Overview

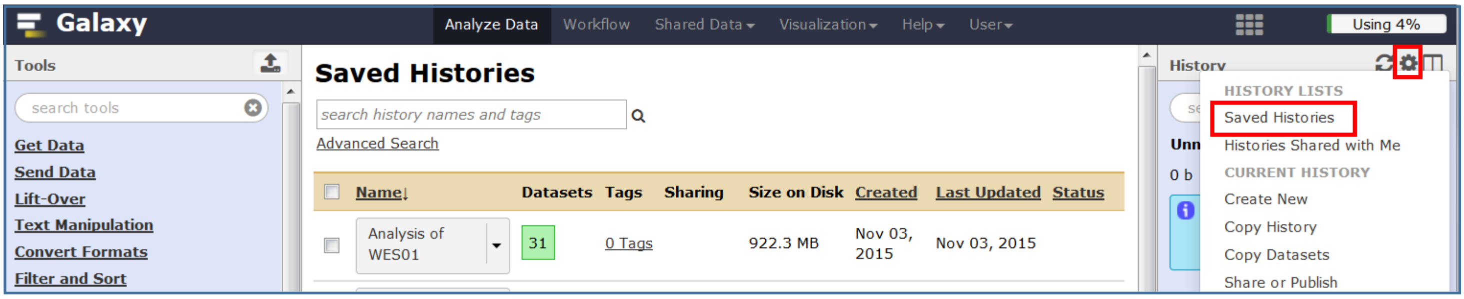
At this point in the analysis of whole exome data from patient WES01, we have assessed the quality of raw sequence data, removed low quality reads, aligned sequence data to the reference genome, removed poorly mapped reads and duplicate reads, assessed the alignment process, called variants and evaluated variant calling. The aim of this practical is to complete the analysis of WES01 by annotating the variants and using a filtering strategy to generate a shortlist of potentially pathogenic variants. At the end of this exercise you will be able to:

1. Use ANNOVAR to annotate variants with respect to genes, databases of normal variation, and predictors of pathogenicity
2. Use phenomizer (http://compbio.charite.de/phenomizer/) to generate a list of candidate genes from human phenotype ontology (HPO) terms
3. Interpret deleteriousness scores such as SIFT and Polyphen
4. Investigate phenotype-genotype relationships using OMIM, ClinVar, and PubMed
5. Create and apply a workflow from your Galaxy history

Let's begin

Open the saved history for the analysis of WES01.

1. Click the cog icon in the history pane and select Saved Histories



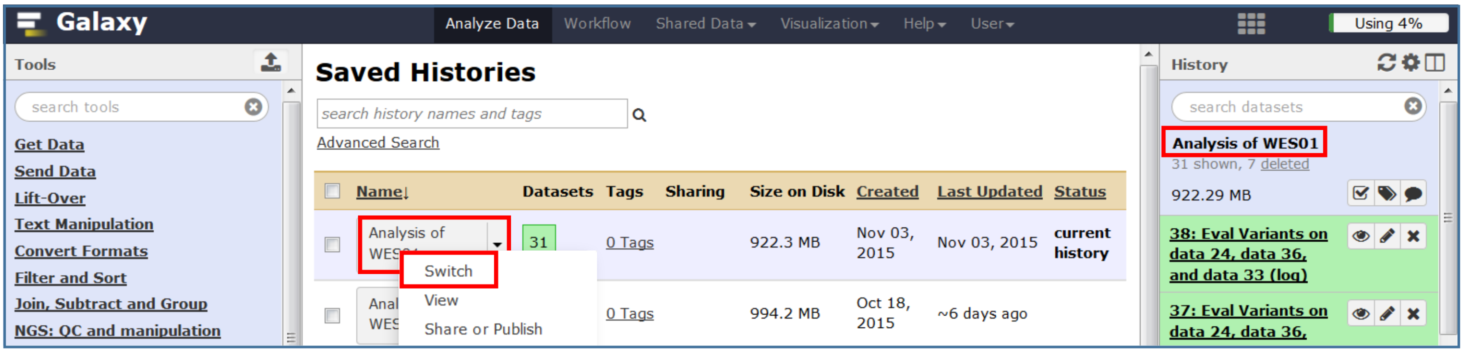
Saved Histories

search history names and tags

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
Analysis of WES01	31	0 Tags		922.3 MB	Nov 03, 2015	Nov 03, 2015	

2. Click the saved history for 'Analysis of WES01' and select switch



Saved Histories

search history names and tags

Advanced Search

Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated	Status
Analysis of WES01	31	0 Tags		922.3 MB	Nov 03, 2015	Nov 03, 2015	current history
Anal… WES…		0 Tags		994.2 MB	Oct 18, 2015	~6 days ago	

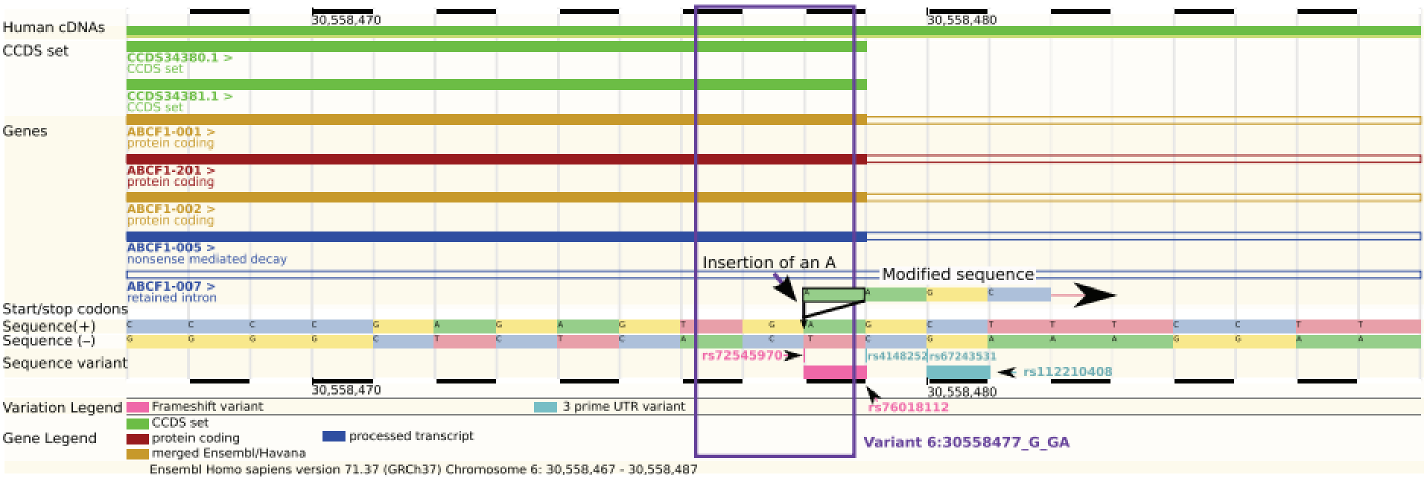
When the correct history has been loaded, the history pane should be titled 'Analysis of WES01'.

Annotate variants

The result of variant calling is a variant call file (VCF) which describes the location and data quality of each variant. However, the initial VCF file does not provide any information about the functional relevance of variants and their potential contribution to disease. To gain these insights we will use ANNOVAR (Wang et al 2010) to annotate each variant in the VCF file with respect to their location in relation to genes and coding sequences (exon, intron or intergenic), whether they change the coding sequence and if so how (missense, stop gain, synonymous, frameshift, amino acid consequence etc). In addition, we will cross reference the variants with databases of known variation (1000 genomes, dbSNP, Exome Sequencing Project and COSMIC) and predictions of functional significance (avsift and conservation scores).

At this stage, it is important to be aware that the annotation result will vary according to the choice of annotation software (alternative software: SnpEff Cingolani et al 2012 and Variant Effect Predictor VEP McLaren et al 2010) and definition of transcripts (Ensemble Flicek et al 2012, RefSeq Pruitt et al 2012, Gencode Harrow et al 2012). This variability occurs because a single variant may have different effects in different transcripts (isoforms of the gene) and in some cases may effect different genes (genes can overlap, one on forward strand the other on the reverse strand) each with multiple transcripts. It is also possible to interpret a single variant as having multiple effects on a single transcript (Figure 1). The annotation software must therefore apply a set of rules to determine which variant takes precedence (see here for ANNOVAR precedence rules) but these rules vary between programs so that they generate different results. For example, the overall agreement for Loss of Function (LoF) variants annotated by ANNOVAR and VEP is just 64% (McCarthy et al 2014).

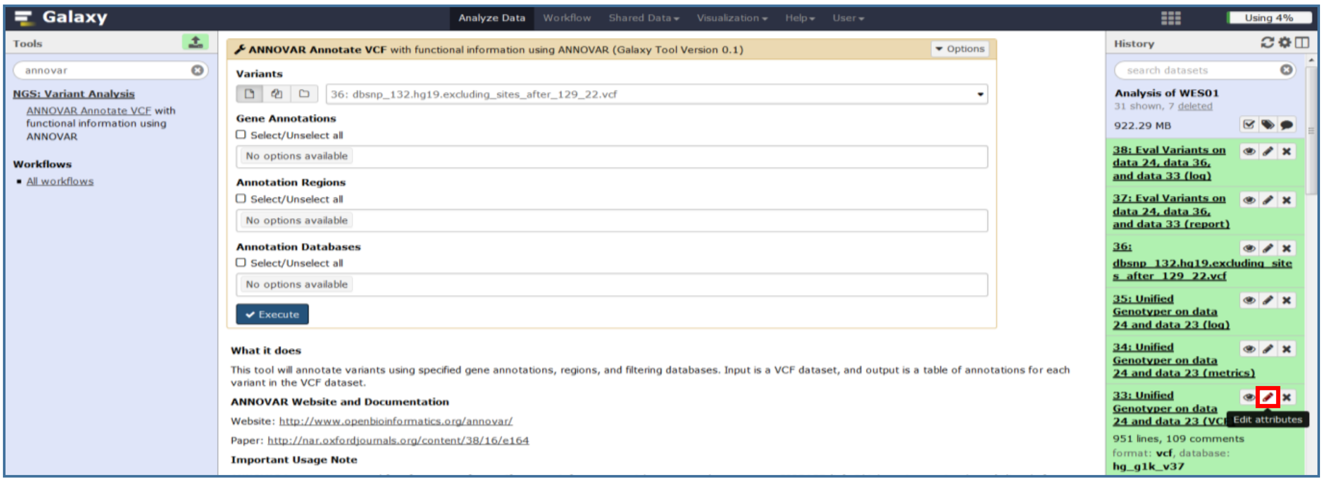
Figure 1. Annotation examples from McCarthy et al 2014



In figure 1, the variant is an insertion of an A in a stop codon (TGA) in the penultimate base of the exon in all transcripts except one. It is possible to interpret this variant as a frameshift insertion or a stop-loss but it is actually a synonymous variant because the stop codon remains in the same place.

The choice of which transcript set to use for annotation has an even bigger effect on the annotation results. For example, the overlap between LoF variants is only 44% when the same software is used to annotate against transcripts from Ensembl or RefSeq (McCarthy et al 2014). This is because RefSeq transcripts are based on a collection of non-redundant mRNA models that have strong support and are manually curated. As a result, the RefSeq database is highly accurate but does not contain all possible transcripts or gene models (n=41,501 transcripts from RefSeq release 57 that were used by ANNOVAR). In comparison, Ensembl provides a less curated but more comprehensive list of transcripts (n=115,901 transcripts from Ensembl version 69 that were used by ANNOVAR) based on information from the Consensus Coding Sequence (CCDS, Pruitt et al 2009), Human and Vertebrate Analysis and Annotation (HAVANA), Vertebrate Genome Annotation (Vega, Ashurst et al 2005), ENCODE (The ENCODE Project Consortium 2012) and GENCODE (Searle et al 2010).

1. In Galaxy, ANNOVAR will only work with hg19 VCF files. We therefore need to use the 'Edit attributes' icon in the history pane to change the database from hg_g1k_v37 to hg19 for the VCF file from GATK Unified Genotyper.



ANNOVAR Annotate VCF with functional information using ANNOVAR (Galaxy Tool Version 0.1)

Variants: 36: dbsnp_132.hg19.excluding_sites_after_129_22.vcf

Gene Annotations: □ Select/Unselect all — No options available

Annotation Regions: □ Select/Unselect all — No options available

Annotation Databases: □ Select/Unselect all — No options available

What it does

This tool will annotate variants using specified gene annotations, regions, and filtering databases. Input is a VCF dataset, and output is a table of annotations for each variant in the VCF dataset.

ANNOVAR Website and Documentation

Website: http://www.openbioinformatics.org/annovar/

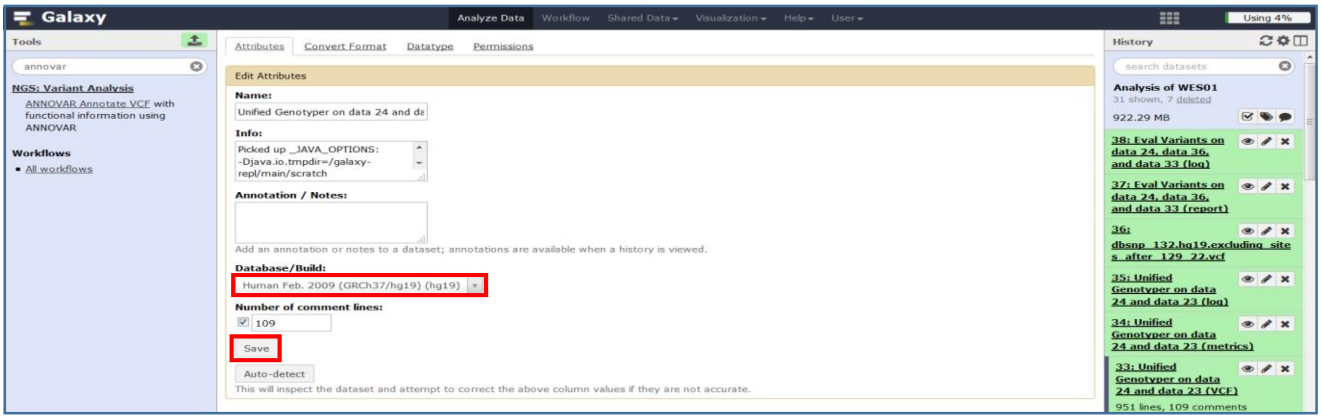
Paper: http://nar.oxfordjournals.org/content/38/16/e164

Important Usage Note

Analysis of WES01
31 shown, 7 deleted
922.29 MB

38: Eval Variants on data 24, data 36, and data 33 (log)
37: Eval Variants on data 24, data 36, and data 33 (report)
36: dbsnp_132.hg19.excluding_sites_after_129_22.vcf
35: Unified Genotyper on data 24 and data 23 (log)
34: Unified Genotyper on data 24 and data 23 (metrics)
33: Unified Genotyper on data 24 and data 23 (VCF)
951 lines, 109 comments
format: vcf, database: hg_g1k_v37

2. Select the hg19 Database/Build and click save.



Edit Attributes

Name: Unified Genotyper on data 24 and da

Info: Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-repl/main/scratch

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Human Feb. 2009 (GRCh37/hg19) (hg19)

Number of comment lines: ☑ 109

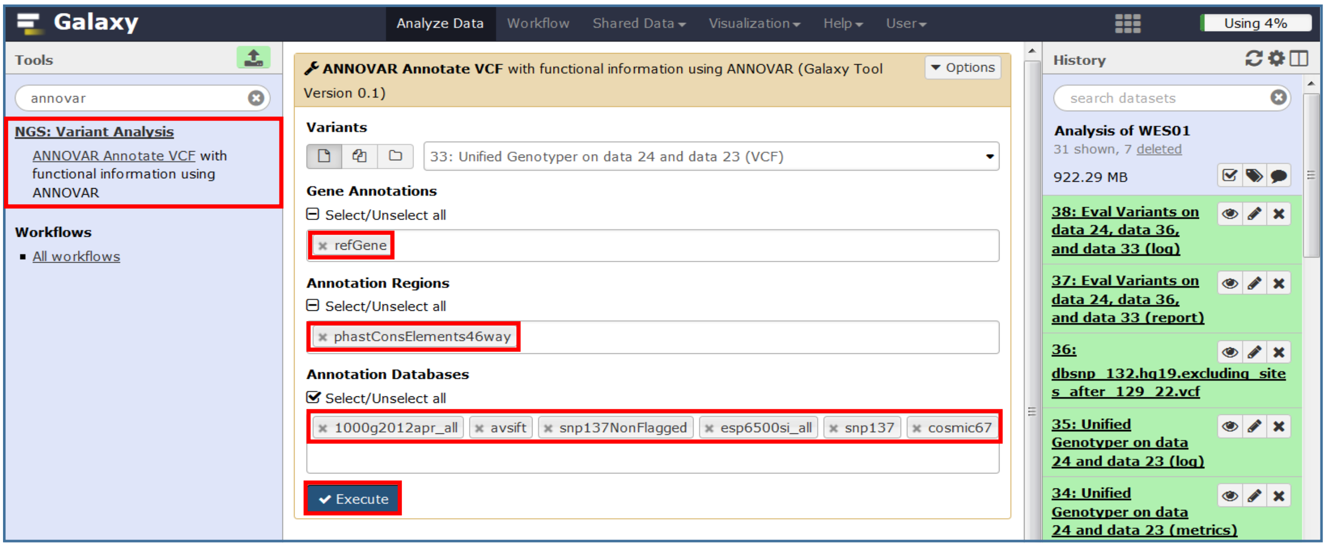
This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Analysis of WES01
31 shown, 7 deleted
922.29 MB

38: Eval Variants on data 24, data 36, and data 33 (log)
37: Eval Variants on data 24, data 36, and data 33 (report)
36: dbsnp_132.hg19.excluding_sites_after_129_22.vcf
35: Unified Genotyper on data 24 and data 23 (log)
34: Unified Genotyper on data 24 and data 23 (metrics)
33: Unified Genotyper on data 24 and data 23 (VCF)
951 lines, 109 comments

3. In **Tool Pane**: Go to **NGS: Variant Analysis** > ANNOVAR Annotate VCF

For Gene Annotation we will use refGene only, which is based on refSeq, since this is the most curated dataset with fewest errors. Segmental duplications (genomicSuperDups) are associated with false positives but we will ignore them in this analysis. Select all Annotation Databases.



ANNOVAR Annotate VCF with functional information using ANNOVAR (Galaxy Tool Version 0.1)

Variants: 33: Unified Genotyper on data 24 and data 23 (VCF)

Gene Annotations: ⊟ Select/Unselect all — × refGene

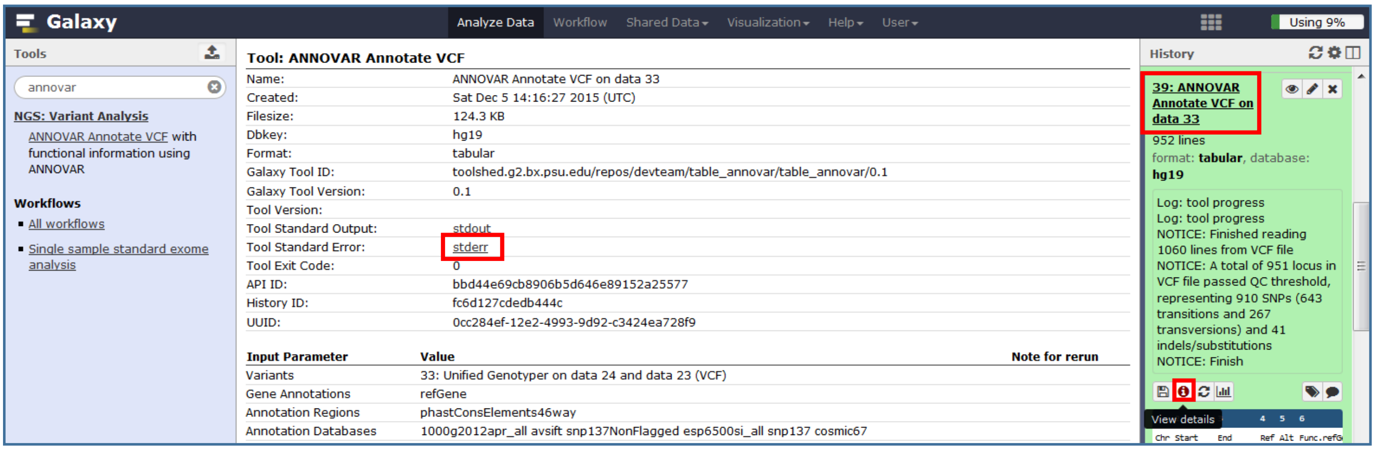
Annotation Regions: ⊟ Select/Unselect all — × phastConsElements46way

Annotation Databases: ☑ Select/Unselect all — × 1000g2012apr_all × avsift × snp137NonFlagged × esp6500si_all × snp137 × cosmic67

Analysis of WES01
31 shown, 7 deleted
922.29 MB

38: Eval Variants on data 24, data 36, and data 33 (log)
37: Eval Variants on data 24, data 36, and data 33 (report)
36: dbsnp_132.hg19.excluding_site s_after_129_22.vcf
35: Unified Genotyper on data 24 and data 23 (log)
34: Unified Genotyper on data 24 and data 23 (metrics)

4. Before we look at the ANNOVAR results, let's check the log file. Click on the info icon 'i'. ANNOVAR prints its log file information to the Tool Standard Error: stderr, click this link to view the log file.



Tool: ANNOVAR Annotate VCF

Name:	ANNOVAR Annotate VCF on data 33
Created:	Sat Dec 5 14:16:27 2015 (UTC)
Filesize:	124.3 KB
Dbkey:	hg19
Format:	tabular
Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/devteam/table_annovar/table_annovar/0.1
Galaxy Tool Version:	0.1
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	bbd44e69cb8906b5d646e89152a25577
History ID:	fc6d127cdedb444c
UUID:	0cc284ef-12e2-4993-9d92-c3424ea728f9

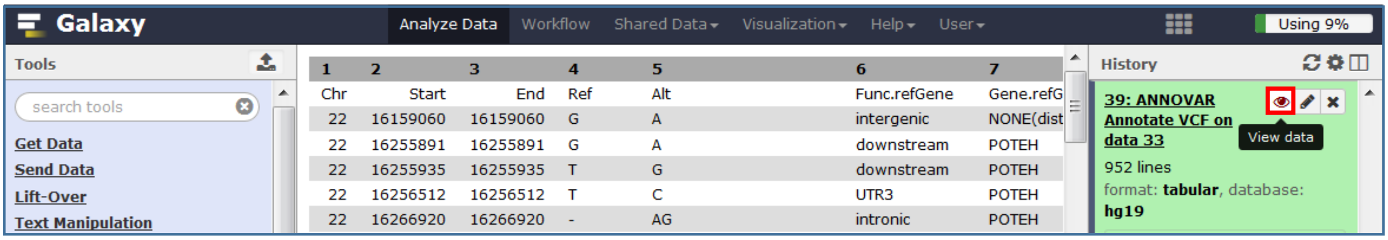
Input Parameter	Value	Note for rerun
Variants	33: Unified Genotyper on data 24 and data 23 (VCF)	
Gene Annotations	refGene	
Annotation Regions	phastConsElements46way	
Annotation Databases	1000g2012apr_all avsift snp137NonFlagged esp6500si_all snp137 cosmic67	

39: ANNOVAR Annotate VCF on data 33
952 lines
format: **tabular**, database: **hg19**

```
Log: tool progress
Log: tool progress
NOTICE: Finished reading 1060 lines from VCF file
NOTICE: A total of 951 locus in VCF file passed QC threshold, representing 910 SNPs (643 transitions and 267 transversions) and 41 indels/substitutions
NOTICE: Finish
```

Q1: From the ANNOVAR log file, how many refGene transcripts were used for annotation?

