands can occur within the same cell concurrently and independently, rather than all arising from the import of a single large molecule of DNA, as has been observed in *S. agalactiae* (Brochet *et al*., 2008) and inferred from *C. difficile* genome sequences (He *et al*., 2010).

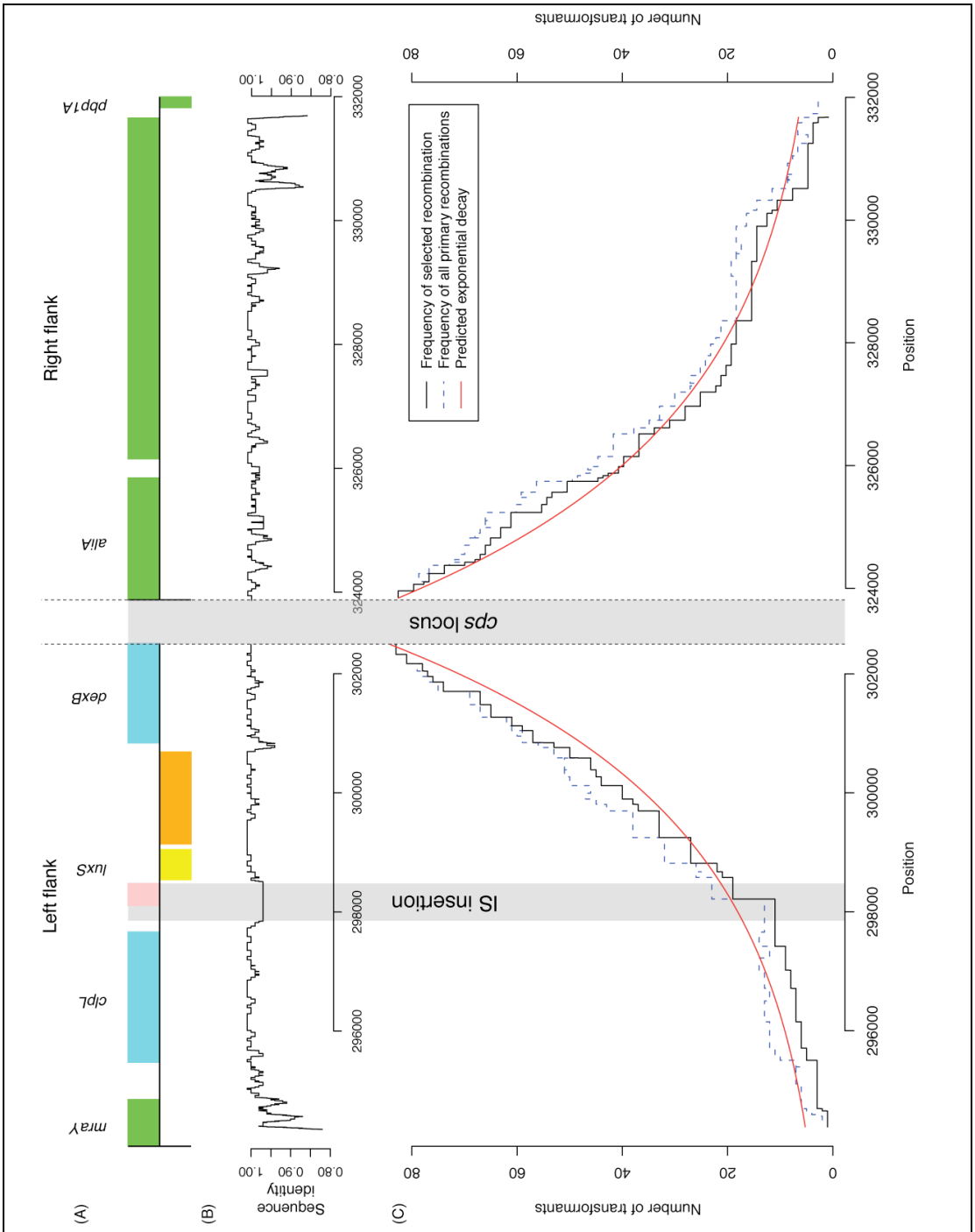

Figure 5.2 Structure of recombinations within the primary locus. (A) Annotation of two sections immediately flanking the *cps* locus, within the primary locus. The position of the *cps* locus, lying between the two flanks, is marked by the grey bar. Similarly, the extent of an IS element insertion in *S. pneumoniae* ATCC 700669[lab], not present in *S. pneumoniae* TIGR4**Δ***cps*, is marked by a grey bar within the left flank. (B) A plot of sequence identity showing the level of similarity between the donor and recipient throughout the locus. The value on the graph represents the proportion of the aligned flanking 100 bp that is identical in both strains. (C) Distribution of recombination sizes. At the *cps* locus itself, all 84 isolates have the donor allele. The black line traces the rate at which this selected recombination ends relative to the extent of the locus. This process can be modelled as an exponential decay, which is represented by the red line for each flank. Other recombinations occur within the primary locus, and the overall prevalence of these, along with the selected recombination, is indicated by the blue dashed line. This shows the total number of isolates in the recombinant state throughout the displayed locus.

