4. Dynamics and impact of homologous recombination on the evolution of *L. pneumophila*

Declaration of work contributions

Julian Parkhill, Simon Harris and Timothy Harrison supervised this work. Massimo Mentasti performed culture and DNA extraction of all newly sequenced isolates. Leonor Sánchez-Busó contributed to the detection of MGEs and the inference of recombination donors. Jukka Corander performed the Bayesian analysis of population structure (BAPS) clustering. I conducted the remaining bioinformatics analyses and generated all the figures.

Publication

The following work has been prepared for publication:

David, S., Sánchez-Busó, L., Harris, S. R., Harrison, T. G. & Parkhill, J. Dynamics and impact of homologous recombination on the evolution of *Legionella pneumophila*.

4.1 Introduction

While all bacteria reproduce clonally, some also import DNA from other organisms into their chromosomes in a process known as recombination or horizontal gene transfer. There are three known mechanisms through which this can take place including transduction (*via* phage infection), conjugation (*via* direct contact), and transformation (*via* the uptake of naked DNA from the environment). The imported DNA can comprise either novel genes that are new to the recipient genome (non-homologous recombination), or can replace an equivalent segment of the genome (homologous recombination). The latter, which is the main focus of this thesis chapter, results in the replacement of genes with alternative allelic variants and requires the DNA to be highly similar, and possibly identical, at both ends of the fragment (Majewski & Cohan, 1998). For this reason, homologous recombination usually occurs between closely related bacteria.

The importance of recombination in bacterial evolution first became clear through the analysis of MLST data, which showed that phylogenetic trees constructed from individual MLST genes were often incongruent (Feil *et al.*, 2001). These analyses also predicted that the rate of homologous recombination varies considerably between different species (Perez-Losada *et al.*, 2006). There are a number of hypotheses regarding why bacteria engage in homologous recombination (Vos, 2009). One explanation is that recombination is used as a mechanism by which DNA damage, such as double-strand breaks, can be repaired using foreign DNA as a template (Michod *et al.*, 2008). Another is that it is a side effect of DNA uptake for use as an energy source or for DNA synthesis from nucleotide precursors (Redfield, 1993). Finally, the ability of recombination events to remove deleterious mutations and rapidly introduce combinations of advantageous mutations could mean it increases the efficiency of natural selection and is selectively maintained (Narra & Ochman, 2006).

In recent years, the availability of WGS data from multiple closely related isolates has enabled homologous recombination to be studied in great detail in species such as *Streptococcus pneumoniae* (Croucher *et al.*, 2011; Chewapreecha *et al.*, 2014), *Chlamydia trachomatis* (Harris *et al.*, 2012) and *Neisseria meningitidis* (Kong *et al.*, 2013). These

studies have confirmed that homologous recombination plays an important role in the evolution and adaptation of important bacterial pathogens, for example by facilitating vaccine escape (Croucher *et al.*, 2011) and antibiotic resistance (Chewapreecha *et al.*, 2014) in *S. pneumoniae*.

L. pneumophila was first reported to have a clonal population structure based on mult-locus enzyme electrophoresis (MLEE) analysis (Selander *et al.*, 1985). However, the three primary mechanisms of bacterial recombination (conjugation, transduction and transformation) have since all been described in *L. pneumophila* (Dreyfus & Iglewski, 1985; Mintz & Shuman, 1987; Stone & Abu Kwaik, 1999), and thus it was unsurprising when later studies reported its occurrence. Indeed, an early genomic study of the first sequenced genomes of *L. pneumophila* showed that recombination events are frequent and predicted that it can involve large chromosomal fragments of over 200kb (Gomez-Valero *et al.*, 2011). More recently, a study of closely related genomes belonging to ST578 demonstrated that recombination accounts for over 98% of the SNPs detected within the lineage and is therefore a dominant force in *L. pneumophila* evolution (Sanchez-Buso *et al.*, 2014). These findings are concordant with those made in the first results chapter of this thesis whereby over 96% of the SNPs found in STs 1, 23, 37 and 62 were found to be imported *via* recombination. Interestingly though, no recombination was detected in the ST47 lineage. In both the published study by Sanchez-Buso *et al.* (2014) and *Chapter 3*, the relative contribution of homologous and non-homologous recombination is not disentangled, nor is the impact of recombination on the adaptation and evolution of *L. pneumophila* studied in detail.

Therefore, the first aim of this results chapter is quantify the relative impact of homologous and non-homologous recombination on the evolution of major disease-associated lineages of *L. pneumophila* including STs 1, 23, 37, 62 (as studied in *Chapter 3*), ST578 (as studied by Sanchez-Buso *et al*., 2014) and an additional disease-associated lineage comprising ST42 isolates. The chapter subsequently focuses solely on homologous recombination, and seeks to characterise the regions that have been imported *via* this process. It explores whether there are "hotspots" in the genome where homologous recombination events are more likely to be selectively maintained, which could provide insight into important selection pressures of *L. pneumophila*. Finally, by

inferring the donor lineages of recombined regions, the last aim of the chapter is to examine the extent to which homologous recombination occurs between the two *L. pneumophila* subspecies and between and within major clades of the *L. pneumophila pneumophila* subspecies. This will provide further insight into the dynamics of genomic flux within the *L. pneumophila* species and reveal the extent to which major disease-associated STs from different clades are able to exchange DNA. As was suggested in *Chapter 3*, this process could represent an important mechanism by which diverse strains can adapt rapidly to new niches.

4.2 Materials & Methods

4.2.1 Bacterial isolates

L. pneumophila isolates belonging to six major disease-associated lineages are primarily used in this study (*n*=290). These include 81 ST1, or ST1-derived, isolates (including 71 used in *Chapter 3* and 10 from a study by Sanchez-Buso *et al.* (2014)), 42 ST23 isolates (including 37 used in *Chapter 3* and 5 from another study by Sanchez-Buso *et al.* (2016)), 72 ST37 and 35 ST62 isolates (all of which were used in *Chapter 3*), 46 ST578 (including one published by D'Auria *et al.* (2010) and 45 published by Sanchez-Buso *et al.* (2014)) and 15 ST42 isolates (including 2 previously published isolates (Schroeder *et al*., 2010; Underwood *et al.*, 2013) and 13 newly sequenced isolates). Those additional isolates belonging to these six STs that are not used in *Chapter 3* are listed in **Appendix Table 3**. A further 246 *L. pneumophila* isolates, listed in **Appendix Table 4** and which belong to a range of STs, were also used in the inference of recombination donors. Importantly, these include a set of previously published genomes, which were selected for sequencing using MLST data with the aim of encompassing as much of the species diversity as possible (Underwood *et al.*, 2013). Culture, DNA extraction and sequencing of all newly sequenced isolates were performed as described in *Chapter 2 (Materials & Methods).* Accession numbers or references for all sequence data are provided in **Appendix Tables 3 and 4**.

4.2.2 Reference genomes

Isolates belonging to each of the six disease-associated STs (1, 23, 37, 42, 62 and 578) were mapped to a reference genome of the same ST to enable each lineage to be studied at a high resolution. The complete genomes of Paris (Cazalet *et al.*, 2004) and Alcoy (D'Auria *et al.*, 2010) were already available for ST1 and ST578, respectively. Reference genomes were generated for the remaining four STs by sequencing a representative isolate from each ST on the PacBio RSII sequencer at the WTSI. The isolates chosen were EUL 28, EUL 120, EUL 165 and H044120014 belonging to STs 23, 42, 37 and 62, respectively. 1-2μg of DNA from each isolate was sheared using a 26G blunt-ended needle (ThermoFisher, UK) and used in library preparation according to the manufacturer's protocol. The P4 DNA polymerase was used with C2 chemistry to perform the sequencing. *De novo* assemblies were produced from the sequence reads using HGAP.3 (Pacific Biosciences). Assemblies that consisted of a single chromosomal contig were circularised using the overlapping sequence at the two ends and the start of the genome was set to the beginning of the *dnaA* gene. Each genome was subsequently confirmed by mapping Illumina sequence reads from each of the isolates to the PacBio assembly. Sequencing statistics for the four PacBio reference genomes are provided in **Appendix Table 5**. They were annotated using an in-house pipeline at the WTSI, which uses Prokka (Seemann, 2014), together with the complete genomes of Paris (ST1) and Alcoy (ST578).

Repetitive regions over 100bp were detected in the six reference genomes using repeat-match from MUMmer v3.0 (Kurtz *et al.*, 2004) (**Appendix Table 6**). This was performed by Leonor Sánchez-Busó.

4.2.3 Mapping, recombination detection, phylogenetic analysis and BAPS clustering

Sequence reads from all isolates belonging to the six major disease-associated STs were mapped to the appropriate reference genome of the same ST using SMALT v0.7.4 (available at: http://www.sanger.ac.uk/science/tools/smalt-0). All isolates used in the study (n=536) were also mapped to the Paris (ST1) reference genome (Cazalet *et al.*, 2004) in order to study the species-wide phylogenetic structure. An in-house pipeline at

the WTSI was used to call bases and identify SNPs as described in *Chapter 2 (Materials & Methods)*.

Recombined regions were detected in the alignments of the six disease-associated STs using Gubbins (Croucher *et al.*, 2015). Phylogenetic trees of these lineages were generated as described in *Chapter 2 (Materials & Methods)*, firstly using all SNPs to later allow ancestral sequence reconstruction (see *4.2.6 Inference of recombination donors*), and secondly using only the vertically inherited SNPs. A phylogenetic tree of the total 536 isolates was constructed using all the detected SNPs, as the high diversity renders recombination detection impossible. The alignment of all 536 genomes against the Paris reference genome was also used to group the isolates into clusters using hierBAPS (Cheng *et al.*, 2013), which was performed by Jukka Corander.

4.2.4 Detection of MGEs

The annotation files for each of the six reference genomes were parsed to detect genes annotated as "integrase", "transposase", "recombinase", "phage", "lvrA", "csrA", "HTX", "helix-turn-helix", "xre", "conjugal", "conjugation", "tra", "trb", "vir" and "mobile". Both the published annotation files of the Paris (ST1) and Alcoy (ST578) complete genomes and those generated using the in-house pipeline at the WTSI were used. However, the new annotations were only considered when the original one was a "hypothetical protein" in order to respect experimentally proven annotations. Plots showing the mapping coverage of all isolates in the six STs against the corresponding reference genome were also evaluated. Regions over 8kb with no coverage and that did not match repetitive regions were considered as potential mobile regions. These analyses were performed by Leonor Sánchez-Busó.