5. Now view the annotated VCF file by clicking 'View data' and familiarise yourself with the contents.



1	2	3	4	5	6	7
Chr	Start	End	Ref	Alt	Func.refGene	Gene.refG
22	16159060	16159060	G	A	intergenic	NONE(dist
22	16255891	16255891	G	A	downstream	POTEH
22	16255935	16255935	T	G	downstream	POTEH
22	16256512	16256512	T	C	UTR3	POTEH
22	16266920	16266920	-	AG	intronic	POTEH

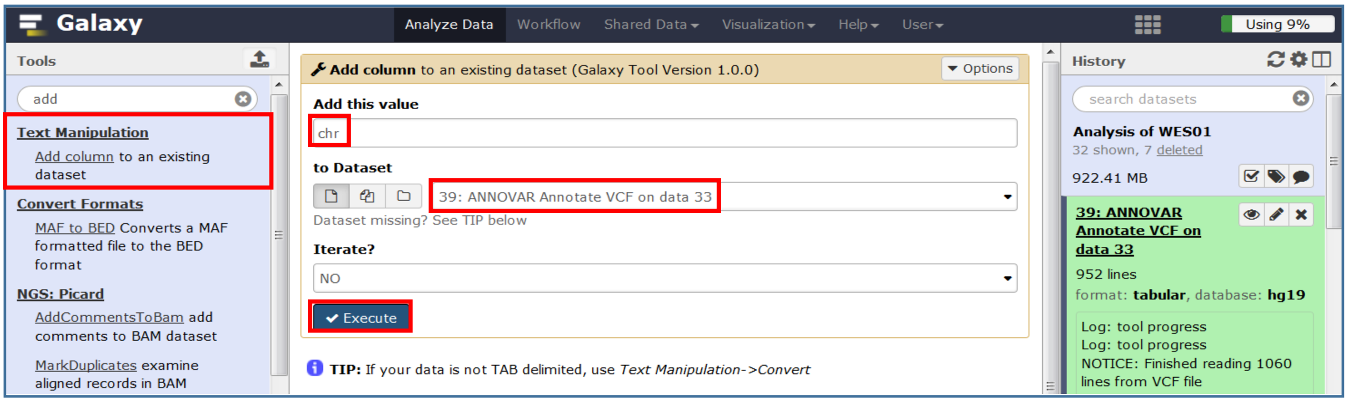
39: ANNOVAR Annotate VCF on data 33
952 lines
format: **tabular**, database: **hg19**

Note that missing data is listed as NA in some columns and -1 in others.

Q2: Is there anything else that would be useful for annotation?

To gather more information we will use SIFT to annotate the variants further. To run SIFT, the VCF file needs to be edited so that it includes four columns in the format (chr#, base pair location, strand and alleles as A/T).

6. In **Tool Pane**: Go to **Text Manipulation** > Add column
Enter 'chr' to Add this value and do not iterate. This will add a column to the end of the VCF file with chr in each row.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9%
Tools
add
Text Manipulation
Add column to an existing dataset
Convert Formats
MAF to BED Converts a MAF formatted file to the BED format
NGS: Picard
AddCommentsToBam add comments to BAM dataset
MarkDuplicates examine aligned records in BAM
Options
Execute
History
search datasets

Add column to an existing dataset (Galaxy Tool Version 1.0.0)

Add this value
chr

to Dataset
39: ANNOVAR Annotate VCF on data 33
Dataset missing? See TIP below

Iterate?
NO

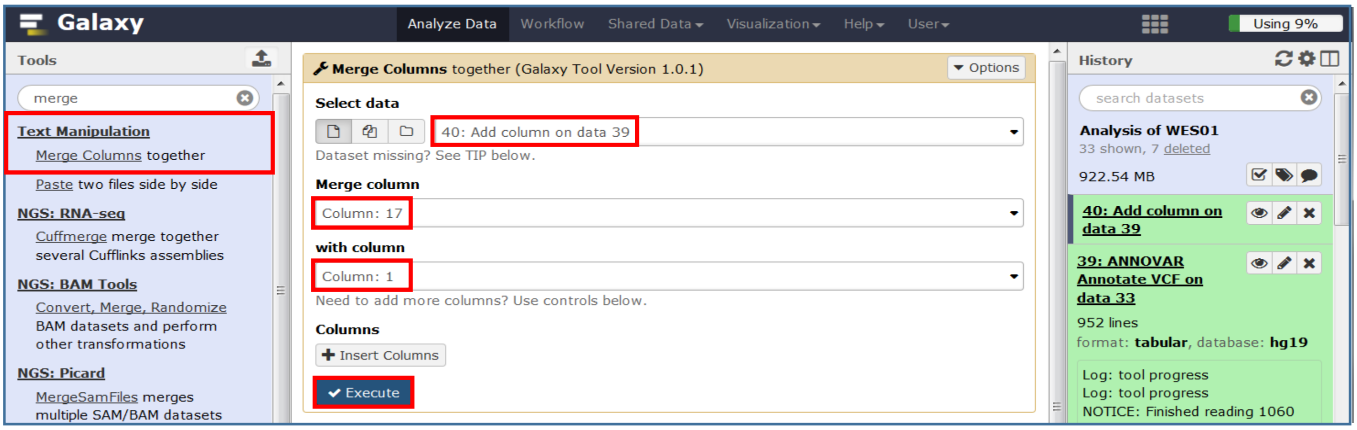
TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Analysis of WES01
32 shown, 7 deleted
922.41 MB

39: ANNOVAR Annotate VCF on data 33
952 lines
format: **tabular**, database: **hg19**

```
Log: tool progress
Log: tool progress
NOTICE: Finished reading 1060 lines from VCF file
```

7. In **Tool Pane**: Go to **Text Manipulation** > Merge Columns
Merge columns 17 and 1 to make a new column of chromosome locations in the format chr#.



Merge Columns together (Galaxy Tool Version 1.0.1)

Select data: 40: Add column on data 39
Dataset missing? See TIP below.

Merge column: Column: 17

with column: Column: 1
Need to add more columns? Use controls below.

Columns

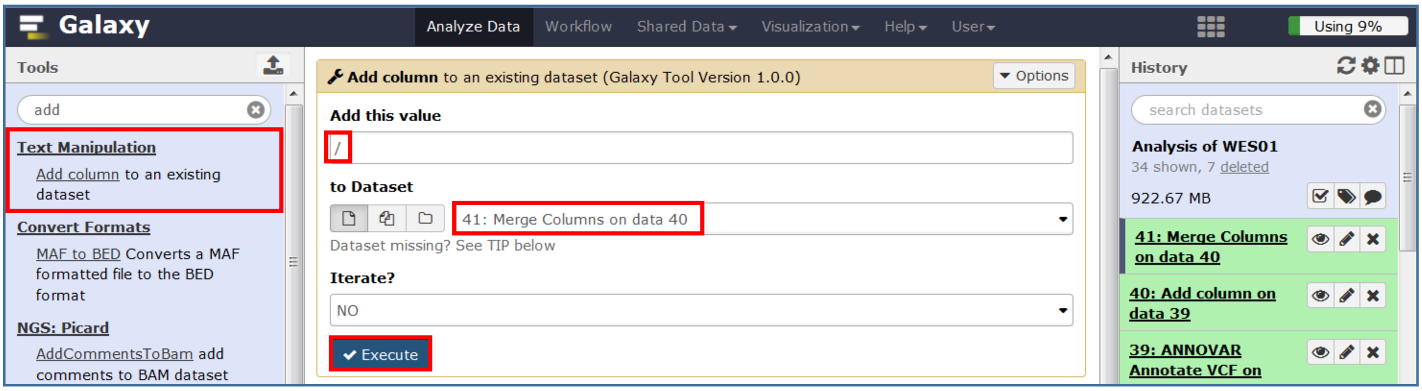
Analysis of WES01
33 shown, 7 deleted
922.54 MB

40: Add column on data 39

39: ANNOVAR Annotate VCF on data 33
952 lines
format: tabular, database: hg19

Log: tool progress
Log: tool progress
NOTICE: Finished reading 1060

8. In **Tool Pane**: Go to **Text Manipulation** > Add column
Enter '/' to Add this value and do not iterate. This will add a column to the end of the VCF file with / in each row.



Add column to an existing dataset (Galaxy Tool Version 1.0.0)

Add this value: /

to Dataset: 41: Merge Columns on data 40
Dataset missing? See TIP below

Iterate?: NO

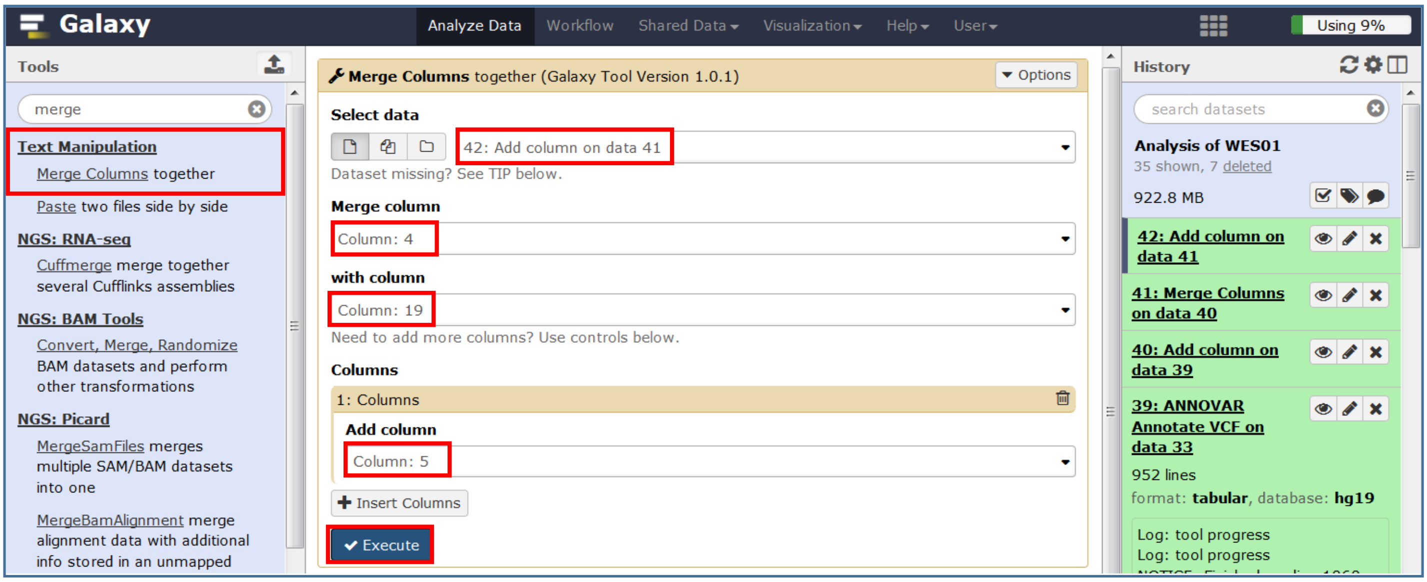
Analysis of WES01
34 shown, 7 deleted
922.67 MB

41: Merge Columns on data 40

40: Add column on data 39

39: ANNOVAR Annotate VCF on

9. In **Tool Pane**: Go to **Text Manipulation** > Merge Columns
Merge columns 4 and 19 to make a new column of reference and alternate alleles in the format ref/alt.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9%
Tools
merge
Text Manipulation
Merge Columns together
Paste two files side by side
NGS: RNA-seq
Cuffmerge merge together several Cufflinks assemblies
NGS: BAM Tools
Convert, Merge, Randomize BAM datasets and perform other transformations
NGS: Picard
MergeSamFiles merges multiple SAM/BAM datasets into one
MergeBamAlignment merge alignment data with additional info stored in an unmapped
Options
Insert Columns
Execute
History
search datasets

Merge Columns together (Galaxy Tool Version 1.0.1)

Select data: 42: Add column on data 41
Dataset missing? See TIP below.

Merge column: Column: 4

with column: Column: 19
Need to add more columns? Use controls below.

Columns

1: Columns
Add column: Column: 5

Analysis of WES01
35 shown, 7 deleted
922.8 MB

42: Add column on data 41

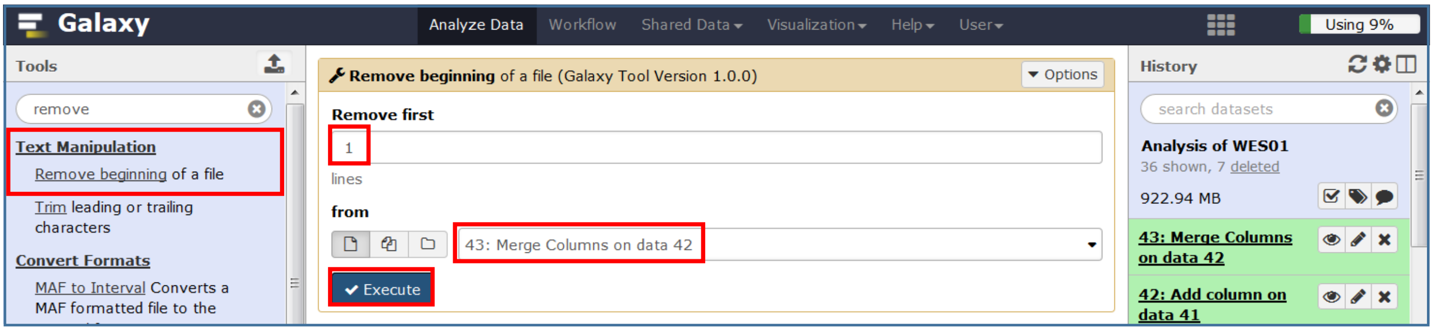
41: Merge Columns on data 40

40: Add column on data 39

39: ANNOVAR Annotate VCF on data 33
952 lines
format: tabular, database: hg19

Log: tool progress
Log: tool progress

10. In **Tool Pane**: Go to **Text Manipulation** > Remove beginning
Delete the first line of the VCF file.



Remove beginning of a file (Galaxy Tool Version 1.0.0)

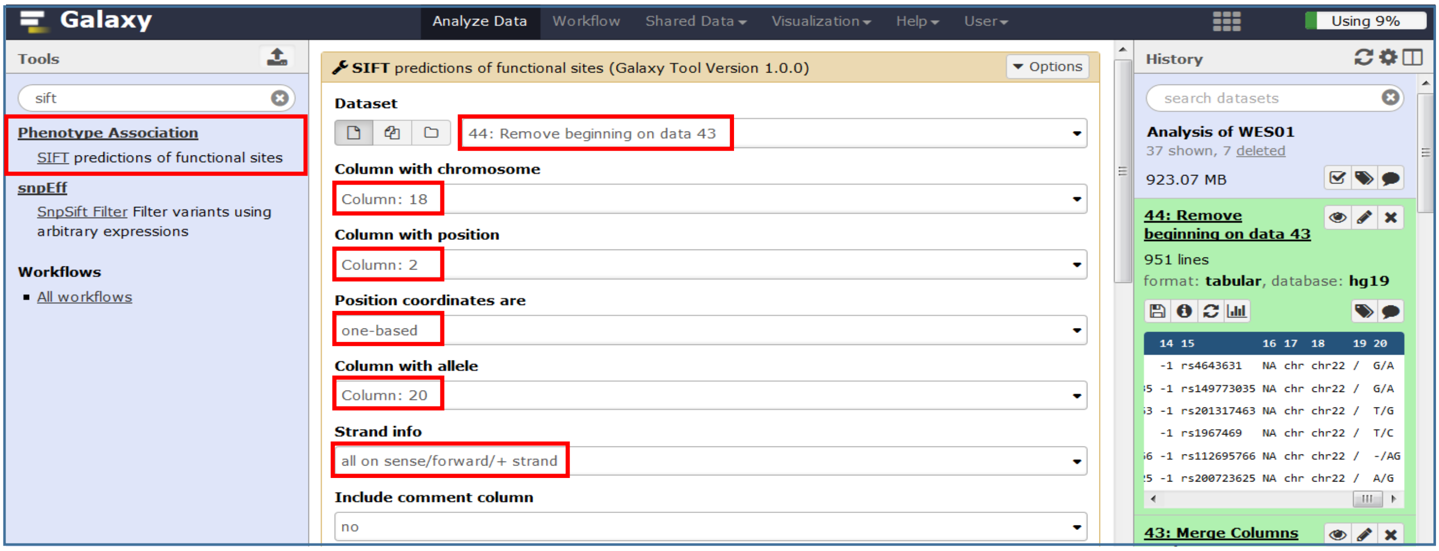
Remove first
1
lines

from
43: Merge Columns on data 42

✔ Execute

We can now use the newly formatted columns of data to run SIFT.

11. In **Tool Pane**: Go to **Phenotype Association** > SIFT



SIFT predictions of functional sites (Galaxy Tool Version 1.0.0)

Dataset
44: Remove beginning on data 43

Column with chromosome
Column: 18

Column with position
Column: 2

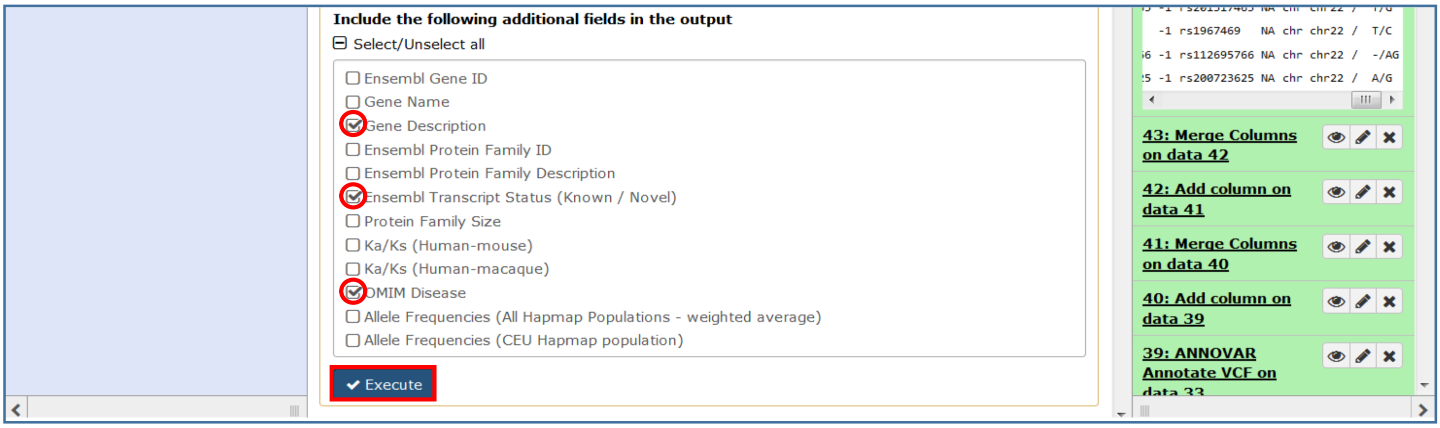
Position coordinates are
one-based

Column with allele
Column: 20

Strand info
all on sense/forward/+ strand

Include comment column
no

```
14  15         16  17  18      19  20
   -1  rs4643631    NA  chr  chr22  /   G/A
5  -1  rs149773035  NA  chr  chr22  /   G/A
3  -1  rs201317463  NA  chr  chr22  /   T/G
   -1  rs1967469    NA  chr  chr22  /   T/C
6  -1  rs112695766  NA  chr  chr22  /   -/AG
5  -1  rs200723625  NA  chr  chr22  /   A/G
```



Include the following additional fields in the output
⊟ Select/Unselect all

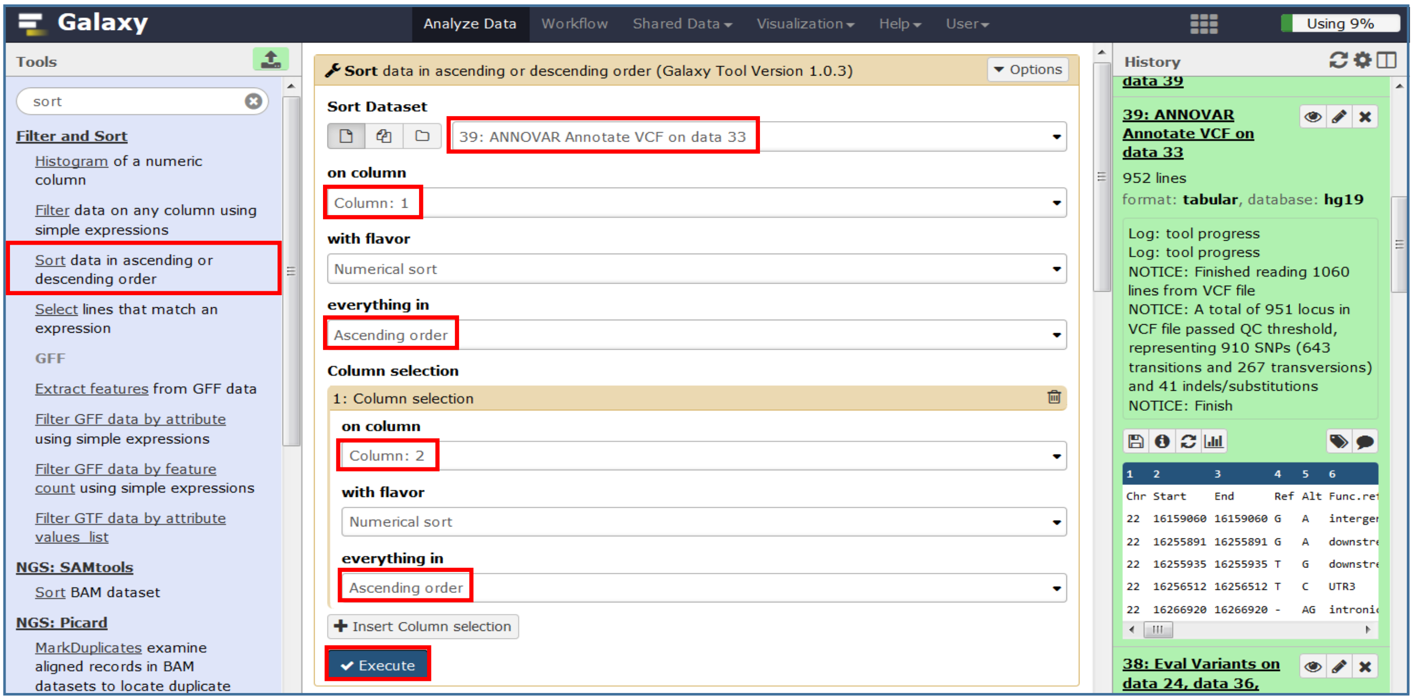
☐ Ensembl Gene ID
☐ Gene Name
☑ Gene Description
☐ Ensembl Protein Family ID
☐ Ensembl Protein Family Description
☑ Ensembl Transcript Status (Known / Novel)
☐ Protein Family Size
☐ Ka/Ks (Human-mouse)
☐ Ka/Ks (Human-macaque)
☑ OMIM Disease
☐ Allele Frequencies (All Hapmap Populations - weighted average)
☐ Allele Frequencies (CEU Hapmap population)

✔ Execute

43: Merge Columns on data 42
42: Add column on data 41
41: Merge Columns on data 40
40: Add column on data 39
39: ANNOVAR Annotate VCF on data 33

After annotating with ANNOVAR and SIFT the results need to be merged which requires both files to be sorted.

12. In **Tool Pane**: Go to **Filter and Sort** > Sort
Sort the ANNOVAR result by chromosome (column 1) and base pair location (column 2) in ascending order.



