DATE: 2nd SEPTEMBER, 2024

NAME: HASSANA ABDULKADIR

EMAIL: abdulkadirhassana97@gmail.com

TITLE: SEARCHING FOR MUTATIONS LEADING TO ISONIAZID RESISTANCE IN TUBERCULOSIS

AIM: To identify and characterise genetic mutations associated with isoniazid resistance in *Mycobacterium tuberculosis*, the causative agent of TB.

INTRODUCTION

Tuberculosis (TB) remains a global health concern and a continuous threat to humanity. *Mycobacterium tuberculosis*, a large, nonmotile, rod-shaped, obligate aerobic bacterium is the pathogen that causes tuberculosis (TB). Commonly introduced into the body by inhaling droplet nuclei, which are present in the lungs' well-aerated upper lobes. In addition to the lungs, *Mycobacterium tuberculosis* can affect the liver, bones, genitourinary tract, central nervous system, lymph nodes, and gastrointestinal tract. Tuberculosis causes severe symptoms, such as a persistent cough, blood in the sputum, and weight loss. The illness claims the lives of almost 1.5 million people each year.

Those suffering from HIV/AIDS and other immune-compromised conditions are especially vulnerable to contracting tuberculosis. Four front-line medications are used to treat tuberculosis: pyrazinamide, ethambutol, rifampicin, and isoniazid.

The development of drug-resistant bacteria makes treatment considerably more difficult. An important first-line medication for tuberculosis treatment is isoniazid, a prodrug that inhibits the formation of the mycobacterial cell wall, and drug resistance to this medication can make treatment less effective.

KatG is a major gene associated with isoniazid resistance. The catalase-peroxidase enzyme, which KatG encodes, is in charge of changing isoniazid into its bactericidal form. Isoniazid resistance would arise from mutations in the KatG gene, thus decreasing the enzyme's efficiency.

The aim of this project is to identify and characterise genetic mutations associated with isoniazid resistance in *Mycobacterium tuberculosis*, the causative agent of Tuberculosis by performing variant calling and also explore associated biological pathways to understand the molecular mechanisms of isoniazid resistance.

Variant calling is a very popular analysis pipeline. Variant calling is a kind of analysis done on whole genome or whole exom sequencing data, basically to determine the variants present in the sequenced reads. Variations are usually caused by mutations.

METHOD

The workflow for this project includes;
Sequenced reads→ Trim reads (Sickle)→ Index Reference Genome → Mapping sequence reads to downloaded referenced genome (bwa-mem)→ SAM files→ BAM files→ Sorted BAM files→ Variant calling(bcftools) → Variant annotation (SnpEff)

The Sequence reads were downloaded and trimmed using **sickle** (as a form of quality control). The trimmed reads were mapped or aligned to a Reference genome using BWA-Burrows-wheeler Aligner and the information was stored in a SAM file. After mapping, the SAM files were converted to BAM files and the BAM files were sorted.
VCF tools such as **mpileup** command was used to generate a coverage information for all the bases and the variance was called up using bcftools and then finally a post vcf analysis and Variant annotation was performed using SnpEff. The Variant calling algorithm applied, basically calls variance, identifies these variants and store them in a file called Variant call format file (VCF). The VCF (Variant call format) file generated stores variants present in the data. SnpEff is used to annotate variants by providing functional information about them.

STEP 1

- The softwares necessary for this project were already installed via conda and homebrew. Dedicated directories were created for the different file types.
- The sample datasets (a total of 19), usually stored in fasta or fastq format were downloaded and the reference genome was downloaded from the NCBI database.

STEP 2

- The sample datasets were trimmed using sickle installed via homebrew. Trimming helps to, improve the quality of the reads.

#Trimm reads

sickle pe -f ERR8774458_1.fastq -r ERR8774458_2.fastq -t sanger -q 20 -l 20 -g -o trimmed_R1.fastq -p trimmed_R2.fastq -s trimmed_S.fastq

sickle pe -f ERR8774464_1.fastq -r ERR8774464_2.fastq -t sanger -q 20 -l 20 -g -o trimmed2_R1.fastq -p trimmed2_R2.fastq -s trimmed2_S.fastq

sickle pe -f ERR8774480_1.fastq -r ERR8774480_2.fastq -t sanger -q 20 -l 20 -g -o trimmed3_R1.fastq -p trimmed3_R2.fastq -s trimmed3_S.fastq

sickle pe -f ERR8774482_1.fastq -r ERR8774482_2.fastq -t sanger -q 20 -l 20 -g -o trimmed4_R1.fastq -p trimmed4_R2.fastq -s trimmed4_S.fastq

sickle se -t sanger -q 20 -l 20 -g -f ERR8774487_1.fastq -o trimmed5_R1.fastq

sickle pe -f ERR8774512_1.fastq -r ERR8774512_2.fastq -t sanger -q 20 -l 20 -g -o trimmed6_R1.fastq -p trimmmed6_R2.fastq -s trimmed6_S.fastq

sickle pe -f ERR8774514_1.fastq -r ERR8774514_2.fastq -t sanger -q 20 -l 20 -g -o trimmed7_R1.fastq -p trimmed7_R2.fastq -s trimmed7_S.fastq

sickle pe -f ERR8774522_1.fastq -r ERR8774522_2.fastq -t sanger -q 20 -l 20 -g -o trimmed8_R1.fastq -p trimmed8_R2.fastq -s trimmed8_S.fastq

sickle pe -f ERR8774524_1.fastq -r ERR8774524_2.fastq -t sanger -q 20 -l 20 -g -o trimmed9_R1.fastq -p trimmed9_R2.fastq -s trimmed9_S.fastq

sickle pe -f ERR8774534_1.fastq -r ERR8774534_2.fastq -t sanger -q 20 -l 20 -g -o trimmed10_R1.fastq -p trimmed10_R2.fastq -s trimmed10_S.fastq

STEP 3

- The sample datasets were aligned to the downloaded reference genome. This is referred to as Genome mapping and is usually done using the BWA tool (Burrows-Wheeler Aligner). The purpose of genome mapping is to identify regions or locations

in the reference genome where the reads map uniquely to or corresponds to. The output generated by bwa-mem were stored in SAM files. The reference genome was first indexed before alignment.

#Index the reference genome

bwa index sequence.fasta

#Perform genome mapping

bwa mem -t 8 ref/sequence.fasta trimmed_R1.fastq trimmed_R2.fastq > output.sam
bwa mem -t 8 ref/sequence.fasta trimmed2_R1.fastq trimmed2_R2.fastq > 2output.sam
bwa mem -t 8 ref/sequence.fasta trimmed3_R1.fastq trimmed3_R2.fastq > 3output.sam
bwa mem -t 8 ref/sequence.fasta trimmed4_R1.fastq trimmed4_R2.fastq > 4output.sam
bwa mem -t 8 ref/sequence.fasta trimmed5_R1.fastq > 5output.sam
bwa mem -t 8 ref/sequence.fasta trimmed6_R1.fastq trimmed6_R2.fastq > 6output.sam
bwa mem -t 8 ref/sequence.fasta trimmed7_R1.fastq trimmed7_R2.fastq > 7output.sam
bwa mem -t 8 ref/sequence.fasta trimmed8_R1.fastq trimmed8_R2.fastq > 8output.sam
bwa mem -t 8 ref/sequence.fasta trimmed9_R1.fastq trimmed9_R2.fastq > 9output.sam
bwa mem -t 8 ref/sequence.fasta trimmed10_R1.fastq trimmed10_R2.fastq > 10output.sam

STEP 4

- The SAM files were converted into BAM files using **samtools** as they are usually large. BAM files are compressed and usually have smaller size.

#converting SAM files to BAM files using samtools

samtools view -S -b output.sam > output.bam
samtools view -S -b 2output.sam > 2output.bam
samtools view -S -b 3output.sam > 3output.bam
samtools view -S -b 4output.sam > 4output.bam
samtools view -S -b 5output.sam > 5output.bam
samtools view -S -b 6output.sam > 6output.bam
samtools view -S -b 7output.sam > 7output.bam
samtools view -S -b 8output.sam > 8output.bam
samtools view -S -b 9output.sam > 9output.bam

samtools view -S -b 10output.sam > 10output.bam

STEP 5

- The bam files were sorted using samtools sort, which was used for the variant calling.

#sorting the bam files using samtools sort

samtools sort output.bam -o../sort/output.sort.bam

samtools sort 2output.bam -o 2output.sort.bam

samtools sort 3output.bam -o 3output.sort.bam

samtools sort 4output.bam -o 4output.sort.bam

samtools sort 5output.bam -o 5output.sort.bam

samtools sort 6output.bam -o 6output.sort.bam

samtools sort 7output.bam -o 7output.sort.bam

samtools sort 8output.bam -o 8output.sort.bam

samtools sort 9output.bam -o 9output.sort.bam

samtools sort 10output.bam -o 10output.sort.bam

STEP 6

- The sorted bam files were indexed

#index sorted bam files

samtools index output.sort.bam

samtools index 2output.sort.bam

samtools index 3output.sort.bam

samtools index 4output.sort.bam

samtools index 5output.sort.bam

samtools index 6output.sort.bam

samtools index 7output.sort.bam

samtools index 8output.sort.bam

samtools index 9output.sort.bam

samtools index 10output.sort.bam

STEP 7

- The statistics was checked using flagstat on one of the reads to check how well the genome mapped;

#To view the statistics or output of the reads

-samtools flagstat output.sort.bam

793440 + 0 in total (QC-passed reads + QC-failed reads)

783766 + 0 primary

0 + 0 secondary

9674 + 0 supplementary

0 + 0 duplicates

0 + 0 primary duplicates

786124 + 0 mapped (99.08% : N/A)

776450 + 0 primary mapped (99.07% : N/A)

783766 + 0 paired in sequencing

391883 + 0 read1

391883 + 0 read2

754788 + 0 properly paired (96.30% : N/A)

774778 + 0 with itself and mate mapped

1672 + 0 singletons (0.21% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)

This means that 99.08% of the reads for sample 1 mapped to the reference genome. Majority of the reads mapped

-samtools flagstat 2output.sort.bam

1426249 + 0 in total (QC-passed reads + QC-failed reads)

1411836 + 0 primary

0 + 0 secondary

14413 + 0 supplementary

0 + 0 duplicates

0 + 0 primary duplicates

1417810 + 0 mapped (99.41% : N/A)
1403397 + 0 primary mapped (99.40% : N/A)
1411836 + 0 paired in sequencing
705918 + 0 read1
705918 + 0 read2
1370854 + 0 properly paired (97.10% : N/A)
1400382 + 0 with itself and mate mapped
3015 + 0 singletons (0.21% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

STEP 8

VARIANT CALLING

VCF tools such as **mpileup** command was used to generate a coverage information for all the bases and the variance was called up using bcftools.

Bcftools is used to do the mpileup and this generates coverage information about the mapped bases. The mpileup sub-command of bcftools is used to count the read coverage at each sequence position.

```
bcftools mpileup -O b -o raw.bcf -f ref/sequence.fasta output.sort.bam
bcftools mpileup -O b -o 2raw.bcf -f ref/sequence.fasta 2output.sort.bam
bcftools mpileup -O b -o 3raw.bcf -f ref/sequence.fasta 3output.sort.bam
bcftools mpileup -O b -o 4raw.bcf -f ref/sequence.fasta 4output.sort.bam
bcftools mpileup -O b -o 5raw.bcf -f ref/sequence.fasta 5output.sort.bam
bcftools mpileup -O b -o 6raw.bcf -f ref/sequence.fasta 6output.sort.bam
bcftools mpileup -O b -o 7raw.bcf -f ref/sequence.fasta 7output.sort.bam
bcftools mpileup -O b -o 8raw.bcf -f ref/sequence.fasta 8output.sort.bam
bcftools mpileup -O b -o 9raw.bcf -f ref/sequence.fasta 9output.sort.bam
bcftools mpileup -O b -o 10raw.bcf -f ref/sequence.fasta 10output.sort.bam

#Call single nucleotide variants
bcftools call --ploidy 1 -m -v -o raw.vcf raw.bcf
bcftools call --ploidy 1 -m -v -o 2raw.vcf 2raw.bcf
```

```
bcftools call --ploidy 1 -m -v -o 3raw.vcf 3raw.bcf
bcftools call --ploidy 1 -m -v -o 4raw.vcf 4raw.bcf
bcftools call --ploidy 1 -m -v -o 5raw.vcf 5raw.bcf
bcftools call --ploidy 1 -m -v -o 6raw.vcf 6raw.bcf
bcftools call --ploidy 1 -m -v -o 7raw.vcf 7raw.bcf
bcftools call --ploidy 1 -m -v -o 8raw.vcf 8raw.bcf
bcftools call --ploidy 1 -m -v -o 9raw.vcf 9raw.bcf
bcftools call --ploidy 1 -m -v -o 10raw.vcf 10raw.bcf
```

#Filter the snvs for the final output in vcf format using vcfutilis

```
vcfutils.pl varFilter raw.vcf >  raw.filter.vcf
vcfutils.pl varFilter 2raw.vcf >  2raw.filter.vcf
vcfutils.pl varFilter 3raw.vcf >  3raw.filter.vcf
vcfutils.pl varFilter 4raw.vcf >  4raw.filter.vcf
vcfutils.pl varFilter 5raw.vcf >  5raw.filter.vcf
vcfutils.pl varFilter 6raw.vcf >  6raw.filter.vcf
vcfutils.pl varFilter 7raw.vcf >  7raw.filter.vcf
vcfutils.pl varFilter 8raw.vcf >  8raw.filter.vcf
vcfutils.pl varFilter 9raw.vcf >  9raw.filter.vcf
vcfutils.pl varFilter 10raw.vcf >  10raw.filter.vcf
```

#To check the number of snps and indels (variant types); the number of variants that are snps and indels

```
grep -v -c '^#' raw.vcf
1047
bcftools view -v snps raw.vcf | grep -v -c '^#'
956
bcftools view -v indels raw.vcf | grep -v -c '^#'
91

grep -v -c '^#' 2raw.vcf
```

1025
bcftools view -v snps 2raw.vcf | grep -v -c '^#'
940
bcftools view -v indels 2raw.vcf | grep -v -c '^#'
85

grep -v -c '^#' 3raw.vcf
1067
bcftools view -v snps 3raw.vcf | grep -v -c '^#'
983
bcftools view -v indels 3raw.vcf | grep -v -c '^#'
84

grep -v -c '^#' 4raw.vcf
1049
bcftools view -v snps 4raw.vcf | grep -v -c '^#'
964
bcftools view -v indels 4raw.vcf | grep -v -c '^#'
85

grep -v -c '^#' 5raw.vcf
951
bcftools view -v snps 5raw.vcf | grep -v -c '^#'
870
bcftools view -v indels 5raw.vcf | grep -v -c '^#'
81

grep -v -c '^#' 6raw.vcf
1027
bcftools view -v snps 6raw.vcf | grep -v -c '^#'
937
bcftools view -v indels 6raw.vcf | grep -v -c '^#'

90

grep -v -c '^#' 7raw.vcf
1031
bcftools view -v snps 7raw.vcf | grep -v -c '^#'
941
bcftools view -v indels 7raw.vcf | grep -v -c '^#'
90

grep -v -c '^#' 8raw.vcf
1003
bcftools view -v snps 8raw.vcf | grep -v -c '^#'
918
bcftools view -v indels 8raw.vcf | grep -v -c '^#'
85

grep -v -c '^#' 9raw.vcf
1027
bcftools view -v snps 9raw.vcf | grep -v -c '^#'
938
bcftools view -v indels 9raw.vcf | grep -v -c '^#'
89

grep -v -c '^#' 10raw.vcf
1019
bcftools view -v snps 10raw.vcf | grep -v -c '^#'
932
bcftools view -v indels 10raw.vcf | grep -v -c '^#'
87

#final filtered datasets
grep -v "##" raw.vcf
grep -v "##" 2raw.vcf
grep -v "##" 3raw.vcf

```
grep -v "##" 4raw.vcf
grep -v "##" 5raw.vcf
grep -v "##" 6raw.vcf
grep -v "##" 7raw.vcf
grep -v "##" 8raw.vcf
grep -v "##" 9raw.vcf
grep -v "##" 10raw.vcf


#To view results
samtools tview output.sort.bam ref/sequence.fasta
samtools tview 2output.sort.bam ref/sequence.fasta
samtools tview 3output.sort.bam ref/sequence.fasta
samtools tview 4output.sort.bam ref/sequence.fasta
samtools tview 5output.sort.bam ref/sequence.fasta
samtools tview 6output.sort.bam ref/sequence.fasta
samtools tview 7output.sort.bam ref/sequence.fasta
samtools tview 8output.sort.bam ref/sequence.fasta
samtools tview 9output.sort.bam ref/sequence.fasta
samtools tview 10output.sort.bam ref/sequence.fasta
```

STEP 9

VARIANT ANNOTATION

Variant annotation provides more information to the variants/mutations called from the sequence data. Synonymous or nonsynonymous substitution, start-gain codon, start-loss codon, stop-gain codon, stop-loss codon, or frameshifts are some of the variant effects that are predicted using SnpEff.

```
#build database

scripts/buildDbNcbi.sh NC_000962.3 Downloading genome NC_000962.3

#download database
java -jar snpEff.jar download -v NC_000962.3
```

#using snpEff to perform variant annotation

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 raw.filter.vcf > mmySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 2raw.filter.vcf > mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 3raw.filter.vcf > 3mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 4raw.filter.vcf > 4mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 5raw.filter.vcf > 5mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 6raw.filter.vcf > 6mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 7raw.filter.vcf > 7mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 8raw.filter.vcf > 8mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 9raw.filter.vcf > 9mySNPanot.vcf

java -Xmx8g -jar snpEff/snpEff.jar NC_000962.3 10raw.filter.vcf > 10mySNPanot.vcf

RESULTS

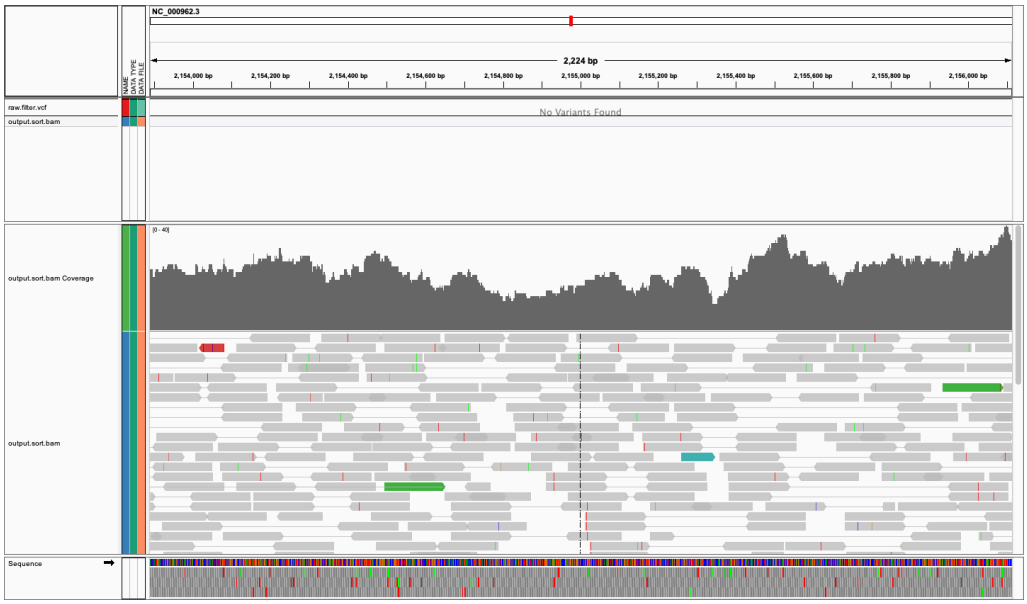
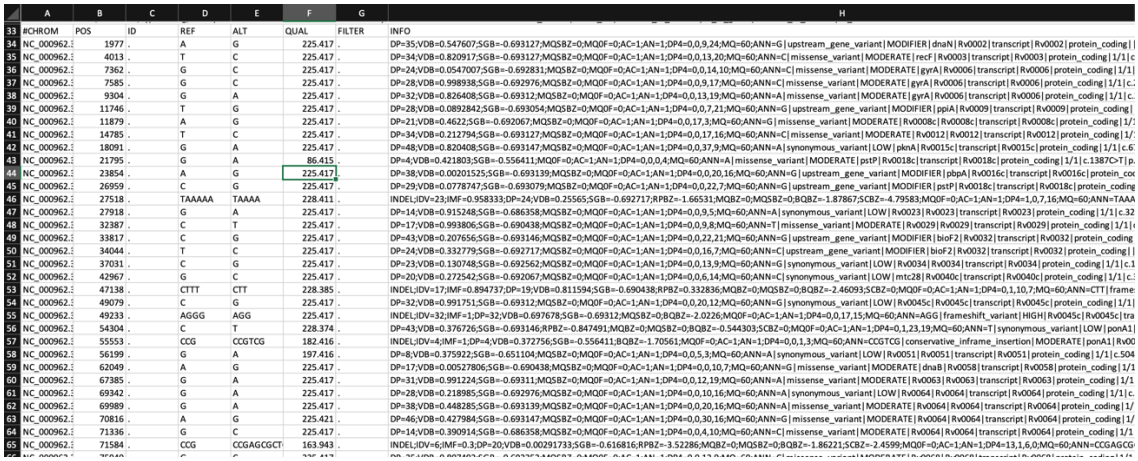


Fig 1.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 1.



