8. Methods

This section includes all bioinformatics methods used in this thesis. Laboratory techniques and experiments carried out by others are not included. Many of the programs described below form part of in-house scripts and pipelines made available for use by the informatics team and group members at the WT Sanger Institute, and are stated as such.

8.1. Illumina sequencing

The DNA pipeline teams at the WT Sanger Institute carried out all library preparation and sequencing unless stated otherwise. All sequencing was carried out in a paired end and multiplex (12-96 samples per run) fashion on the GAIIx, HiSeq2000 or MiSeq platforms.

8.2. Mapping of sequencing data to a reference sequence

Paired-end reads (fastq file format (Cock, Fields *et al*. 2010)) were mapped to a suitable reference (Table 10) using the program SMALT (v0.5.8) (Ponstingl 2011) uniquely (reads with multiple best matches are discarded) using default parameters except for:

- Maximum insert size of 1000 (-i)
- Minimum insert size of 50 (-j)
- Turn on exhaustive search were each mate of paired end reads are mapped independently (-x)
- Filter out aligned reads that don't have a certain proportion of exact matches (-y – used default of 0 unless stated otherwise)

Table 10 - Description of reference sequences used for mapping analyses. *Illumina reads were generated at a later date and were used to iteratively correct the reference using iCORN which resulted in 32 single base pair corrections.

Species	**Strain name**	**Type**	**Accession or source**
M. tuberculosis	H37Rv	Finished genome (AL123456) with corrections*.	Casali, Nikolayevskyy *et al*. 2012
M. tuberculosis	2535G	PacBio assembly	Currently unavailable
M. abscessus	CU458896	Finished	NC_010394.1
M. abscessus	3_k, subspecies *bolletii* representative	*De novo* assembly of Illumina reads	ERR115028 (raw reads)
M. abscessus	22_e, subspecies *massiliense* representative	*De novo* assembly of Illumina reads	ERR115082 (raw reads)

8.3. Calling and filtering variants from mapping data

Samtools and bcftools (Li, Handsaker *et al*. 2009) were used to call bases as part of an in-house pipeline (written by Simon Harris). Appropriate filters were used to reduce the number of false positive SNP calls to a level estimated to be less than one SNP per genome (Harris, Feil *et al*. 2010). Filters were:

- The minimum base quality to call a base is 50 (Phred score)
- The minimum mapping quality to call a base is 30
- The minimum number of high quality reads mapping to call a base is 4.
- The minimum number of high quality reads mapping on each strand to call a base is 2.
- The minimum proportion of high quality mapped reads that must match the called base is 0.75
- Minimum P value for strand bias, base quality bias, mapping quality, and end bias of 0.001

Any positions that failed the quality criteria were called as 'N' in the final alignment.

8.4. *De novo* assembly of sequencing reads

Raw sequencing reads were assembled using Velvet v1.2.03 (Zerbino and Birney 2008) using an in-house script which optimises the k-mer (hash length) with an expected depth of coverage of 20.

8.5. Multiple sequence alignment

Sequences (FASTA format) were aligned using Muscle v3.8.31 (Edgar 2004) or MAFFT vs. 7 (Katoh and Standley 2013) with default settings.

8.6. Construction of maximum likelihood phylogenetic trees

Maximum likelihood phylogenetic trees were constructed using RAxML v. 7.0.4 (Stamatakis 2006) using the GTR+GAMMA method for among site rate variation and 100 bootstrap replicates. The final tree was created using the maximum likelihood

tree topology with the calculated bootstrap values drawn onto the bipartitions. In some cases it was useful to map the variants back onto the resultant tree topology in order to scale the branch lengths by number of SNPs. This was carried out using an in-house script (written by Simon Harris) which utilised ACCTRAN parsimony algorithms (Farris 1970).

