

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 History

ENU-mutant RH clone resistant to IBET-151 IC6

52.95 GB

17: SnpEff on data 15
16: SnpEff on data 15
13: BAM to BigWig on data 12
9: FastQC on data 2: Webpage
3: FastQC on data 1: Webpage
2: SRR5123637_2.fastq.gz
1: SRR5123637_1.fastq.gz

2

History Options

HISTORY LISTS

- Saved Histories
- Histories Shared with Me

CURRENT HISTORY

- Create New
- Copy History
- Copy Datasets
- Share or Publish**
- Extract Workflow
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Unhide Hidden Datasets
- Delete Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export Citations
- Export to File
- Delete
- Delete Permanently

OTHER ACTIONS

3

Share or Publish History 'ENU-mutant RH clone resistant to IBET-151 IC6'

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:

Make History Accessible via Link

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

Make History Accessible and Publish

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

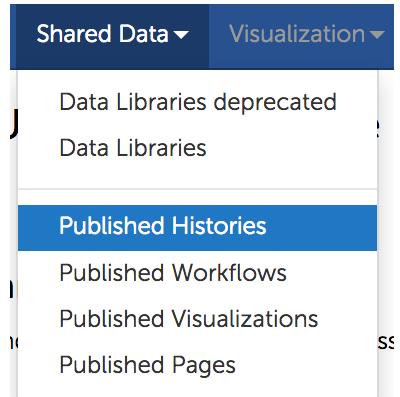
You have not shared this history with any users.

Share with a user

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. At the top, there is a navigation bar with links for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". Below the navigation bar, the title "Published Histories | hb394 | Group 1 results" is displayed. On the far right of the header, there is a button labeled "Import history" which is circled in red. The main content area shows a table titled "Group 1 results" with a total size of "47.44 GB". The table has two columns: "Dataset" and "Annotation". There are four rows, each containing a dataset name and an eye icon in the "Annotation" column. The datasets listed are: "5: SRR1041268_1.fastq.gz", "6: SRR1041268_2.fastq.gz", "7: SRR1041270_1.fastq.gz", and "8: SRR1041270_2.fastq.gz".

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

The screenshot shows two panels of the Galaxy History interface. The left panel displays a list of workflow steps with one dataset hidden. The right panel shows the same steps after clicking the 'hidden' link, revealing all datasets and displaying a warning message for each hidden dataset.

Left Panel (Initial State):

- B. micro Wisconsin single (4 shown, 7 hidden)
- 11: SnpEff on data 9
- 10: SnpEff on data 9
- 3: FastQC on data 1:
RawData
- 1: ERR1349056.fastq.gz

Right Panel (After Clicking 'hidden'):

- B. micro Wisconsin single (11 shown, hide hidden)
- 11: SnpEff on data 9
- 10: SnpEff on data 9
- 3: FastQC on data 1:
RawData
- 1: ERR1349056.fastq.gz

Each of the hidden datasets (11 total) now has a yellow warning box with the message "This dataset has been hidden" and a blue "Unhide it" link.

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.

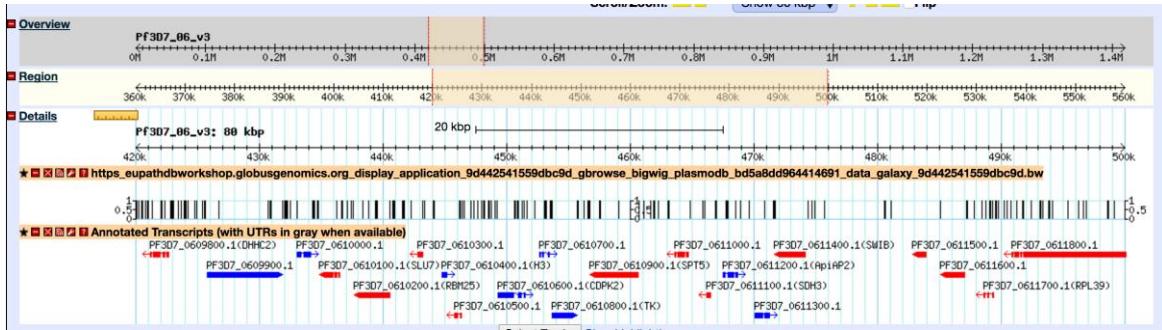
```

8: FreeBayes on data 7
(variants)
3,511 lines, 56 comments
format: vcf, database: PiroplasmaDB-
32_BmictriRI_Genome
display with IGV local

1. Chrom          2. Pos
##fileformat=VCFv4.1
##fileDate=20170617
##source=freeBayes v0.9.21-19-gc003c1e
##reference=/mnt/galaxyIndices2/genome
##phasing=none
##commandline="freebayes --bam localba

```

5. Once the conversion is done, you can export the BigWig file(s) to you EuPathDB dataset page



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.