5.2.2 Characterisation of 'capsule switching' recombinations

The high density of SNPs within the primary locus allows a high-resolution view of the boundaries of the selected RSSs that span the *cps* locus. For each distance from the edge of the selected *cps* locus, on the two sides independently, the number of isolates with a selected RSS extending to, or beyond, that point can be plotted (Figure 5.2C). This outlines the rate at which the primary RSSs end relative to their distance from the selected marker, best modelled as an exponential decay, with similar estimates of the decay constant on both sides: 3.42×10^{-4} (95% confidence interval, $3.41\text{-}3.43 \times 10^{-4}$) on the left flank, and 3.40×10^{-4} (95% confidence interval, $3.39\text{-}3.41 \times 10^{-4}$). The symmetrical nature of the decays on the two flanks is not disrupted by the presence of an IS element insertion in the recipient distinguishing it from the donor. Correspondingly, the exponential decay on the left hand side fits the position relative to the recipient (residual sum-of-squares, 172,966) better than that relative to the donor (residual sum-of-squares, 264,836). Such observations imply that the size of the RSSs is dictated by a Poisson process that involves the recipient's DNA, hence may represent a process such as resolution of the heteroduplex, rather than events during the pre-processing of the donor strand, such as endonucleolytic cleavage during DNA import.

The symmetry is also in spite of the low level of correlation in terms of sequence identity between the donor and recipient between these two regions, either side of the *cps* locus (Pearson correlation, $R^2 = 0.0011$), suggesting the density of SNPs observed in this region is not enough to significantly affect the distribution of recombination events (Figure 5.2B). In the absence of sequence identity affecting the exponential declines, the number of isolates in the recombinant state should halve over each 2 kb stretch of sequence. On the basis of this rate, of the recombinations that directly affect the *cps* locus, 7.4% will affect *pbp1A* and 5.8% will affect *pbpX*. Less than 0.1% of recombinations encompassing the *cps* locus would be expected to replace both *pbpX* and *pbp1A* in their entirety, explaining the lack of a significant inhibition of *cps* transfer by ampicillin selection.

5.2.3 Mosaic recombinations in the *cps* locus

Thirty-six further RSSs occur in the primary locus but do not span the *cps* gene cluster (Figure 5.2C). This high density of unselected primary recombinations suggests that they are associated with that spanning the *cps* locus. Supporting this hypothesis, strains transformed with the lower concentration of DNA actually have a larger mean number of primary RSSs (1.52 per strain) than those exposed to the higher concentration (1.33 per strain), although this difference is not significant (Wilcoxon rank sum test, p = 0.33). This suggests that the frequency of recombinations in close proximity to the selected event is independent of the external DNA concentration, indicating that these mosaic recombinations (MRs) are not a consequence of the locus acting as a hotspot for integrations by several imported strands, but rather reflects multiple RSSs originating from the same piece of donor DNA.

A histogram of the $L50_R$ of the recombination events within the primary locus reveals a bimodal distribution (Figure 5.3A). While the selected recombinations spanning the *cps* locus, 21,373 bp long in *S. pneumoniae* ATCC 700669^{lab}, have a modal $L50_R$ of around 27 kb, the nearby flanking events are mainly 5 kb or less in size. The shapes of the two distributions are also distinct. The smaller events form an approximate exponential distribution, supporting the suggestion that RSSs are generated through a Poisson process with a per base probability of strand exchange, λ_R, of 5.0×10^{-4} bp^{-1}. The theoretical mean length, λ_R^{-1}, is therefore 2 kb. By contrast, the lengths of the events that span the *cps* locus are not exponentially distributed. This is an artefact of selection; although the longer events are less frequent, they are more likely to span the selectable marker, and hence are observed at an unusually high frequency relative to shorter events when compared with unselected recombinations.

