Variant Calling analysis, Part 2: Analyzing Variant Call results (Group Exercise)

In this exercise, we will work in groups to examine the results from the SNP analysis workflow that we started yesterday. **Groups sharing files: 1&2, 3&4, and 5&6.**

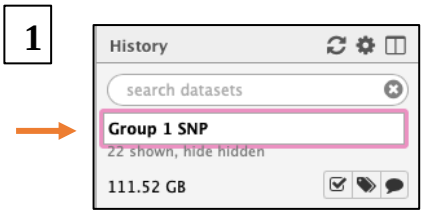
Learning objectives:

- Share and publish your workflow results
- Examine your results and the outputs of the workflow
- View VCF files in JBrowse
- Examine the filtered VFC file, extract Gene IDs, and create a Venny diagram

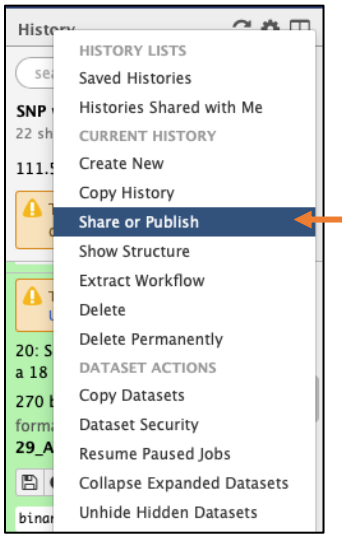
1. Share and Publish your workflow results.

- Give your workflow a meaningful name, *e.g*. Group 1 Afumgiatus Af293.
- Click on the on the *History options* link and select the *Share* or *Publish* option.
- On the next page click on the *Make History Accessible* and *Publish* link.
- To import a shared history into your workspace follow these steps:

1

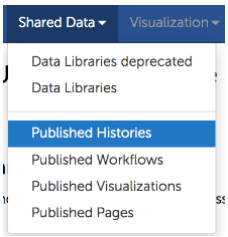


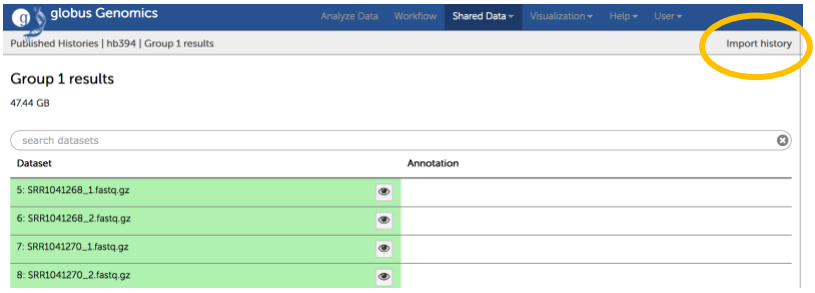
2



3

- Select *Published Histories* from the *Shared data* menu.
- From the list of shared histories click on the one you want to import and on the next page select the *Import* link in the upper right-hand side.
- ***Groups 1&2, 3&4, and 5&6 must be able to download each other histories before proceeding***





VEuPathDB Galaxy workflow has three major components: (1) mapping of raw reads to the reference genome, (2) calling variants, and (3) annotating variants. This workflow can be used to call single nucleotide polymorphisms, insertions and deletions (also defined as indels), and multiple nucleotide polymorphisms.

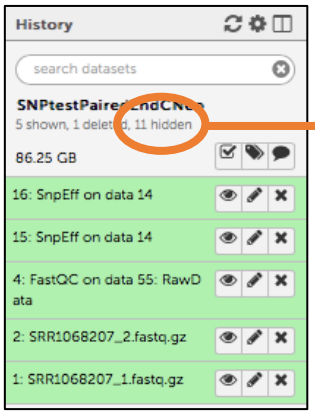
In this workflow, we used Bowtie2 to align and map sequences to a reference genome. Once they are aligned it may be worth checking the quality of this process because misalignments lead to false SNP calls. SAM or BAM files provide sore this information and you can find these files to export in the hidden workflow steps.

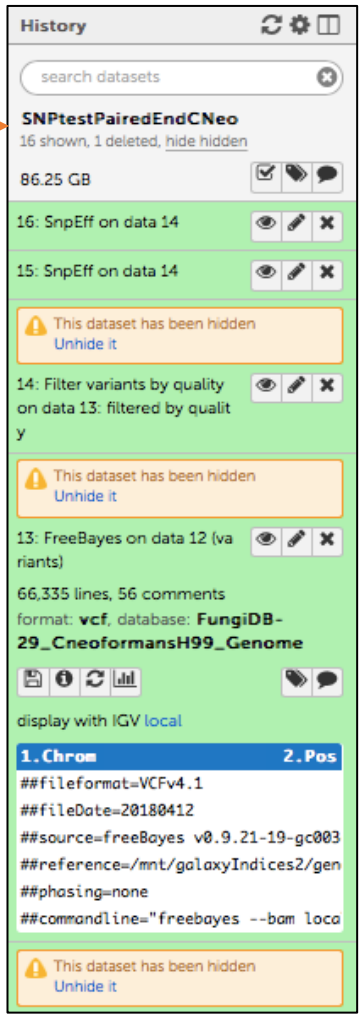
After reads have been aligned, they are sorted based on the chromosomal position. The tool that we are using is called Sort and it belongs to the suite of SAMtools. The sorted file is an input for downstream FreeBayes that calls SNPs and outputs into SnpEff that annotates variants.