	A	B	C	D	E	F	G	H
33	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
34	NC_000962.3	1977	.	A	G	225.417	.	DP=35;VDB=0.547607;SGB=-0.693127;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,9,24;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|dnaN|Rv0002|transcript|Rv0002|protein_coding||
35	NC_000962.3	4013	.	T	C	225.417	.	DP=34;VDB=0.820917;SGB=-0.693127;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,13,20;MQ=60;ANN=C|missense_variant|MODERATE|recF|Rv0003|transcript|Rv0003|protein_coding|1/1|c
36	NC_000962.3	7362	.	G	C	225.417	.	DP=24;VDB=0.0547007;SGB=-0.692831;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,14,10;MQ=60;ANN=C|missense_variant|MODERATE|gyrA|Rv0006|transcript|Rv0006|protein_coding|1/1|
37	NC_000962.3	7585	.	G	C	225.417	.	DP=28;VDB=0.998938;SGB=-0.692976;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,9,17;MQ=60;ANN=C|missense_variant|MODERATE|gyrA|Rv0006|transcript|Rv0006|protein_coding|1/1|c.:
38	NC_000962.3	9304	.	G	A	225.417	.	DP=32;VDB=0.826408;SGB=-0.69312;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,13,19;MQ=60;ANN=A|missense_variant|MODERATE|gyrA|Rv0006|transcript|Rv0006|protein_coding|1/1|c.
39	NC_000962.3	11746	.	T	G	225.417	.	DP=28;VDB=0.0892842;SGB=-0.693054;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,21;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|ppiA|Rv0009|transcript|Rv0009|protein_coding|
40	NC_000962.3	11879	.	A	G	225.417	.	DP=21;VDB=0.4622;SGB=-0.692067;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,17,3;MQ=60;ANN=G|missense_variant|MODERATE|Rv0008c|Rv0008c|transcript|Rv0008c|protein_coding|1/1
41	NC_000962.3	14785	.	T	C	225.417	.	DP=34;VDB=0.212794;SGB=-0.693127;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,17,16;MQ=60;ANN=C|missense_variant|MODERATE|Rv0012|Rv0012|transcript|Rv0012|protein_coding|1/
42	NC_000962.3	18091	.	G	A	225.417	.	DP=48;VDB=0.820408;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,37,9;MQ=60;ANN=A|synonymous_variant|LOW|pknA|Rv0015c|transcript|Rv0015c|protein_coding|1/1|c.6
43	NC_000962.3	21795	.	G	A	86.415	.	DP=4;VDB=0.421803;SGB=-0.556411;MQ0F=0;AC=1;AN=1;DP4=0,0,0,4;MQ=60;ANN=A|missense_variant|MODERATE|pstP|Rv0018c|transcript|Rv0018c|protein_coding|1/1|c.1387C>T|p.
44	NC_000962.3	23854	.	A	G	225.417	.	DP=38;VDB=0.00201525;SGB=-0.693139;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,20,16;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|pbpA|Rv0016c|transcript|Rv0016c|protein_coc
45	NC_000962.3	26959	.	C	G	225.417	.	DP=29;VDB=0.0778747;SGB=-0.693079;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,7;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|pstP|Rv0018c|transcript|Rv0018c|protein_coding
46	NC_000962.3	27518	.	TAAAAA	TAAAA	228.411	.	INDEL;IDV=23;IMF=0.958333;DP=24;VDB=0.25565;SGB=-0.692717;RPBZ=-1.66531;MQBZ=0;MQSBZ=0;BQBZ=-1.87867;SCBZ=-4.79583;MQ0F=0;AC=1;AN=1;DP4=1,0,7,16;MQ=60;ANN=TAAA
47	NC_000962.3	27918	.	G	A	225.417	.	DP=14;VDB=0.915248;SGB=-0.686358;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,9,5;MQ=60;ANN=A|synonymous_variant|LOW|Rv0023|Rv0023|transcript|Rv0023|protein_coding|1/1|c.32
48	NC_000962.3	32387	.	C	T	225.417	.	DP=17;VDB=0.993806;SGB=-0.690438;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,9,8;MQ=60;ANN=T|missense_variant|MODERATE|Rv0029|Rv0029|transcript|Rv0029|protein_coding|1/1|
49	NC_000962.3	33817	.	C	G	225.417	.	DP=43;VDB=0.207656;SGB=-0.693146;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,21;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|bioF2|Rv0032|transcript|Rv0032|protein_coding
50	NC_000962.3	34044	.	T	C	225.417	.	DP=24;VDB=0.332779;SGB=-0.692717;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,16,7;MQ=60;ANN=C|upstream_gene_variant|MODIFIER|bioF2|Rv0032|transcript|Rv0032|protein_coding||
51	NC_000962.3	37031	.	C	G	225.417	.	DP=23;VDB=0.130748;SGB=-0.692562;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,13,9;MQ=60;ANN=G|synonymous_variant|LOW|Rv0034|Rv0034|transcript|Rv0034|protein_coding|1/1|c.1
52	NC_000962.3	42967	.	G	C	225.417	.	DP=20;VDB=0.272542;SGB=-0.692067;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,6,14;MQ=60;ANN=C|synonymous_variant|LOW|mtc28|Rv0040c|transcript|Rv0040c|protein_coding|1/1|c.:
53	NC_000962.3	47138	.	CTTT	CTT	228.385	.	INDEL;IDV=17;IMF=0.894737;DP=19;VDB=0.811594;SGB=-0.690438;RPBZ=0.332836;MQBZ=0;MQSBZ=0;BQBZ=-2.46093;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,1,10,7;MQ=60;ANN=CTT|frame
54	NC_000962.3	49079	.	C	G	225.417	.	DP=32;VDB=0.991751;SGB=-0.69312;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,20,12;MQ=60;ANN=G|synonymous_variant|LOW|Rv0045c|Rv0045c|transcript|Rv0045c|protein_coding|1/1|
55	NC_000962.3	49233	.	AGGG	AGG	225.417	.	INDEL;IDV=32;IMF=1;DP=32;VDB=0.697678;SGB=-0.69312;MQSBZ=0;BQBZ=-2.0226;MQ0F=0;AC=1;AN=1;DP4=0,0,17,15;MQ=60;ANN=AGG|frameshift_variant|HIGH|Rv0045c|Rv0045c|tra
56	NC_000962.3	54304	.	C	T	228.374	.	DP=43;VDB=0.376726;SGB=-0.693146;RPBZ=-0.847491;MQBZ=0;MQSBZ=0;BQBZ=-0.544303;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,1,23,19;MQ=60;ANN=T|synonymous_variant|LOW|ponA1|
57	NC_000962.3	55553	.	CCG	CCGTCG	182.416	.	INDEL;IDV=4;IMF=1;DP=4;VDB=0.372756;SGB=-0.556411;BQBZ=-1.70561;MQ0F=0;AC=1;AN=1;DP4=0,0,1,3;MQ=60;ANN=CCGTCG|conservative_inframe_insertion|MODERATE|ponA1|Rv00
58	NC_000962.3	56199	.	G	A	197.416	.	DP=8;VDB=0.375922;SGB=-0.651104;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,5,3;MQ=60;ANN=A|synonymous_variant|LOW|Rv0051|Rv0051|transcript|Rv0051|protein_coding|1/1|c.504
59	NC_000962.3	62049	.	A	G	225.417	.	DP=17;VDB=0.0052780 6;SGB=-0.690438;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,10,7;MQ=60;ANN=G|missense_variant|MODERATE|dnaB|Rv0058|transcript|Rv0058|protein_coding|1/1
60	NC_000962.3	67385	.	G	A	225.417	.	DP=31;VDB=0.991224;SGB=-0.69311;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,12,19;MQ=60;ANN=A|missense_variant|MODERATE|Rv0063|Rv0063|transcript|Rv0063|protein_coding|1/1
61	NC_000962.3	69342	.	G	A	225.417	.	DP=28;VDB=0.218985;SGB=-0.692976;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,10,16;MQ=60;ANN=A|synonymous_variant|LOW|Rv0064|Rv0064|transcript|Rv0064|protein_coding|1/1|c.
62	NC_000962.3	69989	.	G	A	225.417	.	DP=38;VDB=0.448285;SGB=-0.693139;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,20,16;MQ=60;ANN=A|missense_variant|MODERATE|Rv0064|Rv0064|transcript|Rv0064|protein_coding|1/
63	NC_000962.3	70816	.	A	G	225.421	.	DP=46;VDB=0.427984;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,30,16;MQ=60;ANN=G|missense_variant|MODERATE|Rv0064|Rv0064|transcript|Rv0064|protein_coding|1/
64	NC_000962.3	71336	.	G	C	225.417	.	DP=14;VDB=0.390914;SGB=-0.686358;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,4,10;MQ=60;ANN=C|missense_variant|MODERATE|Rv0064|Rv0064|transcript|Rv0064|protein_coding|1/1
65	NC_000962.3	71584	.	CCG	CCGAGCGCT	163.943	.	INDEL;IDV=6;IMF=0.3;DP=20;VDB=0.00291733;SGB=-0.616816;RPBZ=-3.52286;MQBZ=0;MQSBZ=0;BQBZ=-1.86221;SCBZ=-2.4599;MQ0F=0;AC=1;AN=1;DP4=13,1,6,0;MQ=60;ANN=CCGAGCG

Fig 1.1 A VCF annotated with SnpEff.

Summary

Genome	NC_000962.3
Date	2024-08-08 17:48
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 raw.filter.vcf
Warnings	3,102
Errors	0
Number of lines (input file)	999
Number of variants (before filter)	999
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	999
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,617
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,415 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	999	4,415
Total	**4,411,532**	**999**	**4,415**

Fig 1.2 SnpEff annotation summary.

Number variants by type

Type	Total
SNP	910
MNP	0
INS	39
DEL	50
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	999

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	46	0.478%
LOW	320	3.327%
MODERATE	468	4.866%
MODIFIER	8,783	91.328%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	450	57.841%
NONSENSE	8	1.028%
SILENT	320	41.131%

Missense / Silent ratio: 1.4062

Fig 1.3. Variant effect summary.

Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	8	0.083%
conservative_inframe_insertion	7	0.073%
disruptive_inframe_deletion	2	0.021%
disruptive_inframe_insertion	4	0.042%
downstream_gene_variant	4,294	44.641%
frameshift_variant	35	0.364%
intergenic_region	127	1.32%
intragenic_variant	41	0.426%
missense_variant	447	4.647%
splice_region_variant	1	0.01%
start_lost	3	0.031%
stop_gained	8	0.083%
stop_lost	1	0.01%
synonymous_variant	320	3.327%
upstream_gene_variant	4,321	44.922%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,294	44.65%
EXON	834	8.672%
INTERGENIC	127	1.321%
TRANSCRIPT	41	0.426%
UPSTREAM	4,321	44.931%

Fig 1.4 The Number of effects by type and region

Fig 2.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 2

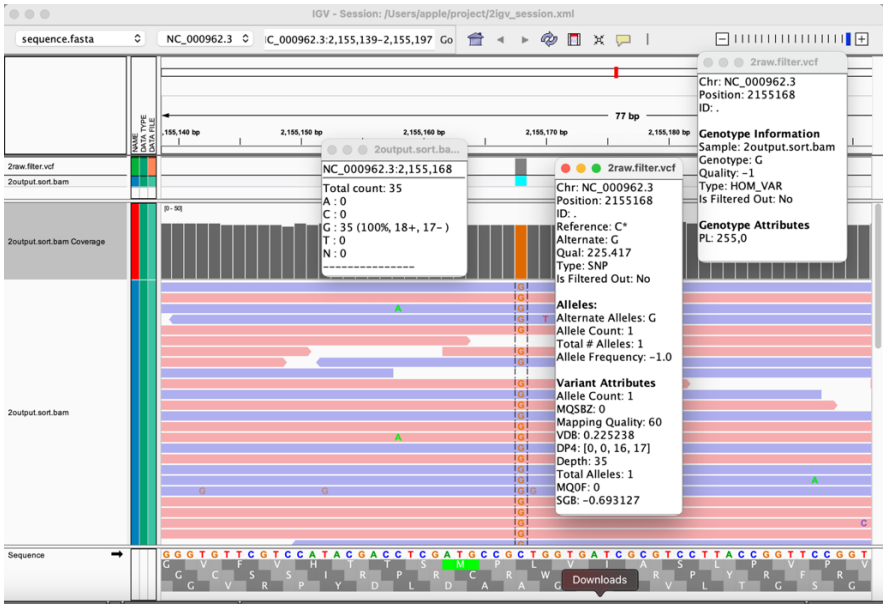


Fig 2.0.1

Fig 2.1 A VCF annotated with SnpEff.

Summary

Genome	NC_000962.3
Date	2024-08-08 16:13
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 2raw.filter.vcf
Warnings	2,992
Errors	0
Number of lines (input file)	984
Number of variants (before filter)	984
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	984
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,345
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,483 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	984	4,483
Total	**4,411,532**	**984**	**4,483**

Fig 2.2 SnpEff annotation summary.

Number variants by type

Type	Total
SNP	901
MNP	0
INS	35
DEL	48
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	984

Number of effects by impact

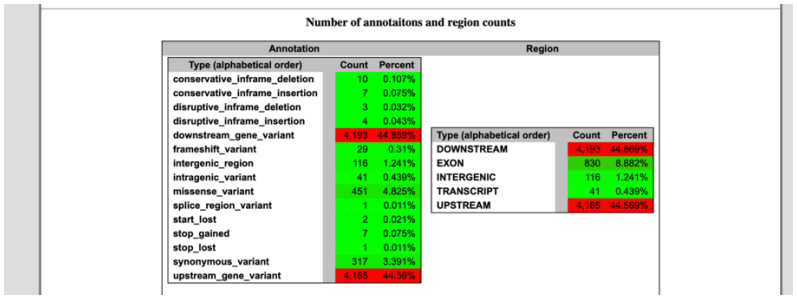
Type (alphabetical order)	Count	Percent
HIGH	38	0.407%
LOW	317	3.392%
MODERATE	475	5.083%
MODIFIER	8,515	91.118%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	453	58.901%
NONSENSE	7	0.901%
SILENT	317	40.798%

Missense / Silent ratio: 1.429

Fig 2.3 . Variant effect summary



Number of annotaitons and region counts

Annotation

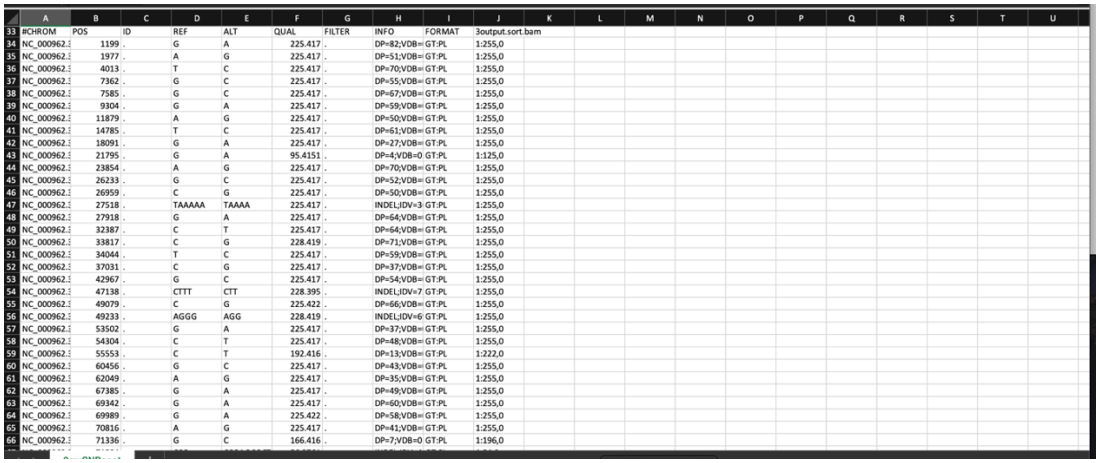
Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	10	0.107%
conservative_inframe_insertion	7	0.075%
disruptive_inframe_deletion	3	0.032%
disruptive_inframe_insertion	4	0.043%
downstream_gene_variant	4,193	44.869%
frameshift_variant	29	0.31%
intergenic_region	116	1.241%
intragenic_variant	41	0.439%
missense_variant	451	4.825%
splice_region_variant	1	0.011%
start_lost	2	0.021%
stop_gained	7	0.075%
stop_lost	1	0.011%
synonymous_variant	317	3.391%
upstream_gene_variant	4,165	44.56%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,193	44.869%
EXON	830	8.882%
INTERGENIC	116	1.241%
TRANSCRIPT	41	0.439%
UPSTREAM	4,165	44.569%

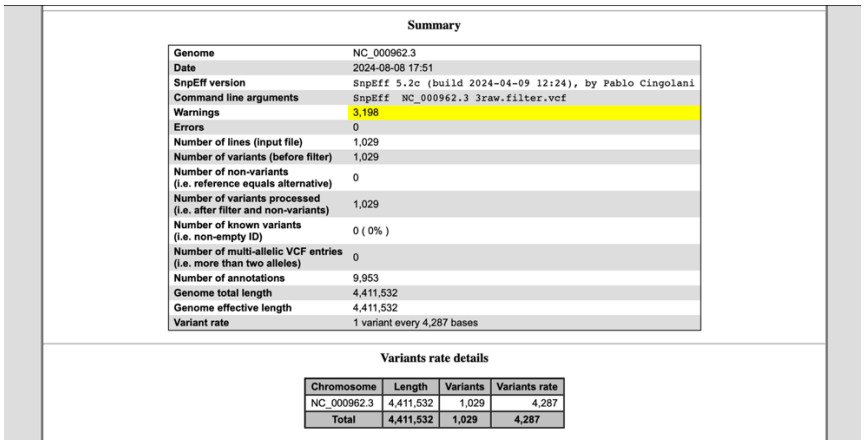
Fig 2.4 The Number of effects by type and region

Fig 3.0 IGV visualisation of a Mycobacterium tuberculosis H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 3



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	3output.sort.bam
NC_000962.3	1199	.	G	A	225.417	.	DP=82;VDB=0	GT:PL	1:255,0
NC_000962.3	1977	.	A	G	225.417	.	DP=51;VDB=0	GT:PL	1:255,0
NC_000962.3	4013	.	T	C	225.417	.	DP=70;VDB=0	GT:PL	1:255,0
NC_000962.3	7362	.	G	C	225.417	.	DP=55;VDB=0	GT:PL	1:255,0
NC_000962.3	7585	.	G	C	225.417	.	DP=67;VDB=0	GT:PL	1:255,0
NC_000962.3	9304	.	G	A	225.417	.	DP=59;VDB=0	GT:PL	1:255,0
NC_000962.3	11879	.	A	G	225.417	.	DP=50;VDB=0	GT:PL	1:255,0
NC_000962.3	14785	.	T	C	225.417	.	DP=61;VDB=0	GT:PL	1:255,0
NC_000962.3	18091	.	G	A	225.417	.	DP=27;VDB=0	GT:PL	1:255,0
NC_000962.3	21795	.	G	A	95.4151	.	DP=4;VDB=0	GT:PL	1:125,0
NC_000962.3	23854	.	A	G	225.417	.	DP=70;VDB=0	GT:PL	1:255,0
NC_000962.3	26233	.	G	C	225.417	.	DP=52;VDB=0	GT:PL	1:255,0
NC_000962.3	26959	.	C	G	225.417	.	DP=50;VDB=0	GT:PL	1:255,0
NC_000962.3	27518	.	TAAAAA	TAAAA	225.417	.	INDEL;IDV=3	GT:PL	1:255,0
NC_000962.3	27918	.	G	A	225.417	.	DP=64;VDB=0	GT:PL	1:255,0
NC_000962.3	32387	.	C	T	225.417	.	DP=64;VDB=0	GT:PL	1:255,0
NC_000962.3	33817	.	C	G	228.419	.	DP=71;VDB=0	GT:PL	1:255,0
NC_000962.3	34044	.	T	C	225.417	.	DP=59;VDB=0	GT:PL	1:255,0
NC_000962.3	37031	.	C	G	225.417	.	DP=37;VDB=0	GT:PL	1:255,0
NC_000962.3	42967	.	G	C	225.417	.	DP=54;VDB=0	GT:PL	1:255,0
NC_000962.3	47138	.	CTTT	CTT	228.395	.	INDEL;IDV=7	GT:PL	1:255,0
NC_000962.3	49079	.	C	G	225.422	.	DP=66;VDB=0	GT:PL	1:255,0
NC_000962.3	49233	.	AGGG	AGG	228.419	.	INDEL;IDV=6	GT:PL	1:255,0
NC_000962.3	53502	.	G	A	225.417	.	DP=37;VDB=0	GT:PL	1:255,0
NC_000962.3	54304	.	C	T	225.417	.	DP=48;VDB=0	GT:PL	1:255,0
NC_000962.3	55553	.	C	T	192.416	.	DP=13;VDB=0	GT:PL	1:222,0
NC_000962.3	60456	.	G	C	225.417	.	DP=43;VDB=0	GT:PL	1:255,0
NC_000962.3	62049	.	A	G	225.417	.	DP=35;VDB=0	GT:PL	1:255,0
NC_000962.3	67385	.	G	A	225.417	.	DP=49;VDB=0	GT:PL	1:255,0
NC_000962.3	69342	.	G	A	225.417	.	DP=60;VDB=0	GT:PL	1:255,0
NC_000962.3	69989	.	G	A	225.422	.	DP=58;VDB=0	GT:PL	1:255,0
NC_000962.3	70816	.	A	G	225.417	.	DP=41;VDB=0	GT:PL	1:255,0
NC_000962.3	71336	.	G	C	166.416	.	DP=7;VDB=0	GT:PL	1:196,0

Fig 3.1 A VCF annotated with SnpEff.