8.7. Path-O-Gen analysis

Path-o-gen (Rambaut 2007) is a program used to assess the presence of molecular clock in a dataset. Using a phylogenetic (often maximum likelihood) tree as input, it plots the relationship between root-to-tip distance for each taxa with its isolation date. Under the assumption of a strict molecular clock there should be a strong positive linear relationship between root-to-tip distance and time. The program can be used to root the tree in the position most consistent with this relationship. This power of this analysis is limited by its non-independence as many taxa will share evolutionary history and therefore their root-to-tip distance will not be independent. Therefore where this has been used, P values have not been stated.

8.8. Bayesian molecular evolution analysis

The BEAST package (v1.7.5), a program used for Bayesian Markov chain Monte Carlo (MCMC) analysis of genetic sequences, was used to estimate the mutation rate and the age of phylogenetic nodes (Drummond and Rambaut 2007). BEAST requires XML files as input where all the priors are set, which were created was using the GUI BEAUTi (Drummond and Rambaut 2007) and nucleotide alignments and the associated dates of isolation as input. For all analyses three independent MCMC chains of 100,000,000 states were run using a GTR model of evolution and a variety of different clock and population size models. Tracer (v1.5) (Drummond and Rambaut 2007) was used to assess convergence (after an initial burn-in period of 10,000,000), agreement between the three runs and that all effective sample size (ESS) values were greater than 200. When the uncorrelated lognormal relaxed clock was used, there was no appreciable probability mass in the marginal posterior distribution of the standard deviation of the clock rate (ucld.stdev) that overlapped with zero, so a strict clock was not deemed appropriate. For each dataset tested, one

run with the best ESS values was used to produce a maximum clade credibility tree in TreeAnnotator v1.7.1 (Drummond and Rambaut 2007), from which the estimated age (and the 95% higher posterior density intervals) of the internal nodes were extracted.

8.9. Statistical analyses and figures

Statistical analyses were implemented in R version 3.0.0 (R Core Team 2013). Figures generated using R or Microsoft Excel version 14.3.9 (Microsoft Corporation 2011).

8.10. Detection of heterogeneous sites / minority variants

Many heterogeneous or minority variants found in mapping data are due to sequencing error or mismapping, therefore strict filters are required to distinguish these errors from true minority variants which are the result of a mixed infection or clonal variation. After mapping to a reference, positions where two or more possible variants were called were extracted, and then each variant had to pass certain quality criteria:

- Supported by at least two reads on both forward and reverse strands
- Overall quality greater than 100
- Minimum P value for strand bias of 0.05
- Minimum P value for base quality bias, mapping quality, and end bias of 0.001
- Depth of coverage within normal range (+-50% of the average calculated from bam file)
- At least 200bp from another variant.

These criteria were chosen on the basis that these are the requirements for calling consensus SNPs (section 8.3), however it was found that the strand bias and mapping parameters were required to be more stringent when considering minority variants. Using the standard mapping parameters (see section 8.2), many of the detected

minority variants were found in GC rich regions clustered together (Figure 44). These were due to sequence specific errors that frequently appear after a GGC/GCC motif in Illumina reads (Nakamura, Oshima *et al*. 2011). In addition the mutation spectrum of the minority variants detected closely matched the error profile of these regions (Figure 45).