Analysis and annotation of the genomic variants are carried out by the SnpEff tool. It uses reference genome to annotate genomic variants based on their genomic location and also predicts SNP coding effects. The genomic location features are intronic regions, 5’ and 3’ UTRs, and upstream, downstream, splice site and intergenic regions. SNP coding effects are categorised based on the effect of the amino acid change and are classified into synonymous and non-synonymous, gain or loss of start codons, gain of loss of stop codon, and frame shifts.

The SnpSift tool annotates, filters, and manipulates genomic annotated variants. Once you annotated your files using SnpEff, you can use SnpSift to help you filter large genomic datasets (e.g. sort on high or moderate impact SNPs, etc.).

2. Examine your results.





- Click on the *hidden* files link in the history panel to reveal all workflow output files.

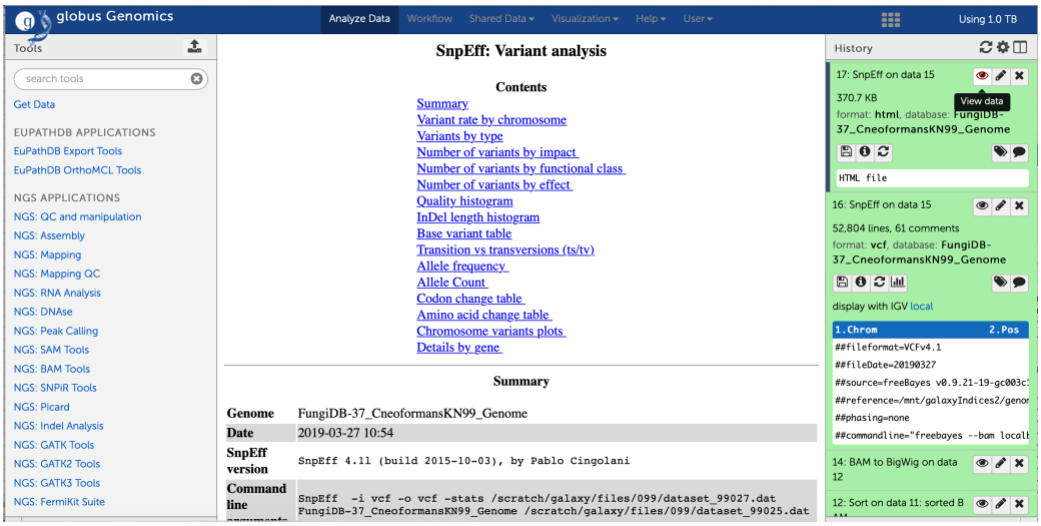
- Examine the output files.
- What does the tool FASTQC do?
- What about Sickle?

The output of Sickle is used by a program called Bowtie2. Bowtie generates a file called a BAM file. Whenever dealing with sequence alignment files you will likely hear of file formats called SAM or BAM. SAM stands for Sequence Alignment/Map format, and BAM is the binary version of a SAM file.

- Many of the downstream analysis programs that use BAM files require a sorted BAM file. This allows for more efficient analysis.
- The sorted BAM file is the input for a program called FreeBayes. This program is a Bayesian genetic variant detector designed to find small polymorphisms, specifically SNPs (single-nucleotide polymorphisms), indels (insertions and deletions), MNPs (multi-nucleotide polymorphisms), and complex events (composite insertion and substitution events) smaller than the length of a short-read sequencing alignment. The output for many variant callers is a file called a VCF file. VCF stands for variant interchange format.
- Examine the VCF file in your results (click on the *eye* icon to view its contents). Detailed information about VCF file content is available here: https://samtools.github.io/hts-specs/VCFv4.2.pdf

3. Examine SnpEff & FastQC summaries (html)

- Click on the *View data icon* (eye) in the SnpEff output file that has the html format.



This will open the html file right in galaxy where you can view it.
The header contains a short summary and information about the run and it has several major components:

Summary table that warns about possible genomic annotation errors or inconsistencies identified in the reference genome. If there are many, use caution interpreting results and examine associated gff files for any issues (*e.g.* missing feature values in gff files, incomplete gene sequences, more than one stop codon per gene, *etc.*).

Summary

Genome	FungiDB-29_AfumigatusAf293_Genome
Date	2017-05-02 17:10
SnpEff version	SnpEff 4.11 (build 2015-10-03), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /scratch/galaxy/files/005/dataset_5945.dat FungiDB-29_AfumigatusAf293_Genome /scratch/galaxy/files/005/dataset_5942.dat
Warnings	107,927
Errors	0
Number of lines (input file)	69,480
Number of variants (before filter)	71,442
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	71,442
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	1,962
Number of effects	330,861
Genome total length	29,420,142
Genome effective length	29,420,142
Variant rate	1 variant every 411 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
1_A_fumigatus_Af293	4,918,979	10,998	447
2_A_fumigatus_Af293	4,844,372	11,685	414
3_A_fumigatus_Af293	4,079,167	10,964	372
4_A_fumigatus_Af293	3,927,224	8,771	447
5_A_fumigatus_Af293	3,948,441	11,157	353
6_A_fumigatus_Af293	3,778,736	9,142	413
7_A_fumigatus_Af293	2,058,334	4,632	444
8_A_fumigatus_Af293	1,833,124	4,088	448
mito_A_fumigatus_Af293	31,765	5	6,353
Total	**29,420,142**	**71,442**	**411**

