Chapter 2: Materials and methods

2. Materials and methods

Materials and methods documented in this chapter are commonly used in all of the following results chapters. Detailed methods specific to particular parts of the analyses will be described in the relevant chapter.

2.1 Pneumococcal collections

Genotypes and phenotypes of *Streptococcus pneumoniae* used in this thesis comprise the densely sampled carriage collection from Maela community and data from elsewhere to give a comparative view for comparison.

2.1.1 Maela whole genome sequencing collection

2.1.1.1 overviews

The Maela cohort study, the main subject of this thesis, was conducted by Drs Claudia and Paul Turner with the original aims of investigating the natural history and outcome of *S. pneumoniae* colonisation in infancy in the presence and absence of simultaneous colonisation with other bacteria and viruses. The study was based in Maela refugee camp, which is located in North West Thailand adjacent to the Myanmarese border. It is a densely populated area of 46,133 registered refugees (mostly of Karen ethnic group) with 12% of the population under the age of 5 years (The Border Consortium 2012). In 2005, infant mortality was estimated to be 21/1000 live births (similar to Thailand) and approximately 14% of deaths in children aged less than five years were due to pneumonia.

2.1.1.2 study population

The Maela cohort study included specimens collected from infants born at Maela camp as well as a quarter of their mothers. Pregnant women were recruited in the camp antenatal clinic and subsequently randomised into the routine follow-up cohort (n=750) or the immunology follow-up cohort (n=250). Monthly nasopharyngeal swabs were taken from infants in both routine and immunology cohorts from birth and terminated at 24 months. The immunology cohort also included additional sampling for maternal colonisation. Overall, 8,386 swabs were taken and were part of the study described in (Turner, Turner *et al.* 2012).

2.1.1.3 Laboratory aspects

The swabs were collected and processed according to the WHO pneumococcal carriage detection protocol (O'Brien and Nohynek 2003). All isolates were serotyped and then tested for antibiotic susceptibilities. Serotyping was performed using latex-agglutination and Quellung reaction, while penicillin and co-trimoxazole susceptibilities were determined by disk diffusion following current **C**linical and **L**aboratory **S**tandard **I**nstitute (CLSI) guidelines. Microbiology work was done entirely at Shoklo Malaria Research Unit (SMRU) microbiology laboratory in Maesot, Thailand.

2.1.1.4 Sequenced isolates

A collection of over 3,000 single-colony isolates was randomly selected for whole genome sequencing in such a way that at least 100 isolates were recovered from each of the 30 consecutive months of studied period. DNA extraction for each isolate was performed using a RBC Bioscience MagCore HF16 platform. Our collection is tabulated in **Appendix A**.

2.1.2 PMEN14 whole genome sequencing collection

A collection of PMEN14 isolates (CC320) was also included in the study. The lineage represents a highly successful clone predominantly detected in South East Asia. The collection is fully tabulated in **Appendix B**, comprising 175 isolates, 40 of which were isolated from invasive disease. The collection came from multiple geographical regions including the Middle East, Europe, USA and other South East Asian countries between 1997 – 2009; thereby providing a view on multiple introductions of this particular global clone into a local community like Maela in chapter 3.

2.1.3 Other global MLST collections

Large pneumococcal carriage cohorts conducted elsewhere were included either to enable comparison of population structure from ranges of different geographical areas between 2006-2010 (chapter 3); or for cross-validation in the genome-wide association studies (chapter 6). Only studies with sampling size larger than 100 isolates were chosen to allow robust statistical analyse.

Table 2.1 Other pneumococcal carriage collections used in the studies

Type of study	Locations	Sample size (no. isolates)	Sampling	References
Comparison of global pneumococcal population structure from infants and children (<7 years) between 2006-2010 by MLST (chapter 3)	Massachusetts, USA	280	2007	(Croucher, Finkelstein *et al.* 2013)
	Southampton, UK	310	2006 - 2009	(Tocheva, Jefferies *et al.* 2013)
	Kilifi, Kenya	316	2006 - 2008	(Brueggemann, Muroki *et al.* 2013)
	The Gambia	445	2010	Unpublished*
	Maela, Thailand	2,492	2007-2010	Described in this thesis
Genome-wide association study (chapter 6)	Massachusetts, USA	616	2001-2007	(Croucher, Finkelstein *et al.* 2013)
	Maela, Thailand	3,085	2007-2010	Described in this thesis

*With kind permission from Dr Sarah Burr, Medical Research Council Unit, The Gambia for the data to be used in this thesis.

A global collection of PMEN14 (Taiwan^{19F}-14) used in Chapter 3 was fully list in (Croucher, Chewapreecha *et al.* 2014) as supplementary table 1.

2.2 Whole-genome sequencing

All processing and sequencing of genomic DNA for Maela pneumococcal collection was performed by the Wellcome Trust Sanger Institute’s core sequencing teams. Illumina sequencing approach was employed to generate data used in this thesis. see (Bentley, Balasubramanian *et al.* 2008) for review, All samples were sequenced as multiplexed libraries using the Illumina HiSeq 2000 analyzers on 75bp paired end runs as described in (Croucher, Harris *et al.* 2011) giving a mean coverage of 276.67 reads per nucleotide. Short reads from this study have been deposited in the European Nucleotide Archive under study numbers ERP000435, ERP000483, ERP000485, ERP000487, ERP000598 and ERP000599.

