Simon Kneringer, BSc

Genome characterization of multi resistant *Mycobacterium tuberculosis* strains

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Information and Computer Engineering

submitted to

Graz University of Technology

Supervisor

Dr. Gerhard Thallinger

Institute of Computational Biotechnology

Institute of Neural Engineering

Graz, October 2017

AFFIDAVIT[1]

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Graz, ______________________ ______________________

Date Signature

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Tuberculosis (TB) is the most lethal bacterial pathogen on the planet and treatment is getting harder due to emerging drug resistances. Without treatment, 10% of TB infections develop into an active form, 50% of which are lethal. Recently a number of TB strains evolved which are resistant to either multiple or all available antibiotics. Reasons for this resistance remain still elusive.

This thesis proposes a workflow to detect known causes for drug resistances by identifying associated single nucleotide polymorphisms (SNPs) based on whole genome analysis. Additionally phylogenetical relations and evolutionary origin of the examined pathogens are investigated.

High quality SNPs are identified by combining the results of two variant calling pipelines. SNPs found via local re-assembly of haplotypes in a reference based mapping and SNPs found by *de-novo* assembly and reference comparison are combined. SNPs that both methods agree on are subsequently filtered and compared with the tuberculosis drug resistance mutation database. For the phylogenetic analysis a maximum likelihood based tree and a Minimum Spanning Tree are generated.

The implemented workflow was applied to data from 64 newly sequenced, drug resistant *Mycobacterium tuberculosis* strains, which were collected and sequenced at the University of Medicine and Pharmacy in Iasi, Romania. SNPs related to drug resistances were found in all but one strain; Two strains were resistant to six of the seven investigated drugs. The phylogenetic analysis assigned four different lineages to the examined clinical isolates.

Contents

1. Introduction

Mycobacterium tuberculosis (MTB), a Gram-positive bacterium, is the main causative agent of tuberculosis (TB). Tuberculosis is a lung disease that presents millions of new cases and deaths every year and is considered a major threat to public health [1].

Fast and reliable diagnosis is a major factor in the treatment of TB. Emerging drug resistances worsen the situation, as the treatment duration gets longer and access to the right drugs harder. Many people lack access to the required facilities for an adequate diagnosis. Research for a better understanding of the mechanisms of antibiotic resistance in general has to be carried out and a fast and reliable analysis method to diagnose known drug resistances has to be found.

Statistical analysis of the evolution and spreading of the disease is a vital part in better understanding the global TB epidemic.

There are many ways for analyzing microbial DNA in context to phylogenetical analysis. This section of the thesis shows a overview of recent methods for phylogenetic inference based on whole genome sequencing. There is a mass of steps involved in preprocessing, sequencing and post-processing a DNA sample, both on the molecular biology and the bioinformatic level. As this thesis focuses on bioinformatics, the involved molecular biology preprocessing and the loading of the sequencer will not be discussed in detail. However, the sequencing technology, as well as the underlying technological implementation for generating reads of DNA, are very important for understanding the digital output of those machines. Therefore a brief overview of the used next-generation sequencing technology will be included in this thesis.

The processing of the generated sequencing output, from raw reads to a human readable representation of the results, involves a variety of different programs and algorithms. Each step will be described in detail during the development of this thesis.

1.1. *Mycobacterium tuberculosis* and tuberculosis

Tuberculosis is a highly contagious disease that generally affects the lungs and is transmitted in airborne particles. Worldwide TB is one of the top ten causes of death with an estimated 1.8 million deaths in 2015 [1].

Mycobacterium tuberculosis is a the causative agent of most cases of tuberculosis. Without treatment, 10% of TB infections develop into an active form, 50% of which lead to death.

The first MTB strain (H37Rv) was sequenced in 1998 by Cole *et al.* [2,3] and since then around 5000 MTB genome sequences were made publicly available. Recent phylogenetic inferences reveals that human-adapted MTB has a strong phylogeographical structure and covers seven major lineages that descended from a common ancestor [4]:

- Lineage 1: Indo-Oceanic
- Lineage 2: East-Asian (including the Beijing family)
- Lineage 3: East Africa, Central and South Asia
- Lineage 4: Euro-American
- Lineage 5: *M. africanum* West African 1
- Lineage 6: *M. africanum* West African 2
- Lineage 7: Ethiopian lineage

Samples used in this thesis are expected to belong to lineage 4 (also known as Euro-American), as it is mainly distributed in Europe, but also in America, Africa and the Middle-Eeast. Stucki *et al.* [5] have shown that lineage 4 comprises ten separate sublineages, including the virulent Haarlem strain.

1.2. Mtb drug resistances

Recently, a number of TB strains evolved which are resistant to either multiple or all available antibiotics. Reasons for this resistances remain still elusive. The WHO defines two groups of drug resistant TB [1]:

- Multidrug-resistanct tuberculosis (MDR-TB) is resistant against the two most powerful anti-TB antibiotics: Isoniazid and Rifampicin.
- Extensively drug-resistant tuberculosis (XDR-TB): in addition to the Isoniazid and Rifampicin resistance, the pathogens of this group is also resistant to any of the Fluoroquinolones, as well as resistant to a injectable seond-line drug like Streptomycin.

Treatment of patients with either MDR-TB or XDR-TB uses costly backup drugs, that can have drastic side-effects, like blindness and psychological disability [6]. In 2015, there were an estimated 480,000 new cases of multidrug-resistant TB with a treatment success rate of 52% [7].
Fast and accurate detection of drug resistances can help to improve the success rate, as well as the treatment duration and costs.

1.3. Ion Torrent sequencing

Ion Torrent is a next generation sequencing (NGS) method based on the ion semiconductor sequencing technology. The technology was developed by Torrent Systems Inc and is now owned by Thermo Fischer Scientific. The Ion Torrent method of sequencing detects hydrogen ions that are released during the polymerisation of DNA. A special semiconductor chip with up to 5 millions microscopic wells is filled with copies of single-stranded template DNA molecules and DNA polymerase. During the sequencing run, the chip is sequentially flooded with unmodified A, C, G or T deoxyribonucleotide triphosphates (dNTPs) [8]. When the introduced dNTP is complementary to the next nucleotide on the template, it is attached by the polymerase. In this process a hydrogen ion is released and detected by the chip.

Except for single molecule sequencing technologies like nanopore sequencing [9] and the Pacific Biosciences zero-mode waveguides [10], all other NGS sequencing methods do not directly detect nucleotides. For the Ion Torrent method, the change in pH level is used as the indicator if a certain nucleotide is present.

This procedure introduces a possible error source known as homopolymer error. If a homopolymer stretch of a base is present in the sequence, the DNA polymerase in the chip well will incorporate multiple complementary dNTPs. The pH level in the well rises proportional to the number of consecutive bases and must be quantized by the sequencer to determine the correct number of bases in the homopolymer stretch. If the correct number of bases cannot be detected correctly an indel error occurs. In the Ion Torrent PGM homopolymer-associated indel errors can be as frequent as 1.5 errors per 100 bases [11].

A modern Ion PGM can output up to 2 Gb (2 billion bases) in a single sequencing run of around 7 hours [12].

1.4. Whole genome analysis

Next generation sequencing technologies made it possible to sequence whole genomes cost efficiently. This enabled a broad range of new genomic analysis methods, including variant detection in whole microbial genomes.

A typical workflow for variant detection based on whole genome sequencing of microbial DNA [13], is shown in Figure 1.1. After extracting and preparing the DNA library, the sequencing is carried out on a high throughput NGS system like Ion Torrent or Illumina [14]. To detect variants in the sample, its sequence must be compared to a known reference. Choosing the right reference is as crucial as the quality of the reference sequence. To get the DNA sequence of the sample, two possible methods are available. The raw reads from the sequencing experiment can be mapped directly to the reference, or the reads are first assembled into a sequence using a process called *de-novo* assembly.

The mapping approach relies on a previously assembled genome as a reference. Raw reads are aligned independently to the reference on the most likely position. Problems can occur when repeated regions are present in the reference genome and the read length is short.

In comparison to the mapping approach, *de-novo* sequencing tries to reconstruct a novel genome without the use of a reference sequence.

The *de-novo* assembler tries to recreate the original sequence by overlapping the raw reads into long coherent sequences. These larger parts are called contigs. Multiple contigs are then aligned into scaffolds by sorting them into the right order. These scaffolds would ideally represent the full chromosome. For a perfect assembly high coverage, good read quality and high read length would be necessary.

Genomic variants like single nucleotide polymorphisms, insertions, deletions of larger structural variations can then be extracted from the results of the mapping or assembly stage. The obtained variants are post-processed and used for further analysis.

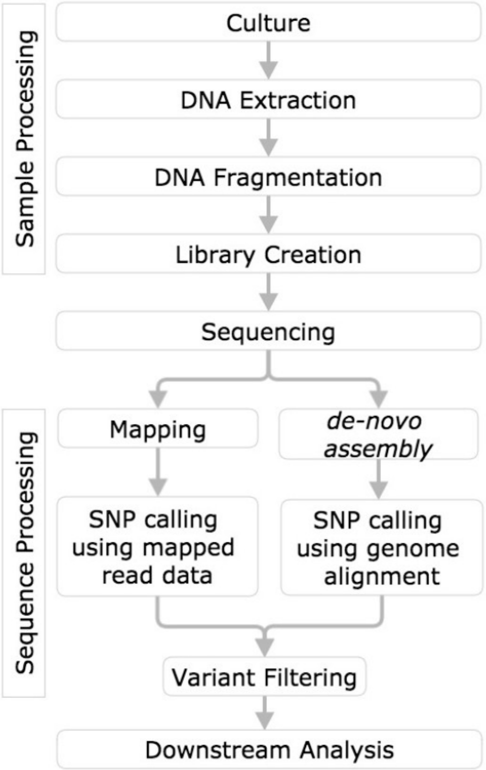

Figure 1.1.: *Variant detection workflow diagram*: Sample and sequence processing of a typical SNP calling workflow. Either a mapping or a *de-novo* assembly approach can be used (figure adapted from [13]).

Executing each step of this workflow can be done manually by choosing tools from an established toolbox, or by using a program that combines all necessary steps. Gardner and

Hall [15] built a tool called *kSNP* that can detect SNPs of hundreds of microbial genomes using k-mer analysis that requires no multiple sequence alignment. The *Harvest* suite from Treangen *et al.* [16] can perform sequence comparisons of multiple sequences simultaneously utilizing a fast core-genome multi-aligner.

1.5. Thesis Objectives

The overall goal of this thesis is to compare the genomic sequence of newly sequenced and publicly available TB strains, to identify the genetic basis of the resistances and their evolutionary relation [17,18].

The specific goals of this thesis are:

- Literature research on resistant MTB strains and methods for the phylogenetic analysis and comparison of whole genomes.
- Publicly available MTB genomes representing all currently known subgroups should be downloaded and combined with available metadata
- Development of a workflow for assembling the newly sequenced strains into contigs using *newbler* and into a single scaffold using an *R* script.
- Identification of common and distinct SNV (single nucleotide variations) and larger structural variations in the genomes.
- Finding the closest subgroup and building an ancestry of the newly assembled strains.
- Development of scripts for the annotation of the variants and the generation of a phylogeny (ancestry).
- Finding the relation of the variations distinct to the multi resistant strains to biological functions.

2. Methods

In this chapter all methods that are needed to go from raw read data to a full genome characterization, including a SNP based phylogenetic analysis and a drug resistance report, are described. The sections of this chapter follow the sequence in which the data was processed. Starting with the sourcing of raw experimental data from the sequencer to the generation of figures.

2.1. File manipulation

Working with a lot of different samples causes some problems in terms of data-management on a computer system. The large raw data files were stored and processed on the Institute's computing cluster. A project folder (`mtb`) was created on the 22 TB network attached storage (`/nas2`). Raw sequencing data for each strain is was downloaded to the `raw` folder. The output of the mapping pipeline is stored in the `out` folder. *De-novo* assembly data is stored in the `newbler` folder. All further calculation for SNP processing, phylogenetic and drug resistance analysis were less computationally intensive and therefore carried out on a local workstation. All scripts and output data is stored on the Institute's SVN [19] server.

There are many cases in which a lot of data needs to be processed similarly. The following section shows some helpful tips and tricks for data and file manipulation. All examples work on Linux and are either single bash commands or small scripts in varying scripting languages.

2.1.1. bash *find* and *execute*

This is a very useful CLI command combination, which can help you to execute programs on a set of similar files. It utilizes the very extensive feature-set of the *find* command on UNIX-like operating systems (for example Linux). For better understanding of these command a real-world example is shown in Listing 2.1. This command sequentially starts the *GATK* [20] HaplotypeCaller for each input file in the current folder.

```
find -path ./test -prune -o -type f -name '*.bam' -exec sh -c 'java -
    jar GenomeAnalysisTK.jar -T HaplotypeCaller -nct 16 -R ref.fasta -
    I $0 -ploidy 1 -stand_call_conf 30 -stand_emit_conf 10 -o ${0%.bam
    }.vcf' {} \;
```

Listing 2.1: find-exec example.

The command searches for all files ending with `.bam` in the current folder and all recursively in all sub-folders, except in a folder named `test` (-prune). It then starts a shell (-exec sh) for each found entry and passes the path and name of the entry in the shell. Now the full path with the name of the file can be accessed via the variable `$1`. To omit the existing file extension and replace it with something else the variable can be extended to `${0%.bam}.vcf`. This replaces the `bam` extension with `vcf` and is used to define the output filename. Further explanations can be found at the corresponding man-page [21].

2.1.2. bash script

For a better documentation and traceability of the used commands, it is best practice to write every command that is needed in an analysis dedicated file on the hard-disk.

2.2. Sequence data acquisition

Raw data from the sequenced MTB strains was made available over the IonTorrent Server from the Institutul Regional de Oncologie, Iasi, Romania. As this server is located in

Romania, accessing and downloading the raw sequencing data was only possible over the Internet. Data was downloaded in the unmapped BAM (UBAM) file format and converted locally into FASTQ and Standard flowgram format (SFF) for further analysis. The reason for this approach was (i) the slow internet connection and (ii) a bug in the IonTorrent tool *bam2sff*, which resulted in corrupted SFF files. As this tool had to be recompiled after fixing the error, all file format conversions were carried out locally. For more details on how this bug was solved see: 3.1. Data acquisition and format conversion.

To download all sequenced strains automatically, an *R* script was developed. This script downloads the raw sequencing reads as well as the report statistics generated by the IonTorrent Server. Throughout this thesis, *R* version 3.3.2 [22] was used.

Another program was written for the automatic file conversion of the raw data. The UBAM data is converted to FASTQ using the *Picard* tool SamToFastq. The output is directly piped into gzip to compress the resulting file. This is repeated for each strain in the raw data folder. Each raw strain data file is also converted to the Standard Flowgram Format (SFF [23]) using a modified version of The IonTorrent *bam2sff* tool. The resulting `fq.gz` files are used as input files for the *bwa-mem* mapping and the SFF files as input for the *newbler* assembly.

2.3. Data format conversion

To convert every bam file in a folder to gzipped FASTQ the command in listing 2.2 can be used. It utilizes the *SamToFastq* utility form *Picard* tools [24] and directly pipes the resulting output data-stream in gzip which compresses the FASTQ output to save disk space.

```
find *.bam -maxdepth 1 -exec sh -c ’java -Xmx8g -jar picard.jar
   SamToFastq I=$1 F=/dev/stdout | gzip > fq/${1%.bam}.fq.gz’ _ {} \;
```

Listing 2.2: Batch conversion of BAM files into compressed FASTQ format.

A similar workflow was used to also convert each bam file to SFF. This was done using the *bam2sff* tool from the IonTorrent Torrent Suite Software. These SFF files are needed as input for the *newbler* genome assembler. Listing 2.3 shows the used command for the conversion.

```
find *.bam -maxdepth 1 -exec sh -c ’bam2sff -o sff/${1%.bam}.sff $1’ _
    {} \;
```

Listing 2.3: Batch conversion of BAM files into SFF format.

2.4. Zero coverage analysis

Analysis of non-covered regions in the reference mapping was carried out with the sequencing of a known reference strain: ATCC® 25177™ H37Ra. The sequenced reference strain was mapped against its NCBI sequence (`Accession:NC_009525/CP000611`) [25] entry using *bwa-mem* v0.7.5a [26] and *CLC Genomics Workbench* v6.5.2 [27].

Missing regions were extracted from the mapping using the genomecov command of *bedtools* v2.17.0 [28]. Mapping statistics were collected using *QualiMap* v2.0 [29]. Visualization of the qualimap results were generated using the *NGI* Visualizations tool [30].

For visualizing the non-covered regions in the ATCC mapping, the web based tool *CiVi* [31] was used. A custom coloring file was generated out of the extracted regions.

2.5. Variant calling

After gathering the raw data files from the sequencing experiments and converting them into the required input formats, variant calling can be carried out.

As the main objective of this thesis is a SNP based phylogenetic analysis, the single nucleotide polymorphisms for each strain must be found with high accuracy.
Two different variant calling approaches were implemented. The first one follows the recommendations of the Broad Institutes Genome Analysis Toolkits (*GATK*) [32]. This analysis uses a reference based alignment (mapping) to find structural variants. For the

second approach, a *de-novo* assembly is created. The resulting set of contigs is aligned to the reference genome and variants are called by comparison.

By calculating the intersection of the two sets, the output of both pipelines can be combined to a set of high quality variants.

2.5.1. Mapping pipeline

The following software pipeline is based on the *GATK* best practice guide [32,33]. Figure 2.1 shows the workflow proposed by the Genome Analysis Toolkit from the Broad Institute.

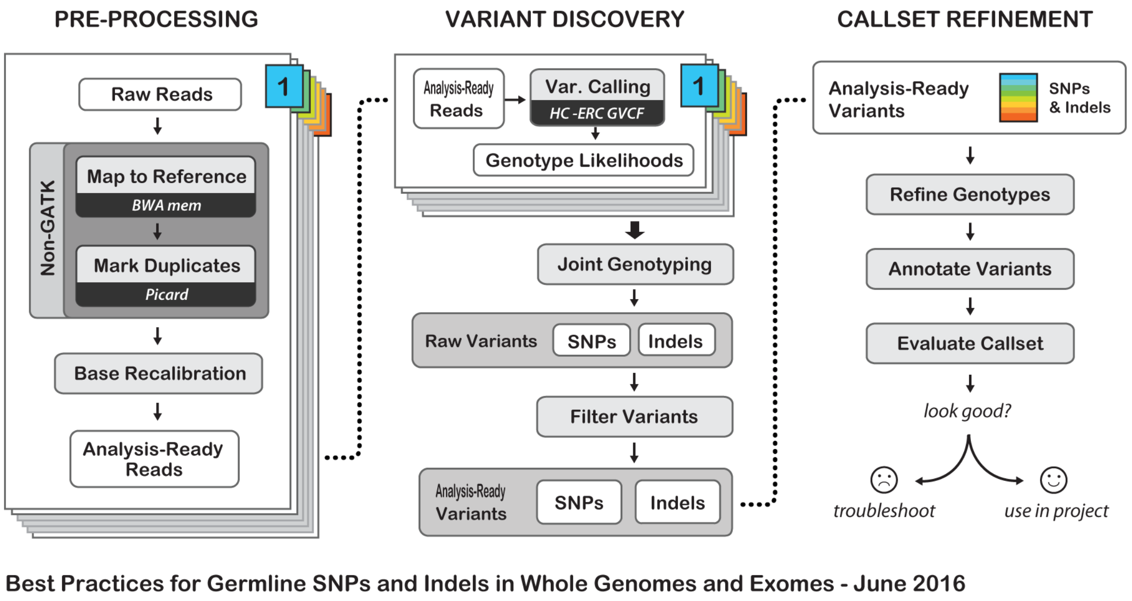

Figure 2.1.: *GATK* best practice pipeline for SNP discovery in whole genomes. Raw reads are pre-processed by reference mapping with *bwa-mem*. Duplicate reads are marked and base quality scores are recalibrated. Variants are discovered using the HaplotypeCaller. Genomic VCFs are passed to the joint genotyping tool. SNPs and Indels are called and filtered. Optionally the callset can be refined by variant annotation, Genotype refinement and callset evaluation (figure taken from [33]).