- Number of line (input file) - number of lines in vcf file
- Number of not variants: 0 - some packages report non-variant observations for nt positions between reference genome and vcf file generate.
- Number of known variants and multi-allelic VCF entries - if you work with a model organism where some variants were given an accession number (most commonly in mice and human projects) any recognised variants will be listed here
- Number of effects - SNP effects summary by type and regions
- Genome total length - number of bp in the reference genome
- Genome effective length - how many nucleotides can be mapped back to the

genome

- Variant rate - higher frequency of variants before samples can indicate selective pressure

Number variantss by type

Type	Total
SNP	114,034
MNP	12,864
INS	6,907
DEL	7,304
MIXED	2,180
INTERVAL	0
Total	143,289

Summary statistics for variant types

Here is an example of variant calls and what they mean in terms of nucleotide changes:

Type	What is means	Example
SNP	Single-Nucleotide Polymorphism	Reference = 'A', Sample = 'C'
Ins	Insertion	Reference = 'A', Sample = 'AGT'
Del	Deletion	Reference = 'AC', Sample = 'C'
MNP	Multiple-nucleotide polymorphism	Reference = 'ATA', Sample = 'GTC'
MIXED	Multiple-nucleotide and an InDel	Reference = 'ATA', Sample = 'GTCAGT'

Statistics for the variant effects and impacts:

- High impact normally refers to frame shift or new stop codon detections as those changes will generate profound effects on gene function.

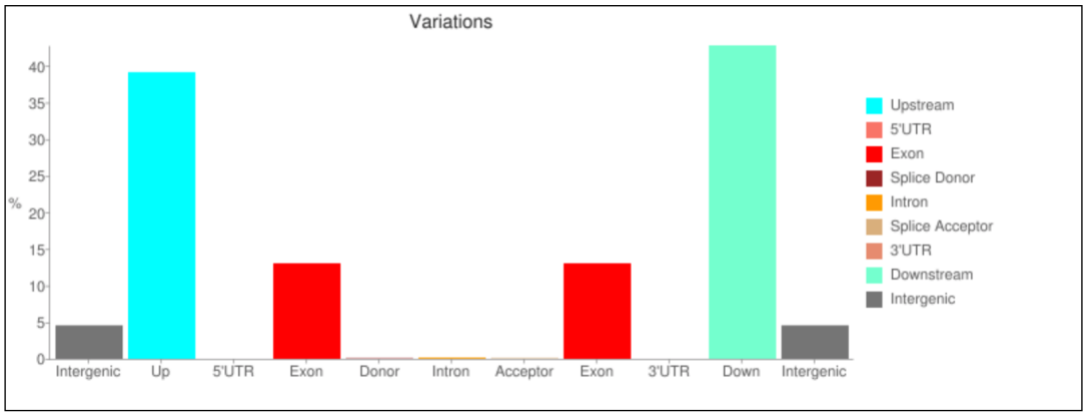
- Modifier SNPs can affect promoter function, while low and moderate SNPs are most commonly identified inside genes and are either non-coding or non-synonymous SNPs.

- Base changes summary. SnpEff html files provide a breakdown of SNPs across gene features:

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	21,588	35.949%
NONSENSE	131	0.218%
SILENT	38,332	63.832%

Type (alphabetical order)	Count	Percent
DOWNSTREAM	321,858	40.292%
EXON	67,505	8.451%
INTERGENIC	74,749	9.358%
INTRON	1,064	0.133%
NONE	1	0%
SPLICE_SITE_ACCEPTOR	5	0.001%
SPLICE_SITE_DONOR	4	0.001%
SPLICE_SITE_REGION	176	0.022%
TRANSCRIPT	12	0.002%
UPSTREAM	333,432	41.741%



Additionally, you may see several SNPs being reported at the same time: missense variant + splice region variant. This means that some SNPs that are found within certain splice sites also contain a missense variant. SNPs in the splice sequences may affect intron splicing and lead to read through.

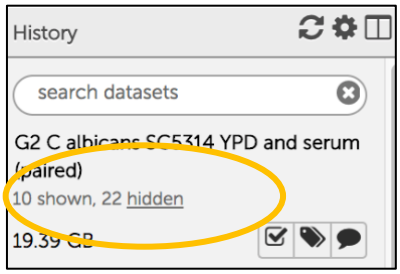
- Quality of reads is indicated in Phred's scale and is a good indicator of the quality of your datasets and results. Quality scores are normally represented by a bar graph where count = number of SNPs and X axis is quality score (higher score mean better p-values and high confidence of the results)

- Base changes: Reflects the frequency of base changes (purine-purine, purine-pyrimidine, pyrimidine-purine, pyrimidine-pyrimidine).

- Transition and transversion ratio helps to identify if you may have a selective pressure on certain alleles (high ratio suggests that genes may be under selective pressure).

