4 The evolution of the PMEN1 lineage

4.1 Introduction

4.1.1 Global collection of PMEN1 strains

The complete genome of *S. pneumoniae* ATCC 700669 provides some insight into the evolution of features that are common to almost all members of PMEN1, such as penicillin, chloramphenicol and tetracycline resistance, but offers no information on the clinically important mechanisms of serotype switching and the acquisition of resistance to macrolides and fluoroquinolones. In order to study how this lineage has evolved as it has spread, Illumina sequencing of multiplexed genomic DNA libraries was used to characterise a global collection of 240 PMEN1 strains isolated between 1984 and 2008. These were identified either by using MLST or on the basis of serotype, drug-resistance profile, and targeted polymerase chain reaction (see Materials and Methods). Selected isolates were distributed among Europe (seven countries, 81 strains); South Africa (37 strains); America (six countries, 54 strains); and Asia (eight countries, 68 strains) and included a variety of drug-resistance profiles, as well as five serotypes distinct from the ancestral 23F: namely, 19F, 19A, 6A, 15B, and 3 (Appendix II: PMEN1 strains).

4.1.2 Detecting recombination in sequence data

Algorithms for constructing phylogenies assume that the entire alignment being studied has a single, common ancestry; this condition is violated in recombined sequences (Posada *et al*., 2002). Hence in order to reconstruct the history of a naturally transformable species, it is necessary to distinguish vertically inherited point mutations, which are informative about relationships between taxa, from horizontally acquired sequences, which may introduce many polymorphisms simultaneously but actually represent just a single mutational event. In the context of a phylogeny, recombinations can be defined as either ‘imports’, if the level of sequence divergence between the recombination donor and recipient is large compared to the diversity of taxa within the tree, or ‘exchanges’, if the divergence between donor and recipient is comparable to (or indeed, smaller than) the diversity within the tree. The major

impact of imports is on branch lengths, due to the increased divergence between the recipient and those taxa that retain the ancestral sequence. In addition, if multiple independent imports occur at the same locus in the studied population, then they can lead to distortions in the tree topology. This can either be the consequence of a genuine relationship between the donors in each case, such that independent imports lead to separate groups of taxa appear to look similar to one another through convergent evolution, or simply because two unrelated, but divergent, sequences can spuriously appear homologous due to long branch attraction effects. Exchanges have relatively little impact on branch lengths, unless they involve the transmission of a recent import through a population; instead, they primarily affect the topology of the tree, because distinct segments of the sequence will support different branching patterns, depending their recent history of transfer among the taxa. Hence phylogenies of populations in which recombinations, in particular exchanges, occur frequently will include a high level of homoplasy.

Early statistical tests for detecting recombination generally focussed on detecting exchanges through identifying patterns of polymorphisms in alignments that could not be explained by each mutation occurring only once in the maximally parsimonious reconstruction of the sequences' evolution. One of the first applications of this approach, using protein sequences, was an investigation of cytochrome evolution by Sneath *et al* (Sneath *et al*., 1975). Early examples using DNA sequences include tests for 'incongruent phylogenetic partititions' (Stephens, 1985) and distributions of discordant sites (Sawyer, 1989). Algorithms were also proposed for the detection of imports, such as the maximum χ^2 test (Maynard Smith, 1992), which essentially identified boundaries between regions of low, and high, SNP density. These tests all rely on detecting recombinations as uninterrupted runs of polymorphisms. However, when a population exists in complete linkage equilibrium, such contiguous segments of sequence with common ancestry no longer exist, as they are broken up by continual horizontal exchange (Maynard Smith, 1999). In such a situation, methods such as the 'homoplasy test' can be used to identify whether the population is recombining or not, but the lack of extended regions with common ancestry means identification of recombined loci themselves is no longer possible (Maynard Smith and Smith, 1998).

Another productive approach that was developed involved scanning the alignment using a moving window, identifying recombination boundaries as positions at which the surrounding upstream and downstream regions of the alignment implied different relationships between the taxa. At first, phylogenies were used to evaluate relationships: maximum parsimony (Fitch and Goodman, 1991; Hein, 1993), distance-based (McGuire *et al*., 1997) and maximum likelihood methods (Grassly and Holmes, 1997) were each implemented. An alternative involving comparing the distance matrices themselves along the alignment, thereby avoiding the need to construct trees, was also developed (Weiller, 1998). However, these methods often suffered from having a poor ability to detect imports, either because they did not consider branch lengths and therefore could only detect recombinations that changed the topology of the tree (Fitch and Goodman, 1991; Hein, 1993), or else struggled to distinguish imports from regions of the alignment where selection was relaxed, and hence diverged more quickly than the rest of the sequences (McGuire *et al*., 1997). Hence amendments to such algorithms were employed to improve the accuracy of import identification (McGuire and Wright, 2000).

Such approaches were subsequently adapted for Bayesian methods; rather than a moving window, a Hidden Markov Model (HMM), the states of which were phylogenies of different topologies, was used to scan the alignment for segments of sequence with different ancestries (McGuire *et al*., 2000). This was later modified to improve the distinction between heterogeneity in the rate of point mutation accumulation and imports, as for the scanning window approaches (Husmeier, 2005). These approaches were extended following the exposition of the concept of a 'clonal frame' in bacteria: that sufficiently closely related isolates would share a clonally descended fraction of their chromosomes, interrupted by a number of dispersed loci that had undergone recombination since their divergence (Milkman and Bridges, 1993). A Bayesian algorithm was developed that removed recombinant segments of sequences from the alignment, such that a final phylogeny was produced, based only on the clonal frame of each taxon (Didelot and Falush, 2007). The recombinations themselves were identified by an HMM on the basis of their elevated density of polymorphisms, rather than considering homoplasy; however, this analysis was not performed on the whole alignment, but instead independently on each branch of the

tree, using the patterns of reconstructed SNPs. This improves the sensitivity and resolution for predicting the occurrence of recombinations, and ameliorates, to an extent, the problem of distinguishing imports from loci accumulating mutations at a relatively high rate. This is because relaxed selection at a locus leads to mutations accumulating at a generally elevated rate throughout the tree, whereas in the event of a recombination, a large number of polymorphisms are introduced simultaneously. However, while this approach has been successfully applied to MLST datasets, its computational intensity means it is impractical for large, whole genome alignments.

4.2 Analysis of the PMEN1 population

4.2.1 Construction of the phylogeny

Sequence reads were mapped against the complete reference chromosome of *S. pneumoniae* ATCC 700669, identifying 39,107 polymorphic sites. Maximum likelihood analysis of these data produced a phylogeny with a high proportion of homoplasic sites (23%) and a weak correlation between the date of a strain's isolation and its distance from the root of the tree (Pearson correlation, N = 222, $R^2 = 0.05$, $p = 0.001$) (Figure 4.2). This suggested that variation was primarily arising through recombination and not through steady accumulation of base substitutions. In order to generate a phylogeny based on the clonal frame of the isolates, a maximum likelihood-based algorithm was designed to remove the recombinations from each taxon through analysis of the patterns of polymorphisms occurring on each branch, analogous to the Bayesian implementation of Didelot and Falush (Didelot and Falush, 2007). All recombinations were assumed to be imports; as PMEN1 constitutes a small fraction of the overall carried pneumococcal population, it was assumed that the rate of exchange between members of the lineage would be negligible.

Using the starting phylogeny constructed on the basis of all SNPs using RAxML (Stamatakis *et al*., 2005), the pattern of polymorphic events occurring on each branch of the tree was reconstructed using PAML (Yang, 2007). The positions of the SNPs occurring on each branch across the reference chromosome were analyzed using a one dimensional spatial scan statistic (Kulldorf, 1997) in order to detect clusters of polymorphisms that would indicate recombination events. The null hypothesis for

branch B, $H_{0,B}$, assumed the absence of any recombination events, therefore implying the SNPs occurring on the branch should be evenly distributed across the chromosome. This was considered a reasonable axiom considering the closely related nature of these isolates, as there should be minimal opportunity for selection to cause any significant spatial heterogeneity in the level of observed base substitutions. Hence $H_{0,B}$ was modelled as a binomial distribution, with SNPs uniformly distributed throughout the chromosome, of length g, occurring at a mean frequency of $d_{0,B}$, the mean number of SNPs per base, calculated separately for each branch B. This was tested using a moving window, which was altered in length, w, such that, given the number of polymorphisms occurring on the branch, the mean number of SNPs in a window, N, would be at least 10 according to $H_{0,B}$ (up to a maximum window length of 10 kb; Equation 4.1).

$$H_{0,B}: N \sim \text{Bin}(w, d_{0,B})$$

Equation 4.1

The test statistics were only calculated for the moving window at polymorphic sites, hence the threshold for significance was set as 0.05 divided by the number of SNPs occurring on the branch. Each region of the chromosome, r, where $H_{0,B}$ could be rejected at this threshold was treated as containing a recombination, and hence conforming to an alternative hypothesis, $H_{1,B,r}$, that it contained a higher density of SNPs, $d_{1,B,r}$, calculated as the mean number of SNPs within the region (Equation 4.2).

$$H_{1,B,r}: N \sim \text{Bin}(w, d_{1,B,r})$$

Equation 4.2

However, the size of these identified regions exceeded the length of the recombination they contained. In order to more precisely delineate the borders of the recombination event within the regions identified by the moving window, the block was first reduced in size such that its boundaries were the outermost SNPs within the region. Each end of the block was then progressively moved inwards until the density of SNPs within the block was more likely under $H_{1,B,r}$ than $H_{0,B}$. Once the boundaries of the putative recombination had been identified, the inequality Equation 4.3 had to

be satisfied as a final test for rejection of $H_{0,B}$ on the basis of the length of the block, b, and the number of SNPs it contained, N.

$$\frac{0.05}{g/b} > 1 - \sum_{i=0}^{i=N-1} \binom{b}{i} d_{0,B}{}^{i} (1 - d_{0,B})^{b-i}$$

Equation 4.3

This condition was required to eliminate false positive events generated as artifacts of the window length spanning the edges of neighboring, but separate, clusters of SNPs. The block identified in this manner as having the smallest likelihood ratio, calculated as the probability of the block under $H_{0,B}$ divided by its probability under $H_{1,B,r}$, was then removed from the dataset, and $d_{0,B}$ was recalculated as the mean density of SNPs across the remainder of the chromosome outside of this recombination block. The identification of recombinations was then repeated, with the process iterating until either no more loci deviated from $H_{0,B}$ or the minimum number of SNPs within a window required to identify a recombination fell below three. This approach was taken to avoid SNP-dense regions reducing the power to detect other recombinations occurring on the same branch.

The loci corresponding to these putative recombination events were then treated as missing data in all taxa downstream of the branch on which the recombination was estimated to occur when redrawing the phylogeny with RAxML. Subsequently all mutations, including those occurring within putative recombination events, were reconstructed on the new tree and recombinations re-identified as described above. This process was repeated for five iterations to produce the final dataset. The algorithm rapidly converges on a topology, as assessed by comparing the phylogenies produced by each iteration using ftreedist (Rice *et al*., 2000) (**Table 4.1**). Additionally, extending the analysis for a further four iterations resulted in few changes in the tree (**Table 4.1**), suggesting that the output of the algorithm in the case of this study is robust. The only alterations between iterations involve rearrangements concerning very short branches near the base of the tree that are difficult to resolve, with the annotated clades identified in Figure 4.1 consistently identified in all phylogenies from iteration 2 onwards.

From this analysis (Figure 4.1, Figure 4.3), a total of 57,736 single-nucleotide polymorphisms (SNPs) were reconstructed as occurring during the history of the lineage, 50,720 (88%) of which were introduced by 702 recombination events. This gives a per site *r/m* ratio (the relative likelihood that a polymorphism was introduced through recombination rather than point mutation) of 7.2, less than the previously calculated value of ~66 from MLST data (Feil *et al*., 2000). By removing recombination events from the phylogeny, the number of homoplasic sites is reduced by 97%, and the tree has significantly shortened branches, such that root-to-tip distance more strongly correlates with date of isolation ($R^2 = 0.46$, $p = <2.2 \times 10^{-16}$; Figure 4.2). The rate at which base substitutions occur outside of recombinations suggests a mutation rate of 1.57×10^{-6} substitutions per site per year (95% confidence interval 1.34 to 1.79×10^{-6}), close to the estimate of 3.3×10^{-6} substitutions per site per year from *Staph. aureus* ST239 (Harris *et al*., 2010) and much higher than that of $\sim 5 \times 10^{-9}$ substitutions per site per year found between more distantly related isolates (Ochman *et al*., 1999). Furthermore, by excluding SNPs introduced through recombinations, the date of origin of the lineage implied by the tree moved from about 1930—which predates the introduction of penicillin, chloramphenicol, and tetracycline—to about 1970 (Figure 4.2).

