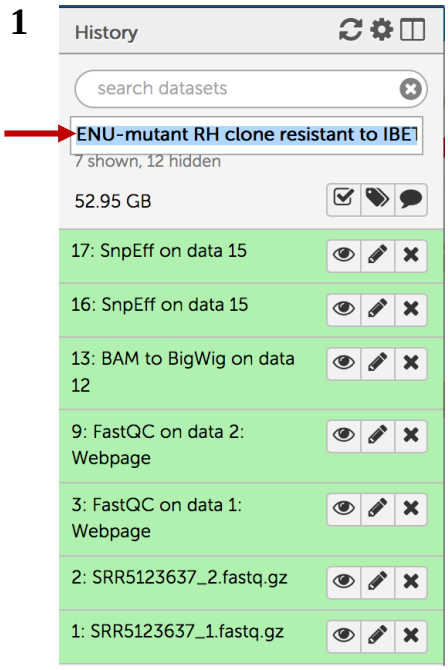
Analyzing Variant Call results using EuPathDB Galaxy, Part II

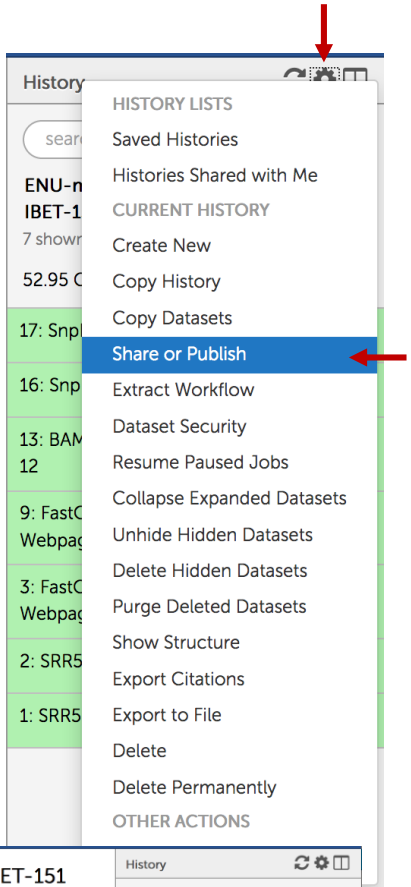
In this exercise, we will work in groups to examine the results from the SNP analysis workflow that we started yesterday. *The first step is to share your SNP workflow histories with the rest of the workshop participants:*

1. Give your workflow a meaningful name, eg. The sample or group name.
2. Click on the on the 'History options' link and select the 'share or Publish option'.
3. On the next page click on the 'Make History Accessible and Publish' link.