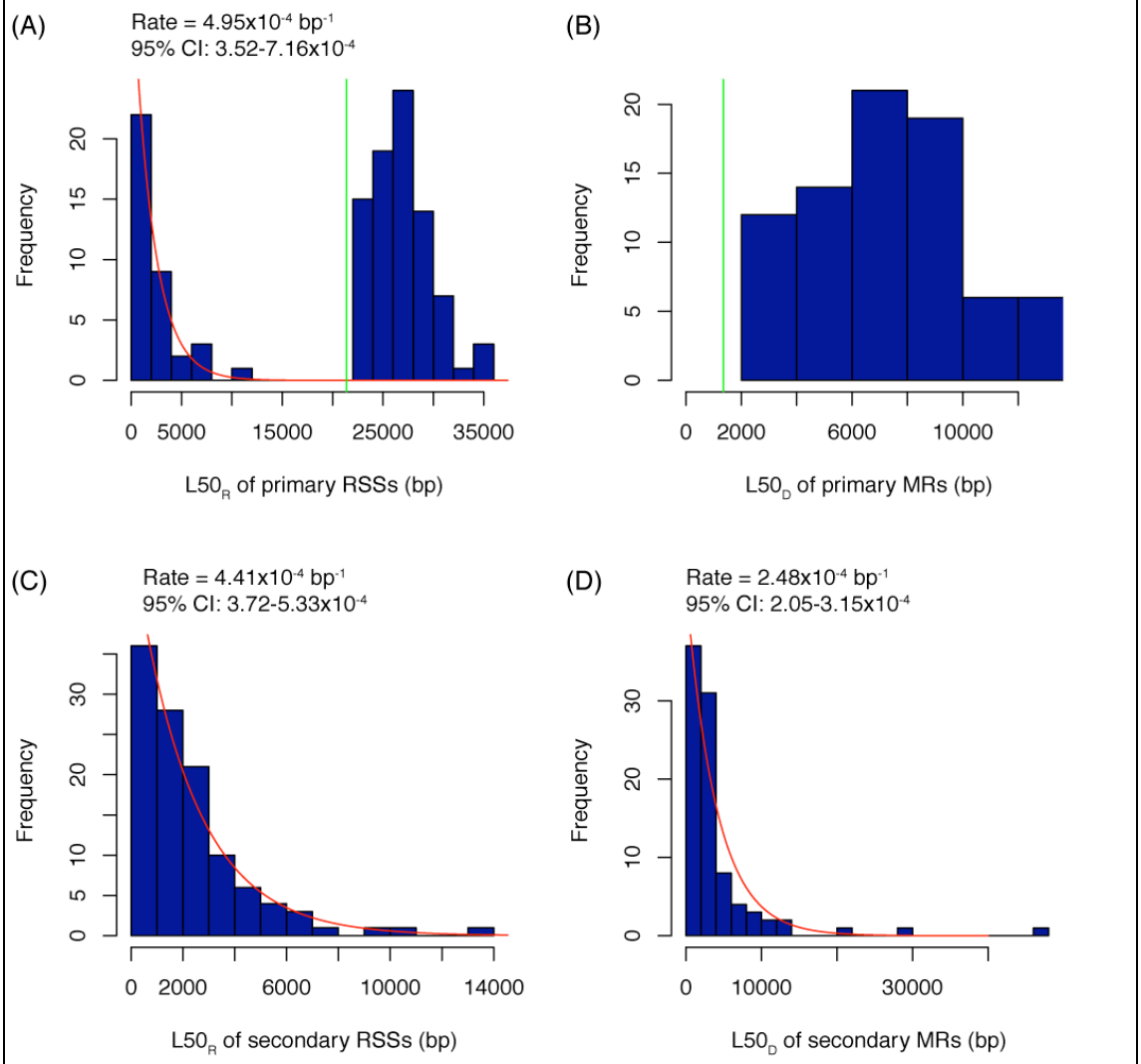

Figure 5.3 Distribution of recombination sizes. (A) Histogram showing the distribution of RSS sizes as $L50_R$ lengths. The vertical green line shows the length of the *cps* locus relative to the recipient (21,372 bp). The smaller, unselected recombinations are modelled as an exponential distribution (red line), the rate parameter of which is displayed. (B) Histogram showing the distribution of donor strand lengths participating in primary recombinations, as estimated using the $L50_D$ lengths of the MR boundaries. The vertical green line shows the length of the *cps* locus relative to the donor (1,354 bp). (C) Histogram showing the distribution of secondary RSS sizes as $L50_R$ lengths. These are modelled as an exponential distribution, indicated by the red line. (D) Histogram estimating the sizes of donor strands participating in secondary recombinations, showing the distribution of MR $L50_R$ lengths. These are modelled as an exponential distribution, indicated by the red line.

As all the primary RSSs in each strain, forming a single MR, originate from the same molecule of DNA, the length of the donor strand participating in the recombinations around the primary locus can be estimated (Figure 5.3B). The median $L50_D$ of these measurements is 7.3 kb, with no significant differences between strains transformed with high or low levels of DNA were detected (Wilcoxon rank sum test, $p = 0.93$). This reflects the consequences of selection at this locus: donor strands smaller than

the kanamycin resistance gene evidently will not be observed, and furthermore the selection process will bias the range towards larger MRs. Despite these constraints, the $L50_D$ is still comparable in length to the median size of the DNA strands imported inside the cell (~6.6 kb) (Morrison and Guild, 1972). Within the primary locus boundaries, the mosaic pattern of transfer could arise either through multiple exchanges involving the same donor strand, which may be enforced by cleavage of the original molecule into several pieces on entry, or through localised action of repair processes acting on a single, larger transfer.

