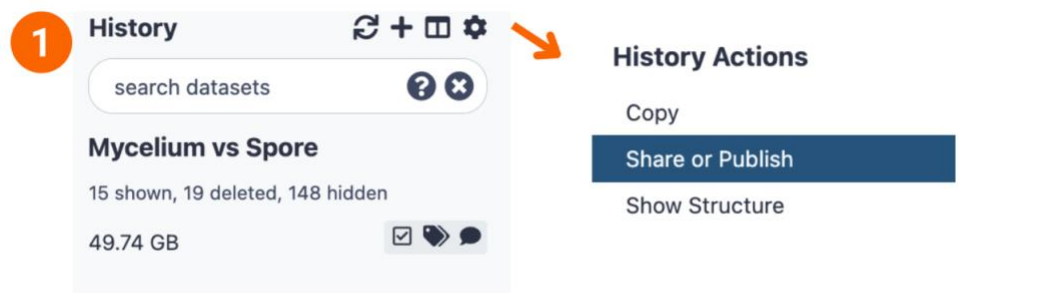


Variant Calling analysis, Part 2: Analyzing results (Group Exercise)

Learning objectives:

- Share and publish your workflow histories.
 - Examine the outputs.
 - View VCF files in JBrowse.
 - Examine the filtered VCF file, extract Gene IDs, and create a Venny diagram.
- **Share workflow histories with others.**
 1. Make sure your history has a useful name (e.g., Group3 SNPs, etc.) and click on the history action menu icon.
 2. Select the “Make History Accessible and Publish” option and check to make sure that all objects within History are accessible.



1 History

search datasets

Mycelium vs Spore

15 shown, 19 deleted, 148 hidden

49.74 GB

History Actions

- Copy
- Share or Publish**
- Show Structure

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

☐ Also make all objects within the History accessible.

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

2 **Make History Accessible and Publish**

☐ Also make all objects within the History accessible.

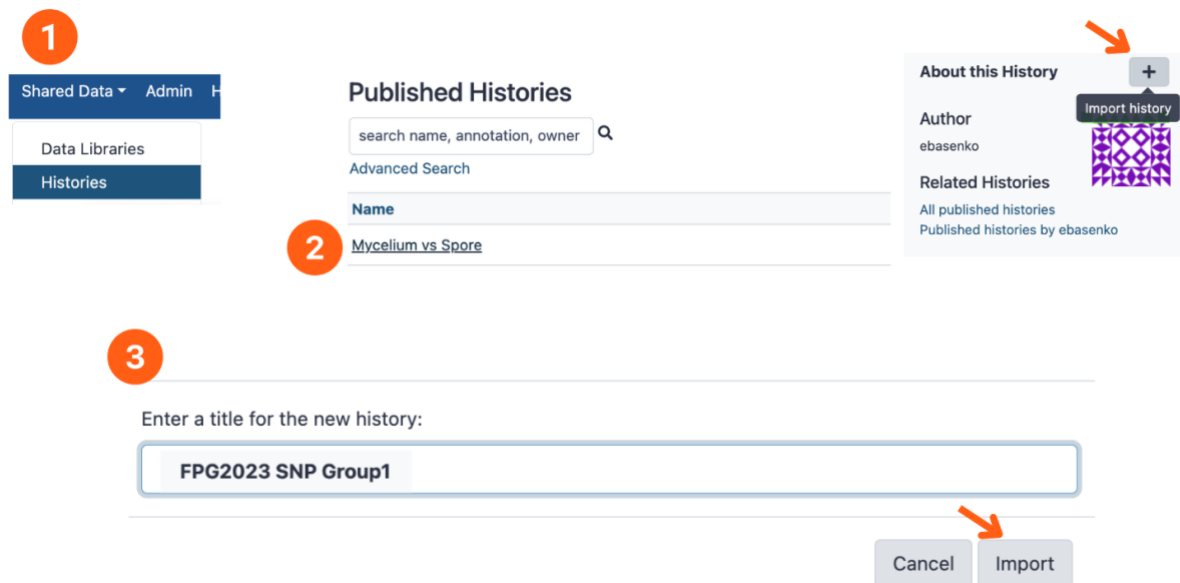
Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, v

Share History with Individual Users

You have not shared this history with any users.

Share with a user

- **Importing workflow histories and output files into your own Galaxy workspace.**
 1. Click on “Shared Data” at the top and select “Histories”.
 2. Click on the history shared by your colleague, click on the plus icon on the far right and choose to import the history.
 3. You can give it a descriptive name if you prefer or leave it as is.