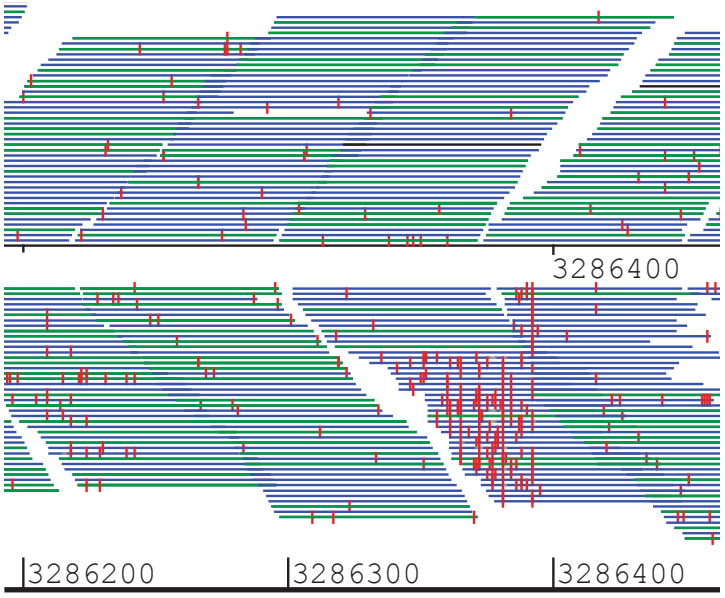
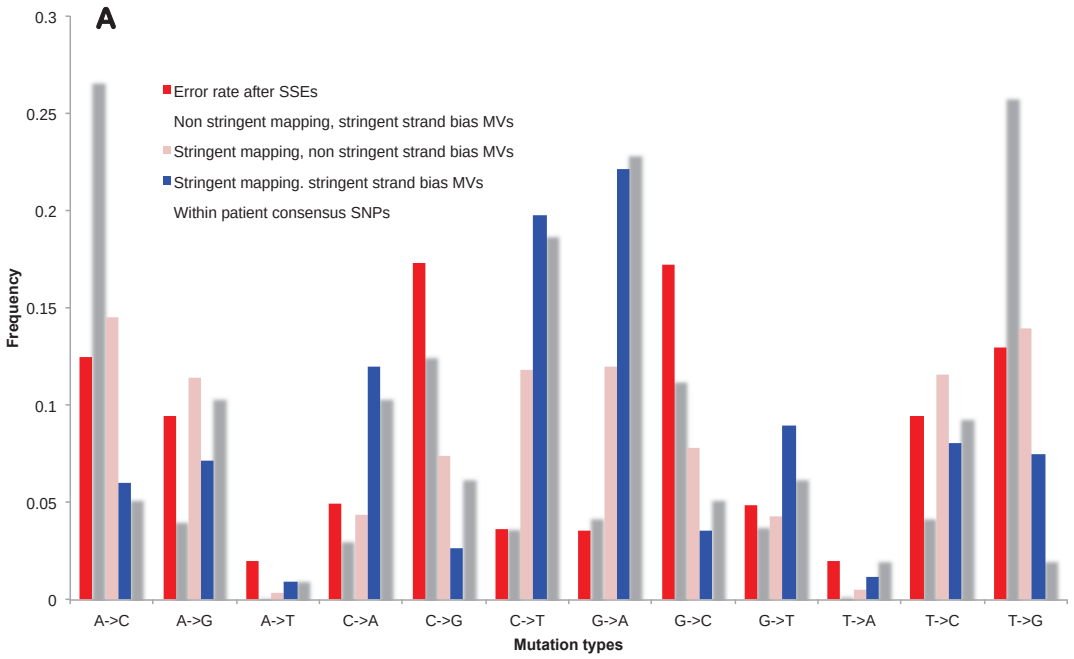


Figure 44 - Example of sequencing errors that occur after GGC/GCC motifs on the reverse strand. Each read is represented by a blue or green line, which are split into those that map to the forward and reverse strands. Red marks indicate bases which have been called as a different nucleotide to the reference genome. Numbers along the bottom indicate the position on the *M. abscessus* (CU458896) reference. Strand specific errors are found clustered between 3266343 and 3286392. Figure made and adapted from Artemis.

In order to avoid these variants, the stringency of the mapping was improved by using a 0.95 value for minimum match (section 8.2). This means any reads of length 75bp with 4 or more mismatches were discarded. In addition it was found that an increased stringency for strand bias was required, with a minimum P value of 0.05 rather than 0.001. In combination, these two additional parameters resulted in a mutation spectrum that highly resembled the mutation spectra of consensus SNPs (Figure 45 - blue).