Sort data in ascending or descending order (Galaxy Tool Version 1.0.3)

Sort Dataset: 39: ANNOVAR Annotate VCF on data 33

on column: Column: 1

with flavor: Numerical sort

everything in: Ascending order

Column selection

1: Column selection

on column: Column: 2

with flavor: Numerical sort

everything in: Ascending order

39: ANNOVAR Annotate VCF on data 33

952 lines

format: **tabular**, database: **hg19**

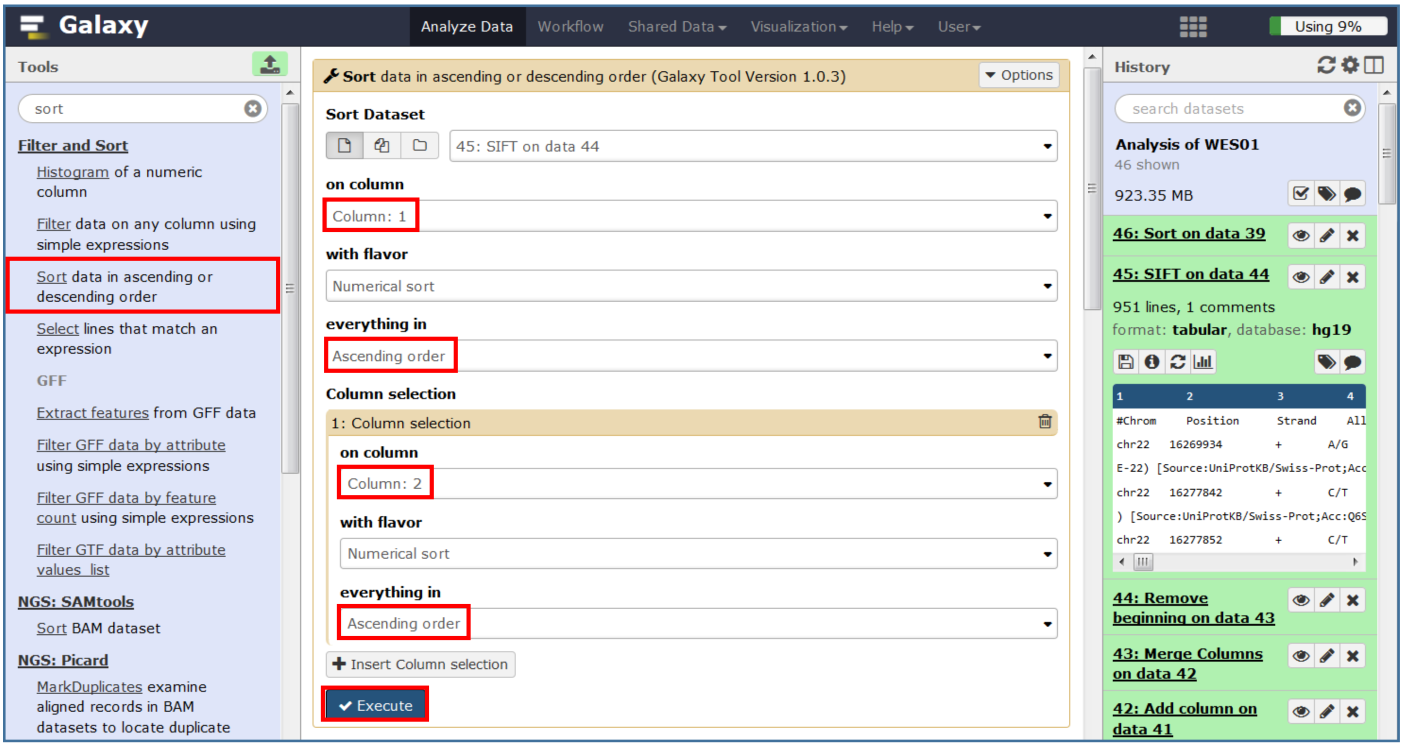
Log: tool progress
Log: tool progress
NOTICE: Finished reading 1060 lines from VCF file
NOTICE: A total of 951 locus in VCF file passed QC threshold, representing 910 SNPs (643 transitions and 267 transversions) and 41 indels/substitutions
NOTICE: Finish

1	2	3	4	5	6
Chr	Start	End	Ref	Alt	Func.re
22	16159060	16159060	G	A	interger
22	16255891	16255891	G	A	downstre
22	16255935	16255935	T	G	downstre
22	16256512	16256512	T	C	UTR3
22	16266920	16266920	-	AG	intronic

38: Eval Variants on data 24, data 36,

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9%
Tools
sort
Filter and Sort
Histogram of a numeric column
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Filter GTF data by attribute values_list
NGS: SAMtools
Sort BAM dataset
NGS: Picard
MarkDuplicates examine aligned records in BAM datasets to locate duplicate
Options
Insert Column selection
Execute
History
data 39

13. In **Tool Pane**: Go to **Filter and Sort** > Sort
Sort the SIFT result by chromosome (column 1) and base pair location (column 2) in ascending order.



Sort data in ascending or descending order (Galaxy Tool Version 1.0.3)

Sort Dataset: 45: SIFT on data 44

on column: Column: 1

with flavor: Numerical sort

everything in: Ascending order

Column selection

1: Column selection

on column: Column: 2

with flavor: Numerical sort

everything in: Ascending order

Analysis of WES01

46 shown

923.35 MB

46: Sort on data 39

45: SIFT on data 44

951 lines, 1 comments

format: **tabular**, database: **hg19**

```
1          2            3          4
#Chrom     Position     Strand     All
chr22   16269934        +          A/G
E-22) [Source:UniProtKB/Swiss-Prot;Acc
chr22   16277842        +          C/T
) [Source:UniProtKB/Swiss-Prot;Acc:Q6S
chr22   16277852        +          C/T
```

44: Remove beginning on data 43

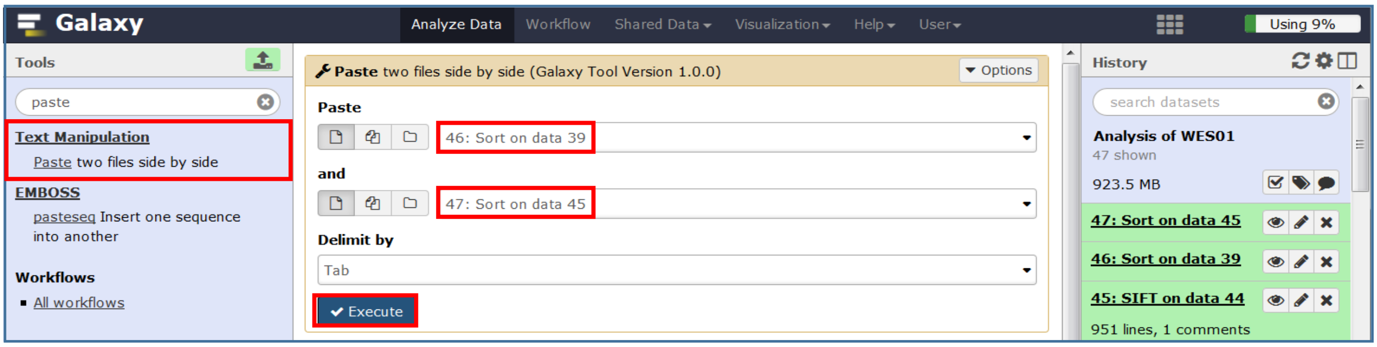
43: Merge Columns on data 42

42: Add column on data 41

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9%
Tools
sort
Filter and Sort
Histogram of a numeric column
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data
Filter GFF data by attribute using simple expressions
Filter GFF data by feature count using simple expressions
Filter GTF data by attribute values_list
NGS: SAMtools
Sort BAM dataset
NGS: Picard
MarkDuplicates examine aligned records in BAM datasets to locate duplicate
Options
Insert Column selection
Execute
History
search datasets

Having sorted the ANNOVAR and SIFT results we can now safely merge them.

14. In **Tool Pane**: Go to **Text Manipulation** > Paste



Tools: paste

Text Manipulation
Paste two files side by side

EMBOSS
pasteseq Insert one sequence into another

Workflows
- All workflows

Paste two files side by side (Galaxy Tool Version 1.0.0)

Paste: 46: Sort on data 39

and: 47: Sort on data 45

Delimit by: Tab

✓ Execute

History: search datasets

Analysis of WES01
47 shown
923.5 MB

47: Sort on data 45
46: Sort on data 39
45: SIFT on data 44
951 lines, 1 comments

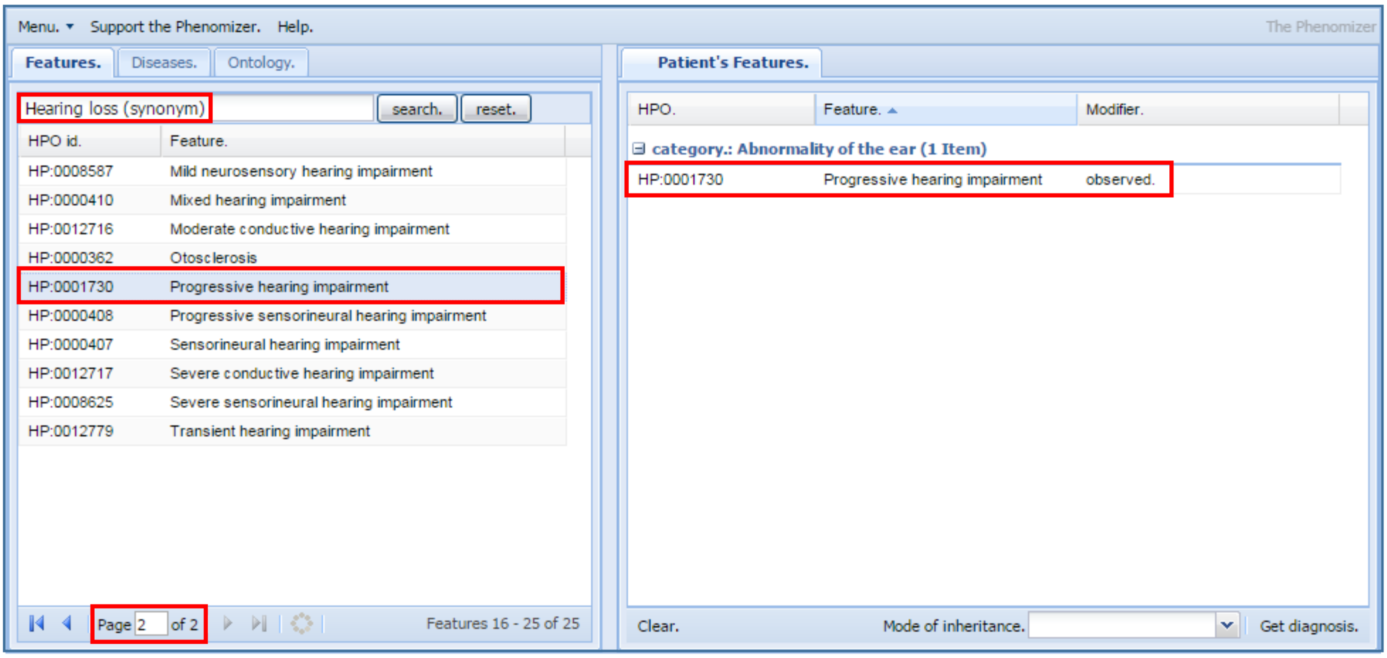
Generate a list of candidate genes

The patient is described as a 25-year-old male with no relevant family history who presented with hearing loss in the left ear, some deterioration in visual acuity especially at night, some numbness of his left arm and some weakness of his left shoulder. An MRI scan showed left acoustic neuromas, a mass under his left scapula and a mass impinging on his left brachial plexus.

We will use Phenomizer (http://compbio.charite.de/phenomizer/ Köhler et al 2009) to generate a list of candidate genes.

1. Open the Phenomizer website (http://compbio.charite.de/phenomizer/)

2. Enter the patient's phenotype, one at a time, click search and select the most relevant HPO term by double clicking the HPO id so that it appears in the Patient's Features window. Note that results can be spread over multiple pages.



Hearing loss (synonym)

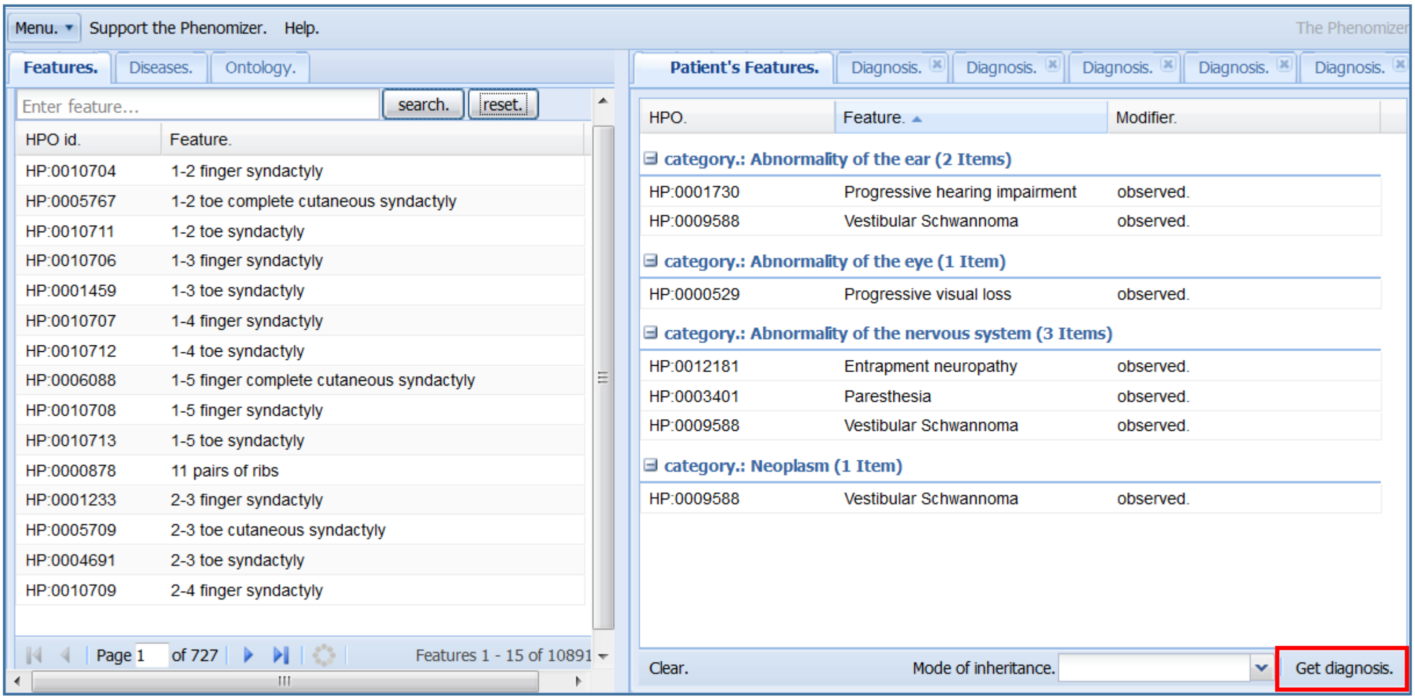
HPO id.	Feature.
HP:0008587	Mild neurosensory hearing impairment
HP:0000410	Mixed hearing impairment
HP:0012716	Moderate conductive hearing impairment
HP:0000362	Otosclerosis
HP:0001730	Progressive hearing impairment
HP:0000408	Progressive sensorineural hearing impairment
HP:0000407	Sensorineural hearing impairment
HP:0012717	Severe conductive hearing impairment
HP:0008625	Severe sensorineural hearing impairment
HP:0012779	Transient hearing impairment

Page 2 of 2 Features 16 - 25 of 25

HPO.	Feature.	Modifier.
category.: Abnormality of the ear (1 Item)		
HP:0001730	Progressive hearing impairment	observed.

For a definition of specific HPO terms, modify the HPO id at the end of this link: http://purl.obolibrary.org/obo/HP_0000365
If you have trouble finding HPO ids for particular phenotypes try entering a more general term or googling for them eg 'human phenotype ontology numbness'.

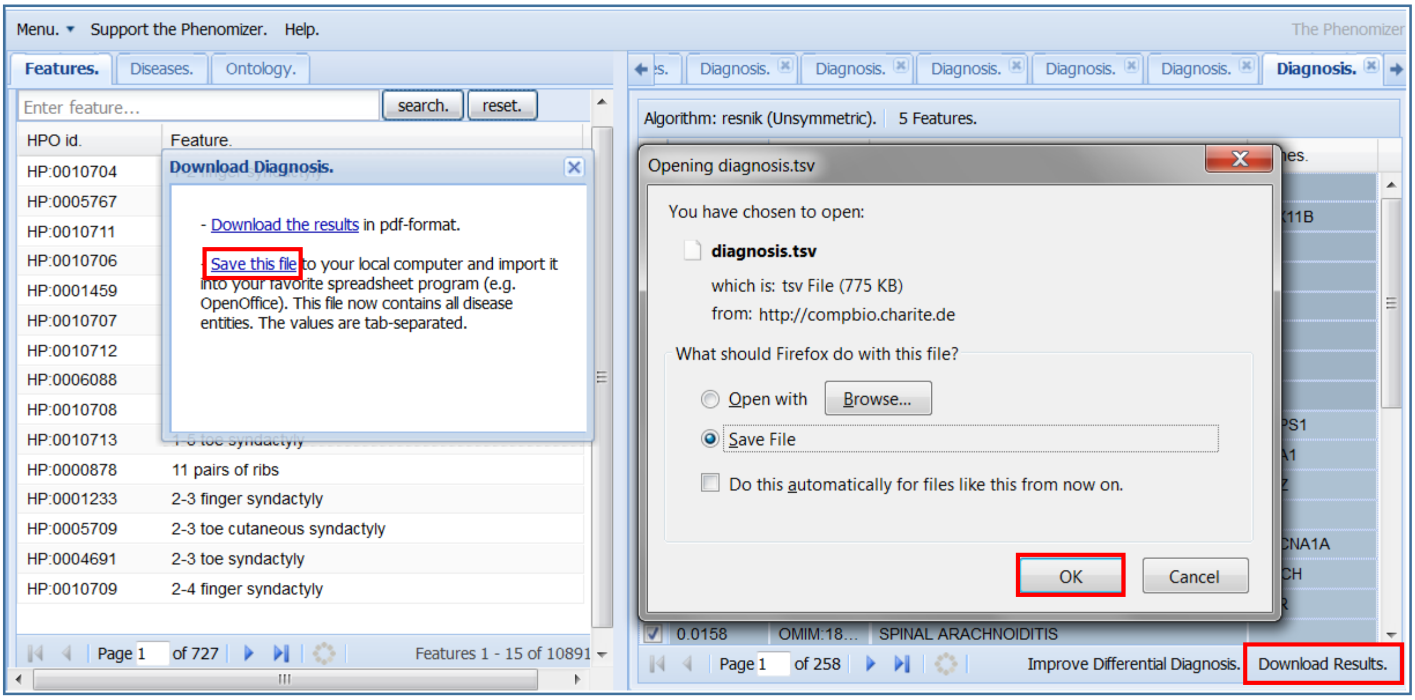
3. Having entered all of the patient's phenotypes click 'Get diagnosis' to generate a list of associated diseases and genes.



HPO id.	Feature.
HP:0010704	1-2 finger syndactyly
HP:0005767	1-2 toe complete cutaneous syndactyly
HP:0010711	1-2 toe syndactyly
HP:0010706	1-3 finger syndactyly
HP:0001459	1-3 toe syndactyly
HP:0010707	1-4 finger syndactyly
HP:0010712	1-4 toe syndactyly
HP:0006088	1-5 finger complete cutaneous syndactyly
HP:0010708	1-5 finger syndactyly
HP:0010713	1-5 toe syndactyly
HP:0000878	11 pairs of ribs
HP:0001233	2-3 finger syndactyly
HP:0005709	2-3 toe cutaneous syndactyly
HP:0004691	2-3 toe syndactyly
HP:0010709	2-4 finger syndactyly

HPO.	Feature.	Modifier.
category.: Abnormality of the ear (2 Items)		
HP:0001730	Progressive hearing impairment	observed.
HP:0009588	Vestibular Schwannoma	observed.
category.: Abnormality of the eye (1 Item)		
HP:0000529	Progressive visual loss	observed.
category.: Abnormality of the nervous system (3 Items)		
HP:0012181	Entrapment neuropathy	observed.
HP:0003401	Paresthesia	observed.
HP:0009588	Vestibular Schwannoma	observed.
category.: Neoplasm (1 Item)		
HP:0009588	Vestibular Schwannoma	observed.

4. Click 'Download Results', then click the 'Save this file' link to save the 'diagnosis.tsv' file to your computer.



Algorithm: resnik (Unsymmetric). 5 Features.

Download Diagnosis.

- Download the results in pdf-format.

Save this file to your local computer and import it into your favorite spreadsheet program (e.g. OpenOffice). This file now contains all disease entities. The values are tab-separated.

Opening diagnosis.tsv

You have chosen to open:

diagnosis.tsv

which is: tsv File (775 KB)
from: http://compbio.charite.de

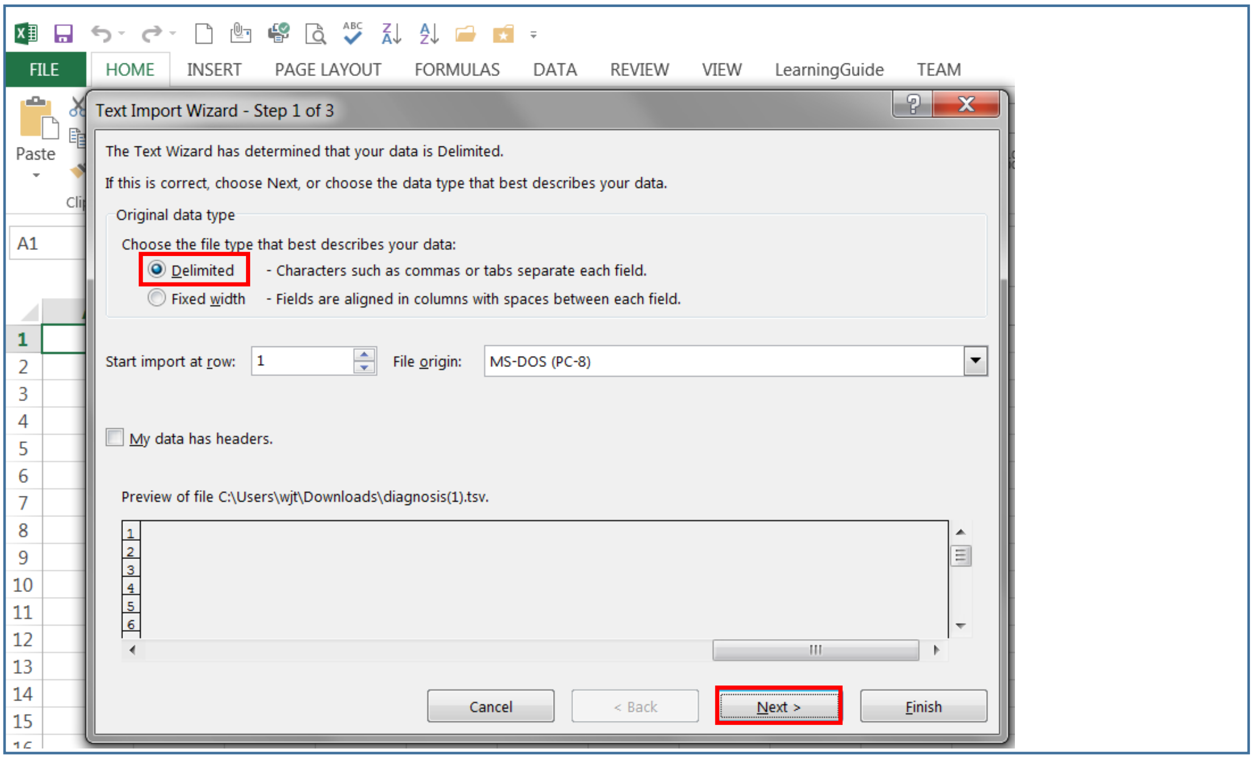
What should Firefox do with this file?