12.04 GB
29: SnpSift Filter on data 27
13,847 lines, 64 comments
format: vcf, database: PlasmoDB-29_Pfalciparum3D7_Genome
Command to execute: java -Xmx4g
-jar /mnt/GalaxyTools/tools/snpEff
/snpEff_4.1/SnpSift.jar filter filter -f
/scratch/galaxy/files
/008/dataset_8076.dat -e /scratch
/galaxy/job_working_directory
/004/4170/tmpAQDb8H
display with IGV local
1.Chrom 2.Pos
##fileformat=VCFv4.1
##fileDate=20170616
##source=freeBayes v0.9.21-19-gc003c1
##reference=/mnt/galaxyIndices2/genom
##phasing=none

```
0;SRP=0;SRR=0;TYPE=mnp;ANN=CC|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
1;SRP=0;SRR=0;TYPE=snp;ANN=T|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
1;SRF=1;SRP=5.18177;SRR=0;TYPE=mnp;ANN=GA|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
11;SRF=1;SRP=5.18177;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
l=0;TYPE=snp;ANN=A|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|1/2;c.446_447  
;SRP=0;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
RR=0;TYPE=snp;ANN=CAA|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
61;SAR=17;SRF=2;SRP=7.35324;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
l;SRP=5.18177;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
AR=25;SRF=0;SRP=0;SRR=0;TYPE=complex;ANN=TCG|missense_variant|MODERATE|PF3D7_0100100|PF3D7_0100100|transcript|PF3D7_0100100.1|Coding|  
5;SAR=0;SRF=5;SRP=24.391;SRR=21;TYPE=snp;ANN=G|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
3;SRF=0;SRP=0;SRR=0;TYPE=complex;ANN=TTGGAG|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
10.8184;SAR=17;SRF=0;SRP=13.8677;SRR=5;TYPE=snp;ANN=A|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
5;SAR=45;SRF=0;SRP=9.52472;SRR=3;TYPE=snp;ANN=G|stop_lost|HIGH|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
F=0;SRP=0;SRR=0;TYPE=snp;ANN=T|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
4.5915;SAR=10;SRF=0;SRP=0;SRR=0;TYPE=complex;ANN=CATGTTCAGCTG|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200.1|Coding|  
SRP=0;SRR=0;TYPE=snp;ANN=G|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
RF=0;SRP=0;SRR=0;TYPE=snp;ANN=A|missense_variant|MODERATE|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|  
779;SAR=20;SRF=0;SRP=0;SRR=0;TYPE=complex;ANN=CCCACCT|stop_gained|HIGH|PF3D7_0100200|PF3D7_0100200|transcript|PF3D7_0100200.1|Coding|
```

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

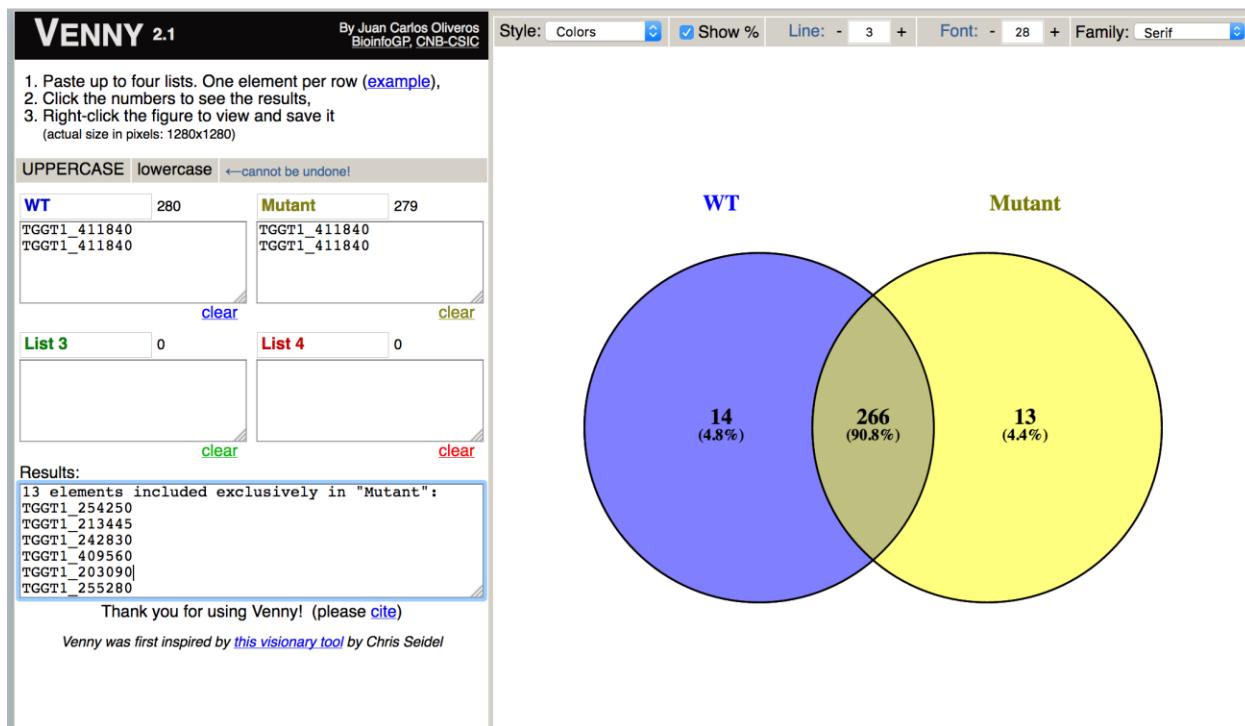
The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the Data Type that best describes your data.
Delimited - Characters such as commas or tabs separate each field.
Fixed width - Fields are aligned in columns with spaces between each field.
Start import at row: 1 File origin: Macintosh

