2 Materials and Methods

2.1 Culturing and transforming *S. pneumoniae*

2.1.1 Culturing of strains

Unless otherwise specified, *S. pneumoniae* was cultured on 5% horse blood agar plates (Oxoid), or grown in Brain-Heart infusion (BHI; Oxoid), statically at 37 °C. *S. pneumoniae* ATCC 700669 and TIGR4, and derivatives thereof, were grown under aerobic conditions, whereas *S. pneumoniae* 99-4038 and 99-4039 required a microaerophilic atmosphere, generated by CampyGen Compact sachets (Oxoid), to grow on blood agar plates, but could be grown aerobically in BHI.

2.1.2 Sampling of PMEN1 and ST180 isolates

For the analyses presented in Chapter 4, all available isolates with a sequence type of 81 (or a single locus variant thereof), or a pulsed field gel electrophoresis profile matching that of known PMEN1 strains, were sampled (Appendix II: PMEN1 strains). Where information on serotype and drug resistance profile was available, strains resistant to both penicillin and tetracycline were screened using PCRs, as described below, targeted to detect the presence of the CDS SPN23F00710, which is not commonly found in the pneumococcal chromosome, and the lantibiotic biosynthesis operon carried on ICE*Sp*23FST81, which has not been identified in any other sequenced pneumococcal lineage. No isolate positive for the SPN23F00710 locus was negative for the lantibiotic biosynthesis locus, and only a single isolate selected on the basis of sequence type information lacked the SPN23F00710 gene, so there is no reason to expect that this strategy significantly biased the sample. For the analyses presented in Chapter 8, the samples received from T Mitchell, selected as diverse representatives from a global collection studied using CGH (Inverarity, 2009), were sequenced using 454 and capillary technologies; all samples received from B Henriques-Normark were sequenced using Illumina technology (Appendix IV: Serotype 3 strains).

2.1.3 Transformation experiments

All strains subject to transformation had the same allele of the *comC* gene, hence were all expected to respond to CSP-2. Ten millilitres of Brain Heart Infusion (Oxoid) was inoculated with 150 μL of an overnight culture of the recipient strain. When the culture reached an OD_{600} of between 0.20-0.25, 1 mL was added to the appropriate amount of donor DNA (20 ng unless specified) in 5 μL water, 10 ng CSP-2 in 2 μL water (Sigma) and 5 μL 500 mM calcium chloride. Another 1 mL was added to the same quantity of donor DNA and calcium chloride in the absence of any CSP as a negative control. These reactions were incubated at 37 °C for 2 h. Samples of these cultures were then serially diluted in phosphate-buffered saline solution and total cell population determined by counting colonies in three 20 μL volumes spotted onto 5% blood agar plates from the appropriate dilution. The number of transformants from each reaction was determined by spreading three 50 μL volumes each onto a 5% blood agar plates supplemented with the appropriate antibiotic selection; unless specified, this was 200 μg mL^{-1} kanamycin (Gibco).

2.1.4 Omnilog experiments

Frozen stocks of *S. pneumoniae* 99-4038 and 99-4039 were passaged twice on blood agar plates overnight in order to prevent contamination of assays with glycerol. Colonies were then scraped off plates using sterile cotton swabs and dispensed into IF-0a solution (Biolog) at room temperature to a cell density corresponding to 81% transmittance. For each Omnilog phenotype microarray plate used (PM9-20) (Bochner, 2009), 120 μL of this cell suspension was added to 10 mL IF-10b solution (Biolog). This was then supplemented with 7.5 mM D-ribose (Sigma), 2 mM magnesium chloride, 1 mM calcium chloride, 2 mM sodium pyrophosphate (Sigma), 25 μM L-arginine (Sigma), 25 μM L-methionine (Sigma), 25 μM hypoxanthine (Sigma), 10 μM lipoamide (Sigma), 5 μM nicotine adenine dinucleotide (Sigma), 0.25 μM riboflavin (Sigma), 0.005% by mass yeast extract (Fluka) and 0.005% by mass Tween 80 (Sigma). The solution was then made up to a volume of 12 mL with distilled water, and 100 μL dispensed into each well on the assay plate. Plates were then allowed to equilibrate in an anaerobic atmosphere (80% N_2, 10% CO_2, 10% H_2) for 5 min prior to being sealed in airtight bags and loaded into the Omnilog machine.

Plates were scanned every 10 min for 48 h while incubated at 37 °C. Two paired replicates were performed for the two strains.

2.2 Extraction and analysis of nucleic acids

2.2.1 Genomic DNA extractions

Isolates were grown in 10 mL BHI (Oxoid) and pelleted through centrifugation (2,594 *g*, 10 min). Pellets were washed in 1 mL 50% glycerol and resuspended in 250 μL Tris-EDTA buffer and 50 μL 30 g L^{-1} lysozyme (Roche) in Tris-EDTA buffer. This mixture was vortexed at room temperature for 15 mins and 400 μL 0.1 M EDTA (Gibco) and 250 μL 10% sarkosyl (BDH) were added. Samples were incubated at 4 °C for 2 h, prior to the addition of 50 μL proteinase K (Roche), 30 μl RNase A (Roche) and 3 mL Tris-EDTA buffer. Samples were incubated at 50 °C overnight. Samples were washed with 5 mL of a 25:24:1 mixture of phenol, chloroform and indole-3-acetic acid (IAA; Fluka) and centrifuged (2,594 *g*, 10 min). The aqueous phase was removed, washed with 5 mL chloroform (Sigma) and centrifuged (2,594 *g*, 10 min). DNA was precipitated from the aqueous phase in 7.5 mL isopropanol, washed in 5 mL 70% ethanol and resuspended in 250 μL Tris-EDTA buffer.