5.2.4 Analysis of secondary recombinations

The lengths of the secondary RSSs are exponentially distributed with a λ_R of $4.4x10^{-4}$ bp and a median of 2.3 kb (Figure 5.3C), very similar to the unselected RSSs within the primary locus but contrasting with the size and distributions of the selected recombination events. They also exhibited a similar pattern of mosaicism to those at the primary locus. A bootstrapping algorithm (see Materials and Methods) was used to organise the 112 secondary RRSs into 90 MRs, each likely to have been derived from a single donor strand. All RSSs less than 8 kp apart were linked into MRs; although the majority of MRs consisted of only one RSS, up to four could be found in significantly close proximity. The stretches of unmodified recipient sequence between linked RSSs were sometimes only identifiable by a single SNP, although they usually contained multiple polymorphic sites; their median length was 509 bp (mean length of 2.5 kb). The most distant sequences joined were 43.6 kb apart; this occurred in a strain where the only three secondary RSSs all fell within a 45 kb region of the genome. Unlike those at the *cps* locus, MRs themselves were exponentially distributed like their RSS components (Figure 5.3D).

The mean sequence divergence across the $L50_R$ of detected RSSs was 0.9%, which falls to 0.7% when considering the $L50_R$ of the entire MRs. Hence the heterogenous pattern of SNP density observed occurring *in vivo* appears to represent an intrinsic property of the mechanism of pneumococcal transformation. However, such estimates are based only on RSSs that can be detected through transfer of SNPs. Using a sliding window analysis (see Materials and Methods), it can be estimated that a quarter of the total number of recombinations outside the primary locus are not detectable. These

are typically short events, which would decrease the overall median length estimate and SNP density values. However, it may be that an even greater proportion is undetectable, if the presence of a single polymorphism is enough to significantly inhibit transfer. Although no such effect was observable within the primary locus, this problem can also be investigated using the detected secondary recombinations.