This method inevitably underestimates the level of recombination occurring in the population, as it only allows for the detection of imports that generate a sufficient level of sequence diversity. Hence the estimate of the *r/m* ratio is effectively a lower bound for the value. In order to quantify a probable upper bound for this parameter, it is necessary to consider how many of the SNPs identified as substitutions may actually have resulted from recombinations. It is possible that the 348 substitutions that are homoplasic with SNPs found in recombinations in the dataset may have originated through short recombinations importing a single polymorphism; this would raise the estimate of *r/m* to 7.7. In addition, if the substitution homoplasies were considered to have arisen once through point mutation, whilst the remaining instances represented horizontal transfer of this SNP, then the value of *r/m* would rise to 8.1. Finally, if all homoplasies were considered to have arisen through recombination, then the value of *r/m* would be 8.2.

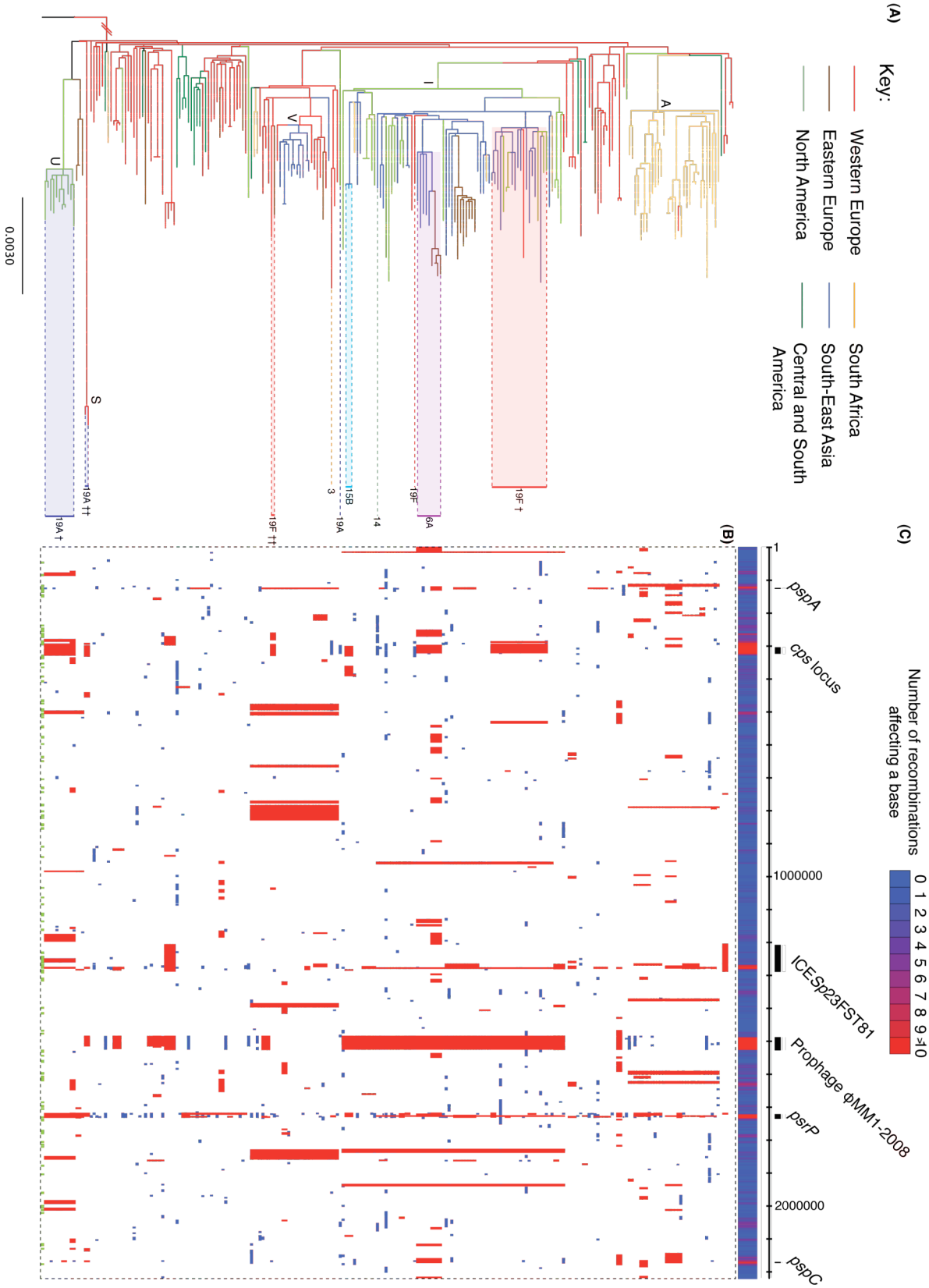
(A)
Key:
Western Europe
Eastern Europe
North America
South Africa
South-East Asia
Central and South America
0.0030
U
S
V
I
A
19A †
19A ††
19F ††
3
19A
15B
14
19F
6A
19F †
(B)
(C)
Number of recombinations affecting a base
0 1 2 3 4 5 6 7 8 9 ≥10
1
pspA
cps locus
1000000
ICESp23FST81
Prophage ϕMM1-2008
psrP
2000000
pspC

Figure 4.1 Phylogeography and sequence variation of PMEN1. (A) Global phylogeny of PMEN1. The maximum likelihood tree, constructed using substitutions outside of recombination events, is coloured according to location, as reconstructed through the phylogeny by using parsimony. Shaded boxes and dashed lines indicate isolates that have switched capsule type from the ancestral 23F serotype. Independent switches to the same serotype are distinguished by annotation with daggers (†). Specific clades referred to in the text are marked on the tree: A (South Africa), I (International), V (Vietnam), S (Spain 19A), and U (USA 19A). (B) Recombinations detected in PMEN1. The panel shows the chromosomal locations of the putative recombination events detected in each terminal taxon. Red blocks are recombinations predicted to have occurred on an internal branch and, therefore, are shared by multiple isolates through common descent. Blue blocks are recombinations predicted to occur on terminal branches and hence are present in only one strain. The green blocks indicate recombinations predicted to have occurred along the branch to the outgroup (*S. pneumoniae* BM4200), used to root the tree. (C) Biological relevance of recombination. The heat map shows the density of independent recombination events within PMEN1 in relation to the annotation of the reference genome. All regions that have undergone 10 or more recombination events are marked and annotated (Tn*916* is encompassed within ICE*Sp*23FST81).

All of these estimates remain considerably lower than the equivalent values estimated from MLST data (Feil *et al*., 2000). It should be noted that the genome-wide data demonstrate that these averages do not apply uniformly across the chromosome, but instead will vary considerably between different loci. Furthermore, it remains possible that an even higher proportion of the substitution SNPs could be the result of recombinations. Despite this underestimation of the rate at which recombination imports variation, the net rate of point mutation remains quite similar to that of *Staph. aureus* ST239, which has a much lower *r/m* value. This emphasises that recombination will overwrite base substitutions, as well as introducing variation.

4.2.2 Recombination and antigenic variation

Even in this sample of a single lineage, 74% of the reference genome length has undergone recombination in at least one isolate, with a mean of 74,097 bp of sequence affected by recombination in each strain. This encompasses both site-specific integrations of prophage and conjugative elements and homologous recombinations mediated by the competence system. The 615 recombinations outside of the prophage and ICE vary in size from 3 bp to 72,038 bp, with a mean of 6.3 kb (Figure 4.4). Within these homologous recombinations, there is a distinct heterogeneity in the density of polymorphisms, although it is unclear whether this represents a consequence of the mechanism by which horizontally acquired DNA is incorporated or a property of the donor sequence.

Recombination hotspots are evident in the genome where horizontal sequence transfers are detected abnormally frequently (Figure 4.1). One of the most noticeable is within Tn*916*, concentrated around the *tetM* gene. Excepting the prophage, the other loci—*pspA*, *pspC*, *psrP*, and the *cps* locus—are all major surface structures. Hence, it seems likely that these loci are under diversifying selection driven by the human immune system, and consequently, the apparent increase in the frequency of recombination in these regions is due to the selective advantage that is offered by the divergent sequence introduced by such recombination events.

In addition to base substitutions, 1,032 small (<6-bp) insertion and deletion events can be reconstructed onto the phylogeny, of which 61% are concentrated in the 13% of the genome that does not encode for CDSs, probably because of selection against the introduction of frameshift mutations. Throughout the phylogeny, 331 CDSs are predicted to be affected by either frameshift or premature stop codon mutations. Modeling these disruptive events as a Poisson distributed process occurring at a rate proportional to the length of the CDS, 11 CDSs were significantly enriched for disruptive mutations after correction for multiple testing (Table 4.2). These included *pspA* and a glycosyltransferase posited to act on *psrP* (SPN23F17730). This again suggests there may be a selective pressure acting either to remove (*pspA*) or alter (*psrP*) two major surface antigens. Furthermore, the longest recombination in the data set spans, and deletes, the *psrP*-encoding island, which shows that such non-essential antigens can be quickly removed from the chromosome. These data imply that the pneumococcal population is likely to be able to respond very rapidly to the introduction of some of the protein antigen–based pneumococcal vaccines currently under development.