If everything worked out, you should see a list of completed workflow steps highlighted in green. The workflow generates many output files, however not all of the output files are visible. You can explore all the hidden files clicking on the word “hidden” (orange circle) – this will reveal all hidden files.

The Variant calling 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 more 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.

Many more output files are available to explore →

filter VCF files using arbitrary expressions →

SnpEff: Analyze and annotate variants, and calculation of the effects →

Bowtie: Align reads to a reference genome →

FPG2023 SNP GROUP5

9 shown, 2 deleted, 7 hidden

11.86 GB

18: SnpSift Filter on data 16	👁️ ✎️ ✕
17: SnpEff on data 15	👁️ ✎️ ✕
16: SnpEff on data 15	👁️ ✎️ ✕
13: BAM to BigWig on data 12	👁️ ✎️ ✕
12: Bowtie2.4.4 on data 8 and data 7: alignments	👁️ ✎️ ✕
10: FastQC on data 4: Webpage	👁️ ✎️ ✕
5: FastQC on data 3: Webpage	👁️ ✎️ ✕
4: SRR10728586_2.fastq.gz	👁️ ✎️ ✕
3: SRR10728586_1.fastq.gz	👁️ ✎️ ✕

Analysis and annotation of the genomic variants are carried out by the SnpEff tool. SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes). 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 categorized based on the effect of the amino

acid change and are classified into synonymous and non-synonymous, gain or loss of start codons, gain or 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.).

- **Examine your results.**

1. Click on the *hidden* files link in the history panel to reveal all workflow output files.
2. Examine the output files.
3. What does the tool FASTQC do?
4. 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.

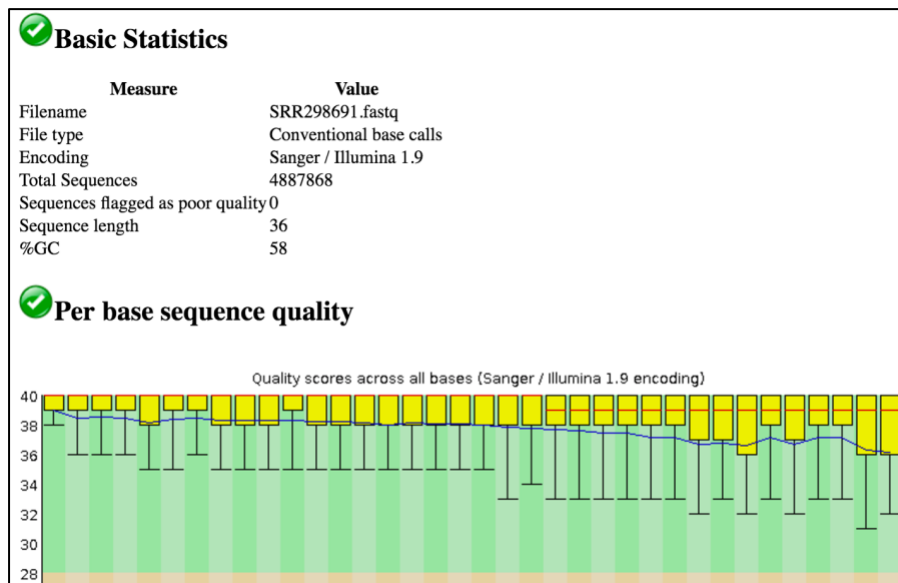
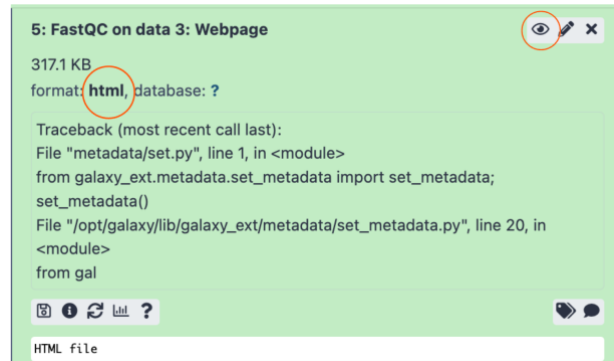
5. Examine the VCF file in your results (click on the *eye* icon to view its contents). Detailed information about VCF file content is available here: <https://samtools.github.io/hts-specs/VCFv4.2.pdf>

The screenshot shows the Galaxy workflow history panel. The left pane lists workflow steps, with 'FPG2023 SNP GROUP5' highlighted and circled in red. An arrow points from this step to the right pane, which shows the details of step 15: 'FreeBayes on data 12 (variants) filtered by quality'. The right pane shows the output of this step, including a traceback and a VCF file.