This screen lets you set the delimiters your data contains.
Delimiters
Tab Semicolon Comma Space Other [I]
Treat consecutive delimiters as one Text qualifier: "

Preview of selected data:
Text Import Wizard - Step 1 of 3
Text Import Wizard - Step 2 of 3
Galaxy29-(SnpSift_Filter_on_data_27).xlsx
Home Insert Page Layout Formulas Data Review View
A1 # ##fileformat=VCFv4.1
49. Genotype Quality, the Phred-scaled marginal (or unconditional) probability of the called genotype*
50. Genomic Likelihood, log-odds scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy*
51. Reference allele observation count*
52. "Sum of quality of the reference observations"
53. "Sum of quality of the alternate observations"
54. "Alternate allele observation count"
55. "Sum of quality of the alternate observations"
56. "Sum of quality of the reference observations"
57. /etc/008/dataset_8077.dat PlasmoDB-29_Pfalciparum3D7_Genome /scratch/galaxy/files/008/dataset_8077.dat*
58. #_ID Feature_Type Feature_ID Transcript_E_Rank HGVS_c HGVS_p cDNA_pos/_/CDS_pos/_/AA_pos/_/AA_Distance ERRORS/_WARNINGS/_INFO/_
59. _idAffected_>
60. _idUnaffected_>
61. _idCyclophilin_>
62. /etc/008/dataset_8076.dat -e /scratch/galaxy/job_working_directory/004/4170/tmpAQDb8H*

QUAL	FILTER	INFO	FORMAT	unknown	INFO	FORMAT	unknown
65. 163.615		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.1729_1730_a Asp157Pro 1729 6492	1729 6492	577 2163
66. 59.2743		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.1773A>T I Asp59Lys 1773 6492	1773 6492	595 2163
67. 112.419		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4420_4421 D Th147Arg 4420 6492	4420 6492	1474 2163
68. 133.945		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4468C>T I Leu144Phe 4468 6492	4468 6492	1446 2163
69. 70.189		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4468C>T I Leu144Phe 4468 6492	4468 6492	1446 2163
70. 203.132		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4655T>A I Leu155Ile 4655 6492	4655 6492	1553 2163
71. 149.708		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4733_4734_A Asp157Ala 4733 6492	4733 6492	1578 2163
72. 102.922		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4742_4743_A Asp158Ala 4742 6492	4742 6492	1587 2163
73. 108.734		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.4742_4743_A Asp158Ala 4742 6492	4742 6492	1587 2163
74. 68.702		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.5873C>G I Thr195Ser 5873 6492	5873 6492	1958 2163
75. 599.479		AB AB AB AB missense var MODERATE	PF3D7_0100100 PF3D7_0100100 transcript PF3D7_0100100 Coding		c.6477_6478_A Ala215Ser 6477 6492	6477 6492	2158 2163

- 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.
- 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