○ Open with
● Save File
□ Do this automatically for files like this from now on.

HPO id.	Feature.
HP:0010704	
HP:0005767	
HP:0010711	
HP:0010706	
HP:0001459	
HP:0010707	
HP:0010712	
HP:0006088	
HP:0010708	
HP:0010713	
HP:0000878	11 pairs of ribs
HP:0001233	2-3 finger syndactyly
HP:0005709	2-3 toe cutaneous syndactyly
HP:0004691	2-3 toe syndactyly
HP:0010709	2-4 finger syndactyly

☑ 0.0158 OMIM:18... SPINAL ARACHNOIDITIS

5. Open the Phenomizer results file ‘diagnosis.tsv’ in Excel, choose ‘Delimited’ in the Text Import Wizard and click next.



Text Import Wizard - Step 1 of 3

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.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

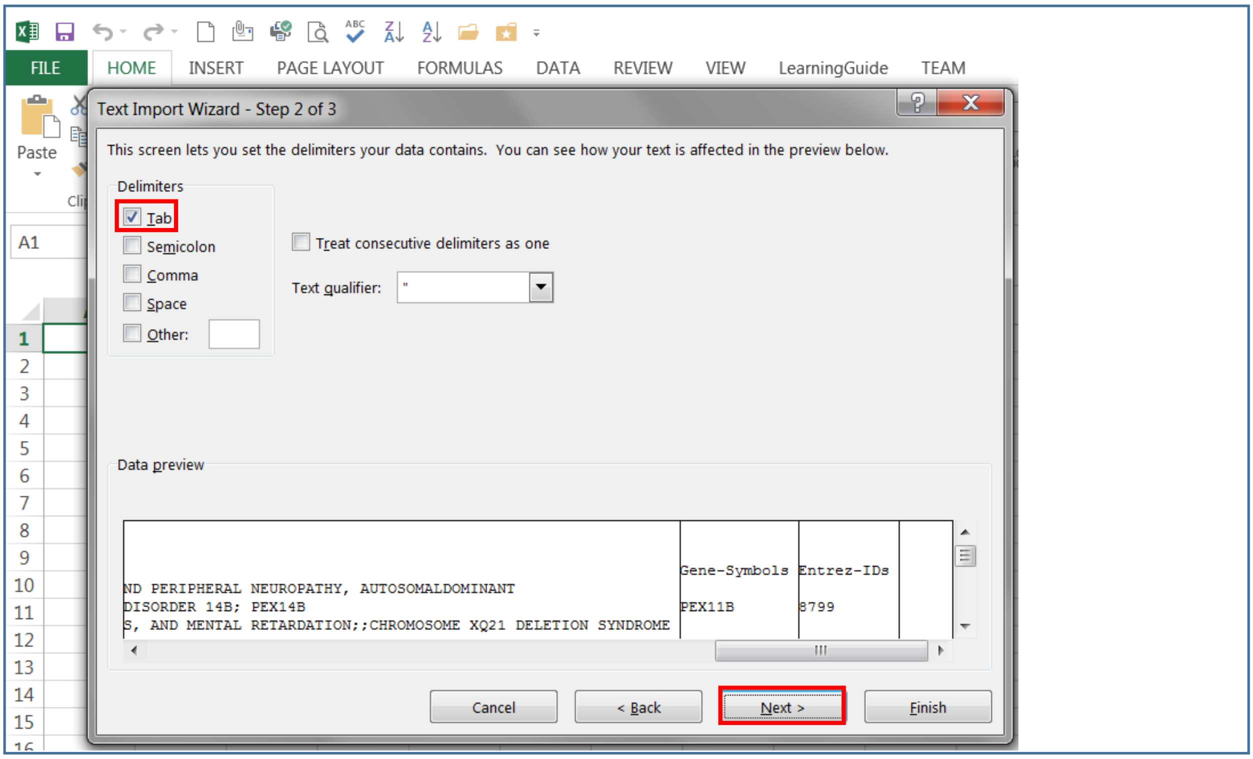
○ Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: MS-DOS (PC-8)

□ My data has headers.

Preview of file C:\Users\wjt\Downloads\diagnosis(1).tsv.

6. Choose ‘Tab’ as the delimiter and click next



Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☑ Tab
□ Semicolon
□ Comma
□ Space
□ Other:

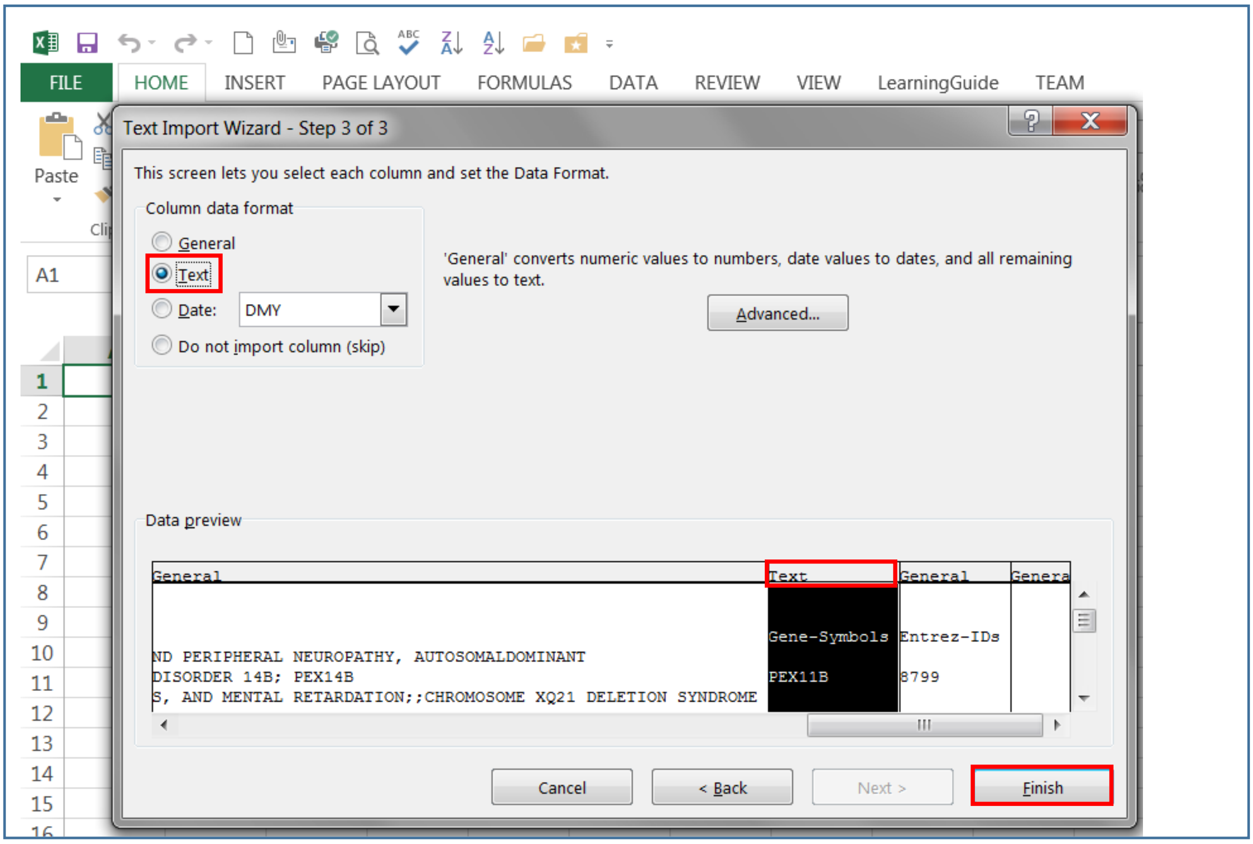
□ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
                                                            Gene-Symbols Entrez-IDs
ND PERIPHERAL NEUROPATHY, AUTOSOMALDOMINANT
DISORDER 14B; PEX14B                                        PEX11B       8799
S, AND MENTAL RETARDATION;;CHROMOSOME XQ21 DELETION SYNDROME
```

7. Use the scroll bar to navigate to the Gene-Symbol column and select ‘Text’ as the column data format, then click Finish. Setting ‘Text’ as the data format for gene symbols is important because it will prevent Excel from automatically interpreting some gene names as dates, eg ‘SEPT12’ will otherwise be erroneously interpreted as ‘Sep-12’.



Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

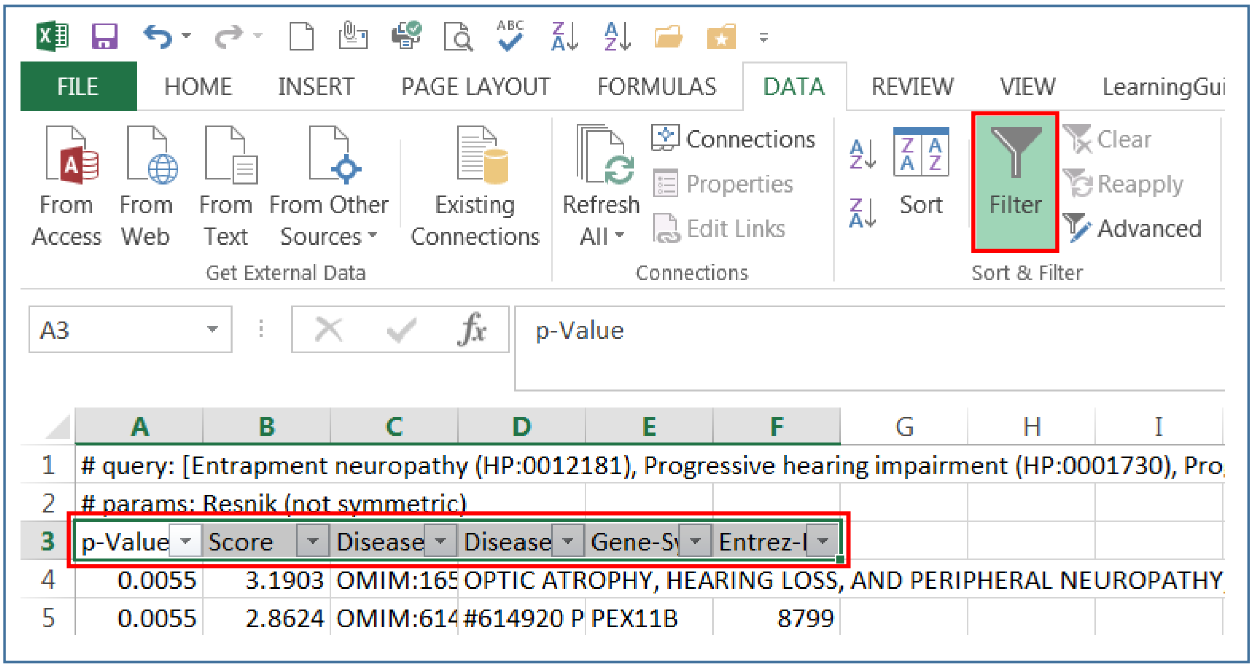
○ General
● Text
○ Date: DMY
○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Data preview