Raw reads were preprocessed by removing adapters with *cutadapt* v1.7.1 [34] and with CLC Assembly Cell *clc_quality_trim* v4.2 (quality trimming) [27] with a quality cutoff of 20. Quality reports for each strain were generated using *FastQC* [35].

Prior to mapping, the reference sequence was indexed to enable random access to the FASTA data which is required by many processing and analysis tools. For the mapper, a Burrows–Wheeler-compressed index is generated.

```
if [ ! -f ../ref/H37Rv.fna.bwt ]; then
  echo "Reference not indexed, start indexing!"
  bwa index ../ref/H37Rv.fna
  samtools faidx ../ref/H37Rv.fna
  java -jar picard.jar CreateSequenceDictionary REFERENCE=../ref/H37Rv.fna OUTPUT=../ref/H37Rv.dict
fi
```

Listing 2.4: Reference indexing with *bwa* index, *samtools* faidx and *Picard* CreateSequenceDictionary.

The raw reads for each strain were mapped with *bwa-mem* [26] to the H37Rv reference (`Accession:NC_000962.3`) [3,36,37].

```
bwa mem -a -M -t 8 -R "@RG\tID:Org\tSM:Sample\tPL:iontorrent\tLB:lib1\tPU:unit1" ../ref/H37Rv.fna $f > tmp/${fwo}.aln.sam
```

Listing 2.5: Mapping the strain raw reads against the reference using *bwa-mem*.

The variable `$f` represents the input file in GZIP compressed FASTQ file format and the variable `${fwo}` stores the strain name.

Aligned reads were coordinate sorted with *Picard* 1.112 [24] SortSam, duplicates were removed with *Picard* MarkDuplicates and indexed with *Picard* BuildBamIndex. The quality of the mappings was accessed with *QualiMap* v2.0 [29].

```
java -jar picard.jar SortSam I=tmp/${fwo}.aln.sam O=tmp/${fwo}.sorted.bam SORT_ORDER=coordinate
java -jar picard.jar MarkDuplicates I=tmp/${fwo}.sorted.bam O=tmp/${fwo}.dedup.bam METRICS_FILE=tmp/${fwo}.metrics.txt
java -jarpicard.jar BuildBamIndex INPUT=tmp/${fwo}.dedup.bam
```

Listing 2.6: Sorting, duplicate removal and indexing of the mapping using *Picard* tools.

Sorting the mapping also performs a file format change from SAM to BAM. The large unsorted SAM file is removed to save disk space.

Creating the realignment targets and Indel realignment was performed using *GATK* v2.7-4 [20] as recommended by the *GATK* documentation.

```
java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R ../ref/H37Rv.fna -I tmp/${fwo}.dedup.bam -o tmp/${fwo}.targetintervals.list
java -jar GenomeAnalysisTK.jar -T IndelRealigner --filter_bases_not_stored -R ../ref/H37Rv.fna -I tmp/${fwo}.dedup.bam -targetIntervals tmp/${fwo}.targetintervals.list -o tmp/${fwo}.realigned.bam
```

Listing 2.7: Indel realignment using *GATK*.

Up to this point, an already established software from the institute could be used. For each strain a configuration file was created and the calculation was carried out on the Institute's computing cluster. The resulting BAM files were copied to a local workstation for further analysis.

Variant calling was carried out by running the *GATK* version `3.3` HaplotypeCaller utilizing the `--ploidy 1` option, since the chromosome of MTB is haploid.

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -nct 8 -R ../ref/
   H37Rv.fna -I tmp/${fwo}.realigned.bam -ploidy 1 -stand_call_conf
   30 -stand_emit_conf 10 -o vcf/${fwo}.raw.vcf
```

Listing 2.8: Variant calling using the *GATK* HaplotypeCaller.

The results of this pipeline are stored in Variant Call Format (VCF [38]). One output file for each clinical isolate is created.

2.5.2. *De-novo* assembly pipeline

Calling variants with the *de-novo* approach was done by assembling the converted raw data in standard flowgram format (SFF) with *newbler* 2.9 [39]. The `-m` option increases the assembly speed by keeping the sequence data in memory and the minimum contig length is set to 1 by the `-a` option.

```
runAssembly -m -a 1 -o outfolder starin.sff
```

Listing 2.9: *De-novo* assembly using *newbler*.

The resulting contigs were realigned towards the H37Rv reference `(Accession:NC_000962.3)` using *nucmer* [40]. After the alignment, the result was filtered using the *delta-filter* of the *MUMmer* package. Variant calling was also carried out with a *MUMmer* tool called *show-snps*.

By using the `-I` option, the show-snps tool can exclude indels and produces output files including only the desired SNPs. This option was not used in this step to preserve all called variants. Indels were later filtered in the SNP extraction script.

```
nucmer --prefix=${1%.fasta} ../ref/H37Rv.fna ${1}
delta-filter -r -q ${1%.fasta}.delta > ${1%.fasta}.filter
show-snps -ClrT ${1%.fasta}.filter > ../snp_ncbiref/${1%.fasta}.snps
```

Listing 2.10: Contig realignment and SNP exctraction from the assembly.

This resulted in a raw tab-delimited file, which included all found indels and SNPs. These steps were repeated for each strain (Fig 2.2).

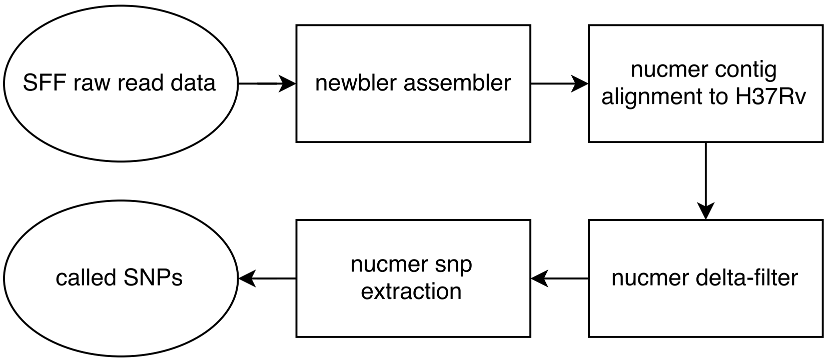

Figure 2.2.: *De-novo* assembly pipeline.

2.6. SNP extraction and variant annotation

For further analysis, out of all found variants, only SNPs were used. The SNPs for each strain from the mapping pipeline were extracted using the *GATK* SelectVariants tool.

These SNPs were prepared for annotation by cleaning up the NCBI reference sequence locus in the VCF file. After searching and replacing the H37Rv reference version number (NC_000962.3) with the accession number (NC_000962) the annotation was carried out by using *SnpEff* v4.2 [41] with its built in database for the H37Rv genome.

SnpEff also predicts possible effects of SNPs on the genes, in which the variations are present. The impact on the amino acid sequence is rated and also stored in the output. This information is used in further research for classifying newly found variants.

The annotated SNPs are stored in an VCF file. For easier processing, all files were also converted to a tab separated file format using the *vcf-toolbox* [42].

Variants from the assembly pipeline were not annotated and were instead used for quality verification in the SNP filtering section of this thesis.

2.7. SNP filtering

Before using the resulting SNP sets in the phylogenetic and drug resistance analysis, variants of low quality need to be filtered. This could be done based on the quality score or `LowQual` flag in the VCF files. In this thesis an alternative approach is proposed.

By using the variants found by both the mapping pipeline, as well as the *de-novo* assembly pipeline, a filter was implemented, which retains a variant only if it is found by both methods. A SNP found by both pipelines does exists with high confidence and is, in this thesis, therefore called a "high quality" SNP.

This method was implemented in an *R* script. The annoated VCF file from the mapping pipeline and the tabulator delimited file containing the SNPs found by the assembly pipeline are loaded and the set of SNPs found by both pipelines is extracted by calculating the intersection. This is done by the *R* function `intersect` which performs a set intersection on two vectors.

Subsequently, SNPs from a known set of repetitive DNA elements were also filtered, because of inaccuracies and mapping problems that can occur when sequencing theses regions using next generation DNA sequencing technologies like IonTorrent [43]. 273 genes were excluded, including members of the PE/PPE and PGRS gene family, mobile elements like transposases and phiRV1 members [44]. A full list can be found in Appendix B.

The resulting sets of high quality SNPs for each strain were saved in VCF format and are ready for further analysis.

2.8. Drug resistance analysis

A collection of all documented mutations associated with drug resistances in MTB was downloaded from the Tuberculosis drug resistance mutation database [45]. The database can be downloaded in Excel, text or CSV file format. As there was a problem with string quotes in the provided CSV file, the Excel file was downloaded and converted locally into a CSV file.

The known mutations were loaded from the database and a script was written, which queries the high quality SNPs for each strain against the documented mutations. For the *rpoB* gene (Rv0667), the position had to be offset by 81, because in the database the position represents - by convention - the one of *rpoB* in *Escherichia coli* [46].

Found resistances were collected in an output file for further visualization and a report was generated for each strain which can be found in Appendix A.

2.9. Phylogenetic analysis

With the processed data from all strains, full genome SNP based phylogenetic analysis was performed. A multi-alignment of all high-quality SNPs was generated using a custom *R* script and a phylogenetic tree was inferred using *RAxML* [47].

2.9.1. SNP based multi-alignment

In preparation for the phylogenetic inference, all found SNPs must be condensed in an multi alignment. This was done by implementing a simple reference based alignment script. All SNPs from each strain are extracted and written back in a matrix. Each column represents a variant event on one reference genome position. Each strain represents a row and either the reference base, or the variant base is written to the matrix. Figure 2.3 shows a schematic representation of the implemented SNP based multi-alignment algorithm.

This new sequence for each strain is then written to a FASTA file and converted to PHYLIP multiple sequence alignment format [48].

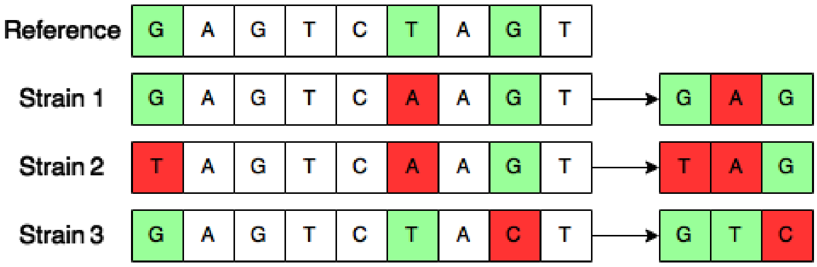

Figure 2.3.: *SNP multi-alignment*: For each SNP position in all strains an output column is generated including either the reference base or the mutation.

2.9.2. Lineage analysis

To get lineage and therefore geographical information about the clinical isolates, a set of known reference strains from the NCBI was downloaded. SNPs from these finished sequences were called using the same approach as in the *de-novo* pipeline used for the raw reads. The NCBI sequences were aligned against the H37Rv reference and SNPs were called using *nucmer* and *MUMmer* tools.

Table 2.1.: List of reference strains used in the lineage analysis.

Strain Name	Accession	Length in bp	Lineage	References
H37Rv	NC_000962.3	4,411,532	4	[3,36,37]
H37Ra	CP000611	4,419,977	4	[25]
Haarlem	NC_022350	4,408,224	4	[49]
S96-129	NZ_CM001225	4,342,996	4	[50]
UT205	HE608151	4,418,088	4	[51]
HN878	NZ_CM001043	4,404,672	2	[52]
NZXDR1	CCBK000000000	4,224,610	2	[53]

The same *R* script that calculates the multi-alignment was used to perform the lineage analysis by including the set of called SNPs from the known reference strains and calculating the pairwise distances from each clinical strain to each known reference.

An output file with the closest reference for each strain was written. Each result was compared to publicly available phylogenetic trees of the references and the nearest main lineage was chosen manually.

2.9.3. Phylogenetic inference

For inferring a phylogenetic tree, the SNP based multi-alignment in PHYLIP file format was used with *RAxML* version 8.1.21 [54]. This program uses a maximum likelihood method for the phylogenetic analysis and offers features like bootstrapping, multi-threaded execution and different models of rate heterogeneity.

The analysis for this thesis was carried out with a CAT based model of rate heterogeneity, 8 threads and using 100 rapid bootstrap inferences. *RAxML* will evaluate and optimize the likelihood of the final tree using a Γ-model [55]. The finished analysis outputs a phylogenetic tree, including all strains and the bootstrap support values for each branch, in Newick tree format [56].

Rooting of the finished tree was done with the `root()` function of the 'ape' package in *R* as part of the data visualization script.

2.9.4. Minimum spanning tree

Based on the assumption that evolution of organisms should be explained with the least amount of events a Minimum Spanning Tree (MST) [57] can be used as a phylogenetic visualization tool.

As MSTs make no assumptions about hypothetical ancestors they can be used as another method to view the connections between microbial specimens. In comparison to a tool specifically designed for phylogenetic analysis like *RAxML*, a MST can give an unbiased look at the similarities among individual strains.

An MST is a subset of a weighted, undirected graph that connects all nodes so that the sum of the distances is minimal. There can be more than one optimal solution. The result of this analysis shows a single, computationally selected MST for the analyzed data [58].

As the MST is by design ignorant of hypothetical ancestors it is not a different method to generate a phylogenetic tree. An MST is a way to represent relationships among genomes only based on their similarities.

For this thesis, the MST was obtained with *PHYLOViZ* 2.0 [59] using the `goeBURST Full MST` algorithm [60]. The SNP multi-alignment in FASTA format was used for this analysis. Meta data from the lineage analysis was converted to a tab separated file format and included in the MST.

2.10. Data visualization

To visualize the raw data quality, *FastQC* [35] was used. Similarly the mapping quality was visualized using *Qualimap* [29]. After generating a single report for each strain individually, *multiqc* [61] can be used to combine all results in a single report for a better overview of global effects. These tools were extensively used in the analysis of non covered regions.

For temporarily visualizing the resulting phylogenetic tree, *FigTree* [62] was used.

The final result of this thesis consists of a combined data visualization of the phylogenetic tree, the lineage analysis and the found drug resistances. To generate this image, a custom *R* script was written which utilizes the *ggtree* 1.8.1 [63] package.

3. Results

3.1. Data acquisition and format conversion

The first objective of this thesis, in respect to getting a better understanding of the raw material, was to download and assemble all data. This was necessary to see if the sequencing yielded good results, as there was the possibility to re-sequence certain strains, if the results were unsatisfactory.

All data was downloaded in SFF file format and each strain was assembled using *newbler* 2.9, according to the procedure described in the methods, 2.5.2 *de-novo* assembly pipeline. The program terminated for around 10% of the assemblies with no meaningful error message.

After some investigation the problem was found in the file conversion tool (*bam2sff*), a component of the Torrent Suite Software. This software is open source and available for download [64].

The file conversion script was downloaded and installed locally. Debugging the program was now possible and the problem could be traced back to an error in an 'if' statement, leading to an array index out of bounds exception.

If a sequenced read has a homopolymer repeat at the last flow an error is produced, as a loop get skipped and variable `flow_index` is not set to `0` at the homopolymers positions (leaving it to whatever value the array had on this positions). This leads to flow indexes greater than the maximum number of flows. Files with this error cannot be processed by *newbler*. Removing the unnecessary 'if' statement in the loop corrects this error.

```
468 -   while(nBase < sff->rheader->n_bases && nFlow < nFlows2)
468 +   while(nBase < sff->rheader->n_bases)
```

Listing 3.1: Changes made in the bam2sff.cpp source file.

Using the *sffinfo* tool from the 454 Sequencing System Software, which can also be obtained through the *Mothur* package [65], the read information can be displayed. Listing 3.2 shows a part of the *sffinfo* output to demonstrate the problem. In this example read NIYW6:02297:02918 was generated by the IonTorrent sequencer using 500 flows of nucleobases. The order of flows is shown under *Flow Chars*. The flowgram is not shown, but the flow indexes indicate which of the flows were valid and are used in the base sequence of this read. If a flow hits a homopolymer, the same flow index is printed multiple times and the corresponding base is also duplicated.

In this example, the last flow index is 502, which does not match a real flow, as there were only 500 flows. This is the reason, why *newbler* exits with a cryptic error message.

```
>NIYW6:02297:02918
Read Header Len:  40
Name Length:      17
# of Bases:       277
...
# of Flows:    500
Flow Chars:
TACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGC
...
TACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGA
...
Flow Indexes:
1   3   6   8   9   9   12  12  13  14  17  19  21  22  23  24  25
28  33  33  33  36  36  37  40  40  40  42  44  47  48  51  54  55
...
457 461 461 463 465 467 470 471 471 473 476 481 483 484 487 488 490
493 495 499 500 502
Bases:
tcagttggagtgtcgatATTTGGTGGGCGCAGCGGCGCGGGCACGCGCCCCAGGGCCATCGCAA
GGGCGAGCGGGGCGGCGCCCAATGCGCTCATGAGCCAGCGCCTGGTTTCGGCTTCGGCCGGCGA
TCGCGGCTTTGCCGTCGCGCATACTTGCCCGTATCGCCGACAGCCGTTCGACCGGACTGGCCAG
GTGCGTCGCCAGCGAACACATCAACGTACCGATCTTGTTGCTGGACCTGGCCGAACACGTCGAT
AACTGCGGTATCGCGCACGAA
```

Listing 3.2: *sffinfo* output of a defective read.

As IonTorrent uses an unmapped BAM format internally for processing, the raw data is unaffected by this problem. For a correct processing of all strain data, the raw read data was re-downloaded in BAM file format and the conversion to SFF was done locally using the fixed version of *bam2sff*.

Listing 3.3 shows the same read after converting the BAM file with the corrected script. The last flow index is now correctly set to 500 and indicates a homopolymer on the last flow of this read.