B

	Error rate SSEs	Non stringent mapping	Stringent mapping, non stringent strand bias	Stringent mapping, stringent strand bias
Error rate SSEs				
Non stringent mapping	0.69			
Stringent mapping, non stringent strand bias	0.45	0.63		
Stringent mapping, stringent strand bias	-0.36	-0.15	0.47	
Consensus SNPs	-0.28	-0.29	0.42	0.93

Figure 45 - Mutation spectra of minority and consensus variants. A) Sequence specific error (SSE) profile obtained from Nakamura *et al* 2011. Other mutation spectra were obtained from *M. a. massiliense* dataset comprised of the two transmission clusters. Stringent and non stringent parameters are described in the text. MV = minority variant. **B)** Pearson's R coefficient values between mutation spectra, with colours representing the strength of the relationship. The mutation spectra of the consensus SNPs was found to correlate highly with the minority variants detected through stringent mapping and stringent strand bias.

Variants were also only included if they were at least 200bp from another possible minority variant; this was to exclude heterogeneity which may be the result of mismapping. This distance may need to be reduced in non-mycobacterial organisms where the mutation rate is sufficiently high enough for mutations to frequently occur within 200bp of one another.

The number of minority variants detected was found to significantly correlate with the depth of coverage (Figure 46), which is expected as deeper sequencing would increase the ability to both detect minority variants and for them to subsequently pass quality filters. Therefore, when comparing the number of minority variants between different samples, the data was normalised to 100 fold depth of coverage using the following calculation:

$$corrected\ no.of\ minority\ variants = \frac{no.of\ minority\ variants}{average\ depth\ of\ coverage} \times 100$$

This normalisation resulted in a loss of any observable correlation between the two variables (P value 0.2, correlation coefficient 0.02).

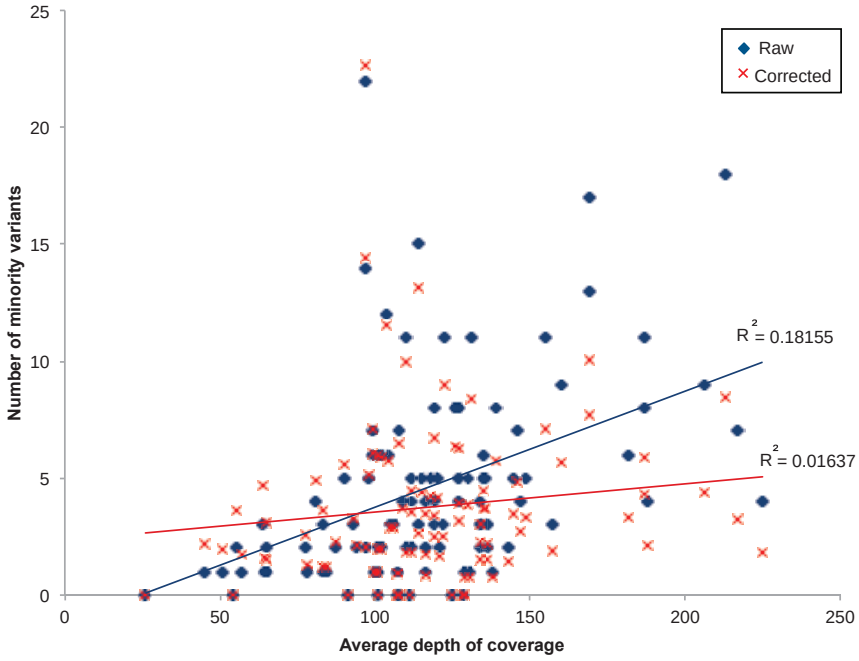


Figure 46 - Correcting the number of minority variants for depth of coverage. Analysis of the *M. abscessus* dataset (Chapter 6, excluding the hypermutator) revealed a highly significant (P value 8.893e-06) positive linear relationship between the raw number of minority variants detected and coverage (blue). When normalised for coverage (red) no significant correlation could be observed (P of 0.2).