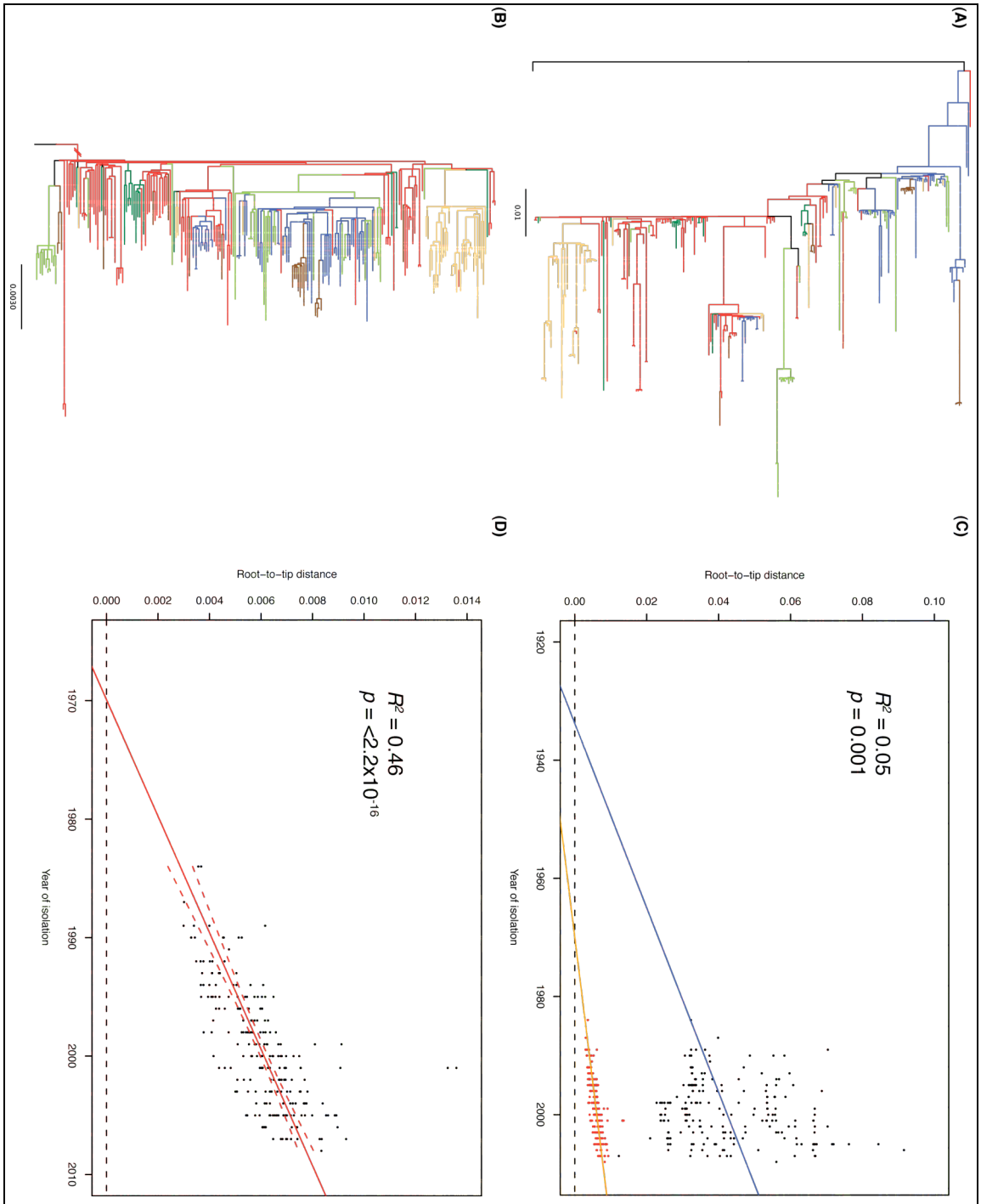

Figure 4.2 Construction of the PMEN1 phylogeny. The maximum likelihood phylogeny constructed on the basis of all SNPs, coloured according to geographical distribution as in Figure 4.1, is shown in (A), and that derived by excluding those SNPs falling within putative recombination events is shown in (B) for comparison. A plot of root-to-tip distance against date of isolation, for taxa for which dates were available, is shown in (C): points in black correspond to tree (A), and those in red correspond to tree (B), with the regression lines coloured blue and orange, respectively. In (D), the points corresponding to tree (B) are shown in greater detail, with the 95% confidence interval indicated by the dashed red lines. The two outlying points correspond to the Spanish 19A isolates (clade 'S' in Fig. 1), which lie on a long branch that indicates they may have accumulated mutations at an unusually high rate at some point in their recent history. This graph suggests the PMEN1 lineage originated around 1970.

4.2.3 Population and serotype dynamics

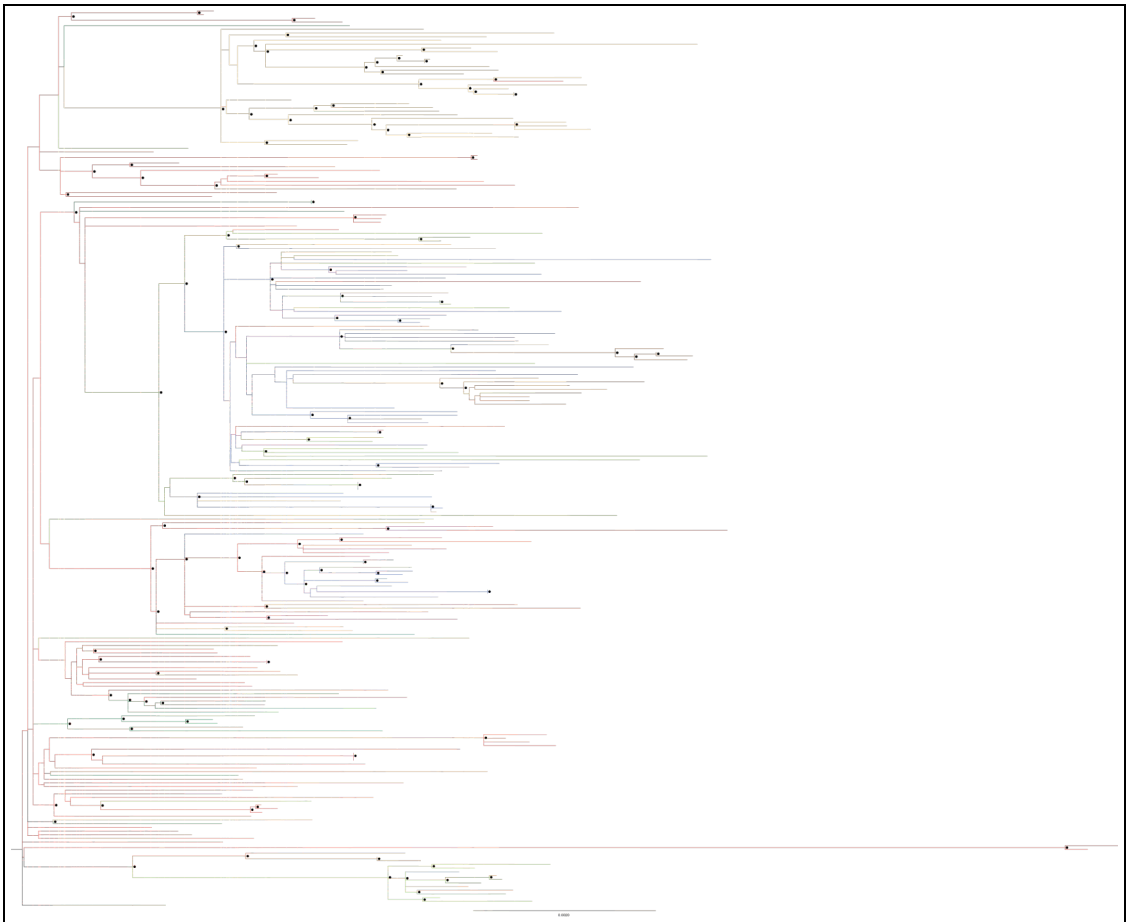

Figure 4.3 Robustness of the PMEN1 phylogeny. All nodes in the tree with bootstrap values greater than, or equal to, 75 are marked with a black dot (•).

The spread of PMEN1 can be tracked using the phylogeography indicated by the tree (Figure 4.1). There are several European clades with their base near the root of the tree, and a parsimony-based reconstruction of location supports a European origin for the lineage. Interspersed among the European isolates are samples from Central and South America, which may represent an early transmission from Spain, where the clone was first isolated, to Latin America, a route previously suggested to occur by data from *Staph. aureus* (Harris *et al*., 2010). One clade (labelled A in Figure 4.1), containing South African isolates from 1989 to 2006, appears to have originated from a single highly successful intercontinental transmission event. There is also a cluster of isolates from Ho Chi Minh City (labeled V), representing a transmission to Southeast (SE) Asia. However, the predominant clade found outside of Europe (labelled I) appears to have spread quite freely throughout North America, SE Asia,

and Eastern Europe, which implies that there are few barriers to intercontinental transmission of *S. pneumoniae* between these regions.

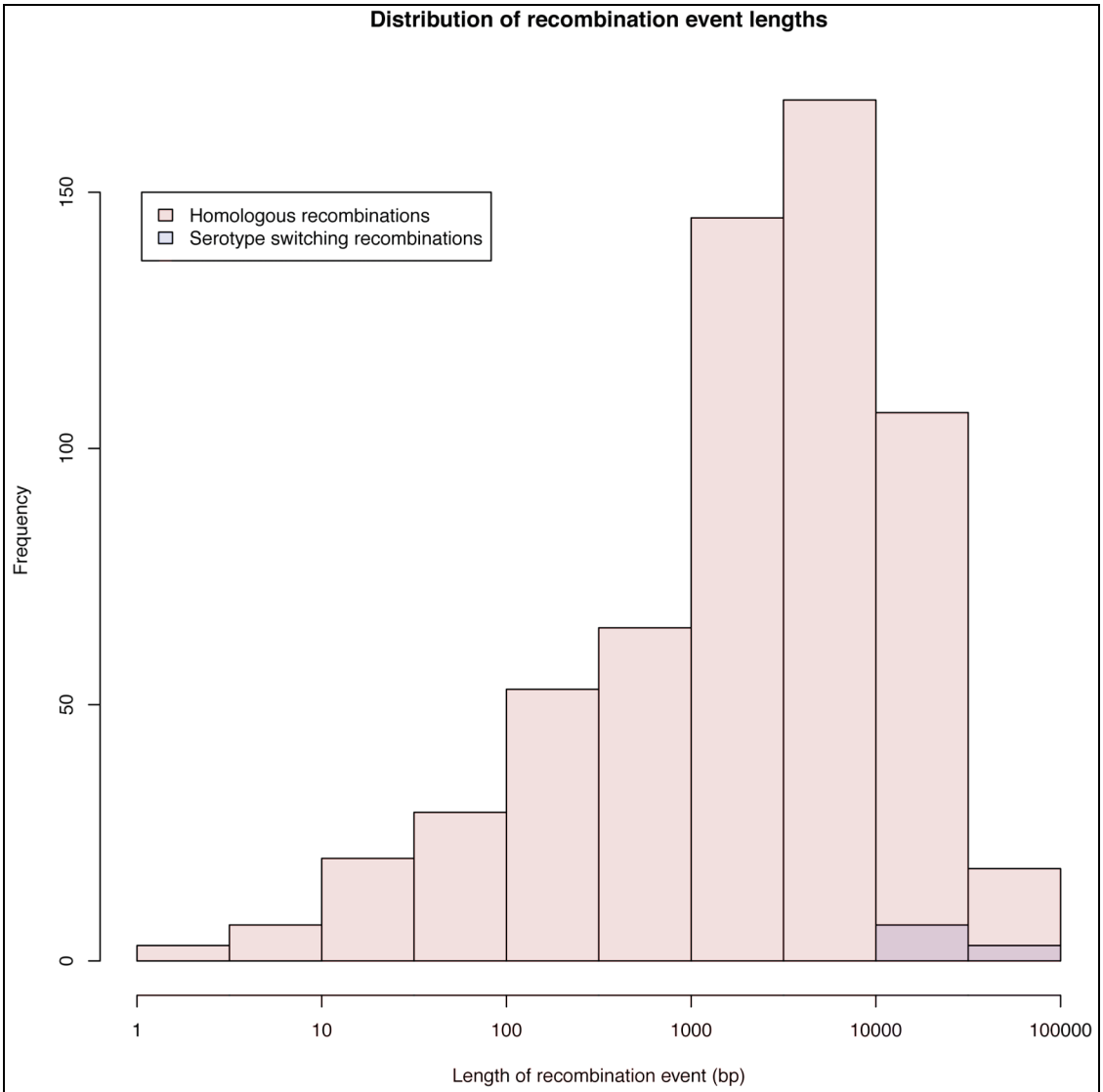

Figure 4.4 Histogram showing the variation of recombination event lengths. The bars in red show the distribution of lengths for the 615 homologous recombinations occurring outside of ICE*Sp*23FST81 and the ϕMM1-2008 prophage on a $\log_{10}$ scale. The bars in blue represent the ten recombinations causing serotype switching. This clearly shows that, although such events are within the range of homologous recombination lengths found elsewhere in the chromosome, they are atypically large. Hence it is unusual that serotype switches are observed so frequently within the PMEN1 population.