```
>NIYW6:02297:02918
Read Header Len:  40
Name Length:      17
# of Bases:       277
...
# of Flows:    500
...
Flow Indexes:
1   3   6   8   9   9   12  12  13  14  17  19  21  22  23  24  25
28  33  33  33  36  36  37  40  40  40  42  44  47  48  51  54  55
...
457 461 461 463 465 467 470 471 471 473 476 481 483 484 487 488 490
493 495 499 500 500
Bases:
tcagttggagtgtcgatATTTGGTGGGCGCAGCGGCGCGGGCACGCGCCCCAGGGCCATCGCAA
GGGCGAGCGGGGCGGCGCCCAATGCGCTCATGAGCCAGCGCCTGGTTTCGGCTTCGGCCGGCGA
TCGCGGCTTTGCCGTCGCGCATACTTGCCCGTATCGCCGACAGCCGTTCGACCGGACTGGCCAG
GTGCGTCGCCAGCGAACACATCAACGTACCGATCTTGTTGCTGGACCTGGCCGAACACGTCGAT
AACTGCGGTATCGCGCACGAA
```

Listing 3.3: *sffinfo* output of same read using the corrected *bam2sff* program.

3.2. Sequencing and assembly results

In total, sequencing data from 68 strain was downloaded and assembled. Sequencing coverage ranged between 0.47 and 98.96 fold.

26 strains (marked green in Table 3.1) were sequenced at least twice, because of bad loading of the IonTorrent chips in the first run. Raw data from re-sequenced strains was combined and analyzed as one big set of data. Table 3.1 shows the raw read count from the sequencer,

the number of large contigs (longer than 500 bases), the number of all contigs, the number of all bases from all contigs summed up and the coverage, for each strain.

The last three strains (8817, 597 and 5361) were discarded from further analysis because of insufficient sequencing and assembly results.

Data from 64 clinical isolates and one sequenced known reference strain (ATCC) were used in the remaining analysis steps.

Table 3.1.: *Newbler de-novo* assembly results, sorted by coverage. Green: multiple sequenced strain, all runs combined. Red: discarded because of bad sequencing runs.

Strain	Total	contigs		Total	Cov-
Name	Reads	Large	All	Bases	erage
14163	4,352,272	244	1310	4,415,127	182.20
5603	3,091,904	265	988	4,359,673	128.20
4619	2,428,509	372	739	4,326,324	109.83
16106	2,613,344	627	2175	4,462,282	108.57
5450	2,915,407	432	1336	4,365,901	106.98
11572	2,288,386	450	763	4,259,219	102.56
4082	2,319,756	507	655	4,231,314	98.96
9435	2,324,466	237	889	4,332,416	92.39
6978	1,961,553	431	856	4,275,882	85.65
1374	2,571,283	223	412	4,320,372	81.66
2196	2,156,966	334	1226	4,373,524	80.54
4538	1,630,048	400	648	4,276,256	75.27
7170	1,634,643	558	814	4,270,016	69.03
ATCC	1,581,166	547	686	4,262,033	68.18
10019	1,532,681	254	900	4,400,124	65.46
10400	1,350,868	611	819	4,274,108	62.67
12400	1,611,628	609	1169	4,363,364	60.71
12650	1,397,243	260	477	4,312,638	59.32
14114	1,362,507	1086	1449	4,189,710	58.21
9820	1,614,579	801	1168	4,246,227	55.52
5303	1,242,952	643	844	4,233,164	53.88
4793	1,265,999	805	1081	4,240,710	53.77
7193	1,268,178	771	1059	4,260,567	52.63
6185	1,191,531	873	1087	4,185,517	50.41
10862	1,153,488	2095	3428	4,340,599	48.23
9784	1,102,280	685	924	4,221,014	47.26
9937	1,214,611	670	869	4,248,795	45.69
2417	1,144,340	987	1218	4,173,564	43.85
2162	1,090,645	1179	1496	4,174,512	42.64
13449	1,043,654	1109	1464	4,173,537	42.56
8759	921,938	948	1268	4,218,144	41.92
10755	896,625	1120	1402	4,176,506	41.41
15696	1,005,011	735	946	4,212,931	40.56
13689	908,449	856	1073	4,193,201	40.31
259	1,203,472	1279	1822	4,199,222	39.85
3702	1,036,338	1031	1323	4,200,600	38.82
10237	955,832	1273	1650	4,157,165	38.72
8208	910,276	882	1084	4,178,488	38.63
2263	988,685	1169	1654	4,222,071	38.30
2435	873,385	1570	2003	3,990,854	38.23
12488	885,847	425	705	4,320,188	36.58
2599	855,039	1654	2305	4,106,012	35.97
6993	801,165	1208	1488	4,154,869	35.81
2380	918,391	1244	1555	4,125,635	35.74
16694	788,405	1331	1649	4,121,100	35.69
7924	1,192,674	1742	2598	4,167,037	35.23
260	910,252	1227	1550	4,113,063	35.08
9480	781,134	1340	1772	4,103,648	34.99
4737	852,140	1205	1510	4,138,156	34.80
7656	1,093,829	1242	1652	4,144,945	34.65
7649	746,930	1254	1535	4,109,864	34.61
12442	791,099	1415	1785	4,059,595	34.42
13050	775,724	1132	1387	4,147,796	34.09
7630	779,047	1626	2101	4,055,898	32.32
7101	724,189	1282	1575	4,117,125	32.12
3537	773,545	1652	2165	4,046,838	32.08
1770	888,623	1231	1548	4,141,865	31.41
3212	693,365	1741	2247	3,975,848	30.19
10851	669,673	1374	1702	4,056,291	29.75
15097	677,107	1168	1445	4,125,726	29.43
2025	822,803	706	1027	4,334,265	29.20
1117	651,921	1299	1625	4,078,013	29.16
10485	618,878	1300	1589	4,109,477	27.91
8139	608,378	2098	3050	3,935,290	22.45
13452	557,692	2726	5339	4,058,495	11.49
8817	155,822	1307	4602	2,119,726	4.55
597	23,070	26	496	152,281	0.69
5361	15,331	19	257	82,950	0.47

3.3. Mapping results

Before starting the time-consuming variant calling pipeline on all 65 strains, a detailed evaluation of the whole workflow was conducted.

To benchmark the used sequencing workflow and to put our own assembly and mapping results into perspective in terms of quality and quantity, the team in Romania performed a separate sequencing run of a known reference strain: ATCC® 25177™ H37Ra. The sequencing run resulted in 1,581,166 reads with a mean length of 191 base pairs yielding an estimated coverage of 68.2x.

Figure 3.1 shows the resulting coverage and GC-content of the *CLC* mapping. Those statistics were generated using *Qualimap* [29]. A large amount (nearly 40,000) of uncovered bases in the reference sequence was observed. This was very surprising considering the high total coverage and amount of raw sequencing data.

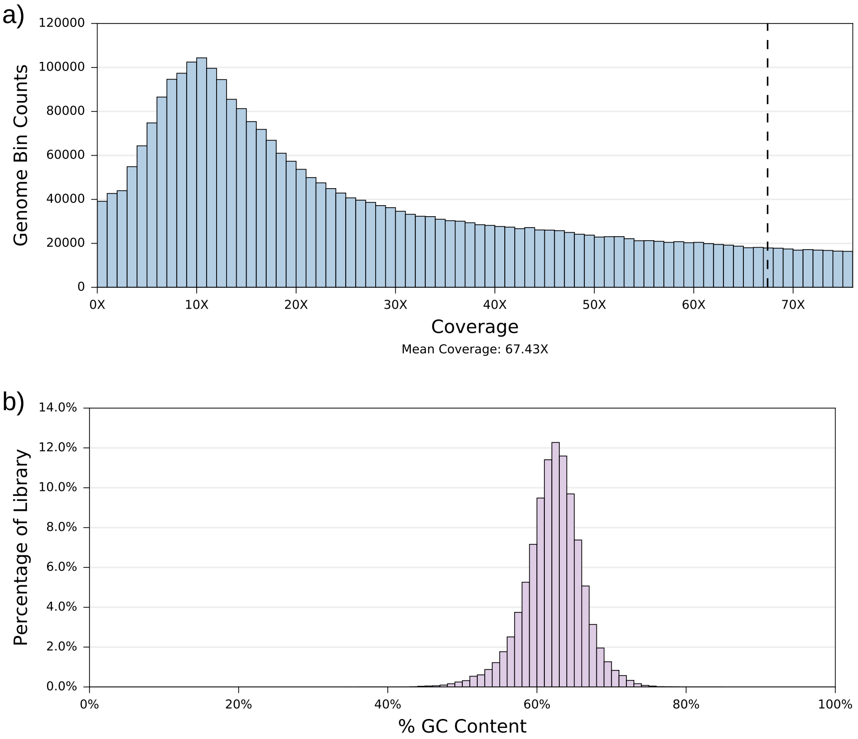

Figure 3.1.: Sequence reads of strain ATCC mapped against H37Ra. Part a) shows the coverage histogram and part b) shows the GC-content distribution. The image was created with the *NGI* Visualizations tool [30] using data from *Qualimap* [29].

For a better understanding of the reasons for the high amount of uncovered bases in the reference sequence in the mapping, further analyses have been conducted.

To rule out potential errors or artifacts in the *CLC* mapping, the same analysis was repeated with the *bwa-mem* mapper (data not shown). Even more genomic loci were uncovered when using the *bwa-mem* mapper and therefore, all further analyses were carried out on the *CLC* mapping.

Table 3.2 shows a summary of the ten longest regions with zero coverage in the *CLC* mapping of the ATCC strain. All shown regions have a very high guanine-cytosine (GC) content in comparison to the mean GC content of around 65%. There is also a strong correlation with certain gene families.

Table 3.2.: The ten longest regions with zero coverage in the mapping of the reference strain ATCC, their GC-content and affected genes.

RangeStart	RangeEnd	Width	GC[%]	Overlap	Gene Name	Product
624479	625334	855	76	855	MRA_0539	PE-PGRS family protein
837528	838149	621	79	621	MRA_0754	PE-PGRS family protein
3949791	3950387	596	77	596	MRA_3551	PE-PGRS family protein
838592	839168	576	79	576	MRA_0754	PE-PGRS family protein
3953767	3954301	534	83			not annotated
1219027	1219547	520	79	490	MRA_1102	PE-PGRS family protein
1619997	1620470	473	76	473	MRA_1449	PE-PGRS family protein
3936695	3937144	449	78	449	MRA_3547	PE-PGRS family protein
3941839	3942286	447	80	447	MRA_3548	PE-PGRS family protein
3958318	3958760	442	81	442	MRA_3553	PE-PGRS family protein

A visualization of non-covered regions is shown in Figure 3.2. There is a strong correlation between the missing coverage and the GC-content of those specific regions. For further interpretation see chapter 4. Discussion.

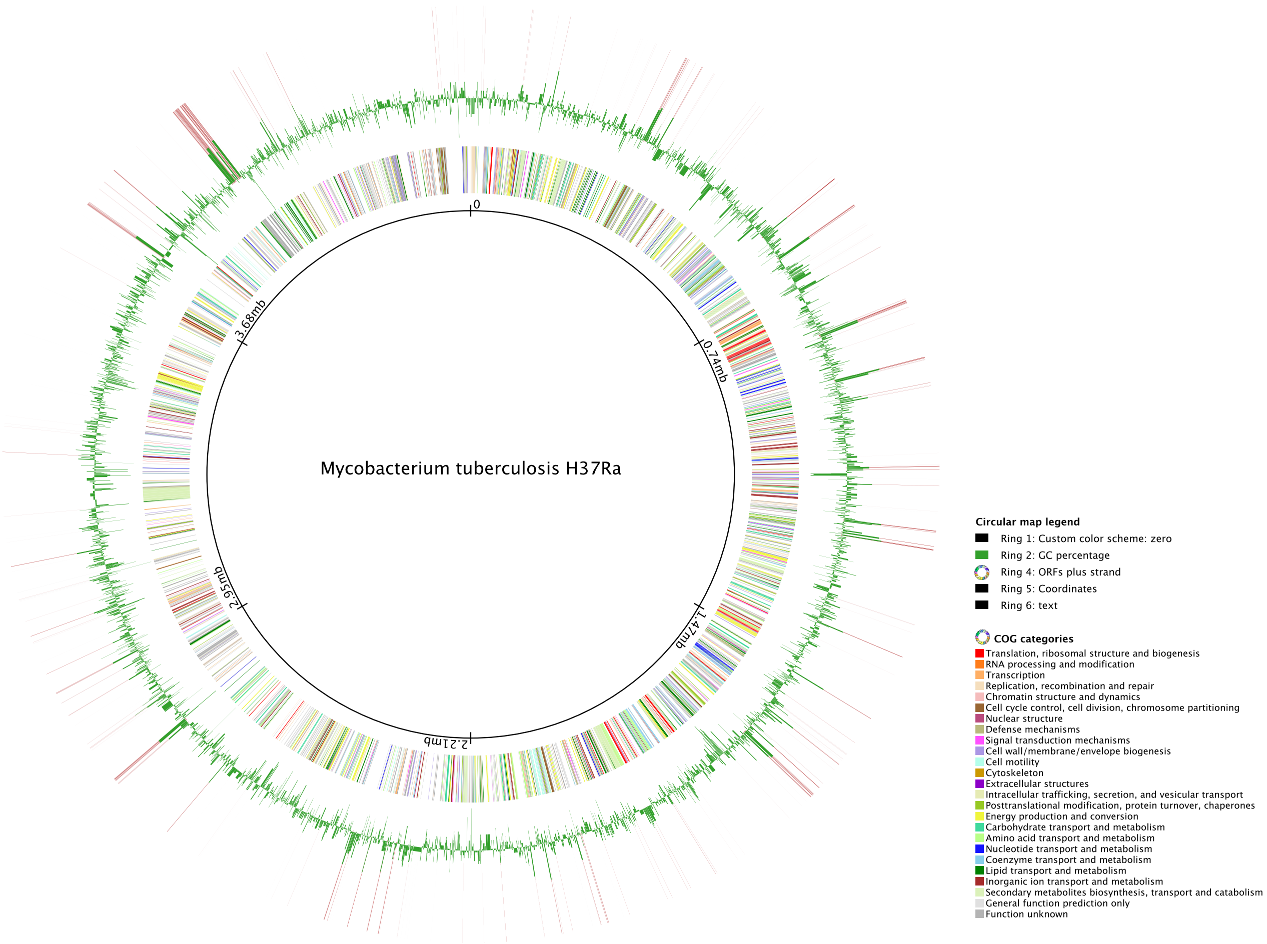

Figure 3.2.: Zero-coverage plot: the red outer segments show the non-covered regions of the mapped reference genome ATCC. The green ring indicates the GC-content of the corresponding bases. The baseline is 50%. The colorful inner ring represents the genes of the reference genome in clusters of orthologous groups (COG) [66] as standardized by the NCBI. Image create with the online tool *CiVi* [31].

The problems with non covered regions in GC rich parts of the genome were addressed by implementing a filter for known problematic elements. All strains were mapped as described in the methods using *bwa-mem*.

Between 88% and 99% of the reads could be mapped to the reference. The percentage of duplicated reads ranges between 6% and just over 50%. On average 3,4% of the reads were clipped, with the noticeable exception of strain 10862 (24,8%). Coverage ranges between 93 to 4,5 fold with an average value over all strains of 28,3x (Table 3.3).

Table 3.3.: *bwa-mem* mapping metrics, sorted by coverage. Mapping efficiency is shown in absolute and relative numbers. Green: multiple sequenced strain, all runs combined. Red: discarded because of bad sequencing runs.

Strain	Total reads	Mapped reads	Unmapped reads	Duplicated reads	Clipped reads	Coverage
14163	3,550,923	3,440,204 / 96.88%	110,719 / 3.12%	1,794,795 / 50.54%	96,808 / 2.73%	93.39
4619	2,088,183	2,035,545 / 97.48%	52,638 / 2.52%	940,692 / 45.05%	34,807 / 1.67%	67.75
5603	2,412,504	2,348,123 / 97.33%	64,381 / 2.67%	1,003,446 / 41.59%	74,977 / 3.11%	64.83
16106	2,266,447	2,228,251 / 98.31%	38,196 / 1.69%	896,069 / 39.54%	176,950 / 7.81%	64.26
11572	1,932,221	1,892,335 / 97.94%	39,886 / 2.06%	785,586 / 40.66%	55,284 / 2.86%	60.22
4082	1,982,434	1,945,379 / 98.13%	37,055 / 1.87%	808,656 / 40.79%	26,754 / 1.35%	59.49
1374	2,536,186	2,501,622 / 98.64%	34,564 / 1.36%	965,230 / 38.06%	30,128 / 1.19%	59.26
5450	2,043,969	1,964,495 / 96.11%	79,474 / 3.89%	786,322 / 38.47%	65,363 / 3.2%	50.53
6978	1,563,770	1,528,497 / 97.74%	35,273 / 2.26%	598,529 / 38.27%	34,089 / 2.18%	45.5
9435	1,655,811	1,598,387 / 96.53%	57,424 / 3.47%	611,571 / 36.93%	41,189 / 2.49%	43.64
4538	1,316,195	1,286,826 / 97.77%	29,369 / 2.23%	467,597 / 35.53%	44,247 / 3.36%	41.49
2196	1,575,721	1,512,670 / 96%	63,051 / 4%	537,351 / 34.1%	55,665 / 3.53%	38.28
ATCC	1,271,462	1,248,561 / 98.2%	22,901 / 1.8%	465,713 / 36.63%	14,371 / 1.13%	36.61
10400	1,124,389	1,100,716 / 97.89%	23,673 / 2.11%	343,213 / 30.52%	54,778 / 4.87%	35.97
7170	1,254,217	1,215,791 / 96.94%	38,426 / 3.06%	401,508 / 32.01%	30,961 / 2.47%	33.74
10019	1,193,273	1,149,678 / 96.35%	43,595 / 3.65%	413,742 / 34.67%	38,848 / 3.26%	33.7
6185	1,018,373	993,159 / 97.52%	25,214 / 2.48%	286,930 / 28.18%	18,671 / 1.83%	32.81
7193	1,064,402	1,035,494 / 97.28%	28,908 / 2.72%	305,124 / 28.67%	60,104 / 5.65%	31.65
12400	1,283,857	1,245,943 / 97.05%	37,914 / 2.95%	386,679 / 30.12%	106,163 / 8.27%	31.6
5303	1,030,804	1,003,081 / 97.31%	27,723 / 2.69%	291,179 / 28.25%	36,661 / 3.56%	31.46
4793	989,808	962,779 / 97.27%	27,029 / 2.73%	243,710 / 24.62%	32,703 / 3.3%	29.55
12650	1,044,545	1,014,040 / 97.08%	30,505 / 2.92%	339,801 / 32.53%	19,684 / 1.88%	29.28
14114	1,066,701	1,036,636 / 97.18%	30,065 / 2.82%	298,992 / 28.03%	46,176 / 4.33%	28.85
10862	1,050,562	1,031,219 / 98.16%	19,343 / 1.84%	260,867 / 24.83%	253,803 / 24.16%	27.49
9820	1,181,461	1,129,536 / 95.61%	51,925 / 4.39%	293,191 / 24.82%	31,412 / 2.66%	27.2
9784	866,641	845,446 / 97.55%	21,195 / 2.45%	214,568 / 24.76%	32,268 / 3.72%	26.32
8759	768,414	753,612 / 98.07%	14,802 / 1.93%	200,533 / 26.1%	54,084 / 7.04%	23.77
13689	751,799	735,014 / 97.77%	16,785 / 2.23%	183,076 / 24.35%	19,736 / 2.63%	23.75
9937	921,894	888,935 / 96.42%	32,959 / 3.58%	200,082 / 21.7%	17,188 / 1.86%	23.46
10755	736,105	719,487 / 97.74%	16,618 / 2.26%	178,622 / 24.27%	36,949 / 5.02%	23.32
259	1,006,781	964,820 / 95.83%	41,961 / 4.17%	227,570 / 22.6%	51,969 / 5.16%	21.69
8208	680,631	662,325 / 97.31%	18,306 / 2.69%	136,790 / 20.1%	10,945 / 1.61%	21.56
2435	728,784	711,737 / 97.66%	17,047 / 2.34%	178,994 / 24.56%	10,929 / 1.5%	21.53
10237	703,127	680,263 / 96.75%	22,864 / 3.25%	158,950 / 22.61%	10,106 / 1.44%	21.06
7649	630,121	619,601 / 98.33%	10,520 / 1.67%	137,438 / 21.81%	13,716 / 2.18%	21.03
12488	727,816	705,857 / 96.98%	21,959 / 3.02%	173,722 / 23.87%	60,541 / 8.32%	20.64
2599	736,315	720,558 / 97.86%	15,757 / 2.14%	154,178 / 20.94%	66,810 / 9.07%	20.58
6993	637,351	621,855 / 97.57%	15,496 / 2.43%	132,421 / 20.78%	13,709 / 2.15%	20.25
2417	820,205	790,610 / 96.39%	29,595 / 3.61%	186,666 / 22.76%	13,071 / 1.59%	19.84
2162	802,903	774,761 / 96.49%	28,142 / 3.51%	203,824 / 25.39%	11,441 / 1.42%	19.79
13449	726,091	698,251 / 96.17%	27,840 / 3.83%	146,906 / 20.23%	17,857 / 2.46%	19.78
16694	601,157	585,088 / 97.33%	16,069 / 2.67%	114,256 / 19.01%	14,879 / 2.48%	19.73
3702	779,938	750,668 / 96.25%	29,270 / 3.75%	153,924 / 19.74%	11,805 / 1.51%	19.62
3537	663,675	649,586 / 97.88%	14,089 / 2.12%	139,314 / 20.99%	10,035 / 1.51%	19.33
13050	589,309	574,198 / 97.44%	15,111 / 2.56%	107,001 / 18.16%	17,085 / 2.9%	19
4737	629,405	610,071 / 96.93%	19,334 / 3.07%	113,703 / 18.07%	15,376 / 2.44%	18.67
9480	590,981	576,215 / 97.5%	14,766 / 2.5%	116,085 / 19.64%	33,260 / 5.63%	18.55
15696	680,991	652,882 / 95.87%	28,109 / 4.13%	123,707 / 18.17%	22,414 / 3.29%	18.39
7924	948,608	884,679 / 93.26%	63,929 / 6.74%	239,705 / 25.27%	21,213 / 2.24%	18.38
12442	589,251	572,824 / 97.21%	16,427 / 2.79%	108,575 / 18.43%	14,036 / 2.38%	17.95

Table 3.3.: *bwa-mem* mapping metrics (continued).

Strain	Total reads	Mapped reads	Unmapped reads	Duplicated reads	Clipped reads	Coverage
2263	725,244	693,719 / 95.65%	31,525 / 4.35%	156,929 / 21.64%	14,769 / 2.04%	17.75
7630	577,279	559,041 / 96.84%	18,238 / 3.16%	109,290 / 18.93%	11,830 / 2.05%	17.04
3212	575,127	561,947 / 97.71%	13,180 / 2.29%	120,493 / 20.95%	12,736 / 2.21%	16.85
2380	672,960	650,282 / 96.63%	22,678 / 3.37%	136,725 / 20.32%	11,047 / 1.64%	16.73
7101	530,852	514,690 / 96.96%	16,162 / 3.04%	93,636 / 17.64%	18,358 / 3.46%	16.69
15097	524,115	510,974 / 97.49%	13,141 / 2.51%	84,907 / 16.2%	13,392 / 2.56%	16.16
7656	807,356	765,040 / 94.76%	42,316 / 5.24%	152,451 / 18.88%	8,806 / 1.09%	16.08
1117	508,830	497,061 / 97.69%	11,769 / 2.31%	81,052 / 15.93%	9,103 / 1.79%	15.93
260	651,552	626,838 / 96.21%	24,714 / 3.79%	119,500 / 18.34%	11,339 / 1.74%	15.8
10485	465,325	452,290 / 97.2%	13,035 / 2.8%	72,247 / 15.53%	8,066 / 1.73%	15.3
10851	504,860	490,049 / 97.07%	14,811 / 2.93%	76,881 / 15.23%	14,448 / 2.86%	15.15
1770	645,280	612,530 / 94.92%	32,750 / 5.08%	103,578 / 16.05%	12,151 / 1.88%	14.17
2025	595,369	567,379 / 95.3%	27,990 / 4.7%	100,247 / 16.84%	41,447 / 6.96%	14
8139	479,116	458,976 / 95.8%	20,140 / 4.2%	75,666 / 15.79%	11,077 / 2.31%	12.6
13452	296,394	260,164 / 87.78%	36,230 / 12.22%	17,659 / 5.96%	18,664 / 6.3%	4.45
8817	122,772	113,825 / 92.71%	8,947 / 7.29%	7,308 / 5.95%	3,318 / 2.7%	2.36
597	18,483	17,269 / 93.43%	1,214 / 6.57%	315 / 1.7%	283 / 1.53%	0.36
5361	12,358	11,511 / 93.15%	847 / 6.85%	185 / 1.5%	219 / 1.77%	0.24

3.4. Variant detection and filtering results

Variants were called using *nucmer* [40] and the *GATK* [20] HaplotypeCaller , in accordance to the process described in the Methods section. Found variants from both processes were filtered for SNPs and combined for each strain. SNPs on which both variant caller agreed were saved and another filter step was executed to filter SNPs found in 273 problematic genes.

On average 583 filtered high quality SNPs were found for each strain. The minimum amount of SNPs (26) was found in the ATCC strain and the maximum of 1102 SNPs was found in strain 11572. Over all strains between 1,5% and 9% of the refernce sequence was not covered (Table 3.4).

A total of 4345 high quality SNPs in 2173 genes were detected in this analysis. Out of those 4345 SNPs 1599 were synonymous and 2746 were nonsynonymus variants. SNPs were found on 4333 unique positions of the reference sequence. Those 4333 position times the 65 examined strain is also the size of the SNP multialignment used for the phylogenetic analysis.

Table 3.4.: Variant detection and SNP filtering results, sorted by strain number. Number of not covered bases of the reference sequence is included as an indicator for library preparation problems. Total count of SNPs found by *newbler* and the *GATK* HaplotypeCaller, the agreed upon SNPs calculated by an intersection filter and the resulting high quality filtered SNPs are shown.

			SNPs			
Strain	not covered bases		GATK	nucmer	intersec.	filt.
10019	85799	1,94%	454	658	432	417
10237	249229	5,65%	707	1029	668	646
10400	129802	2,94%	795	1074	761	739
10485	258324	5,86%	756	1089	715	696
10755	200803	4,55%	745	1064	710	691
10851	296784	6,73%	713	1043	675	655
10862	134495	3,05%	447	661	411	401
1117	270872	6,14%	412	612	396	384
11572	135986	3,08%	1203	1516	1144	1102
12400	98520	2,23%	452	563	422	404
12442	300280	6,81%	409	633	397	388
12488	91438	2,07%	509	684	497	482
12650	114254	2,59%	457	628	430	413
13050	222236	5,04%	431	600	413	394
13449	235046	5,33%	418	629	404	385
13452	394473	8,94%	352	1230	337	325
13689	185914	4,21%	430	598	419	408
1374	66900	1,52%	817	1047	779	725
14114	209218	4,74%	750	1023	712	688
14163	87385	1,98%	794	983	752	726
15097	244750	5,55%	701	1003	659	641
15696	200413	4,54%	731	988	695	675
16106	90678	2,06%	465	631	438	412
16694	247915	5,62%	741	1092	719	702
1770	248121	5,62%	416	654	394	380
2025	108754	2,47%	436	743	425	408
2162	245349	5,56%	719	1069	693	672
2196	98957	2,24%	458	581	430	407
2263	228046	5,17%	715	1017	685	667
2380	275236	6,24%	693	1045	664	645
2417	228987	5,19%	695	996	674	657
