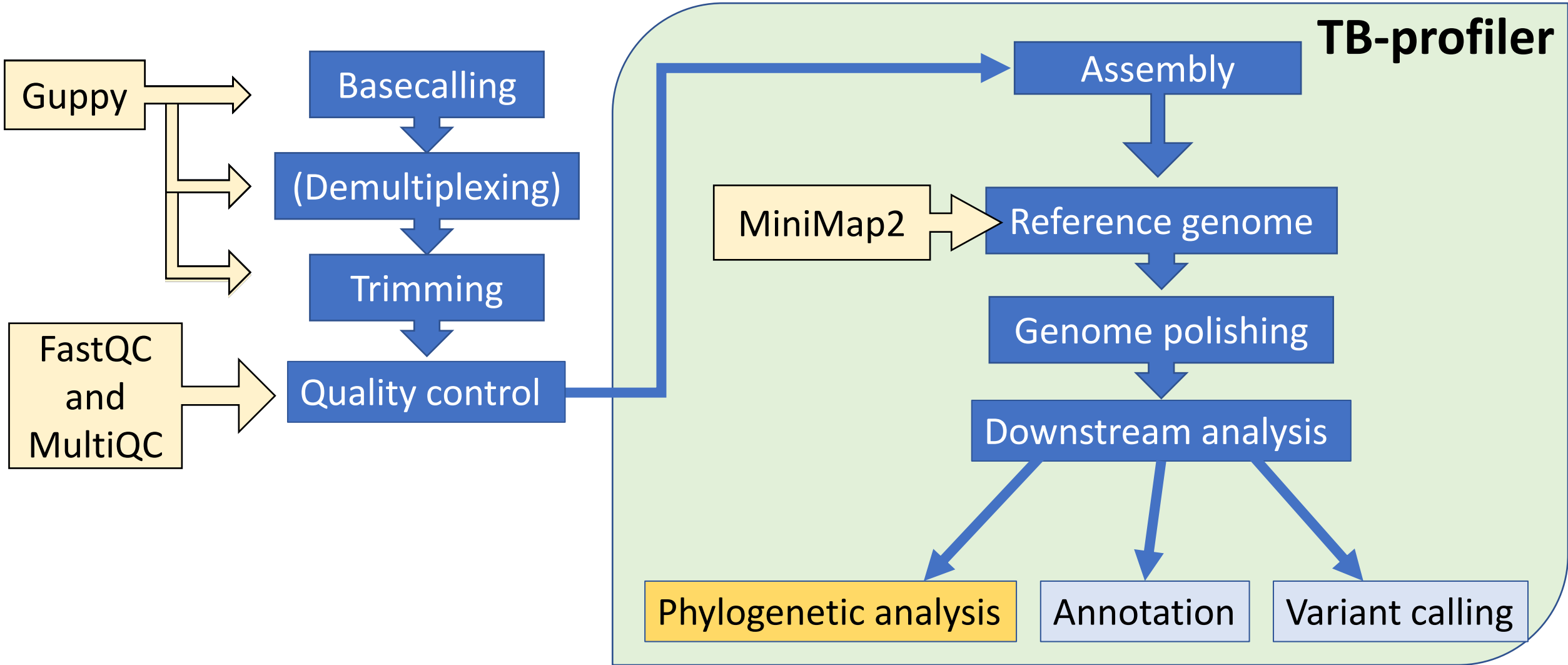
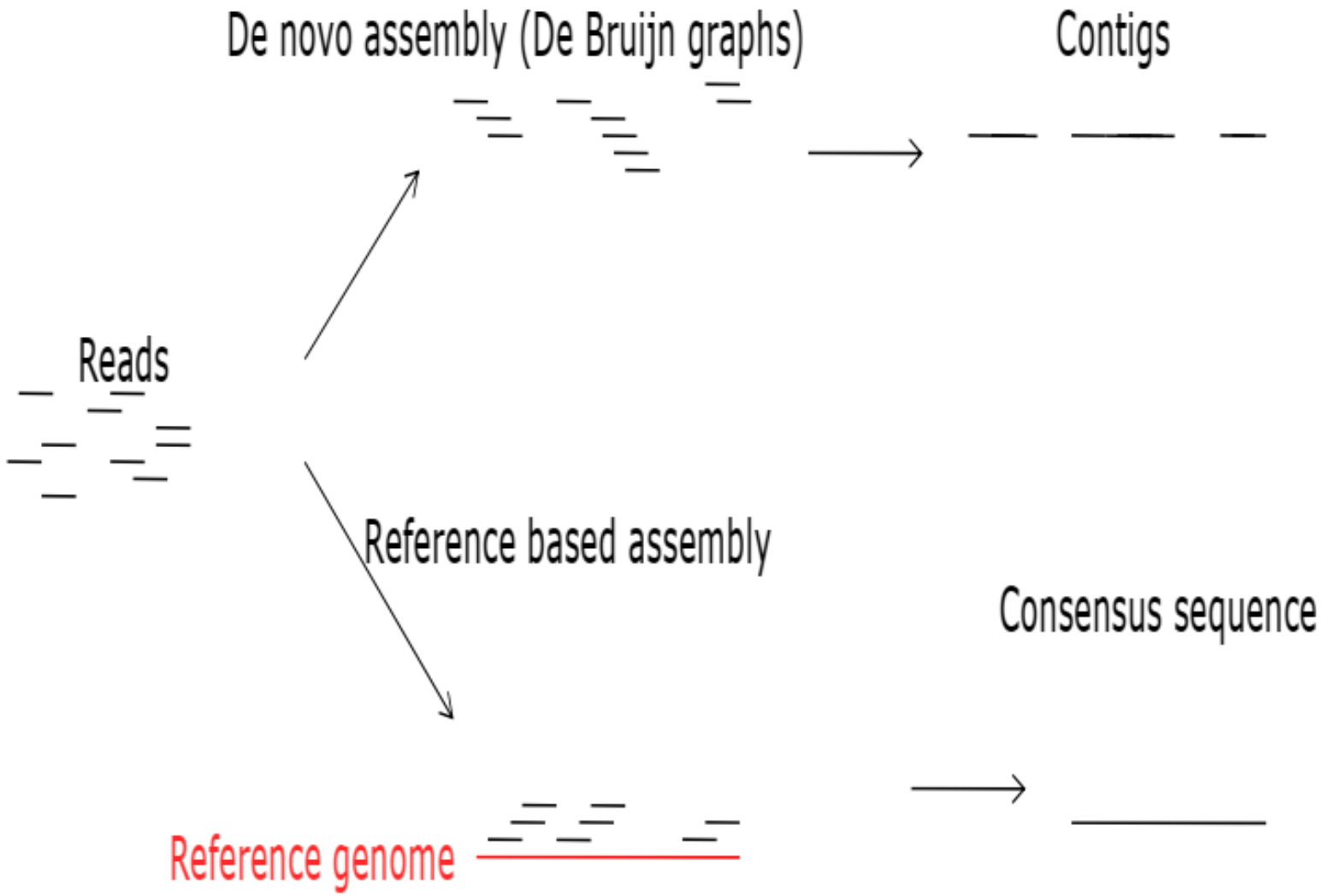
Oxford Nanopore Technologies (ONT) MTB sequencing bioinformatics pipeline 2