2.2.2 RNA sample extractions

For *S. pneumoniae* ATCC 700669, samples were harvested from 10 mL cultures at an OD_{600} of 0.8 through mixing with RNAProtect (Qiagen) in a 1:2 ratio then pelleted through centrifugation (2,594 *g*, 10 min). Cells were resuspended in 1 mg mL^{-1} lysozyme (Roche) in 200 μL Tris-EDTA buffer and lysed at 37 °C for 10 min. The sample volume was then made up to 800 μL with Tris-EDTA and split six equal volumes, each of which was independently processed using the SV Total RNA Extraction System (Promega). The quality of the RNA was then assessed using an Agilent 2100 Bioanalyzer RNA Nano chip (Agilent); any samples with an RNA integrity number below nine were discarded. RNA was then precipitated through mixture with 300% by volume ethanol and 10% by volume 3 M sodium acetate followed by storage at -80 °C overnight. RNA was then resuspended at a

concentration of 0.83 mg mL^{-1} and the 16S and 23S rRNA transcripts depleted through complementary oligonucleotide hybridization (MicrobExpress, Ambion) according to manufacturer's instructions. RNA was then precipitated, then resuspended at a concentration of 0.625 mg mL^{-1} in water and treated with DNase I (Roche) at room temperature for 15 min. Reactions were stopped through washing with an equal volume of phenol, chloroform and IAA mixed in proportions 25:24:1 (Fluka) followed by phase separation through centrifugation at 16,157 *g* for 10 min. A PCR using primers smL and smR, targeting the *rpsL* gene, was performed as described below using the aqueous RNA solution as the template; the DNase I treatment was repeated until no amplification product was detectable. The replicate RNA samples were then pooled and split into two halves, each resuspended in 13.73 μL water and mixed with 1.67 μL of 3 μg μL^{-1} random hexamer oligonucleotides (Sigma), then incubated at 70 °C for 10 min, followed by incubation on ice for 10 min. A reverse transcription reaction was then performed using SuperScript III (Invitrogen), according to manufacturer's instructions, at 42°C for 2 h. Samples were then washed on a G50 spin column (GE Healthcare). For one of the two samples, the second cDNA strand was synthesised through incubating this first strand cDNA with DNA polymerase I (Invitrogen) and RNase H (Invitrogen) in second strand buffer, according to manufacturer's instructions, at 16°C for 2.5 h.

For analysis of *S. pneumoniae* 99-4038 and 99-4039, samples were harvested from 10 mL cultures at an OD_{600} of 0.6 through centrifugation (2,594 *g*, 10 min), then lysed by treatment with 30 mg mL^{-1} lysosyme (Roche) at room temperature for 15 min. RNA was extracted as described above, but no depletion of rRNA was performed. For analysis of *S. pneumoniae* TIGR4 and $TIGR4^{PUS}$, RNA was extracted as described for *S. pneumoniae* 99-4038 and 99-4039 then, following evaluation on the Bioanalyzer, sent to the BμG@S group (St. George's Hospital, London) for microarray analysis.

2.2.3 PCR and RT-PCR

All PCRs were conducted using 1 μL 10 ng μl^{-1} template DNA, when using genomic or plasmid DNA as a template, or 1 μL of the relevant undiluted sample when using RNA or cDNA as a template. Each primer (Appendix I: Primer sequences) was then added as 1 μl of a 10 μM solution (Sigma or IDT), and the reaction made up to a

volume of 50 μl with PCR Platinum Supermix (Invitrogen). The thermocycle in each case used a denaturing temperature of 95 °C (30 s), a hybridization temperature falling from 60 °C to 55 °C over 5 cycles, then remaining at 55 °C for a further 25 cycles (30 s), and an extension temperature of 72 °C (one minute per 1 kb of product length).