Other software to detect MGEs was also used including AlienHunter (Vernikos & Parkhill, 2006) and Island Viewer, the latter of which incorporates IslandPick, IslandPath-DIMOB and SIGI-HMM (Langille & Brinkman, 2009). However, these results were discarded due to major incongruences between them. Finally, manual curation of all predicted MGEs was performed using Artemis v15.0.0 (Carver *et al.*, 2012) (**Appendix Table 6**).

4.2.5 Identification of homologous recombination hotspots

In each of the six lineages, any predicted recombined regions that overlap with either repetitive regions or MGEs in the reference genome were identified and discarded for the majority of the analysis in this study, leaving only putative homologous recombination regions. An in-house script was used to calculate the number of times each gene had been involved in a homologous recombination event. Recombination "hotspots" were defined as genes with a recombination frequency above the 95th percentile observed in that particular ST.

4.2.6 Inference of recombination donors

A custom genome BLAST database (BLAST v2.2.30+) (Camacho *et al.*, 2009) was constructed using *de novo* assemblies from all 536 *L. pneumophila* isolates used in this study. The method by which *de novo* assemblies were generated is described in *Chapter 2 (Materials & Methods)*. Homologous recombination regions were extracted from the ancestral sequences inferred from the nodes of the phylogenetic trees (constructed prior to recombination removal) using PAML 4 (Yang, 2007). The reconstructed recombined regions were used as query sequences in BLAST searches against the custom genome database and the National Center for Biotechnology Information (NCBI) non-redundant nucleotide database. The resulting hits were filtered to remove those against isolates that are descended from the branch in which the recombination event was detected. Of the remaining hits, the one with the highest bit score was considered as the potential donor, provided it had a minimum of 99% nucleotide identity to the recombined fragment, and matched at least 50% of the fragment length.

4.3 Results

4.3.1 Contribution of homologous recombination to *L. pneumophila* diversity

To investigate the relative contribution of homologous recombination to diversity in each of the six major disease-associated STs (1, 23, 37, 42, 62 and 578), isolates were

first mapped to a reference genome of the same ST. Gubbins was used to detect recombined regions in each of the six alignments and construct a phylogenetic tree based on the vertically inherited SNPs located outside of these regions. As found in *Chapter 3* and a study by Sanchez-Buso *et al.* (2014), over 96% SNPs in STs 1, 23, 37, 62 and 578 are predicted to be derived from recombination events (**Table 4.1**). Furthermore, 99.0% SNPs in the ST42 lineage, which has not been studied previously, were also found in predicted recombined regions. The remaining number of vertically inherited SNPs in each of these lineages ranges from just 94 (ST42) to 1006 (ST1) (**Table 4.1**).

Table 4.1. Number of SNPs detected within each of the six disease-associated STs.

ST	Number of isolates	Total number of SNPs	Number (and %) of SNPs in recombined regions	Number (and %) of vertically inherited SNPs
1	81	73,044	72,038 (98.6%)	1006 (1.4%)
23	42	44,886	44,720 (99.6%)	166 (0.4%)
37	72	17,776	17,300 (97.3%)	476 (2.7%)
42	15	9,256	9,162 (99.0%)	94 (1.0%)
62	35	47,684	47,372 (99.3%)	312 (0.7%)
578	46	3,678	3,559 (96.8%)	119 (3.2%)

Any recombined regions that overlapped with either predicted MGE regions or repeat regions were subsequently identified, in order to determine the contribution of only homologous recombination to *L. pneumophila* diversity. It was found that between 33.0% (ST62) and 80.0% (ST578) of all SNPs are predicted to be in regions derived from homologous recombination events (**Table 4.2 and Figure 4.1**). However, the mean length of each individual genome affected by this process varies between just 1.2% (ST42/578) and 3.9% (ST1) (**Table 4.2**). It should be noted that the number of SNPs from homologous recombination might be slightly overestimated (and the number of *de novo* mutations slightly under estimated) since *de novo* mutations may have occurred on top of recombination events. However, the error should be no more than 1.2-3.9%, in

proportion with the average length of genome affected by homologous recombination events.

Table 4.2. Contribution of homologous recombination to the diversity of the six major disease-associated STs.

ST	Number of homologous recombination events	Number of SNPs in homologous recombination regions per vertically inherited SNP	Number of homologous recombination events per vertically inherited SNP (r/m ratio)	Mean length of sequence (and %) of each individual genome affected by homologous recombination (bp)	Total length (and %) of the reference genome affected by homologous recombination across all isolates (bp)
1	198	56.2	0.20	135,208 (3.9%)	1,430,288 (40.8%)
23	44	93.8	0.27	51,242 (1.5%)	520,584 (14.8%)
37	13	20.8	0.03	105,051 (3.0%)	251,988 (7.3%)
42	11	41.3	0.12	41,747 (1.2%)	120,545 (3.51%)
62	48	50.5	0.15	66,559 (1.9%)	456,451 (12.9%)
578	23	24.6	0.19	42,138 (1.2%)	204,114 (5.8%)

In each of the six lineages, the relative number of homologous recombination events to vertically inherited mutations (r/m ratio) was calculated per branch of each phylogenetic tree (**Figure 4.2**). Across all branches, the r/m ratio ranged from 0.03 (ST37) to 0.27 (ST23), indicating that recombination events have occurred less frequently than vertically inherited mutations in all six lineages, despite bringing in between 20.8 (ST37) and 93.8 (ST23) times as many SNPs (**Table 4.2**). The r/m ratios also differ significantly between lineages (Kruskal-Wallis test, $p<0.05$), highlighting different rates of recombination in the six major disease-associated STs.

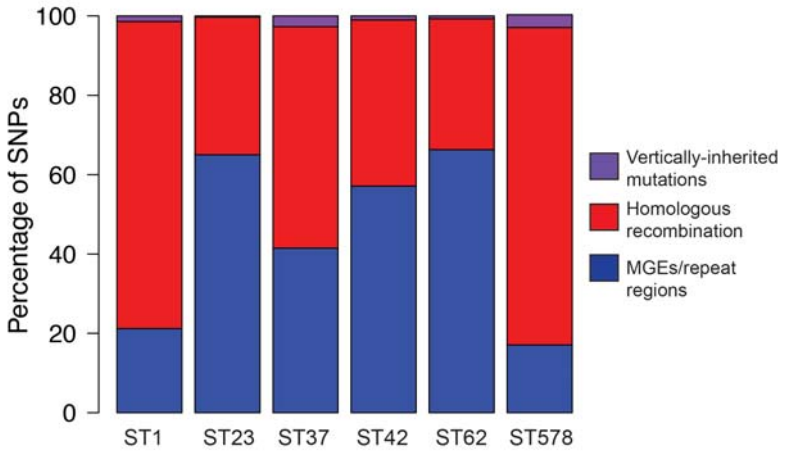

Figure 4.1 Generation of diversity in the six major disease-associated STs. The percentage of SNPs that are predicted to be derived from vertically inherited mutations, homologous recombination or from MGEs (i.e. non-homologous recombination) and repeat regions.

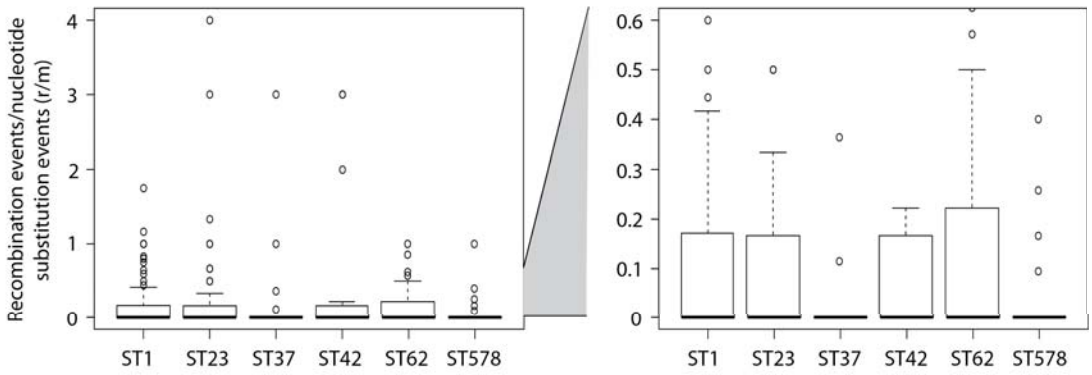

Figure 4.2. Relative frequency of homologous recombination events and vertically inherited mutations. Boxplots showing the number of homologous recombination events detected per vertically inherited SNP (r/m) on each of the branches of the phylogenetic trees belonging to the six STs.

To determine the relative impact of vertically inherited mutations and homologous recombination events on the coding sequence, the types of changes caused by the two processes were analysed. Vertically inherited mutations resulted in approximately

twice as many non-synonymous SNPs than synonymous SNPs, a result that is expected by chance when mutations occur at random in the genome and before selection has time to act on all but the most deleterious mutations (**Figure 4.3**). Interestingly though, the results are reversed for homologous recombination events, which result mostly in synonymous mutations (**Figure 4.3**). However, this observation is also not unexpected given that variants in sequences that are horizontally transferred between different lineages will have been subjected to a longer period of evolution and selection, which has purged harmful, non-synonymous mutations. The same phenomenon has also been observed in a previous study by Castillo-Ramirez *et al.* (2011). Furthermore, fewer SNPs that result in a stop codon are brought in by homologous recombination events than by vertically inherited mutations (**Figure 4.3**), which can also be explained by this process.

