IGSR - Variant calling

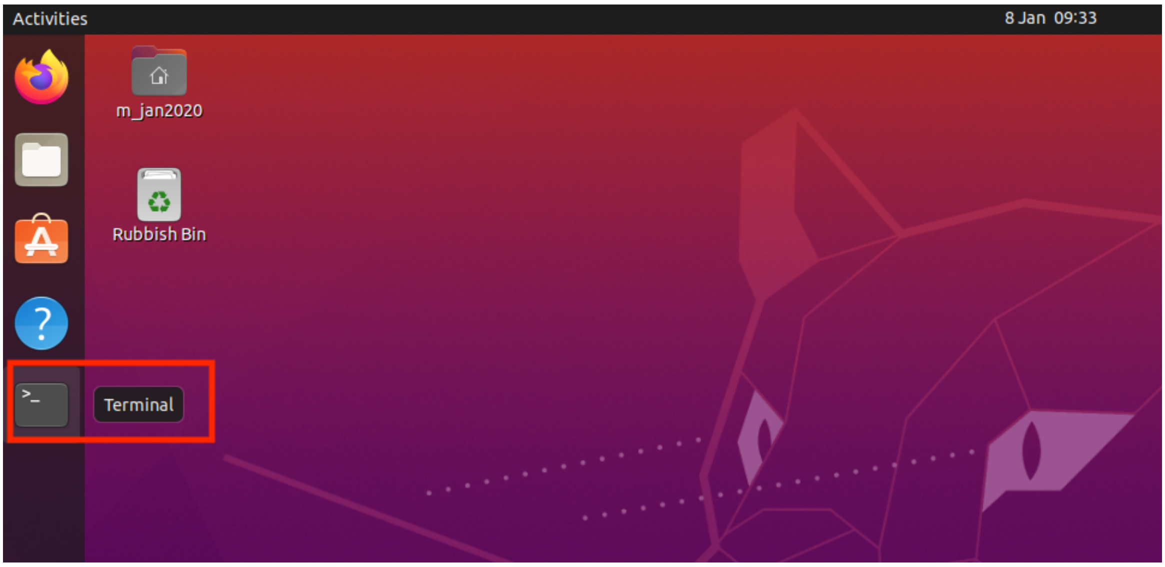
In the first part of the course, we used the sequencing data generated in the 3000 Rice Genomes project for one particular sample with ENA accession id SAMEA2569438, to generate an analysis-ready BAM alignment file.
This part of the course starts from this alignment file and will generate a VCF file containing the germ-line variants identified in chromosome 10.

Log in the Ubuntu virtual machine

```
    user: m_jan2020
    pwd: m_jan2020
```

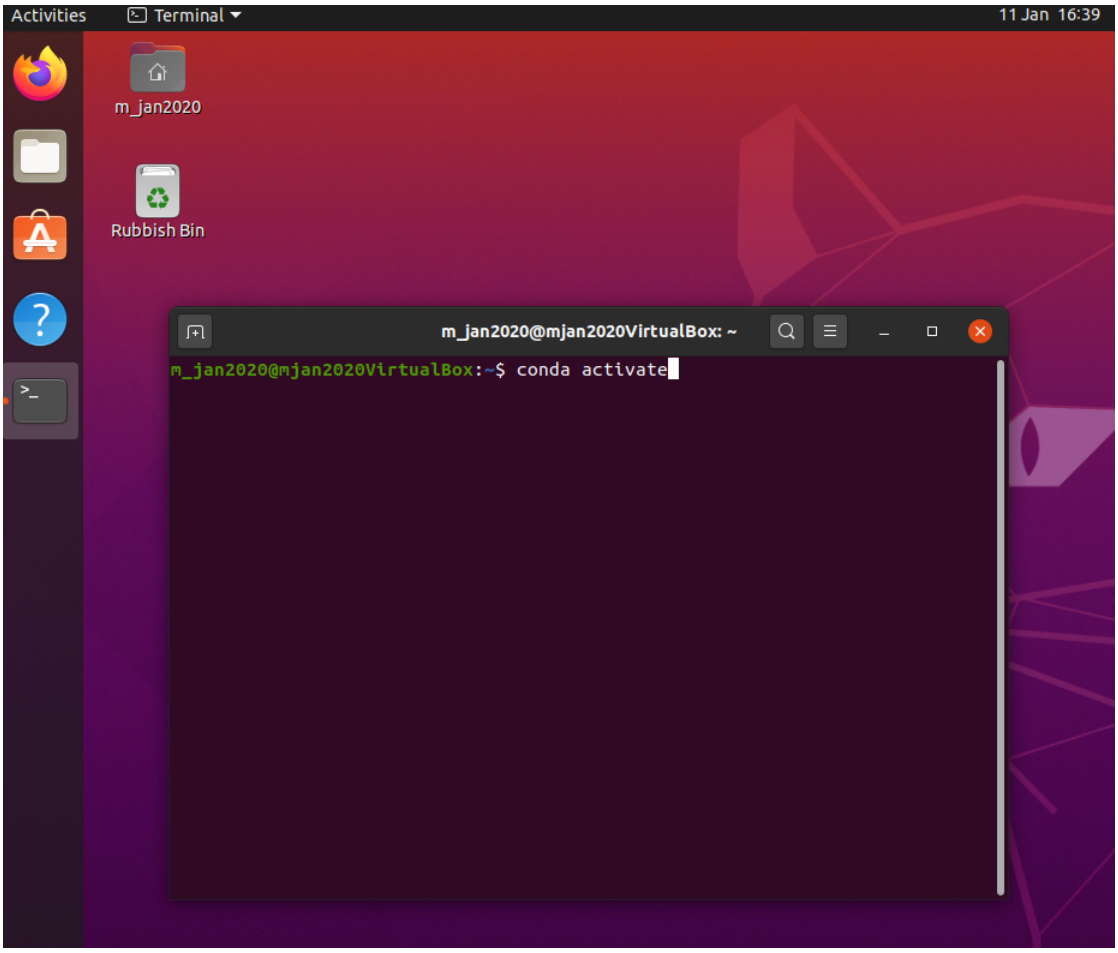
Open a Terminal window in the Ubuntu virtual machine

A terminal in Linux is an interface we can use to enter text commands, it will be the interface we will use to run most of the bioinformatics tools shown in this course. To open a terminal click on the `Terminal` icon you see in your screen:



Activate the conda environment used in this course

Conda is a package and environment management system that we have used in this course to install all the bioinformatics programs.
To start using Conda you need first to activate the course environment by entering the following command in the terminal window you have just opened:



```
m_jan2020@mjan2020VirtualBox:~$ conda activate
```

Unix commands used in this course

Most of the bioinformatics tools used for genomic data analysis run in Unix/Linux, so it is recommended to have a basic knowledge of the commands used to move around the different directories in your system. You should also know how to list the contents of a directory or how to print the contents of a given text file.
Here are some of the basic commands we will use during this course:

- Print the current working directory with the pwd command

```
m_jan2020@mjan2020VirtualBox:~$ pwd
/home/m_jan2020
```

- Change directory (cd)

```
# go to the alignment dir
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/alignment/

# move up one folder
m_jan2020@mjan2020VirtualBox:~$ cd ../

# now, check where you are
m_jan2020@mjan2020VirtualBox:~$ pwd
/home/m_jan2020/course

# get back to original location
m_jan2020@mjan2020VirtualBox:~$ cd
m_jan2020@mjan2020VirtualBox:~$ pwd
/home/m_jan2020
```

- Listing the directory contents (ls)

```
# go to the following directory
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/data

# basic listing
m_jan2020@mjan2020VirtualBox:~/course/data$ ls
SAMEA2569438.chr10_1.fastq.gz  SAMEA2569438.chr10_2.fastq.gz

# ls with -l (long-format) and -h (human readable)
m_jan2020@mjan2020VirtualBox:~/course/data$ ls -ls
total 23M
```

```
-rw-rw-r-- 1 m_jan2020 m_jan2020 11M Nov 26 17:43 SAMEA2569438.chr10_1.fastq.gz
-rw-rw-r-- 1 m_jan2020 m_jan2020 12M Nov 26 17:43 SAMEA2569438.chr10_2.fastq.gz
```

- Open a file to see its contents using the (`less`) UNIX command:

```
# go to the following directory
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/data

# enter the `less` command followed by the file name you want to open:
m_jan2020@mjan2020VirtualBox:~/course/data$ less SAMEA2569438.chr10_2.fastq.gz
# press Ctrl+F to go forward one window
# press Ctrl+B to go back one window
# press 'q' if you want to exit
```

- Now, make (`less`) to print out the line number information by doing:

```
m_jan2020@mjan2020VirtualBox:~/course/data$ less -N SAMEA2569438.chr10_2.fastq.gz
```

- Output redirection:
  In Linux, output redirection means that we can redirect the STDOUT of a given command to a file. For this, we use the (`>`) symbol.
  Example:

```
# go to the following directory
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/data

# the output of 'ls' will be redirected to the file named 'listing'
m_jan2020@mjan2020VirtualBox:~/course/data$ ls -l > listing

# examine the contents of 'listing'
 m_jan2020@mjan2020VirtualBox:~/course/data$ less listing
```

- Piping
  It is a form of redirection that transfers the standard output of one command/program to another command/program for further processing. The different commands in the pipe are connected using the pipe character (`|`). In this way we can combine 2 or more commands. Pipes are unidirectional, i.e. data flows from left to right.
  Examples:

```
# go to the following directory
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/data

# count the number of files in a directory
m_jan2020@mjan2020VirtualBox:~$ ls | wc -l
```

Variant calling

Variant calling is the process that identifies variants from sequence data (see). It begins with the alignment of the sequencing data present in the `FASTQ` files and that has been explained in the first section of the course. Then, the next step in the analysis consists on using a variant discovery tool to identify the germline variants.

There are multiple variant calling tools available, the ones we have the most experience with in our group are SAMTools mpileup, the GATK suite and Freebayes. In this course we are going to use FreeBayes, since it is sensitive, accurate and relatively easy to use.

Freebayes

Freebayes is a haplotype-based variant detector, that uses a joint genotyping method capable of reporting variants in a single sample or in a cohort of samples. It will also be able to detect SNPs (single nucleotide polymorphisms), indels (short insertions and deletions) and MNPs (multi-nucleotide polymorphisms).

Reference Genome

Freebayes requires the reference sequence in the `FASTA` format. In this section of the course we are going to use the same chromosome 10 sequence extracted from the *Oryza_sativa* (rice) genome that we used for the alignment section of the course.

Using Freebayes

To run Freebayes, you will need to specify the ploidy of the genome being analysed, the location of the FASTA reference sequence used for the alignment, and the location of the analysis-ready BAM generated in the first section of the course. Once you have this information, you are ready to run Freebayes.
To do this, go to the directory where the program is going to be run:

```
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/vcalling
```

And enter:

```
    m_jan2020@mjan2020VirtualBox:~$ freebayes -f /home/m_jan2020/course/reference/Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa
/home/m_jan2020/course/alignment/postprocessing/SAMEA2569438.chr10.sorted.reheaded.mark_duplicates.bam --ploidy 2 | bgzip -c >
SAMEA2569438.chr10.vcf.gz
```

This command pipes the Freebayes output to `bgzip`, which is a special compression/decompression program included with Samtools. It is preferable to compress the VCF to decrease the file size and also to use some of the `BCFTools` commands that are discussed later in this course.

Understanding the output VCF

After running Freebayes, you will get a compressed VCF file named `SAMEA2569438.chr10.vcf.gz` containing the identified variants. The full VCF specification with an explanation of each of the components of the file can be found here.

The most relevant sections for us are the meta-information lines (prefixed with `##`), the header line (prefixed with `#`) and then the data lines containing information about the variants. These data lines will contain the following text fields separated by tabs:

Col	Field	Brief description
1	CHROM	Chromosome where the genetic variant was found
2	POS	Position in the chromosome where the genetic variant was found
3	ID	SNP id
4	REF	Reference allele
5	ALT	Alternate allele
6	QUAL	Variant quality
7	FILTER	Filter string (`PASS` i.e. passed all filters)
8	INFO	Semicolon-separated series of variant additional information fields
9	GENOTYPE	Genotype information (if present)

Exploring the VCF file using BCFTools

BCFTools is a set of tools written in C that are quite efficient to manipulate files in the `VCF` format.
In this section we are going to see some of the most useful commmands to manipulate the `VCF` file we have just generated.