The screenshot shows the Galaxy workflow details for step 15: 'FreeBayes on data 12 (variants) filtered by quality'. The step is circled in red. The output shows a traceback and a VCF file.

- 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. What does the report tell you about the quality ?



- Examine SnpEff 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 in Galaxy for your review.

The header contains a short summary and information about the run and it has several major components:

The Summary contains warnings 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.*). Other components:

- 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

Summary	
Genome	FungiDB-34_ZtriticliPO323_Genome
Date	2023-04-11 10:24
SnEff version	SnEff 4.11 (build 2015-10-03), by Pablo Cingolani
Command line arguments	SnEff -i vcf -o vcf -stats /scratch/galaxy/files/000/391/dataset_391424.dat FungiDB-34_ZtriticliPO323_Genome /scratch/galaxy/files/000/391/dataset_391422.dat
Warnings	3,774
Errors	0
Number of lines (input file)	306,885
Number of variants (before filter)	307,538
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	307,538
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	653
Number of effects	1,280,819
Genome total length	39,730,198
Genome effective length	39,730,198
Variant rate	1 variant every 129 bases

Variants rate details			
Chromosome	Length	Variants	Variants rate
Ztri_MitoScaffold	43,947	18	2,441
Ztri_chr_1	6,088,797	44,156	137
Ztri_chr_10	1,682,575	15,039	111
Ztri_chr_11	1,624,292	14,012	115
Ztri_chr_12	1,462,624	12,767	114
Ztri_chr_13	1,185,774	10,694	110
Ztri_chr_14	773,098	2,064	374
Ztri_chr_15	639,501	7,821	81
Ztri_chr_16	607,044	5,094	119

- 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

Summary statistics for variant types

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

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

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:

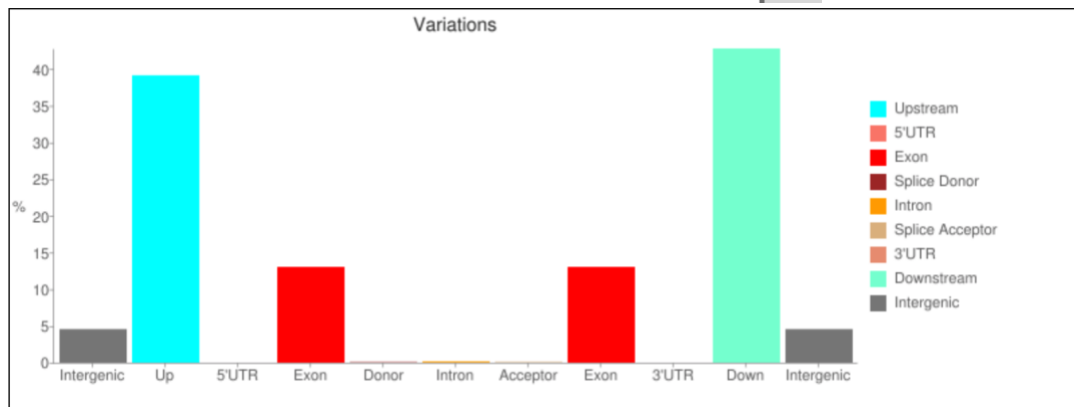
Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	1,857	0.145%
LOW	87,874	6.861%
MODERATE	41,970	3.277%
MODIFIER	1,149,118	89.717%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	29,331	28.472%
NONSENSE	370	0.359%
SILENT	73,317	71.169%

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 in several classes: 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 help 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 help to identify potential sequencing artifacts due to PCR enrichment step (generation of heterozygous counts in a haploid organism).

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).

Examining SNP information.

You can view the SNP information by clicking on the “eye” icon within the SnpEff vcf file.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
Ztr1_chr_1	133	.	CC	GT	59.2437	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	195	.	CATA	CATG	169.043	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	1565	.	A	G	68.5388	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	1603	.	C	T	140.924	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	1651	.	C	T	114.529	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	1927	.	G	A	113.199	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	1985	.	C	T	250.268	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	2168	.	G	A	100.41	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	2272	.	CAATG	TAATG	191.809	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	2293	.	G	A	206.133	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	2367	.	G	A	54.2829	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	2630	.	C	T	112.111	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	2975	.	C	T	62.699	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	3119	.	GAATG	CAATG	58.621	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	3180	.	C	T	80.1965	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	3723	.	G	A	125.847	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	3812	.	T	C	50.3	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	4453	.	G	A	74.7978	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	4465	.	G	A	109.005	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	4479	.	GC	CT	129.602	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	4495	.	T	C	63.6211	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	5145	.	T	C	132.17	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1_chr_1	5265	.	TA	CG	298.39	.	AB=0,ABP=0,AC=2,AF=1,
Ztr1 chr 1	5325	.	G	A	321.168	.	AB=0,ABP=0,AC=2,AF=1,