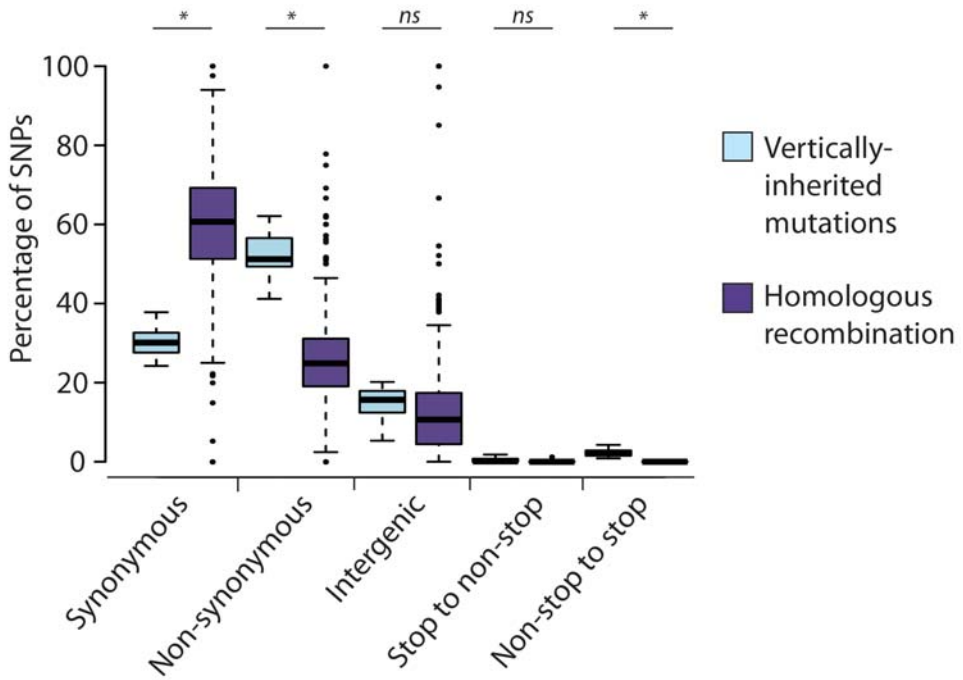

Figure 4.3. Types of change introduced by vertically inherited mutations and homologous recombination. Boxplots showing the percentage of SNPs per branch, derived from either vertically inherited mutations or homologous recombination that are synonymous, non-synonymous, intergenic, or result in a change from a stop to non-stop codon or a non-stop to stop codon. Statistically significant differences as determined by a Student's paired t-test are indicated by an asterisk. ns – not significant

The lengths of the recombined regions are exponentially distributed (rate of decay=7.52 x 10^{-5} bp^{-1}), with the majority of events being small (<10,000bp) and large events occurring relatively infrequently (**Figure 4.4**). The median recombination fragment length in each of the six lineages varies from 5,613bp (ST578) to 12,757bp (ST37), while the largest predicted region is 94,790bp (ST37).

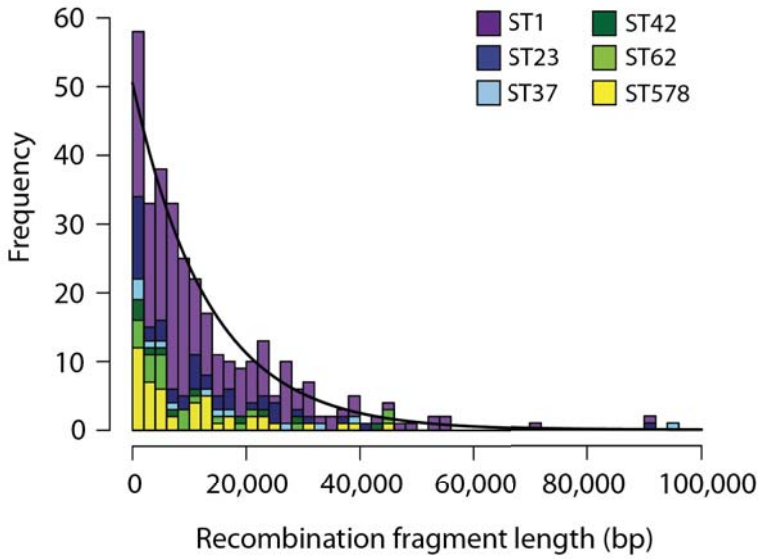

Figure 4.4. Size of detected homologous recombination regions in the six STs. An exponential decay curve (black line) is fitted to the distribution and the rate of decay is 7.52 x 10^{-5} bp^{-1}.

4.3.2 Hotspots of homologous recombination in *L. pneumophila*

To identify genomic regions associated with a high number of homologous recombination events, the number of events that overlap with each gene was calculated with respect to the reference genomes of the six STs. Hotspot regions were defined as genes with equal to or greater than the 95% recombination frequency detected in each lineage, and which were involved in at least two recombination events. This accounted for the recombination frequency, population size and diversity of each lineage. Based on these criteria, the minimum number of recombination events that a gene must have been involved in to be considered within a hotspot region was four events in the ST1

lineage and two events in the remaining five STs. A total of 32 hotspot regions were defined, including at least one in all six STs (**Table 4.3 and Appendix Table 7**). The most notable hotspot regions were observed in the ST1 lineage, which is also predicted to contain the highest number of homologous recombination events. A total of ten hotspot regions were defined in the ST1 lineage and, remarkably, one region contains genes that are predicted to have been involved in up to 27 recombination events and individual bases that have been involved in up to 25 recombination events (**Figure 4.5**). In the other five STs, the highest number of events affecting genes ranges from 2 (ST37/ST578) to 4 (ST42/ST62).

Table 4.3. Recombination hotspots in the six major disease-associated STs.

ST	Number of recombination events affecting genes	Genomic region (with respect to ST-specific reference genome)	Genes (defined in ST-specific reference genome and the Paris genome)
1	4-5	23,666-30,454	*lpp0019-lpp0024*
1	4	405,129-407,048	*lpp0356*
1	4-7	916,526-930,133	*lpp0819-lpp0830*
1	4	1,067,640-1,071,158	*lpp0961-lpp0963*
1	3-4	1,837,986-1,846,399	*lpp1640-lpp1645*
1	4-27	1,981,301-2,028,475	*lpp1761-lpp1794*
1	4	2,529,239-2,532,139	*lpp2198*
1	4	2,894,069-2,902,985	*lpp2543-lpp2550*
1	5-6	2,960,106-2,968,849	*lpp2595-lpp2604*
1	4	3,393,015-3,396,715	*lpp2977-lpp2979*
23	2	451,136-467,120	*ST23_00399-ST23_00417/lpp0453-lpp0471*
23	3	667,828-669,415	*ST23_00625-ST23_00626/lpp0668-lpp0669*
23	2	694,147-696,095	*ST23_00647-ST23_00648/lpp0690-lpp0691*
23	2	779,739-800,490	*ST23_00703-ST23_00713/lpp0748-lpp0758*

23	2-3	1,972,009-1,978,787	*ST23_01779-ST23_01781/lpp1768-lpp1770*
23	2	2,136,974-2,160,755	*ST23_01931-ST23_01947/lpp1925-lpp1942*
23	2	2,202,408-2,203,172	*ST23_01990/lpp1977*
23	2	2,865,124-2,877,303	*ST23_02606-ST23_02617/lpp2517-lpp2528*
23	2	3,365,161-3,367,970	*ST23_03044-ST23_03046/lpp2944-lpp2946*
37	2	1,326,840-1,329,791	*ST37_01205-ST37_01206/lpp1189-lpp1190*
42	2-4	2,830,437-2,845,009	*ST42_02559-ST42_02567/lpp2687-lpp2695*
62	2-4	284,602-299,923	*ST62_00255-ST62_00267/lpp0262-lpp0274*
62	2	310,597-311,994	*ST62_00277/lpp0285*
62	2	329,218-336,516	*ST62_00287-ST62_00292/lpp0305-lpp0310*
62	2	841,221-852,380	*ST62_00754-ST62_00764/lpp0756-lpp0766*
62	3	910,009-911,253	*ST62_00817/lpp0829*
62	2	918,029-919,546	*ST62_00823/lpp0835*
62	2	1,919,344-1,924,240	*ST62_01733-ST62_01736/lpp1667-lpp1670*
578	2	990,286-1,011,913	*lpa_01248-lpa_01273/lpp0880-lpp0902*
578	2	1,021,391-1,021,984	*lpa_01289/lpp0914*
578	2	1,693,186-1,696,080	*lpa_02154/lpp1435*
578	2	3,230,002-3,259,808	*lpa_04035-lpa_04063/lpp2815-lpp2839*

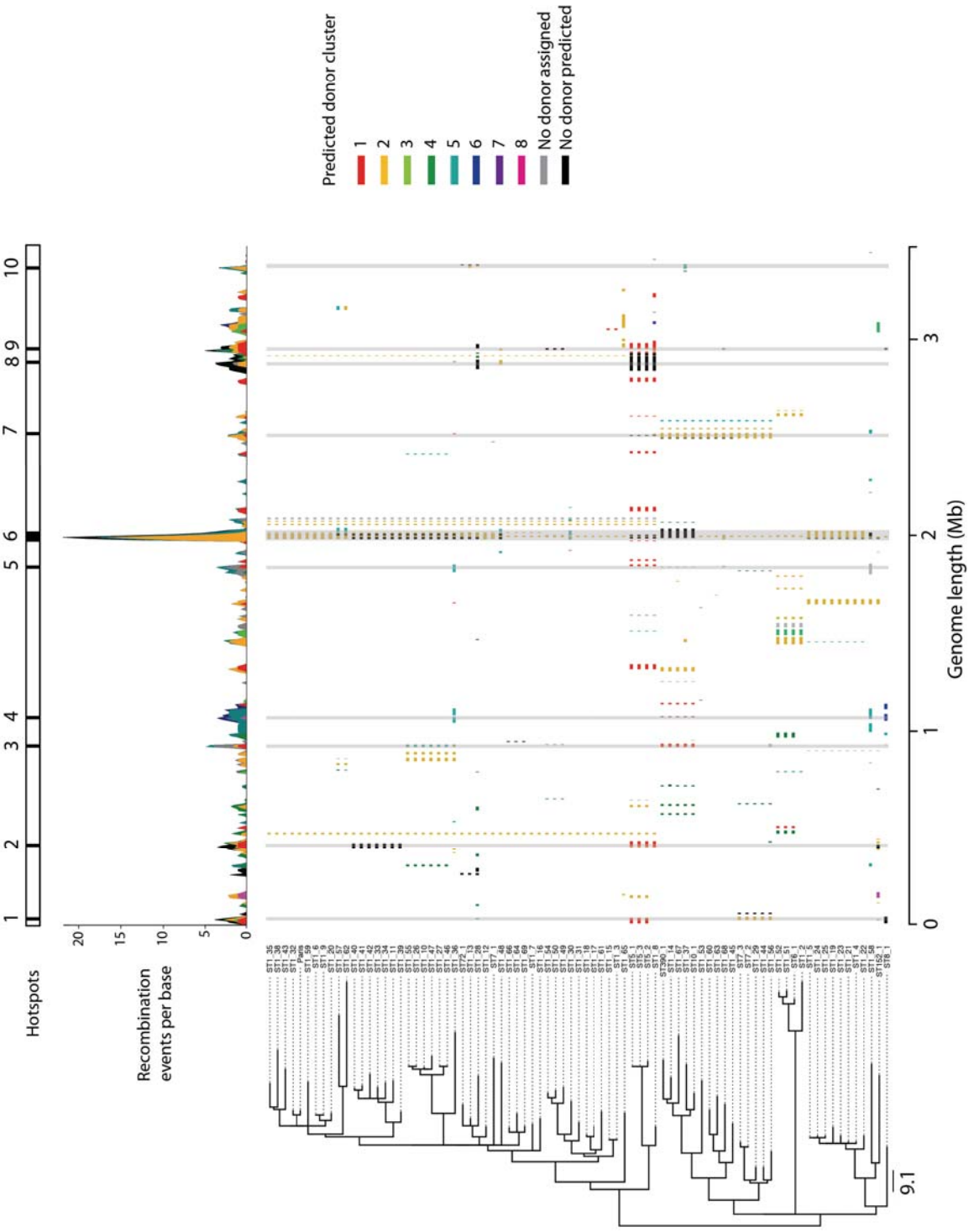
Predicted donor cluster
1
2
3
4
5
6
7
8
No donor assigned
No donor predicted
Hotspots
Recombination events per base
Genome length (Mb)
9.1