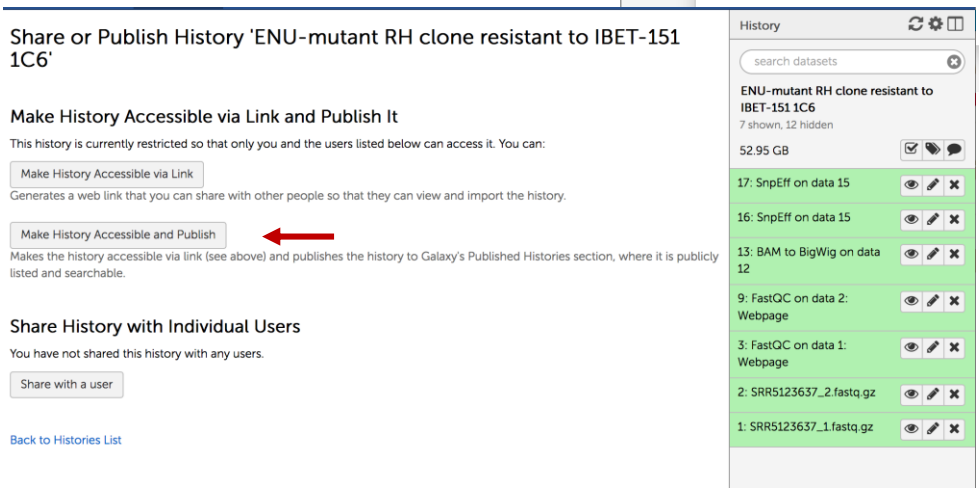
1



2

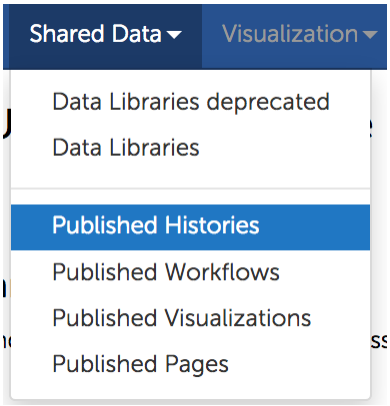


3

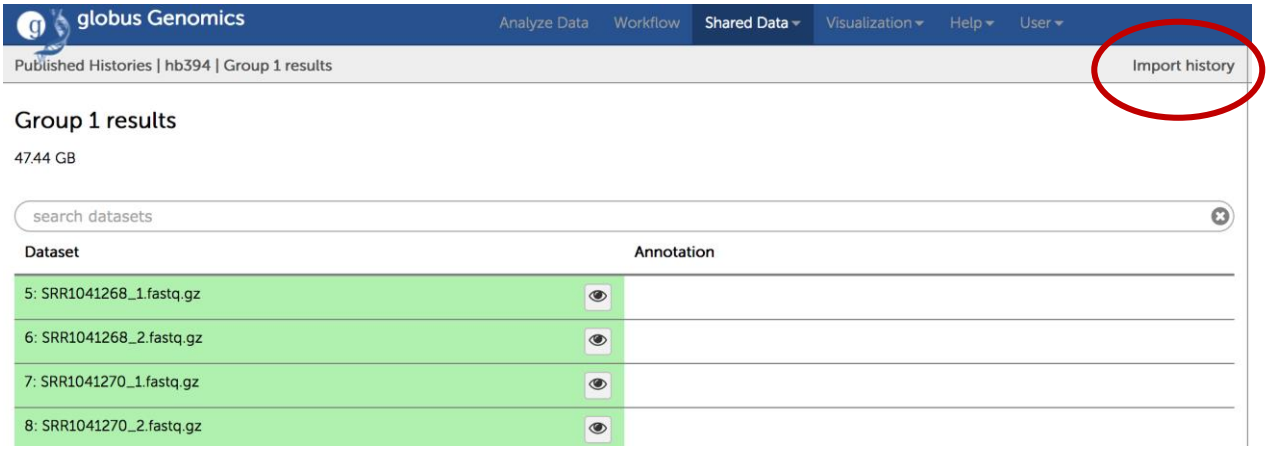


To import a shared history into your workspace follow these steps:

1. Select 'Published Histories' from the Shared data menu.

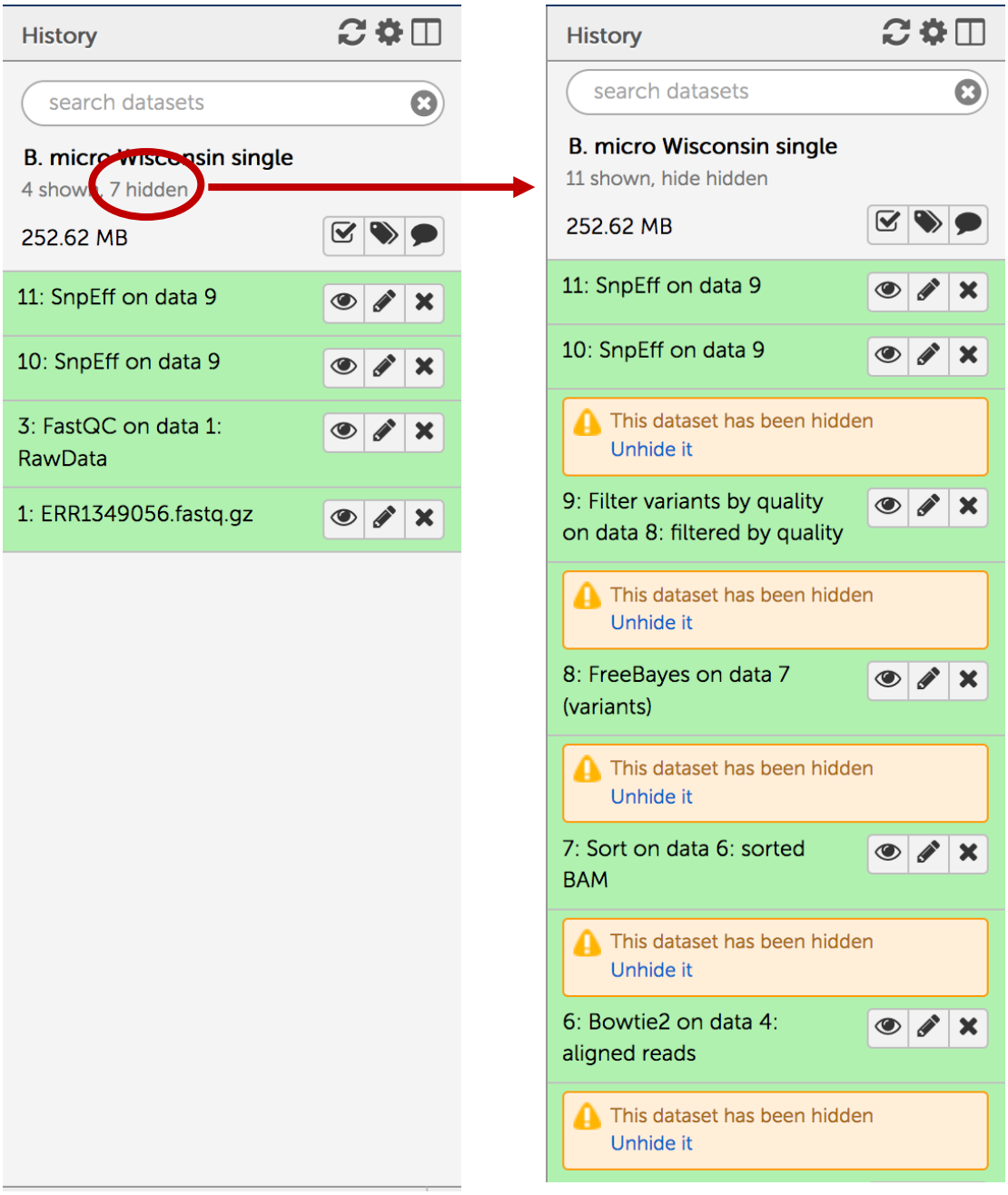


2. From the list of shared histories click on the one you want to import and on the next page select the 'Import' link in the upper right hand side.



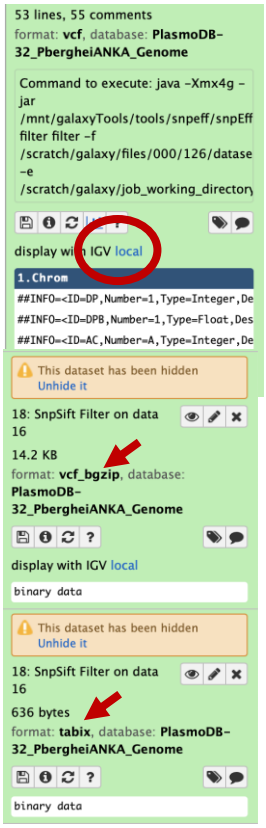
Examining your results:

1. Click on the hidden files link in the history panel to reveal all workflow output files.
2. Examine the output files. What does the tool FASTQC do? What about Sickle?



3. The output of Sickle is used by a program called Bowtie2. What does this tool do? 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.

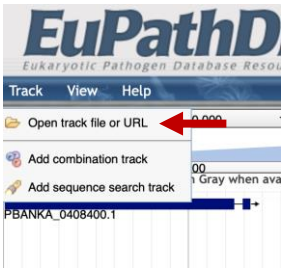
4. Many of the downstream analysis programs that use BAM files require a sorted BAM file. This allows access to reads to be done more efficiently.
5. The sorted BAM file is the input for a program called FreeBayes. This program is a Bayesian genetic variant detector designed to find small polymorphisms, 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.
6. 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
7. What does tool SnpEFF do? SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes).