Dr Linzy Elton, Prof Neil Stoker, Dr Sylvia Rofael

ONT MTB Bioinformatics Pipeline





Reference Genome:
MtbH37Rv_AL123456.embl

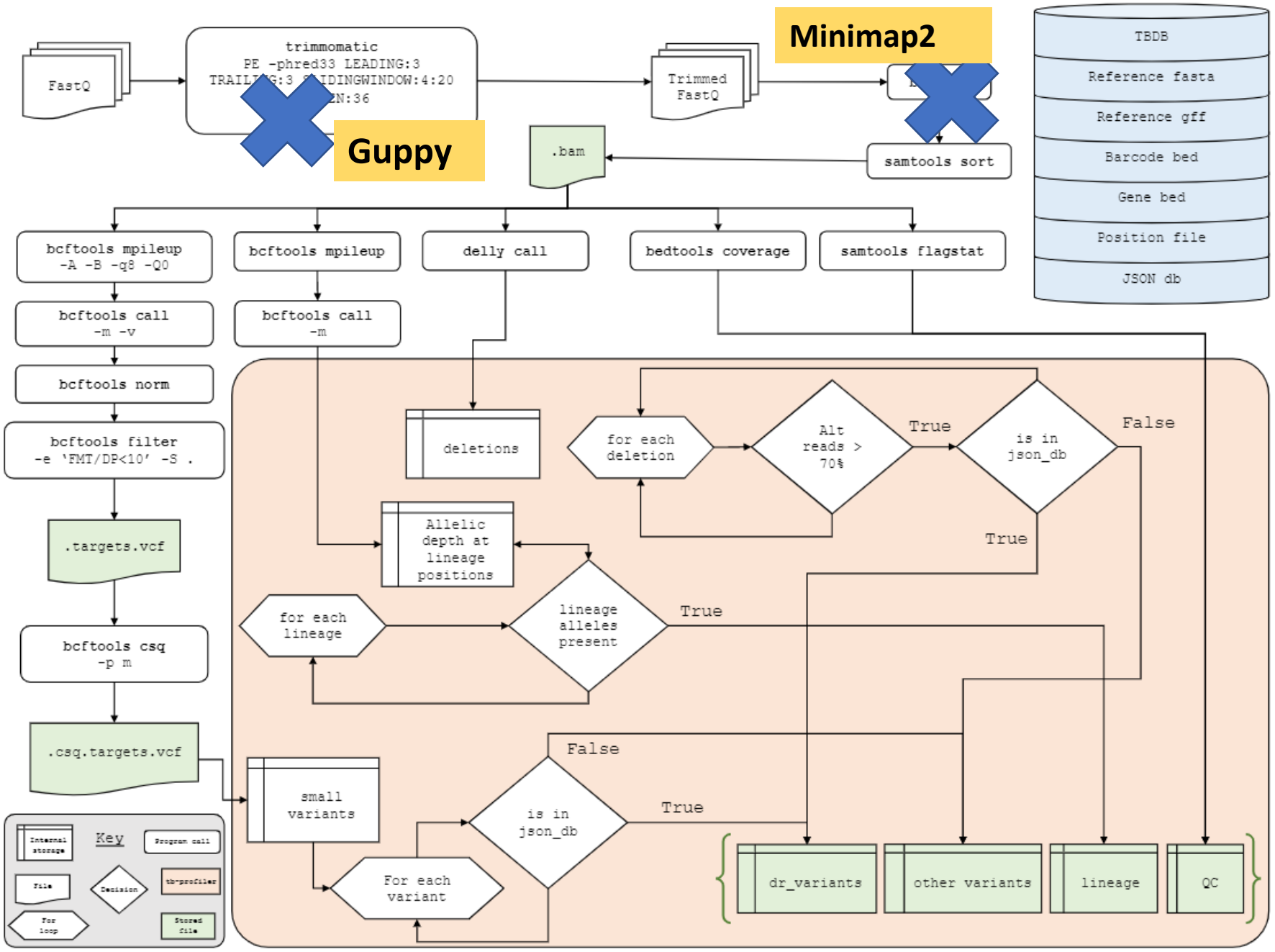
TB-Profiler

- detect drug resistance-conferring mutations and predict drug-resistant phenotypes.
- The pipeline is rabidly developed so make sure you use the latest version (currently v.4.3.0)
- Run