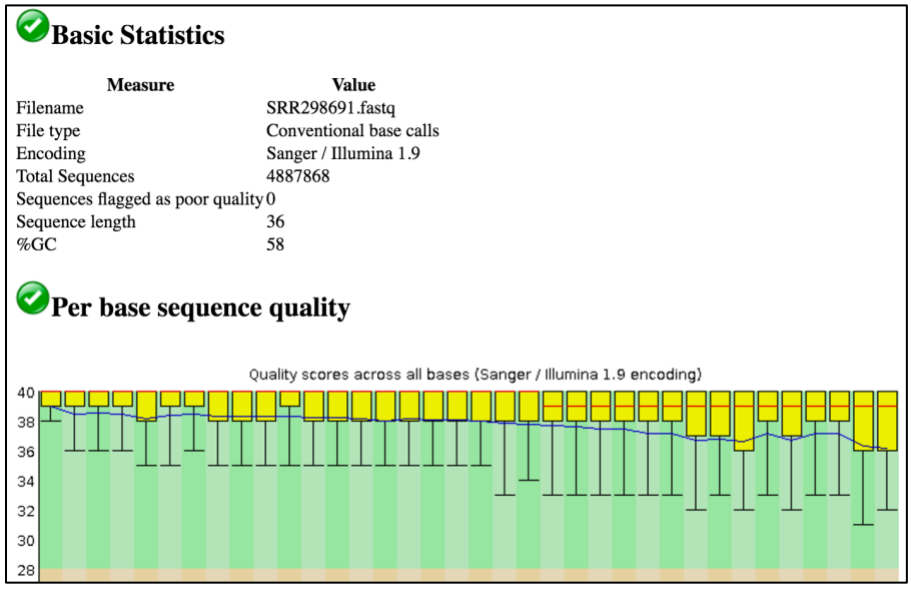
- Allele frequency statistics reports frequency of alleles and also help to identify potential sequencing artifacts due to PCR enrichment step (generation of heterozygous counts in a haploid organism).

The SNP workflow you are using is set up to generate certain files that will provide you with the information you can export and use further in your analysis.
If you select certain options, they will be shown in your history. If you do not select to display these files, you can view the output by clicking on displaying the hidden files from the history menu:



- **Examine sequence quality based on FastQC quality scores.**

FastQC provides an easy-to-navigate visual representation sequencing data quality and distribution of nucleotides per read position.



Basic Statistics

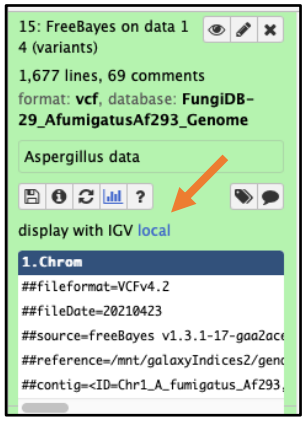
Measure	Value
Filename	SRR298691.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4887868
Sequences flagged as poor quality	0
Sequence length	36
%GC	58

Per base sequence quality

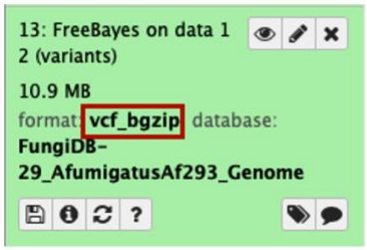
The vcf file generated by SnpEff contains information about SNPs and the genomic location. Post-processing of SNP data is normally required to make sense of thousands of SNPs and to decide which ones have biological and functional importance. Data processing can help you to extract SNP distribution and parse associated data including GeneIDs, protein-coding annotations, and effects in sequence ontology terms such as missense or synonymous variants, stop codon gain, etc. and also link changes to the genome model. SnpSift is among other programs that is often in SNP data post-processing. It can be installed and run locally to manipulate vcf files. Alternatively, you can also visualize vcf files in Artemis (additional steps are required to format the data)

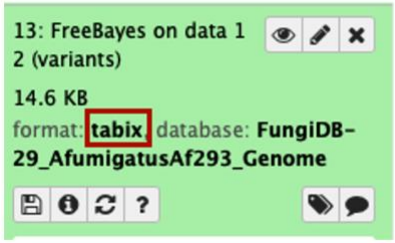
4. **View VCF file results in the JBrowse genome browser:**

In order to view a VCF file in JBrowse, it first has to be indexed and compressed. We will use a trick in galaxy to produce the required files*:*

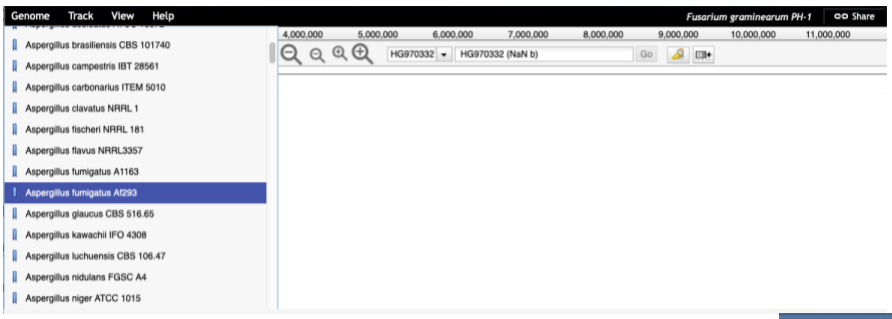


- Find the VCF file you would like to view in JBrowse. This can be the FreeBayes output or the quality filtered VCF. Click on this file to expose the available options.
- Click on "display with IGV local".
- A new window will open up (you are not going to use this window).
- Go back to the galaxy window, you will notice that the number of hidden files has increased by 2 files.
- Show the hidden files by clicking on the word hidden.
- The two new files you want are in the vcf_bgzip and tabix format:

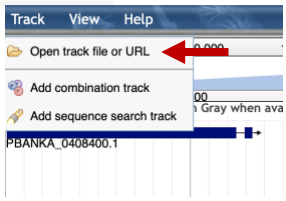




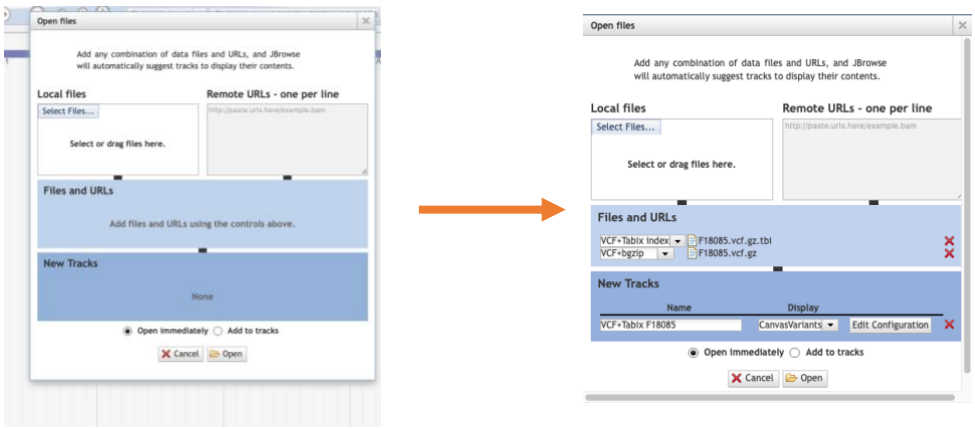
- Download both files by clicking on the download icon. *You will need both files*.
- After the files are downloaded, After the files are downloaded, rename them as follows:
    a. Rename the *Galaxy13-[FreeBayes_on_data_12_(variants)].vcf_bgzip* file to *sample.vcf.gz* (i.e. F18085.vcf.gz)
    b. Rename the *Galaxy13-[FreeBayes_on_data_12_(variants)].tabix* file to sample.vcf.gz.tbi (i.e. F18085.vcf.gz.tbi)
- Next, navigate to the Tool section in FungiDB, click on Genome Browser link and select the reference genome from the Genome drop down list:



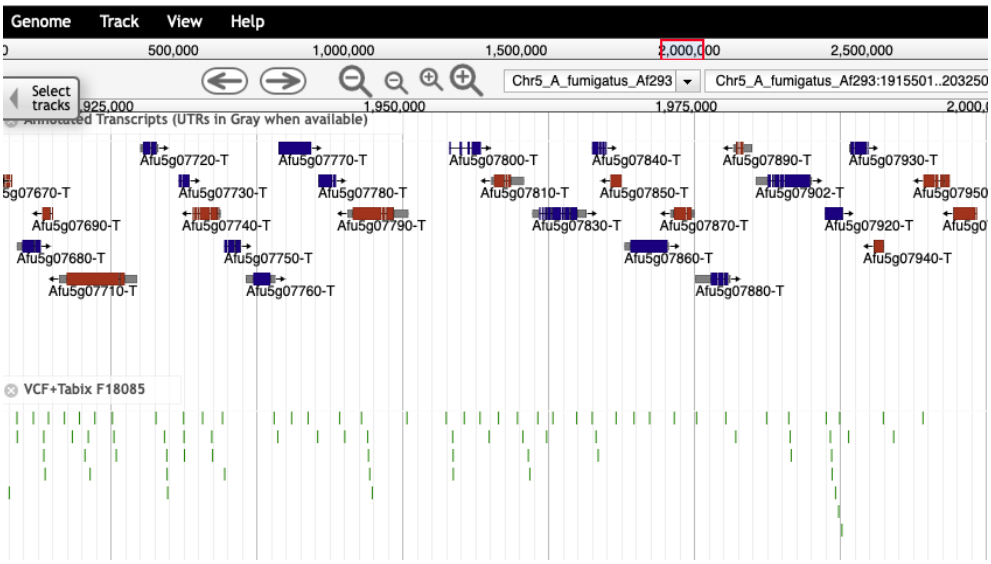
- Click on the Track menu, select "Open track file or URL"



- Drag and drop your files in the window that appears. Notice that the file formats are autodetected.



- Click on “Open”. You should see SNP positions displayed in a new track. You can zoom in and click on the SNPs to get more info.



5. **Download vcf files and evaluate workflow results**.

The vcf file generated by SnpEff contains information about SNPs and the genomic location.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
CM001231	189057	.	AG	CT	787.449	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:143:0:0:143:5341:-207.887,-43.0473,0
CM001231	483825	.	G	A	64.8756	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:4:0:0:4:146:-10.0999,-1.20412,0
CM001231	518226	.	G	C	51.7908	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:8:0:0:7:276:-11.5007,-2.10721,0
CM001231	574021	.	C	G	237.265	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:17:0:0:17:583:-39.079,-5.11751,0
CM001231	609879	.	GAA	CAG	55.2785	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:32:8:277:22:861:-18.1711,-0.694735,0
CM001231	1090073	.	G	T	79.4156	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:8:2:75:6:238:-11.5539,-1.36362,0
CM001231	1090104	.	A	T	70.961	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:6:0:0:6:220:-12.5146,-1.80618,0
CM001231	1153611	.	CCTC	GCTG	111.123	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:8:5:188:3:97:-9.30616,-6.1461,0
CM001231	1159150	.	CT	GC	126.126	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:31:0:0:19:741:-29.7713,-5.71957,0
CM001231	1159438	.	C	G	82.3312	.	AB=0;ABP=0;	GT:DP:RO:QF	0/0:47:30:1092:17:640:0,-9.53002,-3.50705
CM001231	1159465	.	G	C	249.656	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:126:47:1770:79:3013:-53.8644,-25.2134,0
CM001231	1159499	.	T	C	124.95	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:143:32:1167:111:4248:-76.1575,-33.4865,0
CM001231	1181576	.	CC	TG	191.675	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:27:0:0:25:924:-41.7448,-7.52575,0
CM001231	1293309	.	C	G	51.22	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:2:0:0:2:78:-6.92763,-0.60206,0
CM001231	1323058	.	TT	GC	71.3001	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:6:0:0:6:223:-12.5485,-1.80618,0
CM001231	1485397	.	A	G	3558.42	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:499:0:0:497:18671:-804.678,-149.612,0
CM001231	1485429	.	G	A	3783.33	.	AB=0;ABP=0;	GT:DP:RO:QF	1/1:517:1:38:516:20010:-843.425,-151.978,0