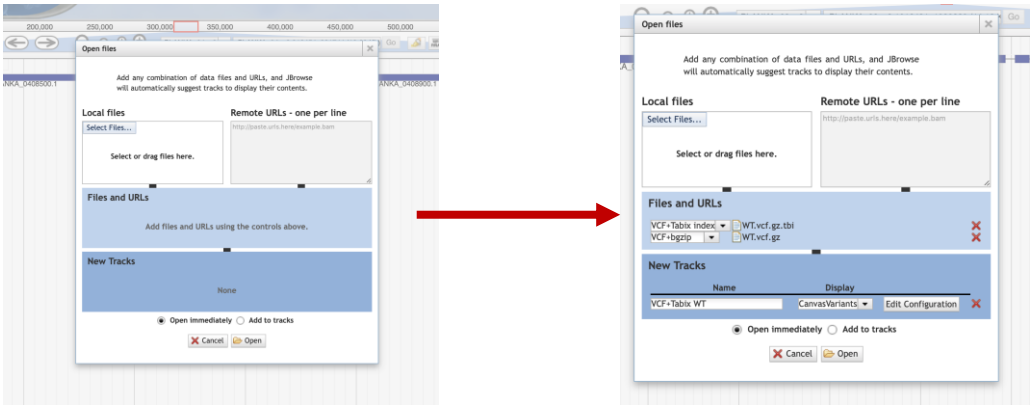


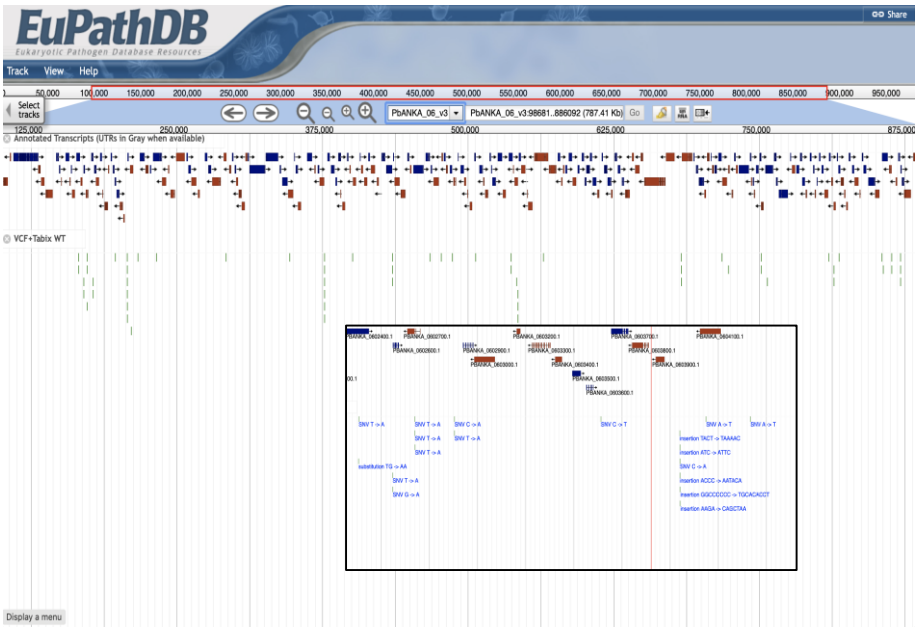
Viewing VCF file results in the JBrowse genome browser:

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:

1. 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.
2. Click on "display with IGV local" (red circle).
3. A new window will open up (you are not going to use this window).
4. Go back to the galaxy window, you will notice that the number of hidden files has increased by 2 files.
5. Show the hidden files by clicking on the word hidden.
6. The two new files you want are in the format of vcf_bgzip and tabix format (red arrows).
7. Download both files by clicking on the download icon. *You will need both files*.
8. After the files are downloaded, rename them as follows:
    a. The vcf_bgzip file to sample.vcf.gz (i.e. WT.vcf.gz)
    b. The tabix file to sample.vcf.gz.tbi (i.e. WT.vcf.gz.tbi)
9. Next go to JBrwose for the organism of interest (*see below for shortcuts*) and click on the Track menu, select "Open track file or URL"
10. Drag and drop your files in the window that appears. Notice that the file formats are autodetected.
11. 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.