1/ shown, 2 deleted, hide hidden

11.86 GB

18: SnpSift Filter on data 16

17: SnpEff on data 15

16: SnpEff on data 15

~340,000 lines

format: vcf, database: FungiDB-34_ZtriticlIPO323_Genome

Traceback (most recent call last):

File "metadata/set.py", line 1, in <module>

from galaxy_ext.metadata.set_metadata import set_metadata;

set_metadata()

File "/opt/galaxy/lib/galaxy_ext/metadata/set_metadata.py", line 20,

in <module>

from gal

display with IGV local

1: Chrom

##INFO=ID=DP,Number=1,Type=Integer,Description="Total read depth at the

##INFO=ID=DP9,Number=1,Type=Float,Description="Total read depth per bp

##INFO=ID=AC,Number=4,Type=Integer,Description="Total number of alterna

##INFO=ID=AN,Number=1,Type=Integer,Description="Total number of alleles

##INFO=ID=AF,Number=4,Type=Float,Description="Estimated allele frequenc

The vcf file generated by SnpEff contains information about SNPs and the genomic location. Here is an example of a file opened in Excel:

#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				

Filtering VCF file data.

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. Your workflow is set up to use an expression that filters VCF files on moderate and high impact SNPs (this setting can be adjusted manually in the workflow editor). Here is the exact expression used:

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

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

For this exercise, two groups will be sharing data SnpSift outputs: group 1 & 2, group 3 & 4, and group 5 & 6. File manipulations should be performed on both SnpSift vcf files.

Look at the filtered vcf file in Galaxy. 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.

History

18: SnpEff Filter on data 16

10,309 lines, 64 columns

format: vcf, database: FungusDB

29_AfumigatusAf293_Genome

View data

Command to execute: java -Xmx4g -jar /mnt/galaxyTools/tools/snpEff/snpEff_4.1 filter filter -f /scratch/galaxy/files/098/dataset_9834 -e /scratch/galaxy/job_working_directory/

display with IGV local

1. Chrom	2. Pos
##fileformat=VCFv4.1	
##fileDate=20190326	
##source=FreeBayes v0.9.21-19-gc003c	
##reference=/mnt/galaxy/Indices2/genome	
##phasing=none	
##commandLine="freebayes --bam local	

1. Download the SnpSift Filter output by clicking on the save icon.
2. Right click and open this file with Excel.

</

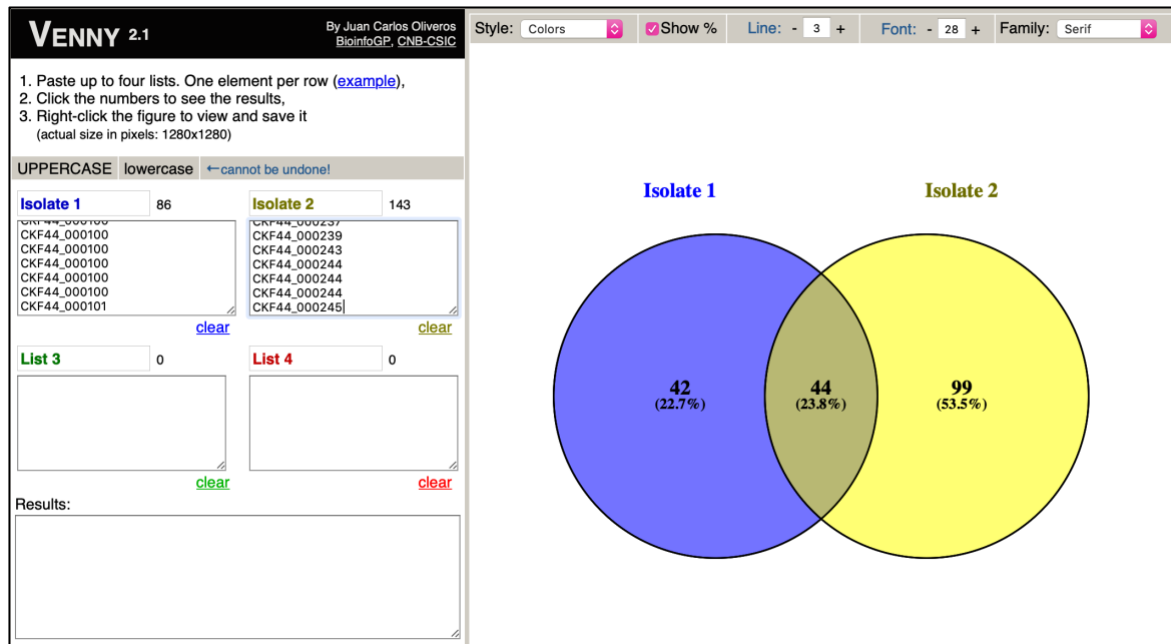
- **Manipulate Excel file to display SNP info in columns.**
 1. Select the “INFO” column.
 2. Navigate to the “Data” tab in Excel and choose “Text to Columns”.
 3. Use the “Delimited” option.
 4. Set delimiters to the “Tab” and “|” in the “Other” and click “Next”
 5. Leave other criteria at default and click on the “Finish” button.
 6. Click “OK” on the Alert pop-up.

