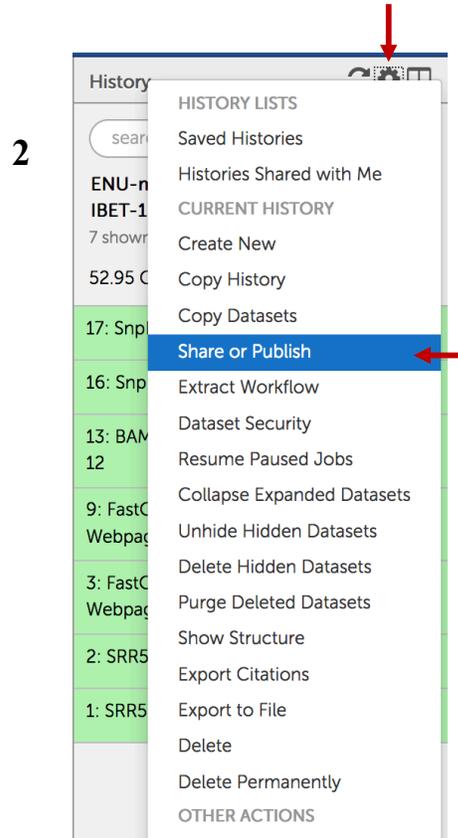
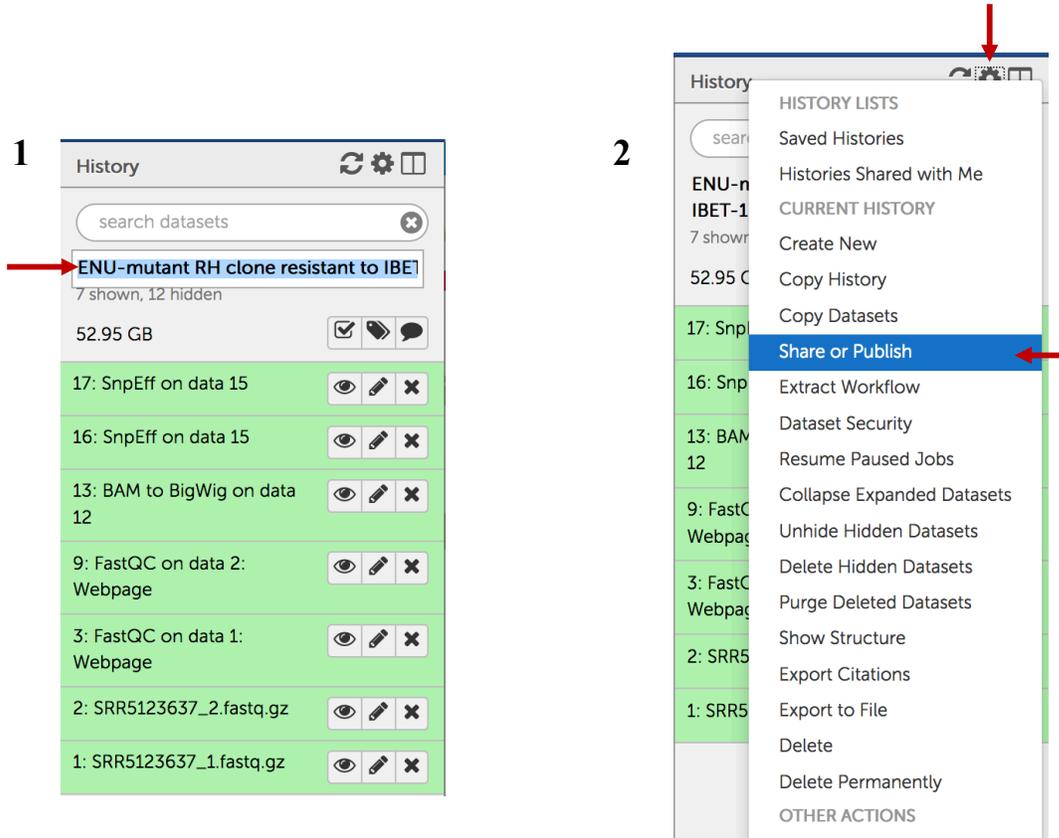


Analyzing Variant Call results using EuPathDB Galaxy, Part II

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.