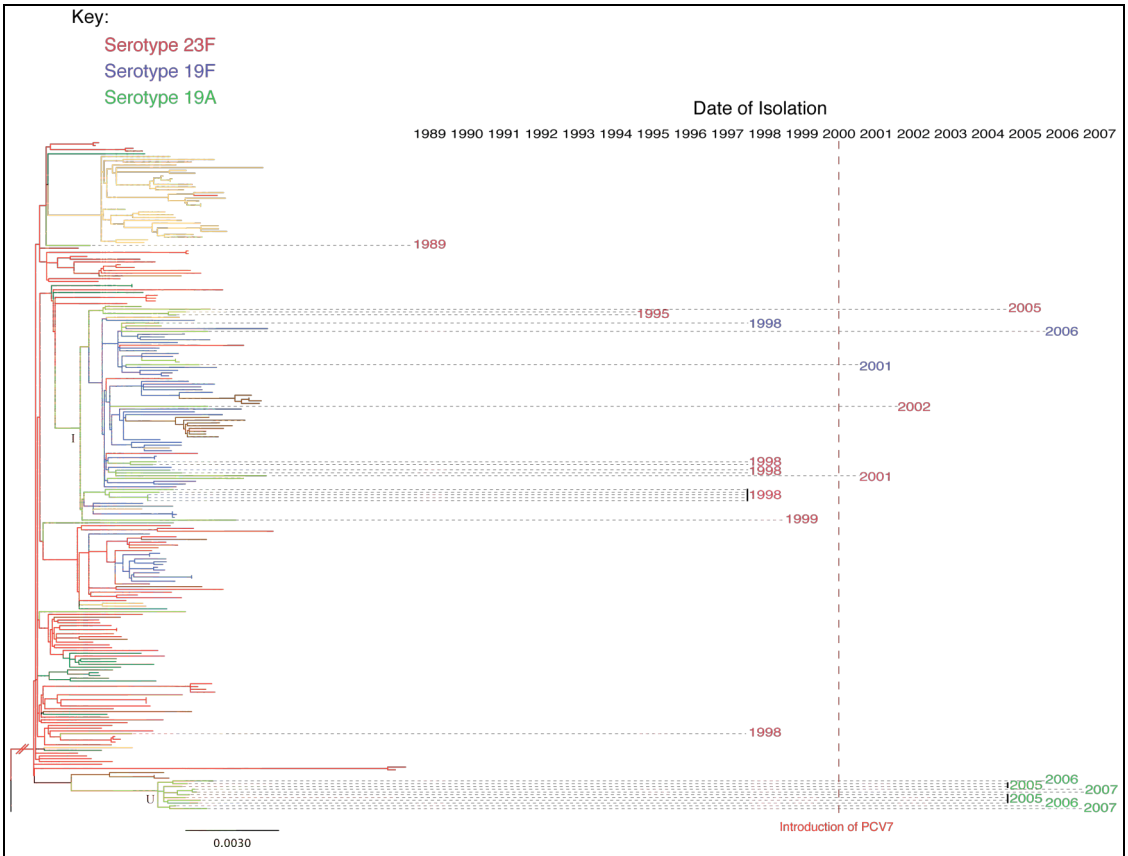

Figure 4.5 Population dynamics of PMEN1 in the USA. All PMEN1 isolates in the phylogeny from the United States, where the PCV7 conjugate polysaccharide vaccine was introduced in 2000, are labelled with their year of isolation and coloured according to their serotype. The earliest isolate, from Cleveland in 1989, appears to represent an initial, limited transmission event. The more successful colonization seems to have come from clade I, which contains all but one of the isolates from the 1990s. This group includes US isolates of serotypes 23F and 19F, both of which correspond to antigens present in PCV7. Hence there is a shift in the population to the serotype 19A isolates (clade U), not targeted by the vaccine, in the later years of this study.

The final non-European group consists of serotype 19A U.S. isolates (labelled U). These all date from between 2005 and 2007 and are distinct from all other U.S. PMEN1 isolates, which have capsular types included in PCV7 (Figure 4.5). This is evidence of a shift in the PMEN1 population in the USA: rather than a change in capsule type occurring among the resident population, it has been eliminated by the vaccine and replaced by a different subpopulation within the lineage that has expanded to fill the vacated niche. Similarly, a pair of Spanish isolates from 2001 (labeled S in Figure 4.1), the year in which PCV7 was introduced in Spain, that have independently acquired a 19A capsule are not closely associated with any other European isolates. The estimated times of origin for clades U (1996; 95% credible interval 1992–1999) and S (1998; 95% credible interval 1996–1999) both predate the introduction of PCV7, and accordingly a third 19A switch, from Canada, was isolated

in 1994. Hence, it appears that these changes in serotype after vaccine introduction result from an expansion of pre-existing capsular variants, which were relatively uncommon and not part of the predominant population, and would have therefore been difficult to detect before the existence of the selection pressure exerted by the vaccine.

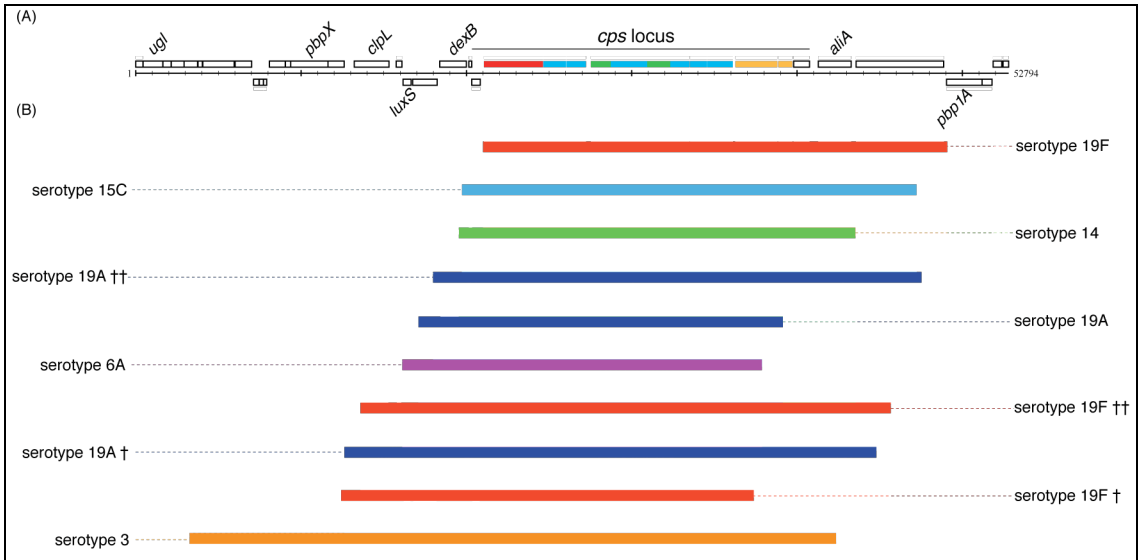

Figure 4.6 Recombinations causing serotype-switching events. (A) The annotated *cps* locus of the reference strain. CDSs involved in capsule biosynthesis are coloured according to their role. Genes in red are regulatory, those in blue synthesize and modify the oligosaccharide subunits, those in green are involved in polymerization and transport, and those in orange are required for the synthesis of rhamnose. (B) Below are delineated the recombinations leading to changes in serotype, coloured according to the serotype of the sequence donor. The different events are labelled to correspond with Figure 4.1.

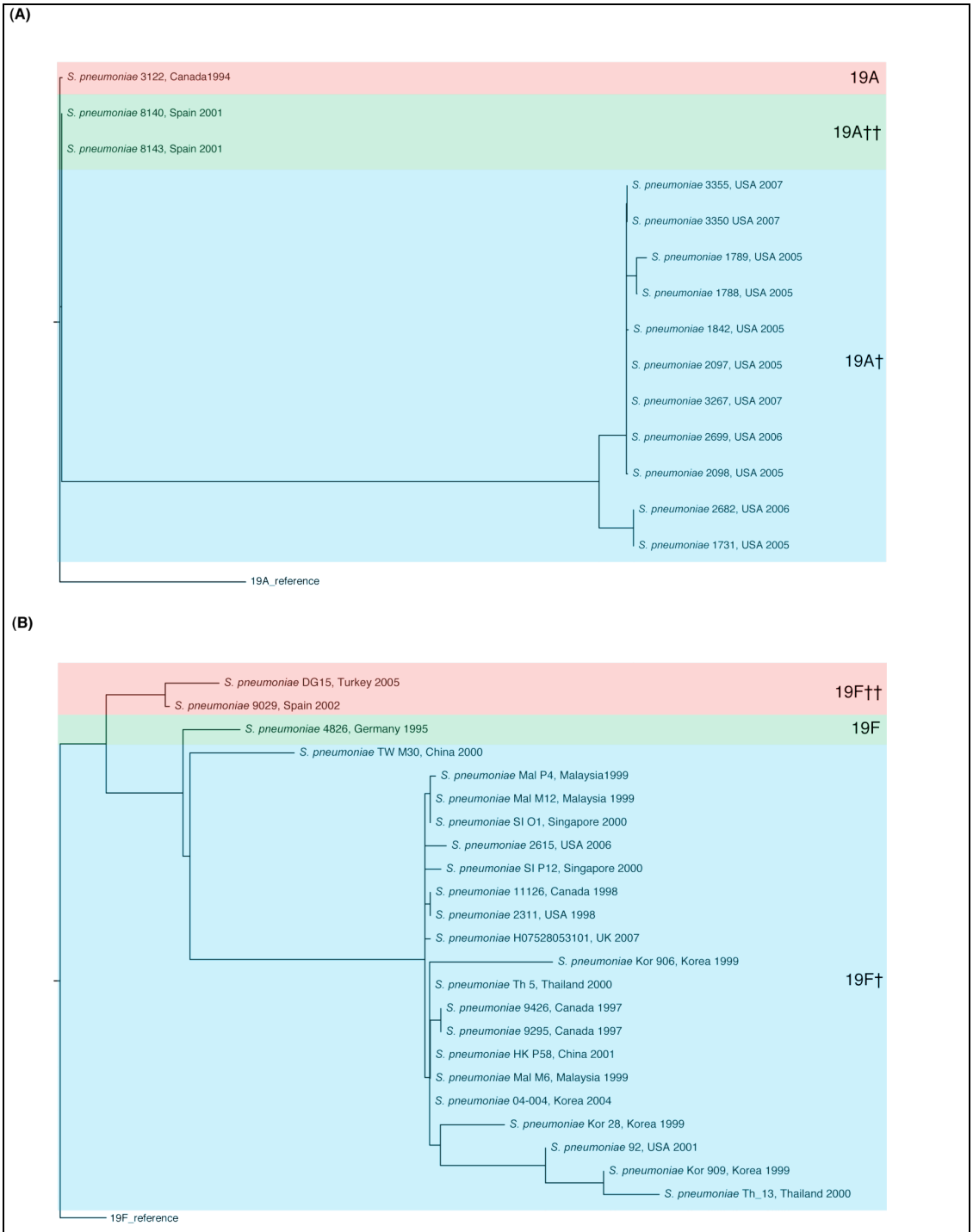

Figure 4.7 Variation in the serogroup 19 *cps* loci. (A) The *cps* loci of the serotype 19A isolates were assembled *de novo*, aligned and a maximum likelihood tree generated. The three acquisition events are coloured differently and marked according to the scheme in Figure 4.1. This shows the *cps* locus acquired by the US isolates is quite distinct from that received by the Spanish or Canadian isolates. There is variation in this region even between the US isolates: *S. pneumoniae* 1731 and 2682 appear to have a second recombination affecting the *wzg* and *wzh* genes at the 5' end of the *cps* locus. (B) The *cps* loci of the serotype 19F isolates were analyzed in the same manner. Here, *S. pneumoniae* TW M30 differs from the other isolates with which it shares its 19A *cps* locus through common descent due to a secondary recombination affecting the *rmlABCD* gene cluster.

Seven further serotype-switching events can be detected in the data (Figure 4.6), including three switches to serotype 19F. The polyphyletic nature of these 19F isolates is supported by the variation observed between the acquired *cps* loci, as is also the case for the 19A isolates (Figure 4.7). The previously known switches to serotypes 3, 6A, and 15B are only found to occur once each in the phylogeny, and in addition, a single Korean sample that had not been typed was identified as a serotype 14 variant by mapping reads to known *cps* loci (Bentley *et al*., 2006). The recombination events leading to these switches ranged from 21,780 bp to 39,182 bp in size, with a mean of 28.2 kb. Only 35 homologous recombinations of an equivalent size or larger occur elsewhere in the genome; most such events are much smaller (Figure 4.4), which makes it surprising that serotype switching occurs with such frequency, indicating a role for balancing selection at this locus. Additionally, the span of these events appears to be limited by the flanking penicillin-binding protein genes, the sequences of which are crucial in determining β-lactam resistance in pneumococci (Trzcinski *et al*., 2004). Only the recombination causing the switch to serotype 3 affects one of these, and it introduces just a single SNP into the *pbpX* CDS, which does not appear to compromise the strain's penicillin resistance (Appendix II: PMEN1 strains). Hence, the positioning of these two genes may hinder the transfer of capsule biosynthesis operons from penicillin-sensitive to penicillin-resistant pneumococci via larger recombinations, although size constraints alone could also cause such a distribution.