The sequence of screenshots shows the following steps:

- Step 1:** An Excel spreadsheet with the 'INFO' column selected. The column contains text like 'FungiDB-34, ZtriticolaPO323_Genome' and 'Annotation_Impact | Gene'.
- Step 2:** The 'Data' tab is active, and the 'Text to Columns' button in the ribbon is highlighted with a red arrow.
- Step 3:** The 'Text to Columns' wizard is open, and the 'Delimited' radio button is selected. The text below the buttons reads: 'Characters such as commas or tabs separate each field.' and 'Fields are aligned in columns with spaces between each field.'
- Step 4:** The 'Delimiters' section of the wizard is shown. The 'Tab' and 'Other' checkboxes are selected. The 'Other' field contains a vertical bar '|'. The 'Text qualifier' is set to double quotes. A preview of the selected data is shown below.
- Step 5:** The 'Column data format' screen is shown. The 'General' radio button is selected. The 'Destination' is set to '\$H:\$I'. A preview of the selected data is shown below.
- Step 6:** An 'Alert' dialog box is shown with the message 'There's already data here. Do you want to replace it?'. The 'OK' button is highlighted with a red circle.

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 sort and examine SNPs based on their characteristics.

If you are comparing two or more strains, you may want to extract gene IDs from all VCF files and identify common signatures across isolates or strains. For this type of analysis, you can use <http://bioinfoGP.cnb.csic.es/tools/venny/> to generate a Venn diagram:



The screenshot above is showing comparison of between lists of GeneIDs. Is it possible to miss some important polymorphisms using this method? Of course, the answer is yes😊
 For example, 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.

- **Analyze 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. Deploy the concatenate function in Excel.
3. Create a unique ID for SNPs by combining 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

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

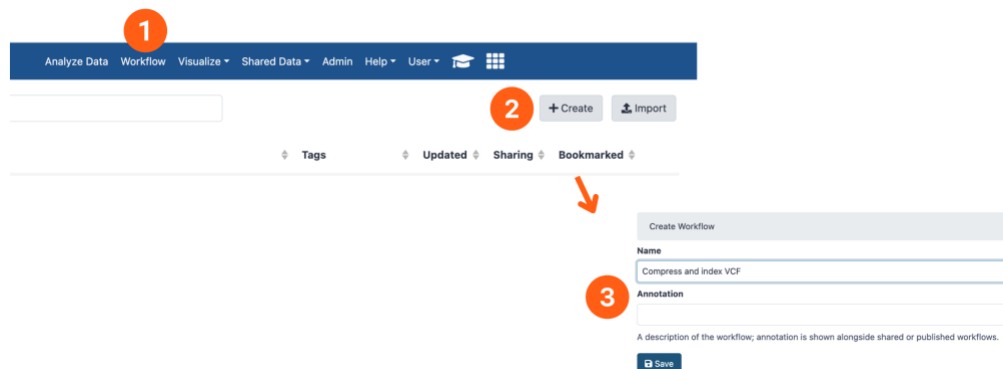
1

Viewing VCF file results in the JBrowse genome browser.