2.3 Construction of mutant strains

2.3.1 Construction of *S. pneumoniae* TIGR4^{PUS}

The region upstream of *patAB* in *S. pneumoniae* 99-4038 and 99-4039 was amplified through PCR using primers IntGL and IntGR (Appendix I: Primer sequences). The ~500 bp PCR products were each purified through agarose gel electrophoresis using a QIAquick Gel Extraction Kit (Qiagen) and then ligated into pGEM-T Easy (Promega) using T4 ligase (Promega) in a 10 μL reaction volume, according to manufacturer's instructions. A 1 μL sample of this reaction was then used to transform electrocompetent *E. coli* TOP10 cells (Invitrogen) through electroporation with a 2.5 kV pulse. These cells were then grown in 250 μL SOC medium (Invitrogen), shaken at 37 °C for 2 h. A 50 μL sample of this culture was then spread on Luria broth (LB) agar plates supplemented with 100 μg mL^{-1} ampicillin (Sigma), 300 μg mL^{-1} S-Gal (Sigma) and 30 μg mL^{-1} isopropyl β-thiogalactoside (Sigma). White colonies were then picked, grown in LB supplemented with 100 μg mL^{-1} ampicillin and stored. The sequences of the plasmid inserts were then amplified by PCR and checked through capillary sequencing as described below. Both plasmids were then extracted from their host *E. coli* using the QIAprep Spin Miniprep Kit (Qiagen) and diluted to 25 μg mL^{-1}. These two stocks was then used to transform three *S. pneumoniae* TIGR4 cultures in parallel as described above; after 2 h growth, a 50 μL sample of each transformation reaction was used to inoculate either BHI or BHI supplemented with 2 μg mL^{-1} ciprofloxacin. Colonies were then isolated from the *S. pneumoniae* TIGR4 culture transformed with the region upstream of *patAB* from *S. pneumoniae* 99-4039 after 20 h growth in the presence of ciprofloxacin. PCR amplification and sequencing of the region upstream of *patAB* in these strains revealed they all shared the PUS; one of these was stored and designated *S. pneumoniae* TIGR4^{PUS}.

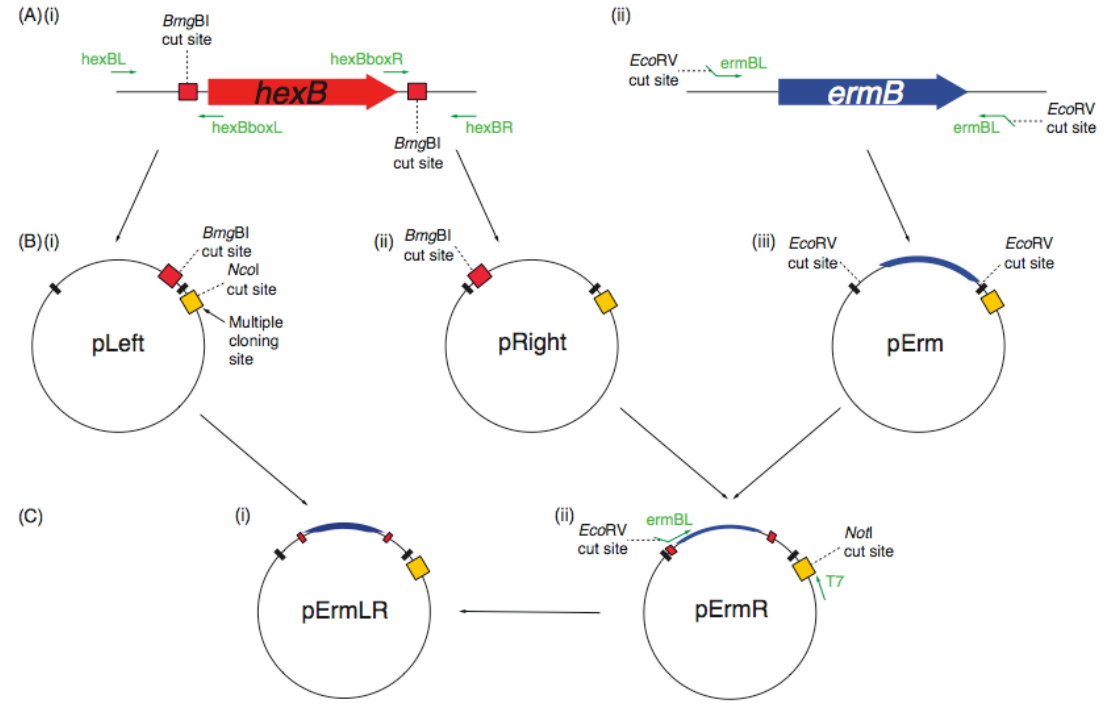

Figure 2.1 Construction of *S. pneumoniae* ATCC 700669Δ*hexB*. See text for details.

2.3.2 Construction of *S. pneumoniae* ATCC 700669Δ*hexB*

The sequences upstream and downstream of *hexB* in *S. pneumoniae* ATCC 700669 each contain similar BOX elements, which encode *Bmg*BI cut sites (Figure 2.1A i). Approximately 500 bp regions flanking *hexB*, both including the *Bmg*BI cut sites, were amplified by PCR and cloned into pGEM-T Easy as described above to give plasmids pLeft and pRight (Figure 2.1A i, B i, ii). The *ermB* erythromycin resistance gene was then amplified from strain *S. pneumoniae* 11930 using the primers ermBL and ermBR (each with an *Eco*RV cut site on the 5' end; Appendix I: Primer sequences), and cloned into pGEM-T Easy as described above to give plasmid pErm (Figure 2.1A ii, B iii). All three plasmids were then transformed into *E. coli* TOP10 cells through electroporation with a 2.5 kV pulse, followed by blue-white selection on ampicillin, as described above. The plasmid carrying the *ermB* gene was then extracted using a QIAprep Spin Miniprep Kit (Qiagen) and digested using *Eco*RV (New England Biolabs), releasing the insert with blunt ends. This fragment was purified through agarose gel electrophoresis as described above. Plasmid pRight was similarly extracted, then digested with *Bmg*BI (New England Biolabs), which also