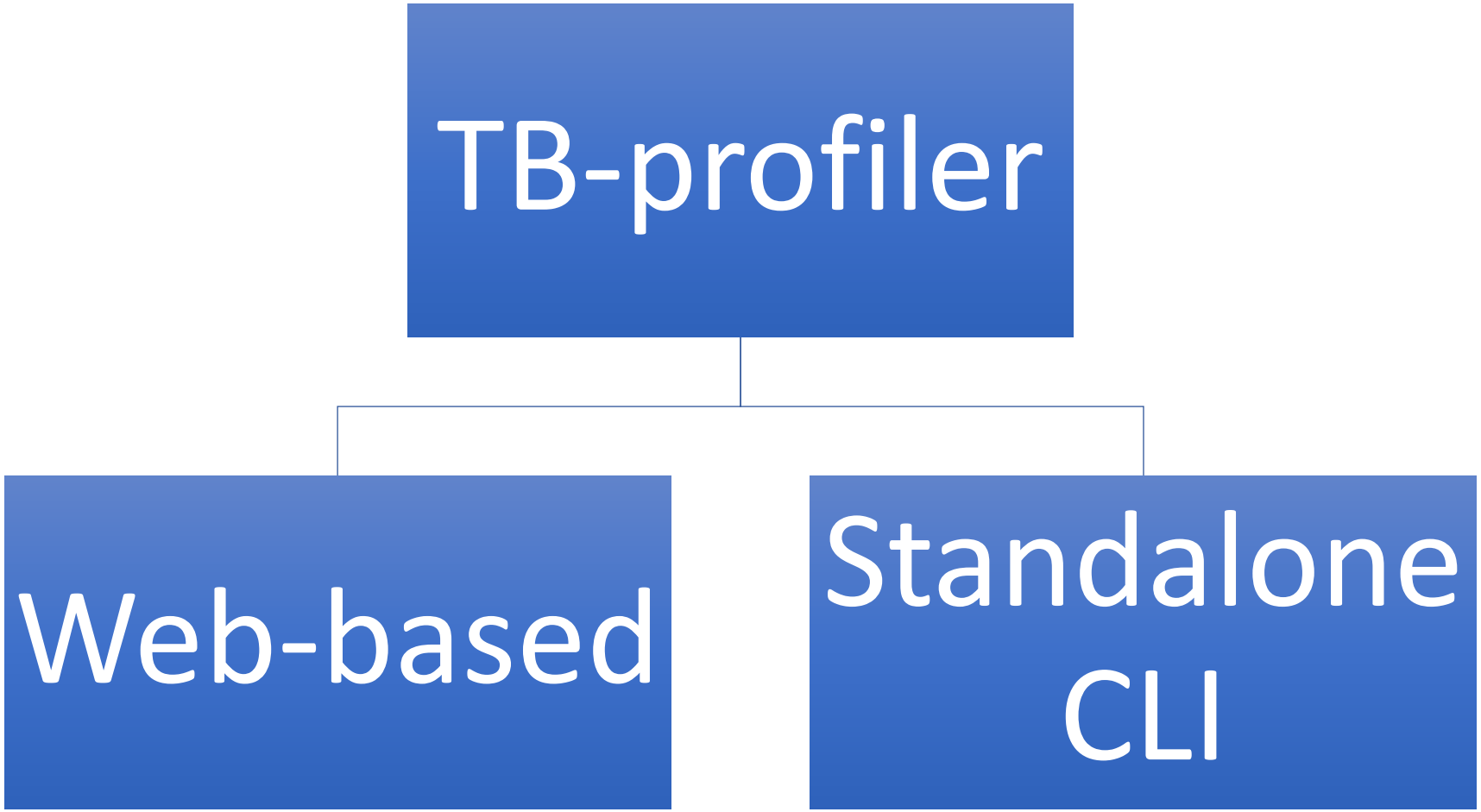
```
conda install -c bioconda tb-profiler
```

- Analysis pipelines is pretty much standardised; the database is manually curated database hosted on separate repository constantly updated, based on emergence of new evidence for the inclusion or exclusion of mutations affect the Accuracy of the AMR prediction.

TB-profiler Pipeline workflow sketch (Illumina)



Important! The mutations and resulting library files are in reference to the H37Rv (NC_000962.3/AL123456.3) reference genome


TB-profiler
Web-based
Standalone CLI

https://jodyphelan.github.io/TBProfiler/

TB Profiler

This repository contains a complete rewrite of the web version of TBProfiler, described here. It allows the use of profiling through a command line interface and contains some additional functionality such as the ability to process minION data.

The pipeline aligns reads to the H37Rv reference using bowtie2, BWA or minimap2 and then calls variants using bcftools. These variants are then compared to a drug-resistance database. We also predict the number of reads supporting drug resistance variants as an insight into hetero-resistance (not applicable for minION data)

Keeping up to date

TBProfiler is under constant rapid development. If you plan to use the program in your work please make sure you are using the most up to date version! Similarly, the database is not static and is continuously being improved so make sure you are using the most latest version. If you use TBProfiler in your work please state the version of both the tool and the database as they are deveoped independantly from each other.

Detailed Manual: https://jodyphelan.gitbook.io/tb-profiler

https://github.com/jodyphelan/tbdb/blob/master/tbdb.csv

master ▾ tbdb / tbdb.csv Go to file ···

jodyphelan remove dups ✕ Latest commit 9f7c75f on 25 Jan History

2 contributors

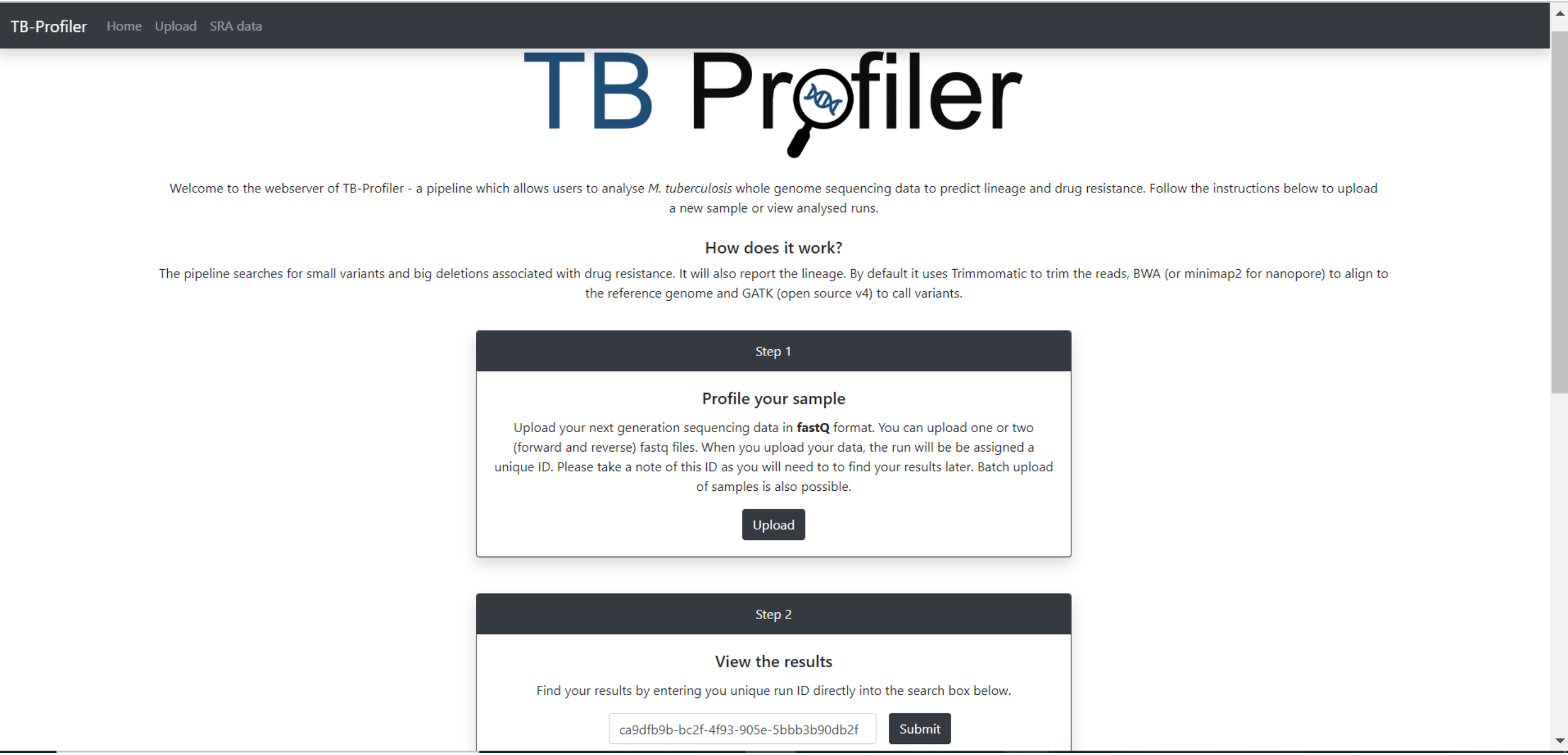
2312 lines (2312 sloc) | 136 KB Raw Blame