4.2.4 Resistance to non–β-lactam antibiotics

The strong selection pressures exerted by antibiotics on the PMEN1 lineage are manifest as multiple examples of geographically disparate isolates converging on common resistance mechanisms. Single base substitutions causing reduced susceptibility to some classes of antibiotics have occurred multiple times throughout the phylogeny, as observed in *Staph. aureus* (Harris *et al*., 2010) and *Salmonella* Typhi (Holt *et al*., 2008) populations, including mutations in *parC*, *parE*, and *gyrA*, which cause increased resistance to fluoroquinolone antibiotics (Pletz *et al*., 2004), and changes in *rpoB* causing resistance to rifampicin (Ferrandiz *et al*., 2005). The S79F, S79Y, and D83N mutations in *parC* are estimated to occur nine, three, and five times, respectively, in PMEN1; additionally, D435N in the adjacent *parE* gene is

found to happen three times. The S81F and S81Y substitutions, in the same position of *gyrA*, are found four and two times, respectively. None of these mutations are predicted to have been introduced by recombination, whereas changes at position H499 of *rpoB* causing rifampicin resistance are introduced twice by horizontal transfer and three times by means of base substitution.

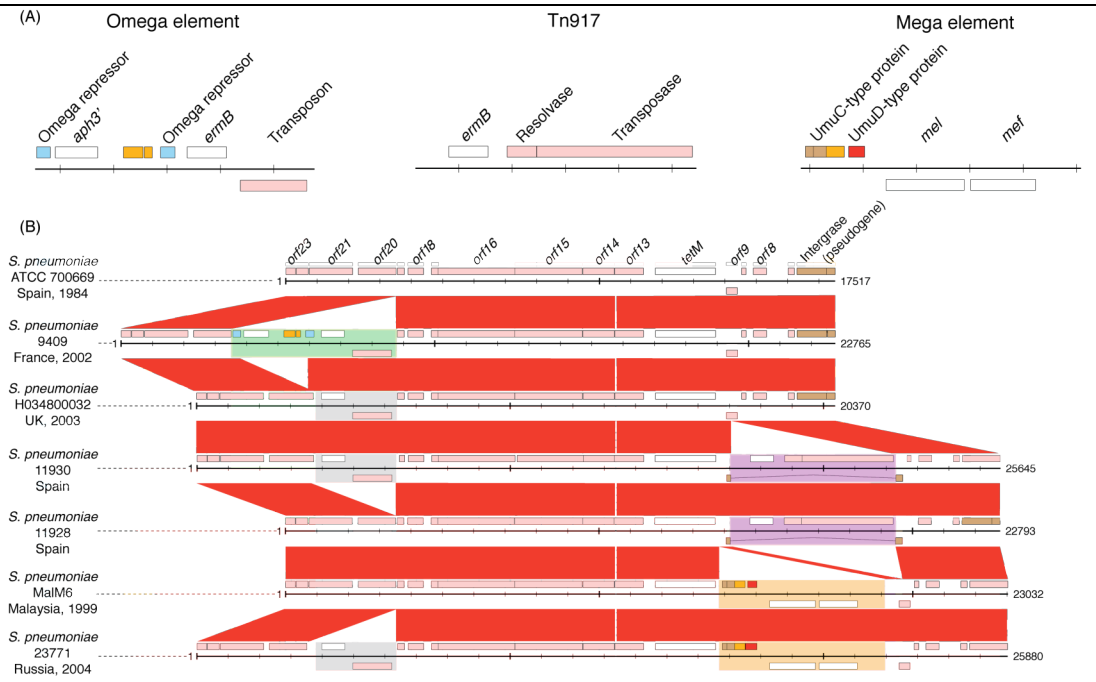

Figure 4.8 Acquisition of macrolide-resistance cassettes. The three full-length resistance cassettes are shown in (A): the omega element, which carries an *aph3′* aminoglycoside-resistance gene and an *ermB* macrolide-resistance gene; Tn*917*, which carries just the *ermB* methylase; and the mega element, which carries the *mel*/*mef* macrolide efflux system. (B) A comparison of the different Tn*916* variants in the PMEN1 lineage. Red bands between the sequences indicate BLASTN matches. The omega element is shaded green when present at full length and shaded grey where present as a remnant resulting from a recombination between the omega repressor–encoding genes, which concomitantly leads to the fusion of an omega transcriptional repressor domain to the 3′ of the *orf20* CDS. Tn*917* is boxed in purple, with the two parts of *orf9*, into which it inserts, indicated on either side. The mega element is boxed in orange.

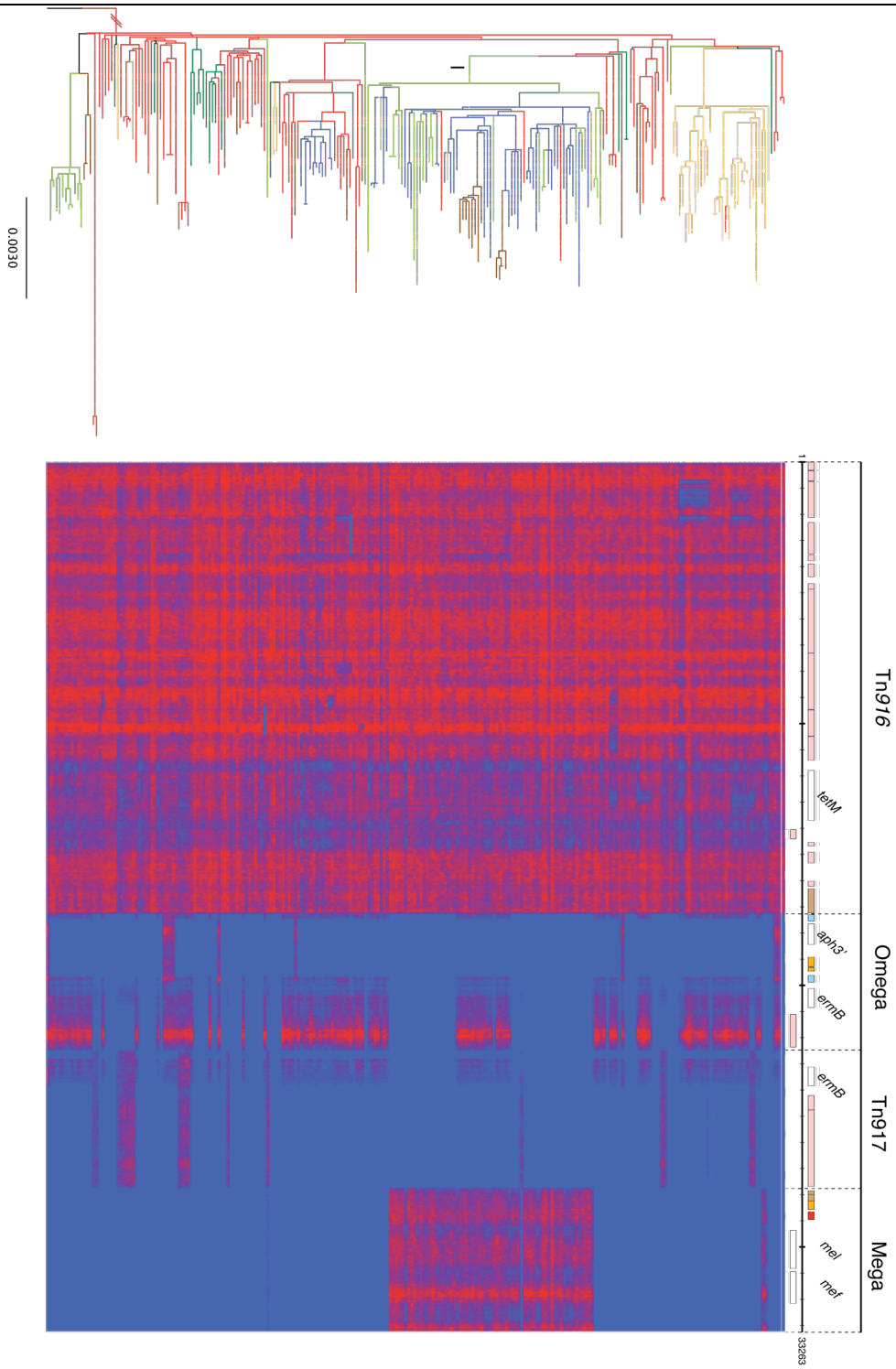

Figure 4.9 Distribution of macrolide resistance cassettes amongst the PMEN1 isolates in relation to the phylogeny, on which the 'International' clade is marked with an 'I'. The reference Tn*916* annotation is shown at the top, along with the Omega, Tn*917* and Mega elements. The level of sequence read mapping from each isolate in the phylogeny is displayed as a heatmap, with blue indicating low mapping and red indicating high coverage. Many isolates only have coverage of the 3' half of the Omega element, because the 5' end has been deleted through a recombination between the Omega repressor encoding genes (Figure 4.8). There is sufficient similarity between the *ermB* nucleotide sequences in the Omega and Tn*917* elements to allow for cross mapping of Illumina reads between them.

Resistance to macrolide antibiotics tends not to derive from SNPs, but from acquisition of CDSs facilitating one of the two common resistance mechanisms: methylation of the target ribosomal RNA by *erm* genes and removal of the drug from the cell by the macrolide efflux (*mef*)-type efflux pumps. Both can be found in the PMEN1 population, and in all cases, the genes appear to be integrated into the Tn*916* transposon (Figure 4.8). They are carried by three different elements. Tn*917*, consisting of an *ermB* gene with an associated transposon and resolvase, inserts into open reading frame *orf9* of Tn*916* (Shaw and Clewell, 1985). A second has been characterized as the macrolide efflux genetic assembly (mega) element (Del Grosso *et al*., 2006), which carries a *mef*/*mel* efflux pump system and, in PMEN1, inserts upstream of *orf9*. A third element (henceforth referred to as an omega element, for omega and multidrug-resistance encoding genetic assembly) carries both an *ermB* gene and an aminoglycoside phosphotransferase, with the latter flanked by direct repeats of omega transcriptional repressor genes, and is found just downstream of *orf20*.

Rather than a single acquisition of these elements occurring, and the resulting clones spreading and replacing macrolide-sensitive isolates, all three elements appear to have been acquired multiple times across the phylogeny (Figure 4.9). The mega element is predominantly shared by isolates in clade I, although the *ermB*-encoding omega element appears to have been subsequently acquired on two occasions, and Tn*917* has entirely superseded the mega element in one isolate. This is congruent with the known advantages of target methylation over drug efflux as a broader-spectrum resistance mechanism (Del Grosso *et al*., 2007). In most instances of the omega element, only the *ermB*-encoding part remains; the aminoglycoside phosphotransferase appears to have been deleted through a recombination between the omega-encoding genes, which leaves only an omega domain–encoding open reading frame fused to *orf20* as a scar. This implies that the benefit of the aminoglycoside-resistance element may have not been sufficient to maintain it on the ICE.