2435	320217	7,26%	700	1040	668	648
259	177136	4,02%	457	1062	427	415
2599	242167	5,49%	706	1020	666	647
260	274541	6,22%	731	987	694	672
3212	330135	7,48%	663	1033	631	614
3537	276200	6,26%	725	1069	695	673
3702	200894	4,55%	736	1006	709	684
4082	145168	3,29%	769	960	733	707
4538	134656	3,05%	775	1006	734	706
4619	112289	2,55%	780	1005	742	718
4737	242589	5,50%	739	1097	705	684
4793	158741	3,60%	786	1080	765	743
5303	159350	3,61%	742	1002	714	691
5450	119323	2,70%	788	1031	746	717
5603	93595	2,12%	673	811	641	616
6185	180754	4,10%	757	981	711	689
6978	122753	2,78%	445	629	428	407
6993	232312	5,27%	731	1066	686	665
7101	262777	5,96%	718	1050	678	658
7170	151961	3,44%	766	1001	731	709
7193	151311	3,43%	752	961	712	685
7630	304729	6,91%	689	1083	656	637
7649	230407	5,22%	407	684	392	382
7656	224400	5,09%	617	869	601	581
7924	239519	5,43%	679	1065	649	624
8139	400090	9,07%	656	1064	625	606
8208	204367	4,63%	730	957	694	671
8759	181031	4,10%	761	997	709	686
9435	112700	2,55%	775	964	746	715
9480	261394	5,93%	392	655	381	372
9784	169400	3,84%	764	987	730	705
9820	158338	3,59%	440	635	427	406
9937	154898	3,51%	459	1051	432	419
ATCC	132377	3,00%	28	117	27	26

3.5. Phylogenetic analysis and drug resistance analysis

To analyse drug resistance, the full TBDReamDB [45] online database was downloaded which consists of around 1200 mutations associated with resistances against 9 drug classes.

The most resistances (86) were detected for Ethambutol. For the other drugs, mutations which are associated with known resistances were found between 63 and 7 times (Table 3.5). A detailed drug resistance report for each strain was generated which can be found in Appendix A.

Table 3.5.: Drugs, for which associated mutations were found.

Abbreviation	Name	Found mutations
EMB	Ethambutol	86
ETH	Ethionamide	7
FLQ	Fluoroquinolones	51
INH	Isoniazid	63
PZA	Pyrazinamide	20
RIF	Rifampicin	63
SM	Streptomycin	25

Phylogentic analysis of the SNPs multialignment grouped the strains into four main clusters (Figure 3.3). The lineage analysis of those clusters identified four reference strains, belonging to two different MTB lineages.

The largest cluster of strains belongs to the virulent lineage 4 Haarlem strain. The documented high virulency and drug resistance of this strain is reflected in the identified mutations. The second largest cluster was associated in the lineage analysis with the H37Rv strain.

There are two significant deviations. For a cluster of strains (shown in blue) the nearest reference strain was identified to be a uncommon strain (UT205) found first in Columbia. This strain also belongs to lineage 4 but shows amino acid changes in one of 430 coding sequences when compared to the H37Rv reference [51].

Strain 11572 is the most significant discrepancy in this analysis. It belongs to a completely different lineage and was identified to be related to the highly virulent HN878 strain which belongs to Beijing lineage 2.

Except for the ATCC reference strain only strain 9480 shows no drug resistance related SNPs. One strain (9937) with a single detected resistance to Rifampicin was found and seven strains are showing a resistance for two different drugs. For 54 strains resistances for 3 to 5 drugs were detected. Haarlem strain 4619 shows a resistance to all but Streptomycin and Beijing strain 11572 a resistance to all but Ethionamide.

Ethambutol resistance was found in almost all strains belonging to the Haarlem lineage, except for a sub-cluster of three strains. These three Haarlem strains are also placed next to the ATCC reference strain in the Minimum Spanning Tree. Isoniazid resistance was detected in all drug resistant European (H37RV) strains.
All but two strains (7193 & 13689) of those with INH resistance were also resistant to Rifampicin.

The MST is a useful tool for investigating the transmission of infection and to analyze relationships among strains only based on their similarities (Figure 3.4). Two of the UT205 strains are placed isolated next to the H37Rv strain 2025. Strain 2196 is identified as the root node for seven other strains. This can be interpreted as a source of transmission but can only be verified by combining the sequencing results to clinical data. Another root node is strain 5450 inside the Haarlem cluster.

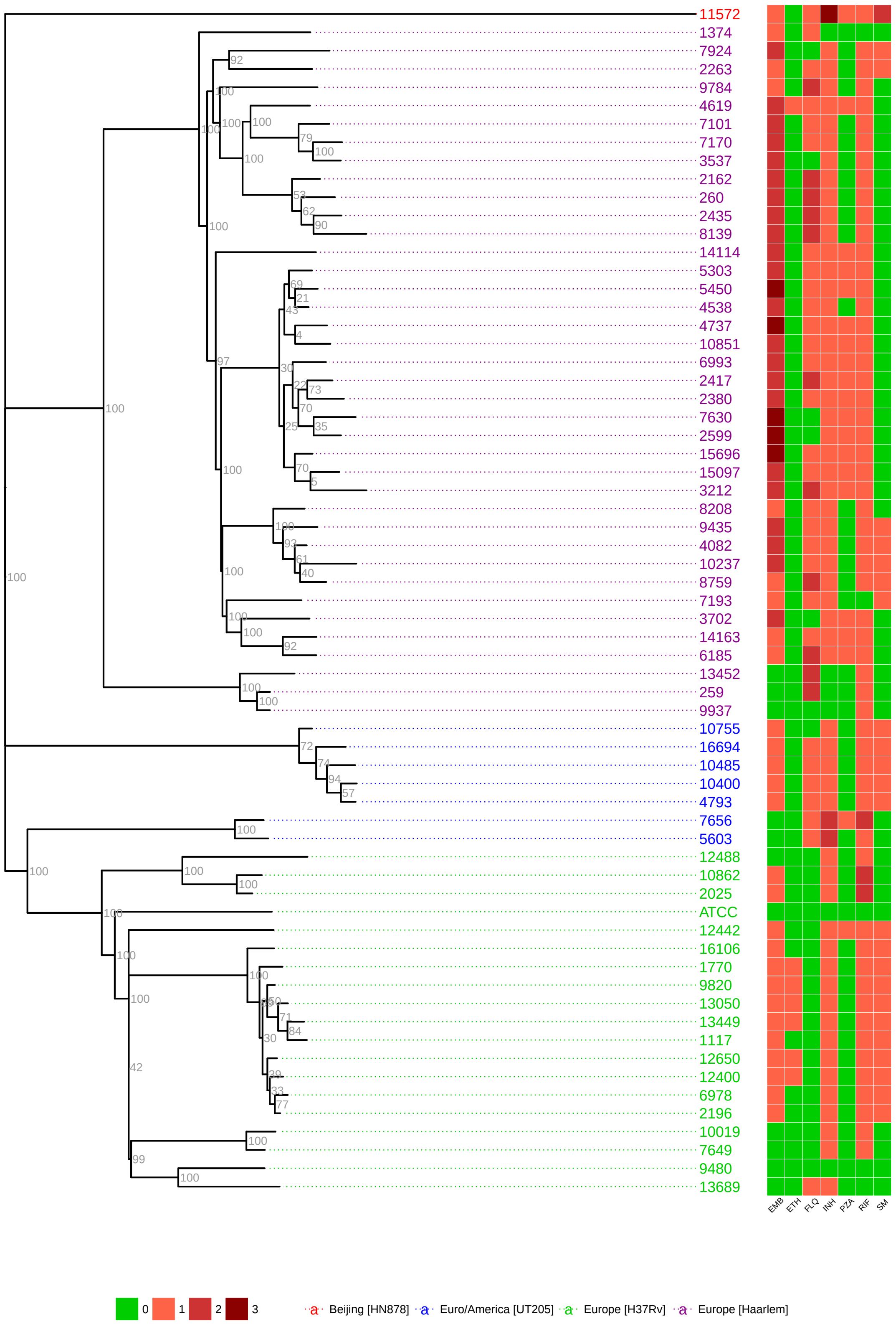

Figure 3.3.: Phylogenetic tree, lineage analysis and found drug resistances. Phylogenetic tree generated by *RAxML*. Bootstrap support values are shown in grey. Strain color indicates the region of origin. The heatmap includes a column for each included drug class. The color indicates the amount of found mutations associated with a resistance for that drug.

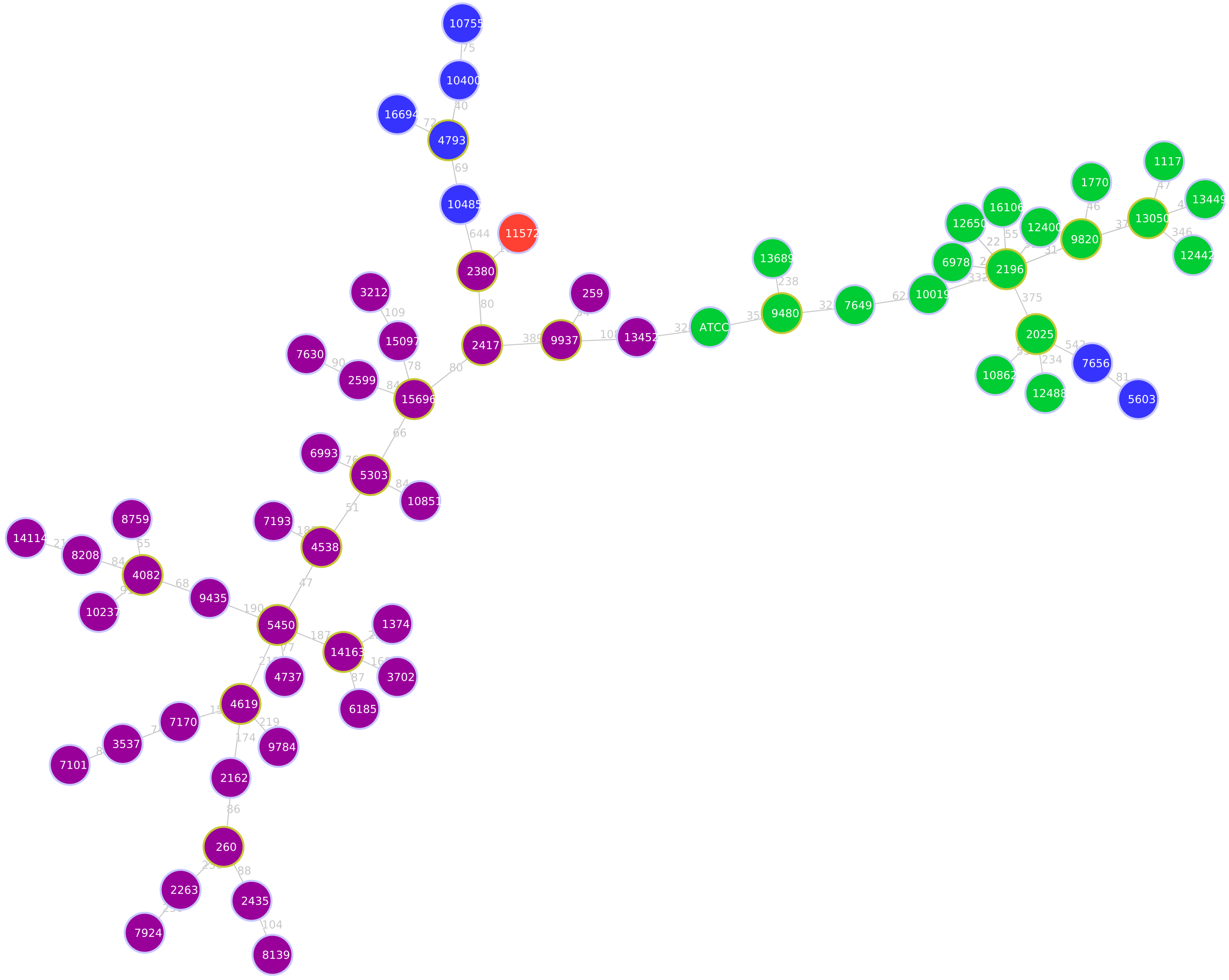

Figure 3.4.: Minimum Spanning Tree generated with Phyloviz. The corresponding main lineage for each strain is color coded in the nodes. Numbers between the nodes denote the calculated goeBURST distance. All distances are scaled logarithmically, to fit the picture on the page. Blue: Euro/America (UT205), Green: Europe (H37Rv), Violet: Europe (Haarlem), Red: Beijing (HN878)

4. Discussion

Mycobacterium tuberculosis strains collected from 64 patients were sequenced by the University of Medicine and Pharmacy in Iasi, Romania. Sequenced data was analyzed within this thesis and a reusable workflow for the computational analysis of whole genome sequencing was implemented. Raw sequencing data is acquired, filtered and preprocessed for genome reconstruction. *De-novo* assembly and reference based mapping is carried out for each strain. Variants are called, combined, filtered and used for phylogenetic, as well as drug resistance analysis.

Data from 68 strains was processed in the pipeline. Three strains were discarded due to insufficient coverage. For workflow evaluation a known reference strain (ATCC) was also sequenced. This leaves a total of 65 strains for which results are presented. The ATCC strain data was used to identify problems with the library preparation regarding the high GC-content of the bacterium. A bug in the IonTorrent software was found, which could be fixed as the software suite is available as open source.

The pipeline identified a total of 4345 SNPs of which 315 were associated with drug resistances in the database. Phylogenetic and lineage analysis yielded four main groups of strains, with all but one strain belonging to the European lineage 4.

4.1. Pipeline implementation

Pior to the bioinformatic analysis some general problems with data handling and computational performance had to be solved. The sequencing data used in this thesis consists of a raw-read data file for each sequenced strain. All of those data files should be processed exactly the same during the analysis stage. If a program needs lots of computing resources, only one data file may be processed at the same time. On the other hand, if an analysis needs little resources but takes a long time, it may be wise to start this program for a number of input data in parallel.

Some programs need a binary version of the data, other programs are not able to work with compressed data. So in the end of an analysis there is not only a single file per strain, but many of them. Keeping an overview over all the data, as well as processing all the files in a sequential manner is a big challenge. Big data analysis was carried out on the Institute's computing cluster. This includes file format conversion as well as mapping and assembly of the genomes. High level analysis with less performance requirements was done on a modern desktop PC.

4.2. Technical limitations and challenges

The IonTorrent sequencing platform has a well documented problem with correctly detecting long homopolymers [11]. To get a better understanding of how this problem affected the SNP detection, the number of indels that are within a homopolymer was collected with *qualimap*. In this analysis an indel is considered homopolymeric when it is found within at least a five base stretch of the same base. The average percentage of indels that are in homoplymer areas is 50,7%. This high number indicates a problem in the sequencing process which could only be solved by resequening the samples on another sequencing platform.

A surprisingly large proportion (1,5 to 9%) of uncovered bases was observed in the generated mappings (Table 3.4). The first attempt to explain the high rate of uncovered bases was a supposed problem in the mapping pipeline. For further investigations on this problem a known reference strain (ATCC) was acquired, sequenced and analyzed. Even with a high coverage of over 68x over 40,000 loci were not covered in the sequenced ATCC reference when compared to the NCBI entry (Figure 3.1). To exclude mapping artifacts different mappers were used in the analysis of the ATCC reference strain. The *bwa-mem* showed an even higher amount of uncovered bases and therefore the *CLC* mapper was used in this analsyis.

Further investigation revealed that the position of uncovered bases correlated with regions of high GC content in the genome (Figure 3.2). The ten widest regions with zero coverage in the ATCC mapping all have a GC content between 76% and 83% (Table 3.2).

This correlation between regions with zero coverage and high GC-content may indicate a problem with the preparation of the sequencing libraries. This has been also repeatedly reported in recent publications [67,68].

The impact of these problems on the conducted research is assumed to be minimal, as the focus lies on detecting single nucleotide polymorphisms in non repetitive elements of the genome. This was also taken into account by implementing a filter for known repetitive elements, including PGRS and PE/PPE genes, transposases, and prophage-like elements. Most of the missing regions in the sample strain ATCC belong to one of those repetitive elements categories. This allows the conclusion, that library problems with high GC-content regions do not affect the presented results of this thesis more than in other studies. Some SNPs may have been missed and some of the excluded may even contribute to a drug resistance, but this is beyond the scope of this thesis and should be investigated in further studies.

A computational solution for this problem does not exist and if variants in regions with high GC content are of interest, the library preparation step before sequencing needs adaption.

Large structural changes in the genome could not be analyzed because of the used sequencing technology and library preparation. Using the short single end reads from the Ion Torrent sequencer as base material for the *de-novo* assembly produced short contigs that could not be assembled into longer scaffolds. This is the main reason why all subsequent analysis was concentrating on single nucleotide variations.

During the initial assembly of the strains, a problem was identified in the file conversion tool (*bam2sff*) of the Torrent Suite Software distributed by Thermo Fisher Scientific. It seems, that this tools is rarely used, because this error was not reported before and was not fixed one year after I reported this bug to the IonTorrent team. This bug only affects people who want to use the IonTorrent server to convert the raw data into SFF file format. Luckily, the

IonTorrent software is open source and available for download from GitHub [64]. One big advantage of open source software is that you can just fix a bug yourself, if you can find it. In this case the main problem was not to find or fix the bug, but to get the tool to recompile, as this file conversion script is only a tiny part of the whole Torrent Suite Software.

4.3. SNP analysis

SNP identification in this thesis is based on the idea that every bit of information counts. This is the reason why both the *de-novo* and the reference based mapping in a combined workflow were used. By executing the variant calling on both mappings independently and calculating the intersection of the results only reliable, high-quality SNPs were used as input for the subsequent analysis. SNPs found in problematic regions like transposases were filtered subsequently to get even higher quality data for the phylogenetic and drug resistance analysis. A total of 4345 high quality SNPs were found on 4333 unique positions in the reference genome.

An average of 590 SNPs were identified for each strain with one major outlier (strain 11572) with 1102 SNPs. No direct correlation between the number of SNPs and the mapping coverage was found. This was suspected to be an issue, as some of the strains were sequenced multiple times due to chip loading problems. Two clusters of strains can be observed when considering the number of high quality SNPs for each strain. The group with 320 to 420 SNPs was later placed next to the H37Rv lineage. For the second cluster between 600 and 750 SNPs were found. Strains belonging to this cluster were assigned to the Haarlem and UT205 linages.

4.4. Phylogenetic and lineage analysis

Results of the phylogenetic analysis show that all but one clinical isolate (11572) belong to the Euro-American lineage (Lineage 4), which was to be expected, as all samples were collected in Romania. The majority of strains (38) are related to the highly virulent and drug resistant Haarlem strain. One isolate of the Haarlem group (4619) even shows resistance to all available drug classes, except Streptomycin.

The second largest group (19 strains) of strains are related to the H37Rv strain, which is the most studied strain of *M. tuberculosis* in the Euro-American lineage. It is also the reference strain for all analysis in this thesis. The sequenced H37Ra reference strain ATCC is closely related to the H37Rv reference strain and shows no drug resistances, as expected. 26 SNPs were detected in the ATCC strain when compared to the H37Rv reference, including previously documented high impact SNPs in the pstA1 (stop gained) and pks3 (stop lost) genes [69].

Seven strains are related to the UT205 strain of the Euro-American lineage [70]. UT205 was first found in a patient in Medellin, Colombia. Four of those seven strains are extensively drug-resistant in terms of the WHO classification [1].

The Euro-American strains are grouped in two sub-trees with very different drug resistance patterns. In the phylogenetic tree as well as in the MST one of those sub-trees (strains 7656 & 5603) is placed closely to the European HR37Rv strains. This could be an indication for a incorrect result in the lineage distance analysis, because of an missing reference strain in the reference set.

The outlier strain 11572 is related to the hypervirulent HN878 [71] strain from the Beijing lineage and shows an extensive drug resistant pattern with a total of 9 SNPs that correspond to a drug resistance in the mutation database.

When comparing the results of the phylogenetic tree generated with *RAxML* with the Minimum Spanning Tree, a good correlation can be observed. Both the MST and the maximum likelihood phylogenetic tree show significant similarities in both the placement of individual strains, as well as the lineage association. As the MST is only based on the distance between the SNP based multialignment, some differences can be observed. The Euro/American group of strains is not present as one whole group but rather split in two and attached to both European groups. The ATCC reference strain is shown as the link between the H37Rv and the Haarlem group.

There are some strains that are directly linked to four or more closely related strains (e.g. strains 2196 or 5450). Those strains can be the source of infection and should be further investigated in combination with clinical data of the patients, to uncover and validate transmission pathways.

4.5. Drug resistance analysis

A total of 86 matches in 54 strains for mutations associated with drug resistance for Ethambutol were observed. This first-line drug was discovered 1961 and is listed on the WHO Model List of Essential Medicines (EML) [72].

For Isoniazid and Rifampicin 63 mutations were detected, affecting 59 strains for Isoniazid resistance and 60 strains for Rifampicin resistance. 57 strains are resistant to both first-line drugs and thereby classified by the WHO as at least Multidrug-resistant [1].

Mutations associated to resistance against Fluoroquinolones were found 51 times, affecting 40 strains. Only one strain from the H37Rv related European group is showing a Fluoroquinolones resistance mutation. Strains from this group are showing the highest percentage of Ethionamide and Streptomycin resistances but are susceptible for Pyrazinamide. Pyrazinamide is not used on its own and is not recommended for the treatment of latent tuberculosis. The main purpose of Pyrazinamide is to shorten the duration of treatment.