3 Share or Publish History 'ENU-mutant RH clone resistant to IBET-151 1C6'

Make History Accessible via Link and Publish It

This history is currently restricted so that only you and the users listed below can access it. You can:

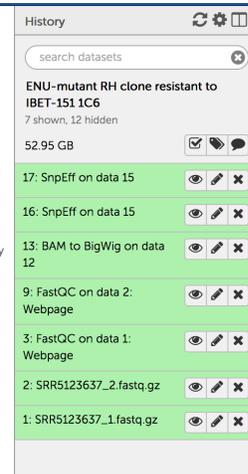
Generates a web link that you can share with other people so that they can view and import the history.

Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Share History with Individual Users

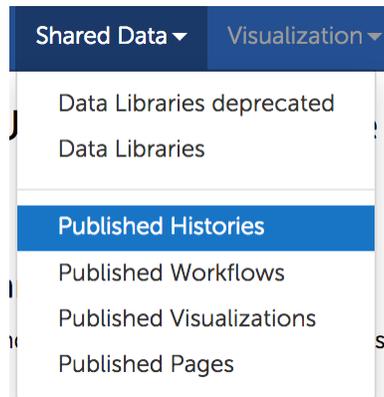
You have not shared this history with any users.

[Back to Histories List](#)

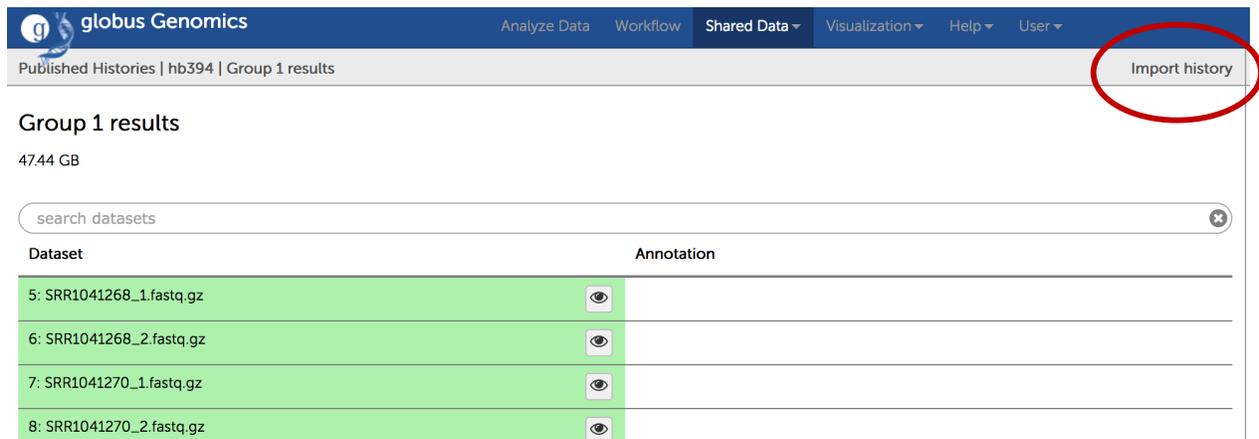


To import a shared history into your workspace follow these steps:

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.

A screenshot of the globus Genomics web interface. The top navigation bar includes 'globus Genomics', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. Below the navigation bar, the breadcrumb 'Published Histories | hb394 | Group 1 results' is visible. On the right side of the page, the 'Import history' link is circled in red. The main content area shows 'Group 1 results' with a size of '47.44 GB'. Below this is a search bar labeled 'search datasets' and a table with two columns: 'Dataset' and 'Annotation'. The table contains four rows of data, each with a green background and an eye icon in the 'Annotation' column.