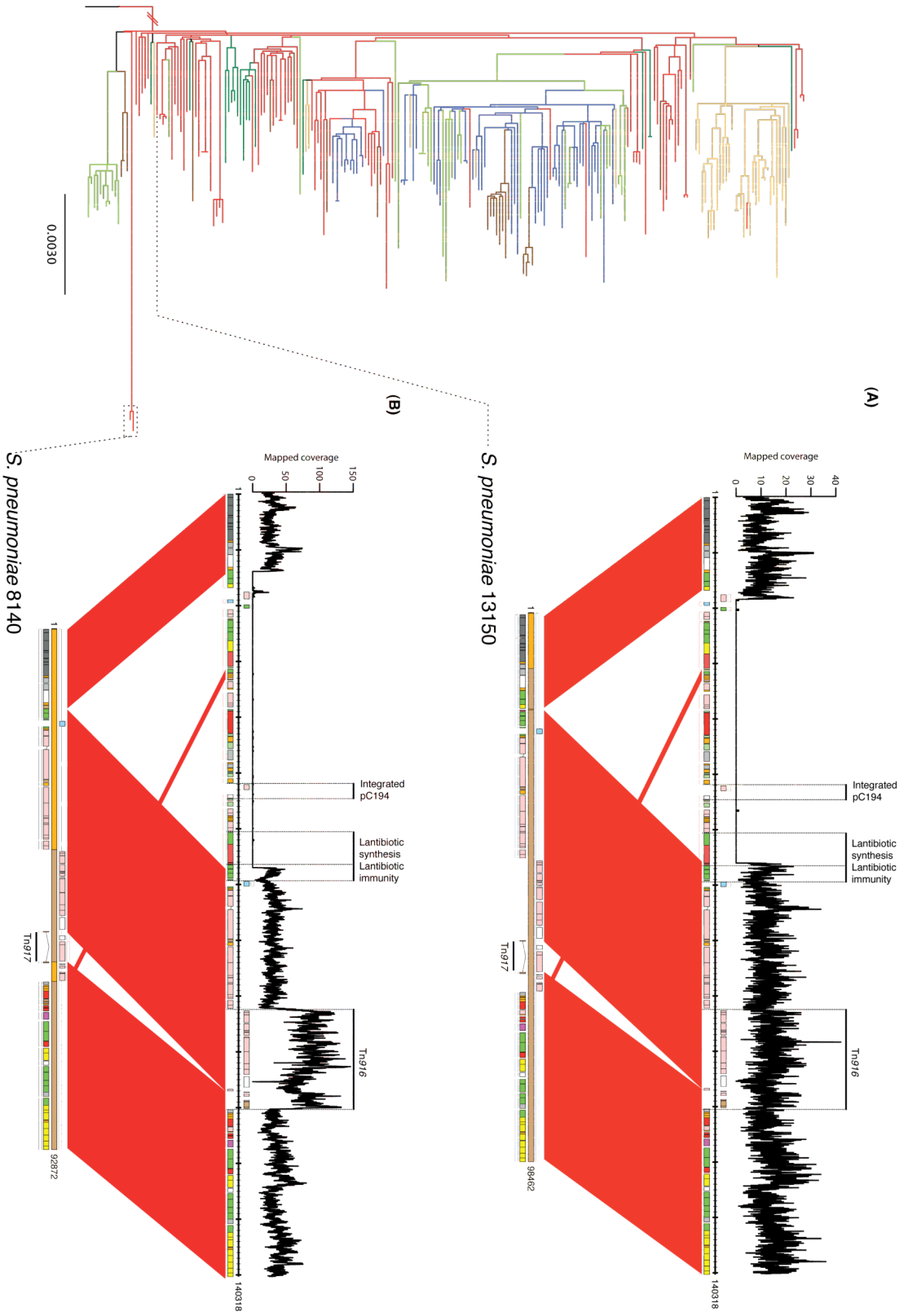
0.0030
(A)
S. pneumoniae 13150
Mapped coverage
0
10
20
30
40
Integrated pC194
Lantibiotic synthesis
Lantibiotic immunity
Tn916
Tn917
1
98462
140318
(B)
S. pneumoniae 8140
Mapped coverage
0
50
100
150
Integrated pC194
Lantibiotic synthesis
Lantibiotic immunity
Tn916
Tn917
1
92872
140318

Figure 4.10 Deletions affecting ICE*Sp*23FST81. Two deletions are displayed; in both cases, *de novo* assemblies are shown relative to the reference annotation, with red bands between the sequences indicating BLASTN matches. Alternate brown and orange boxes covering the scale line indicate scaffolds produced by assembly of Illumina reads, which have been ordered against the reference sequence. A graph illustrating the depth of Illumina read mapping to the reference sequence is shown above the reference annotation to demonstrate that the deletions do not represent deficiencies in the short read assemblies. (A) *S. pneumoniae* 13150, isolated in Germany in 1998, lacks the 5' end of the ICE, with the boundary of the deletion meaning the genes for the biosynthesis of the lantibiotic carried by the ICE are lost, but the genes encoding the ABC transporter presumed to be required for self-immunity are retained. The integrated pC194 plasmid that carries the chloramphenicol acetyltransferase gene of ICE*Sp*23FST81 is also lost. (B) *S. pneumoniae* 8140, isolated in Spain in 2001, has a similar deletion, starting slightly upstream but again ending in a position that retains most of the self-immunity transporter genes but removes the lantibiotic synthesis CDSs and the chloramphenicol acetyltransferase on pC194. The Tn*916* transposon of this strain has a depth of sequence read mapping approximately twice that of the rest of the locus, suggesting that a second copy of this element has been acquired somewhere in the chromosome.

4.2.5 Components of the accessory genome

Other than the insertion of these cassettes, the ICE itself is otherwise relatively unchanged throughout the population. In two cases, the 5′ region of the element up to, and including, the lantibiotic synthesis machinery is deleted, whereas the self-immunity genes are retained (Figure 4.10). This deletion, which also removes the integrated chloramphenicol-resistance plasmid, is analogous to that observed in the PPI–1 of the PMEN1 lineage, in which all that remains are the immunity genes from a once-intact lantibiotic synthesis machinery. In two other cases, the ICE has been supplanted by alternative transposons, both of which are similar composites of Tn*5252*- and Tn*916*-type elements: In *S. pneumoniae* 11876, a wholesale replacement at the same locus entails the gain of an omega element at the expense of losing resistance to chloramphenicol (Figure 4.11), whereas, in isolate 11930, the new ICE inserts elsewhere in the chromosome and carries two *ermB* genes, as well as a chloramphenicol acetyltransferase (Figure 4.12). The only other identified conjugative element was an ICE*St*1-type transposon shared by isolates 8140 and 8143 (Figure 4.13), and the only extrachromosomal element present in the data set was the plasmid pSpnP1 (Romero *et al*., 2007), found in isolate SA8.

The accessory genome is primarily composed of prophage sequence (Figure 4.14), with little evidence of much variation in the complement of metabolic genes. Viral sequences appear to be a transient feature of the pneumococcal chromosome (Figure 4.15), with few persisting long enough to be detected in related isolates. Four of the new prophage that could be assembled were found to insert into the competence pilus structural gene *comYC*, which lies within an operon shown to be essential for competence in *S. pneumoniae* (Pestova and Morrison, 1998). In two cases where such phage appear to be shared through common descent by pairs of isolates, no recombination events can be detected that are unique to either member of the pair, consistent with a nonfunctional competence system in these isolates. Furthermore, assaying the competence of available lysogenic strains in vitro also suggested that these phage insertions abrogate the ability of their host to take up exogenous DNA (Figure 4.16).

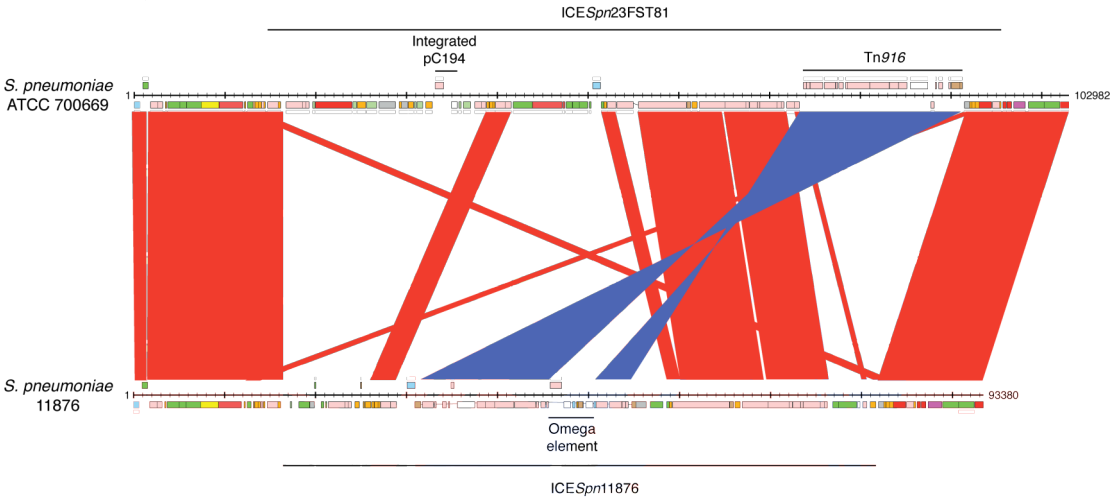

Figure 4.11 The putative ICE ICE*Sp*11876 displayed relative to ICE*Sp*23FST81. Red bands indicate BLASTN matches between the sequences in the same direction, blue bands indicate BLASTN matches between the sequences in the inverse orientation. ICE*Sp*11876 is inserted into the same locus as ICE*Sp*23FST81, shares the majority of the conjugative machinery and also carries a Tn*916* element, albeit at a different position and in the opposite orientation. This Tn*916* element carries an Omega element (Figure 4.8), and has inserted in a metalloprotease gene, which itself is part of a four gene island similar to one found in *S. equi*. The majority of the other ICE*Sp*11876 cargo genes are regulators or transporters, and it lacks the linearised pC194 plasmid that carries the chloramphenicol resistance gene found in ICE*Sp*23FST81.

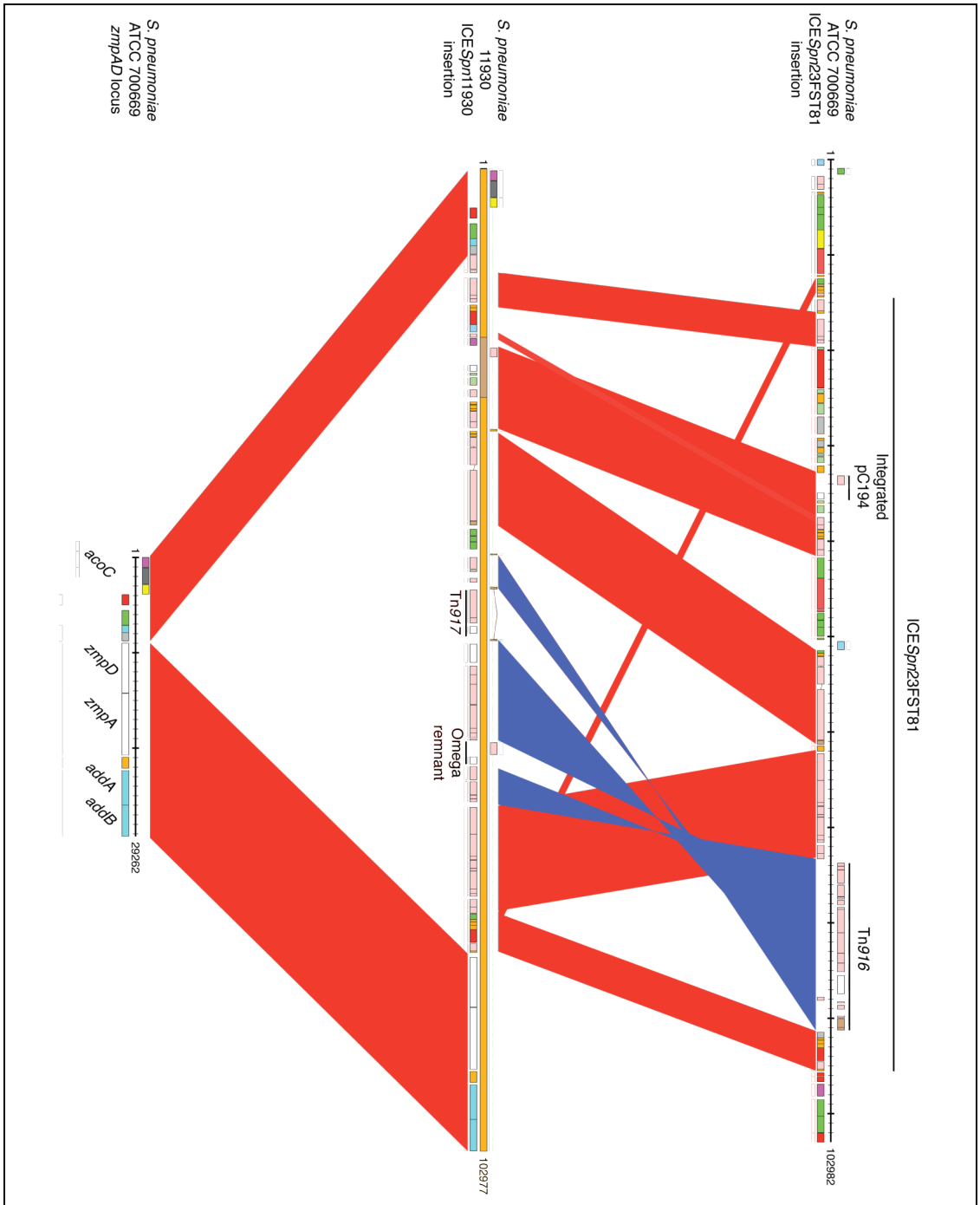

Figure 4.12 The putative ICE ICE*Sp*11930 displayed relative to ICE*Sp*23FST81, as described in Figure 4.11. Alternate brown and orange boxes covering the scale line indicate scaffolds produced by assembly of pyrosequencing and Illumina reads, which have been ordered against the reference sequence. This ICE has inserted downstream of the *zmpAD* genes, in a different position to ICE*Sp*23FST81, which is deleted in this strain. The majority of the conjugative machinery is shared between the ICE and ICE*Sp*11930, and both carry Tn*916*, albeit in different positions and in opposite orientations. The Tn*916* component of ICE*Sp*11930 carries two *ermB* genes, as the Tn*917* and Omega cassettes are both inserted into this transposon. Unlike ICE*Sp*11876 (Figure 4.11), ICE*Sp*11930 appears to carry the linearised pC194 plasmid, which contains the chloramphenicol acetyltransferase gene.