So first go to the directory where you ran Freebayes:

```
m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/vcalling
```

- Print the header section

  ```
  m_jan2020@mjan2020VirtualBox:~$ bcftools view -h SAMEA2569438.chr10.vcf.gz
  ```

You should get something similar to:

```
   ##fileformat=VCFv4.1
   ##FILTER=<ID=PASS,Description="All filters passed">
   ##fileDate=20201204
   ##source=freeBayes v0.9.21
   ##reference=/home/m_jan2020/course/reference/Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa
   ##phasing=none
   ##commandline="freebayes -f /home/m_jan2020/course/reference/Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa
/home/m_jan2020/course/alignment/postprocessing/SAMEA2569438.chr10.sorted.reheaded.mark_duplicates.bam --ploidy 2"
   ##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
   ##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
   ##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the locus; bases in reads overlapping / bases in
haplotype">
   ##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
   ##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
   ##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
   ##INFO=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded
fractionally">
   ##INFO=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations, with partial observations recorded
fractionally">
   ##INFO=<ID=PRO,Number=1,Type=Float,Description="Reference allele observation count, with partial observations recorded
fractionally">
   ##INFO=<ID=PAO,Number=A,Type=Float,Description="Alternate allele observations, with partial observations recorded fractionally">
   ##INFO=<ID=QR,Number=1,Type=Integer,Description="Reference allele quality sum in phred">
   ##INFO=<ID=QA,Number=A,Type=Integer,Description="Alternate allele quality sum in phred">
   ##INFO=<ID=PQR,Number=1,Type=Float,Description="Reference allele quality sum in phred for partial observations">
   ##INFO=<ID=PQA,Number=A,Type=Float,Description="Alternate allele quality sum in phred for partial observations">
   ##INFO=<ID=SRF,Number=1,Type=Integer,Description="Number of reference observations on the forward strand">
   ##INFO=<ID=SRR,Number=1,Type=Integer,Description="Number of reference observations on the reverse strand">
   ##INFO=<ID=SAF,Number=A,Type=Integer,Description="Number of alternate observations on the forward strand">
   ##INFO=<ID=SAR,Number=A,Type=Integer,Description="Number of alternate observations on the reverse strand">
   ##INFO=<ID=SRP,Number=1,Type=Float,Description="Strand balance probability for the reference allele: Phred-scaled upper-bounds
estimate of the probability of observing the deviation between SRF and SRR given E(SRF/SRR) ~ 0.5, derived using Hoeffding's
inequality">
   ##INFO=<ID=SAP,Number=A,Type=Float,Description="Strand balance probability for the alternate allele: Phred-scaled upper-bounds
estimate of the probability of observing the deviation between SAF and SAR given E(SAF/SAR) ~ 0.5, derived using Hoeffding's
inequality">
```

```
   ##INFO=<ID=AB,Number=A,Type=Float,Description="Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference allele to all reads, considering only reads from individuals called as heterozygous">
   ##INFO=<ID=ABP,Number=A,Type=Float,Description="Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observing the deviation between ABR and ABA given E(ABR/ABA) ~ 0.5, derived using Hoeffding's inequality">
   ##INFO=<ID=RUN,Number=A,Type=Integer,Description="Run length: the number of consecutive repeats of the alternate allele in the reference genome">
   ##INFO=<ID=RPP,Number=A,Type=Float,Description="Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between RPL and RPR given E(RPL/RPR) ~ 0.5, derived using Hoeffding's inequality">
   ##INFO=<ID=RPPR,Number=1,Type=Float,Description="Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of observing the deviation between RPL and RPR given E(RPL/RPR) ~ 0.5, derived using Hoeffding's inequality">
   ##INFO=<ID=RPL,Number=A,Type=Float,Description="Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele">
   ##INFO=<ID=RPR,Number=A,Type=Float,Description="Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele">
   ##INFO=<ID=EPP,Number=A,Type=Float,Description="End Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between EL and ER given E(EL/ER) ~ 0.5, derived using Hoeffding's inequality">
   ##INFO=<ID=EPPR,Number=1,Type=Float,Description="End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of observing the deviation between EL and ER given E(EL/ER) ~ 0.5, derived using Hoeffding's inequality">
   ##INFO=<ID=DPRA,Number=A,Type=Float,Description="Alternate allele depth ratio.  Ratio between depth in samples with each called alternate allele and those without.">
   ##INFO=<ID=ODDS,Number=1,Type=Float,Description="The log odds ratio of the best genotype combination to the second-best.">
   ##INFO=<ID=GTI,Number=1,Type=Integer,Description="Number of genotyping iterations required to reach convergence or bailout.">
   ##INFO=<ID=TYPE,Number=A,Type=String,Description="The type of allele, either snp, mnp, ins, del, or complex.">
   ##INFO=<ID=CIGAR,Number=A,Type=String,Description="The extended CIGAR representation of each alternate allele, with the exception that '=' is replaced by 'M' to ease VCF parsing.  Note that INDEL alleles do not have the first matched base (which is provided by default, per the spec) referred to by the CIGAR.">
   ##INFO=<ID=NUMALT,Number=1,Type=Integer,Description="Number of unique non-reference alleles in called genotypes at this position.">
   ##INFO=<ID=MEANALT,Number=A,Type=Float,Description="Mean number of unique non-reference allele observations per sample with the corresponding alternate alleles.">
   ##INFO=<ID=LEN,Number=A,Type=Integer,Description="allele length">
   ##INFO=<ID=MQM,Number=A,Type=Float,Description="Mean mapping quality of observed alternate alleles">
   ##INFO=<ID=MQMR,Number=1,Type=Float,Description="Mean mapping quality of observed reference alleles">
   ##INFO=<ID=PAIRED,Number=A,Type=Float,Description="Proportion of observed alternate alleles which are supported by properly paired read fragments">
   ##INFO=<ID=PAIREDR,Number=1,Type=Float,Description="Proportion of observed reference alleles which are supported by properly paired read fragments">
   ##INFO=<ID=technology.ILLUMINA,Number=A,Type=Float,Description="Fraction of observations supporting the alternate observed in reads from ILLUMINA">
   ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
   ##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the called genotype">
   ##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy">
   ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
   ##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
   ##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of the reference observations">
   ##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
   ##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of the alternate observations">
   ##contig=<ID=10>
   ##bcftools_viewVersion=1.9+htslib-1.9
   ##bcftools_viewCommand=view -h SAMEA2569438.chr10.vcf.gz; Date=Fri Dec  4 11:17:43 2020
   #CHROM   POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  SAMEA2569438
```

- Print some SNPs:

  ```
  m_jan2020@mjan2020VirtualBox:~$ bcftools view -H -v snps SAMEA2569438.chr10.vcf.gz |less
  ```

You get:

```
    10      9000024 .       G       T       52.1811 .       AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP=7.35324;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS=7.37776;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=74;QR=0;RO=0;RPL=0;RPP=7.35324;RPPR=0;RPR=2;RUN=1;SAF=2;SAP=7.35324;SAR=0;SRF=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1    GT:DP:RO:QR:AO:QA:GL    1/1:2:0:0:2:74:-7.02402,-0.60206,0
    10      9000178 .       T       A       93.4005 .       AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS=8.76405;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=119;QR=0;RO=0;RPL=2;RPP=3.73412;RPPR=0;RPR=1;RUN=1;SAF=3;SAP=9.52472;SAR=0;SRF=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1   GT:DP:RO:QR:AO:QA:GL    1/1:3:0:0:3:119:-11.095,-0.90309,0
    10      9000411 .       G       C       93.3954 .       AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS=8.76405;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=119;QR=0;RO=0;RPL=3;RPP=9.52472;RPPR=0;RPR=0;RUN=1;SAF=1;SAP=3.73412;SAR=2;SRF=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1   GT:DP:RO:QR:AO:QA:GL    1/1:3:0:0:3:119:-11.0945,-0.90309,0
    10      9000729 .       G       A       91.6745 .       AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS=8.76405;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=112;QR=0;RO=0;RPL=2;RPP=3.73412;RPPR=0;RPR=1;RUN=1;SAF=1;SAP=3.73412;SAR=2;SRF=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1   GT:DP:RO:QR:AO:QA:GL    1/1:3:0:0:3:112:-10.4453,-0.90309,0
    ...
```

- Print some INDELs

```
m_jan2020@mjan2020VirtualBox:~$ bcftools view -H -v indels SAMEA2569438.chr10.vcf.gz |less
```

You get:

```
    10      9000591 .       TAA     TAAA    97.543  .
AB=0.8;ABP=6.91895;AC=1;AF=0.5;AN=2;AO=4;CIGAR=1M1I2M;DP=5;DPB=6.33333;DPRA=0;EPP=3.0103;EPPR=5.18177;GTI=0;LEN=1;MEANALT=1;MQM=60;
MQMR=60;NS=1;NUMALT=1;ODDS=3.03447;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=160;QR=39;RO=1;RPL=2;RPP=3.0103;RPPR=5.18177;RPR=2
;RUN=1;SAF=4;SAP=11.6962;SAR=0;SRF=0;SRP=5.18177;SRR=1;TYPE=ins;technology.ILLUMINA=1     GT:DP:RO:QR:AO:QA:GL
0/1:5:1:39:4:160:-13.2783,0,-2.39141
    10      9002447 .       TAAAAAAAT       TAAAAAAAAAT     170.777 .
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=6;CIGAR=1M2I8M;DP=6;DPB=7.33333;DPRA=0;EPP=8.80089;EPPR=0;GTI=0;LEN=2;MEANALT=1;MQM=60;MQMR=0;NS=1;NUM
ALT=1;ODDS=12.9229;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=210;QR=0;RO=0;RPL=3;RPP=3.0103;RPPR=0;RPR=3;RUN=1;SAF=2;SAP=4.4579
5;SAR=4;SRF=0;SRP=0;SRR=0;TYPE=ins;technology.ILLUMINA=1  GT:DP:RO:QR:AO:QA:GL    1/1:6:0:0:6:210:-19.2409,-1.80618,0
    10      9003641 .       CTA     CTTA    71.096  .
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1M1I2M;DP=3;DPB=4;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;
ODDS=8.76405;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=95;QR=0;RO=0;RPL=2;RPP=3.73412;RPPR=0;RPR=1;RUN=1;SAF=0;SAP=9.52472;SAR=
3;SRF=0;SRP=0;SRR=0;TYPE=ins;technology.ILLUMINA=1        GT:DP:RO:QR:AO:QA:GL    1/1:3:0:0:3:95:-8.86456,-0.90309,0
    10      9009331 .       GC      GGAC    73.5205 .
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=5;CIGAR=1M2I1M;DP=6;DPB=11.5;DPRA=0;EPP=3.44459;EPPR=0;GTI=0;LEN=2;MEANALT=2;MQM=39.2;MQMR=0;NS=1;NUMA
LT=1;ODDS=12.9229;PAIRED=1;PAIREDR=0;PAO=0.5;PQA=17.5;PQR=17.5;PRO=0.5;QA=129;QR=0;RO=0;RPL=2;RPP=3.44459;RPPR=0;RPR=3;RUN=1;SAF=0;
SAP=13.8677;SAR=5;SRF=0;SRP=0;SRR=0;TYPE=ins;technology.ILLUMINA=1        GT:DP:RO:QR:AO:QA:GL
1/1:6:0:0:5:129:-9.8675,-2.10721,0
    10      9009333 .       GATC    GC      63.2885 .
AB=0.833333;ABP=8.80089;AC=1;AF=0.5;AN=2;AO=5;CIGAR=1M2D1M;DP=6;DPB=3.5;DPRA=0;EPP=3.44459;EPPR=5.18177;GTI=0;LEN=2;MEANALT=1;MQM=3
9.2;MQMR=60;NS=1;NUMALT=1;ODDS=5.84393;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=142;QR=32;RO=1;RPL=3;RPP=3.44459;RPPR=5.18177;
RPR=2;RUN=1;SAF=0;SAP=13.8677;SAR=5;SRF=0;SRP=5.18177;SRR=1;TYPE=del;technology.ILLUMINA=1        GT:DP:RO:QR:AO:QA:GL
0/1:6:1:32:5:142:-9.33308,0,-1.39313
    ...
```

- Print variants located in a specific region

To fetch the variants located in a specific genomic region, first you need to build an index for the VCF, to do this use `bcftools index`:

```
m_jan2020@mjan2020VirtualBox:~$ bcftools index SAMEA2569438.chr10.vcf.gz
```

And then you can use `bcftools view` with the `-r` option to query a specific region:

```
m_jan2020@mjan2020VirtualBox:~$ bcftools view -H -r 10:11000000-12000000 SAMEA2569438.chr10.vcf.gz |less
```

- Print some basic stats for the VCF file

We can use the `stats` command to generate a basic report on the number of variants in a VCF file:

```
m_jan2020@mjan2020VirtualBox:~$ bcftools stats SAMEA2569438.chr10.vcf.gz |grep ^SN
```

We pipe the output of the `stats` command to the UNIX `grep` command to print only the lines starting with `SN`:

```
 SN   0   number of samples:  1
 SN   0   number of records:  31521
 SN   0   number of no-ALTs:  0
 SN   0   number of SNPs: 26352
 SN   0   number of MNPs: 2484
 SN   0   number of indels:   2426
 SN   0   number of others:   341
 SN   0   number of multiallelic sites:   97
 SN   0   number of multiallelic SNP sites:   6
```

- Selecting the multiallelic SNPs

Use the following command to select the multiallelic SNPs:

```
m_jan2020@mjan2020VirtualBox:~$ bcftools view -H -m3 -v snps SAMEA2569438.chr10.vcf.gz |less
```