Summary

Genome	NC_000962.3
Date	2024-08-08 17:51
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 3raw.filter.vcf
Warnings	3,198
Errors	0
Number of lines (input file)	1,029
Number of variants (before filter)	1,029
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	1,029
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,953
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,287 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	1,029	4,287
Total	**4,411,532**	**1,029**	**4,287**

Fig 3.2 SnpEff annotation summary

Number variants by type

Type	Total
SNP	947
MNP	0
INS	37
DEL	45
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	1,029

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	43	0.432%
LOW	336	3.376%
MODERATE	481	4.833%
MODIFIER	9,093	91.359%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	466	57.407%
NONSENSE	9	1.111%
SILENT	336	41.481%

Missense / Silent ratio: 1.3839

Fig 3.3 . Variant effect summary

Number of annotaitons and region counts

Annotation

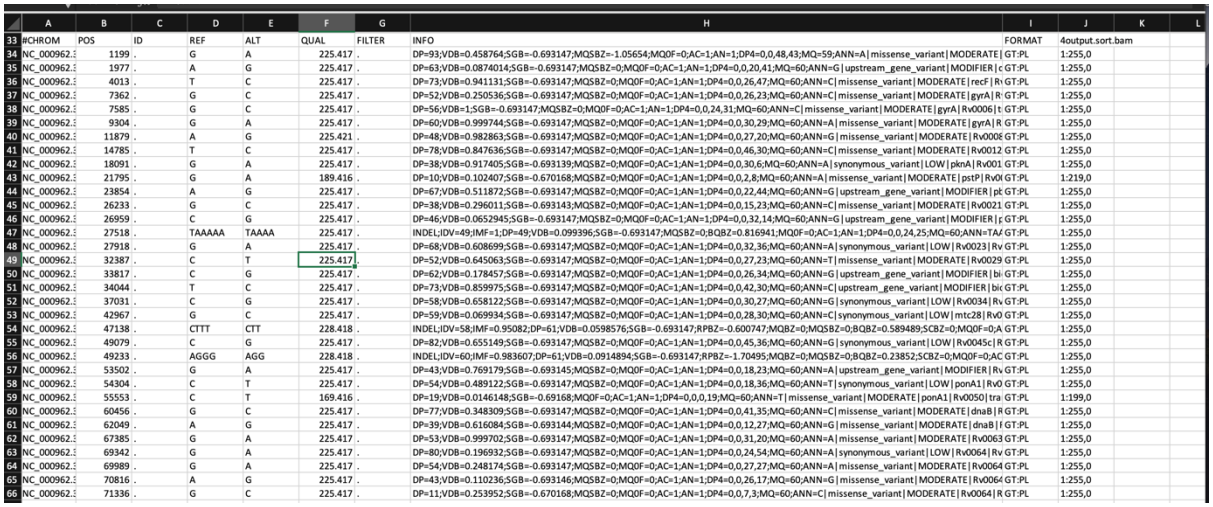
Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	7	0.07%
conservative_inframe_insertion	6	0.06%
disruptive_inframe_deletion	1	0.01%
disruptive_inframe_insertion	4	0.04%
downstream_gene_variant	4,538	45.585%
frameshift_variant	32	0.321%
intergenic_region	132	1.326%
intragenic_variant	40	0.402%
missense_variant	463	4.651%
splice_region_variant	1	0.01%
start_lost	2	0.02%
stop_gained	9	0.09%
stop_lost	1	0.01%
synonymous_variant	336	3.375%
upstream_gene_variant	4,383	44.028%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,538	45.594%
EXON	860	8.641%
INTERGENIC	132	1.326%
TRANSCRIPT	40	0.402%
UPSTREAM	4,383	44.037%

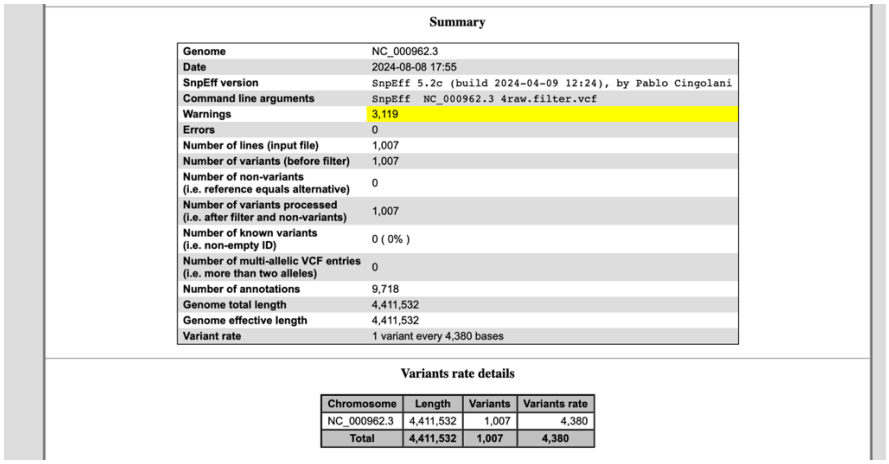
Fig 3.4 The Number of effects by type and region

Fig 4.0 IGV visualisation of a Mycobacterium tuberculosis H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 4



	A	B	C	D	E	F	G	H	I	J	K	L
33	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	4output.sort.bam		
34	NC_000962.3	1199	.	G	A	225.417	.	DP=93;VDB=0.458764;SGB=-0.693147;MQSBZ=-1.05654;MQ0F=0;AC=1;AN=1;DP4=0,0,48,43;MQ=59;ANN=A|missense_variant|MODERATE|	GT:PL	1:255,0		
35	NC_000962.3	1977	.	A	G	225.417	.	DP=63;VDB=0.0874014;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,20,41;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|c	GT:PL	1:255,0		
36	NC_000962.3	4013	.	T	C	225.417	.	DP=73;VDB=0.941131;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,26,47;MQ=60;ANN=C|missense_variant|MODERATE|recF|R	GT:PL	1:255,0		
37	NC_000962.3	7362	.	G	C	225.417	.	DP=52;VDB=0.250536;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,26,23;MQ=60;ANN=C|missense_variant|MODERATE|gyrA|R	GT:PL	1:255,0		
38	NC_000962.3	7585	.	G	C	225.417	.	DP=56;VDB=1;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,24,31;MQ=60;ANN=C|missense_variant|MODERATE|gyrA|Rv0006|t	GT:PL	1:255,0		
39	NC_000962.3	9304	.	G	A	225.417	.	DP=60;VDB=0.999744;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,30,29;MQ=60;ANN=A|missense_variant|MODERATE|gyrA|R	GT:PL	1:255,0		
40	NC_000962.3	11879	.	A	G	225.421	.	DP=48;VDB=0.982863;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,27,20;MQ=60;ANN=G|missense_variant|MODERATE|Rv0008	GT:PL	1:255,0		
41	NC_000962.3	14785	.	T	C	225.417	.	DP=78;VDB=0.847636;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,46,30;MQ=60;ANN=C|missense_variant|MODERATE|Rv0012	GT:PL	1:255,0		
42	NC_000962.3	18091	.	G	A	225.417	.	DP=38;VDB=0.917405;SGB=-0.693139;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,30,6;MQ=60;ANN=A|synonymous_variant|LOW|pknA|Rv001	GT:PL	1:255,0		
43	NC_000962.3	21795	.	G	A	189.416	.	DP=10;VDB=0.102407;SGB=-0.670168;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,8;MQ=60;ANN=A|missense_variant|MODERATE|pstP|Rv0(	GT:PL	1:219,0		
44	NC_000962.3	23854	.	A	G	225.417	.	DP=67;VDB=0.511872;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,44;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|pt	GT:PL	1:255,0		
45	NC_000962.3	26233	.	G	C	225.417	.	DP=38;VDB=0.296011;SGB=-0.693143;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,15,23;MQ=60;ANN=C|missense_variant|MODERATE|Rv0021	GT:PL	1:255,0		
46	NC_000962.3	26959	.	C	G	225.417	.	DP=46;VDB=0.0652945;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,32,14;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|g	GT:PL	1:255,0		
47	NC_000962.3	27518	.	TAAAAA	TAAAA	225.417	.	INDEL;IDV=49;IMF=1;DP=49;VDB=0.099396;SGB=-0.693147;MQSBZ=0;BQBZ=0.816941;MQ0F=0;AC=1;AN=1;DP4=0,0,24,25;MQ=60;ANN=TA/	GT:PL	1:255,0		
48	NC_000962.3	27918	.	G	A	225.417	.	DP=68;VDB=0.608699;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,32,36;MQ=60;ANN=A|synonymous_variant|LOW|Rv0023|Rv	GT:PL	1:255,0		
49	NC_000962.3	32387	.	C	T	225.417	.	DP=52;VDB=0.645063;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,27,23;MQ=60;ANN=T|missense_variant|MODERATE|Rv0029	GT:PL	1:255,0		
50	NC_000962.3	33817	.	C	G	225.417	.	DP=62;VDB=0.178457;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,26,34;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|bi	GT:PL	1:255,0		
51	NC_000962.3	34044	.	T	C	225.417	.	DP=73;VDB=0.859975;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,42,30;MQ=60;ANN=C|upstream_gene_variant|MODIFIER|bi	GT:PL	1:255,0		
52	NC_000962.3	37031	.	C	G	225.417	.	DP=58;VDB=0.658122;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,30,27;MQ=60;ANN=G|synonymous_variant|LOW|Rv0034|Rv	GT:PL	1:255,0		
53	NC_000962.3	42967	.	G	C	225.417	.	DP=59;VDB=0.069934;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,28,30;MQ=60;ANN=C|synonymous_variant|LOW|mtc28|Rv0	GT:PL	1:255,0		
54	NC_000962.3	47138	.	CTTT	CTT	228.418	.	INDEL;IDV=58;IMF=0.95082;DP=61;VDB=0.0598576;SGB=-0.693147;RPBZ=-0.600747;MQBZ=0;MQSBZ=0;BQBZ=0.589489;SCBZ=0;MQ0F=0;A	GT:PL	1:255,0		
55	NC_000962.3	49079	.	C	G	225.417	.	DP=82;VDB=0.655149;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,45,36;MQ=60;ANN=G|synonymous_variant|LOW|Rv0045c|R	GT:PL	1:255,0		
56	NC_000962.3	49233	.	AGGG	AGG	228.418	.	INDEL;IDV=60;IMF=0.983607;DP=61;VDB=0.0914894;SGB=-0.693147;RPBZ=-1.70495;MQBZ=0;MQSBZ=0;BQBZ=0.23852;SCBZ=0;MQ0F=0;AC	GT:PL	1:255,0		
57	NC_000962.3	53502	.	G	A	225.417	.	DP=43;VDB=0.769179;SGB=-0.693145;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,18,23;MQ=60;ANN=A|upstream_gene_variant|MODIFIER|Rv	GT:PL	1:255,0		
58	NC_000962.3	54304	.	C	T	225.417	.	DP=54;VDB=0.489122;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,18,36;MQ=60;ANN=T|synonymous_variant|LOW|ponA1|Rv0	GT:PL	1:255,0		
59	NC_000962.3	55553	.	C	T	169.416	.	DP=19;VDB=0.0146148;SGB=-0.69168;MQ0F=0;AC=1;AN=1;DP4=0,0,0,19;MQ=60;ANN=T|missense_variant|MODERATE|ponA1|Rv0050|tra	GT:PL	1:199,0		
60	NC_000962.3	60456	.	G	C	225.417	.	DP=77;VDB=0.348309;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,41,35;MQ=60;ANN=C|missense_variant|MODERATE|dnaB|R	GT:PL	1:255,0		
61	NC_000962.3	62049	.	A	G	225.417	.	DP=39;VDB=0.616084;SGB=-0.693144;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,12,27;MQ=60;ANN=G|missense_variant|MODERATE|dnaB|R	GT:PL	1:255,0		
62	NC_000962.3	67385	.	G	A	225.417	.	DP=53;VDB=0.999702;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,31,20;MQ=60;ANN=A|missense_variant|MODERATE|Rv0063	GT:PL	1:255,0		
63	NC_000962.3	69342	.	G	A	225.417	.	DP=80;VDB=0.196932;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,24,54;MQ=60;ANN=A|synonymous_variant|LOW|Rv0064|Rv	GT:PL	1:255,0		
64	NC_000962.3	69989	.	G	A	225.417	.	DP=54;VDB=0.248174;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,27,27;MQ=60;ANN=A|missense_variant|MODERATE|Rv0064	GT:PL	1:255,0		
65	NC_000962.3	70816	.	A	G	225.417	.	DP=43;VDB=0.110236;SGB=-0.693146;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,26,17;MQ=60;ANN=G|missense_variant|MODERATE|Rv0064	GT:PL	1:255,0		
66	NC_000962.3	71336	.	G	C	225.417	.	DP=11;VDB=0.253952;SGB=-0.670168;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,3;MQ=60;ANN=C|missense_variant|MODERATE|Rv0064|R	GT:PL	1:255,0		

Fig 4.1 A VCF annotated with SnpEff



Summary

Genome	NC_000962.3
Date	2024-08-08 17:55
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 4raw.filter.vcf
Warnings	3,119
Errors	0
Number of lines (input file)	1,007
Number of variants (before filter)	1,007
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	1,007
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,718
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,380 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	1,007	4,380
Total	**4,411,532**	**1,007**	**4,380**

Fig 4.2 SnpEff annotation summary

Number variants by type

Type	Total
SNP	924
MNP	0
INS	37
DEL	46
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	1,007

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	43	0.442%
LOW	327	3.365%
MODERATE	474	4.878%
MODIFIER	8,874	91.315%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	457	57.629%
NONSENSE	9	1.135%
SILENT	327	41.236%

Missense / Silent ratio: 1.3976

Fig 4.3 . Variant effect summary

Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	8	0.082%
conservative_inframe_insertion	6	0.062%
disruptive_inframe_deletion	1	0.01%
disruptive_inframe_insertion	4	0.041%
downstream_gene_variant	4,417	45.442%
frameshift_variant	32	0.329%
intergenic_region	126	1.296%
intragenic_variant	40	0.412%
missense_variant	455	4.681%
splice_region_variant	1	0.01%
start_lost	2	0.021%
stop_gained	9	0.093%
stop_lost	1	0.01%
synonymous_variant	327	3.364%
upstream_gene_variant	4,291	44.146%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,417	45.452%
EXON	844	8.685%
INTERGENIC	126	1.297%
TRANSCRIPT	40	0.412%
UPSTREAM	4,291	44.155%

Fig 4.4 The Number of effects by type and region

Fig 5.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 5

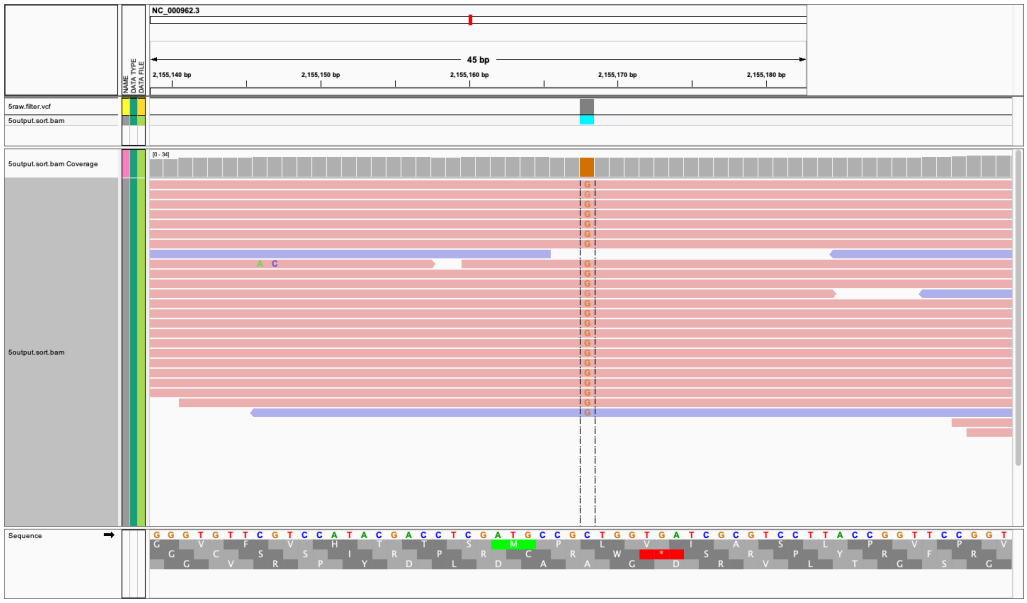


Fig 5.0.1

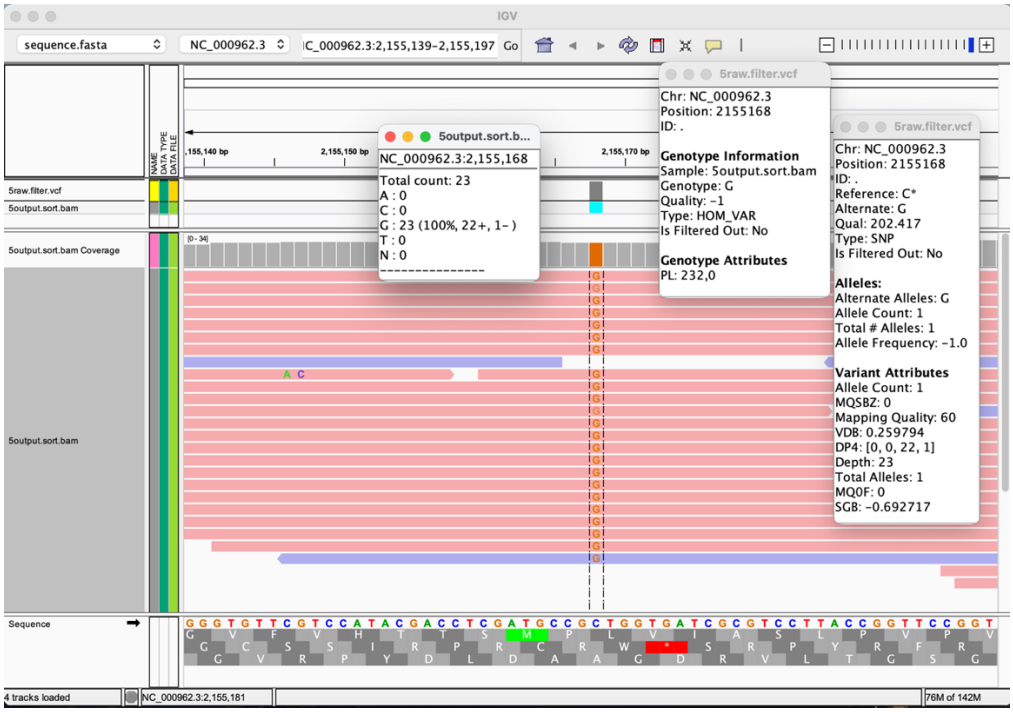
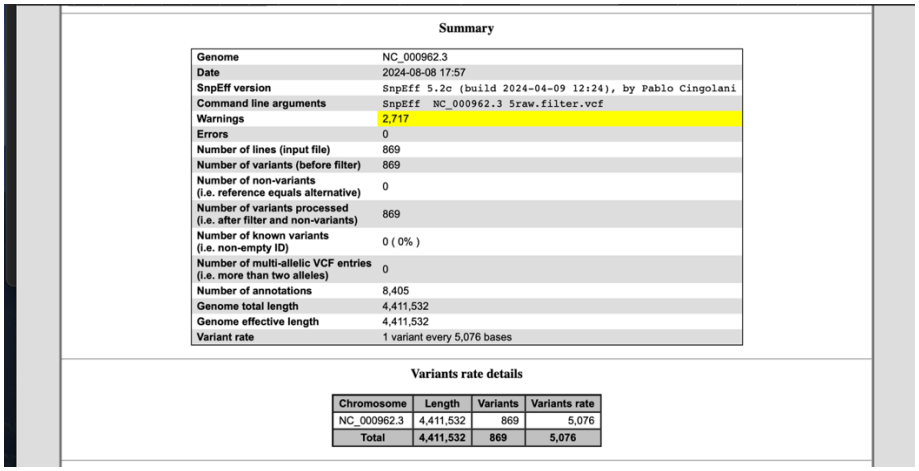