Dataset	Annotation
5: SRR1041268_1.fastq.gz	
6: SRR1041268_2.fastq.gz	
7: SRR1041270_1.fastq.gz	
8: SRR1041270_2.fastq.gz	

Examining your results:

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

The image displays two side-by-side screenshots of a workflow history panel. The left panel shows a workflow named "B. micro Wisconsin single" with 4 shown and 7 hidden files. A red circle highlights the "7 hidden" text, and a red arrow points to the right panel. The right panel shows the same workflow with all 11 files revealed. The files are listed as follows:

- 11: SnpEff on data 9
- 10: SnpEff on data 9
- 3: FastQC on data 1: RawData
- 1: ERR1349056.fastq.gz
- 9: Filter variants by quality on data 8: filtered by quality
- 8: FreeBayes on data 7 (variants)
- 7: Sort on data 6: sorted BAM
- 6: Bowtie2 on data 4: aligned reads

Several files are marked as hidden with orange warning boxes and "Unhide it" links:

- This dataset has been hidden Unhide it

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 a genome browser:

In order to view a VCF file in GBrowse, it first has to be converted to a format that GBrowse can understand like BigWig. To do this follow these steps:

1. Click on the edit attributes icon on the FreeBayes VCF output file.
2. In the central window click on the 'Convert Format' tab.
3. Next select the 'Convert BED, GFF or VCF to BigWig' option and click on the 'Convert' link.
4. Notice a new step will appear in you history for the conversion step.

```
##fileformat=VCFv4.1
##fileDate=20170617
##source=freeBayes v0.9.21-19-gc003c1e
##reference=/mnt/galaxyIndices2/genome
##phasing=none
##commandLine="freebayes --bam localba
```

Attributes Convert Format Datatype Permissions

Convert to new format

Convert VCF to BGZIP

Convert VCF to VCF_BGZIP

Convert Vcf to tabix

Convert BED, GFF, or VCF to BigWig

Contents of this dataset converted to a new format.

- Once the conversion is done, you can click on the view in GBrowse link to go to the appropriate EuPathDB website and view variant locations.

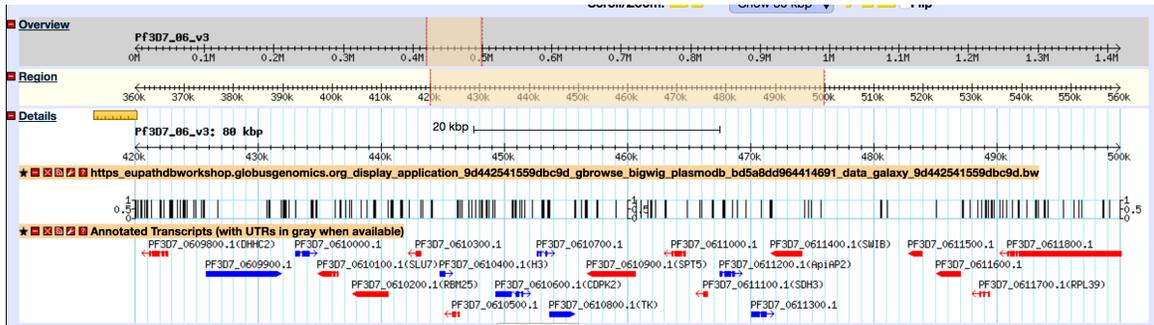
20: Convert BED, GFF, or VCF to BigWig on data 14

758.2 KB

format: bigwig, database: PlasmoDB-29_Pfalciparum3D7_Genome

[Display in PlasmoDB GBrowse](#)