Ethionamide (7) and Streptomycin (25) are confronted with the least amount of resistance mutations. Streptomycin was the first drug discovered to be an effective TB antibiotic [73]. Today it is mainly used for second line treatment of patients that are treated for TB repeatedly. Ethionamide is mainly used to treat active multidrug-resistant tuberculosis. Both drugs can have severe side-effects. Ethionamide can cause persistent nausea and vomiting [74] and the main concern with Streptomycin is its ototoxicity, which can lead to permanent deafness [75].

Strain 13689 has a detected INH resistance but no known RIF resistance mutation in the rpoB gene. Instead, another mutation (V251F) was found in rpoB which was also documented in a recent publication [76]. All other drug resistant strains belonging to the European (H37Rv) lineage, except for the non drug resistant strain (9480 & ATCC), are are showing the typical MDR pattern with INH and RIF resistance.

The durg resistance pattern of the main cluster of Haarlem strains with resistances for at least Rifampicin, Isoniazid and Fluoroquinolones matches the observations of Alvarez *et al.* [77] for their published high-quality draft genome of a Colombian Haarlem XDR-TB clinical isolate.

4.6. Conclusion

Drug resistance testing is a very important examination when treating TB. The therapy duration and the hospitalisation and drug costs can significantly change when a durg resistance is detected early [78].

There are two types of drug susceptibility tests for TB. Growing MTB culture *in vitro* in the presence of a certain drug is the most commonly used method for testing. As MTB grows very slowly, testing for drug susceptibility in culture requires incubation times of eight to 12 weeks [79]. Automated liquid culture systems can speed the process but still need two to four weeks for results [80]. The second type of test are molecular test like the GeneXpert system from Cepheid [81]. It uses a sputum sample and yields results in around 2 hours. But it only tests for the presence of a TB infection and resistance to the drug Rifampicin.

Because of the price drop in per base sequencing costs, sequencing the whole genome of bacterial pathogens gets more accessible. The possibility to detect a drug resistance and therefore select the right treatment for each individual patient will change clinical routine. Instead of choosing an antibiotic on empirical knowledge, an individual therapy can be used for each patient. This will also save costs in terms of efficiently using the available drugs.

This thesis shows a possible way of processing raw sequencing data from a bacterial whole genome sequencing experiment.

It is favorable to cost efficiently sequence a large number of strains in one run. Theses sets can then be analyzed using automated scripts that execute all necessary steps to arrive at the desired results. These results include a phylogenetic tree of all strains that are included in the analysis, as well a heritage lineage analysis and the drug resistance summary.

4.7. Outlook

Combining all described methods of this thesis in an automated workflow would enable a researchers to get detailed ancestry and drug resistance data from raw sequencing data without needing to know the bioinformatical details of the tools that need to be used. Implementing this automated program on a computational cluster can reduce the calculation time. Optimizing the variant detection and filter process would also be of interest.

Acknowledgement

I would first like to thank my thesis advisor Dr. Gerhard Thallinger for his guidance and patience during my studies in Bioinformatics. The door to his office was always open and he steered me in the right direction whenever I ran into trouble.

I would also like to thank the experts of the University of Medicine and Pharmacy in Iasi, Romania, who provided the data for this research project:
Prof. Dr. Grigoriu Bogdan-Dragos from the Regional Institute of Oncology
Drd. Radu Zonda from the Petru Poni Institute of Macromolecular Chemistry

Their hard work with all the sample collection, processing and sequencing made this thesis possible.

Finally, I would like to thank my wife Sabrina, my parents and the rest of my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Bibliography

[1] World Health Organization: **Global tuberculosis report** 2016. URL `http://www.who.int/tb/publications/global_report/en/`. Accessed November, 2016.

[2] Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S, Eiglmeier K, Gas S, Barry C *et al.*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence**. *Nature* 1998. 393(6685):537–544.

[3] Cole S, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S, Eiglmeier K, Gas S, Barry C *et al.*: **Erratum: Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence**. *Nature* 1998. 396(6707):190.

[4] Brites D and Gagneux S: **Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens***. *Immunological Reviews* 2015. 264(1):6–24.

[5] Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T *et al.*: ***Mycobacterium tuberculosis* lineage 4 comprises globally distributed and geographically restricted sublineages**. *Nature Genetics* 2016. 48(12):1535–1543.

[6] Beaubien J: **Moldova grapples with whether to isolate TB patients**. *National Public Radio* 2013.

[7] Günther G, Lange C, Alexandru S, Altet N, Avsar K, Bang D, Barbuta R, Bothamley G, Ciobanu A, Crudu V *et al.*: **Treatment outcomes in Multidrug-Resistant Tuberculosis**. *New England Journal of Medicine* 2016. 375(11):1103–1105.

[8] Rusk N: **Torrents of sequence**. *Nature Methods* 2011. 8(1):44–45.

[9] Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X *et al.*: **The potential and challenges of nanopore sequencing**. *Nature Biotechnology* 2008. 26(10):1146–1153.

[10] Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG and Webb WW: **Zero-mode waveguides for single-molecule analysis at high concentrations**. *Science* 2003. 299(5607):682–686.

[11] Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J and Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms**. *Nature Biotechnology* 2012. 30(5):434–439.

[12] Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER and Pallen MJ: **High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity**. *Nature Reviews Microbiology* 2012. 10(9):599–606.

[13] Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, Keim P, Morrow JB, Salit ML and Zook JM: **Best practices for evaluating single nucleotide variant calling methods for microbial genomics**. *Frontiers in Genetics* 2015. 6.

[14] Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M *et al.*: **Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms**. *The ISME Journal* 2012. 6(8):1621–1624.

[15] Gardner SN and Hall BG: **When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes**. *PLoS ONE* 2013. 8(12):e81760.

[16] Treangen TJ, Ondov BD, Koren S and Phillippy AM: **The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes**. *Genome Biology* 2014. 15(11):524.

[17] Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, Wirth T, Jaenicke S, Schuback S, Rüsch-Gerdes S *et al.*: **Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study**. *PLoS Medicine* 2013. 10(2):e1001387.

[18] Ilina EN, Shitikov EA, Ikryannikova LN, Alekseev DG, Kamashev DE, Malakhova MV, Parfenova TV, Afanas'ev MV, Ischenko DS, Bazaleev NA *et al.*: **Comparative genomic analysis of *Mycobacterium tuberculosis* drug resistant strains from Russia**. *PLoS ONE* 2013. 8(2):e56577.

[19] Pilato CM, Collins-Sussman B and Fitzpatrick B: **Version control with subversion**. O'Reilly Media, Inc., 2008.

[20] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Research* 2010. 20(9):1297–1303.

[21] Free Software Foundation: **find** 2016. URL `http://linuxcommand.org/man_pages/find1.html`. Accessed February, 2016.

[22] R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[23] National Center for Biotechnology Information: **Standard flowgram format (sff)** 2017. URL `https://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=formats&m=doc&s=format#sff`. Accessed April, 2017.

[24] Broad Institute: **Picard Tools** 2016. URL `http://broadinstitute.github.io/picard/`. Accessed February, 2016.

[25] Zheng H, Lu L, Wang B, Pu S, Zhang X, Zhu G, Shi W, Zhang L, Wang H, Wang S *et al.*: **Genetic basis of virulence attenuation revealed by comparative genomic analysis of *Mycobacterium tuberculosis* strain H37Ra versus H37Rv**. *PLoS ONE* 2008. 3(6):e2375.

[26] Li H: **Aligning sequence reads, clone sequences and assembly contigs with bwa-mem**. *arXiv preprint arXiv:1303.3997* 2013.

[27] QIAGEN Bioinformatics: **CLC Genomics Workbench version: 6.5.2** 2016. URL `https://www.qiagenbioinformatics.com`. Accessed February, 2016.

[28] Quinlan AR and Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010. 26(6):841–842.

[29] Okonechnikov K, Conesa A and García-Alcalde F: **Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data**. *Bioinformatics* 2016. 32(2):292–294.

[30] Ewels P: **NGI Visualizations**. *GitHub* 2015. URL `https://github.com/ewels/ngi_visualizations`. Accessed February, 2015.

[31] Overmars L, van Hijum SA, Siezen RJ and Francke C: **CiVi: circular genome visualization with unique features to analyze sequence elements**. *Bioinformatics* 2015. 31(17):2867–2869.

[32] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nature Genetics* 2011. 43(5):491–498.

[33] Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J *et al.*: **From FastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline**. *Current Protocols in Bioinformatics* 2013. pages 11–10.

[34] Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet.journal* 2011. 17(1):pp–10.

[35] Andrews S *et al.*: **FastQC: A quality control tool for high throughput sequence data**. *Reference Source* 2010.

[36] Lew JM, Kapopoulou A, Jones LM and Cole ST: **Tuberculist–10 years after**. *Tuberculosis* 2011. 91(1):1–7.

[37] Camus JC, Pryor MJ, Médigue C and Cole ST: **Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv**. *Microbiology* 2002. 148(10):2967–2973.

[38] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al.*: **The variant call format and VCFtools**. *Bioinformatics* 2011. 27(15):2156–2158.

[39] 454 Life Sciences: **Gs de novo assembler** 2016. URL `http://www.454.com/products/analysis-software/`. Accessed February, 2016.

[40] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C and Salzberg SL: **Versatile and open software for comparing large genomes**. *Genome Biology* 2004. 5(2):R12.

[41] Cingolani P, Platts A, Coon M, Nguyen T, Wang L, Land S, Lu X and Ruden D: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3**. *Fly* 2012. 6(2):80–92.

[42] Cook D: **vcf-toolbox**. *GitHub* 2016. URL `https://github.com/AndersenLab/vcf-toolbox`. Accessed May, 2016.

[43] Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD and Gagneux S: **Human t cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved**. *Nature Genetics* 2010. 42(6):498–503.

[44] Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM, Calver A, Sloutsky A *et al.*: **Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis***. *Nature Genetics* 2013. 45(10):1183–1189.

[45] Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM and Murray MB: **Tuberculosis drug resistance mutation database**. *PLoS Medicine* 2009. 6(2):e1000002.

[46] Telenti A, Imboden P, Marchesi F, Matter L, Schopfer K, Bodmer T, Lowrie D, Colston M and Cole S: **Detection of rifampicin-resistance mutations in *Mycobacterium tuberculosis***. *The Lancet* 1993. 341(8846):647–651.

[47] Stamatakis A: **Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006. 22(21):2688–2690.

[48] Plotree D and Plotgram D: **Phylip-phylogeny inference package (version 3.2)**. *Cladistics* 1989. 5(163):6.

[49] Kremer K, Van Soolingen D, Frothingham R, Haas W, Hermans P, Martin C, Palittapongarnpim P, Plikaytis B, Riley L, Yakrus M *et al.*: **Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility**. *Journal of Clinical Microbiology* 1999. 37(8):2607–2618.

[50] Sandegren L, Groenheit R, Koivula T, Ghebremichael S, Advani A, Castro E, Pennhag A, Hoffner S, Mazurek J, Pawlowski A *et al.*: **Genomic stability over 9 years of an isoniazid resistant *Mycobacterium tuberculosis* outbreak strain in sweden**. *PLoS ONE* 2011. 6(1):e16647.

[51] Isaza JP, Duque C, Gomez V, Robledo J, Barrera LF and Alzate JF: **Whole genome shotgun sequencing of one colombian clinical isolate of *Mycobacterium tuberculosis* reveals dosr regulon gene deletions**. *FEMS Microbiology Letters* 2012. 330(2):113–120.

[52] Ioerger TR, Feng Y, Chen X, Dobos KM, Victor TC, Streicher EM, Warren RM, van Pittius NCG, Van Helden PD and Sacchettini JC: **The non-clonality of drug resistance in Beijing-genotype isolates of *Mycobacterium tuberculosis* from the Western Cape of South Africa**. *BMC Genomics* 2010. 11(1):670.

[53] O'Toole RF, Johari BM, Mac Aogáin M, Rogers TR, Bower JE, Basu I and Freeman JT: **Draft genome sequence of the first isolate of extensively drug-resistant *Mycobacterium tuberculosis* in new zealand**. *Genome Announcements* 2014. 2(3):e00319–14.

[54] Stamatakis A: **Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies**. *Bioinformatics* 2014. 30(9):1312–1313.

[55] Stamatakis A: **Phylogenetic models of rate heterogeneity: A high performance computing perspective**. In **Proceedings of the 20th IEEE International Parallel & Distributed Processing Symposium (IPDPS2006)**, pages 278–286. IEEE Computer Society Press. ISBN 1-4244-0054-6, 2006.

[56] Olsen G: **Gary Olsen's interpretation of the "Newick's 8: 45" tree format standard** 1990. URL `http://evolution.genetics.washington.edu/phylip/newick_doc.html`. Accessed February, 2016.

[57] Kruskal JB: **On the shortest spanning subtree of a graph and the traveling salesman problem**. *Proceedings of the American Mathematical society* 1956. 7(1):48–50.

[58] Salipante SJ and Hall BG: **Inadequacies of minimum spanning trees in molecular epidemiology**. *Journal of Clinical Microbiology* 2011. 49(10):3568–3575.

[59] Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M and Carriço JA: **Phyloviz: phylogenetic inference and data visualization for sequence based typing methods**. *BMC bioinformatics* 2012. 13(1):87.

[60] Francisco AP, Bugalho M, Ramirez M and Carriço JA: **Global optimal eburst analysis of multilocus typing data using a graphic matroid approach**. *BMC bioinformatics* 2009. 10(1):152.

[61] Ewels P, Magnusson M, Lundin S and Käller M: **MultiQC: summarize analysis results for multiple tools and samples in a single report**. *Bioinformatics* 2016. 32(19):3047–3048.

[62] Rambaut A: **Figtree: A graphical viewer of phylogenetic trees, version 1.4** 2014. URL `http://tree.bio.ed.ac.uk/software/figtree/`. Accessed February, 2016.

[63] Yu G, Smith DK, Zhu H, Guan Y and Lam TTY: **ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data**. *Methods in Ecology and Evolution* 2017. 8(1):28–36.

[64] Pan I: **Torrent suite**. *GitHub* 2016. URL `https://github.com/iontorrent/TS/blob/master/Analysis/Converter/bam2sff.cpp`. Accessed March, 2016.

[65] Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ *et al.*: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities**. *Applied and Environmental Microbiology* 2009. 75(23):7537–7541.

[66] Tatusov RL, Koonin EV and Lipman DJ: **A genomic perspective on protein families**. *Science* 1997. 278(5338):631–637.

[67] Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, Graham MR and Sharma MK: **Comparison of sample preparation methods used for the next-generation sequencing of *Mycobacterium tuberculosis***. *PLoS ONE* 2016. 11(2):e0148676.

[68] Kumar A and Kaur J: **Primer based approach for PCR amplification of high GC content gene: *Mycobacterium* gene as a model**. *Molecular Biology International* 2014. 2014:937308.

[69] Ioerger TR, Feng Y, Ganesula K, Chen X, Dobos KM, Fortune S, Jacobs WR, Mizrahi V, Parish T, Rubin E *et al.*: **Variation among genome sequences of H37Rv strains of *Mycobacterium tuberculosis* from multiple laboratories**. *Journal of Bacteriology* 2010. 192(14):3645–3653.

[70] Duque C, Arroyo L, Ortega H, Montúfar F, Ortíz B, Rojas M and Barrera LF: **Different responses of human mononuclear phagocyte populations to *Mycobacterium tuberculosis***. *Tuberculosis* 2014. 94(2):111–122.

[71] Gopal R, Monin L, Slight S, Uche U, Blanchard E, Junecko BAF, Ramos-Payan R, Stallings CL, Reinhart TA, Kolls JK *et al.*: **Unexpected role for IL-17 in protective immunity against hypervirulent *Mycobacterium tuberculosis* hn878 infection**. *PLoS Pathogens* 2014. 10(5):e1004099.

[72] WorldHealth Organization: **Model list of essential medicines** 2017. URL `http://www.who.int/selection_medicines/list/en/`. Accessed September, 2017.

[73] Wainwright M: **Streptomycin: discovery and resultant controversy**. *History and Philosophy of the Life Sciences* 1991. pages 97–124.

[74] Zierski M and Bek E: **Side-effects of drug regimens used in short-course chemotherapy for pulmonary tuberculosis. a controlled clinical study**. *Tubercle* 1980. 61(1):41–49.

[75] Selimoglu E: **Aminoglycoside-induced ototoxicity**. *Current Pharmaceutical Design* 2007. 13(1):119–126.