JBrowse shortcut links:

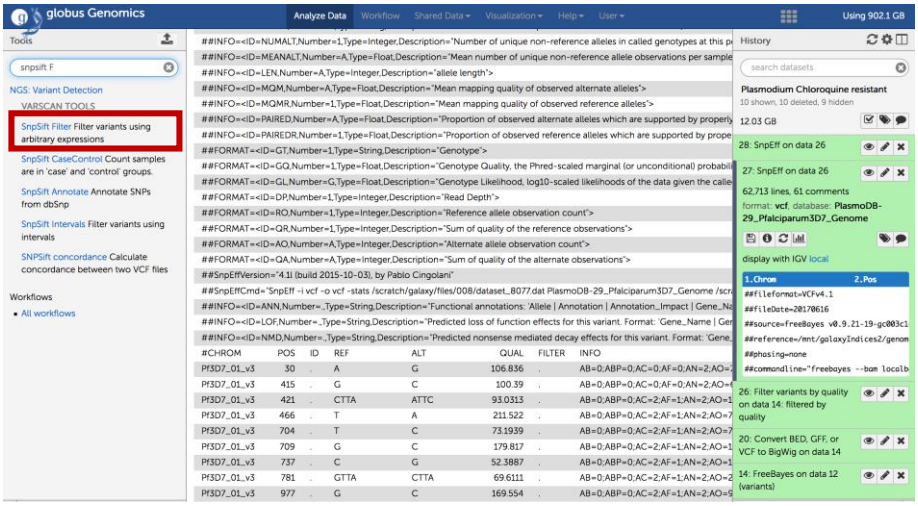
P. berghei:
https://plasmodb.org/a/jbrowse/index.html?data=tracks%2Fpfal3D7&loc=Pf3D7_11_v3%3A1309337..1350904&tracks=gene&highlight=

C. parvum
https://cryptodb.org/a/jbrowse/index.html?data=tracks%2FcparIowaII&loc=CM000431%3A786265..818935&tracks=gene&highlight=

T. gondii
https://toxodb.org/a/jbrowse/index.html?data=tracks%2FtgonME49&loc=TGME49_chrVI%3A468507..509947&tracks=gene&highlight=

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. This tool allows you to write complex expressions to filter a VCF file.

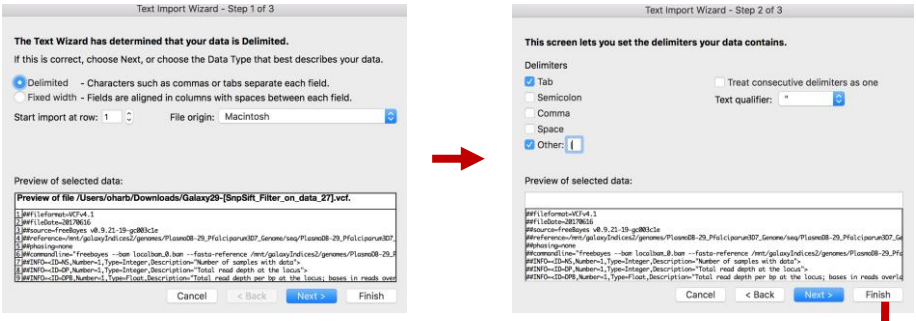


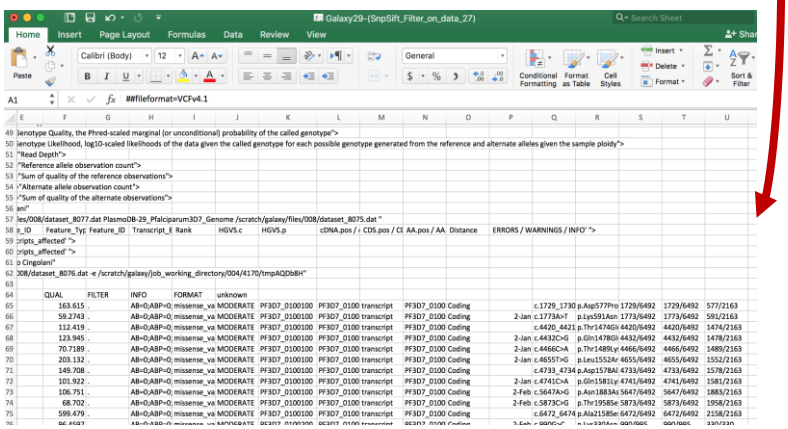
The galaxy workflow we used already includes SnpSift Filter as the final step using the following expression:
(((ANN[*].IMPACT has 'HIGH') | (ANN[*].IMPACT has 'MODERATE')) & ((na FILTER) | (FILTER = 'PASS')))