Fig 5.0.2

	B	C	D	E	F	G	H
459	2094911	.	ACAGGGTCA	ACAGCGTCA	228.368	.	INDEL;IDV=25;IMF=0.833333;DP=30;VDB=0.0237154;SGB=-0.692914;RPBZ=-2.28592;MQBZ=0;MQSBZ=0;BQBZ=-2.36294;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,4,5,20;MQ=60;ANN=ACAGCGTCAG | conservative
460	2096186	.	A	G	225.417	.	DP=16;VDB=0.485546;SGB=-0.689466;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,10,6;MQ=60;ANN=G | synonymous_variant | LOW | bla | Rv1846c | transcript | Rv1846c | protein_coding | 1/1 | c.414T>C | p.Thr138Th
461	2109523	.	CGG	CGGG	225.417	.	INDEL;IDV=16;IMF=1;DP=16;VDB=0.0386424;SGB=-0.689466;MQSBZ=0;BQBZ=0.595844;MQ0F=0;AC=1;AN=1;DP4=0,0,8,8;MQ=60;ANN=CGGG | upstream_gene_variant | MODIFIER | Rv1856c | Rv1856c | transc
462	2113058	.	C	A	225.417	.	DP=22;VDB=0.625112;SGB=-0.692562;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,9,13;MQ=60;ANN=A | upstream_gene_variant | MODIFIER | Rv1863c | transcript | Rv1863c | protein_coding | c.-1697G>T
463	2116903	.	C	T	225.417	.	DP=30;VDB=0.15238;SGB=-0.693097;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,12,18;MQ=60;ANN=T | synonymous_variant | LOW | Rv1867 | Rv1867 | transcript | Rv1867 | protein_coding | 1/1 | c.1140C>T | p.Gly380
464	2122395	.	C	T	225.417	.	DP=18;VDB=0.31448;SGB=-0.691153;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,11,7;MQ=60;ANN=T | missense_variant | MODERATE | IldD2 | Rv1872c | transcript | Rv1872c | protein_coding | 1/1 | c.757G>A | p.Val2
465	2123168	.	T	G	225.417	.	DP=22;VDB=0.123161;SGB=-0.692562;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,10,12;MQ=60;ANN=G | upstream_gene_variant | MODIFIER | Rv1869c | Rv1869c | transcript | Rv1869c | protein_coding | c.-2473A>
466	2123169	.	T	G	225.417	.	DP=22;VDB=0.130153;SGB=-0.692562;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,10,12;MQ=60;ANN=G | upstream_gene_variant | MODIFIER | Rv1869c | Rv1869c | transcript | Rv1869c | protein_coding | c.-2474A>
467	2128870	.	A	G	225.417	.	DP=25;VDB=0.305145;SGB=-0.692914;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,18;MQ=60;ANN=G | synonymous_variant | LOW | glnA3 | Rv1878 | transcript | Rv1878 | protein_coding | 1/1 | c.849A>G | p.Leu283L
468	2133468	.	TTCGCAT	TTCGCATGC(	206.069	.	INDEL;IDV=7;IMF=0.4375;DP=16;VDB=0.372161;SGB=-0.636426;RPBZ=-3.42624;MQBZ=0;MQSBZ=0;BQBZ=0.470162;SCBZ=-2.86285;MQ0F=0;AC=1;AN=1;DP4=4,5,4,3;MQ=60;ANN=TTCGCATGCCGTCACCTCG
469	2135870	.	T	C	225.421	.	DP=19;VDB=0.979035;SGB=-0.69168;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,12;MQ=60;ANN=C | upstream_gene_variant | MODIFIER | cyp140 | Rv1880c | transcript | Rv1880c | protein_coding | c.-4013A>G |
470	2138245	.	CTGGTAATO	CT	225.417	.	INDEL;IDV=28;IMF=1;DP=28;VDB=0.294022;SGB=-0.693054;MQSBZ=0;BQBZ=-2.22375;MQ0F=0;AC=1;AN=1;DP4=0,0,11,17;MQ=60;ANN=CT | upstream_gene_variant | MODIFIER | Rv1883c | Rv1883c | transcript
471	2143328	.	G	C	164.416	.	DP=8;VDB=0.973385;SGB=-0.651104;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,1;MQ=60;ANN=C | missense_variant | MODERATE | Rv1895 | Rv1895 | transcript | Rv1895 | protein_coding | 1/1 | c.808G>C | p.Val27
472	2143958	.	C	T	225.417	.	DP=18;VDB=0.606507;SGB=-0.691153;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,6,12;MQ=60;ANN=T | stop_gained | HIGH | Rv1896c | Rv1896c | transcript | Rv1896c | protein_coding | 1/1 | c.489G>A | p.Trp163* | 48
473	2147022	.	A	C	225.417	.	DP=18;VDB=0.562568;SGB=-0.691153;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,11,7;MQ=60;ANN=C | missense_variant | MODERATE | lipJ | Rv1900c | transcript | Rv1900c | protein_coding | 1/1 | c.612T>G | p.Ile20
474	2155168	.	C	G	202.417	.	DP=23;VDB=0.259794;SGB=-0.692717;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,1;MQ=60;ANN=G | missense_variant | MODERATE | katG | Rv1908c | transcript | Rv1908c | protein_coding | 1/1 | c.944G>C | p.Ser
475	2158327	.	C	T	225.417	.	DP=24;VDB=0.89019;SGB=-0.692831;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,15,9;MQ=60;ANN=T | synonymous_variant | LOW | fadB5 | Rv1912c | transcript | Rv1912c | protein_coding | 1/1 | c.765G>A | p.Leu255
476	2160996	.	C	T	225.417	.	DP=21;VDB=0.314075;SGB=-0.692352;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,14;MQ=60;ANN=T | synonymous_variant | LOW | aceAa | Rv1915 | transcript | Rv1915 | protein_coding | 1/1 | c.534C>T | p.Ala178Al
477	2163444	.	T	C	77.6702	.	DP=58;VDB=2.14616e-11;SGB=-0.693143;RPBZ=-2.69713;MQBZ=2.2559;MQSBZ=0.771145;BQBZ=-1.67527;SCBZ=-0.901416;MQ0F=0.0172414;AC=1;AN=1;DP4=7,13,23,15;MQ=43;ANN=C | missense_variant
478	2163481	.	G	A	12.7751	.	DP=9;VDB=0.979732;SGB=-0.616816;RPBZ=-0.774597;MQBZ=-0.80904;MQSBZ=-0.152894;BQBZ=-0.284268;SCBZ=0;MQ0F=0.111111;AC=1;AN=1;DP4=1,2,1,5;MQ=34;ANN=A | synonymous_variant | LOW | PPE
479	2163484	.	A	C	4.54011	.	DP=7;VDB=0.0298006;SGB=-0.556411;RPBZ=-2.12132;MQBZ=-1.42701;MQSBZ=0.504525;BQBZ=1.38042;SCBZ=0;MQ0F=0.142857;AC=1;AN=1;DP4=1,2,0,4;MQ=27;ANN=C | synonymous_variant | LOW | PPE34
480	2163493	.	A	G	27.4222	.	DP=4;VDB=0.0298006;SGB=-0.556411;MQ0F=0.25;AC=1;AN=1;DP4=0,0,0,4;MQ=16;ANN=G | synonymous_variant | LOW | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3819T>C | p.Ala1273Ala | 3
481	2163494	.	G	C	27.4222	.	DP=4;VDB=0.0277111;SGB=-0.556411;MQ0F=0.25;AC=1;AN=1;DP4=0,0,0,4;MQ=16;ANN=C | missense_variant | MODERATE | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3818C>G | p.Ala1273G
482	2163496	.	A	C	27.4222	.	DP=4;VDB=0.0298006;SGB=-0.556411;MQ0F=0.25;AC=1;AN=1;DP4=0,0,0,4;MQ=16;ANN=C | synonymous_variant | LOW | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3816T>G | p.Gly1272Gly | 3
483	2163504	.	G	A	176.416	.	DP=10;VDB=0.034237;SGB=-0.670168;MQSBZ=-1.1767;MQ0F=0.1;AC=1;AN=1;DP4=0,0,2,8;MQ=42;ANN=A | synonymous_variant | LOW | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3808C>T |
484	2163510	.	T	C	176.416	.	DP=10;VDB=0.0293045;SGB=-0.670168;MQSBZ=-1.1767;MQ0F=0.1;AC=1;AN=1;DP4=0,0,2,8;MQ=42;ANN=C | missense_variant | MODERATE | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3802A
485	2163517	.	A	C	219.417	.	DP=13;VDB=0.012172;SGB=-0.683931;MQSBZ=-1.50772;MQ0F=0.0769231;AC=1;AN=1;DP4=0,0,4,9;MQ=46;ANN=C | synonymous_variant | LOW | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3
486	2163520	.	G	A	225.417	.	DP=14;VDB=0.0100057;SGB=-0.686358;MQSBZ=-1.41714;MQ0F=0.0714286;AC=1;AN=1;DP4=0,0,4,10;MQ=47;ANN=A | synonymous_variant | LOW | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c
487	2163790	.	A	C	221.417	.	DP=15;VDB=0.0308973;SGB=-0.688148;MQSBZ=-1.28895;MQ0F=0;AC=1;AN=1;DP4=0,0,3,12;MQ=47;ANN=C | synonymous_variant | LOW | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.3522T>G
488	2165286	.	A	C	225.417	.	DP=19;VDB=0.973384;SGB=-0.69168;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,7,12;MQ=60;ANN=C | missense_variant | MODERATE | PPE34 | Rv1917c | transcript | Rv1917c | protein_coding | 1/1 | c.2026T>G | p.Se
489	2176101	.	A	G	225.417	.	DP=17;VDB=0.999261;SGB=-0.690438;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,6,11;MQ=60;ANN=G | missense_variant | MODERATE | lipD | Rv1923 | transcript | Rv1923 | protein_coding | 1/1 | c.929A>G | p.Tyr310
490	2181805	.	AG	A	228.417	.	INDEL;IDV=26;IMF=0.962963;DP=27;VDB=0.834135;SGB=-0.692976;RPBZ=-1.67109;MQBZ=0;MQSBZ=0;BQBZ=-1.99094;SCBZ=0.196116;MQ0F=0;AC=1;AN=1;DP4=0,1,12,14;MQ=60;ANN=A | frameshift_varia
491	2183054	.	T	G	225.417	.	DP=24;VDB=0.238507;SGB=-0.692831;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,17,7;MQ=60;ANN=G | synonymous_variant | LOW | Rv1931c | transcript | Rv1931c | protein_coding | 1/1 | c.186A>C | p.Gly
492	2207591	.	T	TC	228.417	.	INDEL;IDV=22;IMF=0.956522;DP=23;VDB=0.018408;SGB=-0.692562;RPBZ=-1.65954;MQBZ=0;MQSBZ=0;BQBZ=-2.32872;SCBZ=-4.69042;MQ0F=0;AC=1;AN=1;DP4=1,0,12,10;MQ=60;ANN=TC | upstream_gene

Fig 5.1 A VCF annotated with SnpEff.



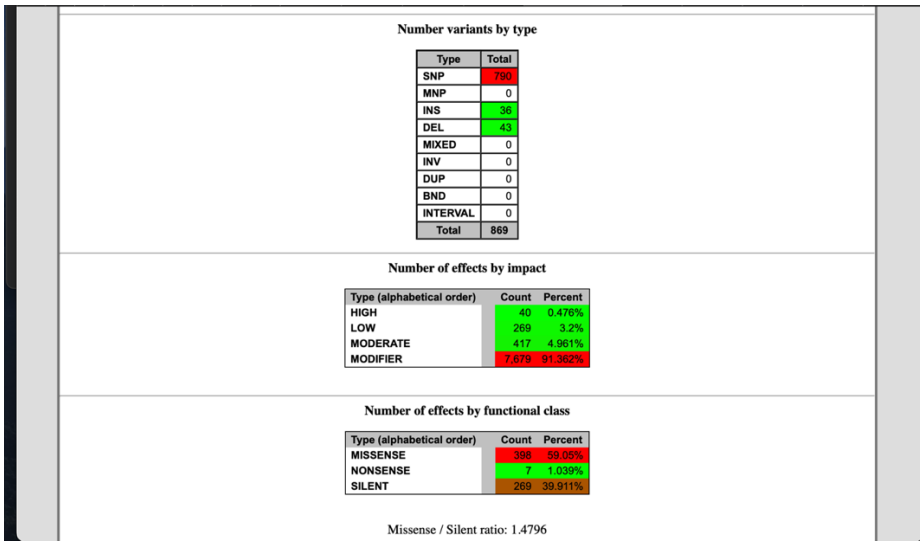
Summary

Genome	NC_000962.3
Date	2024-08-08 17:57
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 5raw.filter.vcf
Warnings	2,717
Errors	0
Number of lines (input file)	869
Number of variants (before filter)	869
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	869
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	8,405
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 5,076 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	869	5,076
Total	**4,411,532**	**869**	**5,076**

Fig 5.2 SnpEff annotation summary



Number variants by type

Type	Total
SNP	790
MNP	0
INS	36
DEL	43
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	**869**

Number of effects by impact

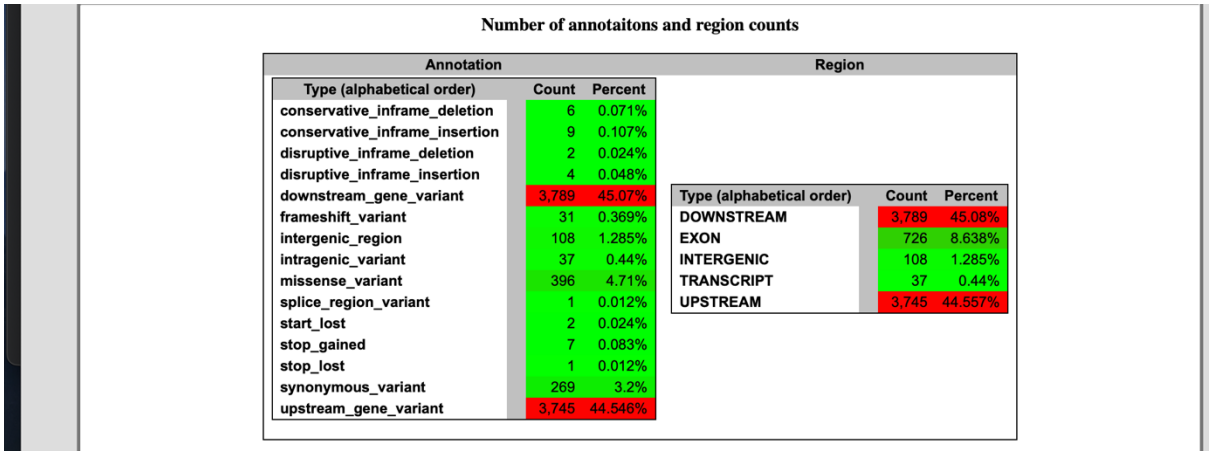
Type (alphabetical order)	Count	Percent
HIGH	40	0.476%
LOW	269	3.2%
MODERATE	417	4.961%
MODIFIER	7,679	91.362%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	398	59.05%
NONSENSE	7	1.039%
SILENT	269	39.911%

Missense / Silent ratio: 1.4796

Fig 5.3 Variant effect summary.



Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	0.071%
conservative_inframe_insertion	9	0.107%
disruptive_inframe_deletion	2	0.024%
disruptive_inframe_insertion	4	0.048%
downstream_gene_variant	3,789	45.07%
frameshift_variant	31	0.369%
intergenic_region	108	1.285%
intragenic_variant	37	0.44%
missense_variant	396	4.71%
splice_region_variant	1	0.012%
start_lost	2	0.024%
stop_gained	7	0.083%
stop_lost	1	0.012%
synonymous_variant	269	3.2%
upstream_gene_variant	3,745	44.546%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	3,789	45.08%
EXON	726	8.638%
INTERGENIC	108	1.285%
TRANSCRIPT	37	0.44%
UPSTREAM	3,745	44.557%

Fig 5.4 The Number of effects by type and region

Fig 6.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 6

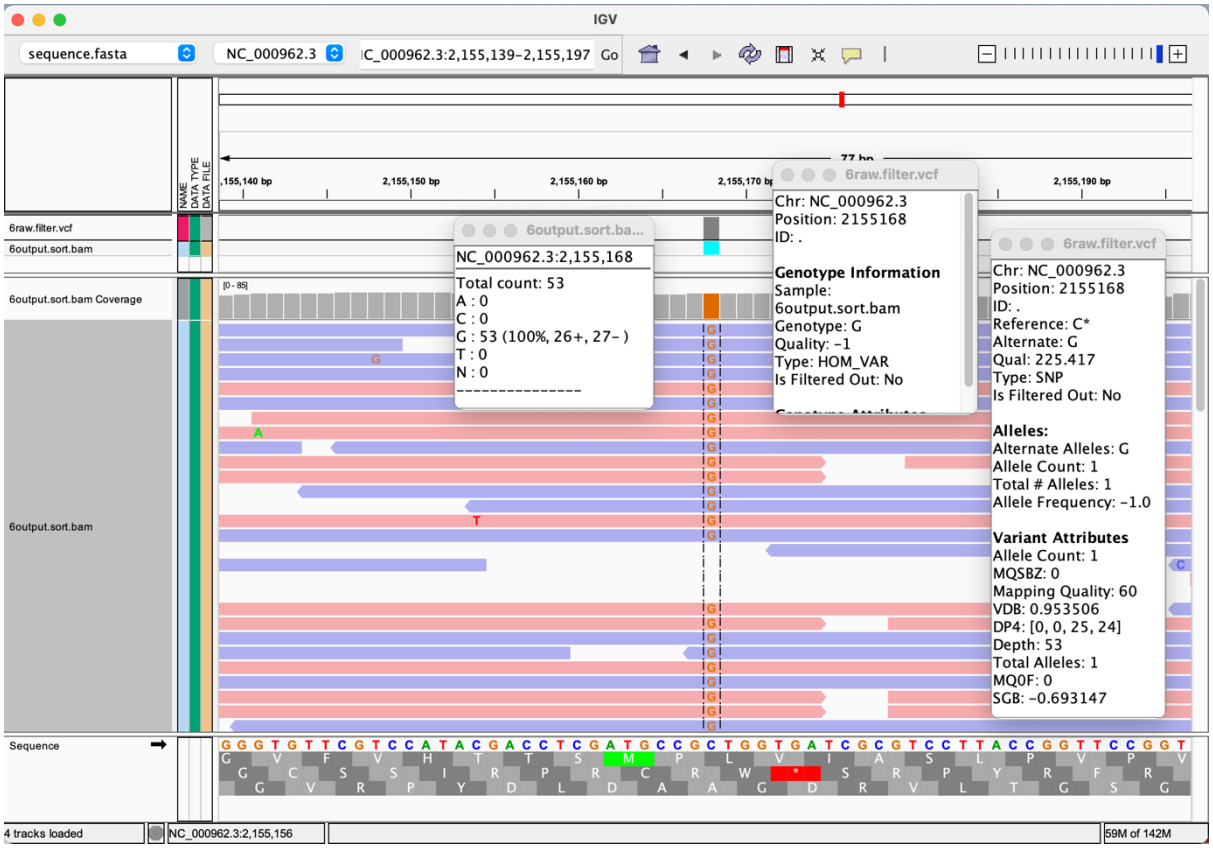

IGV
sequence.fasta
NC_000962.3
C_000962.3:2,155,139-2,155,197
Go
77 bp
2,155,140 bp
2,155,150 bp
2,155,160 bp
2,155,170 bp
2,155,190 bp
6raw.filter.vcf
6output.sort.bam
6output.sort.bam Coverage
6output.sort.bam
Sequence
6output.sort.ba...
NC_000962.3:2,155,168
Total count: 53
A : 0
C : 0
G : 53 (100%, 26+, 27-)
T : 0
N : 0
6raw.filter.vcf
Chr: NC_000962.3
Position: 2155168
ID: .
Genotype Information
Sample: 6output.sort.bam
Genotype: G
Quality: -1
Type: HOM_VAR
Is Filtered Out: No
6raw.filter.vcf
Chr: NC_000962.3
Position: 2155168
ID: .
Reference: C*
Alternate: G
Qual: 225.417
Type: SNP
Is Filtered Out: No
Alleles:
Alternate Alleles: G
Allele Count: 1
Total # Alleles: 1
Allele Frequency: -1.0
Variant Attributes
Allele Count: 1
MQSBZ: 0
Mapping Quality: 60
VDB: 0.953506
DP4: [0, 0, 25, 24]
Depth: 53
Total Alleles: 1
MQ0F: 0
SGB: -0.693147
4 tracks loaded
NC_000962.3:2,155,156
59M of 142M