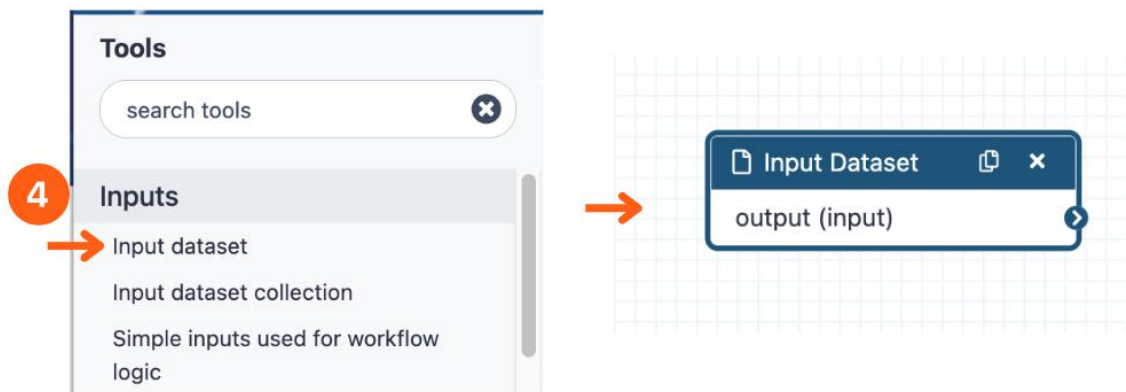
- **Create a workflow to generate a compressed vcf and index files for viewing your data in JBrowse.**

To view a VCF file in JBrowse, it first has to be indexed and compressed. This is done using two tools: bgzip and tabix, respectively. You can run these tools sequentially or you can set up a mini workflow and then run the workflow to generate the output files as follows:

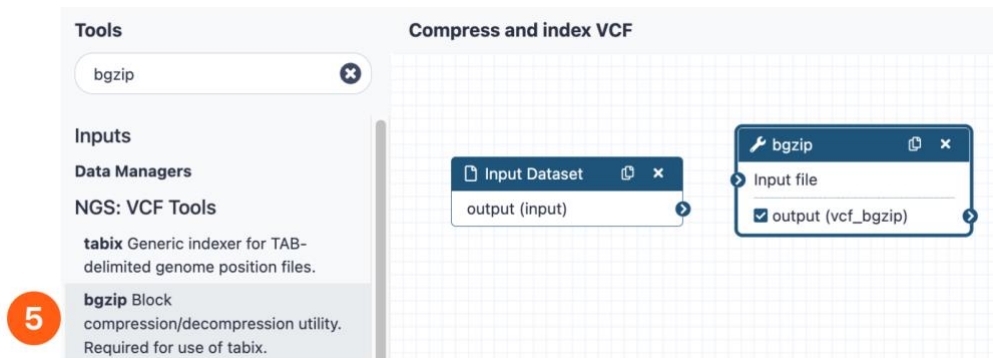
1. Click on the “Workflow” menu.
2. Click on the “Create” button to start a new workflow.
3. Give the workflow a name (e.g. Compress and index VCF) and click on the save button. This will open a workflow canvas.



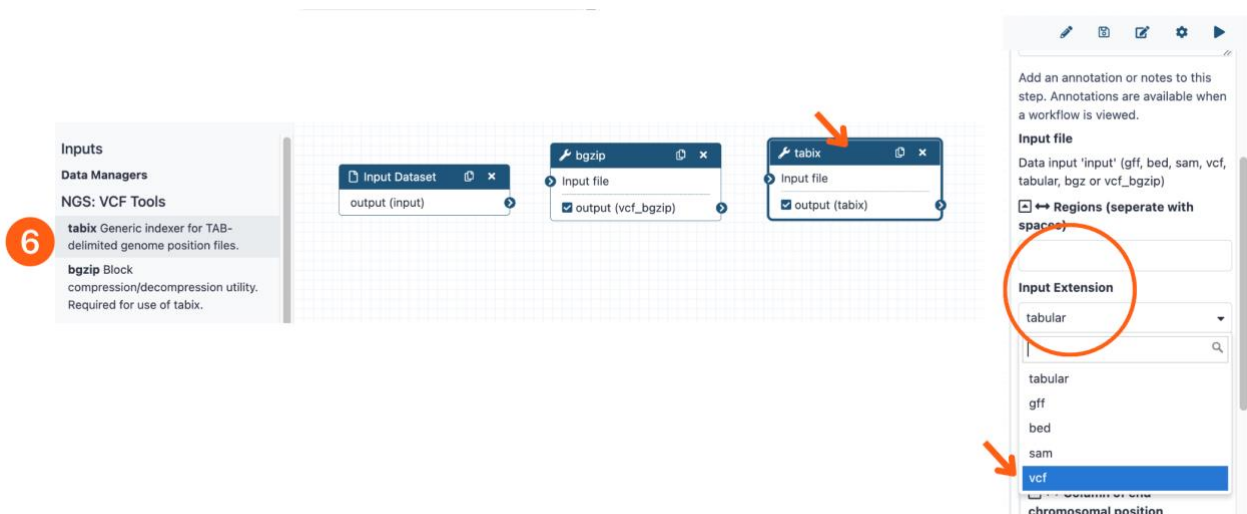
4. All workflows must start with an input file so add the “Input Dataset” step to the workflow using the menu on the left (you must click on the tool for it to appear in the workflow editor canvas).