And you get:

```
    10      9057625 .       GCC     GC,GCA  166.178 .
AB=0.714286,0.285714;ABP=5.80219,5.80219;AC=1,1;AF=0.5,0.5;AN=2;AO=5,2;CIGAR=1M1D1M,2M1X;DP=7;DPB=5.33333;DPRA=0,0;EPP=6.91895,3.01
03;EPPR=0;GTI=0;LEN=1,1;MEANALT=2,2;MQM=60,60;MQMR=0;NS=1;NUMALT=2;ODDS=2.13936;PAIRED=1,1;PAIREDR=0;PAO=0,0;PQA=0,0;PQR=0;PRO=0;QA
=194,74;QR=0;RO=0;RPL=3,2;RPP=3.44459,7.35324;RPPR=0;RPR=2,0;RUN=1,1;SAF=3,1;SAP=3.44459,3.0103;SAR=2,1;SRF=0;SRP=0;SRR=0;TYPE=del,
snp;technology.ILLUMINA=1,1    GT:DP:RO:QR:AO:QA:GL    1/2:7:0:0:5,2:194,74:-22.375,-6.42337,-4.91822,-16.3265,0,-15.7244
```

```
    10      9320227 .       AGCA    GGCG,GGCA       106.742 .
AB=0.5,0.5;ABP=3.0103,3.0103;AC=1,1;AF=0.5,0.5;AN=2;AO=3,3;CIGAR=1X2M1X,1X3M;DP=6;DPB=6.5;DPRA=0,0;EPP=3.73412,3.73412;EPPR=0;GTI=0
;LEN=4,1;MEANALT=2,2;MQM=46,60;MQMR=0;NS=1;NUMALT=2;ODDS=4.98145;PAIRED=1,1;PAIREDR=0;PAO=1,1;PQA=34,34;PQR=0;PRO=0;QA=106,115;QR=0
;RO=0;RPL=0,0;RPP=9.52472,9.52472;RPPR=0;RPR=3,3;RUN=1,1;SAF=1,2;SAP=3.73412,3.73412;SAR=2,1;SRF=0;SRP=0;SRR=0;TYPE=complex,snp;tec
hnology.ILLUMINA=1,1   GT:DP:RO:QR:AO:QA:GL    1/2:6:0:0:3,3:106,115:-16.6863,-9.82069,-7.71348,-8.53081,0,-6.4236
    10      9343463 .       AGGA    GGGG,GGGA       184.726 .
AB=0.4,0.6;ABP=3.87889,3.87889;AC=1,1;AF=0.5,0.5;AN=2;AO=4,6;CIGAR=1X2M1X,1X3M;DP=10;DPB=10;DPRA=0,0;EPP=5.18177,4.45795;EPPR=0;GTI
=0;LEN=4,1;MEANALT=2,2;MQM=27.5,56.6667;MQMR=0;NS=1;NUMALT=2;ODDS=2.15329;PAIRED=1,1;PAIREDR=0;PAO=0,0;PQA=0,0;PQR=0;PRO=0;QA=147,2
26;QR=0;RO=0;RPL=3,5;RPP=5.18177,8.80089;RPPR=0;RPR=1,1;RUN=1,1;SAF=0,5;SAP=11.6962,8.80089;SAR=4,1;SRF=0;SRP=0;SRR=0;TYPE=complex,
snp;technology.ILLUMINA=1,1    GT:DP:RO:QR:AO:QA:GL    1/2:10:0:0:4,6:147,226:-26.2074,-18.6652,-17.4611,-7.83766,0,-6.03148
    10      9362283 .       TGCC    CGCG,CGCC       376.577 .
AB=0.5625,0.4375;ABP=3.55317,3.55317;AC=1,1;AF=0.5,0.5;AN=2;AO=9,7;CIGAR=1X2M1X,1X3M;DP=16;DPB=16;DPRA=0,0;EPP=3.25157,3.32051;EPPR
=0;GTI=0;LEN=4,1;MEANALT=2,2;MQM=55.3333,53.2857;MQMR=0;NS=1;NUMALT=2;ODDS=31.2564;PAIRED=0,0;PAIREDR=0;PAO=0,0;PQA=0,0;PQR=0;PRO=0
;QA=307,256;QR=0;RO=0;RPL=1,2;RPP=14.8328,5.80219;RPPR=0;RPR=8,5;RUN=1,1;SAF=6,3;SAP=5.18177,3.32051;SAR=3,4;SRF=0;SRP=0;SRR=0;TYPE
=complex,snp;technology.ILLUMINA=1,1   GT:DP:RO:QR:AO:QA:GL    1/2:16:0:0:9,7:307,256:-45.476,-20.9531,-18.2438,-24.8727,0,-22.7655
    ...
```

Filtering the artifactual variants

The process for identifiying variants is not perfect, and Freebayes and in general all tools used to identify variants will report variants that are not real. These artifactual variants must be idenfitied and flagged so that users or tools using them do not take them into account, or treat them with caution in any subsequent analysis.

There are several filtering tools and strategies available for variant filtering, with varying degrees of complexity and sophistication. However, in this course we will use a very simple, yet effective approach, which consists of using the quality value assigned by Freebayes as a proxy to estimate the likelihood of a variant being real. The lower the quality value, the less likely it is that a variant is real.

In this course, we will use `bcftools filter` with a hard cut-off value of `<=1` to flag the variants that have a low quality.
For this, first go to the directory where you ran Freebayes if you are not already there:

```
    m_jan2020@mjan2020VirtualBox:~$ cd /home/m_jan2020/course/vcalling
```

And enter the following in your terminal:

```
    m_jan2020@mjan2020VirtualBox:~$ bcftools filter -sQUALFILTER -e'QUAL<1' SAMEA2569438.chr10.vcf.gz -o
SAMEA2569438.chr10.filt.vcf.gz -Oz
```

Where the string passed using the `-s` option sets the label used for the filtered lines in the 7th column of the VCF, while the `-Oz` option is used to generate the output VCF in a compressed format.

Now, use `bcftools view` to verify that the 7th column has 2 new labels: `QUALFILTER` and `PASS`.

```
    m_jan2020@mjan2020VirtualBox:~$ bcftools view -H SAMEA2569438.chr10.filt.vcf.gz |less
```

The `-H` option is used to skip the header section and print the data lines only:

```
    10      9000024 .       G       T       52.1811 PASS
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=1X;DP=2;DPB=2;DPRA=0;EPP=7.35324;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS
=7.37776;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=74;QR=0;RO=0;RPL=0;RPP=7.35324;RPPR=0;RPR=2;RUN=1;SAF=2;SAP=7.35324;SAR=0;SR
F=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1    GT:DP:RO:QR:AO:QA:GL    1/1:2:0:0:2:74:-7.02402,-0.60206,0
    10      9000056 .       CA      TC      41.7389 PASS
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2;CIGAR=2X;DP=2;DPB=2;DPRA=0;EPP=7.35324;EPPR=0;GTI=0;LEN=2;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS
=7.37776;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=63;QR=0;RO=0;RPL=2;RPP=7.35324;RPPR=0;RPR=0;RUN=1;SAF=2;SAP=7.35324;SAR=0;SR
F=0;SRP=0;SRR=0;TYPE=mnp;technology.ILLUMINA=1    GT:DP:RO:QR:AO:QA:GL    1/1:2:0:0:2:63:-5.97977,-0.60206,0
    10      9000178 .       T       A       93.4005 PASS
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS
=8.76405;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=119;QR=0;RO=0;RPL=2;RPP=3.73412;RPPR=0;RPR=1;RUN=1;SAF=3;SAP=9.52472;SAR=0;S
RF=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1   GT:DP:RO:QR:AO:QA:GL    1/1:3:0:0:3:119:-11.095,-0.90309,0
    10      9000411 .       G       C       93.3954 PASS
AB=0;ABP=0;AC=2;AF=1;AN=2;AO=3;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=3.73412;EPPR=0;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=0;NS=1;NUMALT=1;ODDS
=8.76405;PAIRED=1;PAIREDR=0;PAO=0;PQA=0;PQR=0;PRO=0;QA=119;QR=0;RO=0;RPL=3;RPP=9.52472;RPPR=0;RPR=0;RUN=1;SAF=1;SAP=3.73412;SAR=2;S
RF=0;SRP=0;SRR=0;TYPE=snp;technology.ILLUMINA=1   GT:DP:RO:QR:AO:QA:GL    1/1:3:0:0:3:119:-11.0945,-0.90309,0
    ......
```

We can also print only the variants that have been filtered by doing:

```
    m_jan2020@mjan2020VirtualBox:~$ bcftools view -H -f QUALFILTER SAMEA2569438.chr10.filt.vcf.gz |less
```

And you get:

```
    10      9009050 .       A       G       0.292908        QUALFILTER
AB=0.2;ABP=10.8276;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=10;DPB=10;DPRA=0;EPP=7.35324;EPPR=4.09604;GTI=0;LEN=1;MEANALT=1;MQM=39.5;MQMR=
54.75;NS=1;NUMALT=1;ODDS=2.66254;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=81;QR=256;RO=8;RPL=2;RPP=7.35324;RPPR=4.09604;RPR=0;
RUN=1;SAF=0;SAP=7.35324;SAR=2;SRF=4;SRP=3.0103;SRR=4;TYPE=snp;technology.ILLUMINA=1       GT:DP:RO:QR:AO:QA:GL
0/1:10:8:256:2:81:-4.00694,0,-20.267
```

```
    10      9009343 .       A       T       0.142595        QUALFILTER
AB=0.25;ABP=7.35324;AC=1;AF=0.5;AN=2;AO=2;CIGAR=1X;DP=8;DPB=8;DPRA=0;EPP=7.35324;EPPR=3.0103;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=42.6
667;NS=1;NUMALT=1;ODDS=3.39984;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=60;QR=206;RO=6;RPL=0;RPP=7.35324;RPPR=3.0103;RPR=2;RUN
=1;SAF=0;SAP=7.35324;SAR=2;SRF=0;SRP=16.0391;SRR=6;TYPE=snp;technology.ILLUMINA=1 GT:DP:RO:QR:AO:QA:GL
0/1:8:6:206:2:60:-3.29073,0,-13.9981
    10      9009698 .       T       G       0.0499446       QUALFILTER
AB=0;ABP=0;AC=0;AF=0;AN=2;AO=2;CIGAR=1X;DP=3;DPB=3;DPRA=0;EPP=7.35324;EPPR=5.18177;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=1;NUMALT
=1;ODDS=4.46257;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=7;QR=17;RO=1;RPL=2;RPP=7.35324;RPPR=5.18177;RPR=0;RUN=1;SAF=2;SAP=7.3
5324;SAR=0;SRF=1;SRP=5.18177;SRR=0;TYPE=snp;technology.ILLUMINA=1 GT:DP:RO:QR:AO:QA:GL    0/0:3:1:17:2:7:0,-0.238091,-1.03498
    10      9014124 .       A       G       3.66831e-05     QUALFILTER
AB=0;ABP=0;AC=0;AF=0;AN=2;AO=2;CIGAR=1X;DP=6;DPB=6;DPRA=0;EPP=7.35324;EPPR=11.6962;GTI=0;LEN=1;MEANALT=1;MQM=5.5;MQMR=24;NS=1;NUMAL
T=1;ODDS=11.9687;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=74;QR=159;RO=4;RPL=0;RPP=7.35324;RPPR=11.6962;RPR=2;RUN=1;SAF=2;SAP=
7.35324;SAR=0;SRF=4;SRP=11.6962;SRR=0;TYPE=snp;technology.ILLUMINA=1      GT:DP:RO:QR:AO:QA:GL
0/0:6:4:159:2:74:0,-0.761681,-7.48247
    ....
```

- How many variants have been filtered?

We can use the `stats` command together with the `-f` option and the `QUALFILTER` label to generate a report that takes into account only the filtered variants:

```
    m_jan2020@mjan2020VirtualBox:~$ bcftools stats -f QUALFILTER SAMEA2569438.chr10.filt.vcf.gz |grep ^SN
```

And you get:

```
    SN  0   number of samples:  1
    SN  0   number of records:  1042
    SN  0   number of no-ALTs:  0
    SN  0   number of SNPs: 910
    SN  0   number of MNPs: 92
    SN  0   number of indels:   36
    SN  0   number of others:   8
    SN  0   number of multiallelic sites:   6
    SN  0   number of multiallelic SNP sites:   2
```

- How many variants remain after the filtering?