[76] Song T, Park Y, Shamputa IC, Seo S, Lee SY, Jeon HS, Choi H, Lee M, Glynne RJ, Barnes SW *et al.*: **Fitness costs of rifampicin resistance in *Mycobacterium tuberculosis* are amplified under conditions of nutrient starvation and compensated by mutation in the beta subunit of RNA polymerase**. *Molecular Microbiology* 2014. 91(6):1106–1119.

[77] Alvarez N, Haft D, Hurtado U, Robledo J and Rouzaud F: **Whole-genome sequencing of a haarlem extensively drug-resistant mycobacterium tuberculosis clinical isolate from Medellín, Colombia**. *Genome Announcements* 2016. 4(3):e00566–16.

[78] Laurence YV, Griffiths UK and Vassall A: **Costs to health services and the patient of treating tuberculosis: a systematic literature review**. *Pharmacoeconomics* 2015. 33(9):939–955.

[79] Pfyffer GE and Wittwer F: **Incubation time of mycobacterial cultures: how long is long enough to issue a final negative report to the clinician?** *Journal of Clinical Microbiology* 2012. 50(12):4188–4189.

[80] Migliori GB, Matteelli A, Cirillo D and Pai M: **Diagnosis of multidrug-resistant tuberculosis and extensively drug-resistant tuberculosis: Current standards and challenges**. *Canadian Journal of Infectious Diseases and Medical Microbiology* 2008. 19(2):169–172.

[81] Hillemann D, Rüsch-Gerdes S, Boehme C and Richter E: **Rapid molecular detection of extrapulmonary tuberculosis by the automated GeneXpert MTB/RIF system**. *Journal of Clinical Microbiology* 2011. 49(4):1202–1205.

[82] Donnabella V, Martiniuk F, Kinney D, Bacerdo M, Bonk S, Hanna B and Rom WN: **Isolation of the gene for the beta subunit of RNA polymerase from rifampicin-resistant *Mycobacterium tuberculosis* and identification of new mutations.** *American Journal of Respiratory Cell and Molecular Biology* 1994. 11(6):639–643.

[83] Musser JM, Kapur V, Williams DL, Kreiswirth BN, Van Soolingen D and Van Embden JD: **Characterization of the catalase-peroxidase gene (katG) and inhA locus in isoniazid-resistant and-susceptible strains of *Mycobacterium tuberculosis* by automated DNA sequencing: restricted array of mutations associated with drug resistance**. *The Journal of Infectious Diseases* 1996. 173(1):196–202.

[84] Kapur V, Li LL, Hamrick MR, Plikaytis BB, Shinnick TM, Telenti A, Jacobs Jr WR, Banerjee A, Cole S and Yuen KY: **Rapid mycobacterium species assignment and unambiguous identification of mutations associated with antimicrobial resistance in *Mycobacterium tuberculosis* by automated DNA sequencing.** *Archives of Pathology & Laboratory Medicine* 1995. 119(2):131–138.

[85] Finken M, Kirschner P, Meier A, Wrede A and Böttger EC: **Molecular basis of streptomycin resistance in *Mycobacterium tuberculosis*: alterations of the ribosomal protein S12 gene and point mutations within a functional 16S ribosomal RNA pseudoknot**. *Molecular Microbiology* 1993. 9(6):1239–1246.

[86] Sreevatsan S, Stockbauer KE, Pan X, Kreiswirth BN, Moghazeh SL, Jacobs WR, Telenti A and Musser JM: **Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of embB mutations.** *Antimicrobial Agents and Chemotherapy* 1997. 41(8):1677–1681.

[87] Ramaswamy SV, Amin AG, Göksel S, Stager CE, Dou SJ, El Sahly H, Moghazeh SL, Kreiswirth BN and Musser JM: **Molecular genetic analysis of nucleotide polymorphisms associated with ethambutol resistance in human isolates of *Mycobacterium tuberculosis***. *Antimicrobial Agents and Chemotherapy* 2000. 44(2):326–336.

[88] Hirano K, Takahashi M, Kazumi Y, Fukasawa Y and Abe C: **Mutation in pncA is a major mechanism of pyrazinamide resistance in *Mycobacterium tuberculosis***. *Tubercle and Lung Disease* 1998. 78(2):117–122.

[89] Kapur V, Li LL, Iordanescu S, Hamrick MR, Wanger A, Kreiswirth BN and Musser JM: **Characterization by automated DNA sequencing of mutations in the gene (rpoB) encoding the RNA polymerase beta subunit in rifampin-resistant *Mycobacterium tuberculosis* strains from New York City and Texas.** *Journal of Clinical Microbiology* 1994. 32(4):1095–1098.

[90] Chan RC, Hui M, Chan EW, Au T, Chin ML, Yip CK, AuYeang CK, Yeung CY, Kam KM, Yip PC *et al.*: **Genetic and phenotypic characterization of drug-resistant *Mycobacterium tuberculosis* isolates in Hong Kong**. *Journal of Antimicrobial Chemotherapy* 2007. 59(5):866–873.

[91] Nair J, Rouse DA, Bai GH and Morris SL: **The rpsL gene and streptomycin resistance in single and multiple drug-resistant strains of *Mycobacterium tuberculosis***. *Molecular Microbiology* 1993. 10(3):521–527.

[92] Heym B, Alzari PM, Honore N and Cole ST: **Missense mutations in the catalase-peroxidase gene, katG, are associated with isoniazid resistance in *Mycobacterium tuberculosis***. *Molecular Microbiology* 1995. 15(2):235–245.

[93] Ramaswamy SV, Reich R, Dou SJ, Jasperse L, Pan X, Wanger A, Quitugua T and Graviss EA: **Single nucleotide polymorphisms in genes associated with isoniazid resistance in *Mycobacterium tuberculosis***. *Antimicrobial Agents and Chemotherapy* 2003. 47(4):1241–1250.

[94] Okamoto S, Tamaru A, Nakajima C, Nishimura K, Tanaka Y, Tokuyama S, Suzuki Y and Ochi K: **Loss of a conserved 7-methylguanosine modification in 16S rRNA confers low-level streptomycin resistance in bacteria**. *Molecular Microbiology* 2007. 63(4):1096–1106.

[95] Lee H, Cho S, Bang H, Lee J, Bai G, Kim S and Kim J: **Exclusive mutations related to isoniazid and ethionamide resistance among *Mycobacterium tuberculosis* isolates from Korea**. *The International Journal of Tuberculosis and Lung Disease* 2000. 4(5):441–447.

[96] Louw G, Warren R, Donald P, Murray M, Bosman M, Van Helden P, Young D and Victor T: **Frequency and implications of pyrazinamide resistance in managing previously treated tuberculosis patients**. *The International Journal of Tuberculosis and Lung Disease* 2006. 10(7):802–807.

[97] Takiff HE, Salazar L, Guerrero C, Philipp W, Huang WM, Kreiswirth B, Cole ST, Jacobs WR and Telenti A: **Cloning and nucleotide sequence of *Mycobacterium tuberculosis* gyrA and gyrB genes and detection of quinolone resistance mutations.** *Antimicrobial Agents and Chemotherapy* 1994. 38(4):773–780.

[98] Moghazeh SL, Pan X, Arain T, Stover CK, Musser JM and Kreiswirth BN: **Comparative antimycobacterial activities of rifampin, rifapentine, and KRM-1648 against a collection of rifampin-resistant *Mycobacterium tuberculosis* isolates with known rpoB mutations.** *Antimicrobial Agents and Chemotherapy* 1996. 40(11):2655–2657.

[99] Morlock GP, Metchock B, Sikes D, Crawford JT and Cooksey RC: **ethA, inhA, and katG loci of ethionamide-resistant clinical *Mycobacterium tuberculosis* isolates**. *Antimicrobial Agents and Chemotherapy* 2003. 47(12):3799–3805.

[100] Scorpio A, Lindholm-Levy P, Heifets L, Gilman R, Siddiqi S, Cynamon M and Zhang Y: **Characterization of pncA mutations in pyrazinamide-resistant *Mycobacterium tuberculosis*.** *Antimicrobial Agents and Chemotherapy* 1997. 41(3):540–543.

[101] Kim BJ, Kim SY, Park BH, Lyu MA, Park IK, Bai GH, Kim SJ, Cha CY and Kook YH: **Mutations in the rpoB gene of *Mycobacterium tuberculosis* that interfere with PCR-single-strand conformation polymorphism analysis for rifampin susceptibility testing.** *Journal of Clinical Microbiology* 1997. 35(2):492–494.

[102] Haas WH, Schilke K, Brand J, Amthor B, Weyer K, Fourie PB, Bretzel G, Sticht-Groh V and Bremer HJ: **Molecular analysis of katG gene mutations in strains of *Mycobacterium tuberculosis* complex from Africa.** *Antimicrobial Agents and Chemotherapy* 1997. 41(7):1601–1603.

[103] Scorpio A and Zhang Y: **Mutations in pncA, a gene encoding pyrazinamidase/nicotinamidase, cause resistance to the antituberculous drug pyrazinamide in tubercle bacillus**. *Nature Medicine* 1996. 2(6):662–667.

[104] Srivastava S, Ayyagari A, Dhole TN, Nyati KK and Dwivedi SK: **emb nucleotide polymorphisms and the role of embB306 mutations in *Mycobacterium tuberculosis* resistance to ethambutol**. *International Journal of Medical Microbiology* 2009. 299(4):269–280.

[105] Cardoso RF, Cardoso MA, Leite CQF, Sato DN, Mamizuka EM, Hirata RDC, Mello FFd and Hirata MH: **Characterization of *ndh* gene of isoniazid resistant and susceptible *Mycobacterium tuberculosis* isolates from Brazil**. *Memórias do Instituto Oswaldo Cruz* 2007. 102(1):59–61.

[106] Bodmer T, Zürcher G, Imboden P and Telenti A: **Mutation position and type of substitution in the β-subunit of the RNA polymerase influence in-vitro activity of rifamycins in rifampicin-resistant *Mycobacterium tuberculosis***. *Journal of Antimicrobial Chemotherapy* 1995. 35(2):345–348.

[107] Ramaswamy SV, Dou SJ, Rendon A, Yang Z, Cave MD and Graviss EA: **Genotypic analysis of multidrug-resistant *Mycobacterium tuberculosis* isolates from Monterrey, Mexico**. *Journal of Medical Microbiology* 2004. 53(2):107–113.

Appendix A.

SNP Reports

A report for each strain was generated, including all found drug resistances, SNPs in drug related genes and other high impact single nucleotide polymorphisms.

Impact prediction SnpEff reports putative variant impact in order to make it easier to quickly categorize and prioritize variants.

Impact categories must be used with care, they were created only to help and simplify the filtering process. Obviously, there is no way to predict whether a HIGH impact or a LOW impact variant is the one producing a phenotype of interest.

Strain: 10019

Variant Information	SNP count
GATK raw output	454
Intersection GATK - Nucmer	432
resulting high quality SNPs	417

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2288955	Rv2043c	pncA	96	Lys/Met	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4242928	Rv3793	embC	1022	Ser/Ser	LOW	synonymous_variant	EMB
4247595	Rv3795	embB	361	Cys/Ser	MODERATE	missense_variant	EMB
4407712	Rv3919c	gid	164	Gly/Ala	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 10237

Variant Information	SNP count
GATK raw output	707
Intersection GATK - Nucmer	668
resulting high quality SNPs	646

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 10400

Variant Information	SNP count
GATK raw output	795
Intersection GATK - Nucmer	761
resulting high quality SNPs	739

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247402	Rv3795	embB	297	Ser/Ala	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3646964	Rv3266c	rmlD	282	Leu/Leu	LOW	synonymous_variant	EMB
4007595	Rv3566c	nhoA	196	Ala/Ala	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4249594	Rv3795	embB	1027	Thr/Thr	LOW	synonymous_variant	EMB
4326632	Rv3854c	ethA	281	His/Pro	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3811720	Rv3395A	Rv3395A	p.Val1?	p.Val1?	HIGH	start_lost
3867404	Rv3447c	eccC4	279	Glu/*	HIGH	stop_gained

Strain: 10485

Variant Information	SNP count
GATK raw output	756
Intersection GATK - Nucmer	715
resulting high quality SNPs	696

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247402	Rv3795	embB	297	Ser/Ala	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3646964	Rv3266c	rmlD	282	Leu/Leu	LOW	synonymous_variant	EMB
4007595	Rv3566c	nhoA	196	Ala/Ala	LOW	synonymous_variant	INH
4249594	Rv3795	embB	1027	Thr/Thr	LOW	synonymous_variant	EMB
4326632	Rv3854c	ethA	281	His/Pro	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1989850	Rv1759c	wag22	910	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3811720	Rv3395A	Rv3395A	p.Val1?	p.Val1?	HIGH	start_lost
3867404	Rv3447c	eccC4	279	Glu/*	HIGH	stop_gained

Strain: 10755

Variant Information	SNP count
GATK raw output	745
Intersection GATK - Nucmer	710
resulting high quality SNPs	691

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247402	Rv3795	embB	297	Ser/Ala	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2288971	Rv2043c	pncA	91	Glu/*	HIGH	stop_gained	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3646964	Rv3266c	rmlD	282	Leu/Leu	LOW	synonymous_variant	EMB
4007595	Rv3566c	nhoA	196	Ala/Ala	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4249594	Rv3795	embB	1027	Thr/Thr	LOW	synonymous_variant	EMB
4326632	Rv3854c	ethA	281	His/Pro	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3811720	Rv3395A	Rv3395A	p.Val1?	p.Val1?	HIGH	start_lost
3867404	Rv3447c	eccC4	279	Glu/*	HIGH	stop_gained

Strain: 10851

Variant Information	SNP count
GATK raw output	713
Intersection GATK - Nucmer	675
resulting high quality SNPs	655

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4327028	Rv3854c	ethA	149	Pro/Leu	MODERATE	missense_variant	ETH
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 10862

Variant Information	SNP count
GATK raw output	447
Intersection GATK - Nucmer	411
resulting high quality SNPs	401

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761095	Rv0667	rpoB	511	Leu/Pro	RIF	[89]
761284	Rv0667	rpoB	574	Ser/Leu	RIF	[90]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247429	Rv3795	embB	306	Met/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4007761	Rv3566c	nhoA	141	Leu/Pro	MODERATE	missense_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4245316	Rv3794	embA	695	Val/Ala	MODERATE	missense_variant	EMB
4249732	Rv3795	embB	1073	Ala/Ala	LOW	synonymous_variant	EMB
4326497	Rv3854c	ethA	326	Arg/Gln	MODERATE	missense_variant	ETH
4407716	Rv3919c	gid	163	Lys/Glu	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2282611	Rv2036	Rv2036	171	Trp/*	HIGH	stop_gained
2333325	Rv2077c	Rv2077c	324	Ter/Glu	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3870025	Rv3449	mycP4	92	Gln/*	HIGH	stop_gained

Strain: 1117

Variant Information	SNP count
GATK raw output	412
Intersection GATK - Nucmer	396
resulting high quality SNPs	384

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 11572

Variant Information	SNP count
GATK raw output	1203
Intersection GATK - Nucmer	1144
resulting high quality SNPs	1102

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761139	Rv0667	rpoB	526	His/Tyr	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2154724	Rv1908c	katG	463	Arg/Leu	INH	[92]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2289016	Rv2043c	pncA	76	Thr/Pro	PZA	[86]
2521428	Rv2247	accD6	229	Asp/Gly	INH	[93]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]
4407927	Rv3919c	gid	92	Glu/Asp	SM	[94]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6735	Rv0005	gyrB	499	Asn/Thr	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
156608	Rv0129c	fbpC	331	Pro/Pro	LOW	synonymous_variant	INH
763031	Rv0667	rpoB	1156	Ala/Ala	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1793530	Rv1592c	Rv1592c	71	Val/Phe	MODERATE	missense_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
2726300	Rv2428	ahpC	36	Thr/Thr	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4243460	Rv3794	embA	76	Cys/Cys	LOW	synonymous_variant	EMB
4407588	Rv3919c	gid	205	Ala/Ala	LOW	synonymous_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
65150	Rv0061c	Rv0061c	67	Trp/*	HIGH	stop_gained
123454	Rv0104	Rv0104	380	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
392261	Rv0325	Rv0325	75	Ter/Gln	HIGH	stop_lost&splice_region_variant
887668	Rv0794c	lpdB	323	Trp/*	HIGH	stop_gained
932280	Rv0836c	Rv0836c	218	Ter/Trp	HIGH	stop_lost&splice_region_variant
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1502499	Rv1333	Rv1333	301	Gln/*	HIGH	stop_gained
2337179	Rv2079	Rv2079	609	Gln/*	HIGH	stop_gained
2442468	Rv2180c	Rv2180c	249	Trp/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
4383655	Rv3898c	Rv3898c	111	Ter/Gln	HIGH	stop_lost&splice_region_variant

Strain: 12400

Variant Information	SNP count
GATK raw output	452
Intersection GATK - Nucmer	422
resulting high quality SNPs	404

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6575	Rv0005	gyrB	446	Arg/Cys	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 12442

Variant Information	SNP count
GATK raw output	409
Intersection GATK - Nucmer	397
resulting high quality SNPs	388

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2289081	Rv2043c	pncA	54	Pro/Leu	PZA	[96]
4248003	Rv3795	embB	497	Gln/Arg	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9432	Rv0006	gyrA	711	Leu/Met	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
877859	Rv0783c	emrB	194	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1699892	Rv1508c	Rv1508c	p.Val1?	p.Val1?	HIGH	start_lost
2121839	Rv1871c	Rv1871c	16	Gln/*	HIGH	stop_gained
2930369	Rv2602	vapC41	9	Trp/*	HIGH	stop_gained
3047878	Rv2735c	Rv2735c	225	Tyr/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3866717	Rv3447c	eccC4	508	Gln/*	HIGH	stop_gained

Strain: 12488

Variant Information	SNP count
GATK raw output	509
Intersection GATK - Nucmer	497
resulting high quality SNPs	482

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761109	Rv0667	rpoB	516	Asp/Tyr	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4248027	Rv3795	embB	505	Ala/Val	MODERATE	missense_variant	EMB
4249732	Rv3795	embB	1073	Ala/Ala	LOW	synonymous_variant	EMB
4407799	Rv3919c	gid	135	Val/Gly	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1341231	Rv1198	esxL	76	Gln/*	HIGH	stop_gained
2333325	Rv2077c	Rv2077c	324	Ter/Glu	HIGH	stop_lost&splice_region_variant