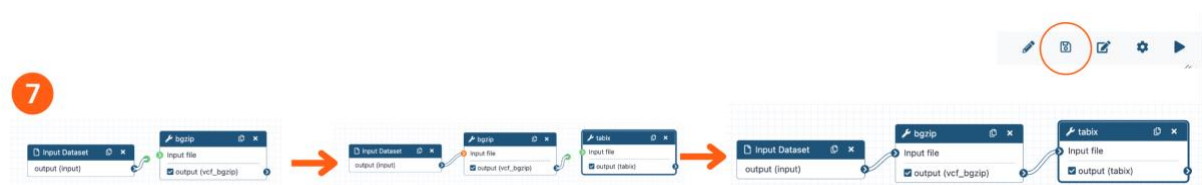
5. Using the menu on the left, search for and add the “bgzip” tool.



- Using the menu on the left, search for and add the “tabix” tool. Left-click on the “tabix” icon and select “vcf” under “input selection” on the right (tool option section)



- Connect each step/tool into a workflow and save it (the button is at the top of the screen)

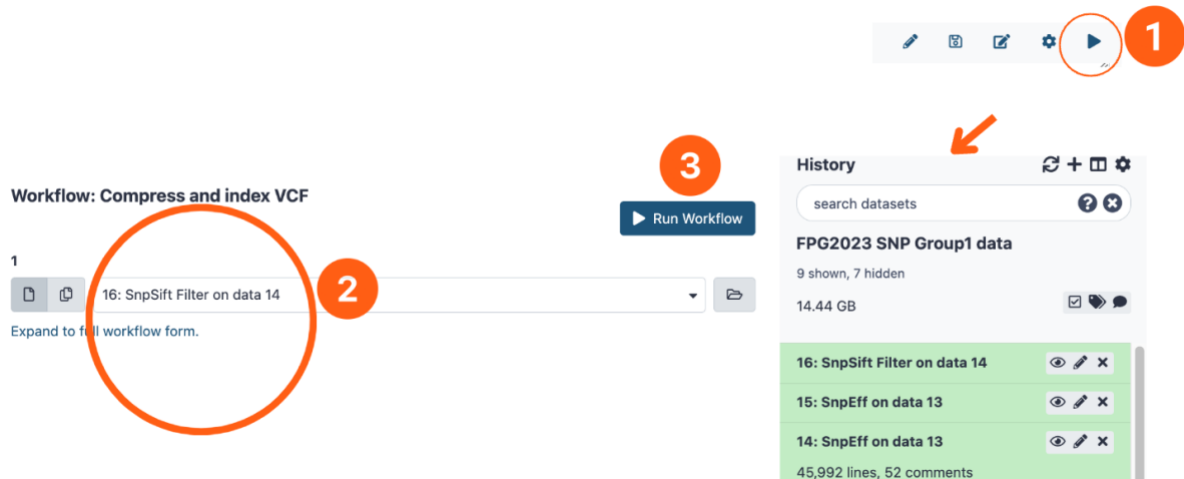


- **Run the newly created workflow to generate a compressed vcf and index files.**

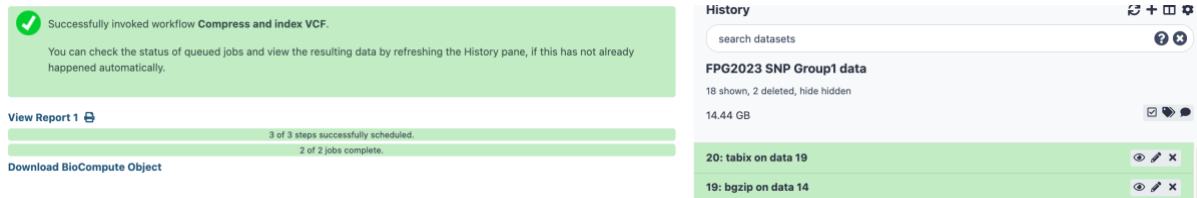
1. Click on the “Play” button to start your workflow.
2. Select the VCF file you want to process.