2.3 Control for sample mix up through determination of serotype and sequence type

Serotype and MLST were derived from Illumina read data as described in (Croucher, Harris *et al.* 2011). Here, any short reads were mapped against pneumococcal *cps* loci reference sequences (Bentley, Aanensen *et al.* 2006) and alternative genes detected between *dexB* and *aliA* (Salter, Hinds *et al.* 2012). Any reference locus with the highest proportion of its length covered by mapped sequence reads was likely to encode the capsule. These *in silico* derived serotypes were compared to serotypes detected from latex-agglutination and Quellung reaction for quality control purposes. Any incompatibility between *in silico* and serologically derived serotypes were re-investigated. This allowed processing and human errors including label swaps and potential sample contaminations to be detected.

Of 3,157 isolates initially sequenced, any suspected contaminations were removed from the data, leaving 3,085 whole genome sequences used in this thesis.

For nontypable (NT) serotypes, methods as explained in Salter *et al.* (Salter, Hinds *et al.* 2012) were used to confirm the absence of a capsule locus or the presence of alternative genes detected between *dexB* and *aliA*. Diversity of the NT category detected in this population is discussed in 3.2.2.3.

2.4 Sequence assembly

De novo assembly was performed through a pipeline developed by the Wellcome Trust Sanger Institute's Pathogen Bioinformatics team (Dr Andrew Page). 3,085 strains were *de novo* assembled multiple times using Velvet (Zerbino and Birney 2008), where the kmer size was varied between 60% and 90% of the read length. The assembly with the best N50 was chosen. Contigs shorter than the insert size length were filtered out because they are most likely misassemblies. The sequencing data were then used to improve further the assembly. The contigs were iteratively scaffolded and extended 16 times using SSPACE (Boetzer, Henkel *et al.* 2011) beginning with the contigs where the greatest number of reads overlap. Gaps, denoted by 1 or more N's were targeted for closure by running 120 iterations of GapFiller (Boetzer and Pirovano 2012), cycling between BWA (Li and Durbin 2009) and Bowtie (Langmead, Trapnell *et al.* 2009), beginning where the greatest number of reads overlapped. A final QC step was performed on each assembly, with the reads mapped back to the assembly using SMALT 0.5.7. The assembly pipeline gave, on average, a total length of 2,161,240 bp from 111.279 contigs with average contig length of 33,191.4 bp and average N50 of 65,656.6. Where appropriate reference genomes were not available in the public databases, reference genomes (with contigs ordered and annotated) were created from *de novo* assembly. The assembly was created as described above and ordered relative to its closest references using ABACAS v2.5.1 (Assefa, Keane *et al.* 2009) and ACT (Carver, Rutherford *et al.* 2005). Annotations were directly transferred from *S. pneumoniae* ATCC 700669 followed by manual curation. These novel references were required for mapping and will be discussed in the next part.

2.5 Sequence mapping

Mapping was used in different parts of the analyses described in this thesis (see table 2.2 for a summary). Short reads from samples were mapped onto different reference genomes using SMALT 0.5.7. Bases were called and aligned using the method described in (Harris, Feil *et al.* 2010). In brief, reads aligning to the reference with a quality score of greater than 30 were considered. For each position, a base was only called if the Phred score (Ewing, Hillier *et al.* 1998) exceeds 50. This theoretically gives an accuracy of 99.999%. The call had to be supported by at least 4 reads, with at least two on each strand. Any calls that failed the criteria were reported as unknown with character “N”.

For lineage-specific analysis (chapter 4), the final alignments also include short insertions and deletions (indels) using the pipeline developed by Dr Simon R. Harris.

Table 2.2 References used for mapping and mapping coverage generated for each dominant cluster

strain	reference	Accession number	serotype	ST	Use in the study	Genome size (bp)	% mapping
Spanish23F (ATCC700669)	Public database	FM211187	23F	81	Map against the Maela and Massachusetts collection to determine coarse population structure and capture variants (chapter 3, 4 and 6)	2221315	82.33
Taiwan19 F-14	Public database	CP000921	19F	236	BC1-19F reference chapter 4	2112148	96.79
INV200	Public database	FQ312029	14	9	BC3-NT reference chapter 4	2093317	91.42
G54	Public database	CP001015	19F	63	BC7-14 reference chapter 4	2078953	96.22
SMRU 1949	Draft genome*	ERR057930	23F	802	BC2-23F reference chapter 4	1935768	96.53
SMRU 2513	Draft genome*	ERR064018	6B	315	BC4-6B reference chapter 4	1991123	95.92
SMRU 1861	Draft genome*	ERR057842	23F	2218	BC5-23A/F reference chapter 4	1896242	94.91
SMRU 1478	Draft genome*	ERR054427	15C	4209	BC6-15B/C reference chapter 4	1933435	96.73

Draft genome* in Table 2.2 indicates references generated from draft genome assemblies.

2.6 Visualisation of phylogenetic trees

Display and manipulation of phylogenetic trees was performed using the online tool Interactive Tree of Life (Letunic and Bork 2011) and the software package Circos (Krzywinski, Schein *et al.* 2009).

2.7 Statistical analyses

All statistical tests and associated diagrams were generated in R version 2.11.1(R Core Team 2014). Statistical analyses were discussed in relevant sections in the text.