cuts to give blunt ends. The *ermB* fragment was cloned into this blunt cut site using T4 ligase (Promega) overnight at 4 °C according to manufacturer's instructions (Figure 2.1C ii). A 1 μL sample of this reaction was then used to transform electrocompetent *E. coli* TOP10 cells as described above. A 50 μL sample of this culture was then spread on LB agar plates supplemented with 100 μg mL^{-1} erythromycin. The plasmid pErmR was isolated from a colony, and its composite insert and the adjacent multiple cloning site (MCS) amplified using primers ermBL (which has an *Eco*RV cut site on the 5' end) and T7. This amplicon was purified through agarose gel electrophoresis using a QIAquick Gel Extraction kit (Qiagen) and digested with with *Eco*RV and *Nco*I (New England Biolabs), the latter of which cuts within the MCS of pGEM-T Easy (Figure 2.1C ii). This was ligated into pLeft following its digestion with *Bmg*BI and *Nco*I (Figure 2.1C i) as described above. This ligation reaction was used to transform *E. coli* TOP10 cells as described above, with strains carrying the plasmids again selected on LB agar supplemented with 100 μg mL^{-1} erythromycin. This gave pErmLR, which was used to transform *S. pneumoniae* ATCC 700669[lab]. Pneumococcal transformants with disrupted *hexB* genes were selected on 5% blood agar plates supplemented with 0.1 μg mL^{-1} erythromycin. One of these colonies was picked and the insert checked through PCR amplification with primers hexBL and hexBR and capillary sequencing.

2.4 DNA and RNA sequencing

2.4.1 Genome sequencing

All processing and sequencing of genomic DNA samples was performed by the Wellcome Trust Sanger Institute's core sequencing teams. DNA sample concentrations were determined using the Qubit system (Invitrogen) and subsequently diluted to 2.5 μg in 100 μl Tris-EDTA buffer. Unless specified otherwise, all samples analysed as part of the PMEN1 and ST180 populations or the *in vitro* transformation experiment were sequenced as multiplexed libraries of up to twelve isolates using the Genome Analyzer II (Illumina). Libraries were constructed according to manufacturer's instructions with a specified insert size of 250 bp. Briefly, DNA samples were first sheared by nebulisation (35 psi, 6 min). Duplexes were then blunt

ended through an end repair reaction using large Klenow fragment, T4 polynucleotide kinase and T4 polymerase. A single 3' adenosine moiety was added to the cDNA using Klenow exo$^-$ and dATP. Illumina adapters, containing primer sites for flow cell surface annealing, amplification and sequencing, along with one of the twelve unique tag sequences in multiplexed libraries, were ligated onto the repaired ends of the DNA. Gel electrophoresis was used to select for DNA constructs around 250 bp in size, which were subsequently amplified by 18 cycles of PCR with Phusion polymerase. These libraries were denatured with sodium hydroxide and diluted to 3.5 pM in hybridization buffer for loading onto a single lane of an Illumina Genome Analyzer II flow cell (Illumina). Cluster formation, primer hybridization and sequencing reactions were according to the manufacturer's recommended protocol. Data for all PMEN1 and ST180 samples was in the form of paired end 54 nt reads, while data for all transformant experiments was in the form of 76 nt paired end reads.

The reference genomes of *S. pneumoniae* ATCC 700669 and *S. pneumoniae* OXC141 were generated through capillary sequencing (Appendix III: EMBL accession codes). Briefly, a shotgun sequence with ~8-fold genome coverage was achieved through sequencing of pUC clones with 1.4- to 2.8-kb inserts and pSMART clones with 8- to 12-kb inserts using a BigDye terminator sequencing kit and AB 3700 sequencers (Applied Biosystems). Sequences from 30- to 40-kb pEpiFOS-5 fosmid clones and 12- to 23-kb pBACe3.6 BAC clones were used to scaffold contigs and bridge repeats. The sequence was finished according to standard criteria (Parkhill *et al*., 2000). Sequence assembly, visualization, and finishing were performed by using PHRAP (www.phrap.org) and Gap4 (Bonfield *et al*., 1995). All repeat sequences were independently verified.

In order to ascertain a better view of the accessory genomes of PMEN1 strains *S. pneumoniae* 11876 and 11930, and ST180 strains *S. pneumoniae* 03-4156, 03-4183, 07-2838, 02-1198, 99-4038 and 99-4039, genomic DNA samples were sequenced using the 454 platform (Roche). Assemblies were generated by using Newbler to produce scaffolds from the 454 data that were subsequently improved by iteratively mapping the Illumina data to the draft assembly using IMAGE (Tsai *et al*., 2010).

2.4.2 Transcriptome sequencing

All processing and sequencing of complementary DNA samples was performed by the Wellcome Trust Sanger Institute's core sequencing teams. Complementary DNA samples were sequenced using the Illumina Genome Analyzer II (Illumina) as described for genomic DNA samples, with the exceptions that the specified insert size was 150 bp and each sample was sequenced as a non-multiplexed library to ensure sufficient coverage. Paired RNA-seq samples for differential expression analyses were sequenced on adjacent lanes on the same flow cell, to avoid variation introduced between cells (Marioni *et al*., 2008). All data was in the form of 54 nt paired end reads.