```
General                                                       Text          General      Genera
                                                              Gene-Symbols  Entrez-IDs
ND PERIPHERAL NEUROPATHY, AUTOSOMALDOMINANT
DISORDER 14B; PEX14B                                          PEX11B        8799
S, AND MENTAL RETARDATION;;CHROMOSOME XQ21 DELETION SYNDROME
```

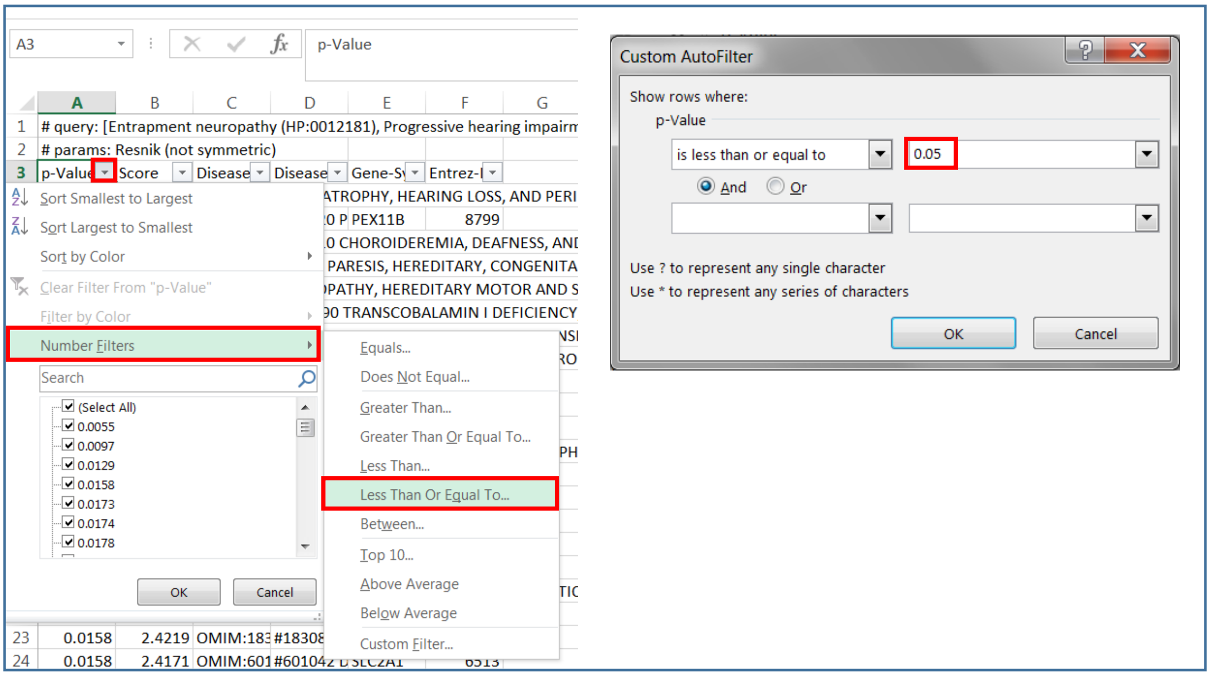
8. Add filters to the columns by selecting the column titles in row 3, hold down left click and mouse over all 6 columns, then select Filter. This will add a drop down box to each column title that enables filtering and sorting based on the column contents.



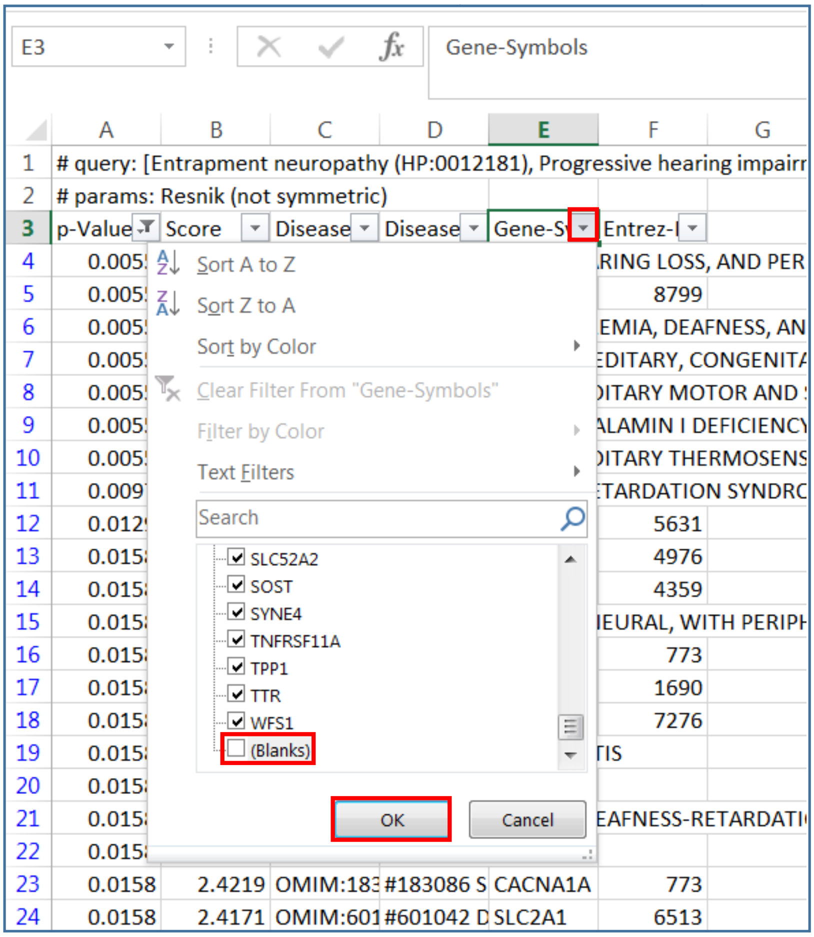
A3 p-Value

	A	B	C	D	E	F
1	# query: [Entrapment neuropathy (HP:0012181), Progressive hearing impairment (HP:0001730), Pro					
2	# params: Resnik (not symmetric)					
3	p-Value	Score	Disease	Disease	Gene-Sy	Entrez-I
4	0.0055	3.1903	OMIM:165	OPTIC ATROPHY, HEARING LOSS, AND PERIPHERAL NEUROPATHY		
5	0.0055	2.8624	OMIM:614	#614920 P	PEX11B	8799

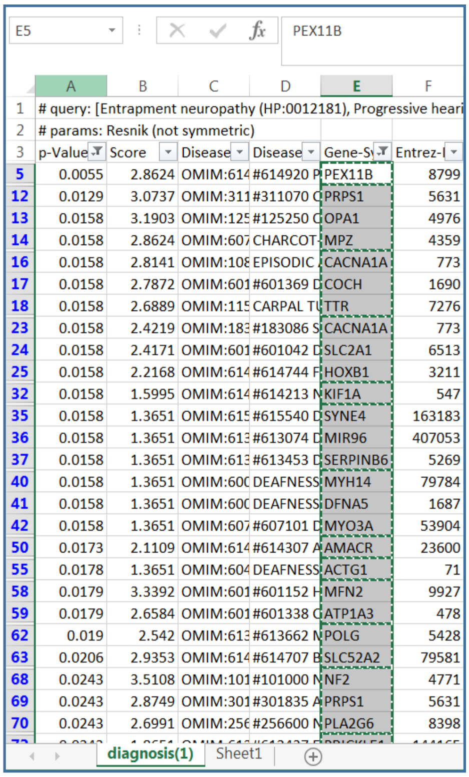
9. Use the p-value filter to select genes that are associated with the phenotype with a p-value less than or equal to 0.05.



10. Now click on the 'Gene-Symbol' filter, scroll to the bottom of the list and uncheck the '(Blanks)' box to exclude rows that do not have a gene symbol.

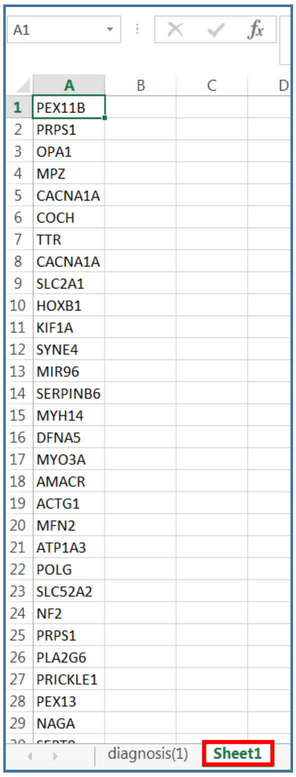


11. Select all of the genes with p-value ≤0.05 by clicking on the first gene and holding down Shift, Ctrl and down arrow. Copy the highlighted list of genes by holding down Ctrl and C then click on a new sheet (bottom left) and paste the gene list into a new column by clicking a cell and holding down Ctrl and V.



E5 fx PEX11B

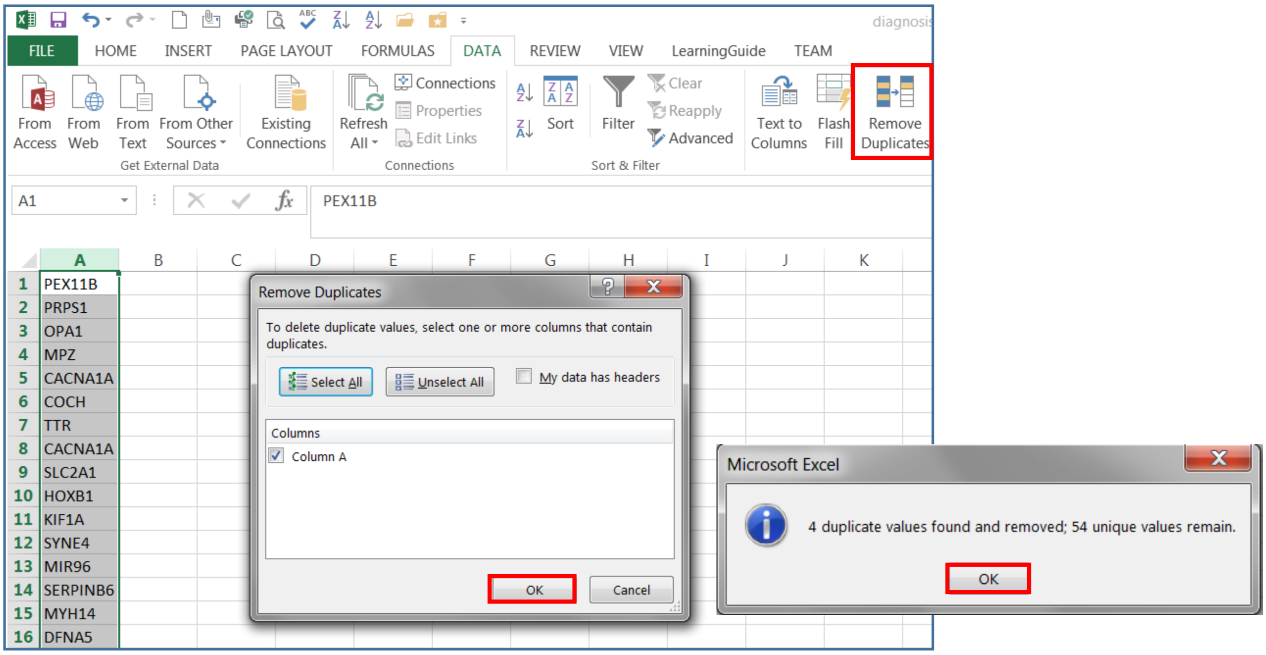
	A	B	C	D	E	F
1	# query: [Entrapment neuropathy (HP:0012181), Progressive heari					
2	# params: Resnik (not symmetric)					
3	p-Value	Score	Disease	Disease	Gene-Sy	Entrez-I
5	0.0055	2.8624	OMIM:614	#614920 P	PEX11B	8799
12	0.0129	3.0737	OMIM:311	#311070 C	PRPS1	5631
13	0.0158	3.1903	OMIM:125	#125250 C	OPA1	4976
14	0.0158	2.8624	OMIM:607	CHARCOT-	MPZ	4359
16	0.0158	2.8141	OMIM:108	EPISODIC A	CACNA1A	773
17	0.0158	2.7872	OMIM:601	#601369 D	COCH	1690
18	0.0158	2.6889	OMIM:115	CARPAL TU	TTR	7276
23	0.0158	2.4219	OMIM:183	#183086 S	CACNA1A	773
24	0.0158	2.4171	OMIM:601	#601042 D	SLC2A1	6513
25	0.0158	2.2168	OMIM:614	#614744 F	HOXB1	3211
32	0.0158	1.5995	OMIM:614	#614213 N	KIF1A	547
35	0.0158	1.3651	OMIM:615	#615540 D	SYNE4	163183
36	0.0158	1.3651	OMIM:613	#613074 D	MIR96	407053
37	0.0158	1.3651	OMIM:613	#613453 D	SERPINB6	5269
40	0.0158	1.3651	OMIM:600	DEAFNESS	MYH14	79784
41	0.0158	1.3651	OMIM:600	DEAFNESS	DFNA5	1687
42	0.0158	1.3651	OMIM:607	#607101 D	MYO3A	53904
50	0.0173	2.1109	OMIM:614	#614307 A	AMACR	23600
55	0.0178	1.3651	OMIM:604	DEAFNESS	ACTG1	71
58	0.0179	3.3392	OMIM:601	#601152 H	MFN2	9927
59	0.0179	2.6584	OMIM:601	#601338 C	ATP1A3	478
62	0.019	2.542	OMIM:613	#613662 N	POLG	5428
63	0.0206	2.9353	OMIM:614	#614707 B	SLC52A2	79581
68	0.0243	3.5108	OMIM:101	#101000 N	NF2	4771
69	0.0243	2.8749	OMIM:301	#301835 A	PRPS1	5631
70	0.0243	2.6991	OMIM:256	#256600 N	PLA2G6	8398



A1 fx

	A	B	C	D
1	PEX11B			
2	PRPS1			
3	OPA1			
4	MPZ			
5	CACNA1A			
6	COCH			
7	TTR			
8	CACNA1A			
9	SLC2A1			
10	HOXB1			
11	KIF1A			
12	SYNE4			
13	MIR96			
14	SERPINB6			
15	MYH14			
16	DFNA5			
17	MYO3A			
18	AMACR			
19	ACTG1			
20	MFN2			
21	ATP1A3			
22	POLG			
23	SLC52A2			
24	NF2			
25	PRPS1			
26	PLA2G6			
27	PRICKLE1			
28	PEX13			
29	NAGA			

12. Now highlight the gene list (column A in Sheet 1), select ‘Remove Duplicates’ and click OK to generate a unique list of genes.



A1 fx PEX11B

	A
1	PEX11B
2	PRPS1
3	OPA1
4	MPZ
5	CACNA1A
6	COCH
7	TTR
8	CACNA1A
9	SLC2A1
10	HOXB1
11	KIF1A
12	SYNE4
13	MIR96
14	SERPINB6
15	MYH14
16	DFNA5

Remove Duplicates

To delete duplicate values, select one or more columns that contain duplicates.

Select All   Unselect All   □ My data has headers

Columns

☑ Column A

OK   Cancel

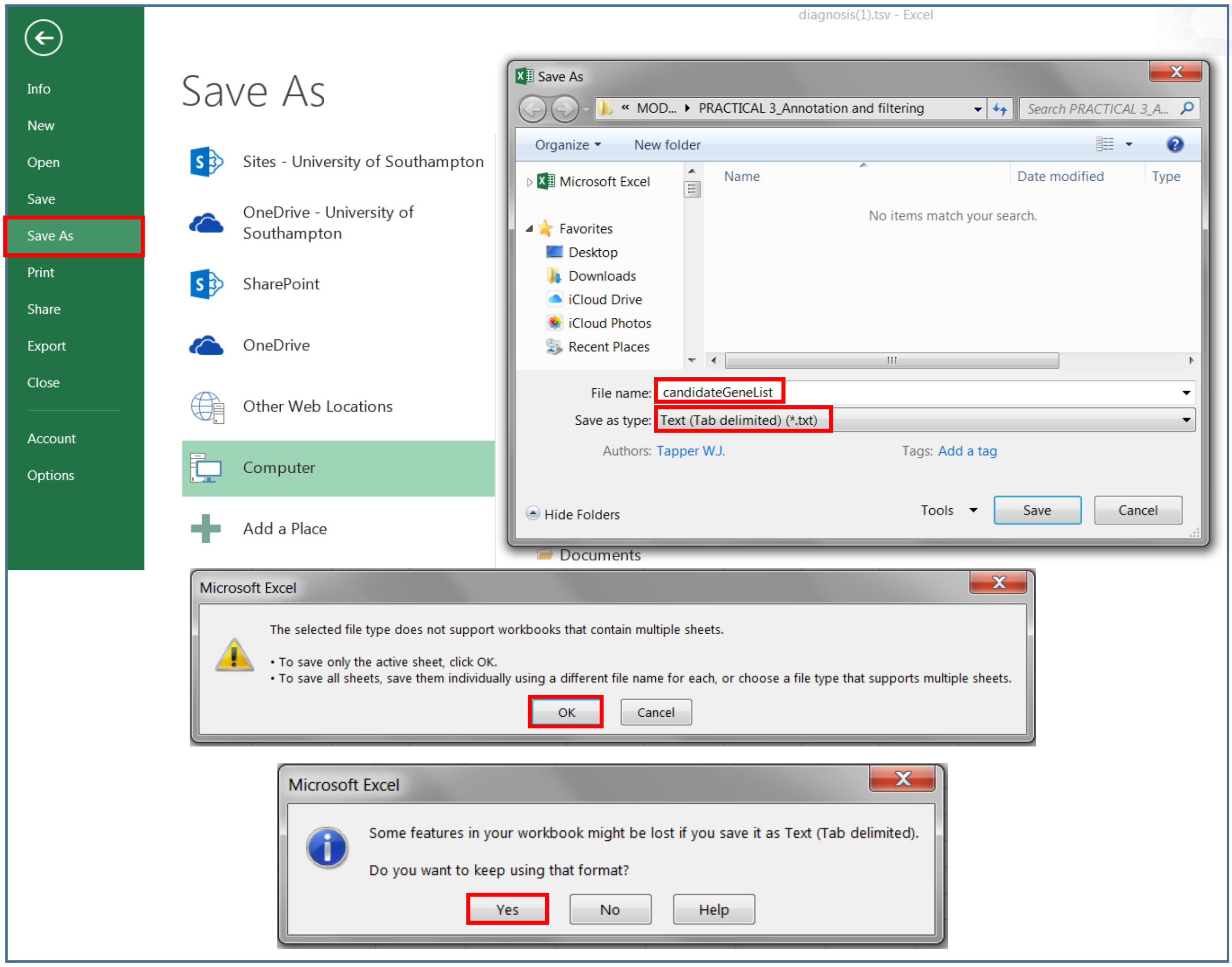
Microsoft Excel

4 duplicate values found and removed; 54 unique values remain.

OK

Depending on the HPO ids that were selected in Phenomizer, this process should yield a unique list of 54 candidate genes that are associated with the patients phenotype (p-value 0.05).

13. Now save the unique list of candidate genes as a tab delimited text file. File > Save as > Save as type ‘Text (Tab delimited) (*.txt), click save ok and yes in the following prompts.



File name: candidateGeneList

Save as type: Text (Tab delimited) (*.txt)

Authors: Tapper W.J.

Tags: Add a tag

Microsoft Excel

The selected file type does not support workbooks that contain multiple sheets.

- To save only the active sheet, click OK.
- To save all sheets, save them individually using a different file name for each, or choose a file type that supports multiple sheets.

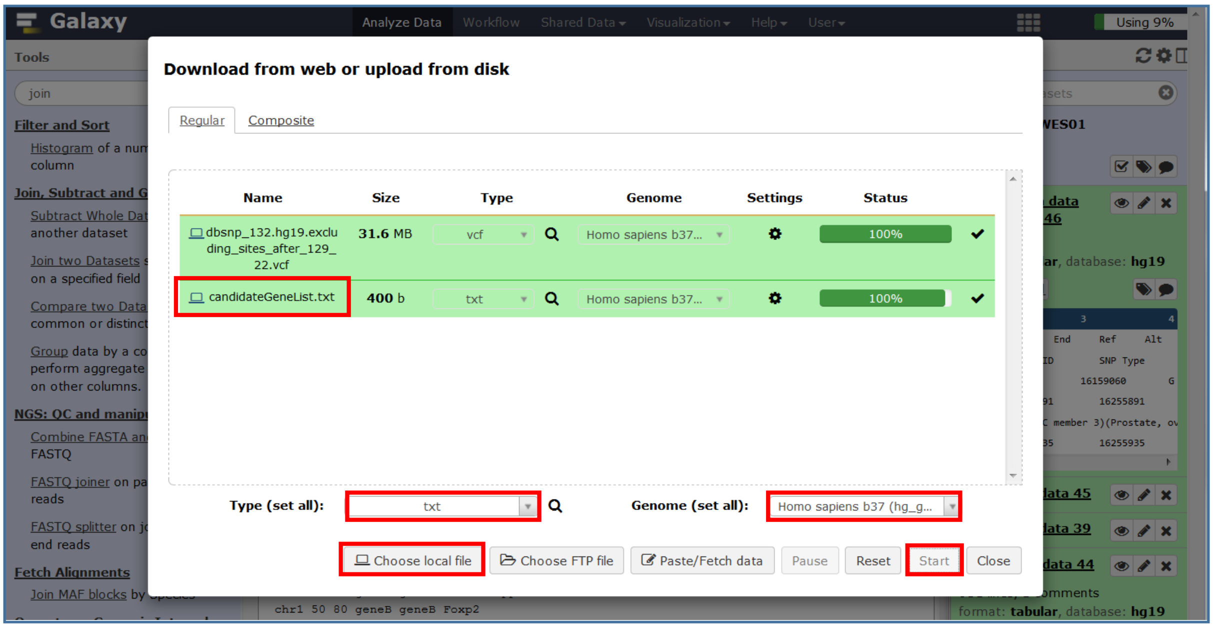
Microsoft Excel

Some features in your workbook might be lost if you save it as Text (Tab delimited).

Do you want to keep using that format?

We will now cross reference the list of candidate genes with the annotated VCF file.

14. Upload the list of candidate genes to Galaxy.


Galaxy
Analyze Data
Workflow
Shared Data ▾
Visualization ▾
Help ▾
User ▾
Using 9%
Tools
join
Filter and Sort
Histogram of a numerical column
Join, Subtract and Group
Subtract Whole Dataset from another dataset
Join two Datasets side by side on a specified field
Compare two Datasets to find common or distinct rows
Group data by a column and perform aggregate operation on other columns.
NGS: QC and manipulation
Combine FASTA and QUAL into FASTQ
FASTQ joiner on paired end reads
FASTQ splitter on joined paired end reads
Fetch Alignments
Join MAF blocks by Species
Regular
Composite
Choose local file
Choose FTP file
Paste/Fetch data
Pause
Reset
Start
Close

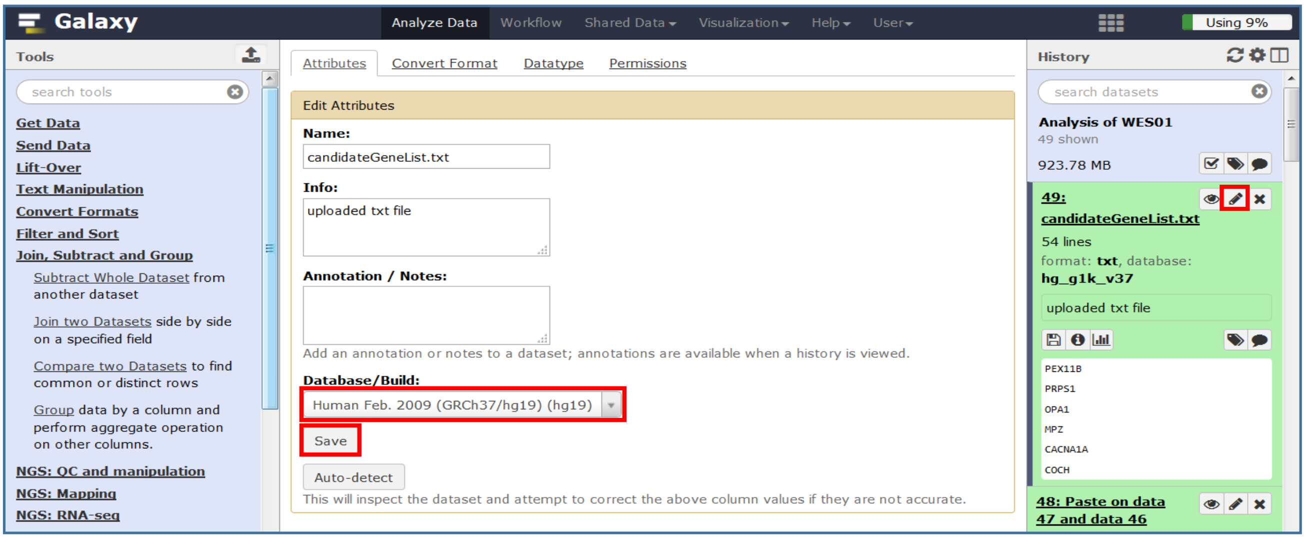
Download from web or upload from disk

Name	Size	Type	Genome	Settings	Status
dbsnp_132.hg19.excluding_sites_after_129_22.vcf	31.6 MB	vcf	Homo sapiens b37...	⚙	100% ✓
candidateGeneList.txt	400 b	txt	Homo sapiens b37...	⚙	100% ✓

Type (set all): txt

Genome (set all): Homo sapiens b37 (hg_g...

15. Click edit attributes and change the database for the candidate gene list to hg19.



Edit Attributes

Name: candidateGeneList.txt

Info: uploaded txt file

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

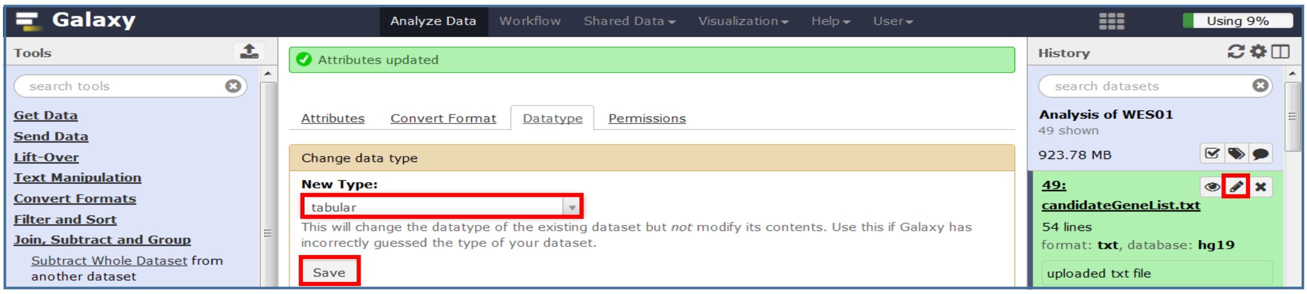
Database/Build: Human Feb. 2009 (GRCh37/hg19) (hg19)

Save

Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

16. Click edit attributes, select the Datatype tab and change the datatype to tabular.



Attributes updated

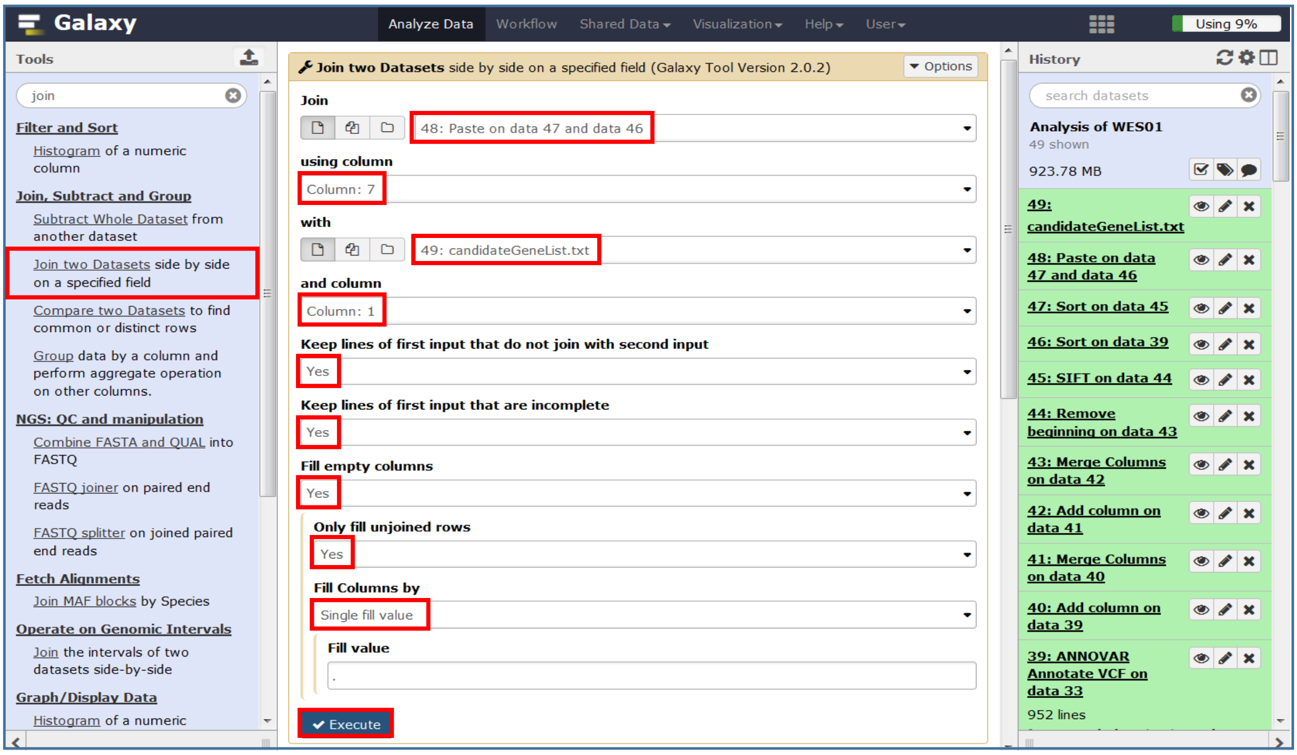
Change data type

New Type: tabular

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Save

17. In **Tool Pane**: Go to **Join, Subtract and Group** > Join two Datasets
Using the options below will add a new column to the annotated VCF file to show if the variants you have called and annotated are in the list of candidate genes or not.


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9%
Tools
join
Filter and Sort
Histogram of a numeric column
Join, Subtract and Group
Subtract Whole Dataset from another dataset
Join two Datasets side by side on a specified field
Compare two Datasets to find common or distinct rows
Group data by a column and perform aggregate operation on other columns.
NGS: QC and manipulation
Combine FASTA and QUAL into FASTQ
FASTQ joiner on paired end reads
FASTQ splitter on joined paired end reads
Fetch Alignments
Join MAF blocks by Species
Operate on Genomic Intervals
Join the intervals of two datasets side-by-side
Graph/Display Data
Histogram of a numeric
Options
History
search datasets
Analysis of WES01
49 shown
923.78 MB
49: candidateGeneList.txt
48: Paste on data 47 and data 46
47: Sort on data 45
46: Sort on data 39
45: SIFT on data 44
44: Remove beginning on data 43
43: Merge Columns on data 42
42: Add column on data 41
41: Merge Columns on data 40
40: Add column on data 39
39: ANNOVAR Annotate VCF on data 33
952 lines

Join two Datasets side by side on a specified field (Galaxy Tool Version 2.0.2)

Join: 48: Paste on data 47 and data 46

using column: Column: 7

with: 49: candidateGeneList.txt

and column: Column: 1

Keep lines of first input that do not join with second input: Yes

Keep lines of first input that are incomplete: Yes

Fill empty columns: Yes

Only fill unjoined rows: Yes

Fill Columns by: Single fill value

Fill value: .

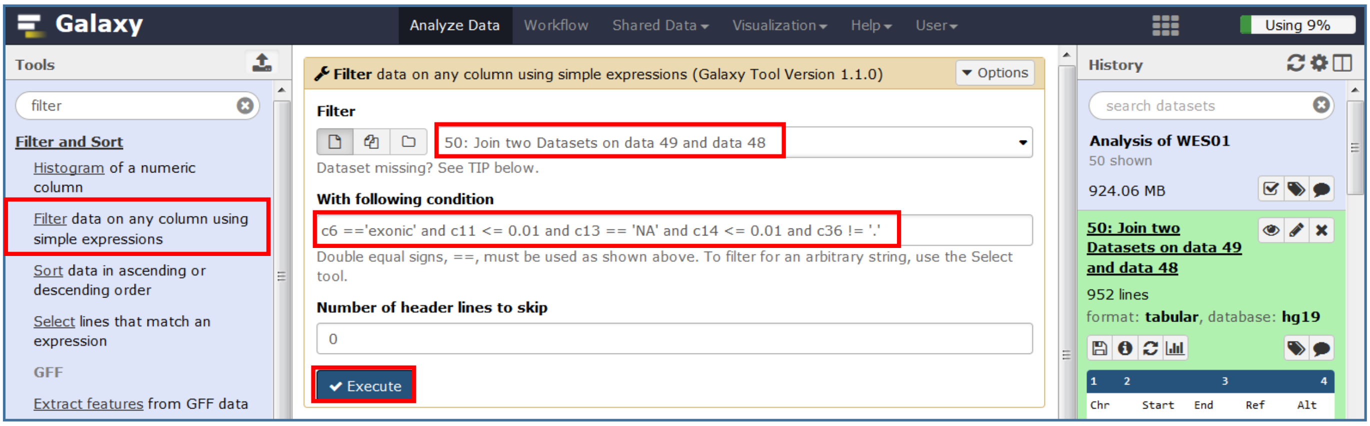
✔ Execute

Filter variants

Having annotated the variants we can now apply filters to identify a shortlist of potentially pathogenic variants.

1. In **Tool Pane**: Go to **Filter and Sort** > Filter

There are many ways to filter the data, one of the most common strategies is to select coding variants (c6 == ‘exonic’) that are either rare or absent from databases of known variation such as 1000 genomes (c11 <= 0.01), the non flagged version of dbSNP (c13 == ‘NA’), or Exome Sequencing Project (c14 <= 0.01) and are located in a candidate gene (c36 != ‘.’).


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9%
Tools
filter
Filter and Sort
Histogram of a numeric column
Filter data on any column using simple expressions
Sort data in ascending or descending order
Select lines that match an expression
GFF
Extract features from GFF data
Options
History
search datasets
Execute

Filter data on any column using simple expressions (Galaxy Tool Version 1.1.0)

Filter

50: Join two Datasets on data 49 and data 48

Dataset missing? See TIP below.

With following condition

c6 =='exonic' and c11 <= 0.01 and c13 == 'NA' and c14 <= 0.01 and c36 != '.'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0

Analysis of WES01
50 shown
924.06 MB

50: Join two Datasets on data 49 and data 48
952 lines
format: tabular, database: hg19

1	2		3	4
Chr	Start	End	Ref	Alt

2. Have a look at the short list of variants and decide which variant or variants may contribute to the disease in your patient.

Q3. Which variant is the most likely to cause the patients disease?

Q4. For the causal variant you have chosen, use the VCF file from Unified Genotyper and information from the variant calling lecture to populate the table below of quality control parameters and categorise the value of each as either good, intermediate or poor.

Quality control parameter	**Value**	**Category**
Depth		
Allelic depth (AD)		
Strand bias (FS)		
Variant confidence (QUAL)		
Quality by depth (QD)		
Mapping quality (MQ)		
Mapping quality bias (MQRankSum)		
Base quality bias (BaseQRankSum)		
Tail bias (ReadPosRankSum)		
Haplotype score		

Q5. Is your variant present in ClinVar and OMIM and if it is does the expected phenotype match your patient?

Q6. Has your variant been published and if so what is the citation?

Q7. Look up the function of your gene. Eg. GeneCards. What is known about it?

Q8. If the variant is real, what disease do you think the patient may have?

Remember that your analysis is relative to the human reference genome build hg19/GRCh37

3. Our initial filtering strategy was fairly stringent which is a good way to minimise incidental findings. However, it is not apparent how many variants remained after each filter and the process would need repeating with fewer filters or less stringent criteria if no variants were left.

Q9. Repeat the filtering process to determine the number of variants that remain after each of the filters in the table is applied.

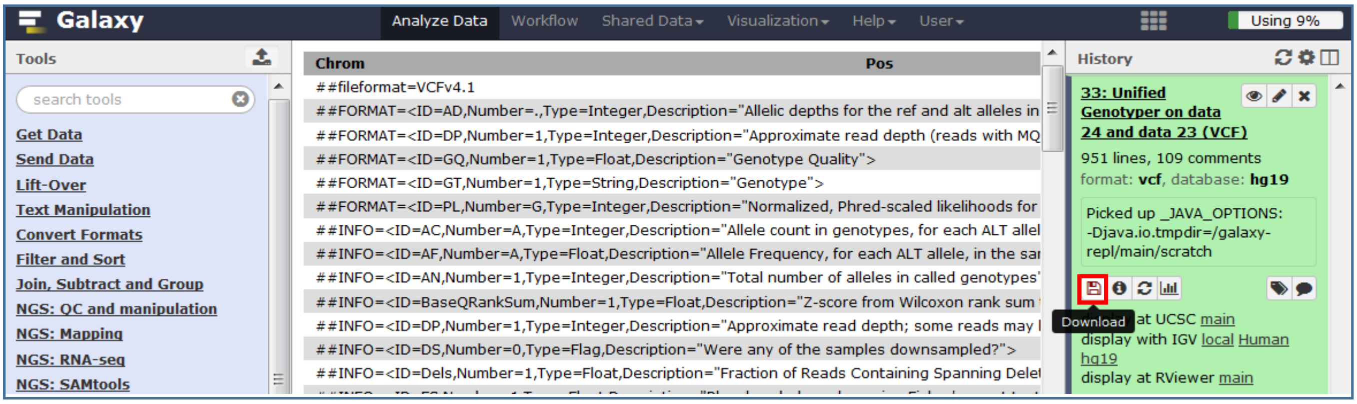
Filtering criteria	No. variants
None	
Exonic	
Exonic and absent from public databases (dbSNP non flagged, 1000 genomes, ESP)	
Exonic, absent from public databases and located in a candidate gene	

Alternatively, we could identify variants that are flagged by dbSNP as clinically associated or are present in OMIM.

Q10. How many variants are flagged by dbSNP as clinically associated?
Q11. How many variants are present in OMIM?

The online version of ANNOVAR offers some additional annotation. Download your VCF file and use wANNOVAR to perform annotation.

4. Click the download icon and save the Unified Genotyper VCF file to your computer