- Examine the filtered VCF file. Notice that the GeneIDs 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.

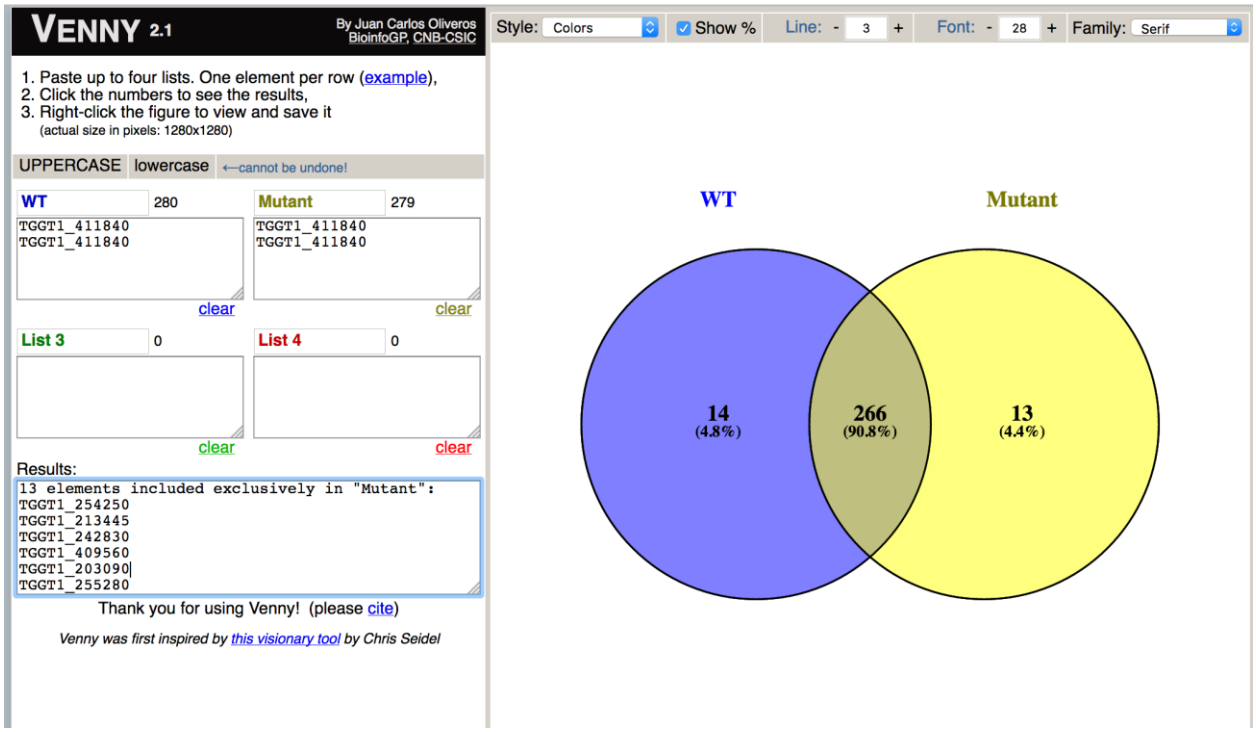
Here are some steps you can take to extract Gene IDs from two VCF files then compare them to identify genes that are in common or that distinguish the two files.