2.4.3 Oligonucleotide sequencing

2.4.3.1 DNA and RNA synthetic oligonucleotide mixtures

All processing and sequencing of DNA and RNA oligonucleotide mixture samples, used for validation of the RNA-seq methodology, was performed by the Wellcome Trust Sanger Institute's research and development sequencing team. A DNA oligonucleotide with the sequence AACATCTGCAAG$(N)_{19}$CAGCGACGCATC$(N)_5$ (Sigma), either alone or in the presence of an equimolar amount of a 3' phosphorylated RNA oligonucleotide of sequence GAUGCGUCGCUG (Sigma), was diluted to a concentration of 120 nM in Tris–EDTA buffer and subjected to standard Illumina library preparation reactions. Following Illumina library construction, as described above, the DNA and RNA oligonucleotides were subjected to both 36 nt paired end sequencing and 54 nt single end sequencing as non-multiplexed libraries on the Illumina Genome Analyzer II.

2.4.3.2 PCR product and plasmid sequencing

All capillary sequencing of PCR products and plasmids was performed by the Wellcome Trust Sanger Institute's faculty sequencing team. Prior to submission, all PCR products were purified through agarose gel electrophoresis and all plasmids purified using the QIAprep Spin Miniprep Kit (Qiagen). These samples were then diluted to a concentration of 10 ng mL^{-1} in 50 μL, measured through the Qubit system

(Invitrogen), in preparation for 3.1 Bigdye sequencing on AB 3730 capillary sequencing machines (Applied Biosystems).

2.5 Alignment and assembly of short sequence reads

2.5.1 Generation of whole genome alignments for phylogenetic analyses

Illumina sequence data were mapped to the appropriate complete reference genome as paired end reads with an insert size between 50 and 400 bp using either SSAHA2 (Ning *et al*., 2001) for PMEN1, or otherwise SMALT (Ponstingl, 2011). The reference sequences used were *S. pneumoniae* ATCC 700669 for PMEN1 (Chapter 4), *S. pneumoniae* ATCC 700669^{lab} for the *in vitro* transformation experiments (Chapter 5) or *S. pneumoniae* OXC141 for ST180 (Chapter 8). For those ST180 strains that had only 454 or capillary data, paired end reads of the appropriate length and insert size were simulated from the best available assembly and analysed in parallel with those isolates sequenced using the Illumina platform. SNPs were identified as described in Harris *et al.* (Harris *et al*., 2010). Briefly, only reads aligning to the reference sequence with a quality score of greater than 30 were considered. For each position, a base was only called if the Phred quality score of the site was above 50 (theoretically equating to an accuracy of 99.999%; www.phrap.org), and the call was supported by 75% of at least four reads, with at least two on each strand. Otherwise, the position was recorded as unknown.

For PMEN1, small indels were identified from the SSAHA2 output. Indels were called where more than 75% of the reads (corresponding to at least five reads) spanning the site supported the change in sequence length. If between 25% and 75% of the reads supported an indel identical to one confidently identified in another strain, the indel was marked as missing data. The sequence of insertions was called as a base if 75% or more of the reads supporting the change agreed on a consensus; otherwise, the position was treated as an unknown base. For the *in vitro* transformation and ST180 samples, indels were identified using bcftools (Danecek *et al*., 2011) and filtered using the same criteria as other sites in the genome.

2.5.2 Generation of a genome alignment for *in vitro* transformation analysis

A draft genome assembly for strain TIGR4Δ*cps* was generated by splicing together the sequence of the disrupted capsule biosynthesis locus (Pearce *et al*., 2002) [EMBL accession code AF160759] with the rest of the TIGR4 genome (Tettelin *et al*., 2001) [EMBL accession code AE005672]. Illumina sequence data generated from the DNA used for transformation was then used to correct this sequence using ICORN (Otto *et al*., 2010). The sequence of the *S. pneumoniae* ATCC 700669 line used in this experiment (*S. pneumoniae* ATCC 700669^{lab}) was derived by correcting the reference sequence with ICORN (Otto *et al*., 2010) using resequencing data; this revealed the presence of four substitutions and the loss of prophage ΦMM1-2008. The finalized genomes were aligned using MUGSY (Angiuoli and Salzberg, 2011) to allow marker polymorphisms to be identified. To avoid the false positive identification of polymorphisms, the hypervariable *hsdS* locus (Tettelin *et al*., 2001) and highly repetitive *psrP* gene were excluded from all analyses.

2.5.3 Assembly of Illumina data

De novo assemblies of prophage, *cps* loci and partially deleted ICE*Sp*23FST81 sequences were produced from multiplexed Illumina sequence data. Assemblies were initially generated from EDENA (Hernandez *et al*., 2008), an overlap graph-based assembler. The overlap parameter was iteratively increased to obtain the highest N_{50} value. Any contigs over 1.5 kb in length were then used to generate a FASTQ file of simulated reads 350 bp in length separated by an insert size of 800 bp. These were then used in conjunction with the original data as an input for Velvet (Zerbino and Birney, 2008), a de Bruijn graph-based assembler. Velvet was optimized to run with the longest k-mer that gave an expected coverage value above 20. Contigs were then ordered against the appropriate reference sequence using ABACAS (Assefa *et al*., 2009) and ACT (Carver *et al*., 2005). Assemblies displayed in Chapter 4 were then annotated and submitted to the EMBL database (Appendix III: EMBL accession codes).