Fig 6.0.1

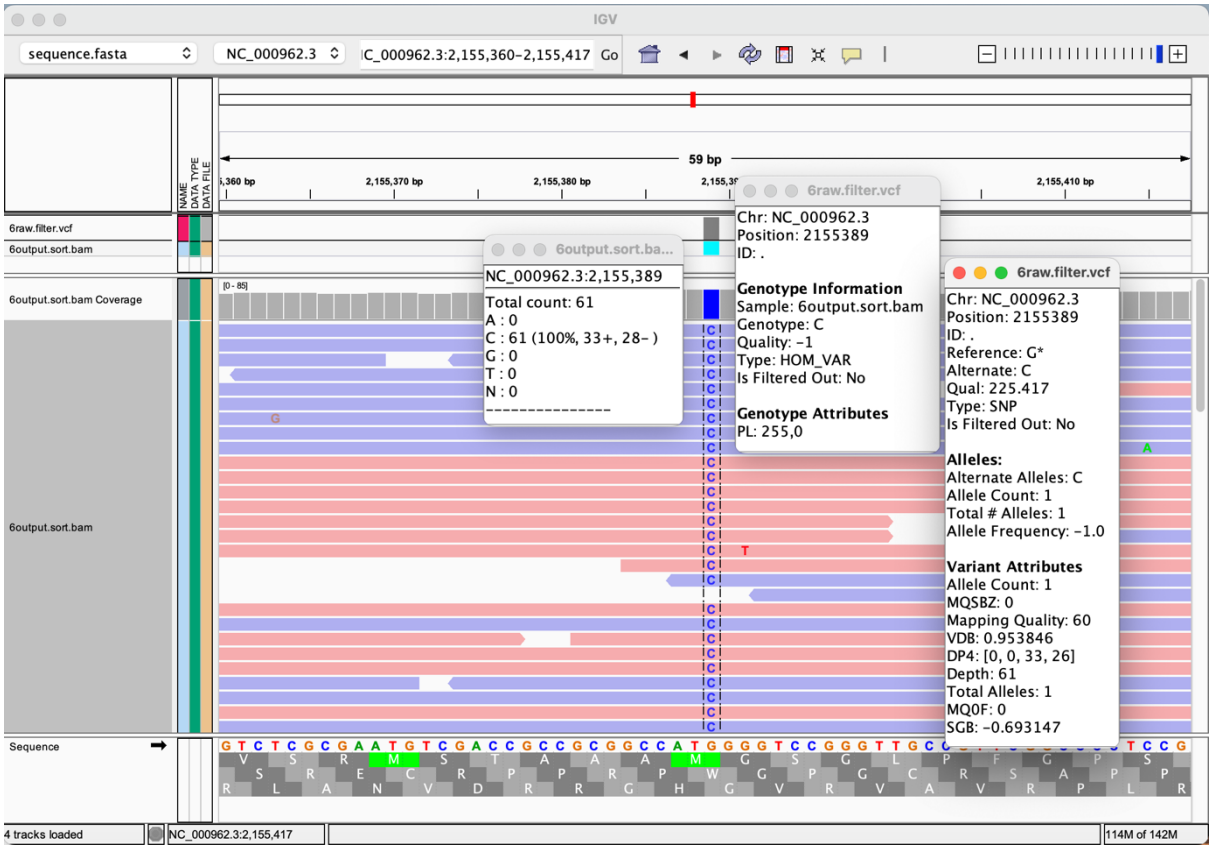
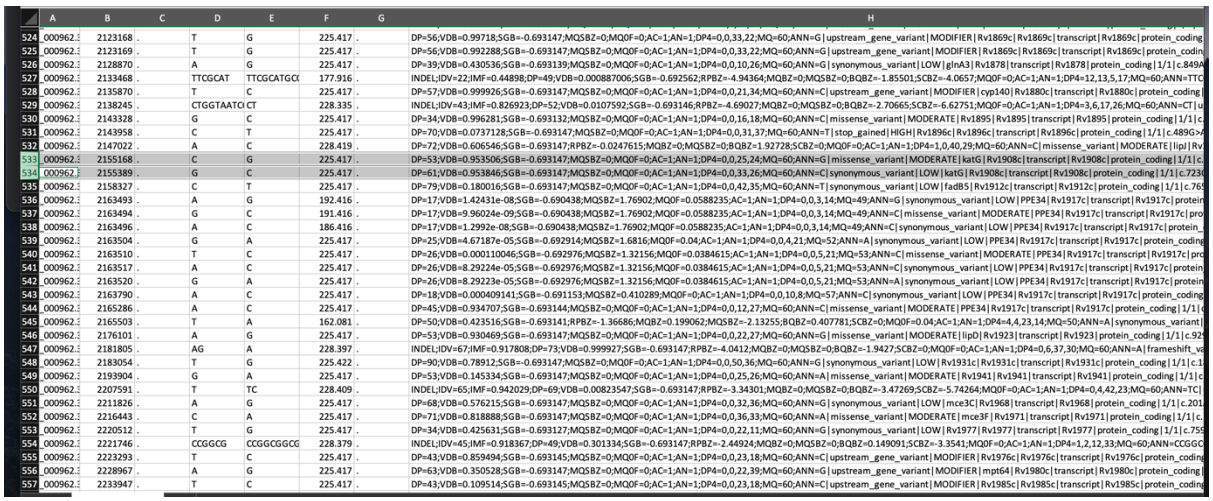
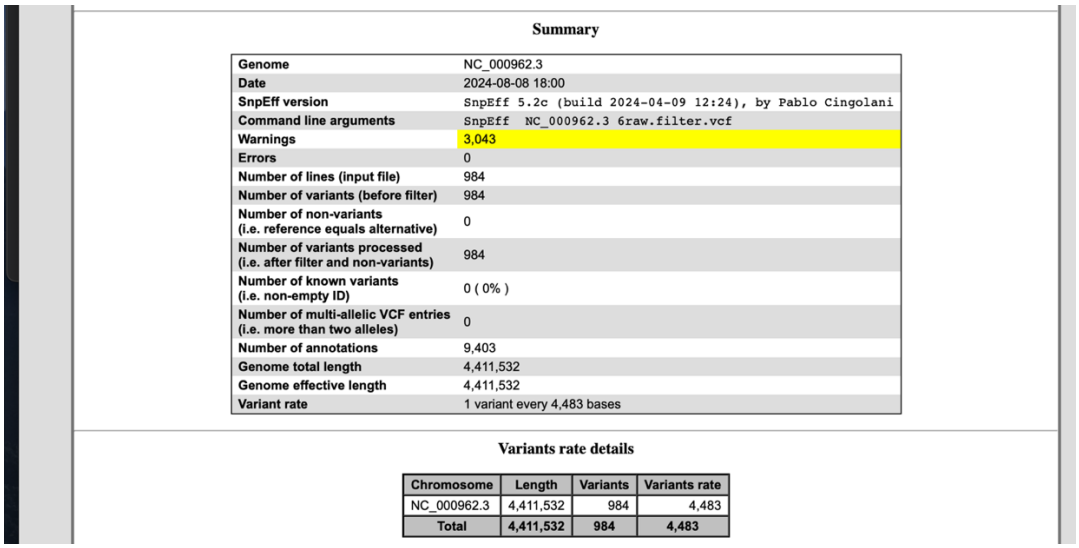

IGV
sequence.fasta
NC_000962.3
C_000962.3:2,155,360-2,155,417
Go
59 bp
2,155,360 bp
2,155,370 bp
2,155,380 bp
2,155,410 bp
6raw.filter.vcf
6output.sort.bam
6output.sort.bam Coverage
6output.sort.bam
Sequence
6output.sort.ba...
NC_000962.3:2,155,389
Total count: 61
A : 0
C : 61 (100%, 33+, 28-)
G : 0
T : 0
N : 0
6raw.filter.vcf
Chr: NC_000962.3
Position: 2155389
ID: .
Genotype Information
Sample: 6output.sort.bam
Genotype: C
Quality: -1
Type: HOM_VAR
Is Filtered Out: No
Genotype Attributes
PL: 255,0
6raw.filter.vcf
Chr: NC_000962.3
Position: 2155389
ID: .
Reference: G*
Alternate: C
Qual: 225.417
Type: SNP
Is Filtered Out: No
Alleles:
Alternate Alleles: C
Allele Count: 1
Total # Alleles: 1
Allele Frequency: -1.0
Variant Attributes
Allele Count: 1
MQSBZ: 0
Mapping Quality: 60
VDB: 0.953846
DP4: [0, 0, 33, 26]
Depth: 61
Total Alleles: 1
MQ0F: 0
SGB: -0.693147
4 tracks loaded
NC_000962.3:2,155,417
114M of 142M

Fig 6.0.2



	A	B	C	D	E	F	G	H
524	000962.2	2123168	.	T	G	225.417	.	DP=56;VDB=0.99718;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,33,22;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|Rv1869c|Rv1869c|transcript|Rv1869c|protein_coding
525	000962.2	2123169	.	T	G	225.417	.	DP=56;VDB=0.992288;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,33,22;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|Rv1869c|Rv1869c|transcript|Rv1869c|protein_codin
526	000962.2	2128870	.	A	G	225.417	.	DP=39;VDB=0.430536;SGB=-0.693139;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,10,26;MQ=60;ANN=G|synonymous_variant|LOW|glnA3|Rv1878|transcript|Rv1878|protein_coding|1/1|c.849A
527	000962.2	2133468	.	TTCGCAT	TTCGCATGC(	177.916	.	INDEL;IDV=22;IMF=0.44898;DP=49;VDB=0.000887006;SGB=-0.692562;RPBZ=-4.94364;MQBZ=0;MQSBZ=0;BQBZ=-1.85501;SCBZ=-4.0657;MQ0F=0;AC=1;AN=1;DP4=12,13,5,17;MQ=60;ANN=TTC
528	000962.2	2135870	.	T	C	225.417	.	DP=57;VDB=0.999926;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,21,34;MQ=60;ANN=C|upstream_gene_variant|MODIFIER|cyp140|Rv1880c|transcript|Rv1880c|protein_coding|
529	000962.2	2138245	.	CTGGTAATC(	CT	228.335	.	INDEL;IDV=43;IMF=0.826923;DP=52;VDB=0.0107592;SGB=-0.693146;RPBZ=-4.69027;MQBZ=0;MQSBZ=0;BQBZ=-2.70665;SCBZ=-6.62751;MQ0F=0;AC=1;AN=1;DP4=3,6,17,26;MQ=60;ANN=CT|u
530	000962.2	2143328	.	G	C	225.417	.	DP=34;VDB=0.996281;SGB=-0.693132;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,16,18;MQ=60;ANN=C|missense_variant|MODERATE|Rv1895|Rv1895|transcript|Rv1895|protein_coding|1/1|c
531	000962.2	2143958	.	C	T	225.417	.	DP=70;VDB=0.0737128;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,31,37;MQ=60;ANN=T|stop_gained|HIGH|Rv1896c|Rv1896c|transcript|Rv1896c|protein_coding|1/1|c.489G>A
532	000962.2	2147022	.	A	C	228.419	.	DP=72;VDB=0.606546;SGB=-0.693147;RPBZ=-0.0247615;MQBZ=0;MQSBZ=0;BQBZ=1.92728;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=1,0,40,29;MQ=60;ANN=C|missense_variant|MODERATE|lipJ|Rv
533	000962.2	2155168	.	C	G	225.417	.	DP=53;VDB=0.953506;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,25,24;MQ=60;ANN=G|missense_variant|MODERATE|katG|Rv1908c|transcript|Rv1908c|protein_coding|1/1|c.
534	000962.2	2155389	.	G	C	225.417	.	DP=61;VDB=0.953846;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,33,26;MQ=60;ANN=C|synonymous_variant|LOW|katG|Rv1908c|transcript|Rv1908c|protein_coding|1/1|c.723
535	000962.2	2158327	.	C	T	225.417	.	DP=79;VDB=0.180016;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,42,35;MQ=60;ANN=T|synonymous_variant|LOW|fadB5|Rv1912c|transcript|Rv1912c|protein_coding|1/1|c.765
536	000962.2	2163493	.	A	G	192.416	.	DP=17;VDB=1.42431e-08;SGB=-0.690438;MQSBZ=1.76902;MQ0F=0.0588235;AC=1;AN=1;DP4=0,0,3,14;MQ=49;ANN=G|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein
537	000962.2	2163494	.	G	C	191.416	.	DP=17;VDB=9.96024e-09;SGB=-0.690438;MQSBZ=1.76902;MQ0F=0.0588235;AC=1;AN=1;DP4=0,0,3,14;MQ=49;ANN=C|missense_variant|MODERATE|PPE34|Rv1917c|transcript|Rv1917c|pro
538	000962.2	2163496	.	A	C	186.416	.	DP=17;VDB=1.2992e-08;SGB=-0.690438;MQSBZ=1.76902;MQ0F=0.0588235;AC=1;AN=1;DP4=0,0,3,14;MQ=49;ANN=C|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_
539	000962.2	2163504	.	G	A	225.417	.	DP=25;VDB=4.67187e-05;SGB=-0.692914;MQSBZ=1.6816;MQ0F=0.04;AC=1;AN=1;DP4=0,0,4,21;MQ=52;ANN=A|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_coding
540	000962.2	2163510	.	T	C	225.417	.	DP=26;VDB=0.000110046;SGB=-0.692976;MQSBZ=1.32156;MQ0F=0.0384615;AC=1;AN=1;DP4=0,0,5,21;MQ=53;ANN=C|missense_variant|MODERATE|PPE34|Rv1917c|transcript|Rv1917c|pro
541	000962.2	2163517	.	A	C	225.417	.	DP=26;VDB=8.29224e-05;SGB=-0.692976;MQSBZ=1.32156;MQ0F=0.0384615;AC=1;AN=1;DP4=0,0,5,21;MQ=53;ANN=C|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein
542	000962.2	2163520	.	G	A	225.417	.	DP=26;VDB=8.29223e-05;SGB=-0.692976;MQSBZ=1.32156;MQ0F=0.0384615;AC=1;AN=1;DP4=0,0,5,21;MQ=53;ANN=A|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein
543	000962.2	2163790	.	A	C	225.417	.	DP=18;VDB=0.000409141;SGB=-0.691153;MQSBZ=0.410289;MQ0F=0;AC=1;AN=1;DP4=0,0,10,8;MQ=57;ANN=C|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_coding
544	000962.2	2165286	.	A	C	225.417	.	DP=45;VDB=0.934707;SGB=-0.693144;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,12,27;MQ=60;ANN=C|missense_variant|MODERATE|PPE34|Rv1917c|transcript|Rv1917c|protein_coding|1/1|c
545	000962.2	2165503	.	T	A	162.081	.	DP=50;VDB=0.423516;SGB=-0.693141;RPBZ=-1.36686;MQBZ=0.199062;MQSBZ=-2.13255;BQBZ=0.407781;SCBZ=0;MQ0F=0.04;AC=1;AN=1;DP4=4,4,23,14;MQ=50;ANN=A|synonymous_variant|
546	000962.2	2176101	.	A	G	225.417	.	DP=53;VDB=0.930469;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,27;MQ=60;ANN=G|missense_variant|MODERATE|lipD|Rv1923|transcript|Rv1923|protein_coding|1/1|c.92
547	000962.2	2181805	.	AG	A	228.397	.	INDEL;IDV=67;IMF=0.917808;DP=73;VDB=0.999927;SGB=-0.693147;RPBZ=-4.0412;MQBZ=0;MQSBZ=0;BQBZ=-1.9427;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,6,37,30;MQ=60;ANN=A|frameshift_va
548	000962.2	2183054	.	T	G	225.422	.	DP=90;VDB=0.78912;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,50,36;MQ=60;ANN=G|synonymous_variant|LOW|Rv1931c|Rv1931c|transcript|Rv1931c|protein_coding|1/1|c.1
549	000962.2	2193904	.	G	A	225.417	.	DP=53;VDB=0.145334;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,25,26;MQ=60;ANN=A|missense_variant|MODERATE|Rv1941|Rv1941|transcript|Rv1941|protein_coding|1/1|c
550	000962.2	2207591	.	T	TC	228.409	.	INDEL;IDV=65;IMF=0.942029;DP=69;VDB=0.00823547;SGB=-0.693147;RPBZ=-3.34301;MQBZ=0;MQSBZ=0;BQBZ=-3.47269;SCBZ=-5.74264;MQ0F=0;AC=1;AN=1;DP4=0,4,42,23;MQ=60;ANN=TC|
551	000962.2	2211826	.	A	G	225.417	.	DP=68;VDB=0.576215;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,32,36;MQ=60;ANN=G|synonymous_variant|LOW|mce3C|Rv1968|transcript|Rv1968|protein_coding|1/1|c.201
552	000962.2	2216443	.	C	A	225.417	.	DP=71;VDB=0.818888;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,36,33;MQ=60;ANN=A|missense_variant|MODERATE|mce3F|Rv1971|transcript|Rv1971|protein_coding|1/1|c.
553	000962.2	2220512	.	T	G	225.417	.	DP=34;VDB=0.425631;SGB=-0.693127;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,11;MQ=60;ANN=G|synonymous_variant|LOW|Rv1977|Rv1977|transcript|Rv1977|protein_coding|1/1|c.759
554	000962.2	2221746	.	CCGGCG	CCGGCGGCG	228.379	.	INDEL;IDV=45;IMF=0.918367;DP=49;VDB=0.301334;SGB=-0.693147;RPBZ=-2.44924;MQBZ=0;MQSBZ=0;BQBZ=0.149091;SCBZ=-3.3541;MQ0F=0;AC=1;AN=1;DP4=1,2,12,33;MQ=60;ANN=CCGGC
555	000962.2	2223293	.	T	C	225.417	.	DP=43;VDB=0.859494;SGB=-0.693145;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,23,18;MQ=60;ANN=C|upstream_gene_variant|MODIFIER|Rv1976c|Rv1976c|transcript|Rv1976c|protein_coding
556	000962.2	2228967	.	A	G	225.417	.	DP=63;VDB=0.350528;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,22,39;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|mpt64|Rv1980c|transcript|Rv1980c|protein_coding|
557	000962.2	2233947	.	T	C	225.417	.	DP=43;VDB=0.109514;SGB=-0.693145;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,23,18;MQ=60;ANN=C|upstream_gene_variant|MODIFIER|Rv1985c|Rv1985c|transcript|Rv1985c|protein_coding

Fig 6.1 A VCF annotated with SnpEff.