Search this file...

Gene	Mutation	Drug	Confers	Interaction	Literature
tlyA	c.100delA	capreomycin	resistance		https://www.
tlyA	c.106_278del	capreomycin	resistance		https://www.
tlyA	c.125dupC	capreomycin	resistance		https://www.
tlyA	c.136dupG	capreomycin	resistance		https://www.
tlyA	c.226delG	capreomycin	resistance		https://www.
tlyA	c.268_269dupGG	capreomycin	resistance		https://www.
tlyA	c.273_274delAT	capreomycin	resistance		https://www.
tlyA	c.311dupG	capreomycin	resistance		https://www.
tlyA	c.315dupC	capreomycin	resistance		https://www.
tlyA	c.328delG	capreomycin	resistance		https://www.

http://varnomen.hgvs.org/recommendations/DNA/variant/substitution/

https://tbdr.lshtm.ac.uk/



Pros

Simple

UFI

Interactive

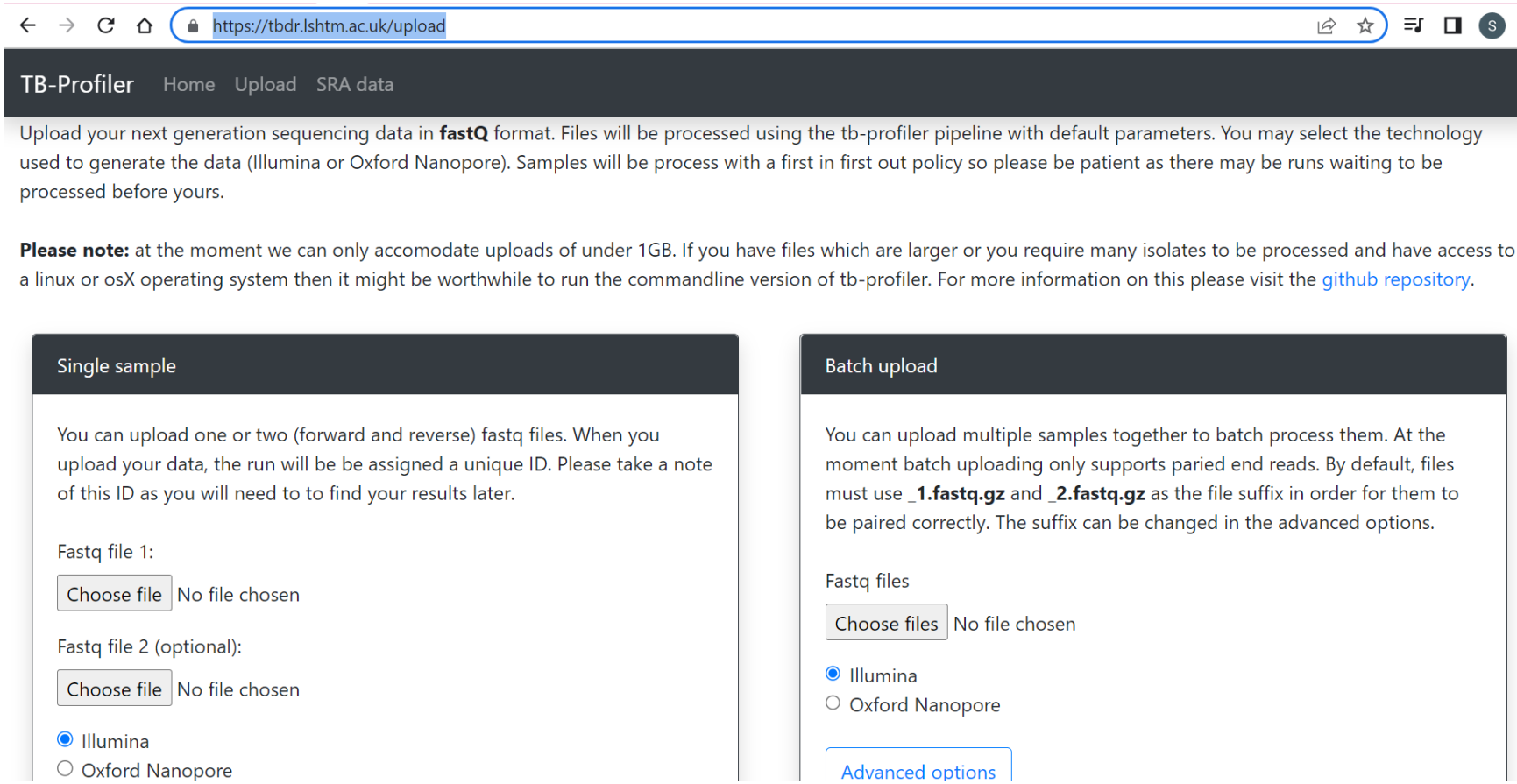
Cons

Limit on the size of the files 1GB

Batch analysis

(ONT not supported)

Stable Internet



Let’s Install TB-profiler Please refer to Script_ONT.txt (line 206)