2884044	Rv2564	glnQ	235	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3870025	Rv3449	mycP4	92	Gln/*	HIGH	stop_gained

Strain: 12650

Variant Information	SNP count
GATK raw output	457
Intersection GATK - Nucmer	430
resulting high quality SNPs	413

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6575	Rv0005	gyrB	446	Arg/Cys	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2159204	Rv1913	Rv1913	5	Trp/*	HIGH	stop_gained
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 13050

Variant Information	SNP count
GATK raw output	431
Intersection GATK - Nucmer	413
resulting high quality SNPs	394

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6575	Rv0005	gyrB	446	Arg/Cys	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 13449

Variant Information	SNP count
GATK raw output	418
Intersection GATK - Nucmer	404
resulting high quality SNPs	385

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6575	Rv0005	gyrB	446	Arg/Cys	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
2159204	Rv1913	Rv1913	5	Trp/*	HIGH	stop_gained
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 13452

Variant Information	SNP count
GATK raw output	352
Intersection GATK - Nucmer	337
resulting high quality SNPs	325

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7582	Rv0006	gyrA	94	Asp/Gly	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761161	Rv0667	rpoB	533	Leu/Pro	RIF	[98]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2153999	Rv1908c	katG	705	Arg/Trp	MODERATE	missense_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2245713	Rv2000	Rv2000	169	Gln/*	HIGH	stop_gained

Strain: 13689

Variant Information	SNP count
GATK raw output	430
Intersection GATK - Nucmer	419
resulting high quality SNPs	408

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7570	Rv0006	gyrA	90	Ala/Val	FLQ	[97]
1674263	Rv1484	inhA	21	Ile/Thr	INH	[99]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
760314	Rv0667	rpoB	251	Val/Phe	MODERATE	missense_variant	RIF
1792785	Rv1592c	Rv1592c	319	Thr/Met	MODERATE	missense_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2288953	Rv2043c	pncA	97	Gly/Cys	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4245553	Rv3794	embA	774	Pro/His	MODERATE	missense_variant	EMB
4247574	Rv3795	embB	354	Asp/Ala	MODERATE	missense_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
279919	Rv0234c	gabD1	408	Lys/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1207386	Rv1082	mca	289	Ter/Trp	HIGH	stop_lost&splice_region_variant
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 1374

Variant Information	SNP count
GATK raw output	817
Intersection GATK - Nucmer	779
resulting high quality SNPs	725

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
39061	Rv0036c	Rv0036c	257	Arg/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2723544	Rv2426c	Rv2426c	214	Arg/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 14114

Variant Information	SNP count
GATK raw output	750
Intersection GATK - Nucmer	712
resulting high quality SNPs	688

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288955	Rv2043c	pncA	96	Lys/Thr	PZA	[100]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247429	Rv3795	embB	306	Met/Val	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760106	Rv0667	rpoB	181	Ser/Ser	LOW	synonymous_variant	RIF
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
972215	Rv0873	fadE10	571	Gln/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4013660	Rv3572	Rv3572	50	Cys/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 14163

Variant Information	SNP count
GATK raw output	794
Intersection GATK - Nucmer	752
resulting high quality SNPs	726

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761139	Rv0667	rpoB	526	His/Tyr	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288826	Rv2043c	pncA	139	Val/Ala	PZA	[86]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
7638	Rv0006	gyrA	113	Gln/Glu	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
410579	Rv0341	iniB	406	Ala/Ala	LOW	synonymous_variant	EMB
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
762973	Rv0667	rpoB	1137	Gln/Arg	MODERATE	missense_variant	RIF
1674769	Rv1484	inhA	190	Ala/Ser	MODERATE	missense_variant	ETH
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4246374	Rv3794	embA	1048	Gly/Arg	MODERATE	missense_variant	EMB
4247574	Rv3795	embB	354	Asp/Ala	MODERATE	missense_variant	EMB
4408084	Rv3919c	gid	40	Glu/Gly	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
249203	Rv0209	Rv0209	56	Gln/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 15097

Variant Information	SNP count
GATK raw output	701
Intersection GATK - Nucmer	659
resulting high quality SNPs	641

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2114135	Rv1866	Rv1866	332	Tyr/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 15696

Variant Information	SNP count
GATK raw output	731
Intersection GATK - Nucmer	695
resulting high quality SNPs	675

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]
4249583	Rv3795	embB	1024	Asp/Asn	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 16106

Variant Information	SNP count
GATK raw output	465
Intersection GATK - Nucmer	438
resulting high quality SNPs	412

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761101	Rv0667	rpoB	513	Gln/Pro	RIF	[101]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247429	Rv3795	embB	306	Met/Val	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289081	Rv2043c	pncA	54	Pro/Gln	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
1524	Rv0001	dnaA	508	Ter/Tyr	HIGH	stop_lost&splice_region_variant
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 16694

Variant Information	SNP count
GATK raw output	741
Intersection GATK - Nucmer	719
resulting high quality SNPs	702

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247402	Rv3795	embB	297	Ser/Ala	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3646964	Rv3266c	rmlD	282	Leu/Leu	LOW	synonymous_variant	EMB
4007595	Rv3566c	nhoA	196	Ala/Ala	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4249594	Rv3795	embB	1027	Thr/Thr	LOW	synonymous_variant	EMB
4326632	Rv3854c	ethA	281	His/Pro	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3811720	Rv3395A	Rv3395A	p.Val1?	p.Val1?	HIGH	start_lost
3867404	Rv3447c	eccC4	279	Glu/*	HIGH	stop_gained

Strain: 1770

Variant Information	SNP count
GATK raw output	416
Intersection GATK - Nucmer	394
resulting high quality SNPs	380

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6575	Rv0005	gyrB	446	Arg/Cys	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 2025

Variant Information	SNP count
GATK raw output	436
Intersection GATK - Nucmer	425
resulting high quality SNPs	408

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761095	Rv0667	rpoB	511	Leu/Pro	RIF	[89]
761284	Rv0667	rpoB	574	Ser/Leu	RIF	[90]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247429	Rv3795	embB	306	Met/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4007761	Rv3566c	nhoA	141	Leu/Pro	MODERATE	missense_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4245316	Rv3794	embA	695	Val/Ala	MODERATE	missense_variant	EMB
4249732	Rv3795	embB	1073	Ala/Ala	LOW	synonymous_variant	EMB
4326497	Rv3854c	ethA	326	Arg/Gln	MODERATE	missense_variant	ETH
4407716	Rv3919c	gid	163	Lys/Glu	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2282611	Rv2036	Rv2036	171	Trp/*	HIGH	stop_gained
2333325	Rv2077c	Rv2077c	324	Ter/Glu	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3870025	Rv3449	mycP4	92	Gln/*	HIGH	stop_gained

Strain: 2162

Variant Information	SNP count
GATK raw output	719
Intersection GATK - Nucmer	693
resulting high quality SNPs	672

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7570	Rv0006	gyrA	90	Ala/Val	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4248003	Rv3795	embB	497	Gln/Arg	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 2196

Variant Information	SNP count
GATK raw output	458
Intersection GATK - Nucmer	430
resulting high quality SNPs	407

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 2263

Variant Information	SNP count
GATK raw output	715
Intersection GATK - Nucmer	685
resulting high quality SNPs	667

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761109	Rv0667	rpoB	516	Asp/Tyr	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Asn	INH	[102]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
761771	Rv0667	rpoB	736	Met/Ile	MODERATE	missense_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2289100	Rv2043c	pncA	48	Lys/Glu	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1044637	Rv0936	pstA2	107	Tyr/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 2380

Variant Information	SNP count
GATK raw output	693
Intersection GATK - Nucmer	664
resulting high quality SNPs	645

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4326893	Rv3854c	ethA	194	Leu/Pro	MODERATE	missense_variant	ETH
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1399239	Rv1251c	Rv1251c	p.Val1?	p.Val1?	HIGH	start_lost
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 2417

Variant Information	SNP count
GATK raw output	695
Intersection GATK - Nucmer	674
resulting high quality SNPs	657

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7582	Rv0006	gyrA	94	Asp/Ala	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
761582	Rv0667	rpoB	673	Glu/Asp	MODERATE	missense_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4249680	Rv3795	embB	1056	Asp/Ala	MODERATE	missense_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 2435

Variant Information	SNP count
GATK raw output	700
Intersection GATK - Nucmer	668
resulting high quality SNPs	648

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7570	Rv0006	gyrA	90	Ala/Val	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4248003	Rv3795	embB	497	Gln/Arg	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 2599

Variant Information	SNP count
GATK raw output	706
Intersection GATK - Nucmer	666
resulting high quality SNPs	647

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]
4249583	Rv3795	embB	1024	Asp/Asn	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6647	Rv0005	gyrB	470	Gly/Ser	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 259

Variant Information	SNP count
GATK raw output	457
Intersection GATK - Nucmer	427
resulting high quality SNPs	415

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7582	Rv0006	gyrA	94	Asp/Gly	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761161	Rv0667	rpoB	533	Leu/Pro	RIF	[98]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2153999	Rv1908c	katG	705	Arg/Trp	MODERATE	missense_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2245713	Rv2000	Rv2000	169	Gln/*	HIGH	stop_gained

Strain: 260

Variant Information	SNP count
GATK raw output	731
Intersection GATK - Nucmer	694
resulting high quality SNPs	672

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7570	Rv0006	gyrA	90	Ala/Val	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4248003	Rv3795	embB	497	Gln/Arg	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4327281	Rv3854c	ethA	65	Arg/*	HIGH	stop_gained	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 3212

Variant Information	SNP count
GATK raw output	663
Intersection GATK - Nucmer	631
resulting high quality SNPs	614

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7582	Rv0006	gyrA	94	Asp/Ala	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
761582	Rv0667	rpoB	673	Glu/Asp	MODERATE	missense_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4249680	Rv3795	embB	1056	Asp/Ala	MODERATE	missense_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 3537

Variant Information	SNP count
GATK raw output	725
Intersection GATK - Nucmer	695
resulting high quality SNPs	673

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761139	Rv0667	rpoB	526	His/Tyr	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247429	Rv3795	embB	306	Met/Val	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4407982	Rv3919c	gid	74	Leu/*	HIGH	stop_gained	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 3702

Variant Information	SNP count
GATK raw output	736
Intersection GATK - Nucmer	709
resulting high quality SNPs	684

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288820	Rv2043c	pncA	141	Gln/Pro	PZA	[103]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247730	Rv3795	embB	406	Gly/Asp	EMB	[104]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6737	Rv0005	gyrB	500	Thr/Ala	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4247013	Rv3795	embB	167	Pro/Arg	MODERATE	missense_variant	EMB
4326900	Rv3854c	ethA	192	Pro/Thr	MODERATE	missense_variant	ETH
4408094	Rv3919c	gid	37	Gly/Arg	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
703839	Rv0609A	Rv0609A	4	Gln/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2241122	Rv1997	ctpF	322	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 4082

Variant Information	SNP count
GATK raw output	769
Intersection GATK - Nucmer	733
resulting high quality SNPs	707

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 4538

Variant Information	SNP count
GATK raw output	775
Intersection GATK - Nucmer	734
resulting high quality SNPs	706

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 4619

Variant Information	SNP count
GATK raw output	780
Intersection GATK - Nucmer	742
resulting high quality SNPs	718

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2289208	Rv2043c	pncA	12	Asp/Asn	PZA	[100]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4249512	Rv3795	embB	1000	Met/Arg	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2518676	Rv2245	kasA	188	Gly/Ser	MODERATE	missense_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4327271	Rv3854c	ethA	68	Pro/Arg	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 4737

Variant Information	SNP count
GATK raw output	739
Intersection GATK - Nucmer	705
resulting high quality SNPs	684

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]
4249583	Rv3795	embB	1024	Asp/Asn	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6529	Rv0005	gyrB	430	Lys/Asn	MODERATE	missense_variant	FLQ
6647	Rv0005	gyrB	470	Gly/Ser	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 4793

Variant Information	SNP count
GATK raw output	786
Intersection GATK - Nucmer	765
resulting high quality SNPs	743

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247402	Rv3795	embB	297	Ser/Ala	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3646964	Rv3266c	rmlD	282	Leu/Leu	LOW	synonymous_variant	EMB
4007595	Rv3566c	nhoA	196	Ala/Ala	LOW	synonymous_variant	INH
4242553	Rv3793	embC	897	Arg/Arg	LOW	synonymous_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4249594	Rv3795	embB	1027	Thr/Thr	LOW	synonymous_variant	EMB
4326632	Rv3854c	ethA	281	His/Pro	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3811720	Rv3395A	Rv3395A	p.Val1?	p.Val1?	HIGH	start_lost
3867404	Rv3447c	eccC4	279	Glu/*	HIGH	stop_gained

Strain: 5303

Variant Information	SNP count
GATK raw output	742
Intersection GATK - Nucmer	714
resulting high quality SNPs	691

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 5450

Variant Information	SNP count
GATK raw output	788
Intersection GATK - Nucmer	746
resulting high quality SNPs	717

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]
4249583	Rv3795	embB	1024	Asp/Asn	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 5603

Variant Information	SNP count
GATK raw output	673
Intersection GATK - Nucmer	641
resulting high quality SNPs	616

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
2102990	Rv1854c	ndh	18	Val/Ala	INH	[105]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
411371	Rv0342	iniA	178	Gly/Gly	LOW	synonymous_variant	EMB
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4007564	Rv3566c	nhoA	207	Gly/Arg	MODERATE	missense_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4326019	Rv3854c	ethA	485	Ala/Ala	LOW	synonymous_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
948270	Rv0851c	Rv0851c	67	Arg/*	HIGH	stop_gained
950955	Rv0853c	pdc	55	Trp/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1769101	Rv1563c	treY	111	Trp/*	HIGH	stop_gained
2054306	Rv1812c	Rv1812c	352	Glu/*	HIGH	stop_gained
3675446	Rv3294c	Rv3294c	184	Arg/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3898809	Rv3479	Rv3479	997	Trp/*	HIGH	stop_gained

Strain: 6185

Variant Information	SNP count
GATK raw output	757
Intersection GATK - Nucmer	711
resulting high quality SNPs	689

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7581	Rv0006	gyrA	94	Asp/Tyr	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761139	Rv0667	rpoB	526	His/Tyr	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2289073	Rv2043c	pncA	57	His/Asp	PZA	[86]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
410579	Rv0341	iniB	406	Ala/Ala	LOW	synonymous_variant	EMB
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
762973	Rv0667	rpoB	1137	Gln/Arg	MODERATE	missense_variant	RIF
1674769	Rv1484	inhA	190	Ala/Ser	MODERATE	missense_variant	ETH
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4246374	Rv3794	embA	1048	Gly/Arg	MODERATE	missense_variant	EMB
4247574	Rv3795	embB	354	Asp/Ala	MODERATE	missense_variant	EMB
4408084	Rv3919c	gid	40	Glu/Gly	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