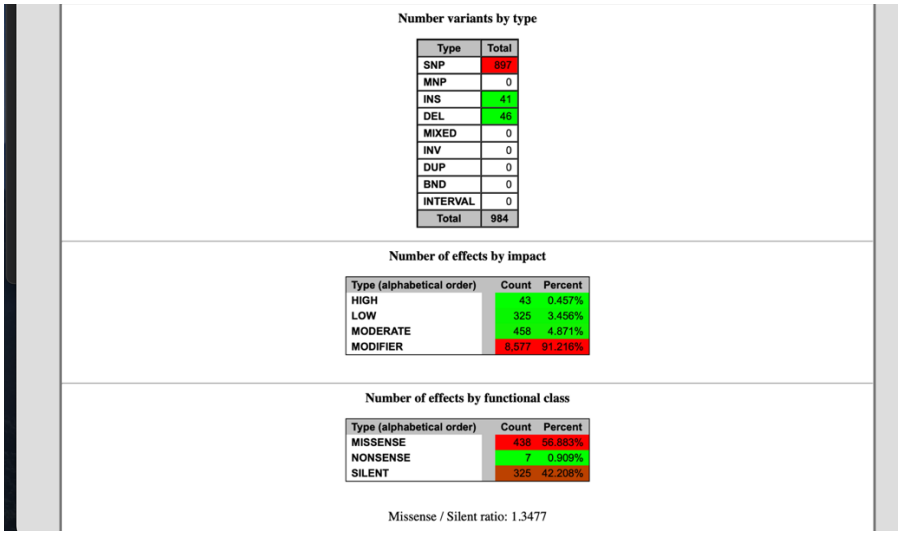
Summary

Genome	NC_000962.3
Date	2024-08-08 18:00
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 6raw.filter.vcf
Warnings	3,043
Errors	0
Number of lines (input file)	984
Number of variants (before filter)	984
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	984
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,403
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,483 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	984	4,483
Total	**4,411,532**	**984**	**4,483**

Fig 6.2 SnpEff annotation summary



Number variants by type

Type	Total
SNP	897
MNP	0
INS	41
DEL	46
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	984

Number of effects by impact

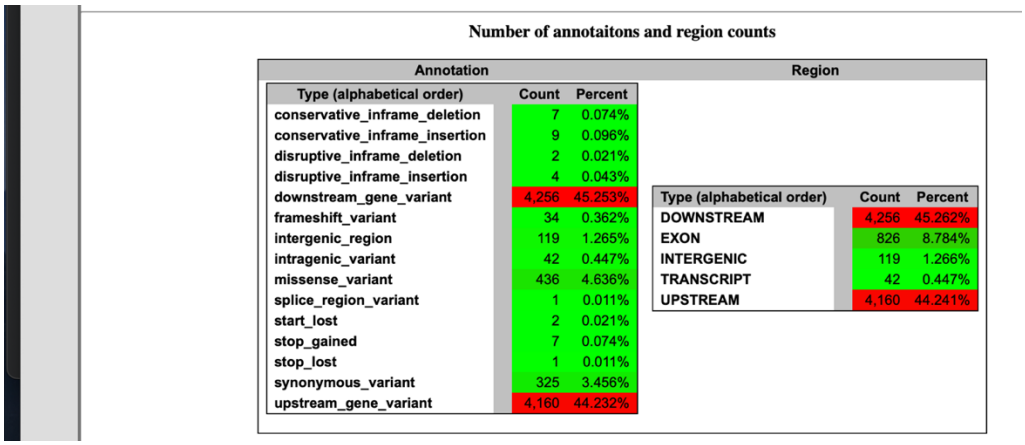
Type (alphabetical order)	Count	Percent
HIGH	43	0.457%
LOW	325	3.456%
MODERATE	458	4.871%
MODIFIER	8,577	91.216%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	438	56.883%
NONSENSE	7	0.909%
SILENT	325	42.208%

Missense / Silent ratio: 1.3477

Fig 6.3 Variant effect summary.



Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	7	0.074%
conservative_inframe_insertion	9	0.096%
disruptive_inframe_deletion	2	0.021%
disruptive_inframe_insertion	4	0.043%
downstream_gene_variant	4,256	45.253%
frameshift_variant	34	0.362%
intergenic_region	119	1.265%
intragenic_variant	42	0.447%
missense_variant	436	4.636%
splice_region_variant	1	0.011%
start_lost	2	0.021%
stop_gained	7	0.074%
stop_lost	1	0.011%
synonymous_variant	325	3.456%
upstream_gene_variant	4,160	44.232%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,256	45.262%
EXON	826	8.784%
INTERGENIC	119	1.266%
TRANSCRIPT	42	0.447%
UPSTREAM	4,160	44.241%

Fig 6.4 The Number of effects by type and region

Fig 7.0 IGV visualisation of a Mycobacterium tuberculosis H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 7

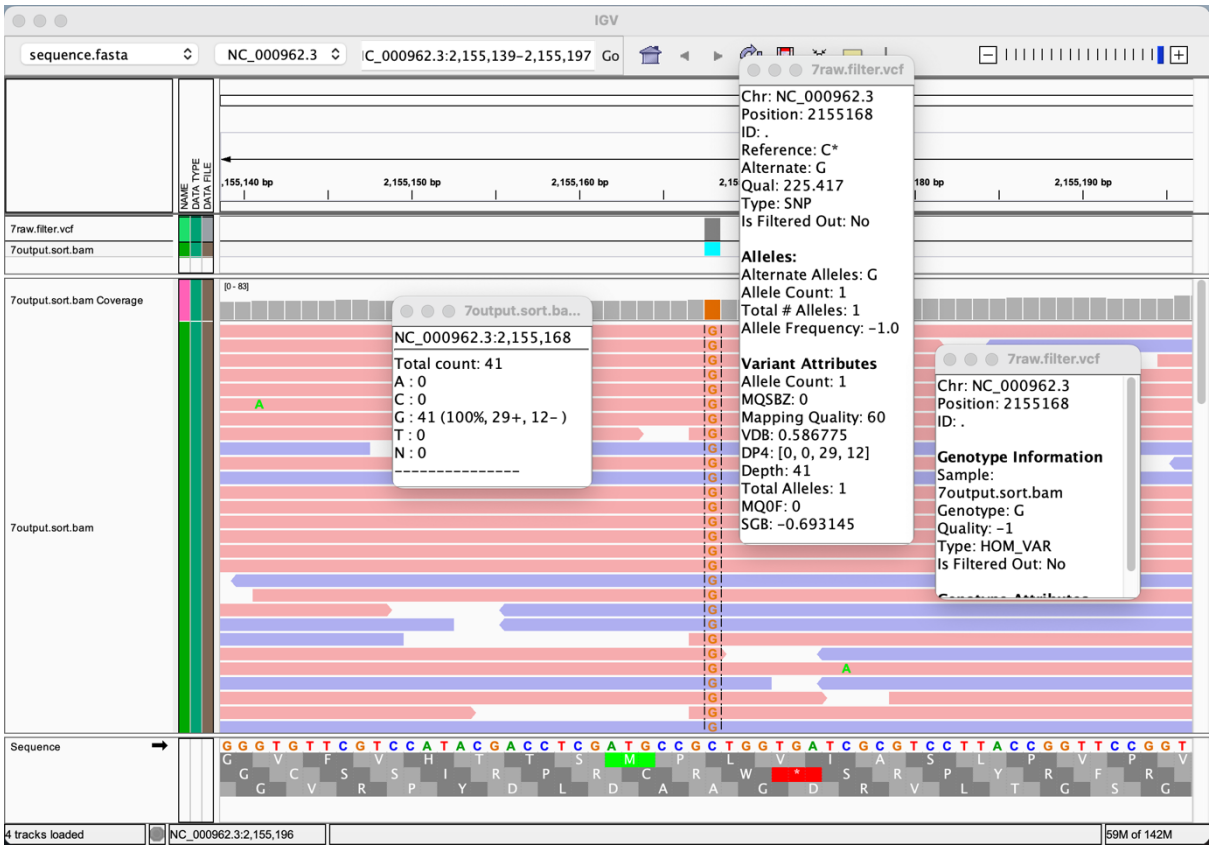


Fig 7.0.1

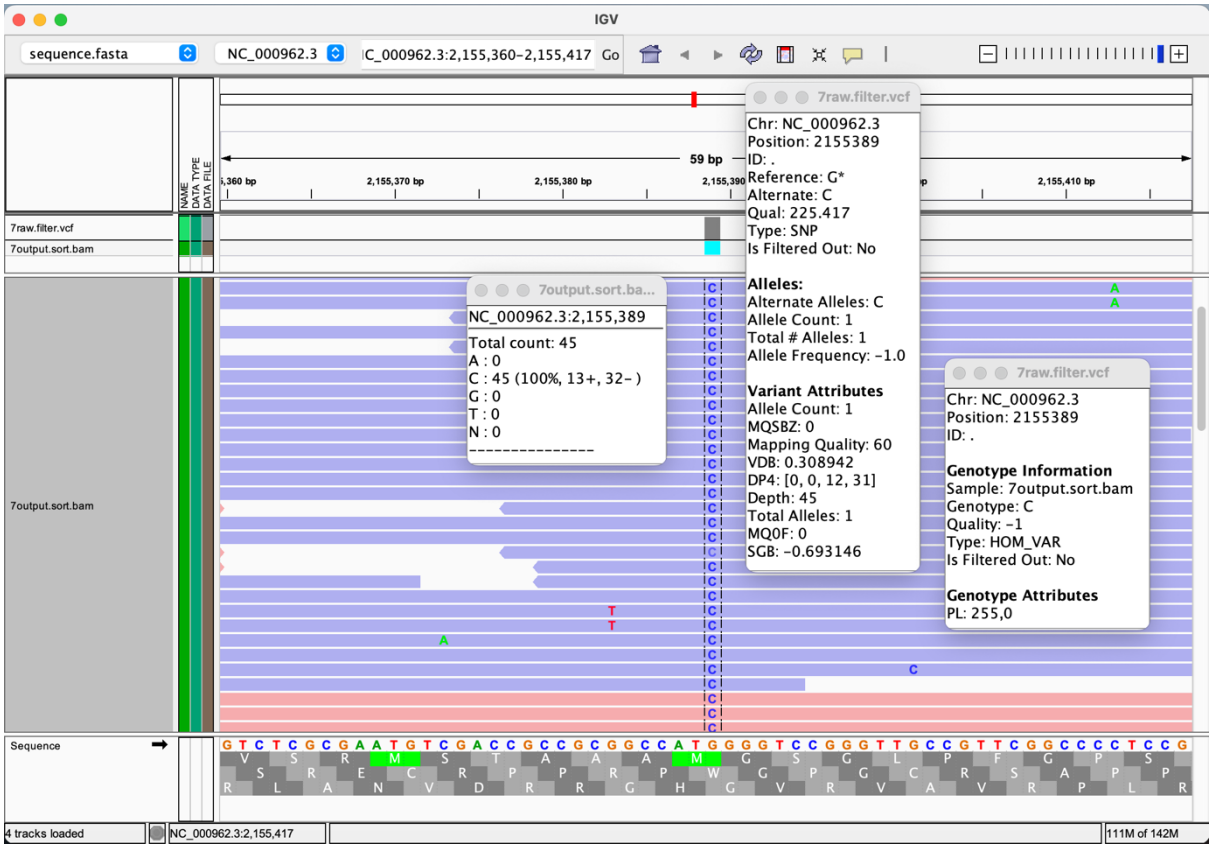


Fig 7.0.2

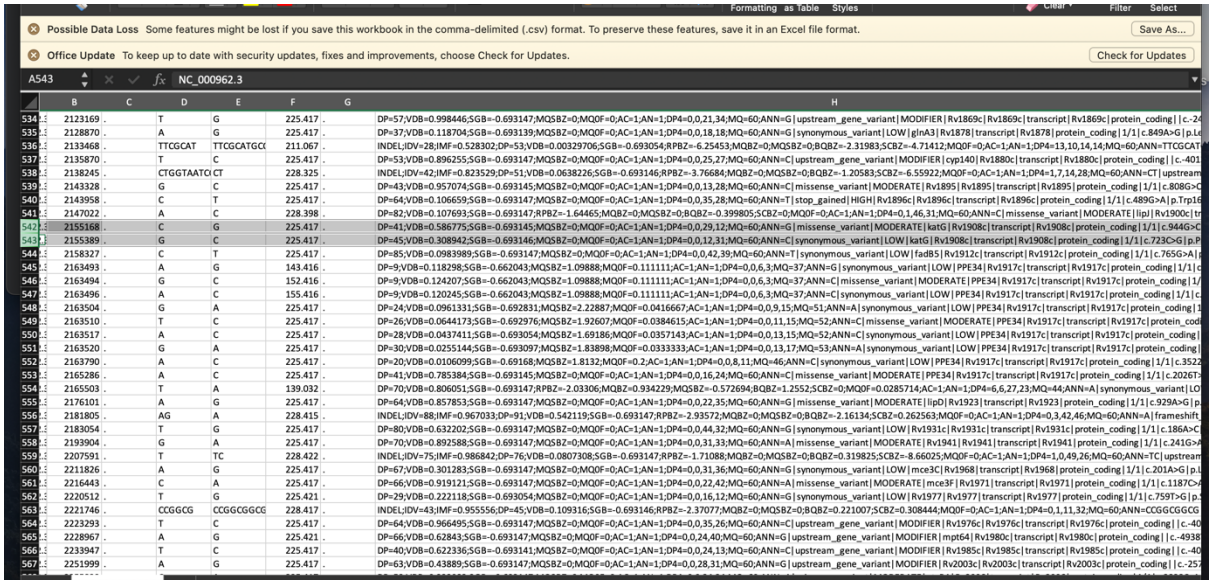
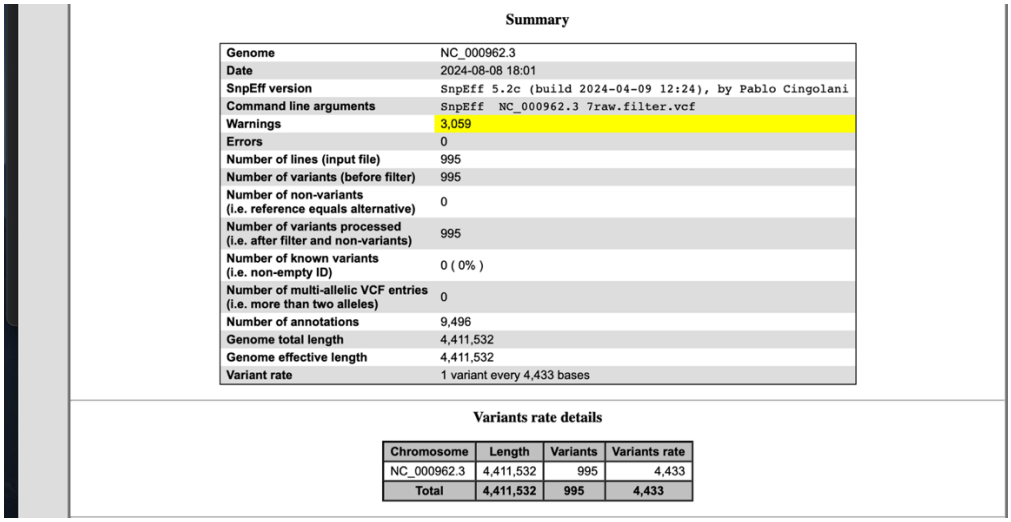


Fig 7.1 A VCF annotated with SnpEff.



Summary

Genome	NC_000962.3
Date	2024-08-08 18:01
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 7raw.filter.vcf
Warnings	3,059
Errors	0
Number of lines (input file)	995
Number of variants (before filter)	995
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	995
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,496
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,433 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	995	4,433
Total	**4,411,532**	**995**	**4,433**

Fig 7.2 SnpEff annotation summary

Number variants by type

Type	Total
SNP	908
MNP	0
INS	43
DEL	44
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	995

Number of effects by impact

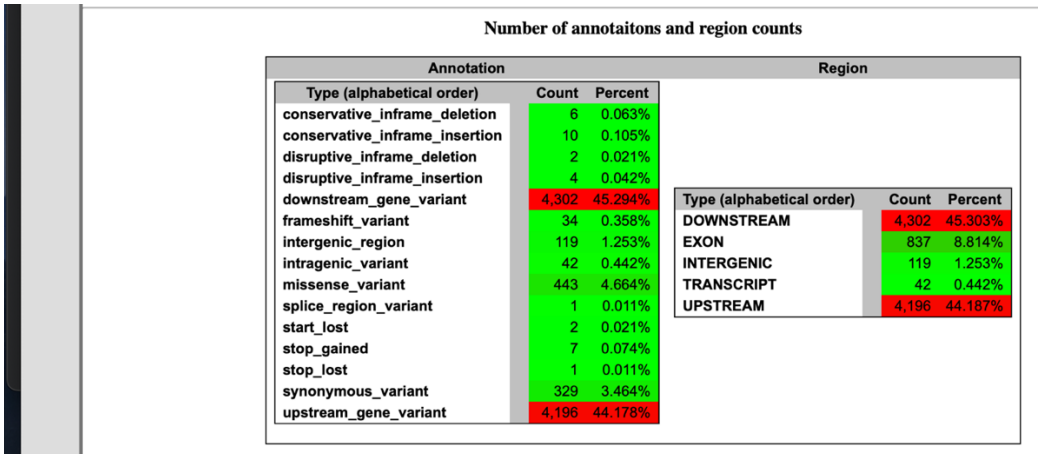
Type (alphabetical order)	Count	Percent
HIGH	43	0.453%
LOW	329	3.465%
MODERATE	465	4.897%
MODIFIER	8,659	91.186%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	445	56.978%
NONSENSE	7	0.896%
SILENT	329	42.125%

Missense / Silent ratio: 1.3526

Fig 7.3 Variant effect summary



Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	0.063%
conservative_inframe_insertion	10	0.105%
disruptive_inframe_deletion	2	0.021%
disruptive_inframe_insertion	4	0.042%
downstream_gene_variant	4,302	45.294%
frameshift_variant	34	0.358%
intergenic_region	119	1.253%
intragenic_variant	42	0.442%
missense_variant	443	4.664%
splice_region_variant	1	0.011%
start_lost	2	0.021%
stop_gained	7	0.074%
stop_lost	1	0.011%
synonymous_variant	329	3.464%
upstream_gene_variant	4,196	44.178%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,302	45.303%
EXON	837	8.814%
INTERGENIC	119	1.253%
TRANSCRIPT	42	0.442%
UPSTREAM	4,196	44.187%

Fig 7.4 The Number of effects by type and region

Fig 8.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 8

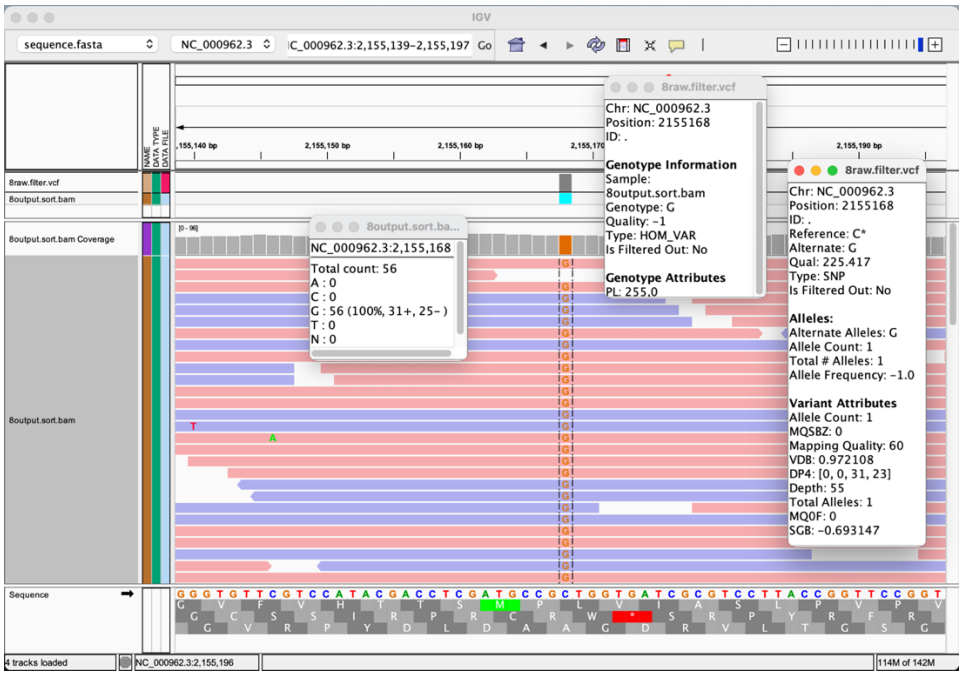


Fig 8.0.1

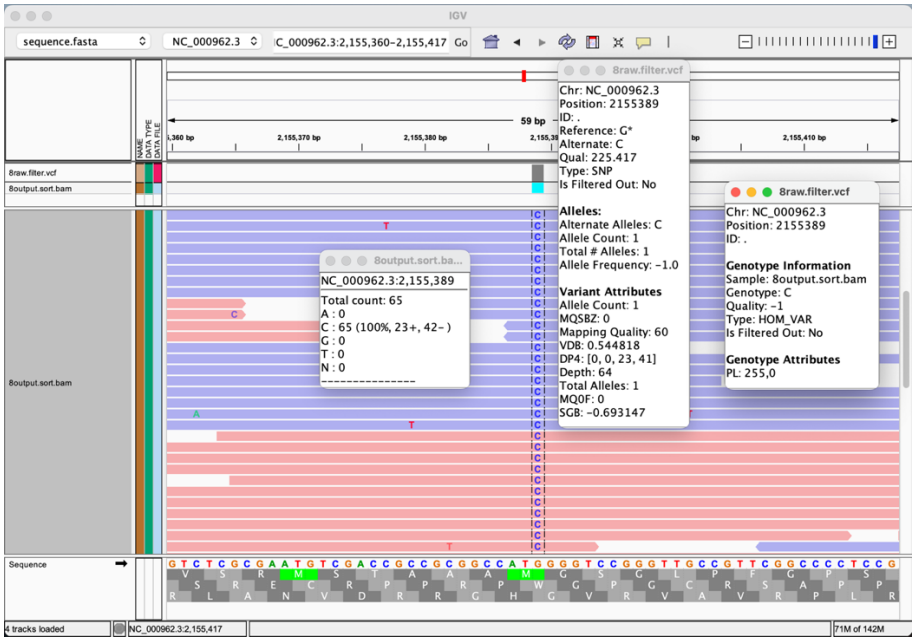


Fig 8.0.2

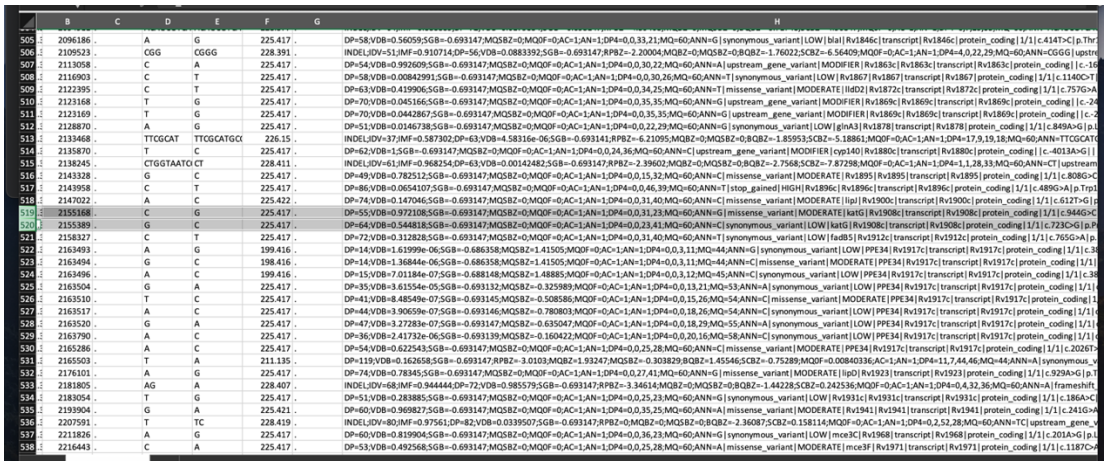
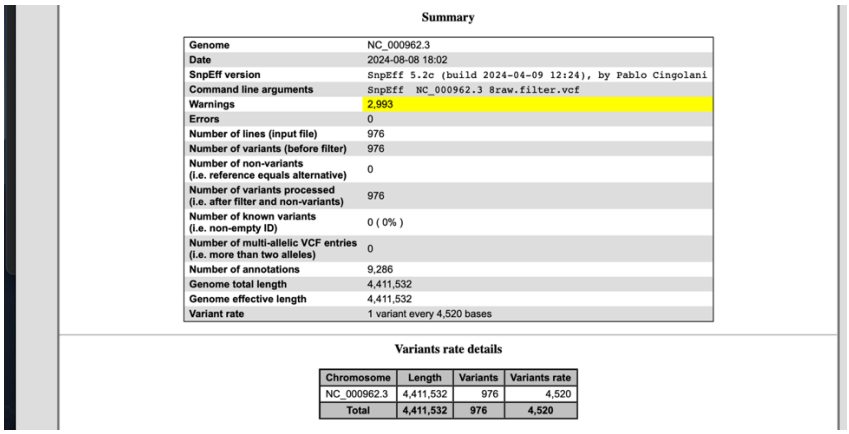


Fig 8.1 A VCF annotated with SnpEff.