Binary UCSC BigWig file



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 GenIDs 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 screenshot shows the globus Genomics interface. On the left, the 'Tools' section has 'SnpSift Filter Filter variants using arbitrary expressions' highlighted with a red box. The main area shows the tool's configuration with various filters and a table of filtered variants.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
PSD7_01_v3	30	.	A	G	106.836	.	AB=0;ABP=0;AC=0;AF=0;AN=2;AO=2
PSD7_01_v3	415	.	G	C	100.39	.	AB=0;ABP=0;AC=0;AF=0;AN=2;AO=2
PSD7_01_v3	421	.	CTTA	ATTC	93.0313	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1
PSD7_01_v3	466	.	T	A	211.522	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=7
PSD7_01_v3	704	.	T	C	73.1939	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=7
PSD7_01_v3	709	.	G	C	179.817	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1
PSD7_01_v3	757	.	C	G	52.3887	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1
PSD7_01_v3	781	.	GTTA	CTTA	69.6111	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2
PSD7_01_v3	977	.	G	C	169.954	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=5

 SnpSift Filter Filter variants using arbitrary expressions (Galaxy Tool Version latest) ▼ Options

VCF input

   27: SnpEff on data 26 ▼

Expression

SnpSift filter

You can filter via vcf file using arbitrary expressions, for instance "(QUAL > 30) | (exists INDEL) | (countHet() > 2)". The actual expressions can be quite complex, so it allows for a lot of flexibility.

Some examples:

I want to filter out samples with quality less than 30:
 (QUAL > 30)
...but we also want InDels that have quality 20 or more:
 ((exists INDEL) & (QUAL >= 20)) | (QUAL >= 30)
...or any homozygous variant present in more than 3 samples:
 (countHom() > 3) | ((exists INDEL) & (QUAL >= 20)) | (QUAL >= 30)
...or any heterozygous sample with coverage 25 or more:
 ((countHet() > 0) & (DP >= 25)) | (countHom() > 3) | ((exists INDEL) & (QUAL >= 20)) | (QUAL >= 30)
I want to keep samples where the genotype for the first sample is homozygous variant and the genotype for the second sample is reference:
 isHom(GEN[0]) & isVariant(GEN[0]) & isRef(GEN[1])

For complete details about this tool and expressions that can be used, please go to <http://snpeff.sourceforge.net/SnpSift.html#filter>

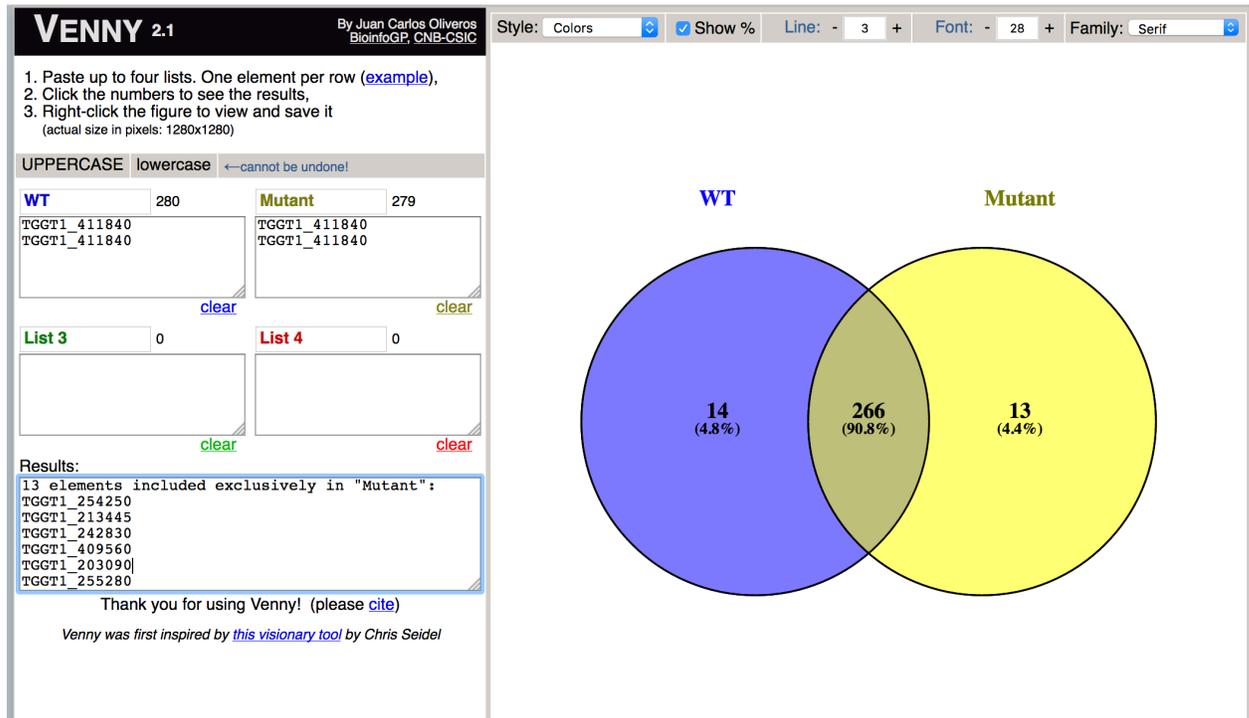
To filter your VCF file based on variant impact do the following:

1. Select the VCF input file – for this exercise select the SnpEff output file, make sure you select the one that is a VCF file not the one that is the html output.
2. In the expression box copy and paste this expressions:

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

3. Click on the 'Execute' button.

- 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.
- 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.

- Start with the same excel files that you opened in the above section.
- To create a unique ID for SNPs we will combine information from multiple columns to create something that looks like this: chromosome:position:geneID
- 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

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
TGGT1_chr1a	227230		A	C	1156.55		AB=0;ABP=0;missense_va	MODERATE	TGGT1_293300
TGGT1_chr1a	1340271		G	C	2387.77		AB=0;ABP=0;missense_va	MODERATE	TGGT1_295040
TGGT1_chr1a	1396177		A	C	387.162		AB=0;ABP=0;missense_va	MODERATE	TGGT1_295125
TGGT1_chr1b	78769		A	G	1780.8		AB=0;ABP=0;missense_va	MODERATE	TGGT1_207440
TGGT1_chr1b	153771		T	G	1414.57		AB=0;ABP=0;missense_va	MODERATE	TGGT1_207480
TGGT1_chr1b	276348		T	G	2066.14		AB=0;ABP=0;missense_va	MODERATE	TGGT1_207750
TGGT1_chr1b	622140		G	C	2335.06		AB=0;ABP=0;missense_va	MODERATE	TGGT1_208310
TGGT1_chr1b	1446003		C	T	60.6579		AB=0;ABP=0;missense_va	MODERATE	TGGT1_209755B
TGGT1_chr1b	1446022		G	T	82.4046		AB=0;ABP=0;missense_va	MODERATE	TGGT1_209755B

- You should get unique SNP IDs that look like this (for example):
TGGT1_chr1b:1446003:TGGT1_209755B
- Copy this function to the rest of the column to replicate the concatenate function.
- 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.

17: Snpeff on data 15

1.6 MB

format: **html**, database: **ToxoDB-29_TgondiiGT1_Genome**

View data

HTML file

- 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:

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 variants 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'

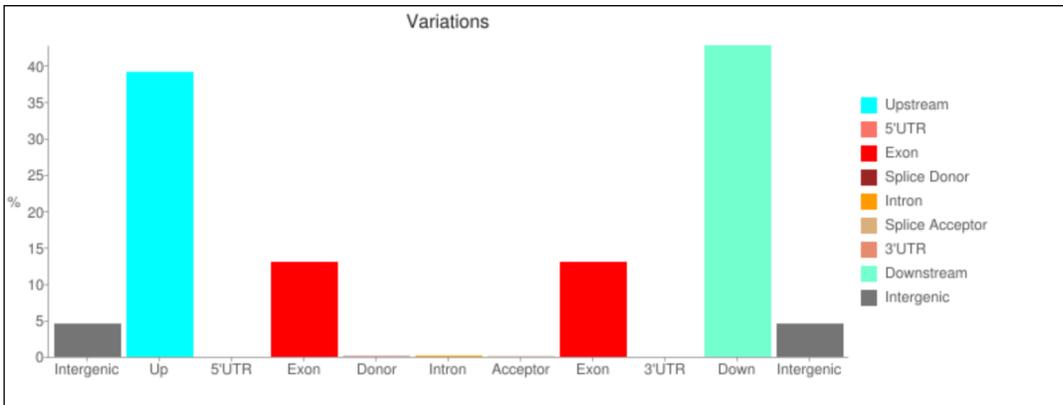
3. Statistics for the variant effects and impacts:

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%

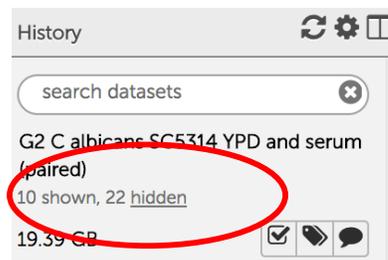
Type (alphabetical order)	Count	Percent
DOWNSTREAM	321,858	40.292%
EXON	67,505	8.451%
INTERGENIC	74,749	9.358%
INTRON	1,064	0.133%
NONE	1	0%
SPLICE_SITE_ACCEPTOR	5	0.001%
SPLICE_SITE_DONOR	4	0.001%
SPLICE_SITE_REGION	176	0.022%
TRANSCRIPT	12	0.002%
UPSTREAM	333,432	41.741%

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:



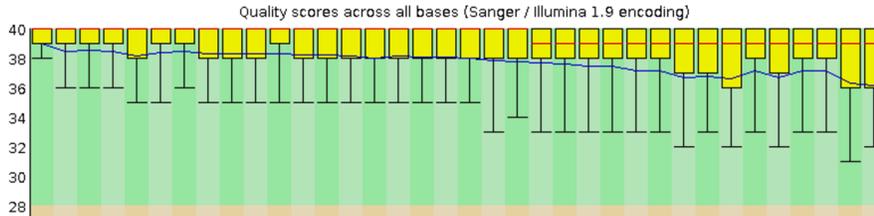
Now, lets take a look at the files generated by the workflow and steps that you can take to further evaluate them.

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: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.

Summary

Genome	ToxoDB-29_TgondiiGT1_Genome
Date	2017-06-17 05:56
Snpeff version	Snpeff 4.11 (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	

SNP result visualization using Ensembl's *Variant Effect Predictor*

Ensembl provides this service for certain organisms including higher eukaryotes, fungi and *Plasmodium falciparum*.

The effect of variants on your genome of interest can be visualized using the ensembl variant effect predictor. You can do this by uploading a VCF file here:

Variant Effect Predictor for Fungi:

http://fungi.ensembl.org/Saccharomyces_cerevisiae/Tools/VEP?db=core

Variant Effect Predictor for *Plasmodium falciparum*:

http://protists.ensembl.org/Plasmodium_falciparum/Tools/VEP?db=core

Go to the Tools section and click on the VEP link

***Note that the upload file size limit is 50MB. Filtered VCF files are smaller than unfiltered ones. **Steps to get a VCF file from galaxy and load to VEP**

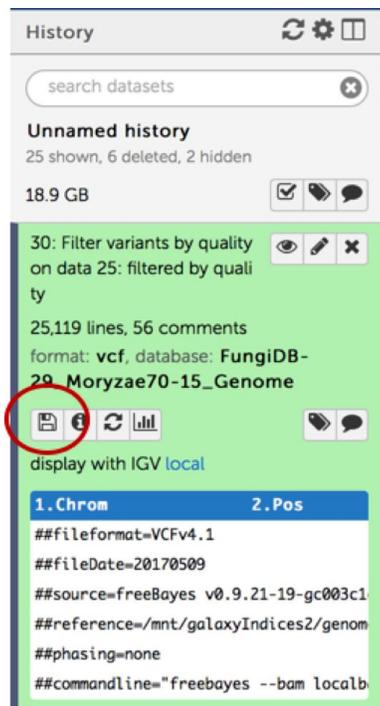
1. Click on on the save icon for the filtered vcf file. This could be any vcf file after (and including) the variant filtering step.