Figure 4.5. Homologous recombination events detected in the ST1 lineage (previous page). A phylogenetic tree, constructed using only vertically inherited mutations, is shown on the left and the scale indicates the number of SNPs. Homologous recombination events are shown by blocks, which are coloured according to the donor cluster from which they are predicted to have been derived (see *4.4.3 Inference of recombination donors*). The plot above shows the number of recombination events that have affected each base in the genome using a stacked visualisation to also indicate the number of events derived from different clusters. The ten genomic regions identified as recombination hotspots are marked at the top of the plot.

To further study the recombination hotspots, the gene content of the regions was analysed and compared between lineages. The most prominent hotspot (hotspot 6) identified in the ST1 lineage that contains genes involved in up to 27 recombination events is a 47,174bp region that ranges from *lpp1761* to *lpp1794* in the Paris (ST1) genome (**Figure 4.6 and Appendix Table 7**). The gene in this region that is predicted to have been involved in 27 events is *hemB*/*lpp1771*, a porphobilinogen synthase (delta-aminolevulinic acid dehydratase), which is an enzyme involved in the biosynthesis of tetrapyrroles. Since there is no obvious reason why this metabolic enzyme should be under a high selective pressure, the genes flanking this locus were also investigated. While the two immediate flanking genes (*lpp1770* and *lpp1772*) both encode "hypothetical proteins", *lpp1773*, which has been involved in 25 recombination events, has been shown to encode an outer membrane protein of *L. pneumophila* in a previous study (Khemiri *et al.*, 2008) and has high homology to the *fadL* gene conserved across many bacterial species. Interestingly, a *fadL*-like gene (*ST62_00760*; *lpp0762*) is also found within a recombination hotspot in the ST62 lineage, where it is involved in two recombination events, although it is found in a different part of the genome to the ST1 hotspot region. Furthermore, a smaller 6,778bp hotspot region in the ST23 region (*ST23_01779-ST23_01781; lpp1768-lpp1770*) overlaps with this hotspot region in the ST1 lineage. However, the region in the ST23 lineage centres on the gene, *ST23_01780*/*lpp1769*, which is involved in three recombination events and encodes the outer membrane protein assembly factor, BamA. Interestingly, *lpp1769* is involved in "just" 18 recombination events in the ST1 lineage, compared with *lpp1771* that is involved in 27.

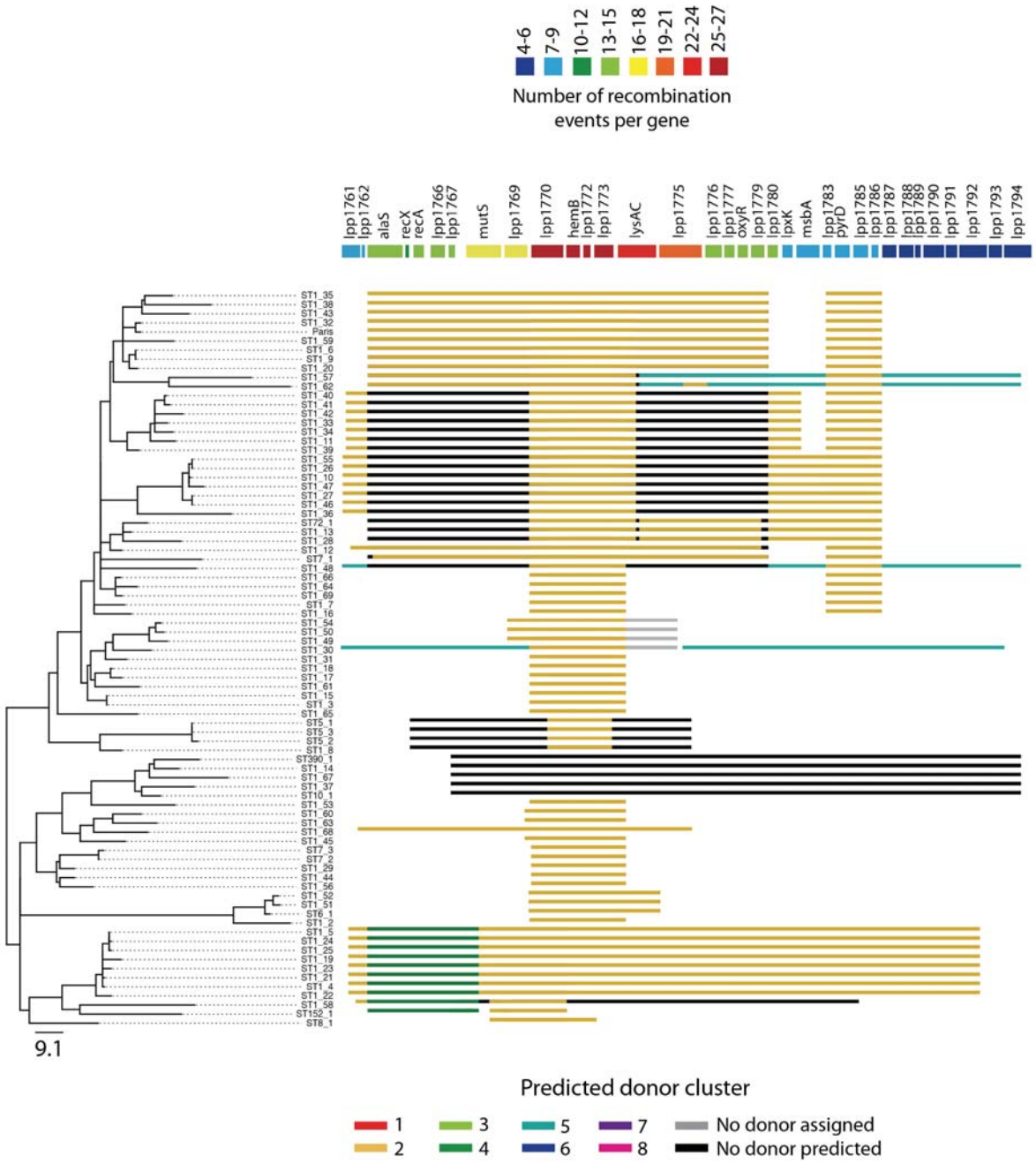

Figure 4.6. Hotspot 6 in the ST1 lineage. A maximum likelihood tree, constructed using only vertically inherited mutations, is shown on the left and the scale indicates the number of SNPs. The homologous recombination events are displayed as blocks, coloured according to the BAPS cluster from which they are predicted to be derived (see *4.4.3 Inference of recombination donors*). The genes shown at the top of the figure are coloured by the number of overlapping recombination regions.

The second most prominent hotspot (hotspot 3) in the ST1 lineage is a 13,607bp region that ranges from *lpp0819* to *lpp0830* in the Paris genome, and which contains genes affected by up to seven recombination events (**Figure 4.7 and Appendix Table 7**). This hotspot is fully contained within the lipopolysaccharide (LPS) locus that spans a region from *lpp0814* to *lpp0843*. Many of the genes in this hotspot region have been implicated in LPS core oligosaccharide biosynthesis including those belonging to the *rml* family, and O-antigen biosynthesis such as *neuA*, *neuB*, *neuC*, *wecA*, *wzt* and *wzm* (Lueneberg *et al.*, 2000). Interestingly, the genes affected by the highest number of recombination events are *wecA* but also *lpp0829a-c*, which are annotated as pseudogenes in the original annotation of the Paris genome (Cazalet *et al.*, 2004). All three genes encode "hypothetical proteins" although *lpp0829a* has a signal peptide and thus may be secreted, while *lpp0829b* has a pectin lyase fold, which has also been found in genes belonging to *L. longbeachae* and is thought to degrade the pectic components of plant cell walls. Furthermore, the ST62 lineage also has two genes from the LPS locus that are in hotspot regions. The first is *ST62_00817*, homologous to the three genes, *lpp0829a-c*, in the Paris genome and which has been involved in three recombination events. The second is *ST62_00823*, homologous to *lpp0835/rmlD* in the Paris genome, which has been involved in two events.

Another notable hotspot in the ST1 lineage is an 8,743bp region comprising genes from *lpp2595* to *lpp2604* (**Appendix Table 7**). The hotspot centres on the *lpp2599/tehB* gene, involved in six recombination events, and which encodes the tellurite resistance protein, TehB.

Across all six disease-associated STs, outer membrane proteins are commonly found within recombination hotspot regions (**Appendix Table 7**). Excluding those mentioned already (i.e. FadL and BamA), these include TolC (encoded by *ST23_00709*/*lpp0754*), involved in two recombination events in the ST23 lineage, and which has been implicated in the virulence of *L. pneumophila* (Ferhat *et al.*, 2009). Another is *lpa_01256*/*lpp0889*, also a TolC-like protein, which has been involved in two recombination events in the ST578 lineage. A small hotspot region in the ST37 lineage is immediately next to a known outer membrane protein (*ST37_01207*/*lpp1191*) described by Khemiri *et al.* (2008), and a hotspot region in the ST23 lineage is also very close to

the major outer membrane protein (*ST23_00628*/*lpp0671*). Furthermore, the *lpp0961* gene, involved in four recombination events in the ST1 lineage, encodes a protein homologous to AsmA, which is known to be involved in the assembly of outer membrane proteins in *E. coli*.