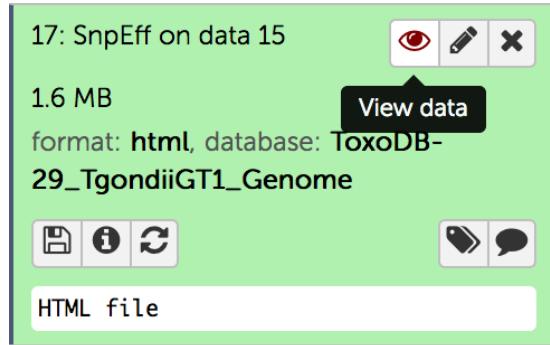
A	B	C	D	E	F	G	H	I	J	K	BS	BT
##SnpEffVersion=4.1 (build 2015-10-03), by Pablo Cingolani												
57 ##SnpEffCmd=SnpEff -vcf -o vcf -stats /scratch/galaxy/files/008/dataset_8107.dat ToxO-DB-29_TgondiiGT1_Genome /scratch/galaxy/files/008/dataset_8105.dat "												
##INFO<ID=ANN,Numt Annotation Annotation_Gene_Name Gene_ID Feature_Type Feature_ID Transcript_E_Rank HGVS.c HGVS.p												
59 ##INFO<ID=LOF,Numb Gene_ID Number_of_Percent_of_transcripts_affected">												
60 ##INFO<ID=NMD,Num Gene_ID Number_of_Percent_of_transcripts_affected">												
61 ##SnpSiftVersion="SnpSift 4.1 (build 2015-10-03), by Pablo Cingolani"												
62 ##SnpSiftCmd=SnpSift filter filter -f /scratch/galaxy/files/008/dataset_8106.dat -e /scratch/galaxy/job_working_directory/004/4169/tmpBopqfU"												
63 ##FILTER=<ID=SnpSift,D (ANN*)>.IMF (FILTER = 'PASS')">												
64 #CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown			
65 TGGT1_chrla	227230.		A	C	1156.55.		AB=0;ABP=0; missense Va MODERATE	TGGT1_293300	[A65,".",K65]			
66 TGGT1_chrla	1340271.		G	C	2387.77.		AB=0;ABP=0; missense Va MODERATE	TGGT1_295040				
67 TGGT1_chrla	1396177.		A	C	387.162.		AB=0;ABP=0; missense Va MODERATE	TGGT1_295125				
68 TGGT1_chrlb	78769.		A	G	1780.8.		AB=0;ABP=0; missense Va MODERATE	TGGT1_207440				
69 TGGT1_chrlb	153771.		T	G	1414.57.		AB=0;ABP=0; missense Va MODERATE	TGGT1_207480				
70 TGGT1_chrlb	276348.		T	G	2066.14.		AB=0;ABP=0; missense Va MODERATE	TGGT1_207750				
71 TGGT1_chrlb	622140.		G	C	2335.06.		AB=0;ABP=0; missense Va MODERATE	TGGT1_208310				
72 TGGT1_chrlb	1446003.		C	T	60.6579.		AB=0;ABP=0; missense Va MODERATE	TGGT1_209755B				
73 TGGT1_chrlb	1446022.		G	T	82.4046.		AB=0;ABP=0; missense Va MODERATE	TGGT1_209755B				

A	B	C	D	E	F	G	H	I	J	K	BS	BT
##SnpEffVersion=4.1 (build 2015-10-03), by Pablo Cingolani												
57 ##SnpEffCmd=SnpEff -vcf -o vcf -stats /scratch/galaxy/files/008/dataset_8107.dat ToxO-DB-29_TgondiiGT1_Genome /scratch/galaxy/files/008/dataset_8105.dat "												
58 ##INFO<ID=ANN,Numt Annotation Annotation_Gene_Name Gene_ID Feature_Type Feature_ID Transcript_E_Rank HGVS.c HGVS.p												
59 ##INFO<ID=LOF,Numb Gene_ID Number_of_Percent_of_transcripts_affected">												
60 ##INFO<ID=NMD,Num Gene_ID Number_of_Percent_of_transcripts_affected">												
61 ##SnpSiftVersion="SnpSift 4.1 (build 2015-10-03), by Pablo Cingolani"												
62 ##SnpSiftCmd=SnpSift filter filter -f /scratch/galaxy/files/008/dataset_8106.dat -e /scratch/galaxy/job_working_directory/004/4169/tmpBopqfU"												
63 ##FILTER=<ID=SnpSift,D (ANN*)>.IMF (FILTER = 'PASS')">												
64 #CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown			
65 TGGT1_chrla	227230.		A	C	1156.55.		AB=0;ABP=0; missense Va MODERATE	TGGT1_293300	TGGT1_chrla:227230:TGGT1_293300			
66 TGGT1_chrla	1340271.		G	C	2387.77.		AB=0;ABP=0; missense Va MODERATE	TGGT1_295040				
67 TGGT1_chrla	1396177.		A	C	387.162.		AB=0;ABP=0; missense Va MODERATE	TGGT1_295125				
68 TGGT1_chrlb	78769.		A	G	1780.8.		AB=0;ABP=0; missense Va MODERATE	TGGT1_207440				
69 TGGT1_chrlb	153771.		T	G	1414.57.		AB=0;ABP=0; missense Va MODERATE	TGGT1_207480				
70 TGGT1_chrlb	276348.		T	G	2066.14.		AB=0;ABP=0; missense Va MODERATE	TGGT1_207750				
71 TGGT1_chrlb	622140.		G	C	2335.06.		AB=0;ABP=0; missense Va MODERATE	TGGT1_208310				
72 TGGT1_chrlb	1446003.		C	T	60.6579.		AB=0;ABP=0; missense Va MODERATE	TGGT1_209755B				
73 TGGT1_chrlb	1446022.		G	T	82.4046.		AB=0;ABP=0; missense Va MODERATE	TGGT1_209755B				