Tools

We provide a number of ready-made tools for processing both our data and yours. We routinely delete results from our servers after 10 days, but if you have an [ensembl account](#) you will be able to save the results indefinitely.

Processing your data

Name	Description	Online tool	Upload limit	Download script	Documentation
Variant Effect Predictor 	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.		50MB*		
HMMER	Quickly search our genomes for your protein sequence.				
BLAST/BLAT	Search our genomes for your DNA or protein sequence.		50MB		
Assembly Converter	Map (liftover) your data's coordinates to the current assembly.		50MB		
ID History Converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.		50MB		



History

search datasets

Unnamed history
25 shown, 6 deleted, 2 hidden
18.9 GB

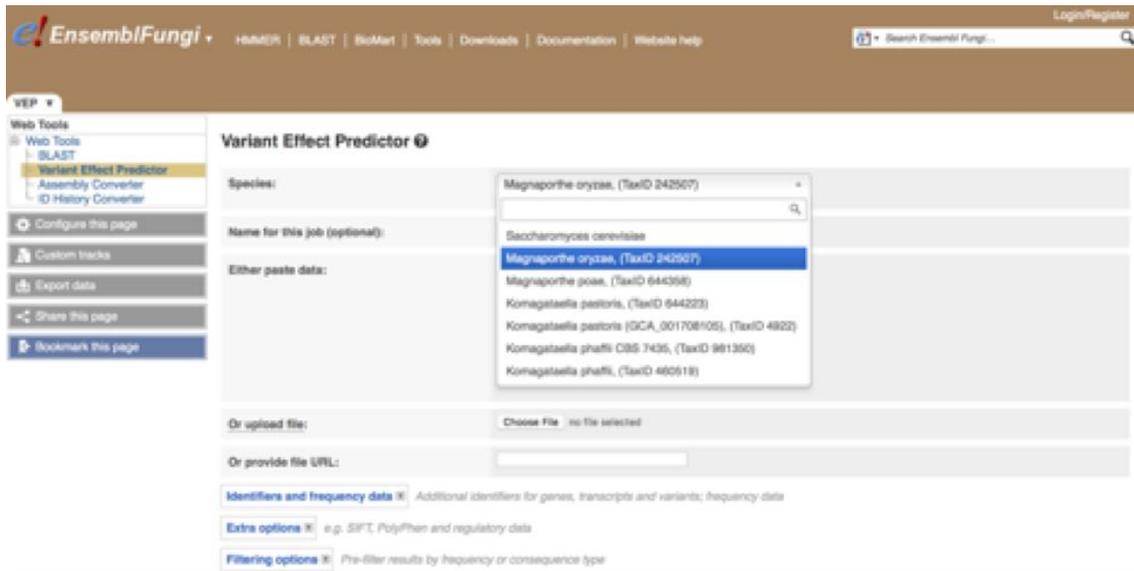
30: Filter variants by quality on data 25: filtered by quality
25,119 lines, 56 comments
format: vcf, database: FungiDB-29_Moryzae70-15_Genome