Post-processing of SNP data is normally required to make sense of thousands of SNPs and to decide which ones have biological and functional importance. Data processing can help you to extract SNP distribution and parse associated data including GeneIDs, protein-coding annotations, and effects in sequence ontology terms such as missense or synonymous variants, stop codon gain, etc. and also link changes to the genome model. We will work with this file in the next section.

Filtering data in VCF files

VCF files contain a lot of data about variants and their positions.
SnpEff generates various analyses/summaries of VCF files (including GeneIDs that overlap variant positions). However, it is often necessary to filter VCF files further to obtain useful information for your specific question. For example, you may want to filter out SNP positions that have an impact on the coding sequence.

One tool that can be used is called SnpSift Filter (look at the last step of the pipeline you just ran). This tool allows you to write complex expressions to filter a VCF file. The following expression has been used to filter the VCF file on moderate and high impact SNPs as a part of the workflow (this setting can be adjusted by opening the workflow in the Galaxy workflow editor):

(((ANN[*].IMPACT has 'HIGH') | (ANN[*].IMPACT has 'MODERATE')) & ((na FILTER) | (FILTER = 'PASS')))

6. **Extract filtered VCF file (SnpSift output) and convert into an Excel document.**

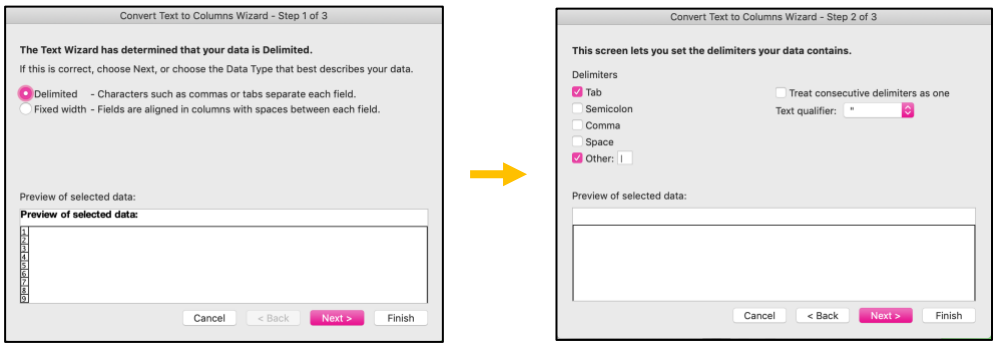
*** Groups 1&2, 3&4, 5&6, you will need VCF files from SnfSift generate by both groups , respectively***

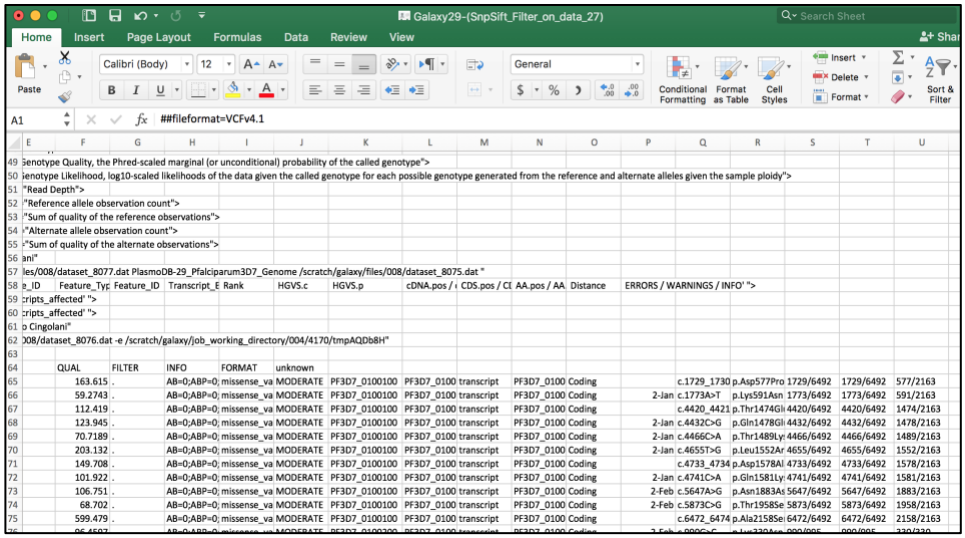
- Examine the filtered VCF file. Notice that the Gene IDs are buried in the file, but the file has some structure which means you can extract them either programmatically or using a program like Excel.

Here are some steps you can take to extract Gene IDs from two VCF files then compare them to identify genes that are in common or that distinguish the two files.