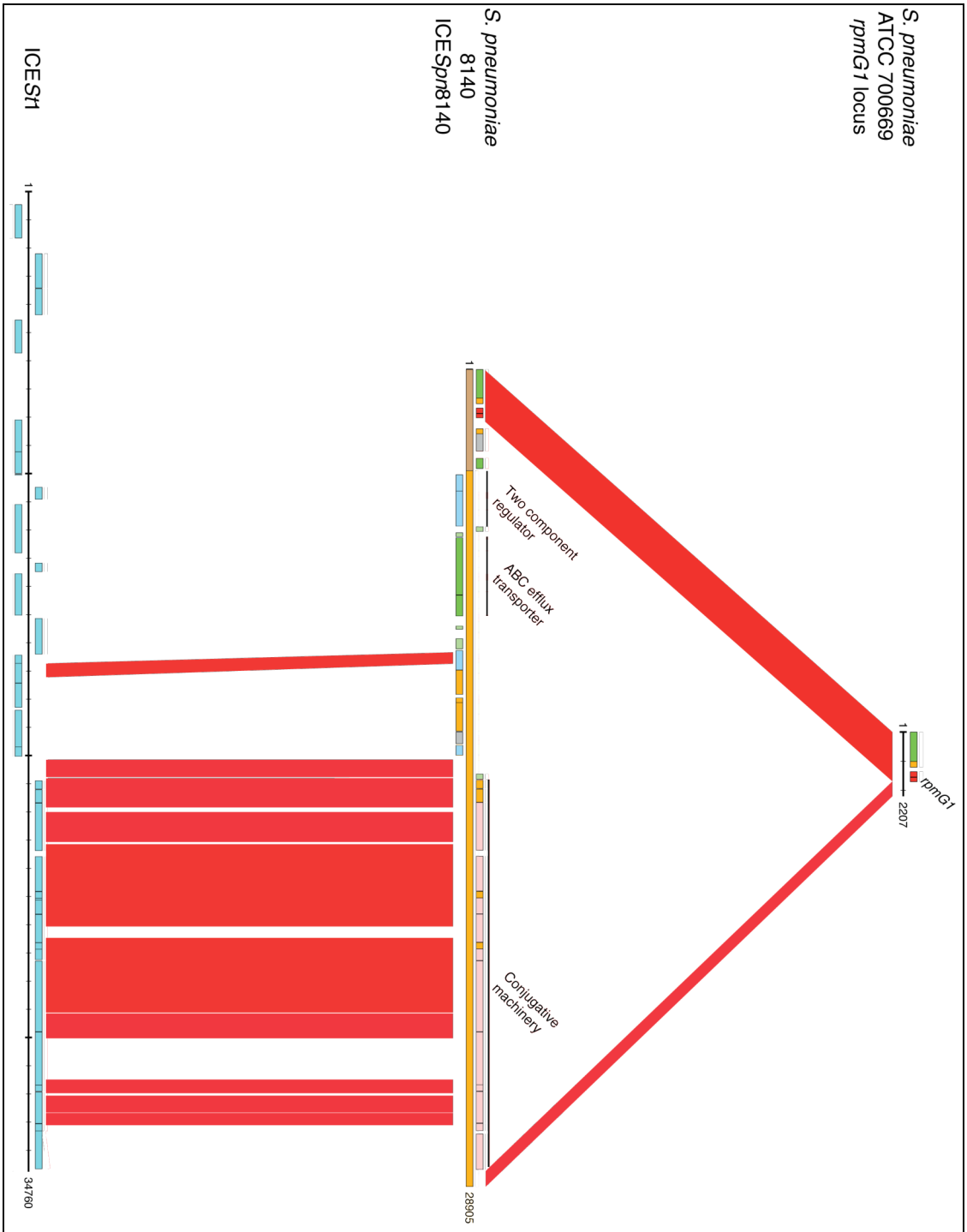

Figure 4.13 The putative ICE*St*1-like element ICE*Sp*8140. The two ends of this small conjugative element were assembled *de novo* from Illumina reads. The brown and orange boxes covering the scale line indicate scaffolds produced by assembly of Illumina reads, which have been ordered against the reference sequence. Similarity between the sequences is indicated as described in Figure 4.11. The element inserts shortly downstream of *rpmG1* and shares the conjugative machinery of ICE*St*1. However, the two elements have very dissimilar cargo complements, with ICE*Sp*8140 carrying a putative two component regulator and ABC transporter.

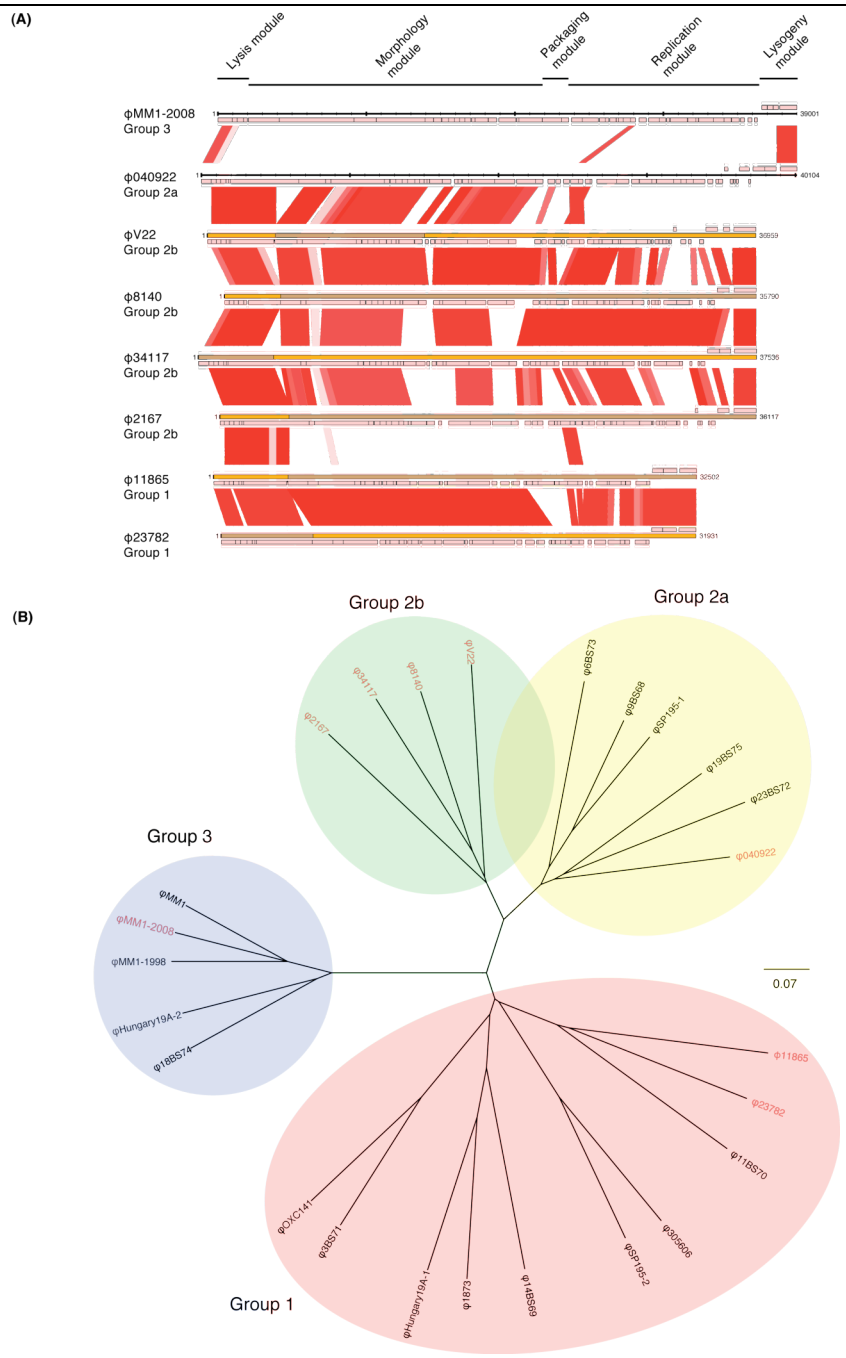

Figure 4.14 Prophage sequences in PMEN1. (A) Comparison of seven prophage assembled from PMEN1 isolates to prophage ϕMM1-2008 from the reference genome, *S. pneumoniae* ATCC 700669. The modular structure of such pneumococcal prophage genomes are indicated along the top. Alternate orange and brown boxes covering the scale line indicate scaffolds produced by assembly of Illumina reads. Red bands between sequences indicate TBLASTX matches between prophage in the same orientation, with the colour intensity reflecting the strength of the sequence similarity. (B) Clustering of pneumococcal prophage. All available identified prophage from *S. pneumoniae* genomes were included in this analysis. The names of prophage assembled from PMEN1 isolates are coloured red, showing that representatives from every group of temperate pneumococcal phage can be identified in the PMEN1 lineage. In addition, the previously unidentified group 2b prophage that insert into the *comYC* gene can be found in the isolates samples in this study.