1. Download the SnpSift Filter output by clicking on the save icon
2. Open this file using excel and make sure you select tabs and | as column delimiters





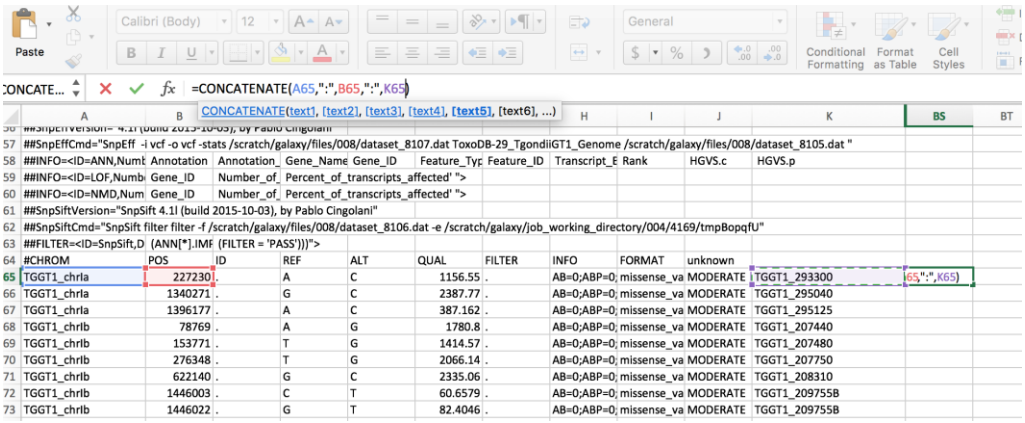
3. Now you can look for Gene IDs of interest in the excel file. For example, if this is a known drug resistant line you can find the gene(s) that might be responsible for the resistance and see what kinds of SNPs are present.
4. If you are comparing a mutant and a wild type or two different strains you can extract gene IDs from both VCF files and use a website like http://bioinfogp.cnb.csic.es/tools/venny/

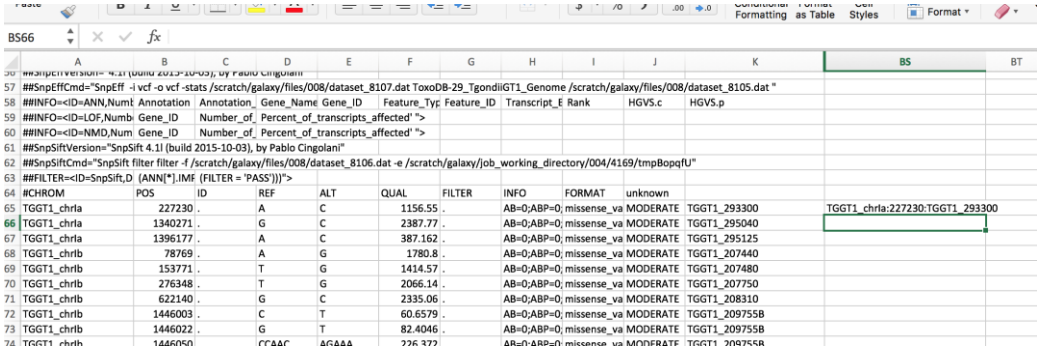


*Note that in the above steps you are ultimately comparing gene IDs – do you think you might be missing some important polymorphisms using this method? Of course, the answer is yes☺

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.

1. Start with the same excel files that you opened in the above section.
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

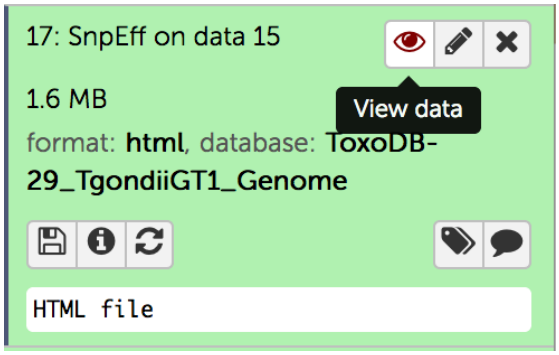




4. You should get unique SNP IDs that look like this (for example): TGGT1_chrIb:1446003:TGGT1_209755B
5. Copy this function to the rest of the column to replicate the concatenate function.
6. Copy the these newly generated unique IDs into Venny and compare the mutant and wild type.