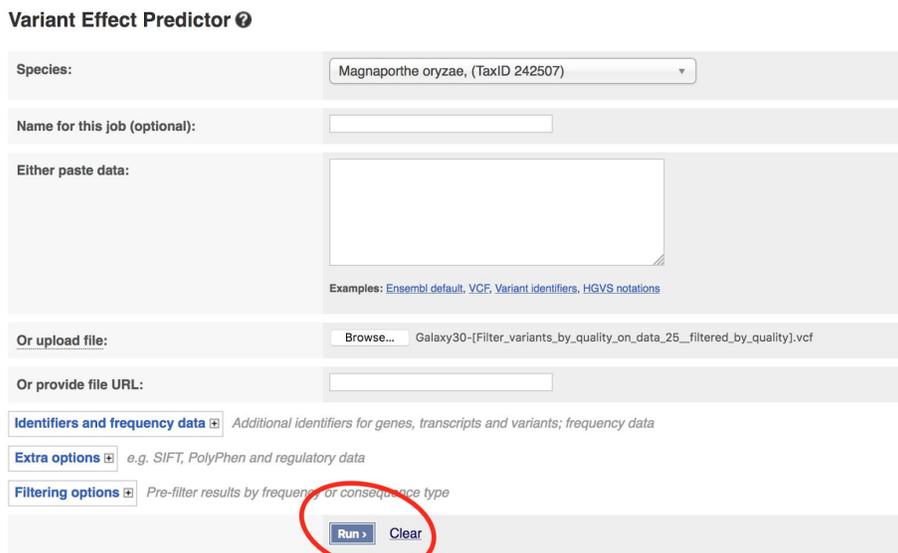
display with IGV local

1. Chrom	2. Pos
##fileformat=VCFv4.1	
##fileDate=20170509	
##source=freeBayes v0.9.21-19-gc003c1	
##reference=/mnt/galaxyIndices2/genom	
##phasing=none	
##commandline="freebayes --bam localb	

Once the file is downloaded, go to the Ensembl fungi VEP page. On this page start by selecting the organism you called SNPs on from the drop down menu.



Next click on the choose file button and select the vcf file you downloaded and click on Run.



The job will start running and will be marked as done when finished.

5. Explore the results (refer to ensembl exercises from earlier today). For example, you can filter the results based on consequence, then sort them in the table to look at ones with High impact.

Variant Effect Predictor results

Job details | Summary statistics

Category: Count
 Variants processed: 24796
 Variants filtered out: 0
 Novel / existing variants: -
 Overlapped genes: 12537
 Overlapped transcripts: 12591
 Overlapped regulatory features: -

Consequences (all)

- splice_acceptor_variant: 43%
- splice_region_variant: 43%
- missense_variant: 3%
- intron_variant: 3%
- synonymous_variant: 2%
- 3_prime_UTR_variant: 2%
- 5_prime_UTR_variant: 2%
- regulatory_variant: 1%
- splice_region_variant: 0%
- Others: 0%

Coding consequences

- missense_variant: 55%
- synonymous_variant: 23%
- frameshift_variant: 3%
- stop_gained: 1%
- inframe_insertion: 1%
- inframe_deletion: 1%
- coding_sequence_variant: 1%
- protein_altering_variant: 1%
- start_lost: 0%

Filters
 Consequence is defined

Impact	Symbol	Gene	Feature	Biotype
High	MODIFIER	IRNA-Pseudo	EFMOG0	00000360 IRNA_pseudoc
High	MODIFIER	-	MGG_01	6T0 protein_coding

Results preview

Navigation: Page: 1 of 2 | Show: 1 5 10 50 All variants

Filters
 Consequence is coding_sequence_variant

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon
CM001231:209683-209683		T	stop_gained	HIGH	MGG_15994	Transcript	MGG_15994T0	protein_coding	8/8	
CM001231:79227-79227		G	synonym							
CM001231:154472-154472		C	synonym							
CM001231:195138-195138		T	synonym							
CM001231:196528-196528		T	synonym							
CM001231:197315-197315		T	synonym							
CM001231:197354-197354		C	synonym							
CM001231:197855-197855		A	synonym							
FM001191:1108002-1108002		C	synonym							

Region in detail
 Location: 1209633-209733
 Gene: MGG_15994

Sequence:
 209 640 209 650 209 660 209 670 209 680 209 690 209 700
 GGGTTGGAGCTGGTGGTGGTATATAGAAAGGGCGGATATATCTTGGTACTCTGGGCAATGGAGGCGCCGATGGTCCGACTTGGGTAATG

Config:
 CCGCAAGTGGAAAGCAACATATAGTGGTGGCGGATATAGAAAGGGCGGATATATCTTGGTACTCTGGGCAATGGAGGCGCCGATGGTCCGACTTGGGTAATG

Gene Legend: protein_coding

7, 07:33