We need to use the `stats` command with the `PASS` label this time to generate a new report with the variants that have not been filtered:

```
    m_jan2020@mjan2020VirtualBox:~$ bcftools stats -f PASS SAMEA2569438.chr10.filt.vcf.gz |grep ^SN
```

You get:

```
    SN  0   number of samples:  1
    SN  0   number of records:  30479
    SN  0   number of no-ALTs:  0
    SN  0   number of SNPs: 25442
    SN  0   number of MNPs: 2392
    SN  0   number of indels:   2390
    SN  0   number of others:   333
    SN  0   number of multiallelic sites:   91
    SN  0   number of multiallelic SNP sites:   4
```

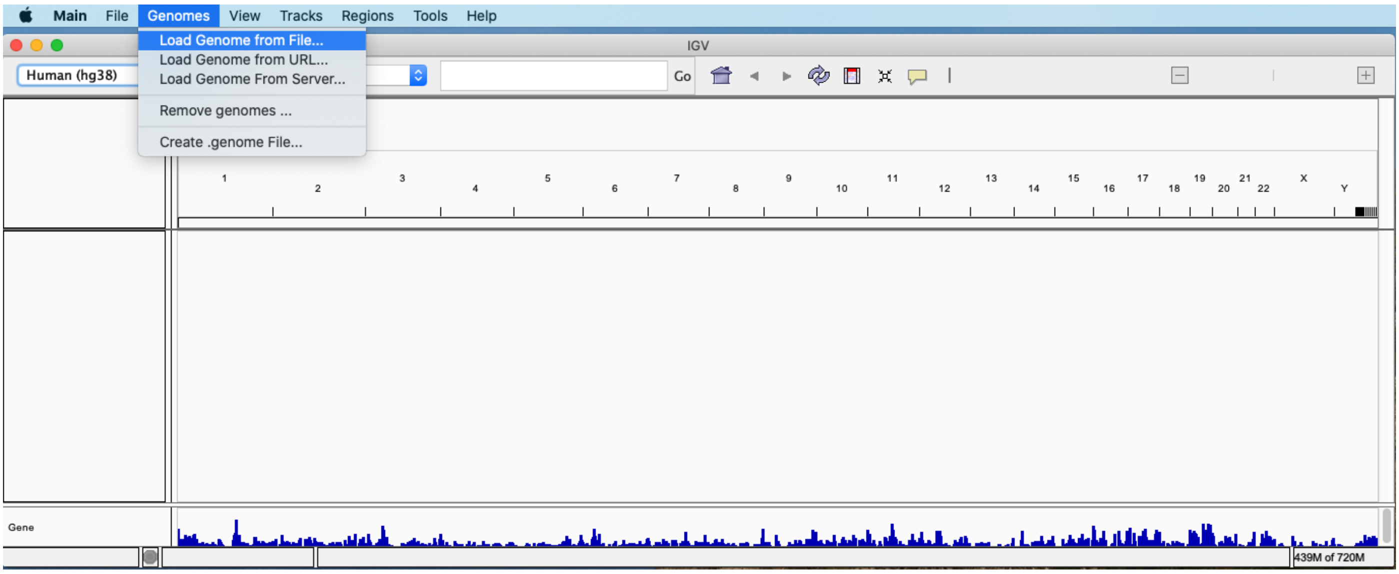
Exploring the identified variants using IGV

The Integrative Genomics Viewer (IGV) is a useful interactive tool that can be used to visually explore your genomic data. We are going to use it here to display the variants we have identified located in a specific region of chromosome 10.

First, open the `igv` program by going to your terminal and typing:

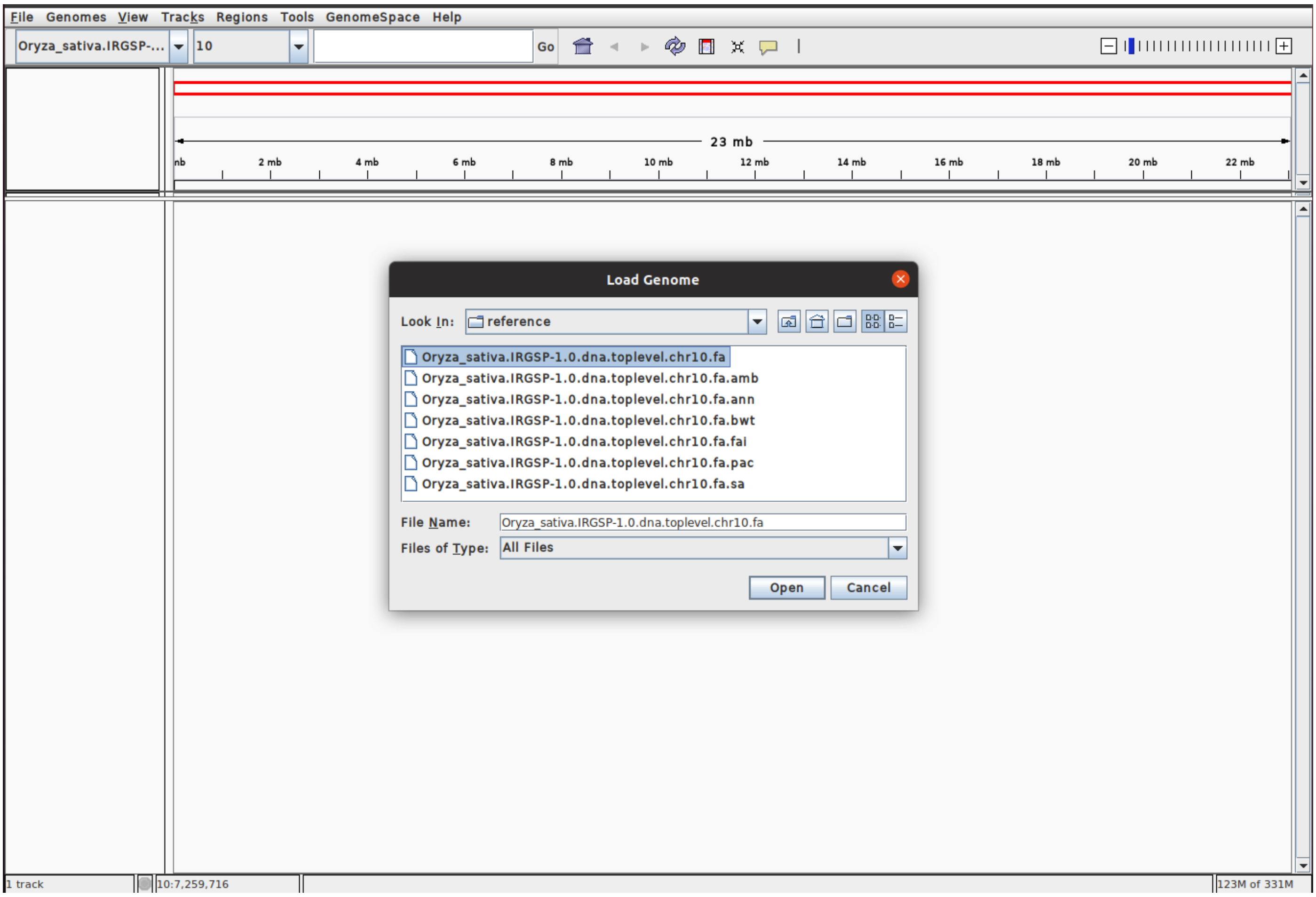
```
    igv
```

Once `IGV` is open, you will need to load the FASTA file containing the chromosome 10 sequence for rice, as this sequence is not included by default in `IGV`:



Look for you file by going to the folder named (`m_jan2020`) and clicking on the different folders until you find the FASTA file named `Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa`:

```
course->reference->Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa
```

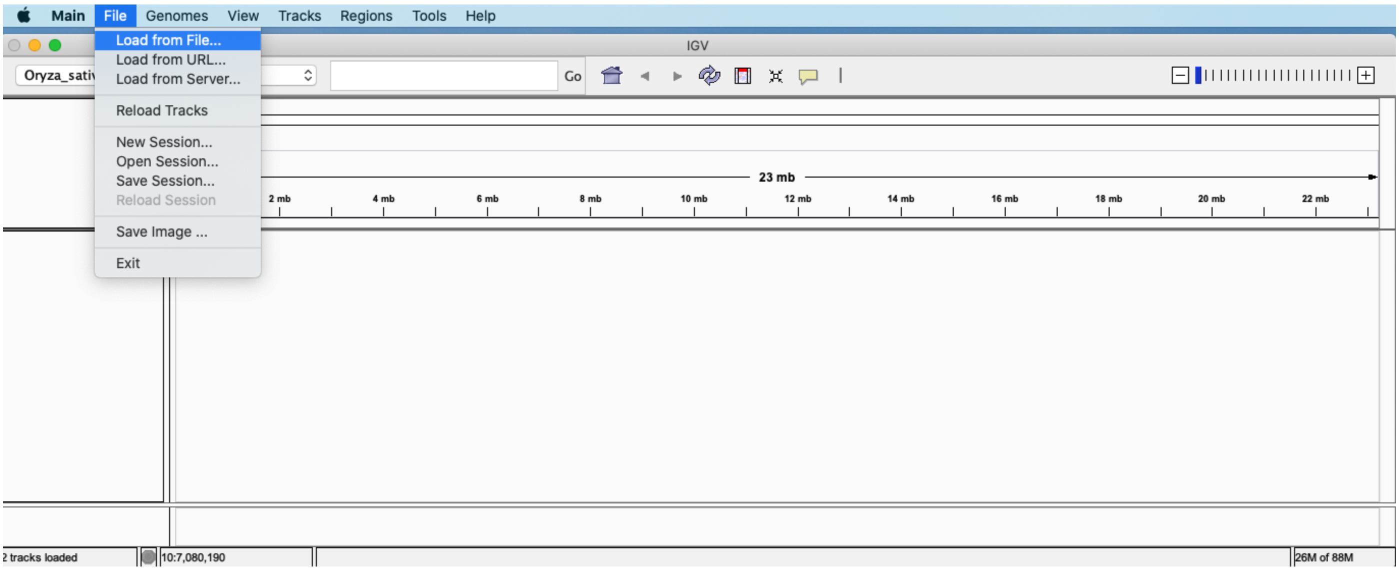

File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Oryza_sativa.IRGSP-...
10
Go
23 mb
2 mb
4 mb
6 mb
8 mb
10 mb
12 mb
14 mb
16 mb
18 mb
20 mb
22 mb
Load Genome
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.amb
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.ann
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.bwt
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.fai
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.pac
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.sa
Open
Cancel
1 track
10:7,259,716
123M of 331M

Look In: reference

File Name: Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa

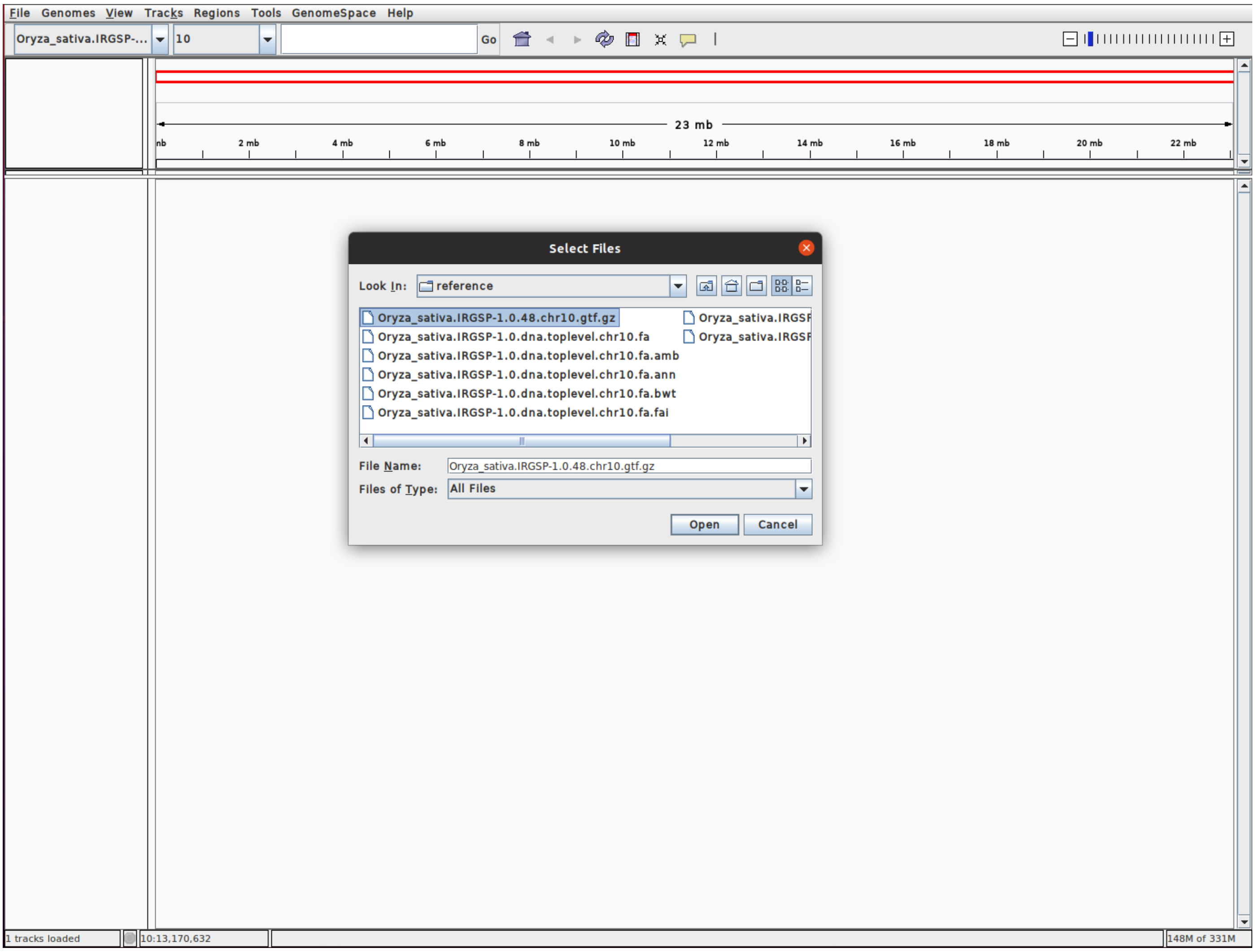
Files of Type: All Files

Now, load the GTF file containing the rice gene annotations for chromosome 10:

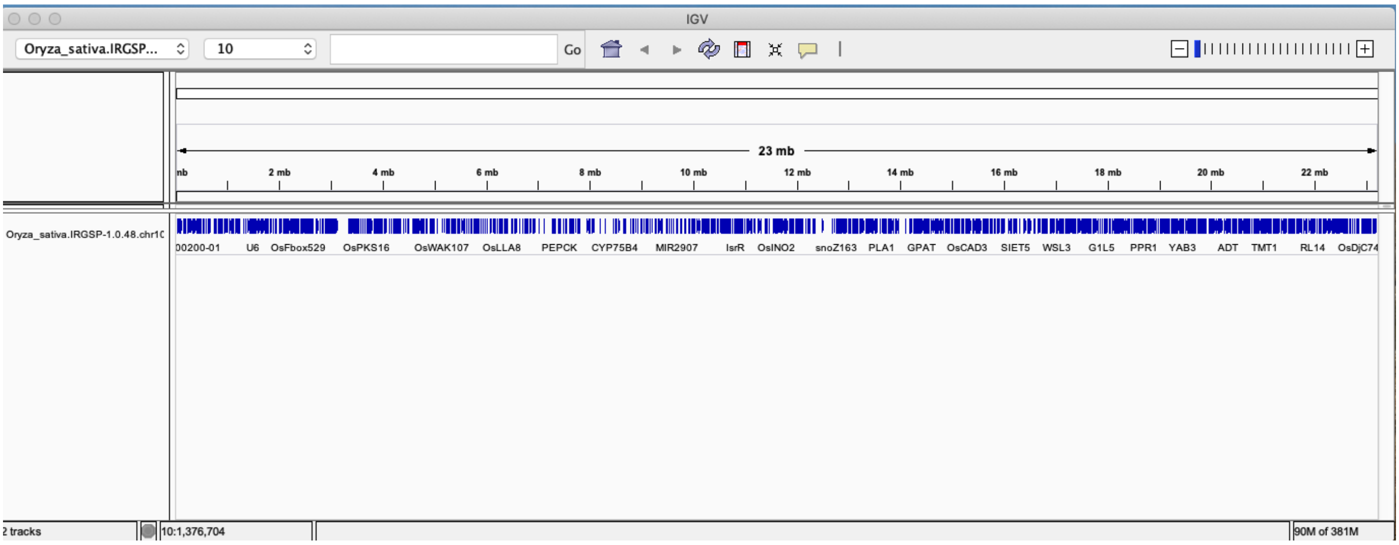


The GTF file can be found by going to the folder named (m_jan2020) and clicking on the different folders until you find the GTF file named Oryza_sativa.IRGSP-1.0.48.chr10.gtf.gz:

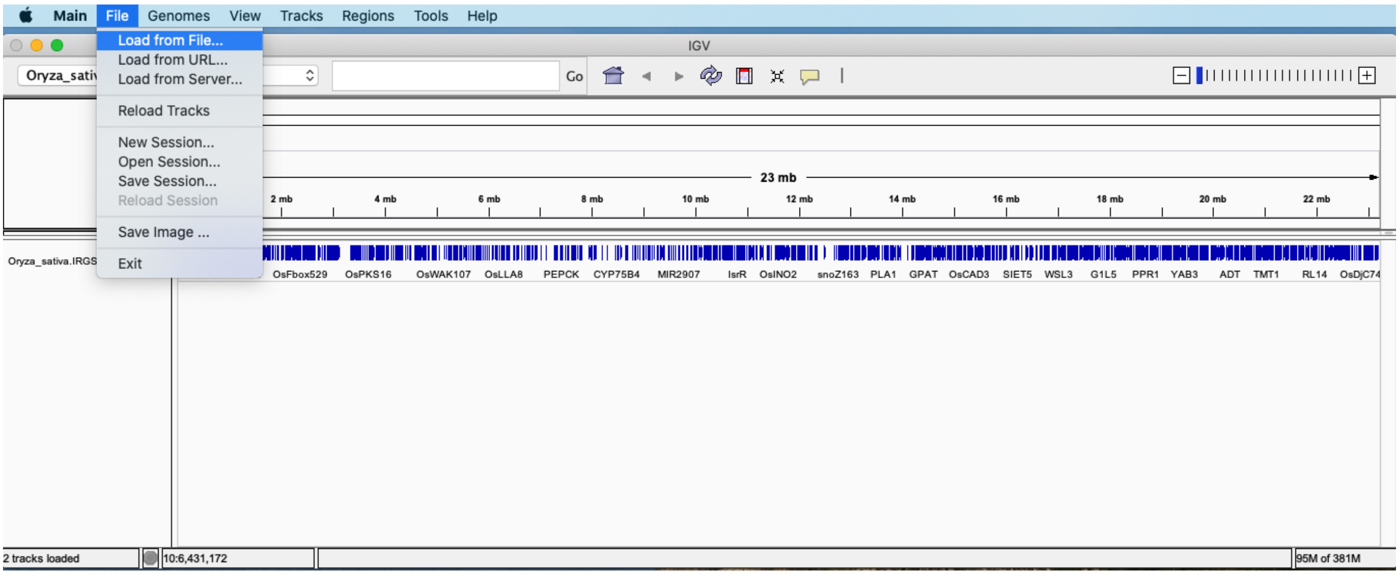
```
course->reference->Oryza_sativa.IRGSP-1.0.48.chr10.gtf.gz
```


File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Oryza_sativa.IRGSP-...
10
Go
23 mb
Select Files
Look In:
reference
Oryza_sativa.IRGSP-1.0.48.chr10.gtf.gz
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.amb
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.ann
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.bwt
Oryza_sativa.IRGSP-1.0.dna.toplevel.chr10.fa.fai
Oryza_sativa.IRGSF
Oryza_sativa.IRGSF
File Name:
Oryza_sativa.IRGSP-1.0.48.chr10.gtf.gz
Files of Type:
All Files
Open
Cancel
1 tracks loaded
10:13,170,632
148M of 331M

You will see a new track with all the genes annotated in chromosome 10:

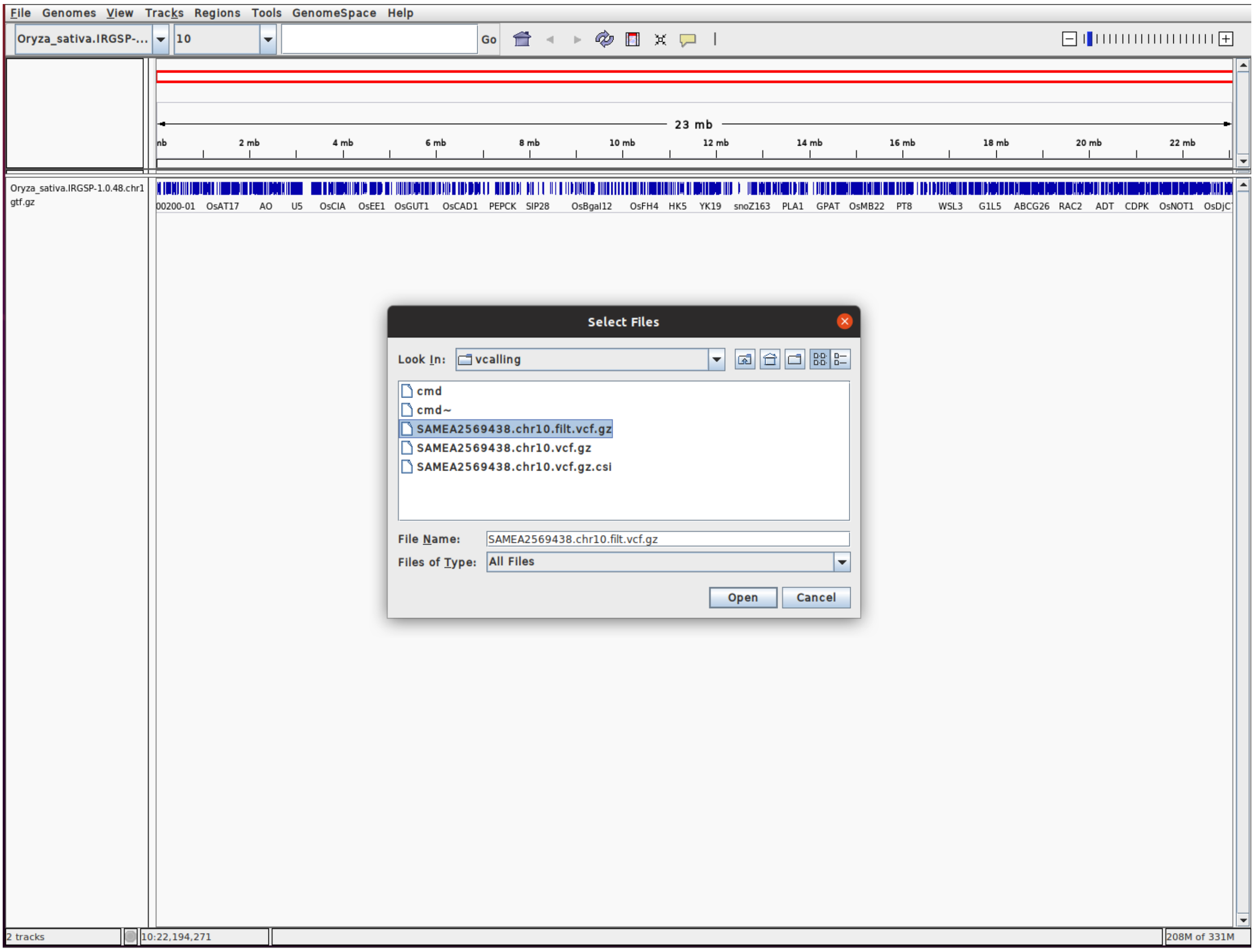


Now, load the filtered VCF file containing the variants:



The VCF file can be found by going to the folder named (`m_jan2020`) and clicking on the different folders until you find the VCF file named `SAMEA2569438.chr10.filt.vcf.gz`:

```
course->vcalling->SAMEA2569438.chr10.filt.vcf.gz
```


File Genomes View Tracks Regions Tools GenomeSpace Help
Oryza_sativa.IRGSP-...
10
Go
23 mb
2 mb 4 mb 6 mb 8 mb 10 mb 12 mb 14 mb 16 mb 18 mb 20 mb 22 mb
Oryza_sativa.IRGSP-1.0.48.chr1 gtf.gz
00200-01 OsAT17 AO U5 OsCIA OsEE1 OsGUT1 OsCAD1 PEPCK SIP28 OsBgal12 OsFH4 HK5 YK19 snoZ163 PLA1 GPAT OsMB22 PT8 WSL3 G1L5 ABCG26 RAC2 ADT CDPK OsNOT1 OsDjC
Open
Cancel
2 tracks
10:22,194,271
208M of 331M

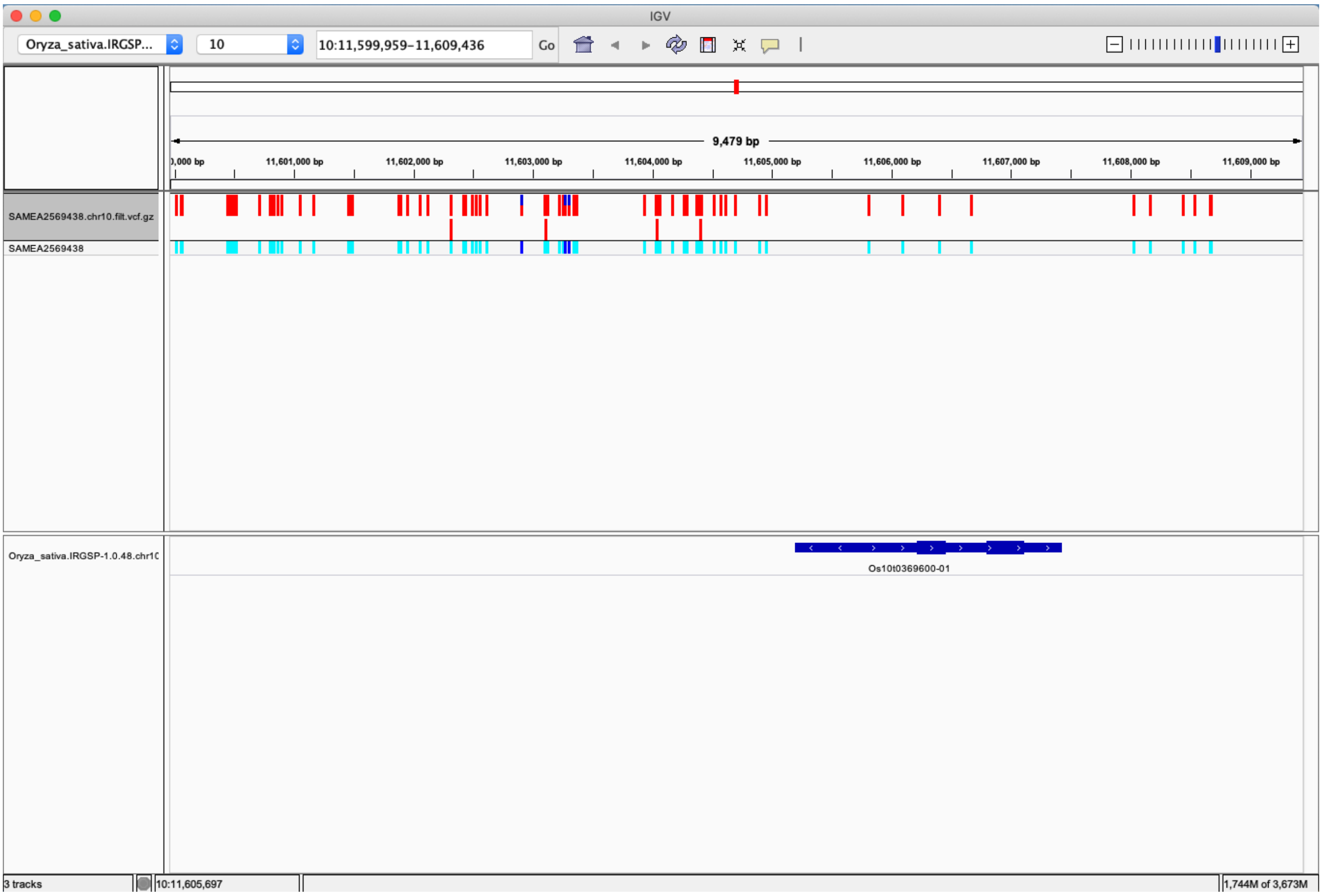
Select Files

Look In: vcalling

```
cmd
cmd~
SAMEA2569438.chr10.filt.vcf.gz
SAMEA2569438.chr10.vcf.gz
SAMEA2569438.chr10.vcf.gz.csi
```