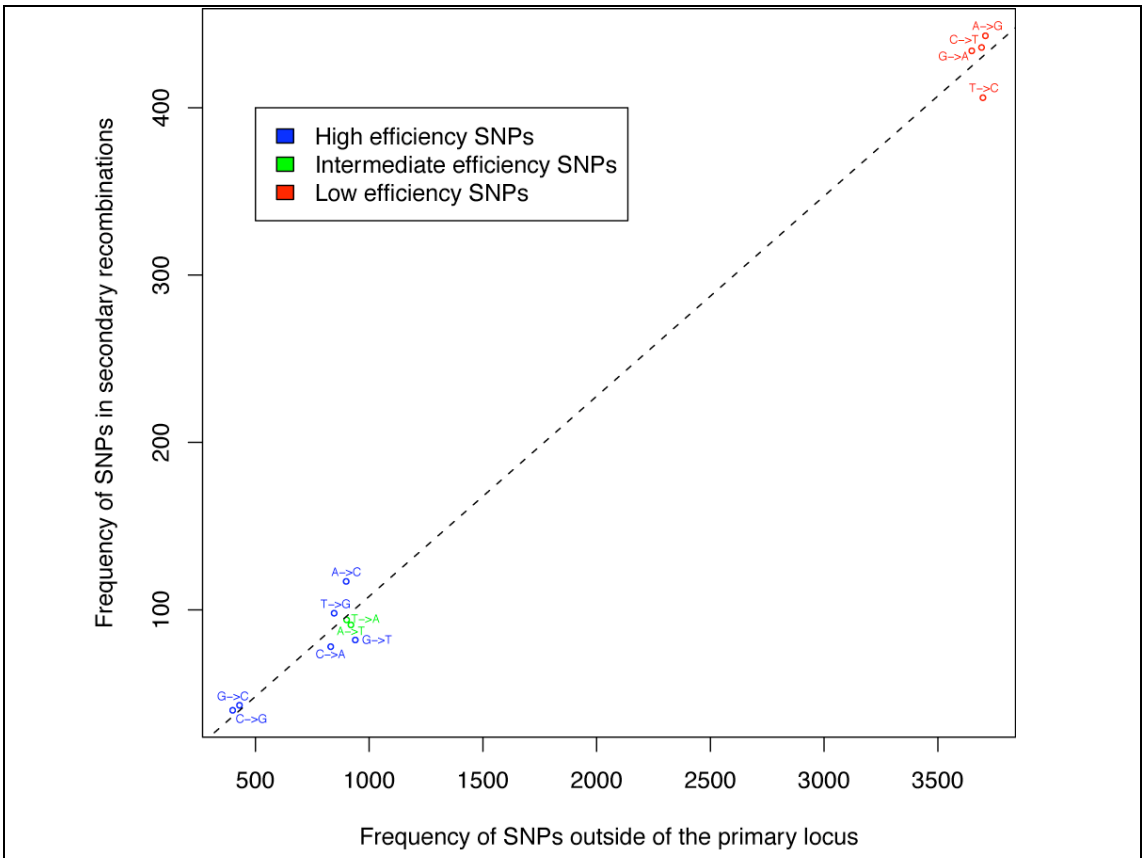

Figure 5.4 Correlation of SNP type frequencies outside the primary locus with those detected within secondary recombinations. Points are coloured according to the efficiency with which they are transferred through transformation when mismatch repair is active. This shows that SNPs repaired effectively by mismatch repair are the most common, and are represented in secondary recombinations at the same level as those SNPs repaired less effectively.

The sliding window analysis found that 62-70% of observed recombinations had a lower mean level of sequence diversity than expected from the distribution of SNPs between the donor and recipient, with the result varying as the length of surrounding sequence considered in calculating the sequence identity was changed between 100 bp and 2 kb. However, this deviation was not significant at any of the tested surrounding sequence lengths (Fisher's exact test, p = 0.080-0.11). Hence no enrichment of RSSs in regions of high sequence similarity can be observed. This may reflect the only constraints, in terms of sequence similarity, being the length of the MEPS. In this

dataset, only one RSS contravenes the suggested 27 bp threshold (Majewski *et al*., 2000); this event transfers a single SNP and has a total flanking identity length of 26 bp. Despite largely conforming to this condition, the population of FRs were not found to be significantly longer than the population of distances between SNPs: the sliding window analysis found half of the 5' and 3' MEPS were larger than the expected length given the $L50_R$ of the RSS (Fisher's exact test, $p = 1.0$). Hence it appears that sequence diversity causes few limitations on the exchange of sequences between *S. pneumoniae* genotypes.

5.2.5 Efficiency of polymorphism transfer

As just a few hundred SNPs appear to be sufficient to overwhelm the pneumococcal MMR system, it is not expected that this form of repair is likely to have impacted on recombinations to a great extent. Correspondingly, the number of each type of SNP outside the primary locus correlates tightly with the frequencies of these mutations in the secondary recombinations ($R^2 = 0.99$, $p = 1.5 \times 10^{-12}$; Figure 5.4). Furthermore, the frequency of each SNP on the outermost position of each RSS is proportional to its prevalence in the nearest flanking unchanged position ($R^2 = 1.0$, $p < 2.2 \times 10^{-16}$). Hence there is no evidence that the low efficiency markers lead to entire recombinations being lost at a higher frequency, or that they trigger localised repair, which might have been a mechanism for the formation of MRs. The alignment of the donor and recipient does show, however, that the most frequent mutations distinguishing them are the transversions, supporting the hypothesis that MMR has evolved to repair the most common mutations most efficiently (Figure 5.4).

Past work has shown that the efficiency with which indels are transferred appears to depend upon their length. The density of deletions in the donor, relative to the recipient, within the secondary recombinations is not significantly different to that outside such regions (Table 5.1), suggesting they have no effect on recombination rates. However, insertions in the donor relative to the recipient are significantly excluded from secondary recombinations (Table 5.1). The mechanism behind this inhibition may be inferred by comparing the sizes of the insertions found in the recombinations; while there is no significant difference in length between the insertions in the recipient within and outside secondary recombinations (Wilcoxon

rank sum test, p = 0.97), the transferred insertions in the donor are significantly smaller than expected (Wilcoxon rank sum test, p = 0.018). This is despite there being no statistically significant difference between the distributions of indels distinguishing the two strains (Wilcoxon rank sum test, p = 0.94). Hence it appears that constraints on the size of the donor strand, likely resulting from the cleavage of the DNA molecule as it is imported, inhibits the acquisition of large insertions in the donor sequence.