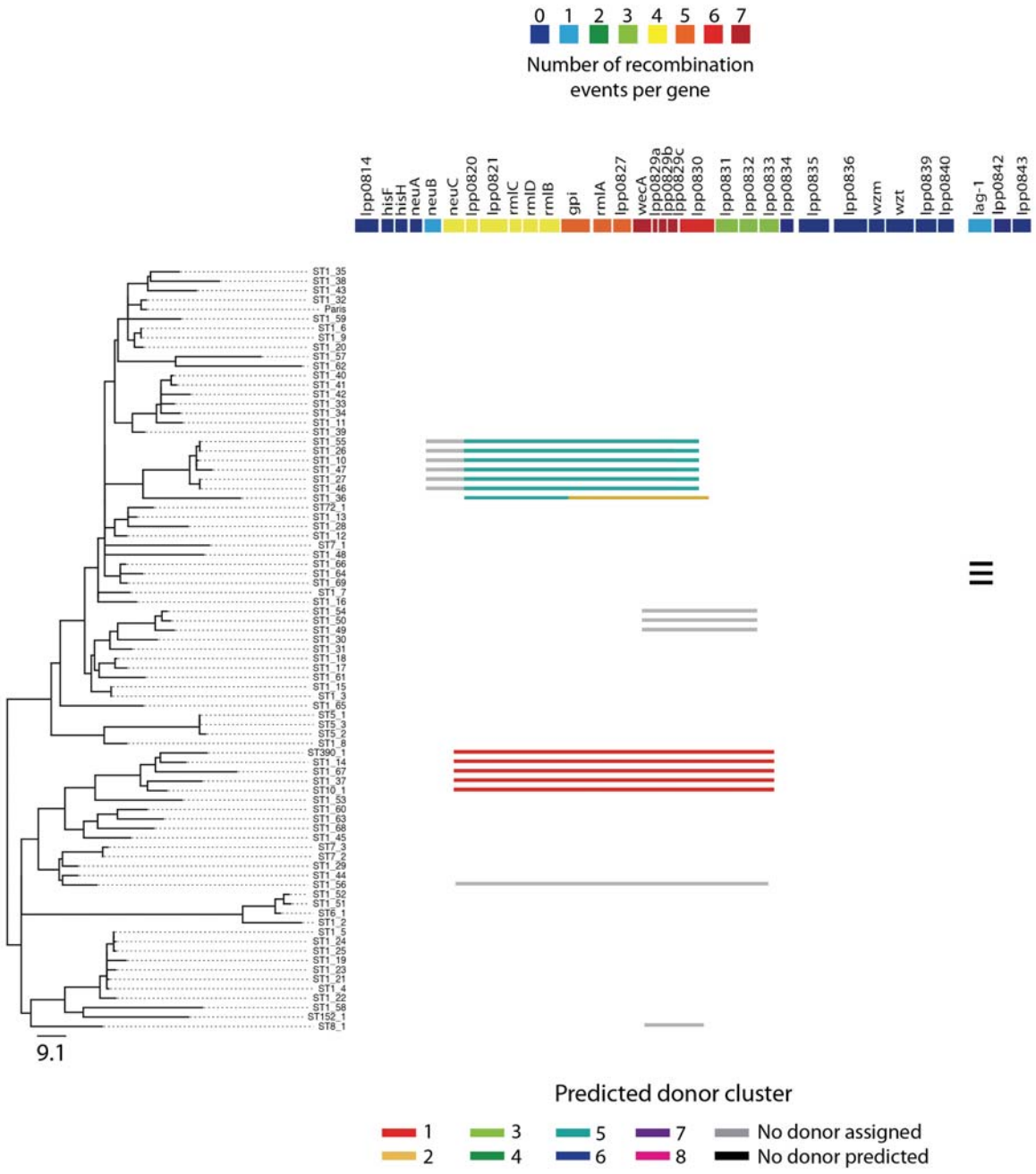

Figure 4.7. The LPS locus comprising hotspot 3 in the ST1 lineage. A maximum likelihood tree, constructed using only vertically inherited SNPs, is shown on the left and the scale indicates the number of SNPs. The recombination events are displayed as blocks, coloured according to

the BAPS cluster from which they are predicted to be derived. The genes within the LPS locus are shown at the top of the figure and coloured by the number of overlapping recombination regions.

A number of genes encoding putative or confirmed Dot/Icm effectors are also found within recombination hotspots across the different lineages (**Appendix Table 7**). These include *lpp0356*, involved in four recombination events in the ST1 lineage, which encodes an ankyrin repeat-containing protein that was originally found only in the Paris genome (Cazalet *et al.*, 2004). The *lpp2546* gene, which encodes the SdbB effector, has also been involved in four recombination events in the ST1 lineage. A further three ankyrin repeat-containing effector genes were identified within ST23 hotspots including *ST23_02606* (encoding LegA14), *ST23_00705* (encoding LegA8) and *ST23_00415* (encoding LegA7), all of which have been involved in two recombination events. Furthermore, the first described Dot/Icm effector, RalF, encoded by *ST23_01938*/*lpp1932*, was also found within a ST23 hotspot and predicted to have been involved in two recombination events.

Finally, while only 11 homologous recombination events were detected within the ST42 lineage, genes within one 14,572bp region have been affected by up to four recombination events. The hotspot region is centred on *ST42_02565*/*lpp2693*, which encodes the enhanced entry protein, EnhB, but also includes genes encoding the other enhanced entry proteins, EnhA and EnhC.

4.3.3 Inference of recombination donors

To predict the origin of the homologous recombination regions, the 536 *L. pneumophila* genomes used in this study were first divided into BAPS clusters, which were mapped onto a phylogenetic tree (**Figure 4.8**). Eight clusters were identified, seven of which comprised isolates from the *L. pneumophila pneumophila* subspecies (BAPS clusters 1-6, 8), and one with isolates from *L. pneumophila fraseri* (BAPS cluster 7).

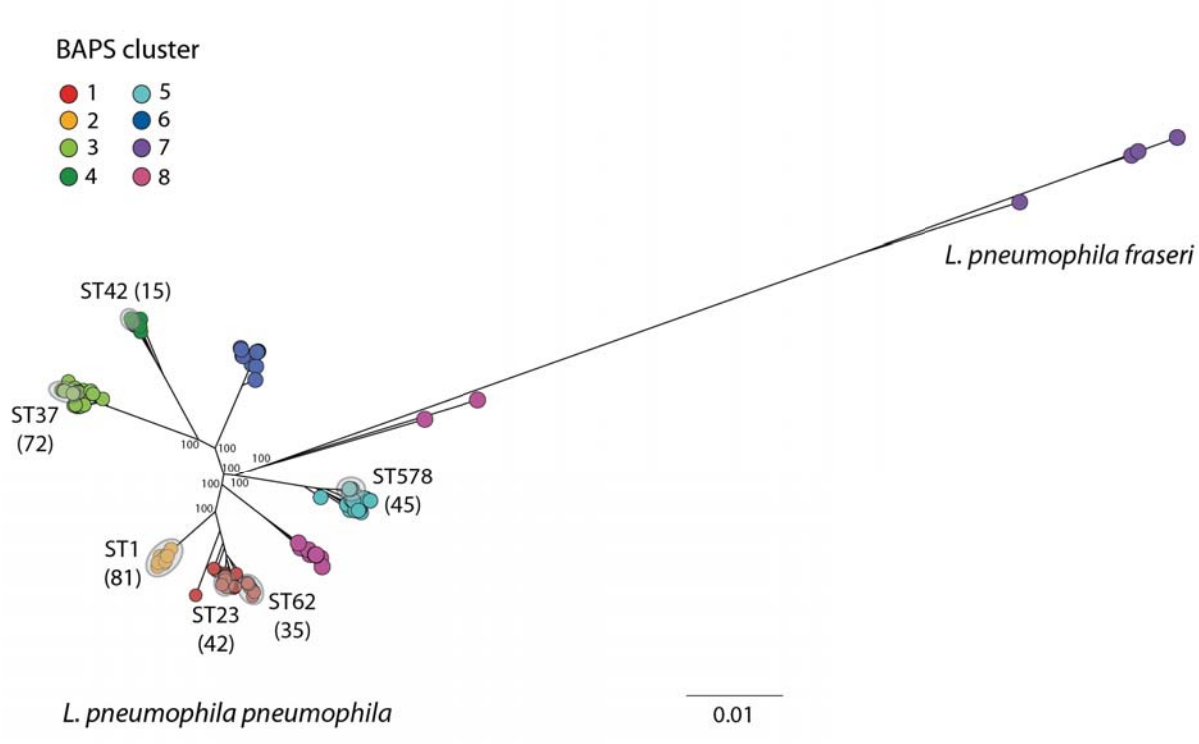

Figure 4.8. Maximum likelihood tree of 536 *L. pneumophila* isolates that are coloured by BAPS cluster. Grey circles also highlight the position of the six major disease-associated STs and the number of isolates belonging to each ST is indicated in brackets. The scale shows the number of SNPs per site. Bootstrap values, based on 1000 resamples, are shown for the major nodes of the tree.

All ancestral recombination sequences were extracted from the node downstream of the phylogenetic tree branch on which the recombination event was predicted to have occurred. Only regions greater than 500bp were used in this analysis, firstly because they were deemed more likely to be a "true" event, and secondly because small regions would likely have high similarity to many genomes. Each of the recombined fragments was used as a query in a BLASTn search against a database comprising all 536 *L. pneumophila* assembled genomes and the NCBI non-redundant database. The isolate with the highest bit score, together with the BAPS cluster from which it is derived, was considered the potential donor, provided that it covers at least 50% of the recombination fragment length and has a minimum of 99% nucleotide identity. Recombination fragments with no hits that met these thresholds were not assigned a donor ("No donor predicted"). Of the total 318 homologous recombination events

greater than 500bp predicted in the six STs, potential donors were predicted for 292 (91.8%) (**Table 4.4**). Many of the hits were almost perfect matches with 122 (41.8%) of the fragments having over 99.9% nucleotide identity, and 155 (53.1%) having hits that cover the full length of the recombination fragment (**Figure 4.9**).

Table 4.4. Number of homologous recombination events with predicted donors in each of the six STs.