4. You should get unique SNP IDs that look like this (for example):
TGGT1_chrlb: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:
 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 it 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'

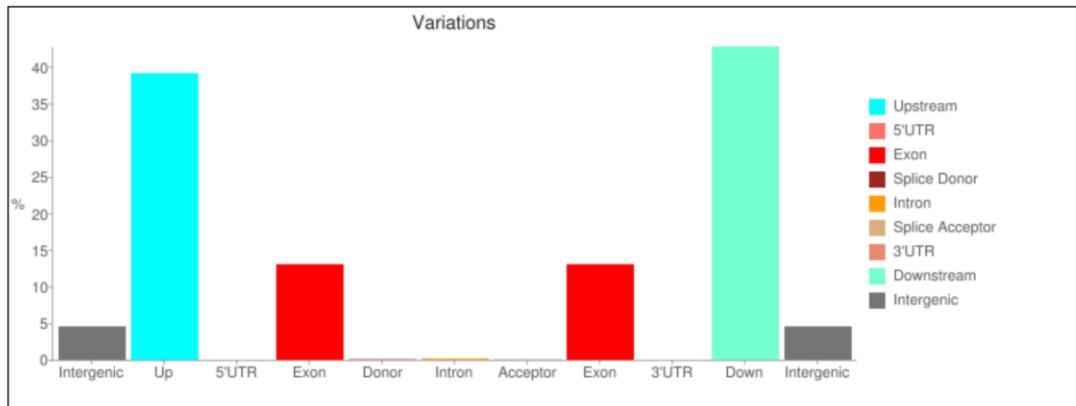
3. Statistics for the variant effects and impacts:

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%

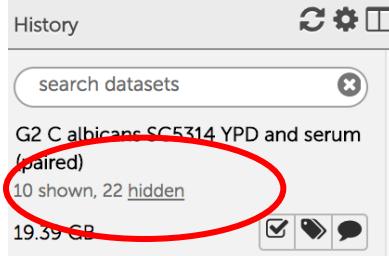
Type (alphabetical order)	Count	Percent
DOWNSTREAM	321,858	40.292%
EXON	67,505	8.451%
INTERGENIC	74,749	9.358%
INTRON	1,064	0.133%
NONE	1	0%
SPlice_SITE_ACCEPTOR	5	0.001%
SPlice_SITE_DONOR	4	0.001%
SPlice_SITE_REGION	176	0.022%
TRANSCRIPT	12	0.002%
UPSTREAM	333,432	41.741%

Base changes summary. SnpEffhtml 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.

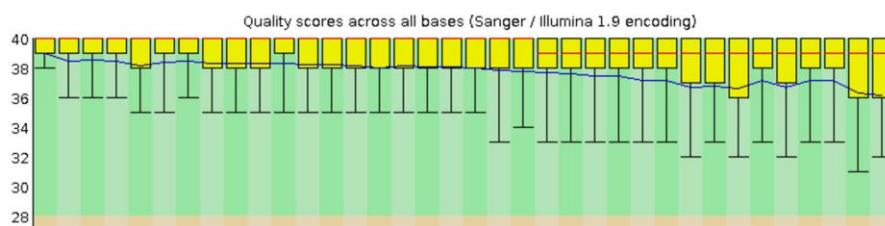
1. Examine sequence quality based on FastQC quality scores.

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 1/1:6:0:6:6:220:-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 1/1:6:0:6:223:-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	-----