2.5.4 Detecting accessory genome components through mapping

Following extraction and assembly, accessory genome loci detected in the PMEN1 and ST180 populations were concatenated into a multiFASTA file. Illumina sequence read data were then mapped against this reference using BWA (Li and Durbin, 2010) to produce an alignment, including redundant mapping of sequences aligned to repeats, that was processed to give a coverage plot using bcftools (Danecek *et al*., 2011). These were then displayed as a heatmap relative to the phylogeny using Biopython (Mangalam, 2002).

2.5.5 Analysis of RNA-seq data

Sequence reads were mapped as paired end data using BWA (Li and Durbin, 2010). The orientation of the second read in correctly mapped pairs was reversed using Samtools (Li *et al*., 2009) before producing coverage plots, in order to maintain the directional fidelity of the data. The 'XA' note in the alignment file was used to identify alternative mapping locations. All reads were used to generated the fully redundant plot in Figures 7.6 and 7.7; reads with alternative mapping loci only within the displayed region were maintained in the set used to generate the 'locally redundant' plot, whilst those that mapped equally well to sequences outside of the displayed region were excluded. This allows reads that come from a specific BOX element, but cannot be unambiguously assigned to a particular boxB module therein, to be retained within the 'locally redundant' plot.

2.6 Bioinformatic analyses

2.6.1 Mathematical analyses

All statistical test *p* values and graphical representations were generated using R (R Development Core Team, 2011).

2.6.2 Annotation of sequences

Coding sequences were initially identified by using Glimmer3 (Delcher *et al*., 2007) and then manually curated using Frameplot (Bibb *et al*., 1984) and Artemis (Carver *et*

al., 2008). All genes were annotated in Artemis using standard criteria (Berriman and Rutherford, 2003). Genome comparisons were performed using BLAST (Altschul *et al*., 1997) and visualised using ACT (Carver *et al*., 2005).

2.6.3 Determination of serotype and sequence type from Illumina data

The sequences of 91 pneumococcal *cps* loci (Bentley *et al*., 2006; Park *et al*., 2007) were concatenated and Illumina sequence reads redundantly aligned against this reference using BWA (Li and Durbin, 2010). The locus with the highest proportion of its length covered by mapped sequence reads was taken to be that encoding the capsule.

The sequences of the seven loci used for sequence typing, along with several hundred base pairs of flanking sequence, were extracted either from the genome of *S. pneumoniae* ATCC 700669 or *S. pneumoniae* OXC141. Five rounds of Illumina read mapping were then used to iteratively transform the reference sequences into those of the sequenced isolate using ICORN (Otto *et al*., 2010). The sequences were then analysed using www.mlst.net (Aanensen and Spratt, 2005). The sequence type of *S. pneumoniae* BM4200 was independently verified from the *de novo* whole genome assembly of the strain.

2.6.4 Recombination and phylogenetic analyses

The algorithm described in Chapter 4 was implemented in order to produce a maximum likelihood phylogeny based on vertically-inherited substitutions occurring outside of recombinations. When applied to PMEN1, the algorithm was run for five iterations, but failed to converge. This was a consequence of the low probability of resolving the short, poorly supported branches at the base of the phylogeny in the same manner in subsequent iterations. Convergence was instead assessed through comparing the Robinson-Foulds distance between the trees (Felsenstein, 1989) produced by each of the iterations, which showed that the phylogenies were highly similar in the three preceding, and following, iterations. By contrast, when applied to the ST180 dataset, the algorithm converged on a stable topology by the third iteration.

2.6.5 Analysis of gene disruption events

Frameshift mutations and premature stop codons that reduced the length of CDSs relative to their annotation in the reference genome were defined as 'disruptive events'. These were reconstructed onto the phylogeny using parsimony, as PAML (Yang, 2007) is unable to reconstruct changes in sequence length. For PMEN1, 537 disruptive events were estimated to occur, giving a mean incidence of 0.278 kb^{-1} disruptive events across the 1,934,819 bp of coding sequence in the reference genome. Modelling the occurrence of these disruptions as a Poisson distributed process occurring at a rate proportional to the length of the gene, 11 CDSs exceeded a *p* value threshold of 0.05 after a Bonferroni correction for multiple testing of 2,135 CDSs. For ST180, 230 disruptive events were identified across 1,756,252 bp of coding sequence, giving a mean incidence of 0.131 kb^{-1}. Assuming the same Poisson model, only two functional CDSs exceeded the Bonferroni corrected 0.05 *p* value threshold: SPNOXC10420 and SPNOXC12950.