Meanwhile Lets try barcode10 on the Web-based version

Common File Types

- ✓Fasta (.fna)
- ✓Fastq (.fastq)
- ✓Fast5

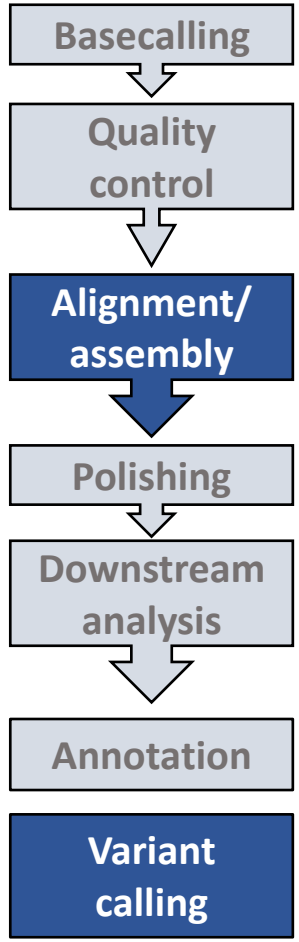
SAM/BAM

SAM

- The Sequence Alignment/Map file (.sam) is a file format to save alignment information of reads mapped against reference genome sequence
- It usually starts with a header section followed by alignment information as tab separated lines for each read

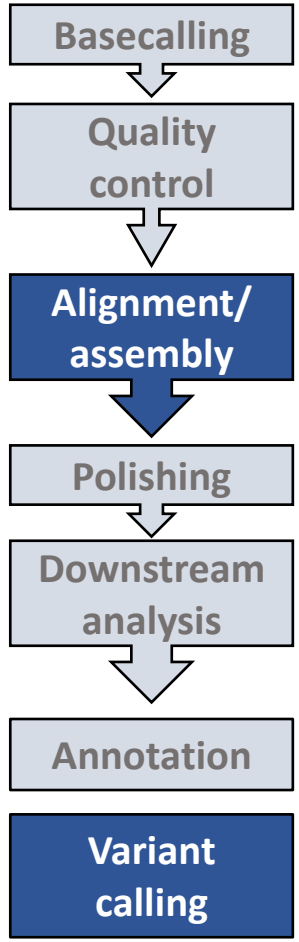
BAM

- The Binary Alignment/Map file (. bam) is the compressed binary version of a SAM file that is used to represent aligned sequences
- BAM file has an index file ".bam.bai"
- Not Human readable but you can open it on Artemis (Genome Viewer by Sanger) line 263 in the Script_ONT.txt



VCF

- Variant calling format (.vcf) (unzip it if it is zipped)
- Used to SNPs (Single Nucleotide Polymorphism) and Insertions and/or deletions (INDELs)
- Can be viewed in genome viewers or in Notepad
- Developed for large scale data storage



```
##bcftools_viewVersion=1.10.2+htslib-1.10.2
##bcftools_viewCommand=view -c1 -a -Oz -o pipeline/callers/tbprofiler_test_run/TB-profiler/4e84cd95-3486-4efc-af21-63affd161753.targets.vcf.gz; Date=Wed
##bcftools_viewCommand=view -v snps pipeline/callers/tbprofiler_test_run/TB-profiler/4e84cd95-3486-4efc-af21-63affd161753.targets.vcf.gz; Date=Wed Sep 2
##bcftools/csqVersion=1.10.2+htslib-1.10.2
##bcftools/csqCommand=csq -p m -f /lustre/projects/CCM_shared_NGS_data/sylvia/.snakemake/conda/2c3a442eb1474c51b4a96ffb80e4b315/share/tbprofiler/tbdb.fa
##INFO=<ID=BCSQ,Number=.,Type=String,Description="Haplotype-aware consequence annotation from BCFtools/csq, see http://samtools.github.io/bcftools/howto
##FORMAT=<ID=BCSQ,Number=.,Type=Integer,Description="Bitmask of indexes to INFO/BCSQ, with interleaved first/second haplotype. Use \"bcftools query -f'[
##bcftools_concatCommand=concat 2c672f8c-6cde-440b-9860-bfa1927f541b.vcf e8ee7b0d-7222-4b88-ad4f-054e2551cfa8.vcf; Date=Wed Sep 29 14:19:50 2021
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  ERR1664619
Chromosome  6140    .   G   T   225 PASS    DP=40;VDB=0.970148;SGB=-0.693145;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,18,22;MQ=60;BCSQ=missense|gyrB|Rv0005|prote
Chromosome  7362    .   G   C   225 PASS    DP=46;VDB=0.682479;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,25,21;MQ=60;BCSQ=missense|gyrA|Rv0006|prote
Chromosome  7582    .   A   C   225 PASS    DP=58;VDB=0.999803;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,27,31;MQ=60;BCSQ=missense|gyrA|Rv0006|prote
Chromosome  7585    .   G   C   225 PASS    DP=58;VDB=0.999706;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,27,31;MQ=60;BCSQ=missense|gyrA|Rv0006|prote
Chromosome  9304    .   G   A   225 PASS    DP=48;VDB=0.261952;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,22,26;MQ=60;BCSQ=missense|gyrA|Rv0006|prote
Chromosome  761155  .   C   T   225 PASS    DP=35;VDB=0.662388;SGB=-0.693136;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,19,16;MQ=60;BCSQ=missense|rpoB|Rv0667|prote
Chromosome  761998  .   T   C   228 PASS    DP=56;VDB=0.159622;SGB=-0.693147;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;AC=2;AN=2;DP4=0,1,32,23;MQ=60;BCSQ=missense
Chromosome  764995  .   C   G   225 PASS    DP=54;VDB=0.998895;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,33,21;MQ=60;BCSQ=synonymous|rpoC|Rv0668|pro
Chromosome  781395  .   T   C   225 PASS    DP=43;VDB=0.7789;SGB=-0.693146;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,19,24;MQ=60  GT:PL:DP:AD 1/1:255,129,0:43:0,4
Chromosome  1473246 .   A   G   225 PASS    DP=47;VDB=0.846133;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,23,24;MQ=60;BCSQ=non_coding|rrs||rRNA  GT:P
Chromosome  1673425 .   C   T   225 PASS    DP=31;VDB=0.442172;SGB=-0.69311;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,21,10;MQ=60 GT:PL:DP:AD 1/1:255,93,0:31:0,31
Chromosome  1674782 .   T   C   225 PASS    DP=39;VDB=0.923491;SGB=-0.693144;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,22,17;MQ=60;BCSQ=missense|inhA|Rv1484|prote
Chromosome  1917972 .   A   G   225 PASS    DP=37;VDB=0.860765;SGB=-0.693141;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,22,15;MQ=60;BCSQ=synonymous|tlyA|Rv1694|pro
Chromosome  2288868 .   A   C   225 PASS    DP=35;VDB=0.788686;SGB=-0.693136;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,17,18;MQ=60;BCSQ=missense|pncA|Rv2043c|prot
Chromosome  3067464 .   A   G   225 PASS    DP=49;VDB=0.35383;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,22,27;MQ=60;BCSQ=missense|thyX|Rv2754c|prote
Chromosome  3067949 .   G   A   225 PASS    DP=42;VDB=0.816864;SGB=-0.693146;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,19,23;MQ=60    GT:PL:DP:AD 1/1:255,126,0:42
Chromosome  3086788 .   T   C   225 PASS    DP=58;VDB=0.516057;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,33,25;MQ=60    GT:PL:DP:AD 1/1:255,175,0:58
Chromosome  3840719 .   T   C   225 PASS    DP=36;VDB=0.377577;SGB=-0.693139;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,24,12;MQ=60;BCSQ=synonymous|alr|Rv3423c|pro
Chromosome  3841433 .   C   G   225 PASS    DP=44;VDB=0.987756;SGB=-0.693146;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,24,20;MQ=60    GT:PL:DP:AD 1/1:255,132,0:44
Chromosome  4242643 .   C   T   225 PASS    DP=47;VDB=0.853588;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,25,22;MQ=60;BCSQ=synonymous|embC|Rv3793|pro
Chromosome  4243217 .   C   T   225 PASS    DP=42;VDB=0.278066;SGB=-0.693146;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,22,20;MQ=60    GT:PL:DP:AD 1/1:255,126,0:42
Chromosome  4247429 .   A   G   225 PASS    DP=51;VDB=0.559817;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,30,21;MQ=60;BCSQ=missense|embB|Rv3795|prote
Chromosome  4247781 .   T   C   225 PASS    DP=48;VDB=0.86667;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,19,29;MQ=60;BCSQ=missense|embB|Rv3795|protei
Chromosome  4328127 .   G   C   225 PASS    DP=45;VDB=0.99981;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,20,25;MQ=60;BCSQ=synonymous|ethR|Rv3855|prot
Chromosome  4407965 .   C   G   225 PASS    DP=33;VDB=0.873953;SGB=-0.693127;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,12,21;MQ=60;BCSQ=missense|gid|Rv3919c|prote
Chromosome  4408156 .   A   C   225 PASS    DP=52;VDB=0.982744;SGB=-0.693147;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,33,19;MQ=60;BCSQ=missense|gid|Rv3919c|prote
```