2094156	Rv1844c	gnd1	345	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 6978

Variant Information	SNP count
GATK raw output	445
Intersection GATK - Nucmer	428
resulting high quality SNPs	407

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4327427	Rv3854c	ethA	16	Gly/Asp	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 6993

Variant Information	SNP count
GATK raw output	731
Intersection GATK - Nucmer	686
resulting high quality SNPs	665

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 7101

Variant Information	SNP count
GATK raw output	718
Intersection GATK - Nucmer	678
resulting high quality SNPs	658

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761139	Rv0667	rpoB	526	His/Tyr	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247429	Rv3795	embB	306	Met/Val	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4407982	Rv3919c	gid	74	Leu/*	HIGH	stop_gained	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 7170

Variant Information	SNP count
GATK raw output	766
Intersection GATK - Nucmer	731
resulting high quality SNPs	709

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761139	Rv0667	rpoB	526	His/Tyr	RIF	[89]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247429	Rv3795	embB	306	Met/Val	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4407982	Rv3919c	gid	74	Leu/*	HIGH	stop_gained	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 7193

Variant Information	SNP count
GATK raw output	752
Intersection GATK - Nucmer	712
resulting high quality SNPs	685

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
29869	Rv0026	Rv0026	50	Glu/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 7630

Variant Information	SNP count
GATK raw output	689
Intersection GATK - Nucmer	656
resulting high quality SNPs	637

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2288932	Rv2043c	pncA	104	Ser/Arg	PZA	[88]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]
4249583	Rv3795	embB	1024	Asp/Asn	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6647	Rv0005	gyrB	470	Gly/Ser	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4326494	Rv3854c	ethA	327	Leu/Pro	MODERATE	missense_variant	ETH
4408050	Rv3919c	gid	51	Asn/Lys	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 7649

Variant Information	SNP count
GATK raw output	407
Intersection GATK - Nucmer	392
resulting high quality SNPs	382

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Trp	RIF	[106]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4242928	Rv3793	embC	1022	Ser/Ser	LOW	synonymous_variant	EMB
4247595	Rv3795	embB	361	Cys/Ser	MODERATE	missense_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 7656

Variant Information	SNP count
GATK raw output	617
Intersection GATK - Nucmer	601
resulting high quality SNPs	581

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761111	Rv0667	rpoB	516	Asp/Glu	RIF	[89]
761139	Rv0667	rpoB	526	His/Asn	RIF	[107]
2102990	Rv1854c	ndh	18	Val/Ala	INH	[105]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
2289192	Rv2043c	pncA	17	Gly/Asp	PZA	[88]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
411371	Rv0342	iniA	178	Gly/Gly	LOW	synonymous_variant	EMB
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4007564	Rv3566c	nhoA	207	Gly/Arg	MODERATE	missense_variant	INH
4247717	Rv3795	embB	402	Leu/Val	MODERATE	missense_variant	EMB
4326019	Rv3854c	ethA	485	Ala/Ala	LOW	synonymous_variant	ETH
4326267	Rv3854c	ethA	403	Cys/Arg	MODERATE	missense_variant	ETH
4407796	Rv3919c	gid	136	Ser/*	HIGH	stop_gained	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
948270	Rv0851c	Rv0851c	67	Arg/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1769101	Rv1563c	treY	111	Trp/*	HIGH	stop_gained
2054306	Rv1812c	Rv1812c	352	Glu/*	HIGH	stop_gained
3675446	Rv3294c	Rv3294c	184	Arg/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3898809	Rv3479	Rv3479	997	Trp/*	HIGH	stop_gained

Strain: 7924

Variant Information	SNP count
GATK raw output	679
Intersection GATK - Nucmer	649
resulting high quality SNPs	624

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247730	Rv3795	embB	406	Gly/Ala	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Assp/Asp	LOW	synonymous_variant	RIF
1674048	Rv1483	mabA	203	Leu/Leu	LOW	synonymous_variant	ETH
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 8139

Variant Information	SNP count
GATK raw output	656
Intersection GATK - Nucmer	625
resulting high quality SNPs	606

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7570	Rv0006	gyrA	90	Ala/Val	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4248003	Rv3795	embB	497	Gln/Arg	EMB	[87]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4327307	Rv3854c	ethA	56	Asp/Ala	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 8208

Variant Information	SNP count
GATK raw output	730
Intersection GATK - Nucmer	694
resulting high quality SNPs	671

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4249518	Rv3795	embB	1002	His/Arg	MODERATE	missense_variant	EMB
4407967	Rv3919c	gid	79	Leu/Ser	MODERATE	missense_variant	SM

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 8759

Variant Information	SNP count
GATK raw output	761
Intersection GATK - Nucmer	709
resulting high quality SNPs	686

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7572	Rv0006	gyrA	91	Ser/Pro	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4326251	Rv3854c	ethA	408	Tyr/Cys	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 9435

Variant Information	SNP count
GATK raw output	775
Intersection GATK - Nucmer	746
resulting high quality SNPs	715

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761161	Rv0667	rpoB	533	Leu/Pro	RIF	[98]
781822	Rv0682	rpsL	88	Lys/Arg	SM	[85]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4326906	Rv3854c	ethA	190	Leu/Phe	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 9480

Variant Information	SNP count
GATK raw output	392
Intersection GATK - Nucmer	381
resulting high quality SNPs	372

No SNPs corresponding to Drug Resistance Mutation Database!

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3645393	Rv3264c	rmlA2	195	Val/Val	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
257071	Rv0214	fadD4	336	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1207386	Rv1082	mca	289	Ter/Trp	HIGH	stop_lost&splice_region_variant
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 9784

Variant Information	SNP count
GATK raw output	764
Intersection GATK - Nucmer	730
resulting high quality SNPs	705

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
7581	Rv0006	gyrA	94	Asp/Asn	FLQ	[97]
7585	Rv0006	gyrA	95	Ser/Thr	FLQ	[84]
761155	Rv0667	rpoB	531	Ser/Leu	RIF	[82]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4242803	Rv3793	embC	981	Val/Leu	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
761248	Rv0667	rpoB	562	Glu/Ala	MODERATE	missense_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289222	Rv2043c	pncA	7	Val/Ala	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
3491105	Rv3125c	Rv3125c	183	Gln/*	HIGH	stop_gained	EMB
3647736	Rv3266c	rmlD	25	Ala/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB
4247574	Rv3795	embB	354	Asp/Ala	MODERATE	missense_variant	EMB
4326786	Rv3854c	ethA	230	Pro/Ala	MODERATE	missense_variant	ETH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained
3791630	Rv3377c	Rv3377c	242	Glu/*	HIGH	stop_gained
3959957	Rv3523	ltp3	143	Tyr/*	HIGH	stop_gained
3966813	Rv3529c	Rv3529c	76	Gln/*	HIGH	stop_gained
4365461	Rv3884c	eccA2	460	Gln/*	HIGH	stop_gained

Strain: 9820

Variant Information	SNP count
GATK raw output	440
Intersection GATK - Nucmer	427
resulting high quality SNPs	406

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761110	Rv0667	rpoB	516	Asp/Val	RIF	[89]
781687	Rv0682	rpsL	43	Lys/Arg	SM	[91]
1674481	Rv1484	inhA	94	Ser/Ala	ETH	[95]
2155168	Rv1908c	katG	315	Ser/Thr	INH	[83]
4247431	Rv3795	embB	306	Met/Ile	EMB	[86]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
6575	Rv0005	gyrB	446	Arg/Cys	MODERATE	missense_variant	FLQ
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2289202	Rv2043c	pncA	14	Cys/Gly	MODERATE	missense_variant	PZA
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH
4241699	Rv3793	embC	613	Met/Val	MODERATE	missense_variant	EMB
4242643	Rv3793	embC	927	Arg/Arg	LOW	synonymous_variant	EMB

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
212244	Rv0180c	Rv0180c	3	Gln/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
3139344	Rv2833c	ugpB	381	Gln/*	HIGH	stop_gained
3435802	Rv3072c	Rv3072c	174	Ser/*	HIGH	stop_gained
3541055	Rv3172c	Rv3172c	104	Glu/*	HIGH	stop_gained
3689523	Rv3303c	lpdA	472	Cys/*	HIGH	stop_gained

Strain: 9937

Variant Information	SNP count
GATK raw output	459
Intersection GATK - Nucmer	432
resulting high quality SNPs	419

SNPs corresponding to Drug Resistance Mutation Database:

Pos	GeneID	Gene	CodonPos	AminoAcid	Drug	Reference
761161	Rv0667	rpoB	533	Leu/Pro	RIF	[98]

SNPs with impact on drug resistance associated genes:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect	Drug
7362	Rv0006	gyrA	21	Glu/Gln	MODERATE	missense_variant	FLQ
9304	Rv0006	gyrA	668	Gly/Asp	MODERATE	missense_variant	FLQ
760115	Rv0667	rpoB	184	Asp/Asp	LOW	synonymous_variant	RIF
1792777	Rv1592c	Rv1592c	322	Ile/Val	MODERATE	missense_variant	INH
1792778	Rv1592c	Rv1592c	321	Glu/Glu	LOW	synonymous_variant	INH
1917972	Rv1694	tlyA	11	Leu/Leu	LOW	synonymous_variant	AMI
2153999	Rv1908c	katG	705	Arg/Trp	MODERATE	missense_variant	INH
2521342	Rv2247	accD6	200	Asp/Asp	LOW	synonymous_variant	INH

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant
1532778	Rv1361c	PPE19	286	Gln/*	HIGH	stop_gained
2245713	Rv2000	Rv2000	169	Gln/*	HIGH	stop_gained

Strain: ATCC

Variant Information	SNP count
GATK raw output	28
Intersection GATK - Nucmer	27
resulting high quality SNPs	26

No SNPs corresponding to Drug Resistance Mutation Database!

No SNPs with impact on drug resistance associated genes!

Other SNPs with high impact:

Pos	GeneID	GeneName	CodonPos	AminoAcid	Impact	Effect
110278	Rv0101	nrp	93	Trp/*	HIGH	stop_gained
234477	Rv0197	Rv0197	749	Tyr/*	HIGH	stop_gained
1037911	Rv0930	pstA1	305	Arg/*	HIGH	stop_gained
1315191	Rv1180	pks3	489	Ter/Tyr	HIGH	stop_lost&splice_region_variant

Appendix B.

Excluded Genes

Table B.1.: List of excluded genes used for high quality SNP extraction

Gene Name	Product	Gene Name	Product
Rv0031	transposase	Rv0797	IS like-2 transposase
Rv1035c	transposase	Rv1036c	transposase
Rv1041c	transposase	Rv1042c	putative transposase
Rv1756c	transposase	Rv1757c	transposase
Rv1763	transposase	Rv1764	transposase
Rv2512c	transposase	Rv2666	IS1533 transposase
Rv2943	transposase	Rv2944	transposase
Rv1054	integrase	Rv2646	integrase
Rv3751	integrase	Rv0151c	PE family protein
Rv0152c	PE family protein	Rv0159c	PE family protein
Rv0160c	PE family protein	Rv0285	PE family protein
Rv0335c	PE family protein	Rv0916c	PE family protein
Rv1040c	PE family protein	Rv1088	PE family protein
Rv1089	PE family protein	Rv1169c	PE family protein
Rv1172c	PE family protein	Rv1195	PE family protein
Rv1214c	PE family protein	Rv1386	PE family protein
Rv1430	PE family protein	Rv1646	PE family protein
Rv1788	PE family protein	Rv1791	PE family protein
Rv1806	PE family protein	Rv2107	PE family protein
Rv2328	PE family protein	Rv2408	PE family protein
Rv2431c	PE family protein	Rv2519	PE family protein
Rv2769c	PE family protein	Rv3018A	PE family protein
Rv3022A	PE family protein	Rv3477	PE family protein
Rv3478	PE family protein	Rv3622c	PE family protein
Rv3650	PE family protein	Rv3746c	PE family protein
Rv3893c	PE family protein	Rv3872	PE family-related protein
Rv0109	PE-PGRS family protein	Rv0124	PE-PGRS family protein
Rv0278c	PE-PGRS family protein	Rv0279c	PE-PGRS family protein
Rv0297	PE-PGRS family protein	Rv0532	PE-PGRS family protein
Rv0578c	PE-PGRS family protein	Rv0742	PE-PGRS family protein
Rv0746	PE-PGRS family protein	Rv0747	PE-PGRS family protein
Rv0754	PE-PGRS family protein	Rv0832	PE-PGRS family protein
Rv0833	PE-PGRS family protein	Rv0834c	PE-PGRS family protein
Rv0872c	PE-PGRS family protein	Rv0977	PE-PGRS family protein
Rv0978c	PE-PGRS family protein	Rv0980c	PE-PGRS family protein

Gene Name	Product	Gene Name	Product
Rv1067c	PE-PGRS family protein	Rv1068c	PE-PGRS family protein
Rv1087	PE-PGRS family protein	Rv1091	PE-PGRS family protein
Rv1243c	PE-PGRS family protein	Rv1325c	PE-PGRS family protein
Rv1396c	PE-PGRS family protein	Rv1441c	PE-PGRS family protein
Rv1450c	PE-PGRS family protein	Rv1452c	PE-PGRS family protein
Rv1468c	PE-PGRS family protein	Rv1651c	PE-PGRS family protein
Rv1768	PE-PGRS family protein	Rv1803c	PE-PGRS family protein
Rv1818c	PE-PGRS family protein	Rv1840c	PE-PGRS family protein
Rv1983	PE-PGRS family protein	Rv2126c	PE-PGRS family protein
Rv2162c	PE-PGRS family protein	Rv2340c	PE-PGRS family protein
Rv2371	PE-PGRS family protein	Rv2396	PE-PGRS family protein
Rv2487c	PE-PGRS family protein	Rv2490c	PE-PGRS family protein
Rv0031	transposase	Rv0797	IS like-2 transposase
Rv1035c	transposase	Rv1036c	transposase
Rv1041c	transposase	Rv1042c	putative transposase
Rv1756c	transposase	Rv1757c	transposase
Rv1763	transposase	Rv1764	transposase
Rv2512c	transposase	Rv2666	IS1533 transposase
Rv2943	transposase	Rv2944	transposase
Rv1054	integrase	Rv2646	integrase
Rv3751	integrase	Rv0151c	PE family protein
Rv0152c	PE family protein	Rv0159c	PE family protein
Rv0160c	PE family protein	Rv0285	PE family protein
Rv0335c	PE family protein	Rv0916c	PE family protein
Rv1040c	PE family protein	Rv1088	PE family protein
Rv1089	PE family protein	Rv1169c	PE family protein
Rv1172c	PE family protein	Rv1195	PE family protein
Rv1214c	PE family protein	Rv1386	PE family protein
Rv1430	PE family protein	Rv1646	PE family protein
Rv1788	PE family protein	Rv1791	PE family protein
Rv1806	PE family protein	Rv2107	PE family protein
Rv2328	PE family protein	Rv2408	PE family protein
Rv2431c	PE family protein	Rv2519	PE family protein
Rv2769c	PE family protein	Rv3018A	PE family protein
Rv3022A	PE family protein	Rv3477	PE family protein
Rv3478	PE family protein	Rv3622c	PE family protein
Rv3650	PE family protein	Rv3746c	PE family protein
Rv3893c	PE family protein	Rv3872	PE family-related protein
Rv0109	PE-PGRS family protein	Rv0124	PE-PGRS family protein
Rv0278c	PE-PGRS family protein	Rv0279c	PE-PGRS family protein
Rv0297	PE-PGRS family protein	Rv0532	PE-PGRS family protein
Rv0578c	PE-PGRS family protein	Rv0742	PE-PGRS family protein
Rv0746	PE-PGRS family protein	Rv0747	PE-PGRS family protein
Rv0754	PE-PGRS family protein	Rv0832	PE-PGRS family protein
Rv0833	PE-PGRS family protein	Rv0834c	PE-PGRS family protein
Rv0872c	PE-PGRS family protein	Rv0977	PE-PGRS family protein
Rv0978c	PE-PGRS family protein	Rv0980c	PE-PGRS family protein
Rv1067c	PE-PGRS family protein	Rv1068c	PE-PGRS family protein
Rv1087	PE-PGRS family protein	Rv1091	PE-PGRS family protein
Rv1243c	PE-PGRS family protein	Rv1325c	PE-PGRS family protein
Rv1396c	PE-PGRS family protein	Rv1441c	PE-PGRS family protein

Gene Name	Product	Gene Name	Product
Rv1450c	PE-PGRS family protein	Rv1452c	PE-PGRS family protein
Rv1468c	PE-PGRS family protein	Rv1651c	PE-PGRS family protein
Rv1768	PE-PGRS family protein	Rv1803c	PE-PGRS family protein
Rv1818c	PE-PGRS family protein	Rv1840c	PE-PGRS family protein
Rv1983	PE-PGRS family protein	Rv2126c	PE-PGRS family protein
Rv2162c	PE-PGRS family protein	Rv2340c	PE-PGRS family protein
Rv2371	PE-PGRS family protein	Rv2396	PE-PGRS family protein
Rv2487c	PE-PGRS family protein	Rv2490c	PE-PGRS family protein
Rv3532	PPE family protein	Rv3533c	PPE family protein
Rv3539	PPE family protein	Rv3558	PPE family protein
Rv3621c	PPE family protein	Rv3738c	PPE family protein
Rv3739c	PPE family protein	Rv3873	PPE family protein
Rv3892c	PPE family protein	Rv0741	transposase
Rv0755A	transposase	Rv0795	transposase IS6110
Rv0796	transposase IS6110	Rv0850	transposase
Rv0920c	transposase	Rv0922	transposase
Rv1034c	transposase	Rv1047	transposase
Rv1149	transposase	Rv1199c	transposase
Rv1313c	transposase	Rv1369c	transposase
Rv1370c	transposase	Rv1765A	transposase
Rv2013	transposase	Rv2014	transposase
Rv2105	transposase	Rv2106	transposase
Rv2167c	transposase	Rv2168c	transposase
Rv2177c	transposase	Rv2278	transposase
Rv2279	transposase	Rv2354	transposase
Rv2355	transposase	Rv2424c	transposase
Rv2479c	transposase	Rv2480c	transposase
Rv2648	transposase IS6110	Rv2649	transposase IS6110
Rv2791c	transposase	Rv2810c	transposase
Rv2812	transposase	Rv2814c	transposase
Rv2815c	transposase	Rv2885c	transposase
Rv2943A	transposase	Rv2961	transposase
Rv2978c	transposase	Rv3023c	transposase
Rv3115	transposase	Rv3184	transposase
Rv3185	transposase	Rv3186	transposase
Rv3187	transposase	Rv3191c	transposase
Rv3325	transposase	Rv3326	transposase
Rv3327	transposase	Rv3348	transposase
Rv3349c	transposase	Rv3380c	transposase
Rv3381c	transposase	Rv3386	transposase
Rv3387	transposase	Rv3427c	transposase
Rv3428c	transposase	Rv3430c	transposase
Rv3474	transposase IS6110	Rv3475	transposase IS6110
Rv3636	transposase	Rv3637	transposase
Rv3638	transposase	Rv3640c	transposase
Rv3798	transposase	Rv3827c	transposase
Rv3844	transposase		

List of Figures

List of Tables