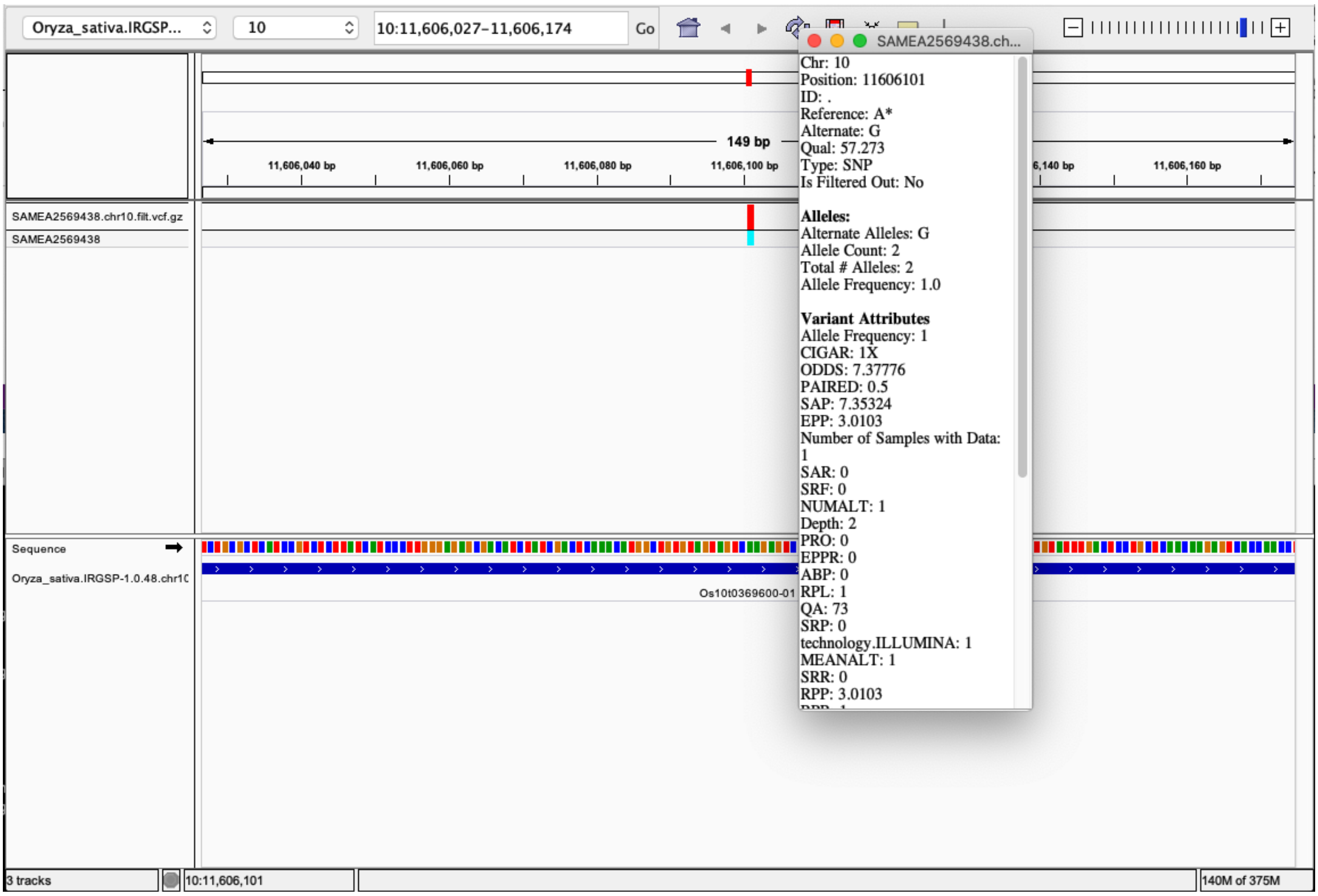
File Name: SAMEA2569438.chr10.filt.vcf.gz

Files of Type: All Files


IGV
Oryza_sativa.IRGSP...
10
10:11,599,959-11,609,436
Go
9,479 bp
),000 bp 11,601,000 bp 11,602,000 bp 11,603,000 bp 11,604,000 bp 11,605,000 bp 11,606,000 bp 11,607,000 bp 11,608,000 bp 11,609,000 bp
SAMEA2569438.chr10.filt.vcf.gz
SAMEA2569438
Oryza_sativa.IRGSP-1.0.48.chr10
Os10t0369600-01
3 tracks
10:11,605,697
1,744M of 3,673M

Now, you can click on a particular variant (red vertical bars) to display information such as:

- Position
- Reference and alternate alleles
- Type of variant (SNPs or INDEL)
- Variant quality
- Filtering status
- Variant attributes
- etc ...

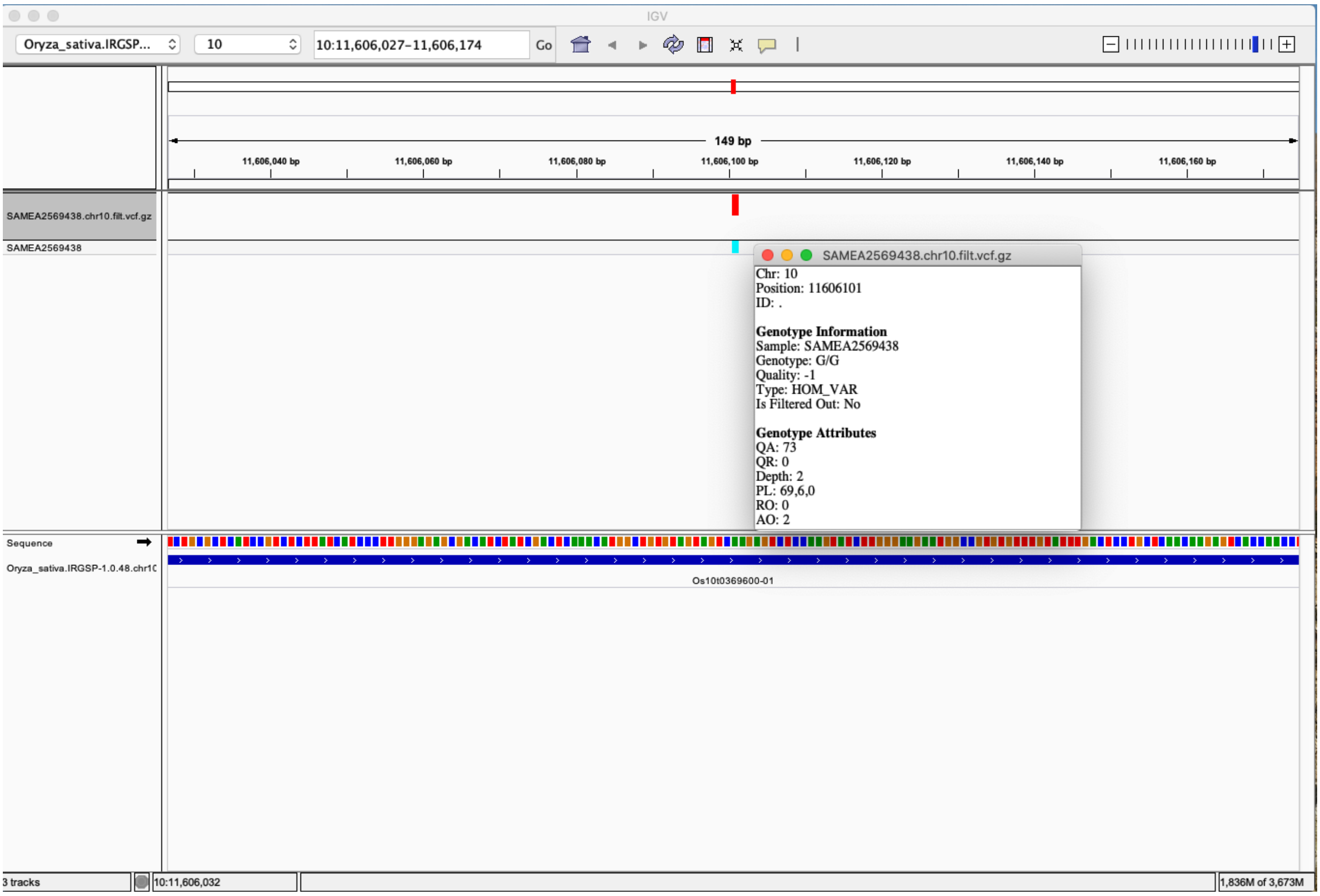


Chr: 10
Position: 11606101
ID: .
Reference: A*
Alternate: G
Qual: 57.273
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: G
Allele Count: 2
Total # Alleles: 2
Allele Frequency: 1.0

Variant Attributes
Allele Frequency: 1
CIGAR: 1X
ODDS: 7.37776
PAIRED: 0.5
SAP: 7.35324
EPP: 3.0103
Number of Samples with Data: 1
SAR: 0
SRF: 0
NUMALT: 1
Depth: 2
PRO: 0
EPPR: 0
ABP: 0
RPL: 1
QA: 73
SRP: 0
technology.ILLUMINA: 1
MEANALT: 1
SRR: 0
RPP: 3.0103

You can also click on the blue vertical bars to display the genotype information and genotype attributes:



SAMEA2569438.chr10.filt.vcf.gz

Chr: 10
Position: 11606101
ID: .

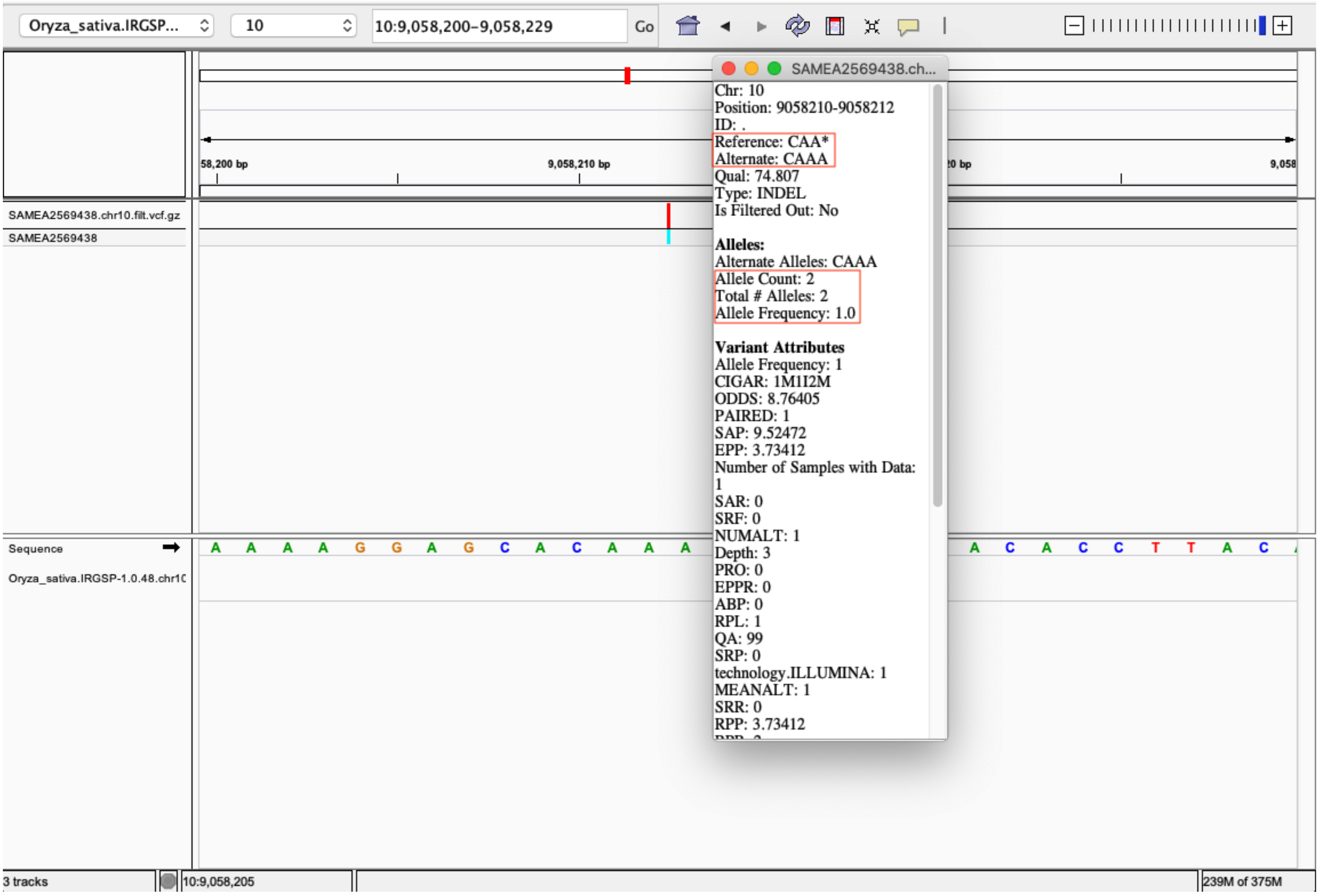
Genotype Information
Sample: SAMEA2569438
Genotype: G/G
Quality: -1
Type: HOM_VAR
Is Filtered Out: No

