2. Materials & Methods

This chapter includes methods that were used in several of the results chapters. Many of these methods form part of in-house scripts or pipelines that have been created by the Pathogen Informatics team or members of the Pathogen Genomics group at the WTSI, and these are indicated as such. Methods that are specific to certain analyses are described in the relevant chapter.

2.1 Culture and DNA extraction

All culture and DNA extraction of isolates for this thesis was performed by collaborators. Isolates were grown at 37°C on BCYE agar for 48-72h prior to DNA extraction. DNA was subsequently extracted using either the Wizard (Promega UK, Southampton, UK), PurElute (VH Bio, Gateshead, UK) or DNease Blood & Tissue (Qiagen) kits, according to the manufacturer's instructions. This was eluted in 1x Tris-EDTA (TE) buffer (pH 8.0) and quantified using a Qubit Fluorometer (Life Technologies Ltd, Paisley, UK).

2.2 Whole genome sequencing

All processing and sequencing of genomic DNA was performed by the core sequencing facilities at either the WTSI or PHE, unless stated otherwise in the relevant results chapters. Paired end libraries were created by these teams as described in previous publications (Quail *et al.*, 2012; Dallman *et al.*, 2014) and most samples were sequenced using the Illumina HiSeq platform and paired-end reads of 100 bases. Any deviations from this are also described in the relevant results chapters.

2.3 *De novo* assembly of Illumina sequence data

All assemblies were produced from the Illumina data using a pipeline developed by the Pathogen Informatics team at the WTSI. This firstly uses Velvet Optimiser (http://bioinformatics.net.au/software.velvetoptimiser.shtml) to determine the optimal kmer size to use before using Velvet to produce the assembly (Zerbino & Birney, 2008). The assembly was further improved using SSPACE (Boetzer *et al.*, 2011) to scaffold the contigs of the assembly and GapFiller (Boetzer & Pirovano, 2012) to close gaps of 1 or more nucleotides.

2.4 Control for sample mix-up through determination of sequence type

The sequence type (ST) of each isolate was derived from the *de novo* assembly using an in-house script at the WTSI. This was compared with the ST that the isolate had previously been designated using the standard laboratory protocol for sequence-based typing (SBT), to help verify that the sample had not been involved in a mix-up during the culture, DNA extraction or sequencing procedures (http://www.hpa-bioinformatics.org.uk/legionella/legionella_sbt/php/sbt_homepage.php).

2.5 Mapping of Illumina sequence data

Illumina sequence reads (in fastq format) were mapped to different reference genomes using SMALT v0.7.4 (http://www.sanger.ac.uk/science/tools/smalt-0) or BWA-MEM (Li & Durbin, 2009) (see results chapters for details). In either case, an in-house pipeline at the WTSI was used to call bases and identify SNPs using SAMtools (Li *et al.*, 2009), mpileup and BCFtools. Various filters were applied to ensure high accuracy base calling (**Table 2.1**). Any positions that did not pass the filters were called as "N" in the alignment. Additionally, any reads that mapped to more than one region equally well were discarded.

Table 2.1. Filters that were applied to the mapping and base calling of Illumina sequence data against a reference genome.

Filtering criteria	Threshold
Minimum base quality	50
Minimum mapping quality	30 (SMALT); 20 (BWA-MEM)
Minimum number of high quality reads matching base	4 (SMALT); 8 (BWA-MEM)
Minimum number of high quality reads on each strand matching base	2 (SMALT); 3 (BWA-MEM)
Minimum proportion of high quality mapped reads matching base	0.75 (SMALT); 0.8 (BWA-MEM)
Allele frequency	Within 0.5 of 1 (for a SNP) or 0 (for a non-variant)
Minimum strand bias p-value	0.001
Minimum mapping quality bias p-value	0.001
Minimum tail distance bias p-value	0.001

2.6 Phylogenetic analysis

Maximum likelihood phylogenetic trees were constructed based on variable positions within the core genome alignment using RAxML v7.0.4 (Stamatakis, 2006), usually after the removal of recombined regions where recombination detection was possible (see individual results chapters). The GTR+GAMMA method for among site rate variation was used and 100 bootstrap replicates were performed to assess support for nodes unless specified otherwise. In order to scale the branch lengths by the number of SNPs, SNPs were reconstructed onto the phylogeny using accelerated transformation parsimony (Farris, 1970), performed with a script written by Dr Simon R. Harris. Phylogenetic trees were visualised using Figtree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

2.7 Statistical analyses and figures

Statistical analyses were performed using R version 3.0.0 (R Core Team, 2013). Figures were also generated in R and using Adobe Illustrator CS5.