```
Chrom                                                    Pos
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allel
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the sa
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes"
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Dele
```

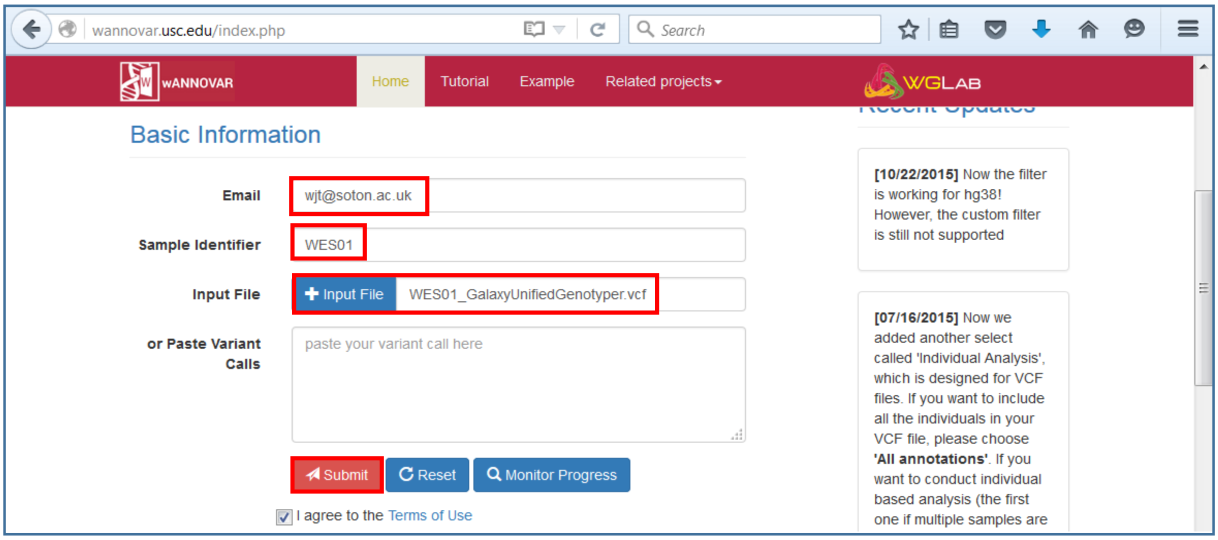
33: Unified Genotyper on data 24 and data 23 (VCF)

951 lines, 109 comments
format: **vcf**, database: **hg19**

Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-repl/main/scratch

display at UCSC main
display with IGV local Human hg19
display at RViewer main

5. Upload the VCF file to wANNOVAR and click Submit



Basic Information

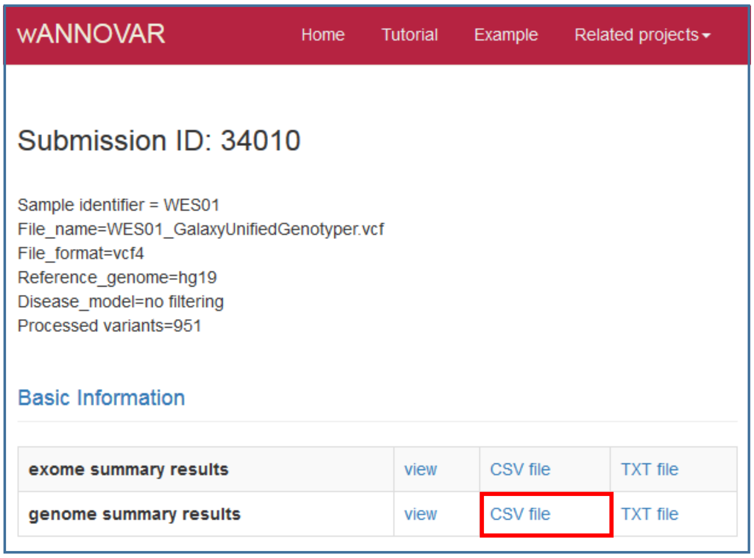
Email: wjt@soton.ac.uk
Sample Identifier: WES01
Input File: + Input File WES01_GalaxyUnifiedGenotyper.vcf
or Paste Variant Calls: paste your variant call here

☑ I agree to the Terms of Use

[10/22/2015] Now the filter is working for hg38! However, the custom filter is still not supported

[07/16/2015] Now we added another select called 'Individual Analysis', which is designed for VCF files. If you want to include all the individuals in your VCF file, please choose **'All annotations'**. If you want to conduct individual based analysis (the first one if multiple samples are

6. Download the genome summary results



Submission ID: 34010

Sample identifier = WES01
File_name=WES01_GalaxyUnifiedGenotyper.vcf
File_format=vcf4
Reference_genome=hg19
Disease_model=no filtering
Processed variants=951

Basic Information

exome summary results	view	CSV file	TXT file
genome summary results	view	CSV file	TXT file

7. Open the wANNOVAR results in excel and compare the annotation for your variant of interest.

Q12. For the causal variant you have chosen, use the result from wANNOVAR to populate the table below of pathogenicity predictors and select a category for each value.

Pathogenicity predictor	Value	Category
SIFT		Deleterious/Tolerated/Missing
Polyphen-2 HDIV		Probably damaging/Possible damaging/Benign/Missing
Polyphen-2 HVAR		Probably damaging/Possible damaging/Benign/Missing
LRT score		Deleterious/Neutral/Unknown/Missing
Mutation Taster		Disease_causing_automatic/Disease_causing/Polymorphism/ Polymorphism_automatic/Missing
Mutation Assessor		Predicted functional (high, medium)/Predicted non-functional (low, neutral)/Missing
FATHMM score		Damaging/Tolerated/Missing
RadialSVM		Deleterious/Tolerated/Missing
LR score		Deleterious/Tolerated/Missing
VEST3 score		None, likelihood of functional effect increase with score
CADD raw		None, likelihood of damaging effect increase with score
CADD phred		None, ranked and phred scaled CADD score
GERP++RS		None, conservation increases with score (range -12.3 to 6.17). Scores >2, high sensitivity for truly constrained sites
phyloP46way placental		None, the larger the score, the more conserved the site
phyloP100way vertebrate		None, the larger the score, the more conserved the site
SiPhy 29way logOdds		The larger the score the more conserved the site

To help shortlist variants further, hard filters could be applied to selected predictors of pathogenicity (eg SIFT score > 0.95, LRT score > 0.95, PolyyPhen2 > 0.85, Mutation Taster predicts the variant to be disease causing automatic or disease causing) or the variants could be sorted by pathogenicity score and followed up in order from most to least deleterious.

NOTE: If you're running out of time to finish the practical you can read the section below in your own time and skip to the last section on how to create a workflow.

SIFT

SIFT score for non-synonymous variants are based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences.

Polyphen-2

Polyphen-2 predicts the impact of an amino acid substitution on the structure and function of a protein. Predictions are based on features that characterise the substitution including the sequence (does the substitution occur within an annotated site in the protein eg active or binding site, non-globular region such as trans-membrane etc), phylogenetics (how often does the substitution occur in a family of related proteins) and structural information (could the substitution affect the proteins 3D structure eg hydrophobic core, electrostatic interactions, interactions with ligands or other important features). These features are assessed by a probabilistic classifier.

Two versions of Polyphen are available, HDIV and HVAR, that used different datasets to train and test the prediction models. HDIV should be used for evaluating rare alleles involved in complex phenotypes where both disease causing and mildly deleterious alleles are treated as damaging. The HDIV dataset consists of all damaging alleles in the UniProtKb database that effect molecular function and cause human Mendelian diseases versus differences between human proteins and their closely related mammalian homologues which are assumned to be non-damaging.

HVAR should be used for diagnostics of Mendelian diseases which requires distinction between mutations with diagnostic effects from all remaining human variation, including a wealth of mildly deleterious alleles. The HVAR dataset used to train and test Polyphen consisted of all human disease causing mutations from UniProtKb together with common human non-synonymous SNPs (MAF>1%) without annotated involvement in disease which were treated as non-damaging.

LRT score

Likelihood ratio test (LRT) for prediction of significantly conserved amino acid positions within the human proteome. Mutations are estimated as 'deleterious' that disrupt highly conserved aminoacid residues, 'neutral' or 'unknown'.

Mutation Taster

MutationTaster uses a Bayes classifier to predict the disease potential of an alteration. For this prediction, the frequencies of all single features for known disease mutations/polymorphisms were studied in a large training set composed of >390,000 known disease mutations from HGMD Professional and >6,800,000 harmless SNPs and Indel polymorphisms from the 1000 Genomes Project.

Mutation Assessor

Mutation assessor scores estimate functional impact based on evolutionary conservation of the affected amino acids in protein homologs. The method has been validated on a large set (60k) of disease associated (OMIM) and polymorphic variants.

FATHMM score

Functional Analysis through Hidden Markov Models. Predicts the functional effects of protein missense mutations by combining sequence conservation within hidden Markov models (HMMs), representing the alignment of homologous sequences and conserved protein domains, with "pathogenicity weights", representing the overall tolerance of the protein/domain to mutations.

RadialSVM

Support vector machine (SVM) based ensemble prediction score, which incorporates 10 scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from -2 to 3 in dbNSFP.

LR score

LR scores evaluate the deleteriousness of missense mutations by using logistic regression to integrate information from:

i) function prediction scores (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, PANTHER, PhD-SNP, SNAP, SNPs&GO, and MutPred)

ii) conservation scores (GERP++, SiPhy and PhyloP)

iii) ensemble scores (CADD, PON-P, KGGSeq and CONDEL)
iv) maximum minor allele frequency

VEST3 score

Variant Effect Scoring Tool (VEST) is a supervised machine learning-based classifier for prioritisation of rare missense variants with likely involvement in human disease. The VEST classifier training set comprised ~ 45,000 disease mutations from the latest Human Gene Mutation Database release and another ~45,000 high frequency (allele frequency >1%) putatively neutral missense variants from the Exome Sequencing Project. VEST estimates variant score p-values against a null distribution of VEST scores for neutral variants not included in the VEST training set. These p-values can be aggregated at the gene level across multiple disease exomes to rank genes for probable disease involvement.

CADD raw

Combined Annotation Dependent Depletion (CADD) is a framework that integrates multiple annotations into one metric by contrasting variants that survived natural selection (differences between 1000 Genomes and the Ensembl Compara inferred human-chimpanzee ancestral genome) with simulated mutations. Raw CADD scores come straight from the model and have no absolute unit of meaning and are incomparable across distinct annotation combinations, training sets, or model parameters. However, raw values have relative meaning with larger scores indicating that a variant is more likely to have a damaging effect.

CADD phred

Phred CADD scores are ranked scores based on the whole genome CADD raw scores. For example, variants at the 10th-% of raw CADD scores are assigned to CADD-10, top 1% to CADD-20, top 0.1% to CADD-30, etc. The results of this transformation are Phred-like scaled CADD scores.

Genomic Evolutionary Rate

Profiling Rejected Substitution scores (GERP RS). GERP identifies constrained elements in multiple alignments by quantifying substitution deficits for SNPs. Scores range from -12.3 to 6.17, with 6.17 being the most conserved. Scores greater than 2 provide a high sensitivity while still strongly enriched for truly constrained sites.

PhyloP46way placental

Phylogenetic conservation score based on a multiple alignment of 46 vertebrate genomes (10 primates, 33 placental mammals).

PhyloP100way vertebrate

Phylogenetic conservation score based on a multiple alignment of 100 vertebrate genomes.

SiPhy 29way logOdds

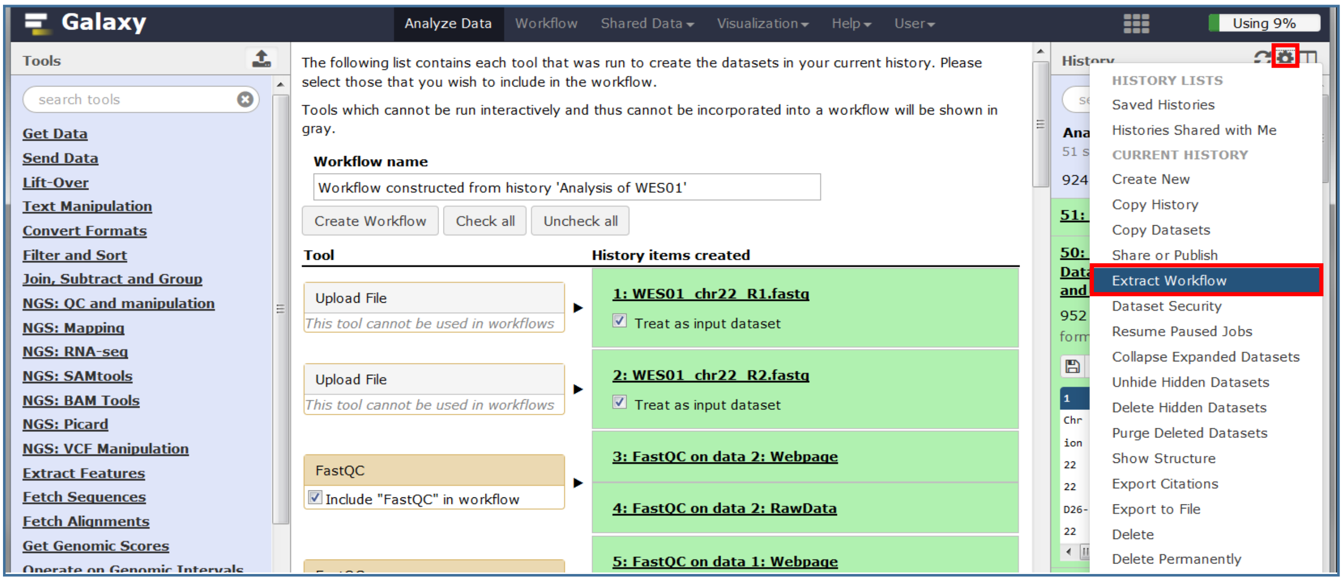
The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.

Creating workflows

Reproducing an experiment on the same data or different datasets and comparing the results is a key feature of research that requires keeping a meticulous set of instructions. In many cases, multiple replicates are required and the best approach is an automated analysis that is consistent and less error prone. In Galaxy, the history acts as a detailed list of experimental instructions that can be reproduced and automated by creating a ‘Workflow’ or bioinformatic pipeline.

1. Before making your workflow delete any unwanted steps from the history so that it only includes the steps needed to make the final result.

2. Go to History panel, click the cog icon and select 'Extract workflow'.



The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

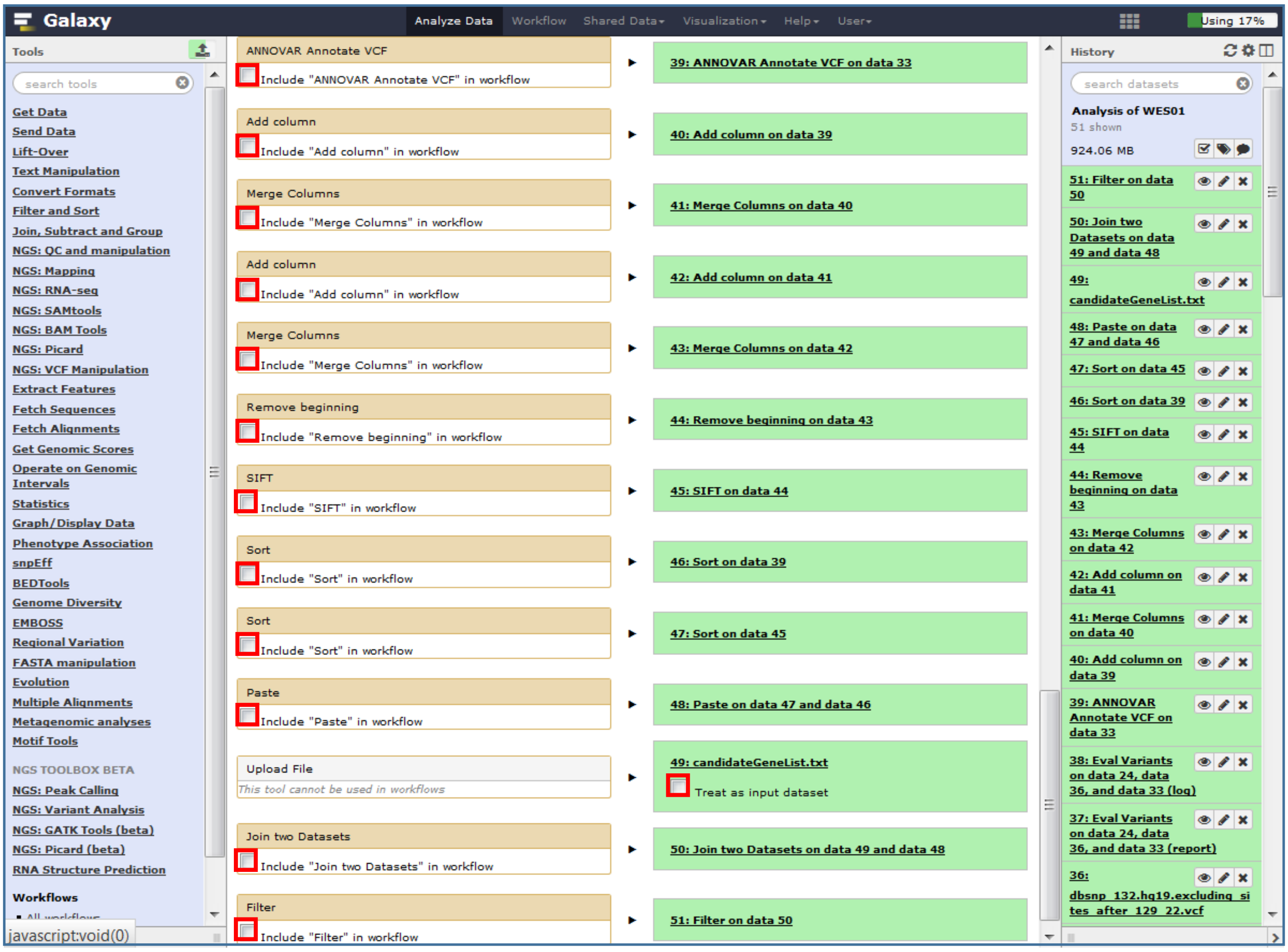
Workflow name

Workflow constructed from history 'Analysis of WES01'