2.6.6 Bayesian phylogenetic analyses

BEAST (Drummond and Rambaut, 2007) was used to date the most recent common ancestors of the two 19A clades in PMEN1 (clade 'U' from the USA and clade 'S' from Spain) and the most recent common ancestors of ST180 and clade I in the analysis of the serotype 3 isolates. The program was used to analyse the final maximum likelihood tree, the topology of which was fixed, and the alignment of base substitutions occurring outside of putative recombinations using an uncorrelated lognormal relaxed molecular clock (Drummond *et al*., 2006). The tree was calibrated using the strains' dates of isolation. The ages of strains for which no precise date of isolation was available were estimated using a uniform distribution spanning the range of years from which the sample could have been isolated. The non-parametric Bayesian skyline plot was used as the tree prior to allow for fluctuation in population size (Drummond *et al*., 2005). A general time reversible model of substitution was used, but no evidence of a requirement of different rate categories was found. Data were combined from multiple runs, with the appropriate 'burn in' removed from each on the basis of parameter traces, such that all values had an effective sample size greater than 200. As validation of the application to PMEN1, the analysis estimated

that the lineage originated around 1969 (95% credibility interval 1958-1977), in concordance with the date calculated from root-to-tip distances (about 1970).

2.6.7 Alignment of assembled sequences

Phylogenies were constructed for the serotype 19A and 19F *cps* loci using RAxML (Stamatakis *et al*., 2005) by extracting these regions from the *de novo* strain assemblies and aligning them as co-linear sequences with progressiveMauve (Darling *et al*., 2010).

2.6.8 Clustering of prophage elements

Prophage sequences were extracted from the host genomes based on the positions of the autolysin and integrase genes at the two ends of such elements. Gene prediction was performed on all the prophage concatenated together using Glimmer 3 (Delcher *et al*., 2007); the putative protein sequences were then extracted and compared using BLASTP (Altschul *et al*., 1997) with an E value cutoff of 10^{-50}. Orthologue clusters were then defined by analysing these sequences with TribeMCL (Enright *et al*., 2002), using an inflation parameter value of 1.5. A Jaccard distance matrix was then constructed on the basis of the number of shared and unique orthologue clusters in each pairwise comparison between phage. This was used to produce a neighbour joining tree using Neighbor in the Phylip package (Felsenstein, 1989).

2.6.9 Analysis of repeat sequences

2.6.9.1 Identification of repeat sequences

The sequence of *S. pneumoniae* ATCC 700669 was searched for repeats longer than 50 bp using RepeatScout (Price *et al*., 2005). For each of the three families identified, multiple sequence alignments were produced with MUSCLE (Edgar, 2004), which were used to generate Hidden Markov Models (HMM) using HMMER1.8 (more recent versions of HMMER have not been optimised for searching long nucleic acid sequences for short motifs) (Eddy, 2008). In order to define the modular nature of BOX elements, HMMs representing boxA, B and C sequences individually were

produced using available sequence data (Martin *et al*., 1992; Koeuth *et al*., 1995). Sequences identified with these initial models were then aligned and used to produce the final HMMs used in this study; cutoff score thresholds were determined empirically from the distribution of scores for all hits throughout the genome. HMM logos were produced using LogoMat-M (Schuster-Bockler *et al*., 2004). Composite BOX elements were defined as two or more adjacent boxA, B or C modules. The same approach was used to generate HMMs for the repeats identified in *S. suis*. Thorough *de novo* searches for novel interspersed repeats in other species were not conducted.

In a number of cases where annotated repeat sequences overlapped, it was evident that one element had inserted into another. In such cases, for each repeat in the pair, a realignment of one repeat with the appropriate HMM was attempted using the concatenated flanking sequences of the other repeat, effectively excluding the sequence of the other element. If one of the elements had a greater bit score when realigned in such a manner, it was reannotated as a split feature into which the other repeat had inserted. The HMMs for the *S. pneumoniae* and *S. suis* repeats, and a program to automate their annotation for viewing in Artemis (Carver *et al*., 2008), are freely available from ftp://ftp.sanger.ac.uk/pub/pathogens/strep_repeats/. The annotation of repeat elements in complete *S. pneumoniae*, *S, mitis and S. suis* genomes is also available from this site.

2.6.9.2 Definition of orthologous repeat sequences

Of the 14 available complete pneumococcal genomes in the EMBL database, all except *S. pneumoniae* R6 (a laboratory derivative of *S. pneumoniae* D39, the sequence of which is also available in the database) were analysed. For each annotated repeat element, 250 bp of upstream and downstream flanking sequence were concatenated into a single 500 bp string. All pairwise sequence comparisons between strings corresponding to repeats of the same type were performed using BLASTN (Altschul *et al*., 1997). The alignments with an E value smaller than 10^{-25} were then used to cluster the strings into groups, corresponding to orthologous repeat insertions, using OrthoMCL (Li *et al*., 2003). The inflationary parameter used in clustering was set to 3, the smallest integral value that did not cluster a pair of

insertions within the same genome together (*i.e.* identify 'paralogous' insertions). The IS elements in Figures 7.3 and 7.4 correspond to all the annotated IS element transposase CDSs in the *S. pneumoniae* ATCC 700669 genome; however, in order to identify orthologous IS elements in different pneumococcal genomes, a consistent annotation across the genomes was required. To automate this, such repeats were identified as BLASTN (Altschul *et al*., 1997) matches to defined elements in the IS database (Siguier *et al*., 2006) with a nucleotide identity >95% and a length >90% of that of the reference sequence. Insertions of the same IS element type were then clustered as described for the small interspersed repeats, and the results for all IS elements subsequently combined to generate the data used in the graph.