Examining SnpEff summary:

- Click on the view 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: SEP

1. 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 (ex. missing feature values in gff files, incomplete gene sequences, more than one stop codon per gene, etc.).
2. Summary statistics for variant types

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

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'

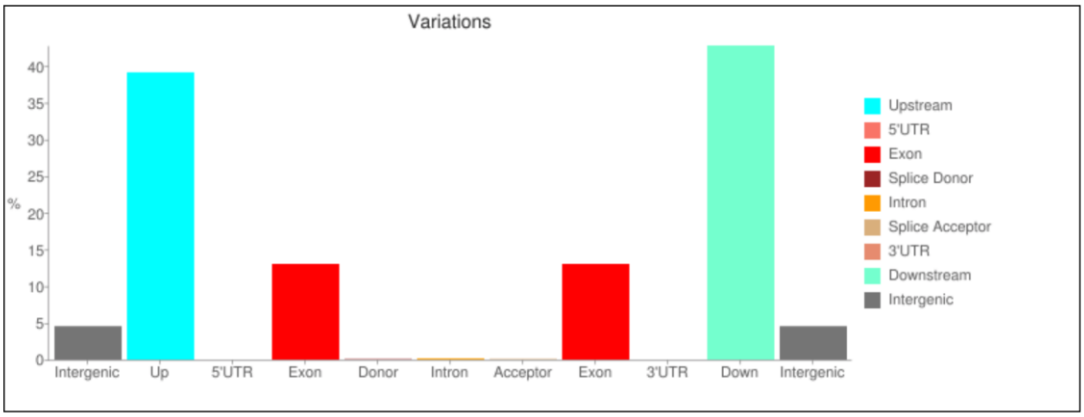
3. Statistics for the variant effects and impacts:

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%

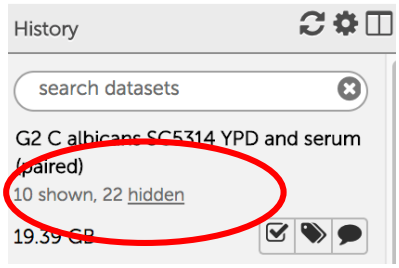
Type (alphabetical order)	Count	Percent
DOWNSTREAM	321,858	40.292%
EXON	67,505	8.451%
INTERGENIC	74,749	9.358%
INTRON	1,064	0.133%
NONE	1	0%
SPLICE_SITE_ACCEPTOR	5	0.001%
SPLICE_SITE_DONOR	4	0.001%
SPLICE_SITE_REGION	176	0.022%
TRANSCRIPT	12	0.002%
UPSTREAM	333,432	41.741%

Base changes summary. SnpEff html files provides a break down of SNPs across gene features:



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 (yellow stars).

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:



Now, lets take a look at the files generated by the workflow and steps that you can take to further evaluate them.

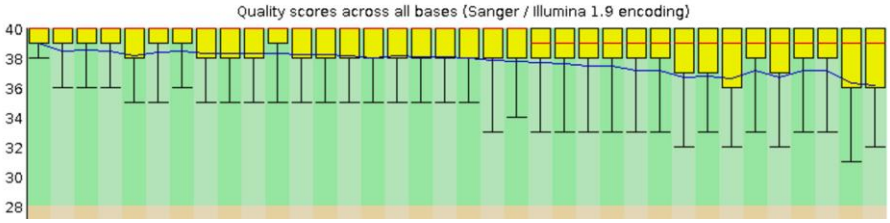
1. Examine sequence quality based on FastQC quality scores.

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



2. 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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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:QI	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.

Summary

Genome	ToxoDB-29_TgondiiGT1_Genome
Date	2017-06-17 05:56
SnpEff version	`SnpEff 4.1l (build 2015-10-03), by Pablo Cingolani`
Command line arguments	`SnpEff -i vcf -o vcf -stats /scratch/galaxy/files/008/dataset_8107.dat ToxoDB-29_TgondiiGT1_Genome /scratch/galaxy/files/008/dataset_8105.dat`
Warnings	3,941
Errors	0
Number of lines (input file)	8,411
Number of variants (before filter)	8,483
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	8,483
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	72
Number of effects	14,149
Genome total length	63,945,332
Genome effective	