Note that the workflow produced several vcf files - SnpSift, SnpEff.. In the screenshot below we will use a vcf filtered for high and moderate impact SNPs but if you are interested in all mutations, you may want to choose another file.

3. Click on the “Run Workflow” button.

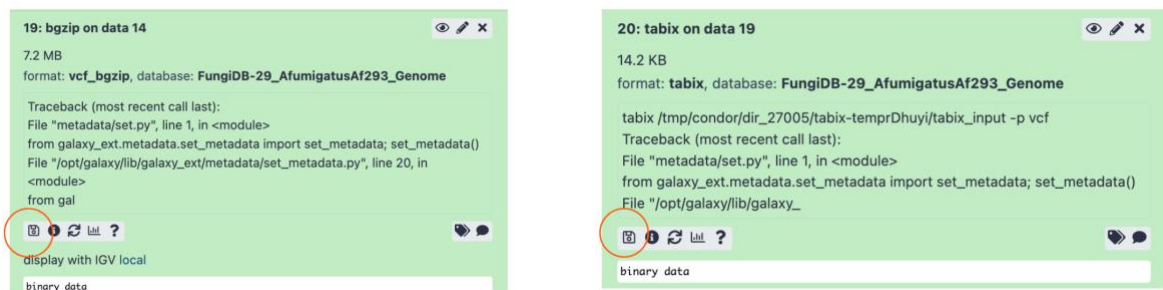


After the workflow completed running, you should have 2 new files in the history on the right (tabix and bgzip).



- **Download compressed vcf (vcf_bgzip) and index (tabix) files and view them in JBrowse.**

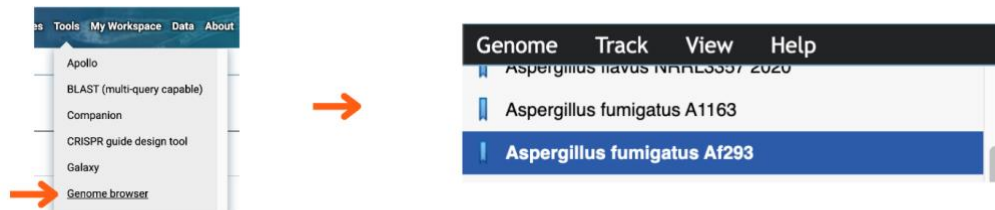
1. Download both files by clicking on the download icon. You will need both files.



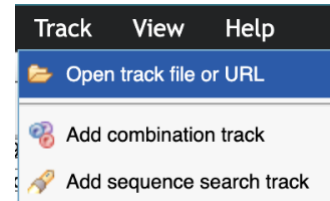
2. After the files are downloaded, rename them as follows:

- a. The **vcf_bgzip** file to “**group#.vcf.gz**” (i.e. **group1.vcf.gz**)
- b. The **tabix** file to “**group#.vcf.gz.tbi**” (i.e. **group1.vcf.gz.tbi**)

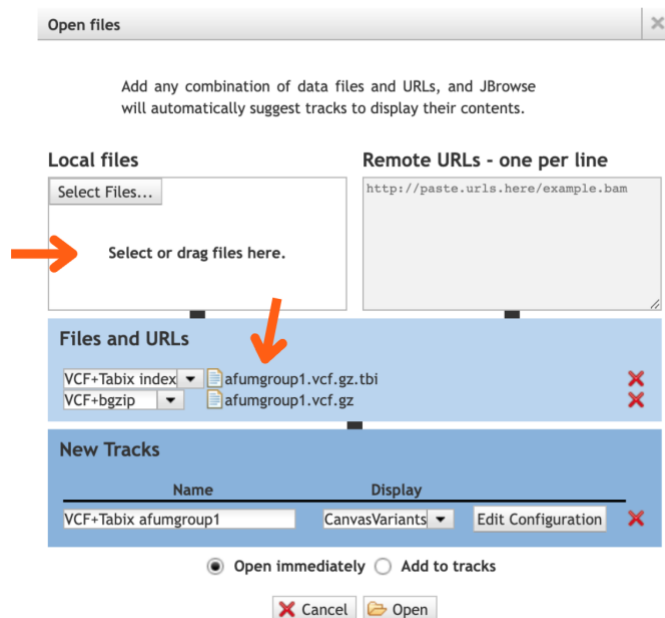
3. Navigate to JBrowse in FungiDB and select the correct genome from the Genome drop-down menu.



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



5. Drag and drop your files in the window that appears. Notice that the file formats are autodetected. Click on the "Open" button at the bottom of the pop-up.



You should now be able to view the SNPs in JBrowse.