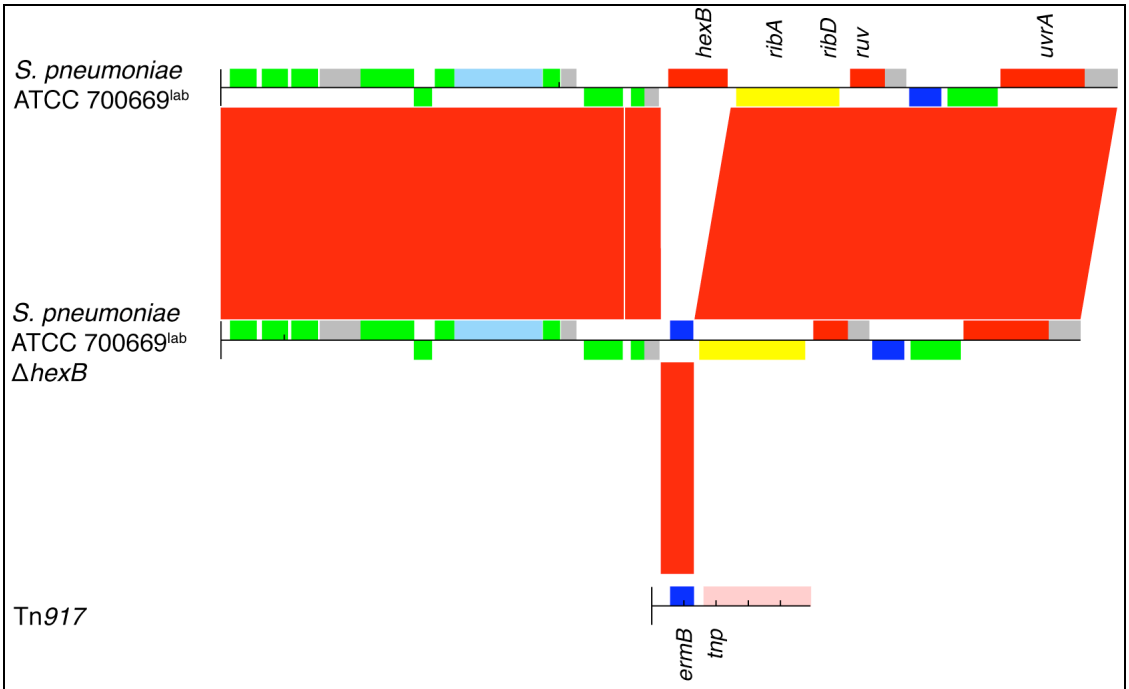

Figure 5.5 Knock out of the *hexB* gene in *S. pneumoniae* ATCC 700669^{lab}Δ*hexB*. This shows an ACT comparison (as described in Figure 3.2) of the *hexB* locus of the *S. pneumoniae* ATCC 700669^{lab} reference strain compared against the orthologous region of an MMR-defective transformant, which has an *ermB* erythromycin resistance in place of *hexB*. The selectable marker was derived from *Tn*917, shown below.

The density of indels in FRs was examined to test the hypothesis that they are capable of stimulating the transfer of adjacent sequence, even when they themselves are not transferred. Both insertions and deletions in the donor were found to be excluded from the FRs, although the result was only significant for deletions (Table 5.1). However, for both types of indels, those in the FRs were significantly larger than expected (Wilcoxon rank sum test, p = 0.00029 for deletions and 0.030 for insertions), although manual inspection of the short read alignments suggested that very few of these indels were actually transferred as part of the associated RSS. Overall, this implies that small indels, of just a few bases, are particularly strongly

excluded from FRs, likely because they interfere with strand exchange in MEPS, though it remains possible that larger indels may trigger the transfer of neighbouring loci.

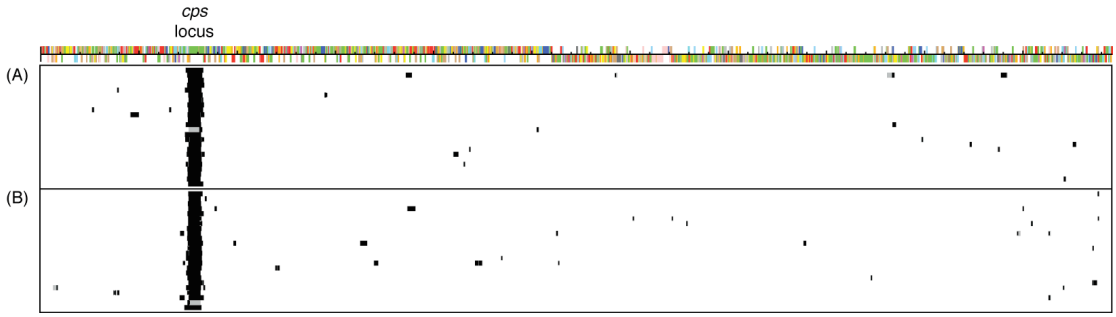

Figure 5.6 Distribution of recombination events across the genome when testing the effect of MMR, displayed as described in Figure 5.1. The transformants are grouped as (A) *S. pneumoniae* ATCC 700669^{lab} transformed at 200 ng mL^{-1} donor DNA and (B) *S. pneumoniae* ATCC 700669^{lab}Δ*hexB* transformed at 200 ng mL^{-1} donor DNA in the second experiment.