Tool	History items created
Upload File *This tool cannot be used in workflows*	1: WES01_chr22_R1.fastq ☑ Treat as input dataset
Upload File *This tool cannot be used in workflows*	2: WES01_chr22_R2.fastq ☑ Treat as input dataset
FastQC ☑ Include "FastQC" in workflow	3: FastQC on data 2: Webpage 4: FastQC on data 2: RawData
	5: FastQC on data 1: Webpage

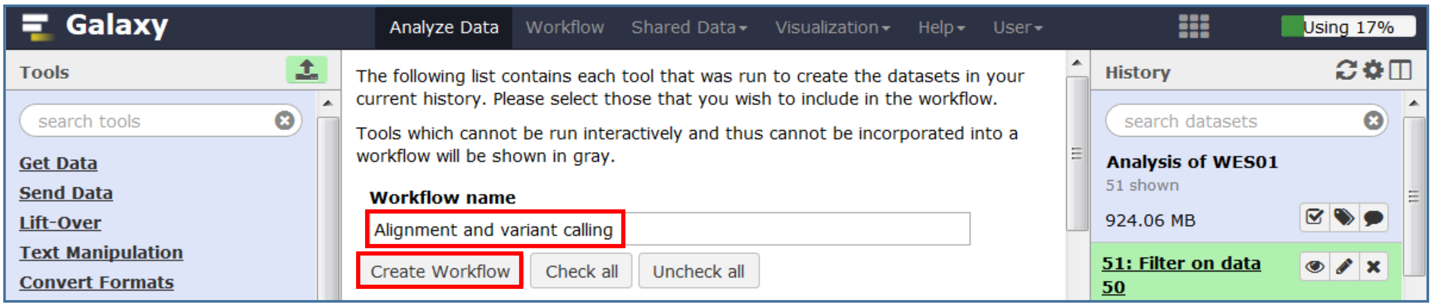
We're going to create two workflows, one for alignment and variant calling and the other for annotation and filtering because the VCF reference genome has to be manually converted from hg_g1k_v37 to hg19.

3. We will create the alignment and variant calling workflow first. Scroll down and uncheck all steps from ANNOVAR onwards.



ANNOVAR Annotate VCF □ Include "ANNOVAR Annotate VCF" in workflow	39: ANNOVAR Annotate VCF on data 33
Add column □ Include "Add column" in workflow	40: Add column on data 39
Merge Columns □ Include "Merge Columns" in workflow	41: Merge Columns on data 40
Add column □ Include "Add column" in workflow	42: Add column on data 41
Merge Columns □ Include "Merge Columns" in workflow	43: Merge Columns on data 42
Remove beginning □ Include "Remove beginning" in workflow	44: Remove beginning on data 43
SIFT □ Include "SIFT" in workflow	45: SIFT on data 44
Sort □ Include "Sort" in workflow	46: Sort on data 39
Sort □ Include "Sort" in workflow	47: Sort on data 45
Paste □ Include "Paste" in workflow	48: Paste on data 47 and data 46
Upload File *This tool cannot be used in workflows*	49: candidateGeneList.txt □ Treat as input dataset
Join two Datasets □ Include "Join two Datasets" in workflow	50: Join two Datasets on data 49 and data 48
Filter □ Include "Filter" in workflow	51: Filter on data 50

4. Rename your workflow as ‘Alignment and variant calling’ and then click ‘Create Workflow’.



The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name

Alignment and variant calling

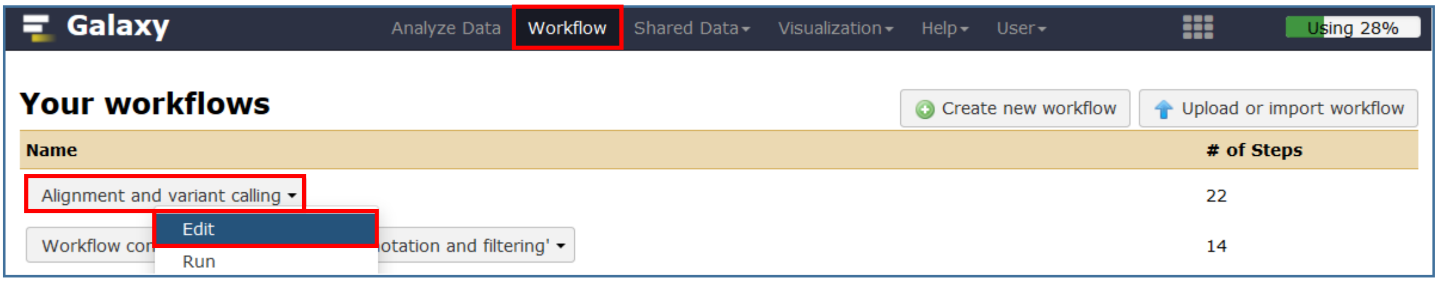
Analysis of WES01

51 shown

924.06 MB

51: Filter on data 50

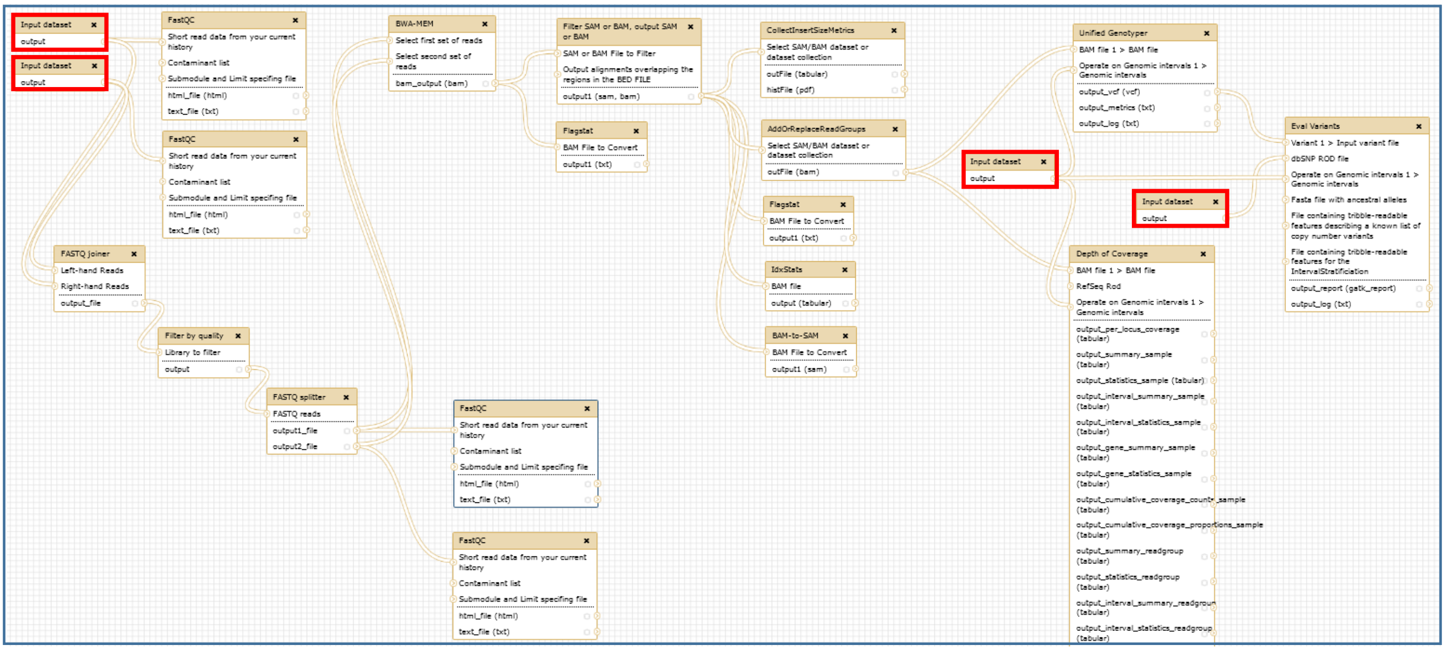
5. Edit the workflow by clicking on the ‘Workflow’ tab, then click on the workflow and select Edit.



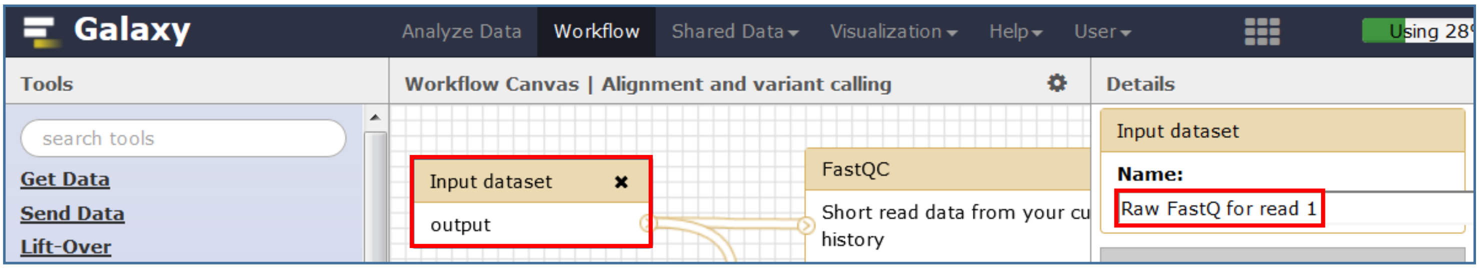
Your workflows

Name	# of Steps
Alignment and variant calling	22
Workflow con… …otation and filtering'	14

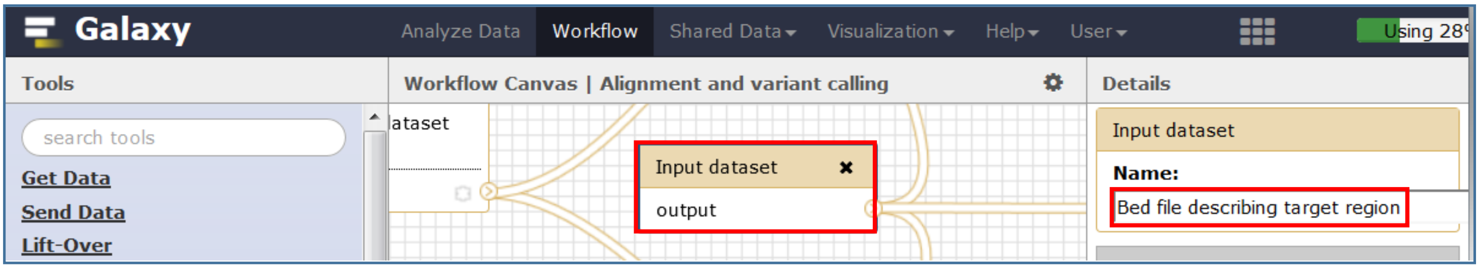
This will bring up the workflow canvas showing a flow diagram of your work. The boxes represent the steps and they are joined together to represent the inputs and outputs of each step. You can drag the boxes around to organise them in the clearest way. If you click on a box, its details will appear in the right hand panel. Here you can add your own annotation or notes for each step. For the Input datasets (highlighted in red below), add a name to describe the input file which will make the workflow much easier to use.


Input dataset
output
Input dataset
output
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
FASTQ joiner
Left-hand Reads
Right-hand Reads
output_file
Filter by quality
Library to filter
output
FASTQ splitter
FASTQ reads
output1_file
output2_file
BWA-MEM
Select first set of reads
Select second set of reads
bam_output (bam)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Filter SAM or BAM, output SAM or BAM
SAM or BAM File to Filter
Output alignments overlapping the regions in the BED FILE
output1 (sam, bam)
Flagstat
BAM File to Convert
output1 (txt)
CollectInsertSizeMetrics
Select SAM/BAM dataset or dataset collection
outFile (tabular)
histFile (pdf)
AddOrReplaceReadGroups
Select SAM/BAM dataset or dataset collection
outFile (bam)
Flagstat
BAM File to Convert
output1 (txt)
IdxStats
BAM file
output (tabular)
BAM-to-SAM
BAM File to Convert
output1 (sam)
Input dataset
output
Unified Genotyper
BAM file 1 > BAM file
Operate on Genomic intervals 1 > Genomic intervals
output_vcf (vcf)
output_metrics (txt)
output_log (txt)
Input dataset
output
Depth of Coverage
BAM file 1 > BAM file
RefSeq Rod
Operate on Genomic intervals 1 > Genomic intervals
output_per_locus_coverage (tabular)
output_summary_sample (tabular)
output_statistics_sample (tabular)
output_interval_summary_sample (tabular)
output_interval_statistics_sample (tabular)
output_gene_summary_sample (tabular)
output_gene_statistics_sample (tabular)
output_cumulative_coverage_counts_sample (tabular)
output_cumulative_coverage_proportions_sample (tabular)
output_summary_readgroup (tabular)
output_statistics_readgroup (tabular)
output_interval_summary_readgroup (tabular)
output_interval_statistics_readgroup (tabular)
Eval Variants
Variant 1 > Input variant file
dbSNP ROD file
Operate on Genomic intervals 1 > Genomic intervals
Fasta file with ancestral alleles
File containing tribble-readable features describing a known list of copy number variants
File containing tribble-readable features for the IntervalStratificiation
output_report (gatk_report)
output_log (txt)

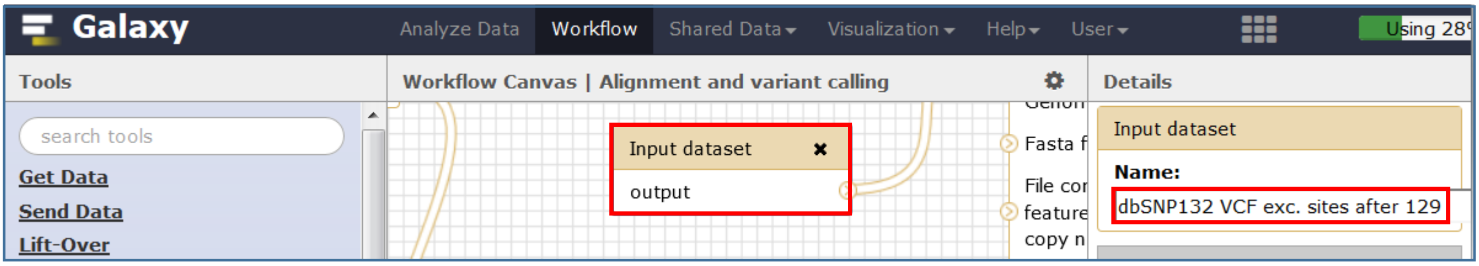
6. Click on the box for each Input dataset and add a name to describe the input file.



Name: Raw FastQ for read 1

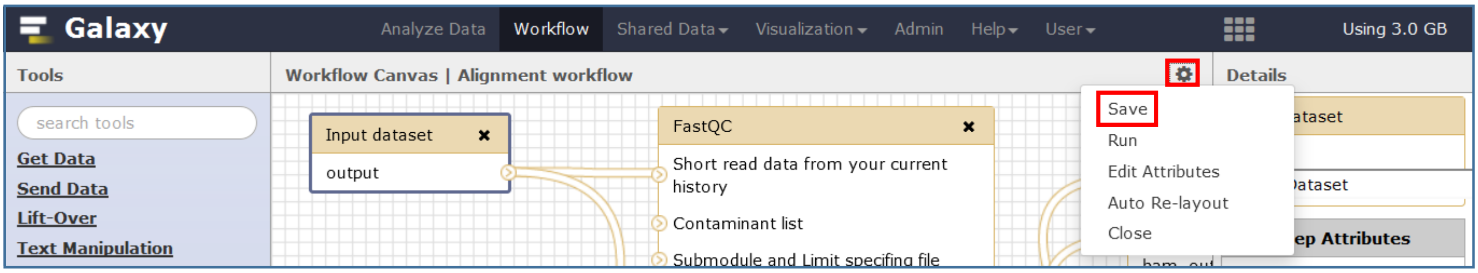


Name: Bed file describing target region



Name: dbSNP132 VCF exc. sites after 129

7. Once you have made all your changes save the workflow by clicking the cog icon and selecting 'Save'.

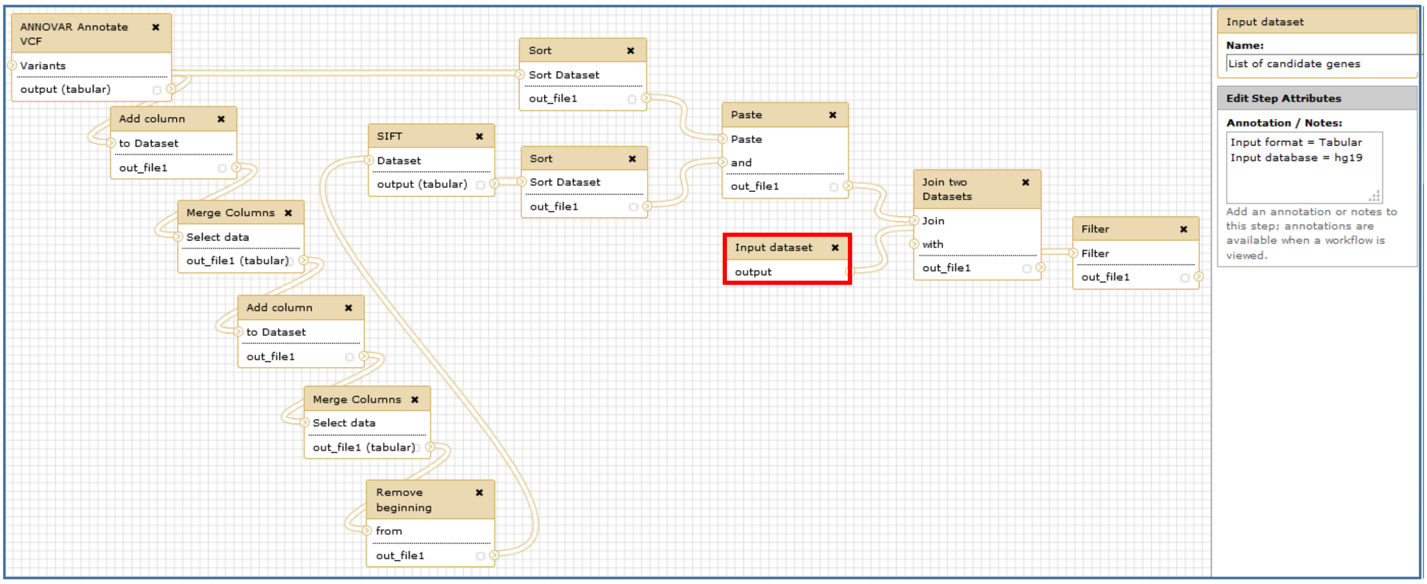


Now close the Workflow editor by clicking the cog icon and selecting 'Close'.

8. Repeat this process to create a workflow for annotation and variant calling. This will involve the following steps:

- Go back to your history for Analysis of WES01, click the cog icon and select 'Extract Workflow'
- Rename the Workflow 'Annotation and filtering'
- Uncheck all the steps and then click the check boxes for all steps from 'ANNOVAR Annotate VCF' onwards
- Click 'Create Workflow'
- Edit the workflow by selecting the workflow tab then clicking the 'Annotation and filtering' workflow and selecting 'Edit'
- Organise the workflow

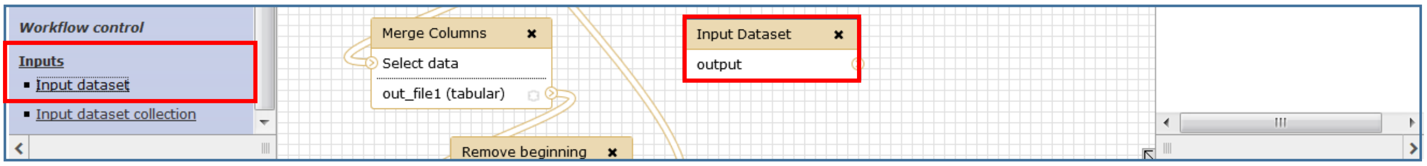
9. The Annotation and filtering workflow should look like this. Give the Input dataset a name to describe the file and its format.



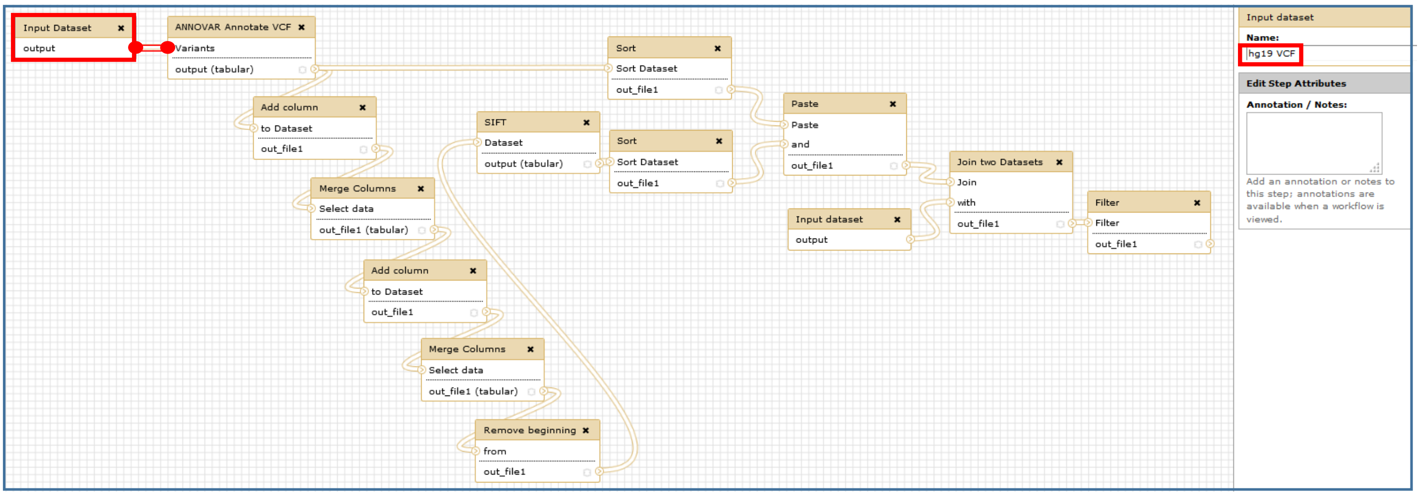
Name: List of candidate genes

Annotation / Notes:
Input format = Tabular
Input database = hg19

10. Now add an Input dataset by clicking the 'Input dataset' link at the bottom left of the tool pane.

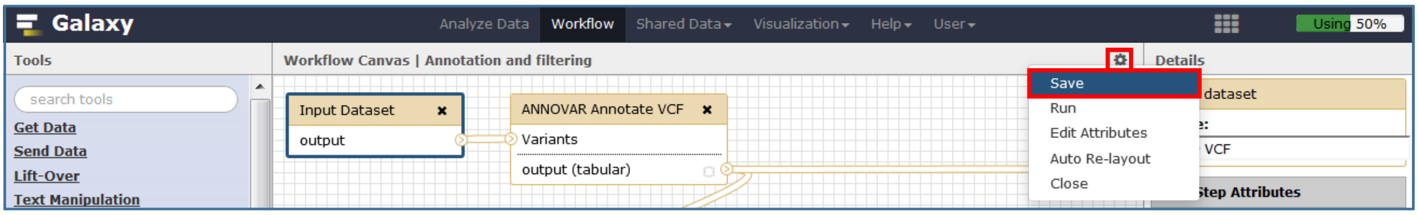


11. Link the new Input Dataset with ANNOVAR by clicking the circle and dragging the noodle to the ANNOVAR step. Add a name to describe the Input Dataset.



Name: hg19 VCF

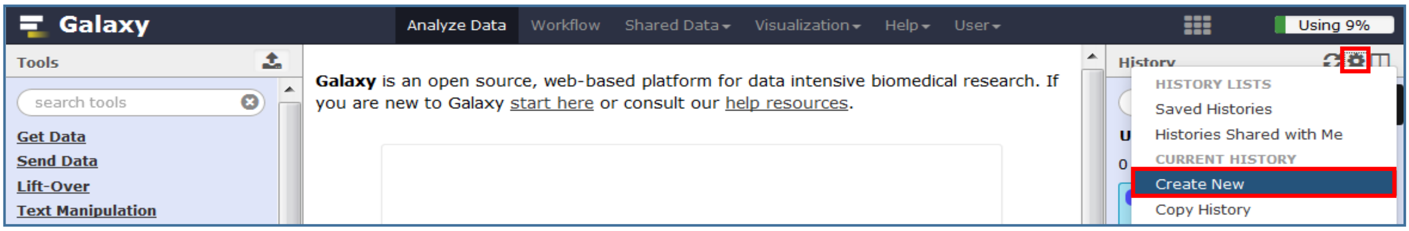
12. Save and Close the annotation and filtering workflow.



Running workflows

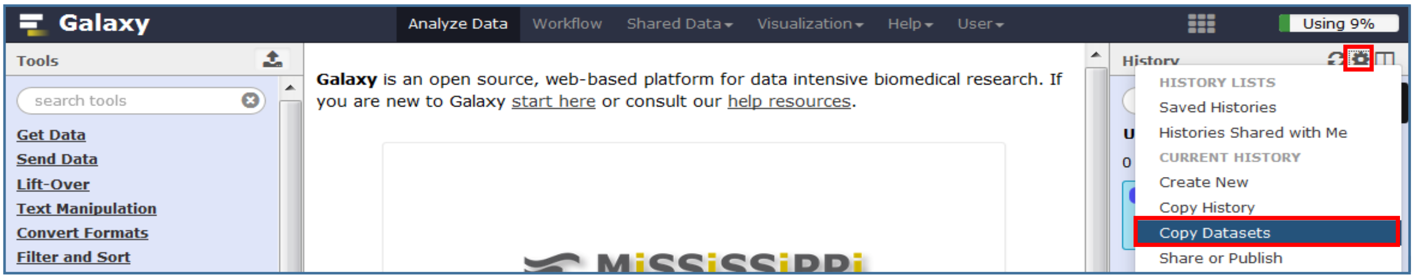
Now that you have made a workflow/bioinformatic pipeline you can apply the whole process to a new dataset with just a few clicks. Let us apply your workflow to the original fastq files.

1. In the history pane click the cog icon and select ‘Create New’ to make a new history.



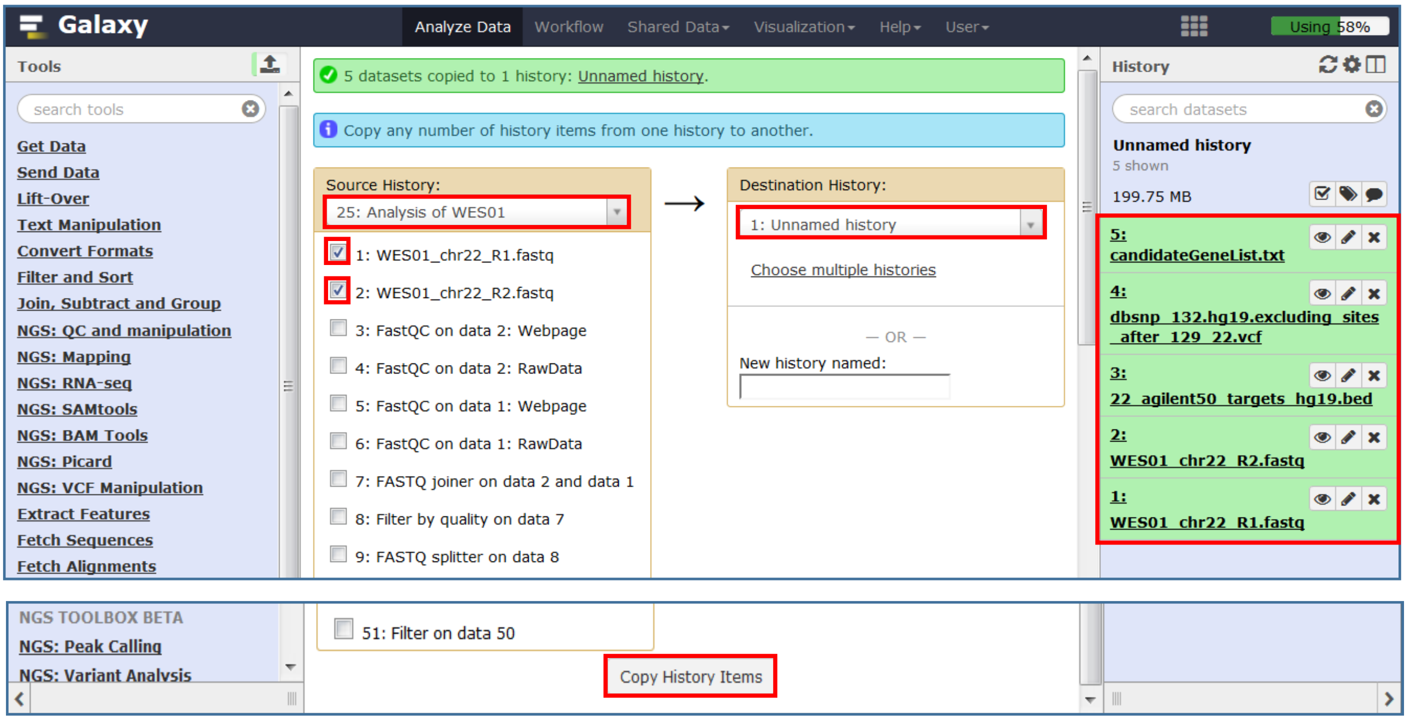
Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

2. Before running the workflows, we need to put the raw data files into the new history. You can do this by uploading or copying the data from an existing dataset. We will copy the input files by clicking the cog icon and selecting ‘Copy Datasets’



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources.

3. Select ‘Analysis of WES01’ as the source history and the new ‘Unnamed history’ as the destination. Check the boxes for the 5 input files that we need (raw fastqs: WES01_chr22_R1.fastq, WES01_chr22_R2.fastq, target bed file: 22_agilent50_targets_hg19.bed, dbSNP rod file: dbsnp_132.hg19.excluding_sites_after_129_22.vcf, list of candidate genes: candidateGeneList.txt). Scroll to the bottom of the page and click ‘Copy History Items’.



5 datasets copied to 1 history: Unnamed history.

Copy any number of history items from one history to another.

Source History: 25: Analysis of WES01

☑ 1: WES01_chr22_R1.fastq
☑ 2: WES01_chr22_R2.fastq
□ 3: FastQC on data 2: Webpage
□ 4: FastQC on data 2: RawData
□ 5: FastQC on data 1: Webpage
□ 6: FastQC on data 1: RawData
□ 7: FASTQ joiner on data 2 and data 1
□ 8: Filter by quality on data 7
□ 9: FASTQ splitter on data 8
□ 51: Filter on data 50

Destination History: 1: Unnamed history

Choose multiple histories

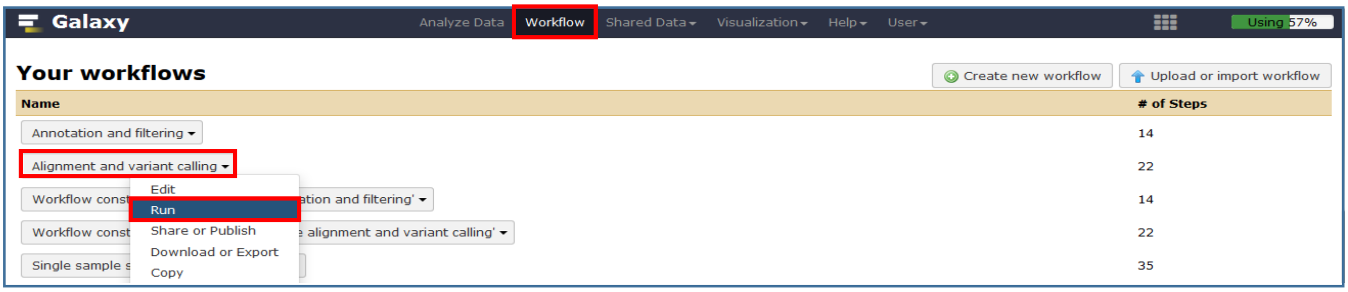
— OR —

New history named:

Unnamed history
5 shown
199.75 MB

5: candidateGeneList.txt
4: dbsnp_132.hg19.excluding_sites_after_129_22.vcf
3: 22_agilent50_targets_hg19.bed
2: WES01_chr22_R2.fastq
1: WES01_chr22_R1.fastq

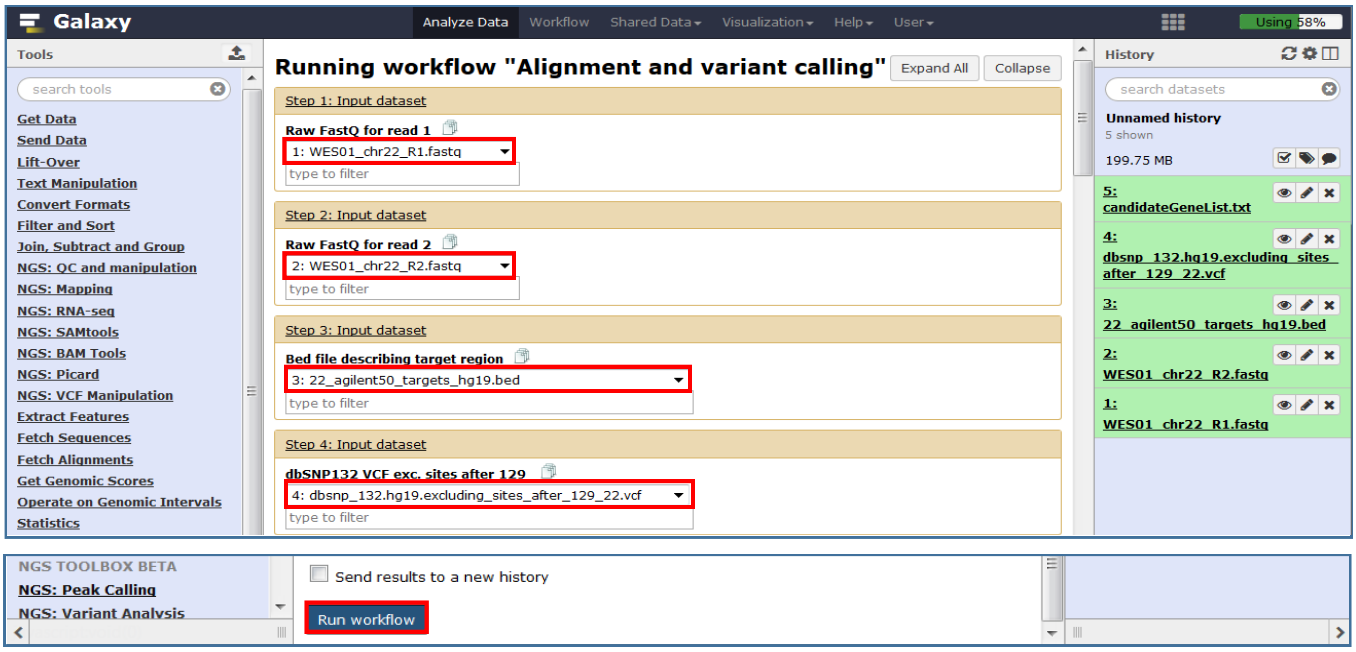
4. Click the workflow tab then click your 'Alignment and variant calling' workflow and select run from the drop down menu.



Your workflows

Name	# of Steps
Annotation and filtering	14
Alignment and variant calling	22
Workflow const… ation and filtering'	14
Workflow const… alignment and variant calling'	22
Single sample s…	35

5. Now choose the relevant input files and click 'Run workflow' to analyse the data.



Running workflow "Alignment and variant calling"

Step 1: Input dataset
Raw FastQ for read 1: 1: WES01_chr22_R1.fastq

Step 2: Input dataset
Raw FastQ for read 2: 2: WES01_chr22_R2.fastq

Step 3: Input dataset
Bed file describing target region: 3: 22_agilent50_targets_hg19.bed

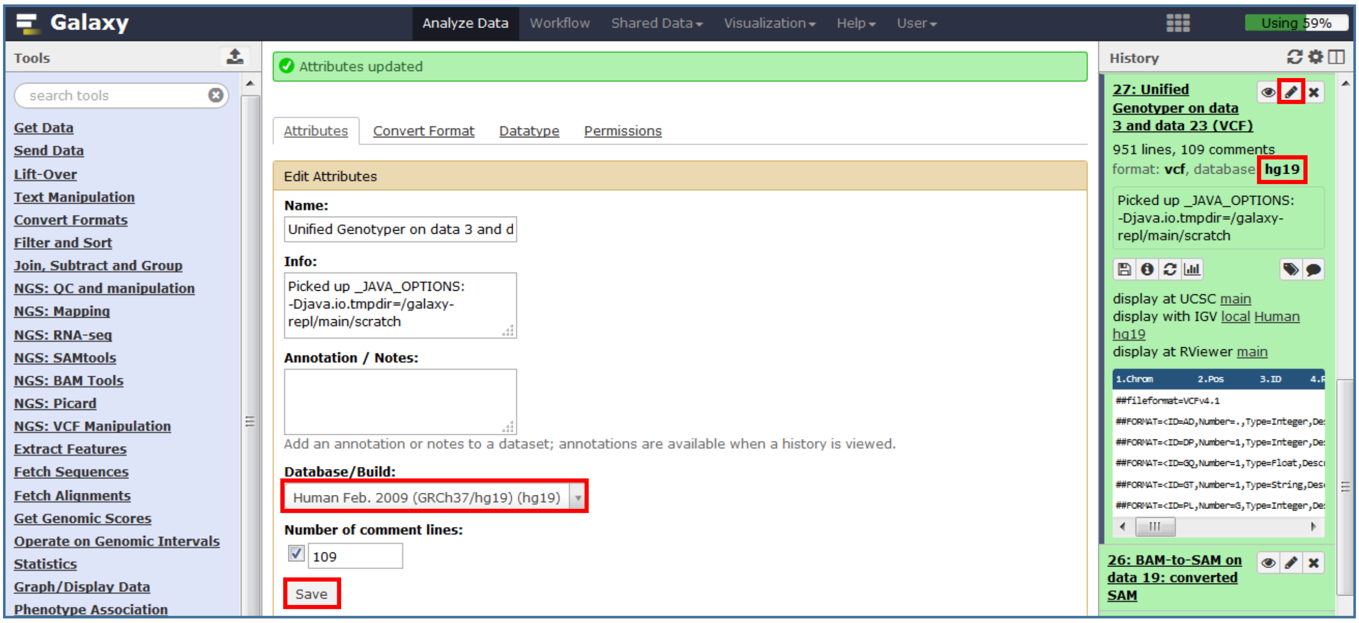
Step 4: Input dataset
dbSNP132 VCF exc. sites after 129: 4: dbsnp_132.hg19.excluding_sites_after_129_22.vcf

□ Send results to a new history

Unnamed history
5 shown
199.75 MB

5: candidateGeneList.txt
4: dbsnp_132.hg19.excluding_sites_after_129_22.vcf
3: 22_agilent50_targets_hg19.bed
2: WES01_chr22_R2.fastq
1: WES01_chr22_R1.fastq

6. When Unified Genotyper has finished, use edit attributes to change the VCF database from hg_g1k_v37 to hg19.



Attributes updated

Edit Attributes

Name: Unified Genotyper on data 3 and d

Info: Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-repl/main/scratch

Annotation / Notes:

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build: Human Feb. 2009 (GRCh37/hg19) (hg19)

Number of comment lines: ☑ 109

27: Unified Genotyper on data 3 and data 23 (VCF)
951 lines, 109 comments
format: vcf, database hg19

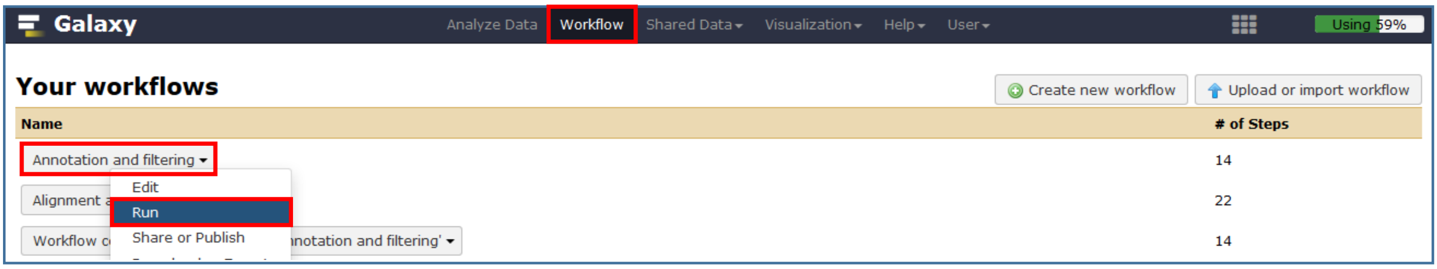
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-repl/main/scratch

display at UCSC main
display with IGV local Human hg19
display at RViewer main