ST	Total number of homologous recombination events	Number of recombination events >500bp	Number (and %) of filtered events with a predicted donor
1	198	193	176 (91.2%)
23	44	42	39 (92.9%)
37	13	12	9 (75.0%)
42	11	10	10 (100%)
62	48	39	36 (92.3%)
578	23	22	22 (100%)
Total	**337**	**318**	**292 (91.8%)**

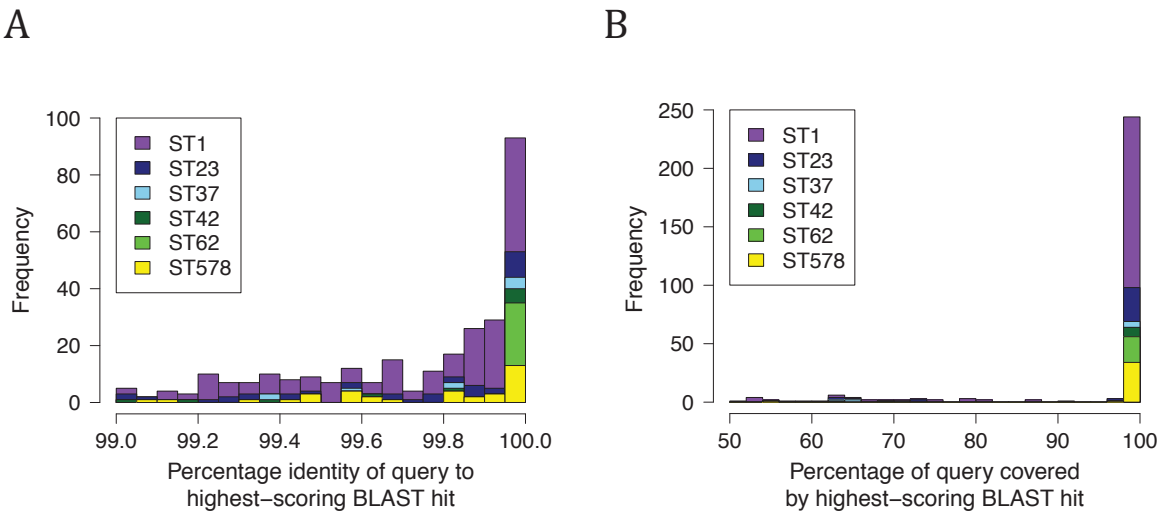

Figure 4.9. Similarity of the recombined regions to the predicted donors. The percentage nucleotide identity of the recombination fragments to the highest-scoring BLAST hit (A) and the percentage length covered by the highest-scoring BLAST hit (B).

The number of homologous recombination events in each of the six STs that are predicted to be derived from each of the eight BAPS clusters was calculated and visualised in a heat plot (**Figure 4.10**). Any events with equally good hits to isolates in more than one BAPS cluster were discarded for this analysis ("No donor assigned"). The heat plot illustrates that, in five of the six STs, recombination donors are most often from the same BAPS cluster as the recipient. The exception is ST37 in which the highest number of recombination fragments is derived from BAPS cluster 4, although its own cluster (BAPS cluster 3) accounts for the second highest number. However, all STs, with the exception of ST578, are also predicted to have acquired recombination fragments from clusters other than their own, demonstrating the occurrence of homologous recombination between major clusters of the *L. pneumophila pneumophila* subspecies. Interestingly, some BAPS clusters act frequently as donors (e.g. BAPS clusters 4 and 5) to other clusters, while others hardly donate except to isolates of their own cluster (e.g. BAPS clusters 2 and 3). Furthermore, just two events (one each in ST23 and ST62) are derived from the *L. pneumophila fraseri* subspecies (BAPS cluster 7).

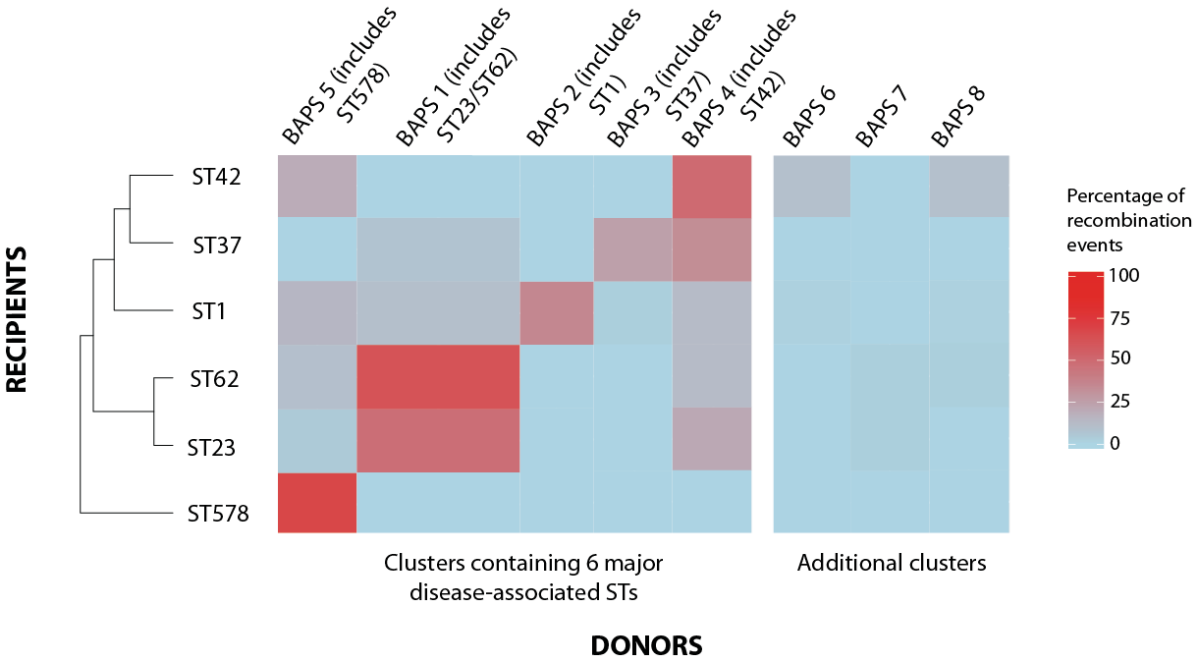

Figure 4.10. Predicted recombination donor clusters. A heat-map showing the percentage of recombination events detected in each of the six lineages that are predicted to be derived from each of the eight BAPS clusters. The six STs are shown in the left dendrogram constructed using

hierarchical clustering and based on the similarity of the predicted recombination donor lineages. The BAPS clusters are ordered from left to right based on the ordering of the six STs in the dendrogram. The column representing BAPS cluster 1, which contains both ST23 and ST62, is given twice the width as the other columns. The three BAPS clusters (6-8) that do not contain one of the six major disease-associated STs are shown on the right.

Recombination hotspot regions were next re-analysed to investigate whether the hotspots were driven by recombination events from the same or different BAPS clusters. The analysis focused on the ST1 lineage, which was previously found to contain the highest number of recombination events and the most prominent hotspots. The most notable hotspot region (hotspot 6), which was found to contain genes involved in up to 27 recombination events, was found to be driven mostly by recombination regions derived from the same BAPS cluster to which ST1 belongs (BAPS cluster 2) (**Figure 4.11**). However, a small number of recombination events that are predicted to be from BAPS cluster 5 were also observed in this region. Meanwhile, although some of the recombination events affecting the LPS locus (hotspot 3) could not be assigned a donor, others were derived from BAPS clusters 1, 2 and 5, suggesting that high diversity in this region may be important. Hotspot 4 appears to be driven by recombination events from BAPS clusters 5, 6 and 8 and contains no events derived from BAPS cluster 2 (to which ST1 belongs). However, the small number of events with predicted donors in most of these hotspots limits the conclusions that can be made.

For all homologous recombination events detected in the six STs, the percentage nucleotide identity between the imported fragment and the replaced fragment was calculated (**Figure 4.12A**). This analysis showed that 70% of homologous recombination events occurred between closely related isolates with >98% nucleotide similarity in the affected region, which agrees with our previous finding that most fragments are derived from the same BAPS cluster as the recipient. Interestingly, two peaks can be observed at ~98% identity and ~99.5-100% identity. These levels of divergence correspond to the nucleotide similarity observed between isolates belonging to different clusters or the same cluster, respectively (**Figure 4.12B**), and thus they represent recombination between and within clusters. Indeed, the recombination events that were predicted to be derived from the same BAPS cluster as the recipient have a

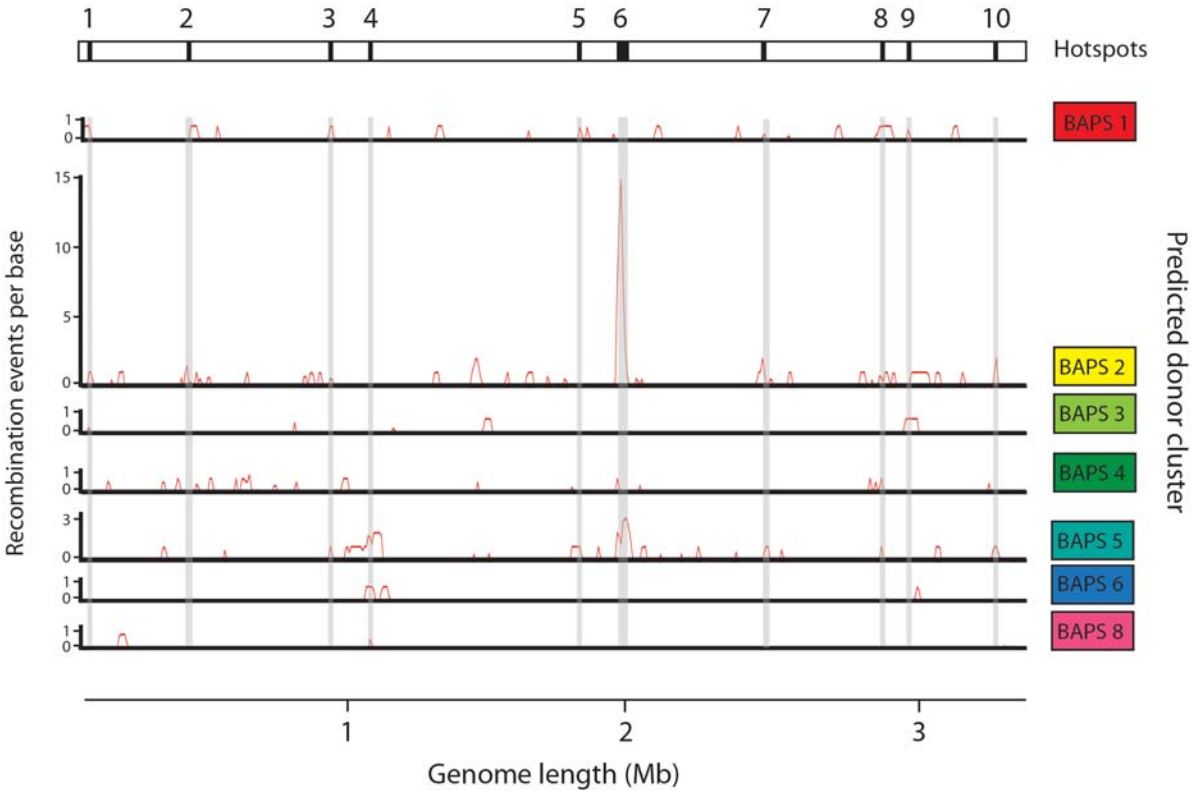

Figure 4.11. Diversity of recombination donors across the genome in the ST1 lineage. The number of recombination events per base that are derived from the different BAPS clusters are plotted. No events were predicted to be derived from BAPS cluster 7, which is thus excluded. The vertical grey bars correspond to the recombination hotspots.