Genotype Attributes
QA: 73
QR: 0
Depth: 2
PL: 69,6,0
RO: 0
AO: 2

Let's take a closer look to an INDEL variant. For this, enter the following genomic coordinate in the search box:

```
10:9,058,200-9,058,229
```

Click on the red vertical bar in the center of the screen:

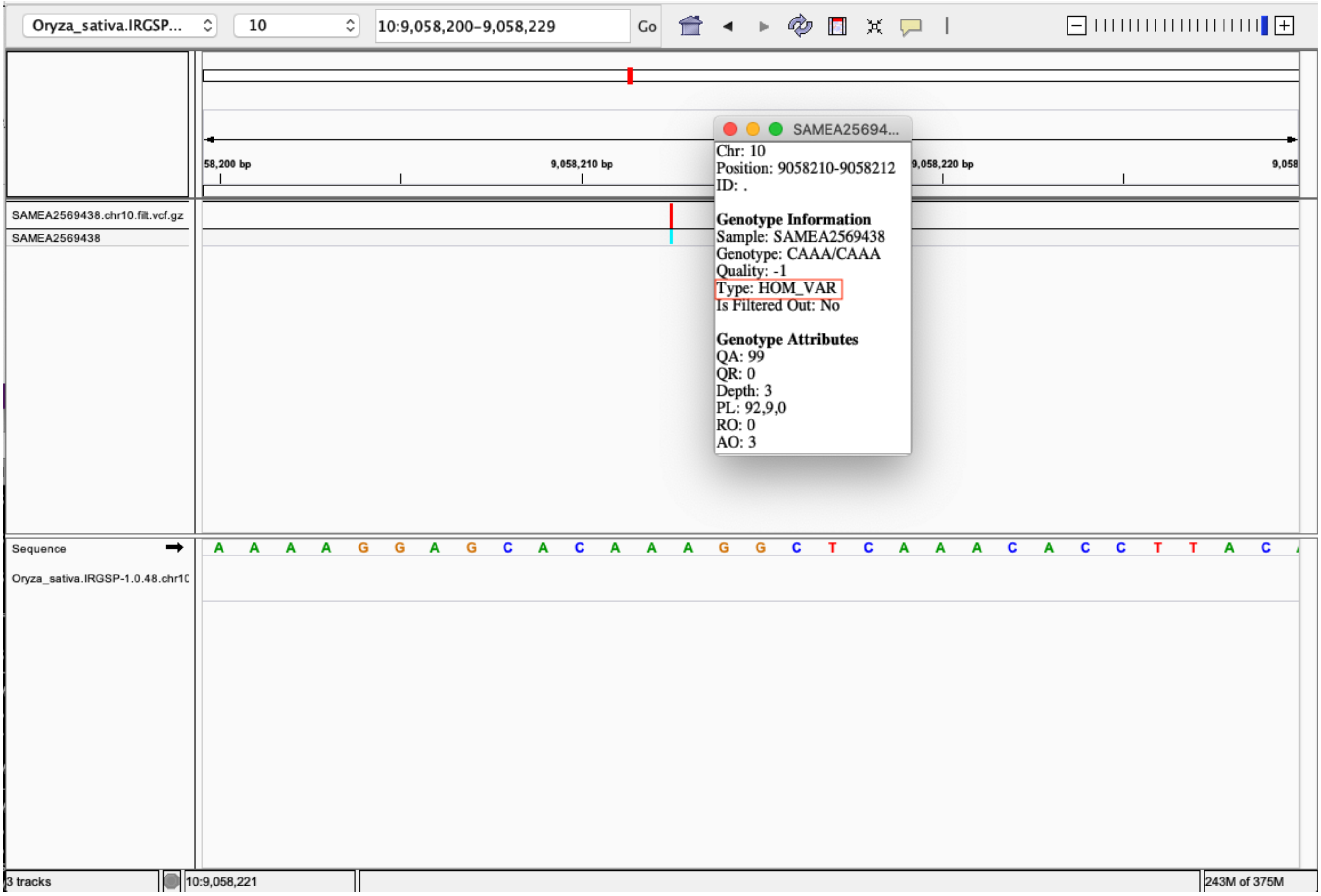


```
Chr: 10
Position: 9058210-9058212
ID: .
Reference: CAA*
Alternate: CAAA
Qual: 74.807
Type: INDEL
Is Filtered Out: No

Alleles:
Alternate Alleles: CAAA
Allele Count: 2
Total # Alleles: 2
Allele Frequency: 1.0

Variant Attributes
Allele Frequency: 1
CIGAR: 1M1I2M
ODDS: 8.76405
PAIRED: 1
SAP: 9.52472
EPP: 3.73412
Number of Samples with Data:
1
SAR: 0
SRF: 0
NUMALT: 1
Depth: 3
PRO: 0
EPPR: 0
ABP: 0
RPL: 1
QA: 99
SRP: 0
technology.ILLUMINA: 1
MEANALT: 1
SRR: 0
RPP: 3.73412
```

You can see that this is a 1bp insertion (CAA->CAAA) that have an alternate allele_count=2. Which means it is a homozygous variant, this can be confirmed by clicking on the genotype information bar:



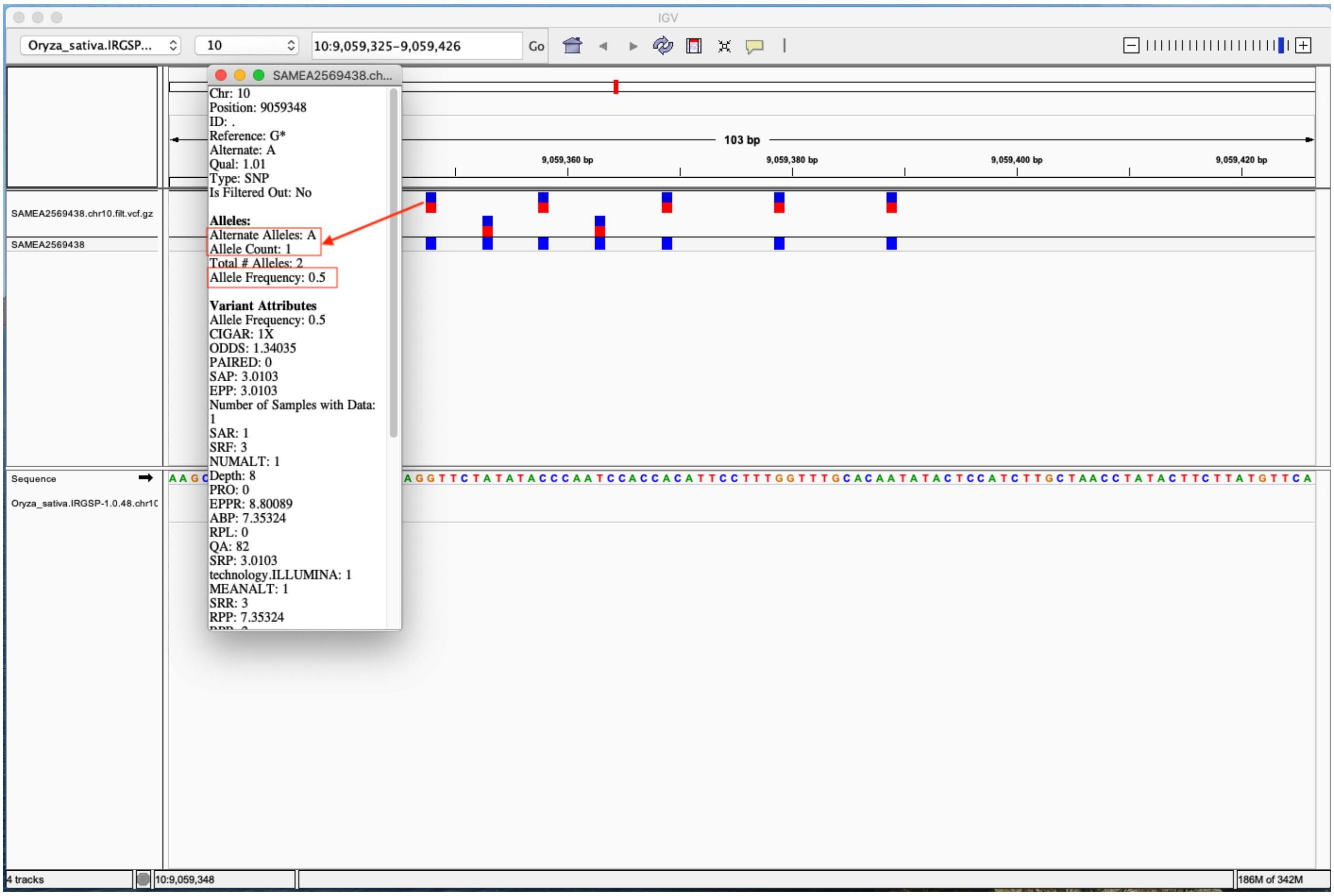
```
Chr: 10
Position: 9058210-9058212
ID: .

Genotype Information
Sample: SAMEA2569438
Genotype: CAAA/CAAA
Quality: -1
Type: HOM_VAR
Is Filtered Out: No

Genotype Attributes
QA: 99
QR: 0
Depth: 3
PL: 92,9,0
RO: 0
AO: 3
```

Now, let's visualize a SNP, for this enter the following genomic coordinate in the search box anc click the leftmost variant:

```
10:9,059,325-9,059,426
```



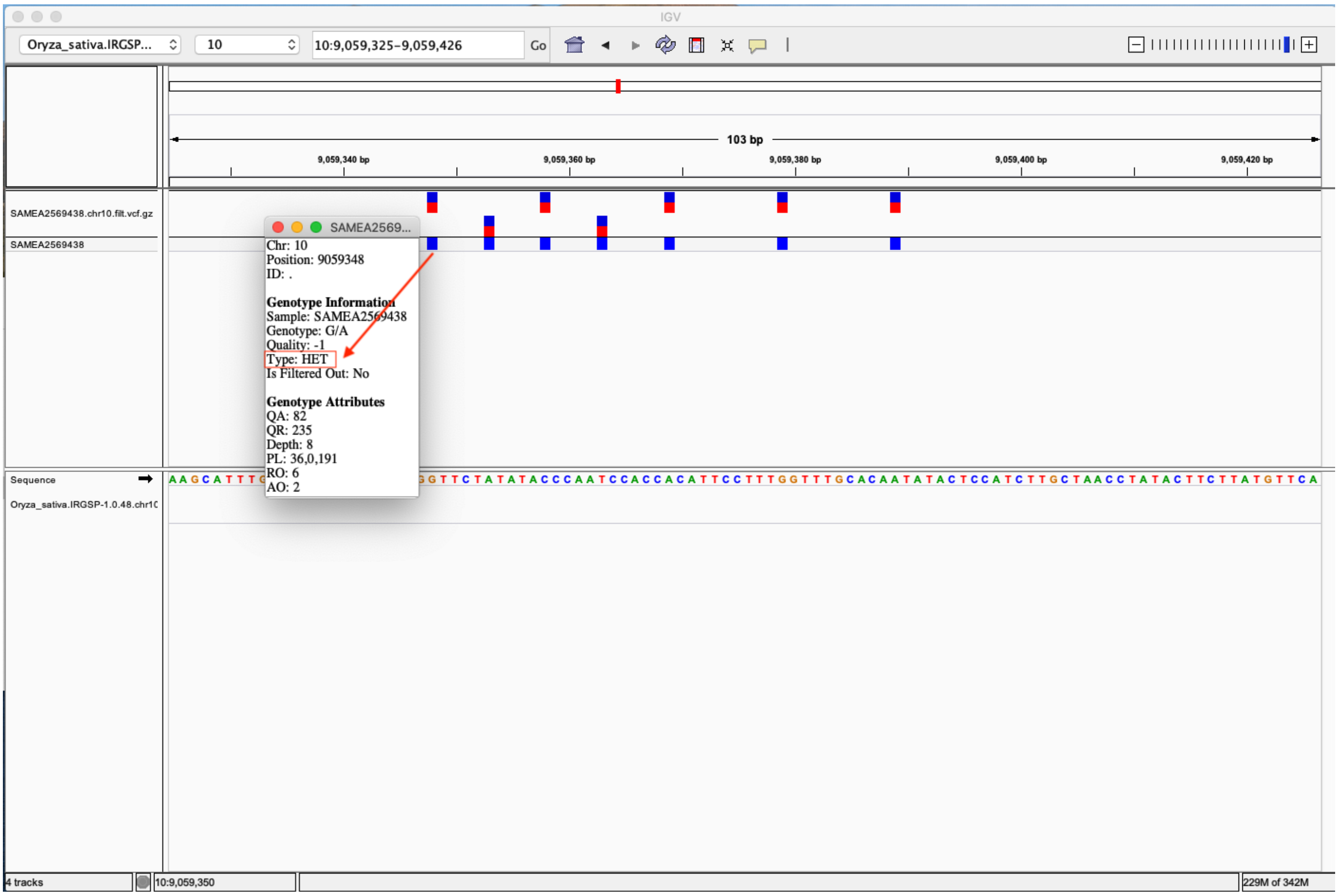
SAMEA2569438.ch...