5.2.6 The role of mismatch repair

In order to confirm whether mismatch repair played any role in the transformation of *S. pneumoniae* with divergent donor DNA, both *S. pneumoniae* ATCC 700669^{lab} and a mutant with the *hexB* MMR gene knocked out (*S. pneumoniae* ATCC 700669^{lab}Δ*hexB*) were transformed with 200 ng mL^{-1} DNA from *S. pneumoniae* TIGR4Δ*cps*. Twenty-four transformants of each background were then sequenced, confirming the knockout (Figure 5.5) and allowing RSSs to be identified (Figure 5.6). This revealed that the mosaicism within the primary locus, redefined in this experiment as lying between coordinates 293,436 and 333,345 using the same criteria as described previously, was observed in both backgrounds: the wild type isolates had a mean of 1.83 primary RSSs per strain, while the Δ*hexB* isolates had a mean of 1.92 (Wilcoxon rank sum test, $p = 0.57$). Hence the observed mosaicism within MRs does not result from localised repair, but instead appears to represent multiple independent invasions of the recipient genome by the same donor strand.

Correspondingly, the number of secondary RSSs also did not differ significantly, with the Δ*hexB* isolates actually having a slightly higher mean (1.96 per genome, as opposed to 1.04; Wilcoxon rank sum test, $p = 0.12$). Although the median secondary $L50_R$ length was smaller in the MMR defective background (386 bp) than in the wild

type strain (1,453 bp), this difference was not significant (Wilcoxon rank sum test, p = 0.084). Therefore, under the tested conditions, MMR does not appear to be responsible for fully, or partially, repairing any recombination events.

5.3 Discussion

Previous studies of transformation have necessarily focused on rates of transfer of particular selectable mutations. The advantage of high-throughput genome sequencing, coupled with a controlled *in vitro* system, is the ability to characterise multiple recombinations occurring across the chromosome in great detail. Using these data, it can be observed that multiple RSSs, with an exponential size distribution, can be generated from a single donor strand of DNA, thereby forming MRs. Furthermore, multiple MRs, each from a different donor strand, can be generated during a single period of competence.

The heterogenous density of SNPs observed within the MRs corresponds with that observed within the recombination events identified in the PMEN1 isolates. While the median length of homologous recombinations in the PMEN1 population was 2.9 kb, the equivalent measurement from this *in vitro* transformation (the minimum size of MRs relative to the recipient's sequence) is 1.7 kb. However, the overall distribution of sizes from the two studies are similar: the recombination events identified in the clinical isolates follow an approximately exponential distribution with λ_R of about 1.6×10^{-4} bp^{-1} (95% confidence interval of 1.5×10^{-4}-1.7×10^{-4} bp^{-1}). Therefore, rather than the discrepancy representing an overestimation of the lengths of events in the PMEN1 population, this difference results from a lack of sensitivity in identifying small events, due to the conservative nature of the algorithm employed. Comparing the events that span the *cps* locus reveals a more accurate correspondence between the datasets, with the median lengths of those from the *in vitro* data being 27.2 kb, while those from the PMEN1 population have a median of 27.9 kb (Wilcoxon rank sum test, p = 1.0). Hence the method used to reconstruct the history of that lineage appears to have been successful in defining transformation events mediated by the competence system.

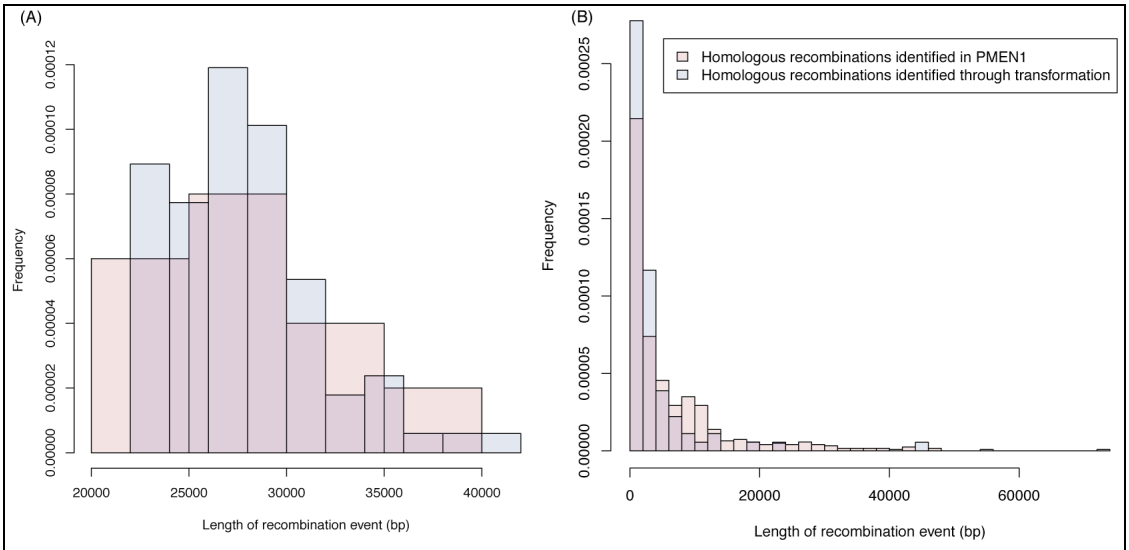

Figure 5.7 Histogram comparing frequencies of detected recombination sizes in PMEN1 isolates and *in vitro* transformants. In both graphs, the blue bars show the data from this experiment, whereas the red bars represent that from the PMEN1 strains. (A) Comparison of the minimum primary MR lengths (*i.e.*, the distance between the outermost polymorphic sites of the donor allele) with the ten serotype switching recombinations characterised in PMEN1. (B) The distribution of secondary MR minimum sizes identified in this experiment with the equivalent size distribution for 615 homologous recombinations (*i.e.* excluding events involving the prophage or conjugative elements) identified as occurring in the PMEN1 isolates.