mean nucleotide identity of 98.9% to the recipient genome while those predicted to be from a different cluster have a mean identity of 98.3%. Furthermore, very few recombination events were observed between isolates with <95% nucleotide identity, which is also concordant with our previous finding that very little recombination occurs between the two subspecies that share less than 95% nucleotide identity (**Figure 4.12A-B**).

Finally, the homologous recombination events that were predicted within the ST1 lineage were mapped onto the phylogenetic tree. This was to search for evidence of multi-fragment recombination, a process in which multiple non-contiguous segments that originate from the same molecule of DNA are imported into a recipient genome, and which is well documented in *S. pneumoniae* (Croucher *et al.*, 2012). Since the recombining fragments are non-contiguous, Gubbins will detect these as separate events

although the events should be predicted to have occurred on the same branch and have the same predicted donor. Indeed, **Figure 4.13** provides good evidence for the occurrence of this process in *L. pneumophila*, since many events with the same predicted donor, down to the BAPS cluster level and even the individual isolate level, are co-localised on branches. For example, over half of all recombination events in the ST1 lineage (100/193) occur on the same branch as another that is predicted to be derived from the same BAPS cluster.

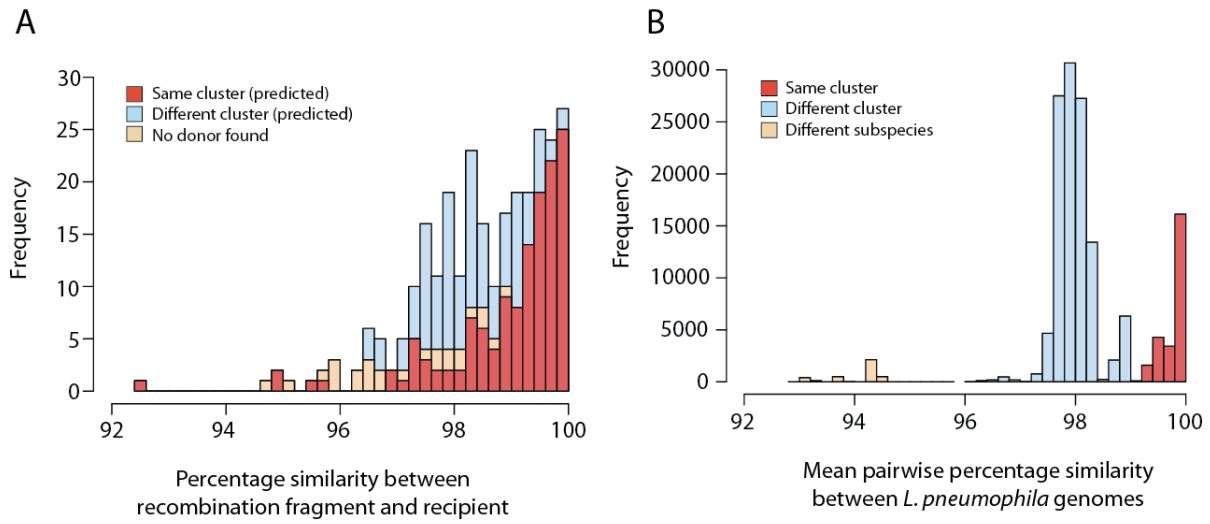

Figure 4.12. Sequence similarity between donors and recipients. A) Distribution of the percentage nucleotide similarities between the imported recombination fragments and the recipient sequence in all of the six STs. The events are categorised as being derived from the same or different BAPS clusters or with no donor lineage identified. B) Distribution of pairwise nucleotide similarities across the core genome amongst the 536 *L. pneumophila* isolates used in this study.

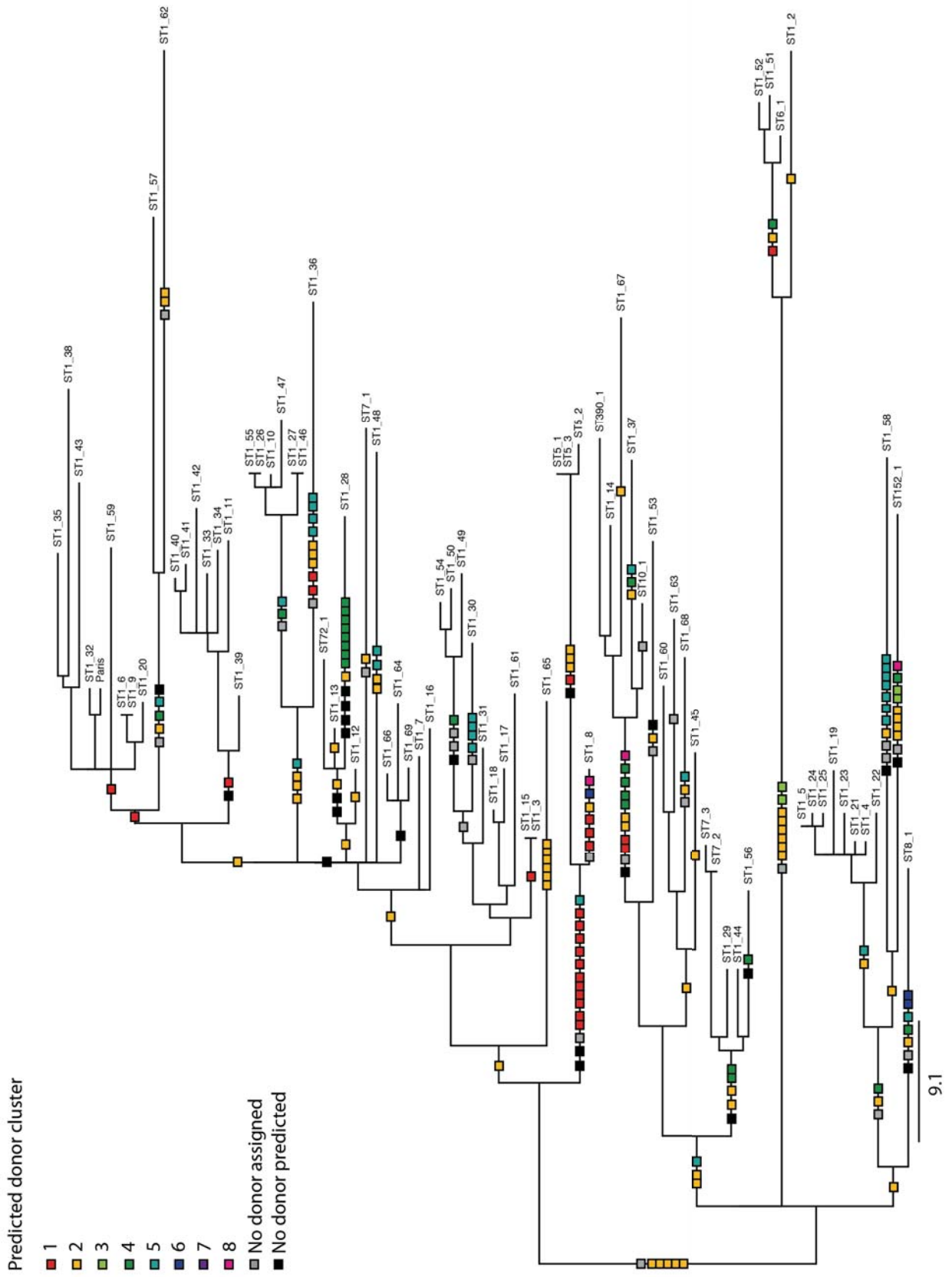
Predicted donor cluster
1
2
3
4
5
6
7
8
No donor assigned
No donor predicted
ST1_35
ST1_38
ST1_43
ST1_32
Paris
ST1_59
ST1_6
ST1_9
ST1_20
ST1_62
ST1_57
ST1_40
ST1_41
ST1_42
ST1_33
ST1_34
ST1_11
ST1_39
ST1_55
ST1_26
ST1_10
ST1_47
ST1_27
ST1_46
ST1_36
ST1_28
ST72_1
ST1_13
ST1_12
ST7_1
ST1_48
ST1_64
ST1_66
ST1_69
ST1_7
ST1_16
ST1_54
ST1_50
ST1_49
ST1_30
ST1_31
ST1_61
ST1_17
ST1_18
ST1_15
ST1_3
ST1_65
ST5_1
ST5_3
ST5_2
ST1_8
ST390_1
ST1_67
ST1_14
ST1_37
ST10_1
ST1_53
ST1_63
ST1_60
ST1_68
ST1_45
ST7_3
ST7_2
ST1_56
ST1_29
ST1_44
ST1_52
ST1_51
ST6_1
ST1_2
ST1_5
ST1_24
ST1_25
ST1_19
ST1_23
ST1_21
ST1_4
ST1_22
ST1_58
ST152_1
ST8_1
9.1

Figure 4.13. Maximum likelihood tree of 81 ST1 isolates with predicted recombination events mapped onto the branches (previous page). The tree was constructed using only vertically inherited SNPs and the scale bar indicates the number of SNPs. Predicted recombination events are represented by squares and coloured according to the BAPS cluster from which they are predicted to have been derived. Squares representing events with the same predicted donor at the isolate level and that have occurred on the same branches are joined together, and possibly represent multi-fragment recombination.