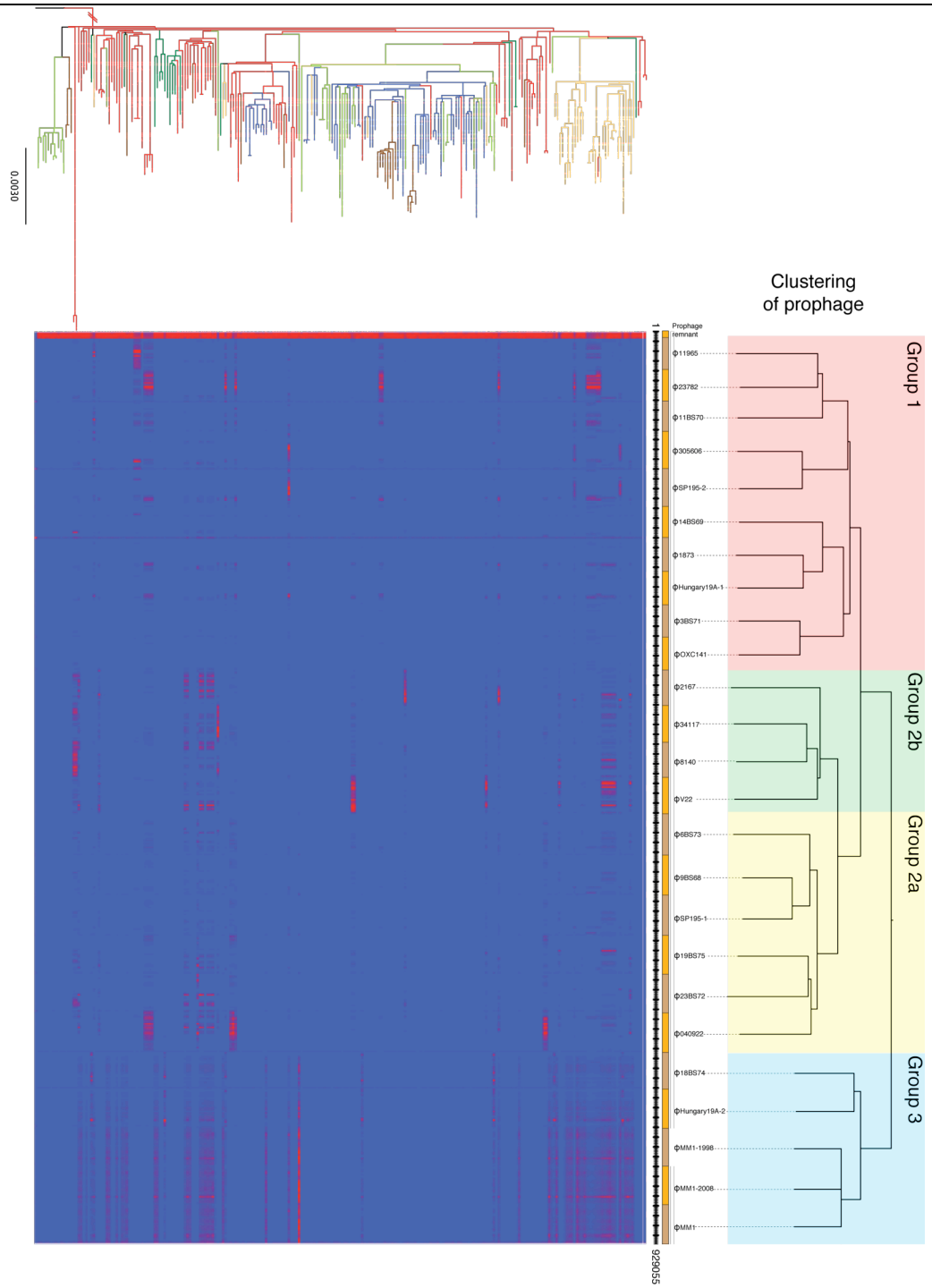

Figure 4.15 Distribution of prophage sequences between PMEN1 isolates. Across the top, the clustering of the prophage sequences is shown (Figure 4.14). The first sequence, not including in the clustering, is the prophage remnant found in the *S. pneumoniae* ATCC 700669 genome. This is conserved among all sequenced members of the lineage, and therefore used to standardize each heatmap, displayed as described in Figure 4.9, to the level of coverage obtained for each sample. The level of nucleotide sequence similarity between the closely related group 3 phage leads to a large amount of cross mapping of reads between these five prophage, hence the mean level of signal across these elements is typically lower than that for prophage in groups 1 or 2.

4.3 Discussion

The ability to distinguish vertically acquired substitutions from horizontally acquired sequences is crucial to successfully reconstructing phylogenies for recombinogenic organisms such as *S. pneumoniae*. Phylogenies are in turn essential for detailed studies of events such as intercontinental transmission, capsule type switching, and antibiotic-resistance acquisition. Although current epidemiological typing methods have indicated that recombination is frequent among the pneumococcal population, they cannot sufficiently account for its impact on relations between strains at such high resolution. Only the availability of such a sample of whole-genome sequences makes it possible to adequately reconstruct the natural history of a lineage. The base substitutions used to construct the phylogeny have accumulated over about 40 years and occur, on average, once every 15 weeks. Recombinations happen at a rate about ten fold more slowly but introduce a mean of 72 SNPs each. The responses to the different anthropogenic selection pressures acting on this variation are distinct. The apparently weak selection by aminoglycosides and chloramphenicol has led to the occasional deletion of loci encoding resistance to these antibiotics. By contrast, resistance to macrolide antibiotics has been acquired frequently throughout the phylogeny, with selection strong enough to drive supplementation or replacement of the resistance afforded by the *mef* efflux pump with the broader-range resistance provided by *ermB*-mediated target modification. The response to vaccine selection is different and involves the depletion of the resident population before it can respond to the selection pressure and thereby opens the niche to isolates that already expressed nonvaccine serotypes. This is likely to reflect the high host population coverage of PCV7 in the USA, as opposed to macrolides or other antibiotics, and the relative likelihood of the recombination events that underlie these responses.

Over a few decades, this single pneumococcal lineage has acquired drug resistance and the ability to evade vaccine pressure multiple times, demonstrating the remarkable adaptability of recombinogenic bacteria such as the pneumococcus. PMEN1 is, nevertheless, only one lineage of this pathogen. Our relative ignorance of the forces that affect bacterial evolution over the long term is illustrated by BM4200 (Buu-Hoi and Horodniceanu, 1980), a multidrug-resistant serotype 23F isolate of ST1010 sequenced as the outgroup for this analysis (Figure 4.1). This isolate dates to 1978 but, despite its apparent similarity to PMEN1 strains, has been found very rarely

since then. Hence, the multidrug-resistant phenotype is not sufficient to guarantee success, suggesting that the nature of the resistances themselves, or other factors in the genotype, may be important for the relative prevalences of these two clones.

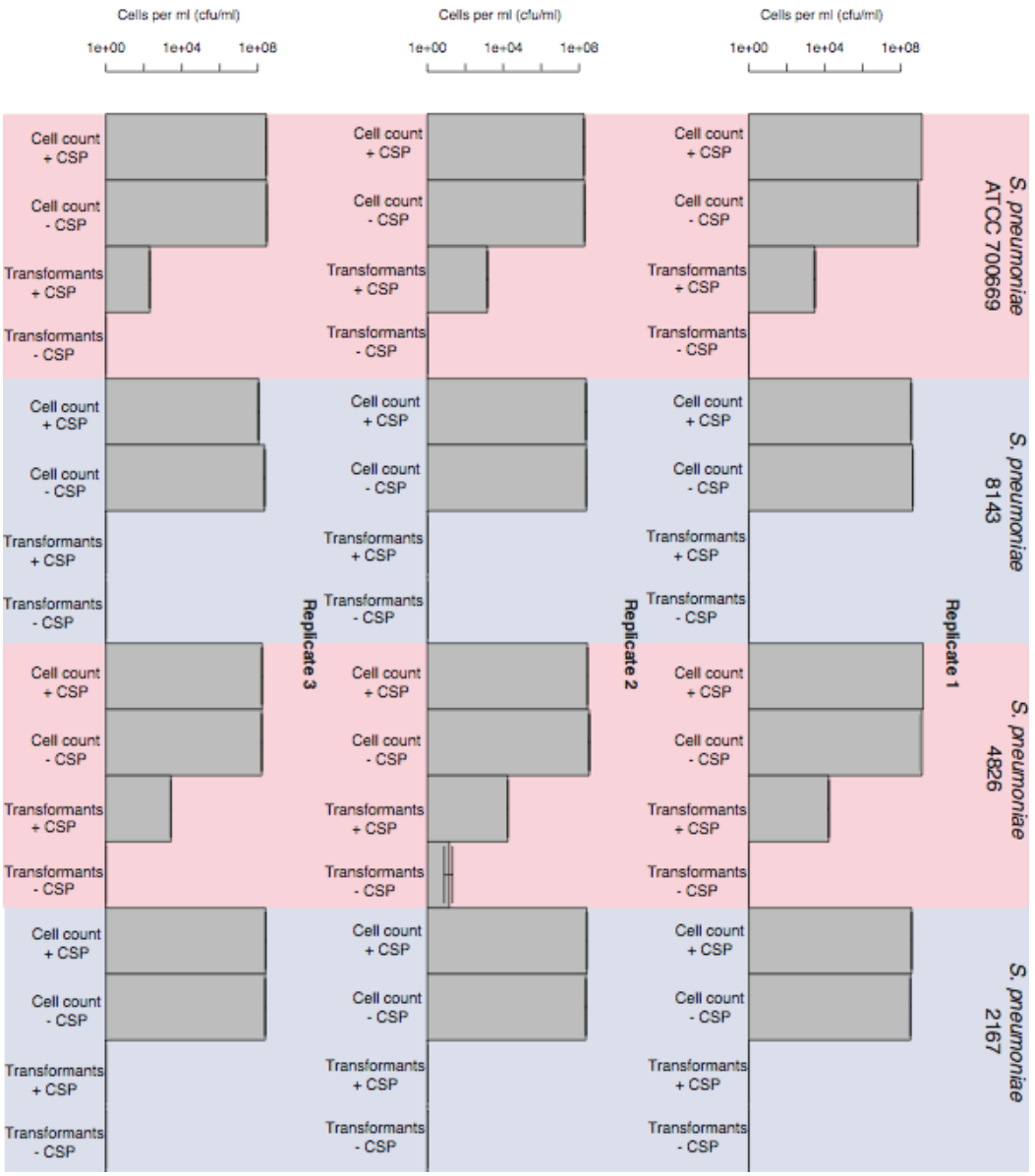

Figure 4.16 Testing the competence of isolates carrying prophage inserted in *comYC*. Four strains (two lysogens carrying group 2b prophage, *S. pneumoniae* 2167 and 8143, shaded blue, and related strains with intact *comYC* genes, *S. pneumoniae* ATCC 700669 and 4826, shaded red) were assayed for their ability to incorporate a kanamycin resistance marker into the *cps* locus in a competence stimulating peptide (CSP)-dependent manner (see Methods). The bars show either the total cell count or the number of transformants, as indicated, in colony forming units per milliliter, with error bars showing the standard error of the mean ($N = 3$ for each measurement). No CSP-dependence in the total population of cells was observed, suggesting that the prophage, if they are active, are not induced by CSP. Kanamycin-resistant colonies were only observed in experiments using the strains with intact *comYC* genes, suggesting that prophage that disrupt the *comYC* gene abrogate the competence of the pneumococcus and prevent it taking up DNA from the environment.

Table 4.1 Convergence of the PMEN1 phylogeny. Quantification of the similarity between phylogenies produced by subsequent iterations of the algorithm used to detect recombination and construct the tree. The branch score difference takes the length of branches into account, while the Robinson-Foulds metric is based only on the topology of the tree.

Comparison	Branch score distance	Robinson-Foulds metric
Iteration 1 vs Iteration 2	1.00	181
Iteration 2 vs Iteration 3	6.49×10^{-3}	41
Iteration 3 vs Iteration 4	1.44×10^{-3}	19
Iteration 4 vs Iteration 5	9.86×10^{-4}	7
Iteration 5 vs Iteration 6	3.87×10^{-4}	5
Iteration 6 vs Iteration 7	5.39×10^{-4}	14
Iteration 7 vs Iteration 8	7.34×10^{-4}	11
Iteration 8 vs Iteration 9	4.83×10^{-4}	14

Table 4.2 CDSs frequently disrupted by mutations in the PMEN1 phylogeny. CDSs affected by a significantly high number of disruptive mutations in the PMEN1 phylogeny.

CDS	Gene Name	Product	Length (bp)	Disruptions	*p* value
SPN23F17730	-	Putative *psrP* glycosyltransferase	905	24	0
SPN23F01290	*pspA*	Pneumococcal surface protein A (pseudogene)	2176	15	3.33×10^{-16}
SPN23F15600	-	Putative phage protein	317	5	4.08×10^{-8}
SPN23F19760	-	Lantibiotic processing protease	1739	8	4.76×10^{-8}
SPN23F15300	*hol*	Antiholin	332	5	5.13×10^{-8}
SPN23F06290	-	Membrane protein	752	6	9.60×10^{-8}
SPN23F05270	-	IS1239 transposase (pseudogene)	449	5	2.26×10^{-7}
SPN23F12860	-	Uncharacterised ICE protein	362	4	3.92×10^{-6}
SPN23F21150	-	Putative DNA binding protein	440	4	8.41×10^{-6}
SPN23F14790	-	IS1239 transposase	1007	5	1.13×10^{-5}
SPN23F17840	-	Transposase	479	4	1.17×10^{-5}