Chr: 10
Position: 9059348
ID: .
Reference: G*
Alternate: A
Qual: 1.01
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: A
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5

Variant Attributes
Allele Frequency: 0.5
CIGAR: 1X
ODDS: 1.34035
PAIRED: 0
SAP: 3.0103
EPP: 3.0103
Number of Samples with Data: 1
SAR: 1
SRF: 3
NUMALT: 1
Depth: 8
PRO: 0
EPPR: 8.80089
ABP: 7.35324
RPL: 0
QA: 82
SRP: 3.0103
technology.ILLUMINA: 1
MEANALT: 1
SRR: 3
RPP: 7.35324

You can see that this is a single nucleotide substitution (G->A) that have an alternate allele_count=1 with an allele frequency of 0.5. Which means it is a heterozygous variant, this can be confirmed by clicking on the genotype information bar:



SAMEA2569...

Chr: 10
Position: 9059348
ID: .

Genotype Information
Sample: SAMEA2569438
Genotype: G/A
Quality: -1
Type: HET
Is Filtered Out: No

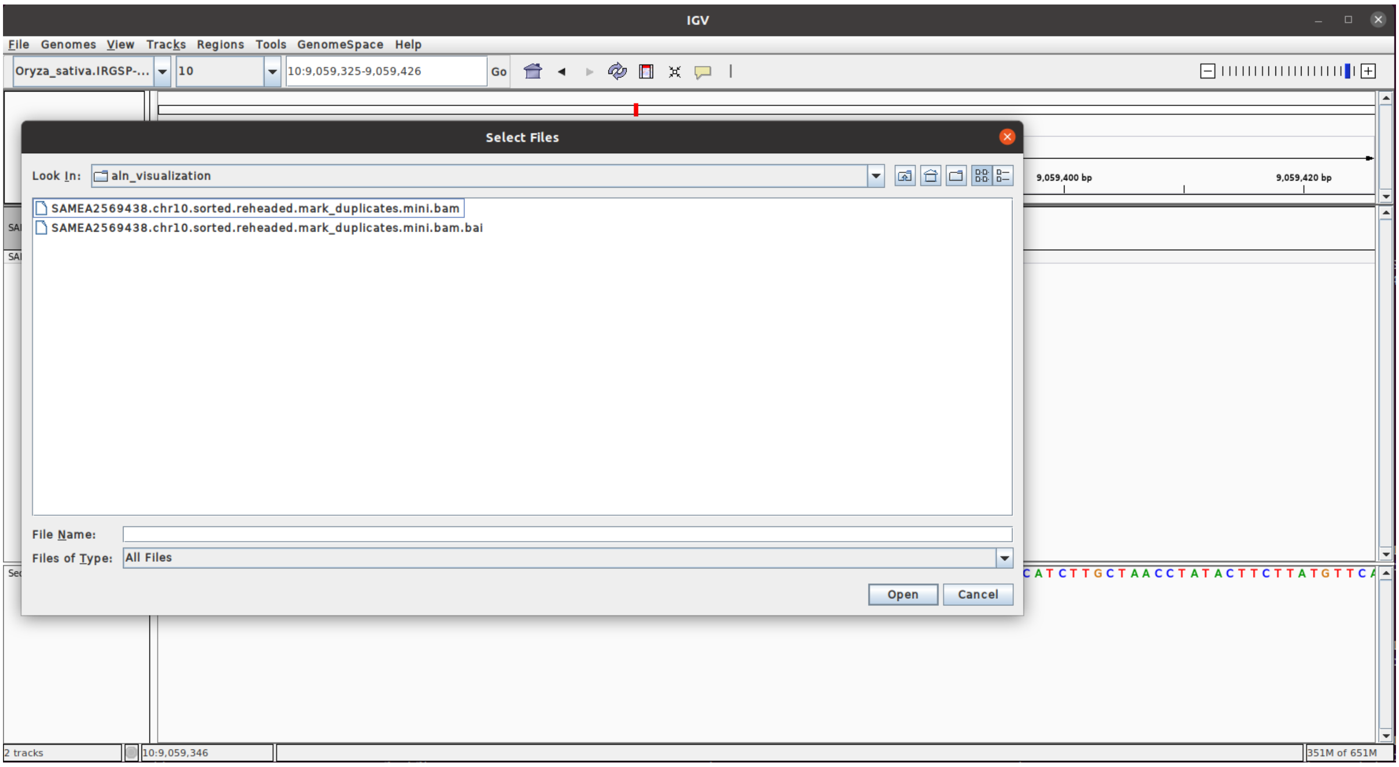
Genotype Attributes
QA: 82
QR: 235
Depth: 8
PL: 36,0,191
RO: 6
AO: 2

Also note that the vertical bars for the heterozygous variants have 2 colours, while the homozygous variants will have only one colour in `IGV`.

- Examining the aligned reads supporting a certain variant

`IGV` is really useful for examining the reads in your alignment file supporting a certain variant. To do this, you need to load the `BAM` file generated in the alignment section of this course that was previously uploaded to `IGV`, by going to the folder named (`m_jan2020`) and clicking on the different folders until you find the BAM file named `SAMEA2569438.chr10.sorted.reheaded.mark_duplicates.mini.bam`:

```
course->alignment->aln_visualization->SAMEA2569438.chr10.sorted.reheaded.mark_duplicates.mini.bam
```



Look In: aln_visualization

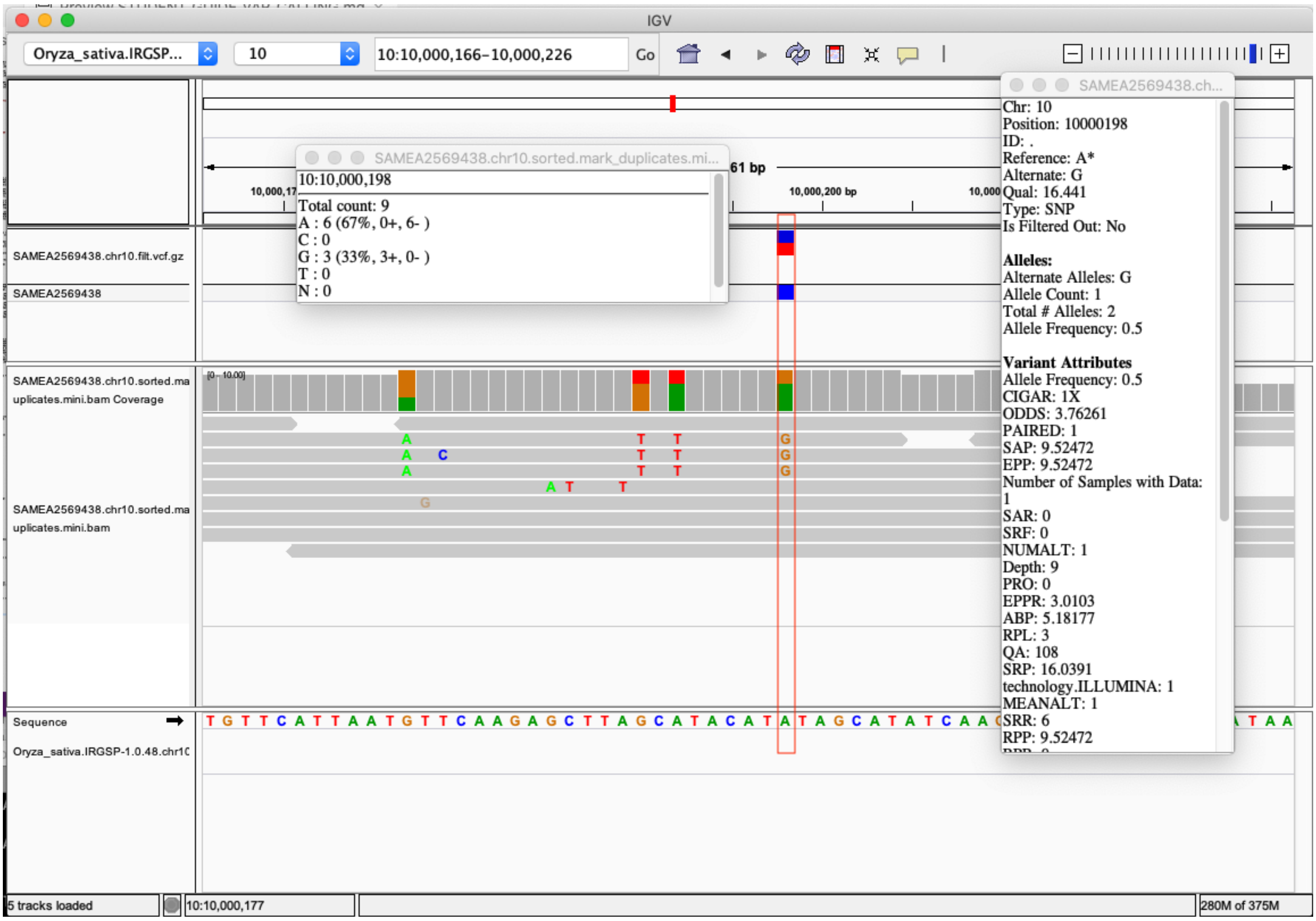
SAMEA2569438.chr10.sorted.reheaded.mark_duplicates.mini.bam
SAMEA2569438.chr10.sorted.reheaded.mark_duplicates.mini.bam.bai

File Name:
Files of Type: All Files

Now, go to the following coordinate:

```
10:10,000,166-10,000,226
```

Take a closer look to the SNP in position `10:10,000,198`. You can see the SNP variant and the reads supporting this SNP, you can also click the coverage track and see that there are a total of 9 reads covering this position, 3 of 9 have the alternate allele `G` and 6 of 9 have the reference allele `A`:


IGV
Oryza_sativa.IRGSP...
10
10:10,000,166-10,000,226
Go
61 bp
10,000,17
10,000,200 bp
10,000
SAMEA2569438.chr10.filt.vcf.gz
SAMEA2569438
SAMEA2569438.chr10.sorted.mark_duplicates.mini.bam Coverage
[0 - 10.00]
SAMEA2569438.chr10.sorted.mark_duplicates.mini.bam
Sequence
Oryza_sativa.IRGSP-1.0.48.chr10
T G T T C A T T A A T G T T C A A G A G C T T A G C A T A C A T A T A G C A T A T C A A
5 tracks loaded
10:10,000,177
280M of 375M

SAMEA2569438.chr10.sorted.mark_duplicates.mi...

10:10,000,198
Total count: 9
A : 6 (67%, 0+, 6-)
C : 0
G : 3 (33%, 3+, 0-)
T : 0
N : 0

SAMEA2569438.ch...

Chr: 10
Position: 10000198
ID: .
Reference: A*
Alternate: G
Qual: 16.441
Type: SNP
Is Filtered Out: No

Alleles:
Alternate Alleles: G
Allele Count: 1
Total # Alleles: 2
Allele Frequency: 0.5

Variant Attributes
Allele Frequency: 0.5
CIGAR: 1X
ODDS: 3.76261
PAIRED: 1
SAP: 9.52472
EPP: 9.52472
Number of Samples with Data: 1
SAR: 0
SRF: 0
NUMALT: 1
Depth: 9
PRO: 0
EPPR: 3.0103
ABP: 5.18177
RPL: 3
QA: 108
SRP: 16.0391
technology.ILLUMINA: 1
MEANALT: 1
SRR: 6
RPP: 9.52472