4.4 Discussion

Previous studies of several major disease-associated STs of *L. pneumophila* including ST578 (Sanchez-Buso *et al.*, 2014) and ST1, ST23, ST37 and ST62 (*Chapter 3* in this thesis) have shown that recombination is a dominant force in *L. pneumophila* evolution. However, the relative impacts of homologous and non-homologous recombination on *L. pneumophila* diversity have not been disentangled in any previous studies, and nor have the dynamics of homologous recombination been studied in detail in this species. Therefore, by studying six major disease-associated STs of *L. pneumophila*, including the five mentioned above and ST42, the aims of this thesis chapter were i) to determine the relative impact of homologous and non-homologous recombination on *L. pneumophila* evolution; ii) to identify homologous recombination hotspots; and iii) to explore the dynamics of recombination flux within the *L. pneumophila* species.

Analysis of all six lineages confirmed findings from *Chapter 3* that over 96% of SNPs in some lineages are found in recombined regions. However, when homologous recombination regions were distinguished from those associated with mobile genetic elements (non-homologous recombination) and repeat regions, the former were found to account for between 33.0% (ST62) and 80.0% (ST578) of the total SNPs. Remarkably, while homologous recombination events were shown to have occurred less frequently than *de novo* mutations in all lineages, they have contributed to between 20.8 and 93.8 times as many SNPs. These results support a very important role for homologous recombination in shaping the population structure and evolution of *L. pneumophila*, and highlight its potential to facilitate rapid adaptation to new niches such as modern, man-made water systems.

The fragments derived from homologous recombination were mostly found to be small (<10kb) although a small number of large events up to ~95kb were identified. A similar distribution of fragment sizes has also been reported in a previous study of recombination in *S. pneumoniae* in which it was suggested that transformation is optimised for exchanging short sequences rather than large features such as complete operons (Croucher *et al.*, 2012). This scenario could be favoured as it allows for larger numbers of potentially advantageous allele combinations to be tested.

Analysis of the genomic distribution of recombination events identified a total of 32 hotspot regions across the six STs. The most prominent hotspots were found in the ST1 lineage, in which the highest number of homologous recombination events was also detected. This is in concordance with the finding from *Chapter 3* that ST1 also has the highest number of vertically inherited mutations. Of particular note is a region containing genes that have been involved in up to 27 recombination events. The region appears to centre on the *hemB*/*lpp1771* gene, a porphobilinogen synthase (delta-aminolevulinic acid dehydratase), which is an enzyme involved in the biosynthesis of tetrapyrroles. However, the surrounding genes were also analysed and it seems more likely that the nearby gene, *lpp1773*, which encodes the outer membrane protein, FadL, may be responsible for driving the selection of recombination events in this region. Outer membrane proteins such as FadL could be under a high selection pressure to vary in order to interact with a highly variable aspect of the environment (e.g. diverse protozoan hosts), to escape protozoan predation, or to cope with an immune response during infection of host cells. However, since protozoa do not have an adaptive immune response, the latter possibility is unlikely unless higher organisms (e.g. humans) are also part of the infection cycle. As suggested in *Chapter 3*, it could be possible that this is indeed the case and that FadL is eliciting an immune response from humans.

Despite the prominence of this hotspot region in the ST1 lineage, it was not identified in any of the other STs apart from ST23, in which the hotspot region appeared to be centred on the outer membrane protein, BamA (encoded by *lpp1769*). BamA is also conserved across Gram-negative bacteria and is required for the assembly and insertion of beta-barrel proteins into the outer membrane (Tommassen, 2010). A *fadL*-like gene was also found within a recombination hotspot in the ST62 lineage, although this hotspot was found in a different part of the genome to the hotspot identified in the ST1

lineage. Further studies, perhaps involving larger number of isolates, would be useful to confirm the gene(s) that are driving these hotspots and to determine whether the prominent hotspot region in the ST1 lineage is also an important hotspot region in other lineages, or whether it represents a unique selection pressure in ST1 isolates.

Across the six STs, a number of other outer membrane proteins such as TolC were also identified within recombination hotspot regions. Of the many outer membrane proteins likely expressed on the surface of *L. pneumophila*, these results provide clues as to which ones are being selected for variation and part of dynamic environmental interactions. Furthermore, the LPS locus was also found in recombination hotspots in both the ST1 and ST62 lineages. Given that the LPS has been shown to be the major immunodominant antigen of *L. pneumophila* in the laboratory (Ciesielski *et al.*, 1986; Petzold *et al.*, 2013), it could be that it is also generating an immune response from humans and is thus under strong selection to vary. Horizontal exchange of the LPS locus also explains a previous observation that sg 1 isolates can have diverse genomic backgrounds, and that serogroups often do not correlate with overall genomic relatedness (Cazalet *et al.*, 2008).

A number of recombination hotspots also contain putative or confirmed effectors of the type IVB Dot/Icm secretion system of *L. pneumophila*. Dot/Icm effectors, of which there are over 300 described, manipulate a wide range of host cell processes and are essential to *L. pneumophila* pathogenesis (Ensminger, 2016). The genes found within recombination hotspots include that which encodes the first described effector, RalF. They likely represent those at the forefront of the arms race between *L. pneumophila* and its protozoan (or even human) hosts. It will be intriguing to decipher whether variation is being selected for within these effectors in order to take advantage of a wide variety of host species, or to counter changes in individual hosts. Larger sets of genomic data would also be useful to confirm the existence of these hotspots and further explore differences between lineages, which could suggest differences in hosts and infection strategies.

A number of other identified hotspots are also worthy of further investigation. These include a region within the ST1 lineage that appears to be centred on the *lpp2599/tehB* gene, which encodes the tellurite resistance protein, TehB. This has been identified in

both Gram-positive and Gram-negative bacterial pathogens and is involved in the detoxification of tellurite (Taylor, 1999). However, due to the apparent rarity of tellurium compounds in the environment, it has been suggested that this may not be the main function of this gene. For example, one study found that when the *tehB* gene from *S. pneumoniae* is expressed in *E. coli*, it results in a filamentous morphology (Liu & Taylor, 1999). The authors hypothesise that it might act as a methyltransferase that can alter the methylation of proteins related to cell division, thereby resulting in the generation of elongated cells (Liu & Taylor, 1999). Several studies have shown that *L. pneumophila* can also form long filamentous cells, particularly in response to stress conditions such as antibiotic exposure (Smalley *et al.*, 1980; Elliott & Rodgers, 1985), but it is unknown whether TehB is involved in this process. Further understanding of the function of TehB is therefore required to understand why this gene is associated with a recombination hotspot in the ST1 lineage.

Just one hotspot region was identified in the ST42 lineage, which appears to be centred on the enhanced entry gene, *enhB*. While little is known about *enhB*, the neighbouring gene, *ST42_02564*/*lpp2692*, which encodes the enhanced entry protein, EnhC, has been shown to be important for entry into host cells (Cirillo *et al.*, 2000) and to facilitate intracellular growth of *L. pneumophila* by evading immune recognition by the pattern recognition receptor (PRR), Nod1, in macrophages (Liu *et al.*, 2012). Further studies are required to understand why variability within the enhanced entry proteins might be advantageous, and also why these genes were found in a hotspot in the ST42 lineage and not others.

Recombination donors were predicted for over 90% of homologous recombination events (over 500bp) identified in the six STs. In all but one lineage, the highest number of recombination events was predicted to be from the same BAPS cluster as the recipient. This is an expected finding since homologous recombination is thought to require high, or even perfect, sequence homology between the donor and recipient at both ends of the recombination fragment (Majewski & Cohan, 1998), a scenario which is more likely between closely-related bacteria. However, all disease-associated STs, with the exception of ST578, have imported regions from BAPS clusters other than their own, thus also demonstrating evidence for homologous recombination between major clades

of the *L. p. pneumophila* subspecies. This suggests that different clades at least partially share the same ecological niche and that new adaptations can be shared freely between these disease-associated STs. This is in concordance with the findings of *Chapter 3*, whereby large regions were found to be transferred from the ST62 lineage to the ST47 lineage, and from the ST1 lineage to the ST37 lineage. Thus, the findings from both *Chapter 3* and the current chapter suggest that one of the disease-associated lineages could have initially acquired mutations or genes facilitating adaptation to human infection and these were subsequently shared with other lineages via homologous recombination (rather than independent acquisition by different lineages). Man-made water systems could provide a mixing vessel in which this process occurs. The findings also highlight the potential risk of more disease-associated strains from wide-ranging genomic backgrounds emerging rapidly in the future after having acquired adaptive features for human infection via homologous recombination.

Interestingly though, some BAPS clusters were predicted to act frequently as donors (e.g. BAPS 4 and 5), while others hardly donate, apart from to isolates of their own cluster (e.g. BAPS 2 and 3). A possible explanation for this could be related to the presence of restriction-modification systems in some lineages that prevent horizontal acquisition of DNA from lineages other than their own. Similar patterns whereby different lineages donate and receive DNA at different rates have also been observed in other species such as *S. pneumoniae* (Chewapreecha *et al.*, 2014), *C. trachomatis* (Harris *et al.*, 2012) and *E. coli* (Didelot *et al.*, 2012).

Only two recombination events detected within the six lineages were predicted to be from the *L. p. fraseri* subspecies. Given that this subspecies shares less than 95% nucleotide identity with the *L. p. pneumophila* subspecies, this was not an unexpected finding, given the high identity required for homologous recombination. It could be that these two subspecies have gradually diverged due to differing ecologies, and that eventually they may become different species that are fully incapable of exchange *via* homologous recombination.

Finally, the detection of multiple recombination events that are derived from the same donor and predicted on the same tree branch suggests the occurrence of multi-fragment

recombination. This is a process in which multiple non-contiguous segments of DNA originating from the same donor molecule are imported into the recipient genome during transformation and which has been documented in several studies of *S. pneumoniae* (Hiller *et al.*, 2010; Golubchik *et al.*, 2012; Croucher *et al.*, 2012). However, it could also be that the recombining isolates have shared a common niche for a prolonged period of time, and that multiple independent recombination events have occurred during this time. Thus further experimental studies will be required to confirm the occurrence of this process in *L. pneumophila*.