1. Download the SnpSift Filter output by clicking on the save icon
2. Open this file using Excel.

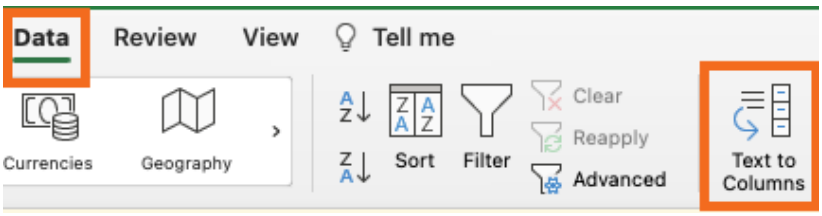
If the file doesn't open automatically, make sure you select tabs and | as column delimiters

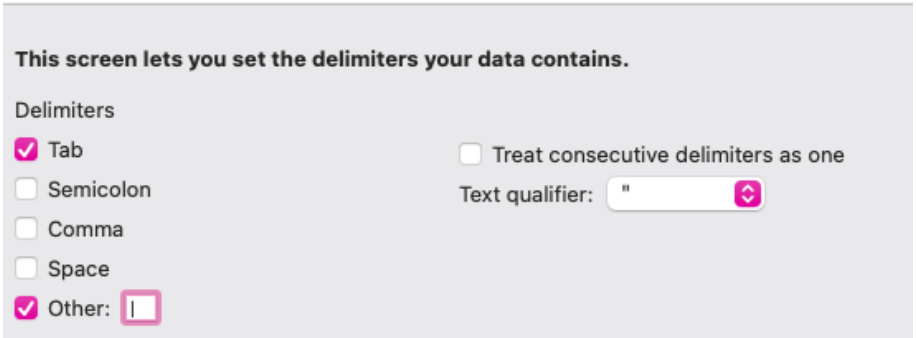


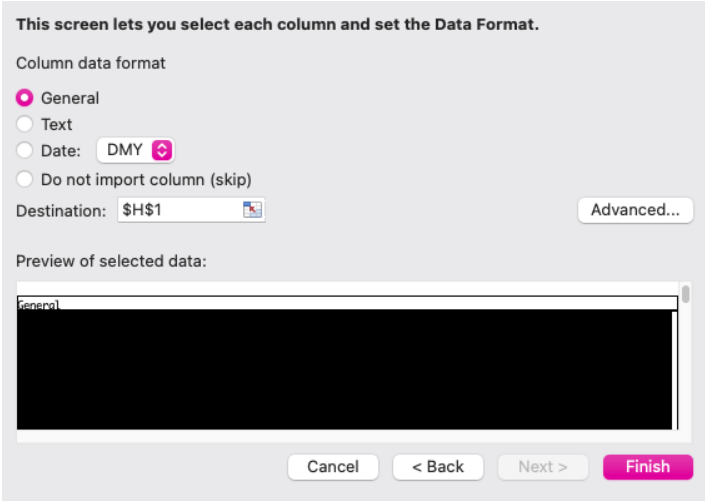


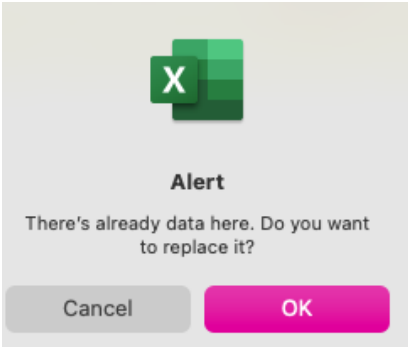
If the file is opened by the Excel automatically:

- select the *INFO* column
- navigate to the *Data* tab in Excel
- select *Text to Column* option
- input | into the "Other" option