The distribution of secondary RSSs observed in this study clearly shows that there are few constraints on the exchange of DNA between pneumococci, congruent with the widespread locations and polymorphism densities observed in the PMEN1 population. As discussed, this is likely to be a consequence of the level of divergence between these strains (a mean of one SNP per 101 bp in this experiment). Assuming 27 bp, or more, of identical sequence is required for a strand exchange to occur (Majewski *et al*., 2000), 95% of the aligned length of the recipient genome is capable of participating in strand exchanges. Such a level of divergence is typical between *S. pneumoniae* chromosomes, as comparing all complete pneumococcal genomes to the recipient reveals a minimum and maximum SNP density of one SNP per 150 bp (*S. pneumoniae* JJA) and one SNP per 81 bp (*S. pneumoniae* Hungary 19A-6), respectively. By contrast, using the same approach to compare *S. mitis* B6 (Denapaite *et al*., 2010) with the recipient sequence, just 54% of the sequence would have a sufficiently low SNP density to permit strand exchanges to occur. Hence, the main determinant on the distribution of sequence exchanges between pneumococci across the chromosome is likely to be the random uptake of sequence by the competence system. This implies the observed patterns of recombination frequencies across the *S.*

pneumoniae PMEN1 genome are determined by selection and the level of detectable sequence divergence in the extant population, not by sequence-based constraints limiting the transfer of sequence.

One factor that does appear to constrain the positioning of transformation events is the incidence of indels. The exponential distribution of sizes implies that, following the formation of a boundary at one end, the extent of the RSS is determined by an event that occurs with a fixed per base probability of λ_R as it becomes further removed from the initial site. The impact of indels is at least partly governed by whether λ_R applies to recipient bases, donor bases or aligned bases only. The evidence from the exponential decay of selected recombination boundary positions suggests that λ_R applies to each recipient base. However, this would predict that deletions, but not insertions, would be excluded from RSSs, whereas the reverse is actually observed; furthermore, larger deletions would be anticipated to be excluded to a greater extent, which is also not the case. That the observations indicate only large insertions being excluded from RSSs suggests that the actual situation *in vivo* may be more complicated.

Instead, I propose a model in which λ_R applies only to aligned bases, which interact within the heteroduplex resulting from invasion of the ssDNA strand. It seems likely that non-aligned bases in indels loop out of the heteroduplex, and thereby destabilise it to some extent; hence the drop in selected recombinations spanning the IS element insertion in the recipient, despite there being no aligned bases, and the exclusion of indels from the FRs, where the heteroduplex must be more stable for the process of strand exchange to occur. This destabilisation is unlikely to be simply related to the length of the indel, given the range of deletion sizes found within RSSs in this experiment. The exclusion of large insertions from RSSs would not, therefore, be a consequence of the process governed by λ_R, but instead relate to the cleavage of ssDNA as it is imported into the cell. Further investigation will be necessary to test this model.

The exclusion of insertions from RSSs, and the exponential distribution of RSS sizes, has the consequence that, at all loci, if there is a difference is size between alleles then

the smaller allele will transfer between cells more quickly. Hence, in the absence of selection, the smallest allele at any given locus will drift to fixation in a population, meaning homologous recombination has a reductive effect on genome size. In this characteristic, it opposes site-specific recombination, which leads to the integration of mobile elements into the chromosome. These two opposing forces are likely to shape the evolution of pneumococcal lineages.

Table 5.1 Association of indels with *in vitro* recombination events. Comparing the density of donor indels, outside of the primary locus, observed within RSSs and FRs relative to that in sequence that does not undergo recombination. For both insertions and deletions, Fisher exact tests were performed comparing the number of events within RSSs or FRs with those outside of both, and the number of aligned bases between the donor and recipient with the same category with those outside of both. The resultant *p* values are displayed in the columns on the right.

	Outside recombinations	**Within RSSs**	**Within FRs**	***p* value, within RSS**	***p* values, within FR**
Sequence length (bp)	1,646,830	200,193	101,784	-	-
Donor insertions	509	40	20	0.023	0.050
Donor deletions	409	40	11	0.22	0.0033