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.10.2+htslib-1.10.2
##bcftoolsCommand=mpileup -f /lustre/projects/CCM_shared_NGS_data/sylvia/.snakemake/conda/2c3a442eb1474c51b4a96ffb80e4b315/share/tbprofiler/tbdb.fasta -
##reference=file:///lustre/projects/CCM_shared_NGS_data/sylvia/.snakemake/conda/2c3a442eb1474c51b4a96ffb80e4b315/share/tbprofiler/tbdb.fasta
##contig=<ID=Chromosome,length=4411532>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of raw reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of raw reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Number of high-quality bases">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths (high-quality bases)">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">
##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.10.2+htslib-1.10.2
##bcftools_callCommand=call -mv; Date=Wed Sep 29 14:19:35 2021
##bcftools_normVersion=1.10.2+htslib-1.10.2
##bcftools_normCommand=norm -f /lustre/projects/CCM_shared_NGS_data/sylvia/.snakemake/conda/2c3a442eb1474c51b4a96ffb80e4b315/share/tbprofiler/tbdb.fasta
##bcftools_filterVersion=1.10.2+htslib-1.10.2
##bcftools_filterCommand=filter -t Chromosome:4998-7267 -e FMT/DP<10 -S . -Oz -o pipeline/callers/tbprofiler_test_run/TB-profiler/4e84cd95-3486-4efc-af2
##bcftools_concatVersion=1.10.2+htslib-1.10.2
##bcftools_concatCommand=concat -aD pipeline/callers/tbprofiler_test_run/TB-profiler/4e84cd95-3486-4efc-af21-63affd161753_Chromosome_4998_7267.vcf.gz pi