2.6.10 Prediction of RNA secondary structures

2.6.10.1 Hypothetical structures of interspersed repeat sequences

Secondary structure predictions were produced from a multiple alignment of 30 repeat sequence examples (a random sample in the base of RUP elements; only BOX elements with the canonical $A_1B_1C_1$ structure were used) using RNAalifold (Bernhart *et al*., 2008).

2.6.10.2 Hypothetical structure of *patAB* leader sequence

Starting at the putative position of the *patAB* transcript's initiation, the downstream sequence, of a length extended by one base at a time, was extracted from the genomes of *S. pneumoniae* 99-4038 and 99-4039. For each sequence length extracted, the most stable folded structure, and its corresponding free energy at 37 °C, were calculated using the Vienna RNA package (Hofacker, 2009). The stability of these structures were then plotted against the length of the sequence.

2.6.11 Analysis of *in vitro* transformation data

2.6.11.1 Identification of recombinant sequences

For each transformant, all sites identified as being polymorphic from the whole genome alignment with a base quality greater than 50 were used to identify sequence characteristic of the donor or recipient in the transformation experiment. Recombinations were initially defined as regions containing donor alleles at polymorphic sites with no intervening recipient alleles. The ambiguous flanking regions around each recombination extended between the outermost donor SNPs identified in the transformant and the nearest flanking sites that were found to have recipient allele SNPs.

Many of the 112 secondary recombinations in the same strain were positioned very close to one another. A bootstrapping approach was used to link recombinant sequences likely to have arisen through the same recombination event. The shortest distance between all pairs of non-overlapping recombinant segments outside the primary recombination locus from all strains, in terms of the donor genome, was calculated to produce the population of test values. For each strain in turn, the shortest distance between two secondary recombinations (d_{test}) was used as the test statistic. Ten thousand distances were then randomly sampled with replacement from the test population, and this distribution then used to test the hypothesis that d_{test} was significantly shorter than expected under the null hypothesis (H_0) of recombinant sequences being positioned at random relative to one another. Multiple testing was accounted for by using a Holm-Bonferroni correction to alter the one-tailed threshold p value according to the number of secondary recombinations in the transformed strain. This test was performed 100 times for each d_{test}, with distances rejecting H_0 on 95 or more of these trials considered to be significantly close to one another and therefore likely to have arisen from the mosaic incorporation of the same strand of donor DNA. If the first d_{test} in a strain was accepted, then the second smallest distance was tested with the appropriate corrected p value threshold; this process continued until d_{test} was accepted under H_0.

2.6.11.2 Sliding window analysis

There are a number of biases that must be taken into account when investigating the impact of sequence identity on recombination distributions. The non-uniform

distribution of SNPs between the donor and recipient sequences mean that there are large numbers of low-identity polymorphisms, and hence short regions of identity between SNPs, concentrated in small regions of the genome. This structuring of the sequence divergence is not commensurate with the requirement of many statistical tests that observations should be independent. However, regions of high identity are also problematic, as recombinations can only be identified and analysed where it is possible to detect them through the transfer of polymorphisms. Hence a sliding window approach was used to ascertain the effect of sequence identity, and infer the proportion of recombinations of a given length that occurred but could not be detected. Each secondary recombinant segment was analysed independently; the $L50_R$ was used as the window size, which was moved along each base of the recipient genome with an orthologous nucleotide in the donor. Using the set of SNPs that could be identified using the Illumina data, at each position it was recorded as to whether the recombination could be detected or not. If it could be identified, then the size of the ambiguous flanking sequences, and the mean identity of the SNPs within, were recorded. This allowed the proportion of possible recombinations of the same size with the same, or greater, mean SNP identity and flanking regions of equal, or greater, length. Fisher's exact test was then used to compare the null hypotheses, that half the 112 recombinations should have a greater than expected mean SNP identity, and half the recombinations should have longer than expected flanking regions, with the observed distribution.

2.6.12 Differential gene expression and phenotype analyses

2.6.12.1 RNA-seq data

Read count and RPKM values for each CDS were calculated using a custom Perl script. Differential expression analysis with the three paired replicate samples from *S. pneumoniae* 99-4038 and 99-4039 was performed using DEseq (Anders and Huber, 2010) to analyse the read count values. The *p* values presented have been corrected through the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

2.6.12.2 Expression microarray data

Microarray analyses were performed by the BμG@S group (St. George's Hospital, London). Three replicates of paired RNA samples from *S. pneumoniae* TIGR4 and TIGR4^{PUS} were analysed using the SPv1.1 PCR product microarray (Hinds *et al*., 2002a; Hinds *et al*., 2002b). Each sample was hybridised independently against a common genomic DNA control sample and the three sets of paired replicates compared using LIMMA (Smyth, 2004; Inverarity, 2009). The *p* values presented have been corrected through the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

2.6.12.3 Phenotype microarray data

The level of respiration occurring in each well of the phenotype microarray plates was quantified as the area under the curve calculated by the Biolog analysis software. Significant differences in respiration rates between strains were assessed using LIMMA (Smyth, 2004). The *p* values presented have been corrected through the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).