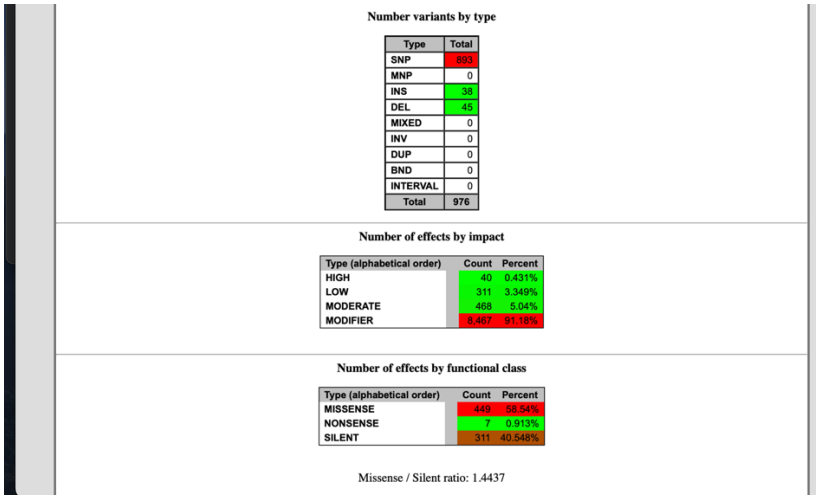
Summary

Genome	NC_000962.3
Date	2024-08-08 18:02
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 8raw.filter.vcf
Warnings	2,993
Errors	0
Number of lines (input file)	976
Number of variants (before filter)	976
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	976
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,286
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,520 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	976	4,520
Total	**4,411,532**	**976**	**4,520**

Fig 8.2 SnpEff annotation summary



Number variants by type

Type	Total
SNP	893
MNP	0
INS	38
DEL	45
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	976

Number of effects by impact

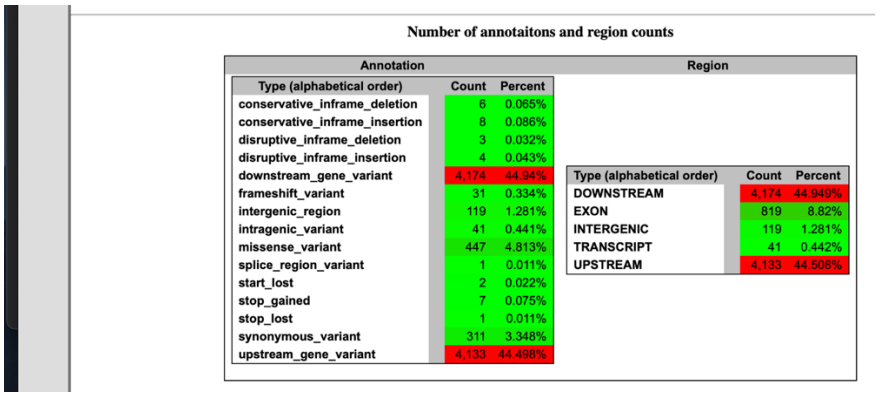
Type (alphabetical order)	Count	Percent
HIGH	40	0.431%
LOW	311	3.349%
MODERATE	468	5.04%
MODIFIER	8,467	91.18%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	449	58.54%
NONSENSE	7	0.913%
SILENT	311	40.548%

Missense / Silent ratio: 1.4437

Fig 8.3 Variant effect summary.



Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	0.065%
conservative_inframe_insertion	8	0.086%
disruptive_inframe_deletion	3	0.032%
disruptive_inframe_insertion	4	0.043%
downstream_gene_variant	4,174	44.94%
frameshift_variant	31	0.334%
intergenic_region	119	1.281%
intragenic_variant	41	0.441%
missense_variant	447	4.813%
splice_region_variant	1	0.011%
start_lost	2	0.022%
stop_gained	7	0.075%
stop_lost	1	0.011%
synonymous_variant	311	3.348%
upstream_gene_variant	4,133	44.498%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,174	44.949%
EXON	819	8.82%
INTERGENIC	119	1.281%
TRANSCRIPT	41	0.442%
UPSTREAM	4,133	44.508%

Fig 8.4 The Number of effects by type and region

Fig 9.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 9

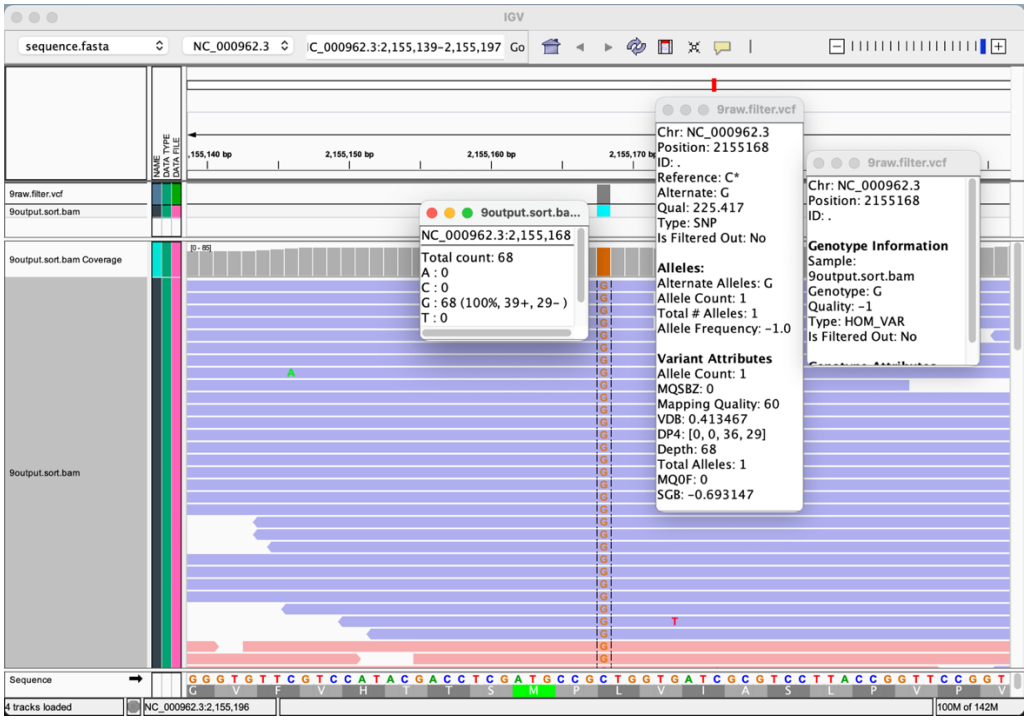


Fig 9.0.1

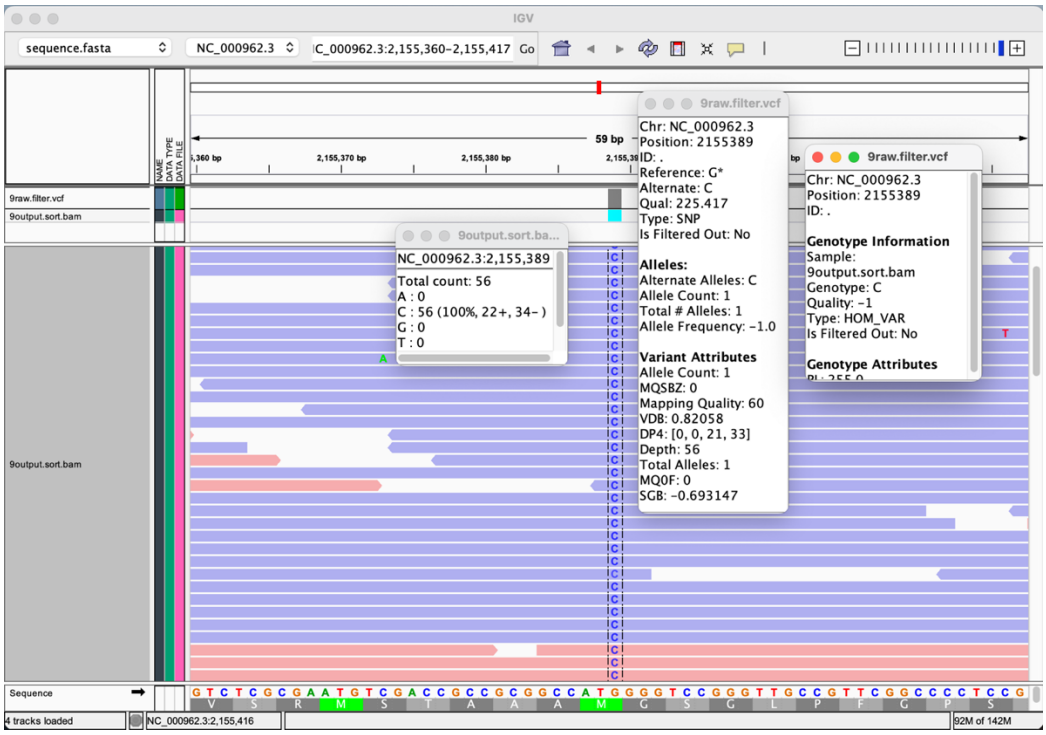
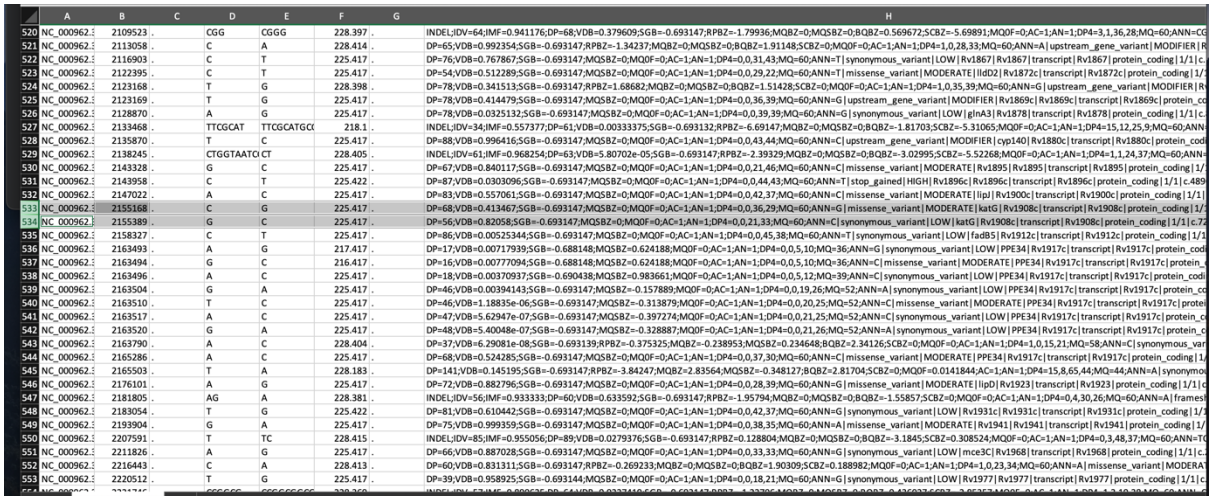
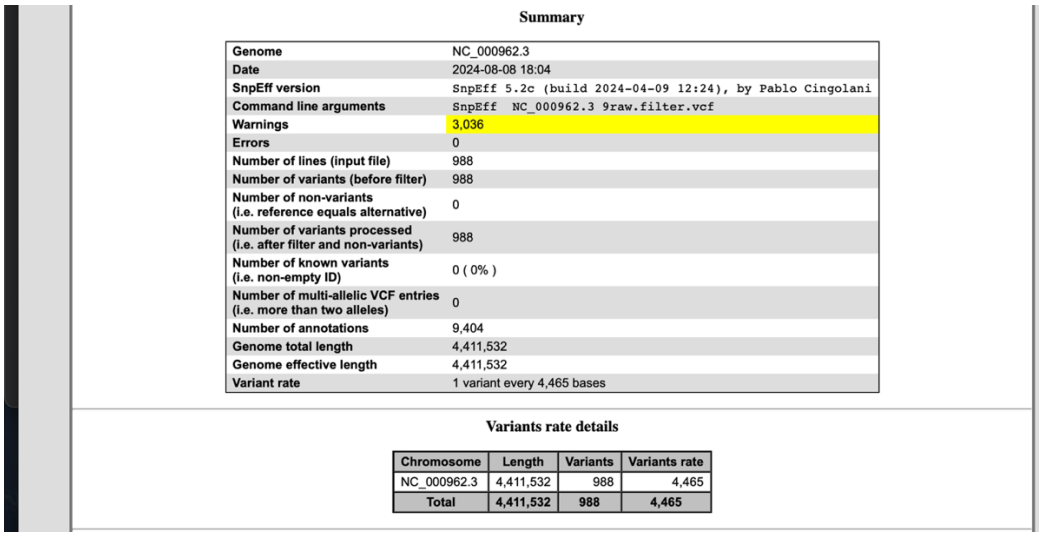


Fig 9.0.2



	A	B	C	D	E	F	G	H
520	NC_000962.3	2109523	.	CGG	CGGG	228.397	.	INDEL;IDV=64;IMF=0.941176;DP=68;VDB=0.379609;SGB=-0.693147;RPBZ=-1.79936;MQBZ=0;MQSBZ=0;BQBZ=0.569672;SCBZ=-5.69891;MQ0F=0;AC=1;AN=1;DP4=3,1,36,28;MQ=60;ANN=C|u
521	NC_000962.3	2113058	.	C	A	228.414	.	DP=65;VDB=0.992354;SGB=-0.693147;RPBZ=-1.34237;MQBZ=0;MQSBZ=0;BQBZ=1.91148;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=1,0,28,33;MQ=60;ANN=A|upstream_gene_variant|MODIFIER|R
522	NC_000962.3	2116903	.	C	T	225.417	.	DP=76;VDB=0.767867;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,31,43;MQ=60;ANN=T|synonymous_variant|LOW|Rv1867|Rv1867|transcript|Rv1867|protein_coding|1/1|c.A
523	NC_000962.3	2122395	.	C	T	225.417	.	DP=54;VDB=0.512289;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,29,22;MQ=60;ANN=T|missense_variant|MODERATE|IldD2|Rv1872c|transcript|Rv1872c|protein_coding|1/1
524	NC_000962.3	2123168	.	T	G	228.398	.	DP=78;VDB=0.341513;SGB=-0.693147;RPBZ=1.68682;MQBZ=0;MQSBZ=0;BQBZ=1.51428;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=1,0,35,39;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|R
525	NC_000962.3	2123169	.	T	G	225.417	.	DP=78;VDB=0.414479;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,36,39;MQ=60;ANN=G|upstream_gene_variant|MODIFIER|Rv1869c|Rv1869c|transcript|Rv1869c|protein_co
526	NC_000962.3	2128870	.	A	G	225.417	.	DP=78;VDB=0.0325132;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,39,39;MQ=60;ANN=G|synonymous_variant|LOW|glnA3|Rv1878|transcript|Rv1878|protein_coding|1/1|c.A
527	NC_000962.3	2133468	.	TTCGCAT	TTCGCATGC(	218.1	.	INDEL;IDV=34;IMF=0.557377;DP=61;VDB=0.00333375;SGB=-0.693132;RPBZ=-6.69147;MQBZ=0;MQSBZ=0;BQBZ=-1.81703;SCBZ=-5.31065;MQ0F=0;AC=1;AN=1;DP4=15,12,25,9;MQ=60;ANN
528	NC_000962.3	2135870	.	T	C	225.417	.	DP=88;VDB=0.996416;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,43,44;MQ=60;ANN=C|upstream_gene_variant|MODIFIER|cyp140|Rv1880c|transcript|Rv1880c|protein_codi
529	NC_000962.3	2138245	.	CTGGTAATC(	CT	228.405	.	INDEL;IDV=61;IMF=0.968254;DP=63;VDB=5.80702e-05;SGB=-0.693147;RPBZ=-2.39329;MQBZ=0;MQSBZ=0;BQBZ=-3.02995;SCBZ=-5.52268;MQ0F=0;AC=1;AN=1;DP4=1,1,24,37;MQ=60;ANN=
530	NC_000962.3	2143328	.	G	C	225.417	.	DP=67;VDB=0.840117;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,21,46;MQ=60;ANN=C|missense_variant|MODERATE|Rv1895|Rv1895|transcript|Rv1895|protein_coding|1/1
531	NC_000962.3	2143958	.	C	T	225.422	.	DP=87;VDB=0.0303096;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,44,43;MQ=60;ANN=T|stop_gained|HIGH|Rv1896c|Rv1896c|transcript|Rv1896c|protein_coding|1/1|c.489
532	NC_000962.3	2147022	.	A	C	225.417	.	DP=83;VDB=0.557061;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,42,37;MQ=60;ANN=C|missense_variant|MODERATE|lipJ|Rv1900c|transcript|Rv1900c|protein_coding|1/1|
533	NC_000962.3	2155168	.	C	G	225.417	.	DP=68;VDB=0.413467;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,36,29;MQ=60;ANN=G|missense_variant|MODERATE|katG|Rv1908c|transcript|Rv1908c|protein_coding|1/1
534	NC_000962.3	2155389	.	G	C	225.417	.	DP=56;VDB=0.82058;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,21,33;MQ=60;ANN=C|synonymous_variant|LOW|katG|Rv1908c|transcript|Rv1908c|protein_coding|1/1|c.72
535	NC_000962.3	2158327	.	C	T	225.417	.	DP=86;VDB=0.00525344;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,45,38;MQ=60;ANN=T|synonymous_variant|LOW|fadB5|Rv1912c|transcript|Rv1912c|protein_coding|1/1
536	NC_000962.3	2163493	.	A	G	217.417	.	DP=17;VDB=0.00717939;SGB=-0.688148;MQSBZ=0.624188;MQ0F=0;AC=1;AN=1;DP4=0,0,5,10;MQ=36;ANN=G|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_codi
537	NC_000962.3	2163494	.	G	C	216.417	.	DP=16;VDB=0.00777094;SGB=-0.688148;MQSBZ=0.624188;MQ0F=0;AC=1;AN=1;DP4=0,0,5,10;MQ=36;ANN=C|missense_variant|MODERATE|PPE34|Rv1917c|transcript|Rv1917c|protein_c
538	NC_000962.3	2163496	.	A	C	225.417	.	DP=18;VDB=0.00370937;SGB=-0.690438;MQSBZ=0.983661;MQ0F=0;AC=1;AN=1;DP4=0,0,5,12;MQ=39;ANN=C|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_codi
539	NC_000962.3	2163504	.	G	A	225.417	.	DP=46;VDB=0.00394143;SGB=-0.693147;MQSBZ=-0.157889;MQ0F=0;AC=1;AN=1;DP4=0,0,19,26;MQ=52;ANN=A|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_co
540	NC_000962.3	2163510	.	T	C	225.417	.	DP=46;VDB=1.18835e-06;SGB=-0.693147;MQSBZ=-0.313879;MQ0F=0;AC=1;AN=1;DP4=0,0,20,25;MQ=52;ANN=C|missense_variant|MODERATE|PPE34|Rv1917c|transcript|Rv1917c|protei
541	NC_000962.3	2163517	.	A	C	225.417	.	DP=47;VDB=5.62947e-07;SGB=-0.693147;MQSBZ=-0.397274;MQ0F=0;AC=1;AN=1;DP4=0,0,21,25;MQ=52;ANN=C|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_c
542	NC_000962.3	2163520	.	G	A	225.417	.	DP=48;VDB=5.40048e-07;SGB=-0.693147;MQSBZ=-0.328887;MQ0F=0;AC=1;AN=1;DP4=0,0,21,26;MQ=52;ANN=A|synonymous_variant|LOW|PPE34|Rv1917c|transcript|Rv1917c|protein_c
543	NC_000962.3	2163790	.	A	C	228.404	.	DP=37;VDB=6.29081e-08;SGB=-0.693139;RPBZ=-0.375325;MQBZ=-0.238953;MQSBZ=0.234648;BQBZ=2.34126;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=1,0,15,21;MQ=58;ANN=C|synonymous_var
544	NC_000962.3	2165286	.	A	C	225.417	.	DP=68;VDB=0.524285;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,37,30;MQ=60;ANN=C|missense_variant|MODERATE|PPE34|Rv1917c|transcript|Rv1917c|protein_coding|1/
545	NC_000962.3	2165503	.	T	A	228.183	.	DP=141;VDB=0.145195;SGB=-0.693147;RPBZ=-3.84247;MQBZ=2.83564;MQSBZ=-0.348127;BQBZ=2.81704;SCBZ=0;MQ0F=0.0141844;AC=1;AN=1;DP4=15,8,65,44;MQ=44;ANN=A|synonymou
546	NC_000962.3	2176101	.	A	G	225.417	.	DP=72;VDB=0.882796;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,28,39;MQ=60;ANN=G|missense_variant|MODERATE|lipD|Rv1923|transcript|Rv1923|protein_coding|1/1|c
547	NC_000962.3	2181805	.	AG	A	228.381	.	INDEL;IDV=56;IMF=0.933333;DP=60;VDB=0.633592;SGB=-0.693147;RPBZ=-1.95794;MQBZ=0;MQSBZ=0;BQBZ=-1.55857;SCBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,4,30,26;MQ=60;ANN=A|frames
548	NC_000962.3	2183054	.	T	G	225.422	.	DP=81;VDB=0.610442;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,42,37;MQ=60;ANN=G|synonymous_variant|LOW|Rv1931c|Rv1931c|transcript|Rv1931c|protein_coding|1/1
549	NC_000962.3	2193904	.	G	A	225.417	.	DP=75;VDB=0.999359;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,38,35;MQ=60;ANN=A|missense_variant|MODERATE|Rv1941|Rv1941|transcript|Rv1941|protein_coding|1/1
550	NC_000962.3	2207591	.	T	TC	228.415	.	INDEL;IDV=85;IMF=0.955056;DP=89;VDB=0.0279376;SGB=-0.693147;RPBZ=0.128804;MQBZ=0;MQSBZ=0;BQBZ=-3.1845;SCBZ=0.308524;MQ0F=0;AC=1;AN=1;DP4=0,3,48,37;MQ=60;ANN=T|C
551	NC_000962.3	2211826	.	A	G	225.417	.	DP=66;VDB=0.887028;SGB=-0.693147;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,33,33;MQ=60;ANN=G|synonymous_variant|LOW|mce3C|Rv1968|transcript|Rv1968|protein_coding|1/1|c.A
552	NC_000962.3	2216443	.	C	A	228.413	.	DP=60;VDB=0.831311;SGB=-0.693147;RPBZ=-0.269233;MQBZ=0;MQSBZ=0;BQBZ=1.90309;SCBZ=0.188982;MQ0F=0;AC=1;AN=1;DP4=1,0,23,34;MQ=60;ANN=A|missense_variant|MODERAT
553	NC_000962.3	2220512	.	T	G	225.417	.	DP=39;VDB=0.958925;SGB=-0.693144;MQSBZ=0;MQ0F=0;AC=1;AN=1;DP4=0,0,18,21;MQ=60;ANN=G|synonymous_variant|LOW|Rv1977|Rv1977|transcript|Rv1977|protein_coding|1/1|c.A