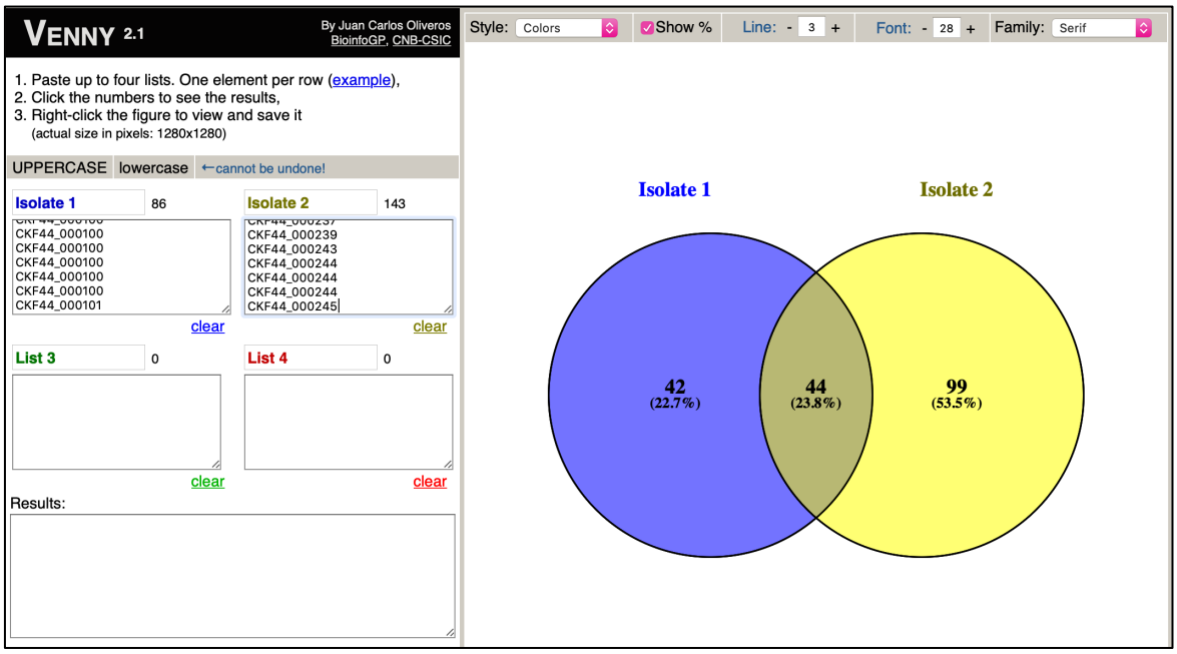






3. Now you can look for Gene IDs of interest in the excel file. For example, if this is a known drug resistant line you can find the gene responsible for the resistance and see what kinds of SNPs are present.

4. If you are comparing a two different strains, you can extract gene IDs from both VCF files and use a website like http://bioinfogp.cnb.csic.es/tools/venny/ to generate a Venn diagram.



*Note that in the above steps you are ultimately comparing gene IDs – do you think you might be missing some important polymorphisms using this method? Of course, the answer is yes☺

It is quite possible that a gene with a SNP in the WT and a SNP in the mutant that will be in the intersection of the two gene lists, contains different SNPs – you will miss this by doing the above steps. Below is a description of steps you can take to create a list of unique IDs for SNPs. This list of unique IDs can then be used in Venny.

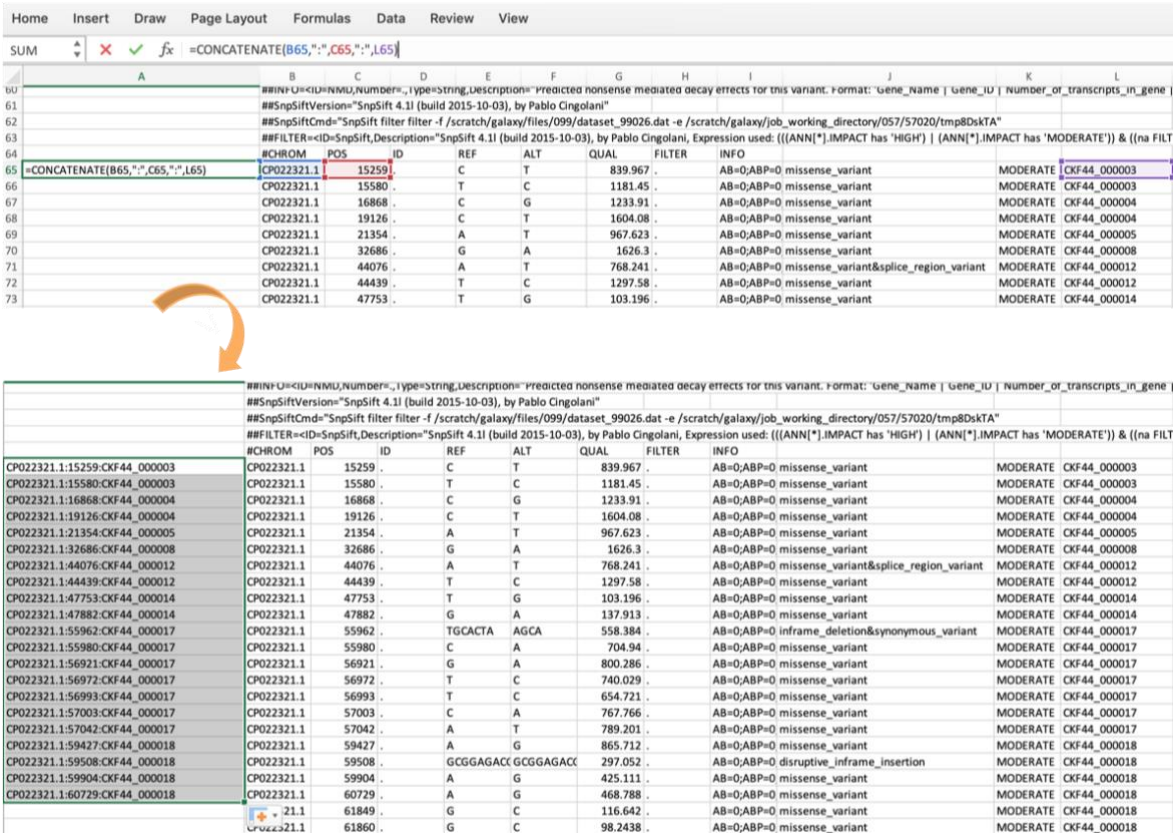
7. Analyse your data in Venny

1. Start with the same excel files that you opened in the above section. Insert an empty column before the data.
2. To create a unique ID for SNPs we will combine information from multiple columns to create something that looks like this: chromosome:position:geneID
3. To do this you will use the concatenate function in Excel:
   =concatenate(cell#1,":",cell#2,":",cell#3)

   Cell#1 = cell with chromosome number

   Cell#2 = cell with position

   Cell#3 = cell with GeneID



4. You should get unique SNP IDs that look like this (for example): CP022321.1:15259:CKF44_000003
5. Copy this function to the rest of the column to replicate the concatenate function.
6. Copy these newly generated unique IDs into Venny and compare your data.