```
1.Chrom          2.Pos        3.ID       4.R
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,De
##FORMAT=<ID=DP,Number=1,Type=Integer,De
##FORMAT=<ID=GQ,Number=1,Type=Float,Desc
##FORMAT=<ID=GT,Number=1,Type=String,Des
##FORMAT=<ID=PL,Number=G,Type=Integer,De
```

26: BAM-to-SAM on data 19: converted SAM

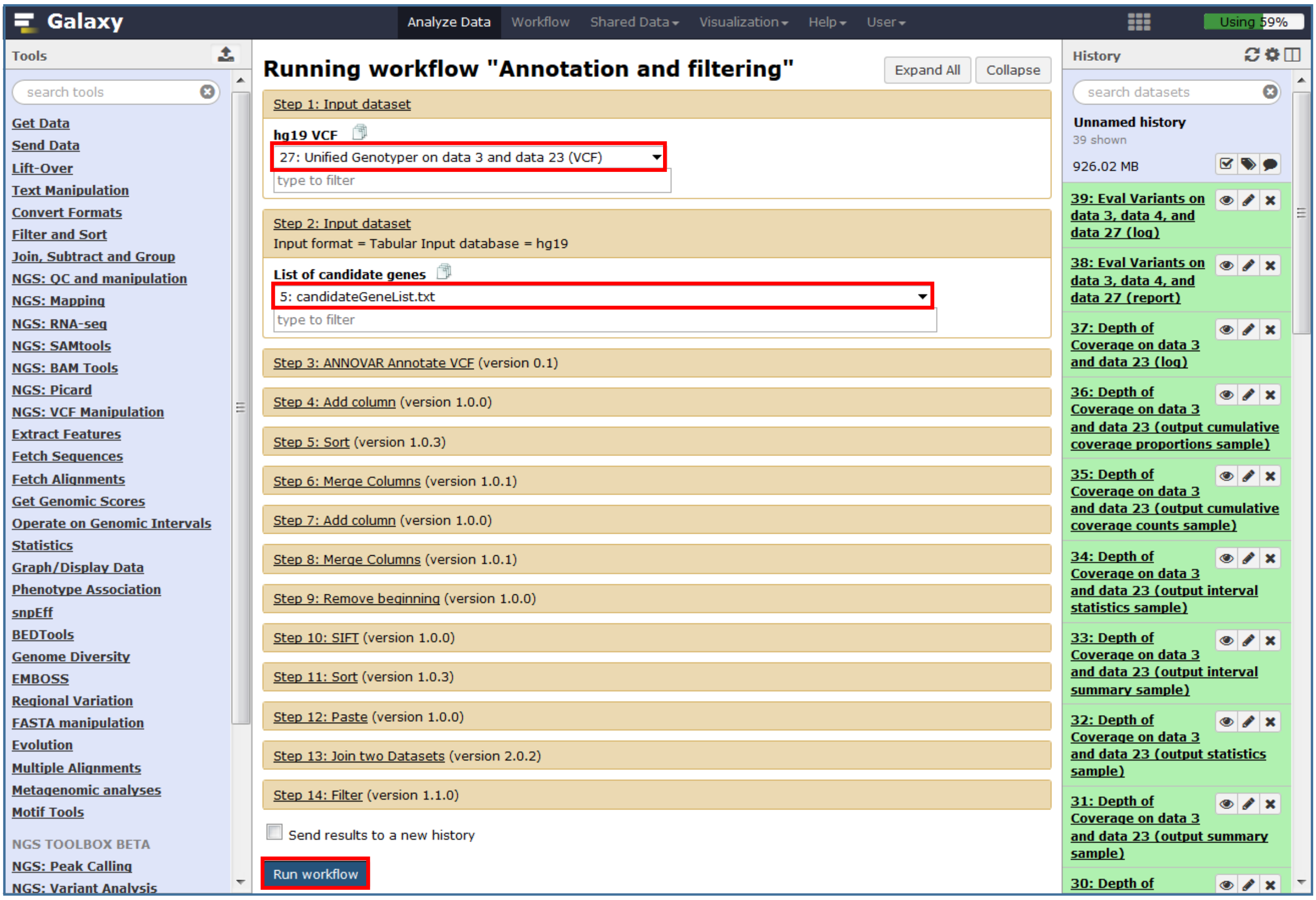
7. When all of the Alignment and variant calling steps have finished, click the workflow tab then click your 'Annotation and filtering' workflow and select run from the drop down menu.



Your workflows

Name	# of Steps
Annotation and filtering	14
Alignment a	22
Workflow c nnotation and filtering'	14

8. Select the appropriate input files and click 'Run workflow' to analyse the data.



Running workflow "Annotation and filtering"

Step 1: Input dataset

hg19 VCF
27: Unified Genotyper on data 3 and data 23 (VCF)
type to filter

Step 2: Input dataset
Input format = Tabular Input database = hg19

List of candidate genes
5: candidateGeneList.txt
type to filter

Step 3: ANNOVAR Annotate VCF (version 0.1)

Step 4: Add column (version 1.0.0)

Step 5: Sort (version 1.0.3)

Step 6: Merge Columns (version 1.0.1)

Step 7: Add column (version 1.0.0)

Step 8: Merge Columns (version 1.0.1)

Step 9: Remove beginning (version 1.0.0)

Step 10: SIFT (version 1.0.0)

Step 11: Sort (version 1.0.3)

Step 12: Paste (version 1.0.0)

Step 13: Join two Datasets (version 2.0.2)

Step 14: Filter (version 1.1.0)

☐ Send results to a new history

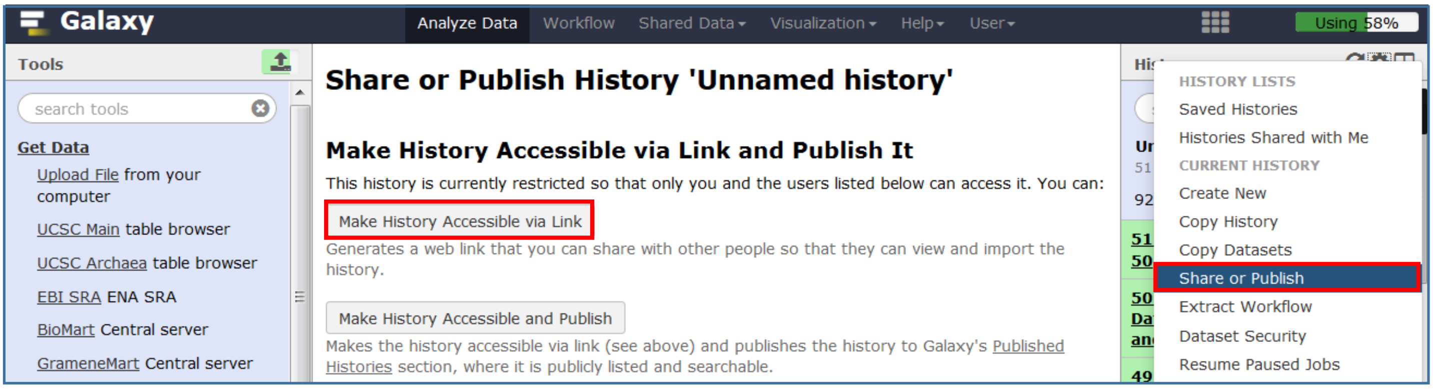
Run workflow

When the annotation and filtering workflow has finished view the output from the last step and check that the result matches your original analysis.

Sharing and publishing your work

For part of the assessment, you will share your history and workflows with the examiner by submitting them as web links in your NGS report. We will now practice how to make these web links by sharing your current history and workflows.

1. To share your history, click the cog icon and select 'Share or Publish' then click 'Make History Accessible via Link'. This will generate a web link that you can share with others so they can import your history into their Galaxy account.



Share or Publish History 'Unnamed history'

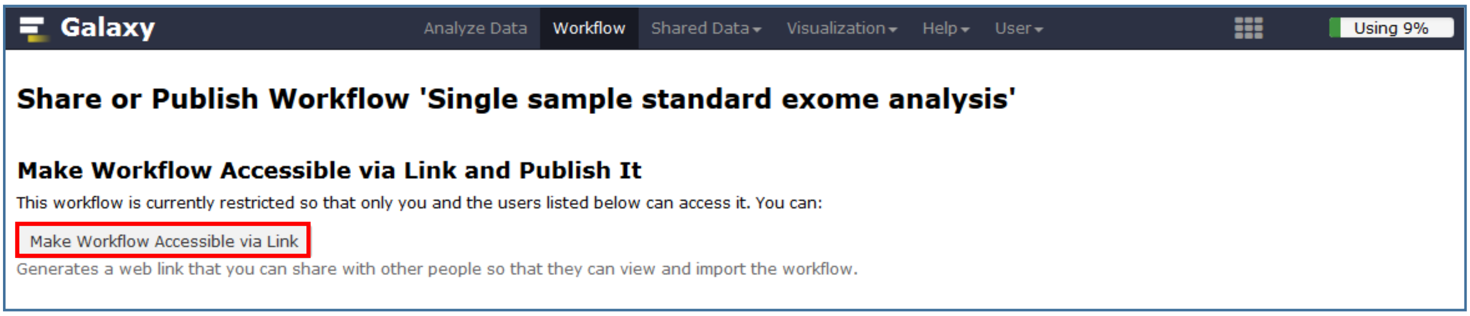
Make History Accessible via Link and Publish It

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.

2. To share your workflow, click the workflow tab, then click your workflow and select 'Share or Publish' from the drop down menu. Click 'Make History Accessible via Link' to generate a web link that can be used by others to import your workflow and apply it to their own data.



Share or Publish Workflow 'Single sample standard exome analysis'

Make Workflow Accessible via Link and Publish It

This workflow 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 workflow.

Well done Bioinformaticians you finished the exercise!