```

Normal text file | length : 15,069 lines : 72 | Ln : 1 Col : 1 Pos : 1 | Unix (LF) | UTF-8 | INS

Drug class definitions

Samples can be classed into different types using the following definitions

Type	Drugs resistance
Sensitive	No drug resistance
Pre-MDR	Rifampicin **or** isonisazid
MDR	Rifampicin **and** isoniazid
Pre-XDR	MDR **and** any fluoriquinolone
XDR	MDR **and** (any fluoriquinolone **and** any group A drug)
Other	Resistance to any drug but none of the above categories

Mutation format

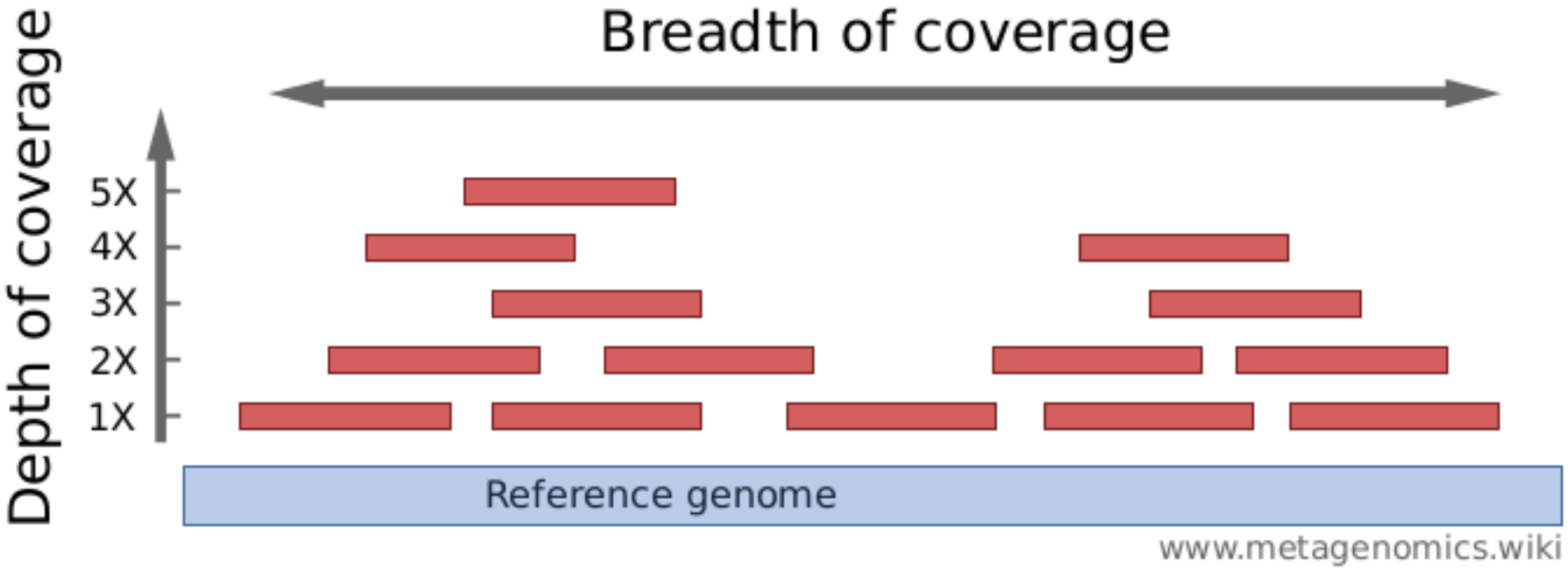
Mutations must follow the HGVS nomenclature. Information on this format can be found here. The following types of mutations are currently allowed:

- Amino acid substitutions. *Example: S450L in rpoB would be p.Ser450Leu*
- Deletions in genes. *Example: Deletion of nucleotide 758 in tlyA would be c.758del*
- Insertion in genes. *Example: Insertion of GT between nucleotide 1850 and 1851 in katG would be c.1850_1851insGT*
- SNPs in non-coding RNAs. *Example: A to G at position 1401 in rrs would be r.1401a>g*
- SNPs in gene promoters. *Example: A to G 7 bases 5' of the start codon in pncA c.-7A>G*

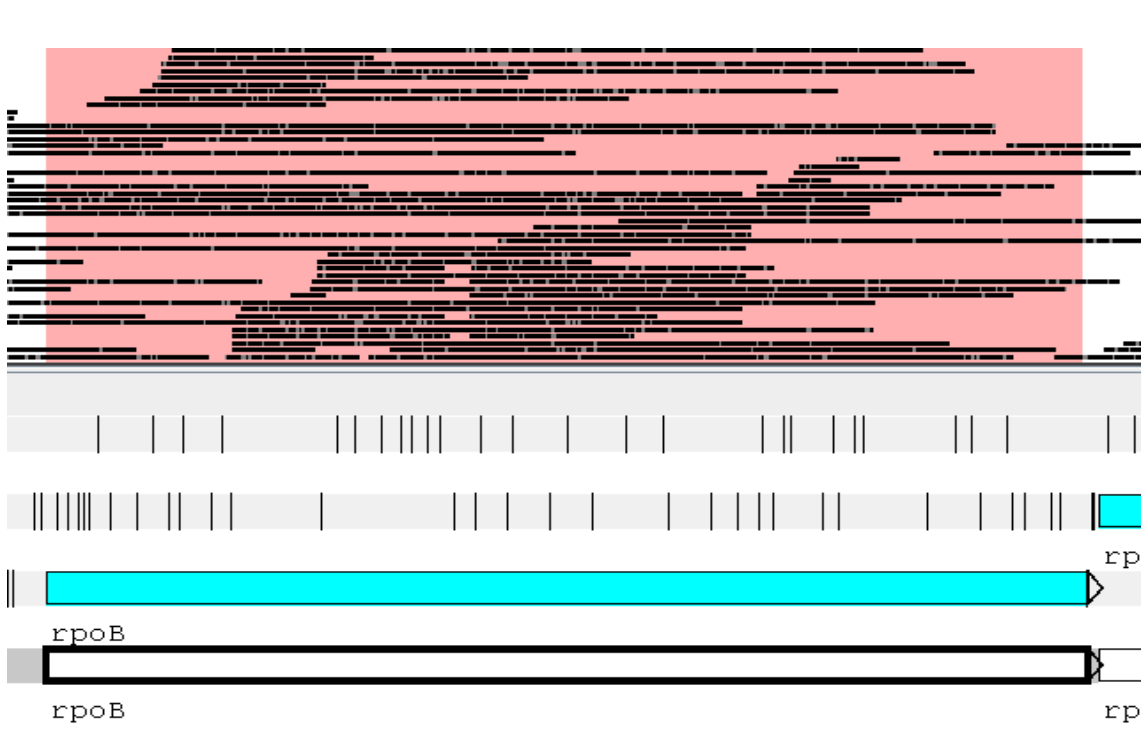
Important! The mutations and resulting library files are in reference to the H37Rv (NC_000962.3/AL123456.3) reference genome

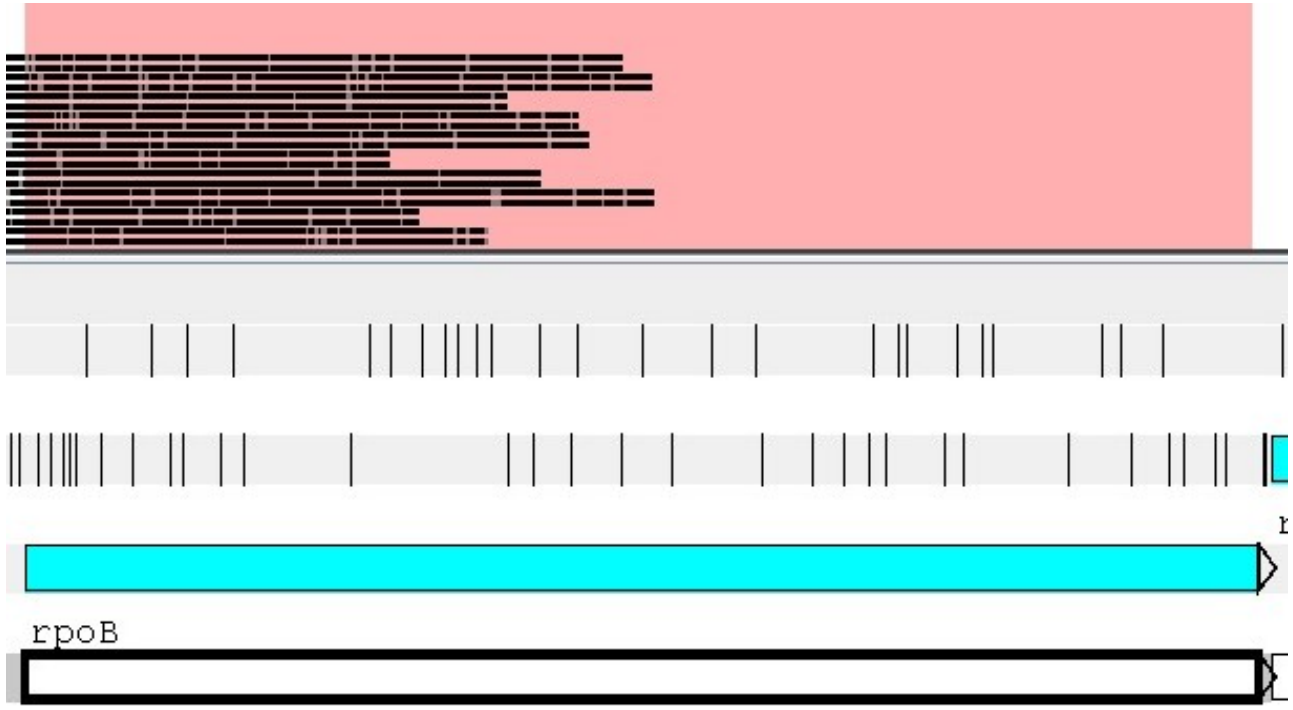