Fig 9.1 A VCF annotated with SnpEff.



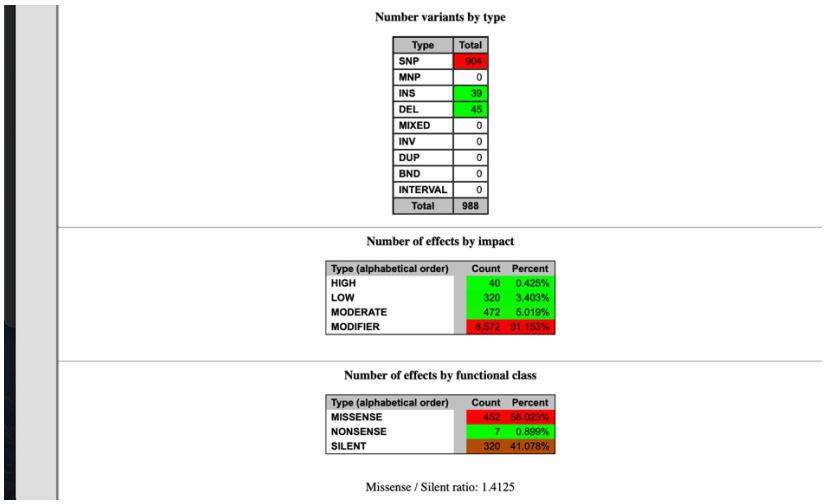
Summary

Genome	NC_000962.3
Date	2024-08-08 18:04
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 9raw.filter.vcf
Warnings	3,036
Errors	0
Number of lines (input file)	988
Number of variants (before filter)	988
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	988
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,404
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,465 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	988	4,465
Total	**4,411,532**	**988**	**4,465**

Fig 9.2 SnpEff annotation summary



Number variants by type

Type	Total
SNP	904
MNP	0
INS	39
DEL	45
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	988

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	40	0.425%
LOW	320	3.403%
MODERATE	472	5.019%
MODIFIER	8,572	91.153%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	452	58.023%
NONSENSE	7	0.899%
SILENT	320	41.078%

Missense / Silent ratio: 1.4125

Fig 9.3 Variant effect summary.

Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	0.064%
conservative_inframe_insertion	9	0.096%
disruptive_inframe_deletion	3	0.032%
disruptive_inframe_insertion	4	0.043%
downstream_gene_variant	4,227	44.939%
frameshift_variant	31	0.33%
intergenic_region	118	1.255%
intragenic_variant	41	0.436%
missense_variant	450	4.784%
splice_region_variant	1	0.011%
start_lost	2	0.021%
stop_gained	7	0.074%
stop_lost	1	0.011%
synonymous_variant	320	3.402%
upstream_gene_variant	4,186	44.504%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,227	44.949%
EXON	832	8.847%
INTERGENIC	118	1.255%
TRANSCRIPT	41	0.436%
UPSTREAM	4,186	44.513%

Fig 9.4 The Number of effects by type and region

Fig 10.0 IGV visualisation of a *Mycobacterium tuberculosis* H37RV variant site (with respect to the NC_000962.3 reference) in a simulated sample 10

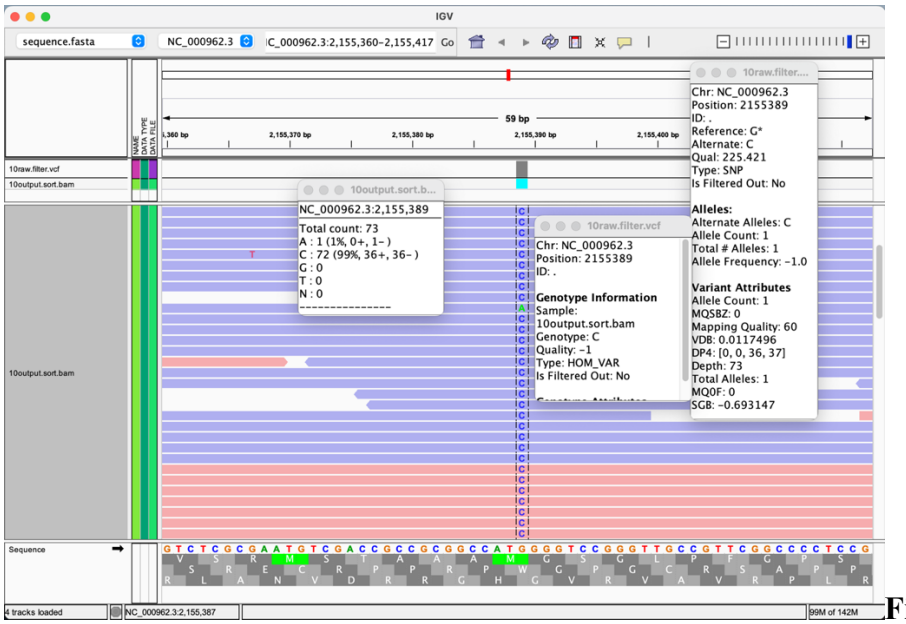
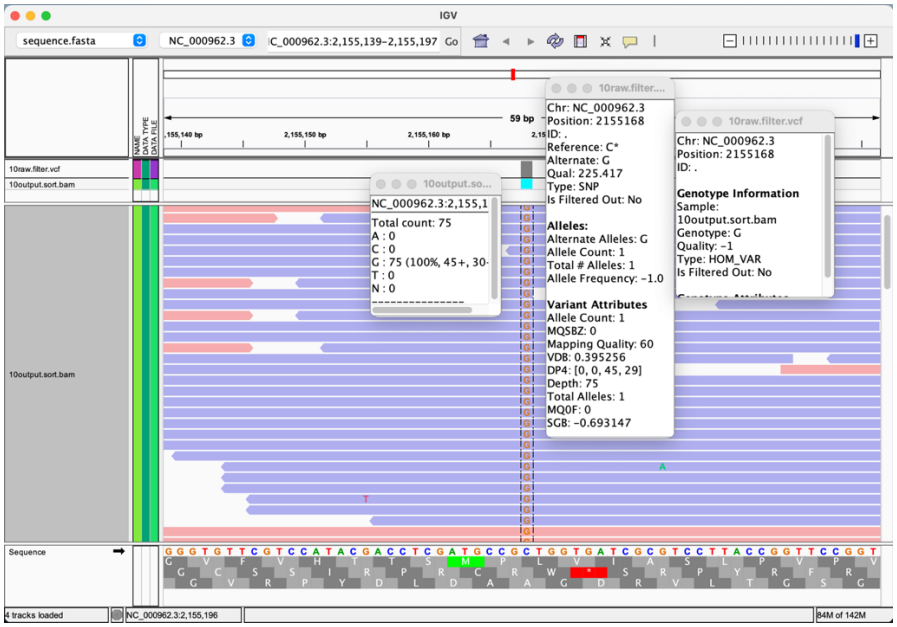
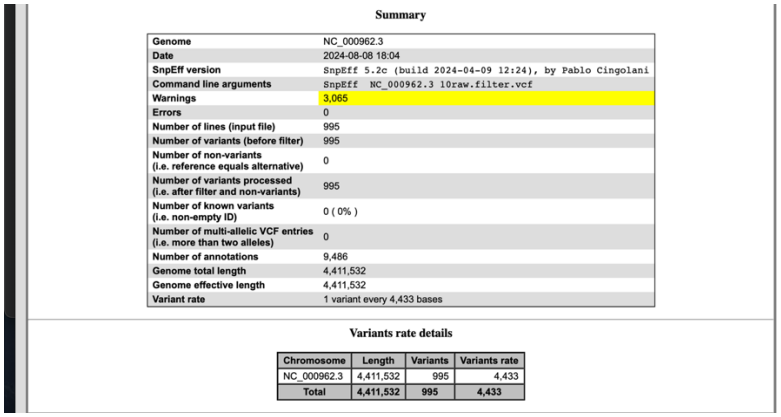


Fig 10.0.1



10.0.2

Fig 10.1 A VCF annotated with SnpEff.



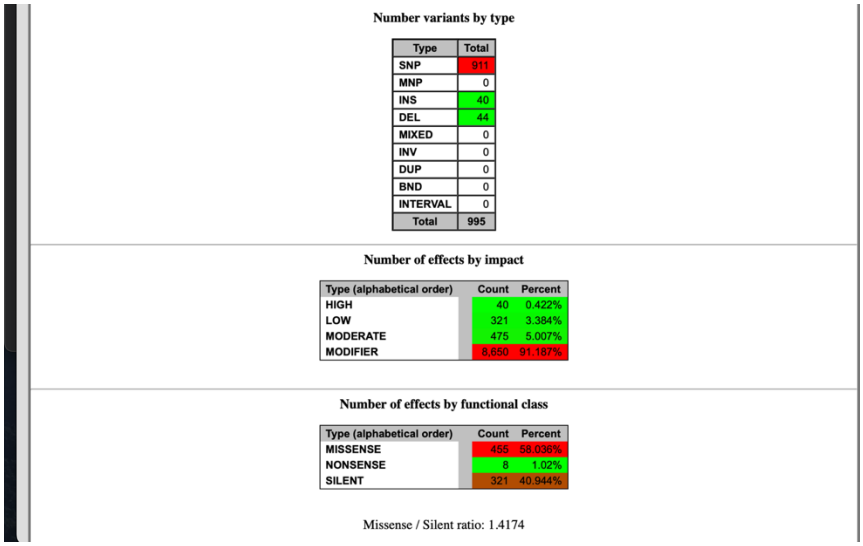
Summary

Genome	NC_000962.3
Date	2024-08-08 18:04
SnpEff version	SnpEff 5.2c (build 2024-04-09 12:24), by Pablo Cingolani
Command line arguments	SnpEff NC_000962.3 10raw.filter.vcf
Warnings	3,065
Errors	0
Number of lines (input file)	995
Number of variants (before filter)	995
Number of non-variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	995
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	0
Number of annotations	9,486
Genome total length	4,411,532
Genome effective length	4,411,532
Variant rate	1 variant every 4,433 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
NC_000962.3	4,411,532	995	4,433
Total	**4,411,532**	**995**	**4,433**

Fig 10.2 SnpEff annotation summary



Number variants by type

Type	Total
SNP	911
MNP	0
INS	40
DEL	44
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	995

Number of effects by impact

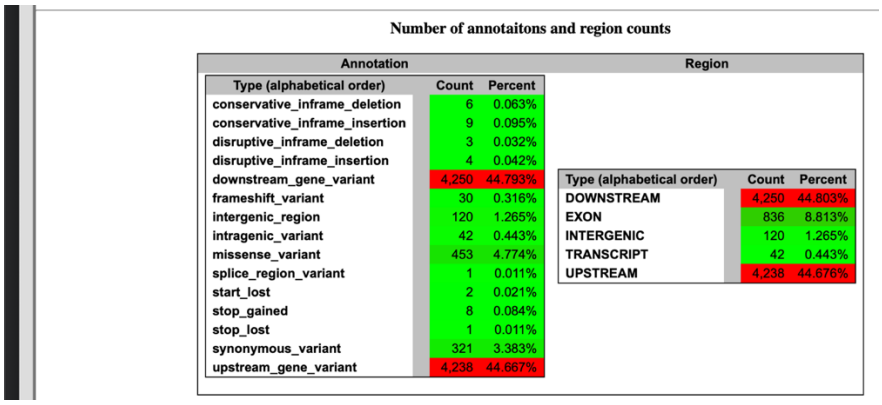
Type (alphabetical order)	Count	Percent
HIGH	40	0.422%
LOW	321	3.384%
MODERATE	475	5.007%
MODIFIER	8,650	91.187%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	455	58.036%
NONSENSE	8	1.02%
SILENT	321	40.944%

Missense / Silent ratio: 1.4174

Fig 10.3 Variant effect summary.



Number of annotaitons and region counts

Annotation

Type (alphabetical order)	Count	Percent
conservative_inframe_deletion	6	0.063%
conservative_inframe_insertion	9	0.095%
disruptive_inframe_deletion	3	0.032%
disruptive_inframe_insertion	4	0.042%
downstream_gene_variant	4,250	44.793%
frameshift_variant	30	0.316%
intergenic_region	120	1.265%
intragenic_variant	42	0.443%
missense_variant	453	4.774%
splice_region_variant	1	0.011%
start_lost	2	0.021%
stop_gained	8	0.084%
stop_lost	1	0.011%
synonymous_variant	321	3.383%
upstream_gene_variant	4,238	44.667%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	4,250	44.803%
EXON	836	8.813%
INTERGENIC	120	1.265%
TRANSCRIPT	42	0.443%
UPSTREAM	4,238	44.676%

Fig 10.4 The Number of effects by type and region

INTERPRETATIONS/DISCUSSION

katG mutations are prevalent in isoniazid resistance. Across all samples; IGV visualisation of mutations in *katG* sequences (Position: 2153889 - 2156111) and IGV browser gives colours to interesting events. Variants that affect a gene's function are the most harmful and

detrimental, yet they can occur anywhere along the genome sequence. A variation or mutation that impacts the gene regulatory region has the potential to repress, inhibit, or activate or deactivate a gene.

At Position: 2153889 – 2156111; Sample 1,3 and 4 showed no variants, While Variants were detected in Sample 2, 5,6,7,8,9, and 10, particularly homozygous variants, where NC_000962.3 is the reference sequence in Genbank

For Sample 2 and Sample 5; The variant NC_000962.3:c.944G>C p.(Ser315Thr), indicating a single nucleotide polymorphism(SNP), where a guanine (G) nucleotide at position 944 has been replaced by a cytosine (C) nucleotide. The p.(Ser315Thr) indicates the amino acid change caused by the SNP, In this case, a serine (Ser) amino acid at position 315 has been replaced by a threonine (Thr) amino acid. This is a missense mutation, as it results in a change in a single amino acid , causing Isoniazid resistance.

Sample 6, 7,8,9 and 10 produced two variants; The variant NC_000962.3:c.944G>C p.(Ser315Thr) and The variant NC_000962.3:c.723C>G p.(Pro241Pro). The former being a missense variation and the Latter being a synonymous variant (the variation doesn't alter the encoded amino acid).

The mutation occurs at a coding region causing a structural and conformational change and preventing it from binding to the target resulting hence resulting in Isoniazid deactivation and resistance. [Larsen et al, 2002].

Isoniazid resistance *in Mycobacterium tuberculosis* can arise from mutations in various genes and alterations in biological pathways. Understanding these pathways is crucial for developing new strategies to combat drug-resistant Tuberculosis. Some key pathways associated with Isoniazid resistance include; Pharamacokinetics/Pharmacodynamics factors, Drug efflux systems, Mycobacterial Cell wall, Drug misuse, generation or use of alternative biochemical pathways.

CONCLUSION

Mutations were identified in 7 samples; with 5 samples having two variants. Isoniazid resistance is due to the mutation of katG gene --Ser315-Thr--. These mutations reduce its activity, making it difficult to treat Tuberculosis.

REFERENCES

1. James T Robinson, Helga Thorvaldsdottir, Douglass Turner, Jill P Mesirov, igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV), Bioinformatics, Volume 39, Issue 1, January 2023, btac830, https://doi.org/10.1093/bioinformatics/btac830

2. HackBio. (2023). Project Section: Searching for Mutations Leading to Isoniazid Resistance in TB (HackBio) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10426436

3. Hamid D. Ismail. Bioinformatics, A Practical Guide to Next Generation Sequencing Data Analysis. First edition published 2023 by CRC Press. Pages 110-160

4. Larsen, M. H., Vilchèze, C., Kremer, L., Besra, G. S., Parsons, L., Salfinger, M., Heifets, L., Hazbon, M. H., Alland, D., Sacchettini, J. C., & Jacobs, W. R., Jr (2002). Overexpression of inhA, but not kasA, confers resistance to isoniazid and ethionamide in Mycobacterium smegmatis, M. bovis BCG and M. tuberculosis. Molecular microbiology, 46(2), 453–466. https://doi.org/10.1046/j.1365-2958.2002.03162.x