Depth is the amount of times a base within a genome has been sequenced. The greater the depth, the greater the confidence in the identity of the sequenced base

Coverage



rpoB gene Coverage in 2 sequencing runs





Artemis: line 251

Let’s Have a look at a TB-profiler output report

TB-profile Collate Script

Very Useful script to collate all data of multiple samples done in batches from .json files in a useful way in order to run queries on the whole batch.

Exercise

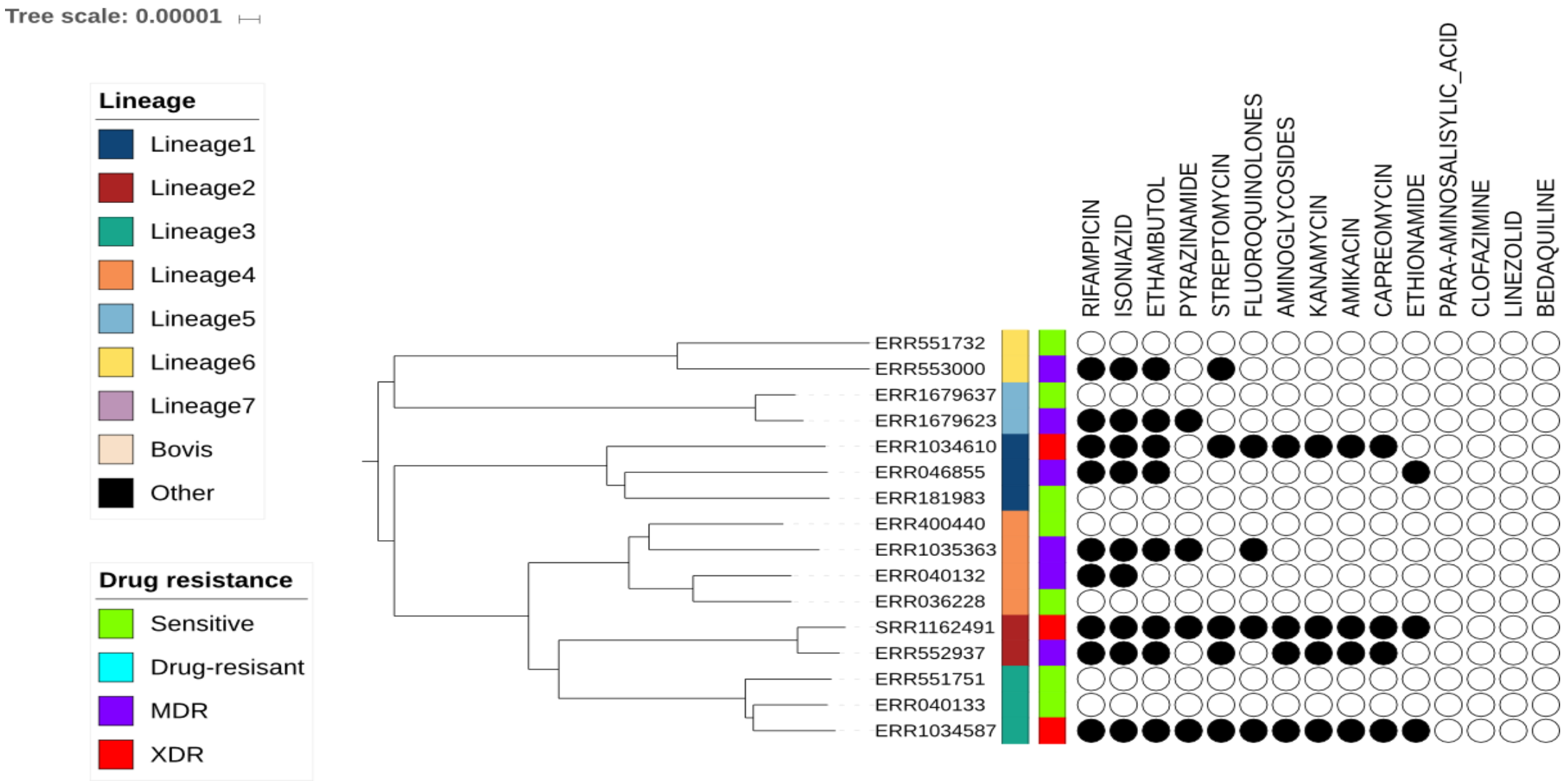
Let's Run the 12 fastq files in the directory fastq_ready through TB-profiler (Please refer to line 225 in Script_ONT.txt)

Then collate the data of the 12 isolates (Please refer to line 259 in Script_ONT.txt)

TB-profile Collate Script

It also produce 3 config files to annotate phylogenetic tree with

- Lineage
- Drug resistance classes (Sensitive, drug-resistant, MDR, XDR)
- Drug resistance calls for individual drugs (filled circles represent resistance)


Tree scale: 0.00001
Lineage
Lineage1
Lineage2
Lineage3
Lineage4
Lineage5
Lineage6
Lineage7
Bovis
Other
Drug resistance
Sensitive
Drug-resisant
MDR
XDR
RIFAMPICIN
ISONIAZID
ETHAMBUTOL
PYRAZINAMIDE
STREPTOMYCIN
FLUOROQUINOLONES
AMINOGLYCOSIDES
KANAMYCIN
AMIKACIN
CAPREOMYCIN
ETHIONAMIDE
PARA-AMINOSALISYLIC_ACID
CLOFAZIMINE
LINEZOLID
BEDAQUILINE
ERR551732
ERR553000
ERR1679637
ERR1679623
ERR1034610
ERR046855
ERR181983
ERR400440
ERR1035363
ERR040132
ERR036228
SRR1162491
ERR552937
ERR551751
ERR040133
ERR1034587

UCL

ASANTE SANA!