Overview

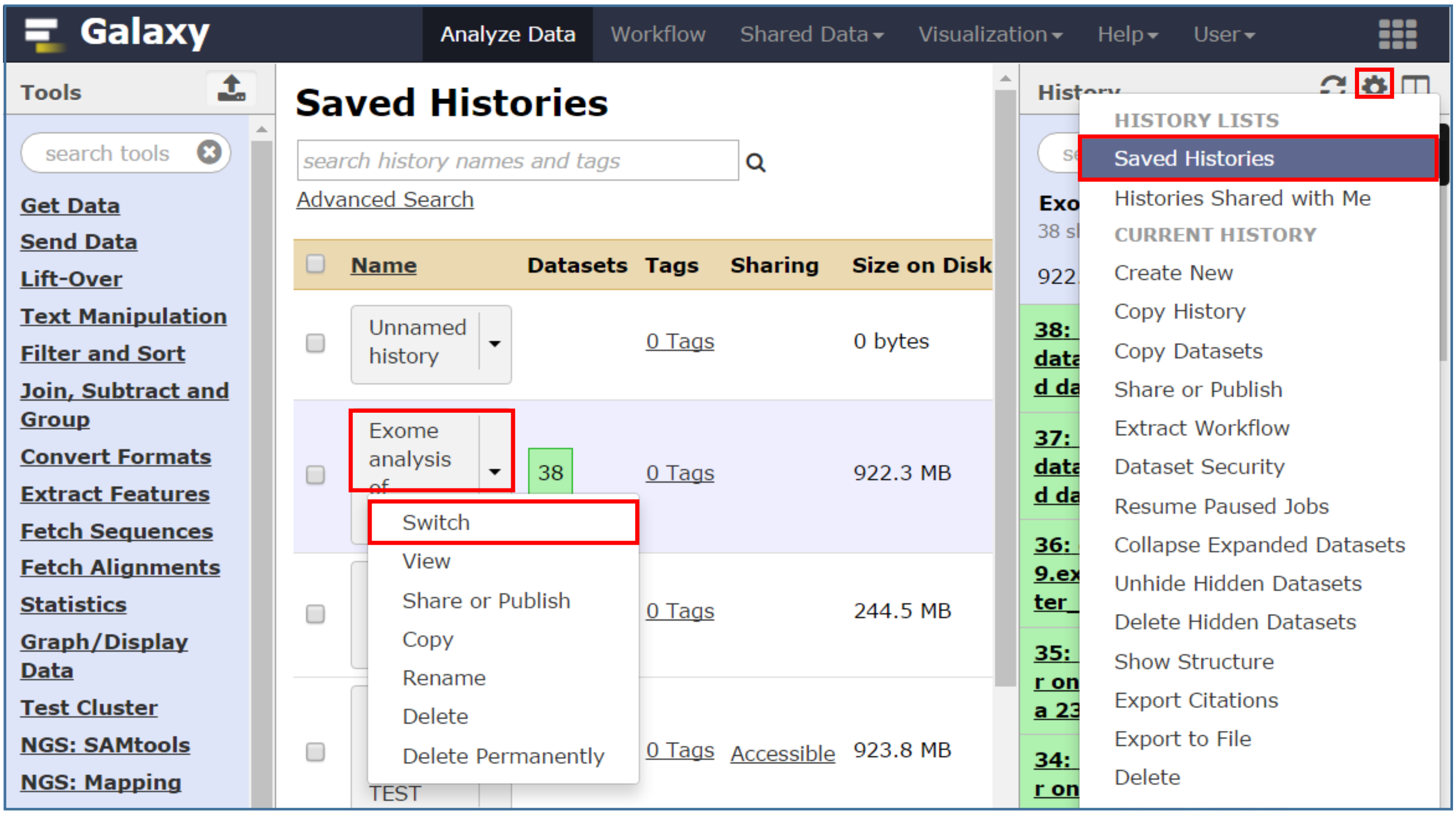
At this point in the analysis of whole exome data from patient WES01, we have assessed the quality of raw sequence data, removed low quality reads, aligned sequence data to the reference genome, removed poorly mapped reads and duplicate reads, assessed the alignment process, called variants and evaluated 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

Login to galaxy (https://galaxy2.pprd.soton.ac.uk/) and open the saved history for the analysis of WES01.

1. Click the cog icon in the history pane and select Saved Histories
2. Click the saved history for 'Exome analysis of WES01' and select switch



Name	Datasets	Tags	Sharing	Size on Disk
Unnamed history		0 Tags		0 bytes
Exome analysis of	38	0 Tags		922.3 MB
		0 Tags		244.5 MB
TEST		0 Tags	Accessible	923.8 MB

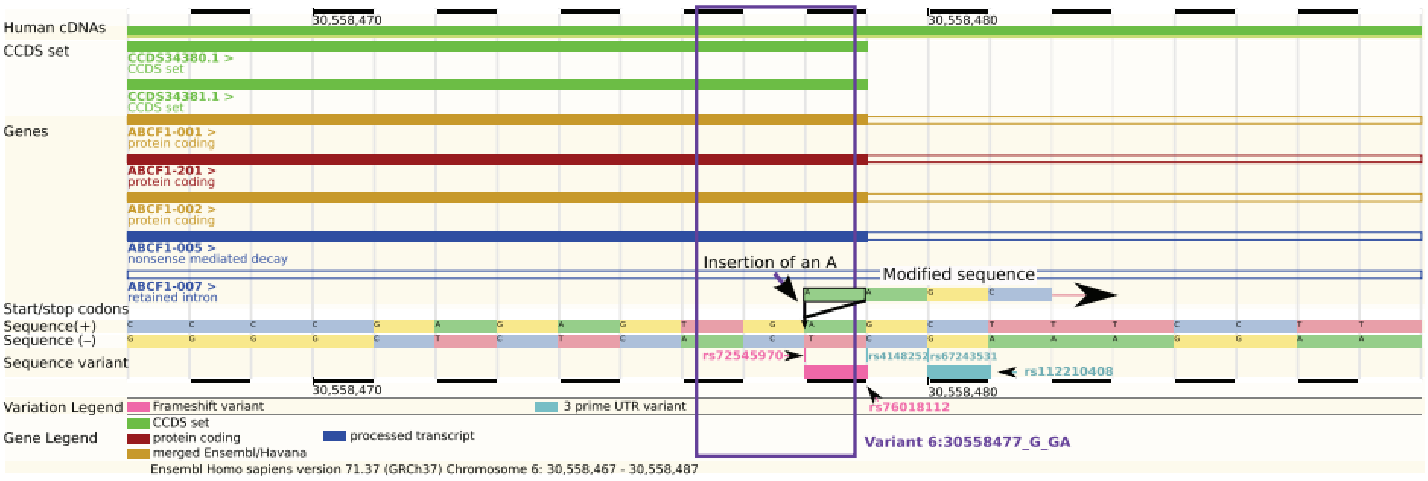
When the correct history has been loaded, the history pane should be titled 'Exome analysis of WES01'.

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).

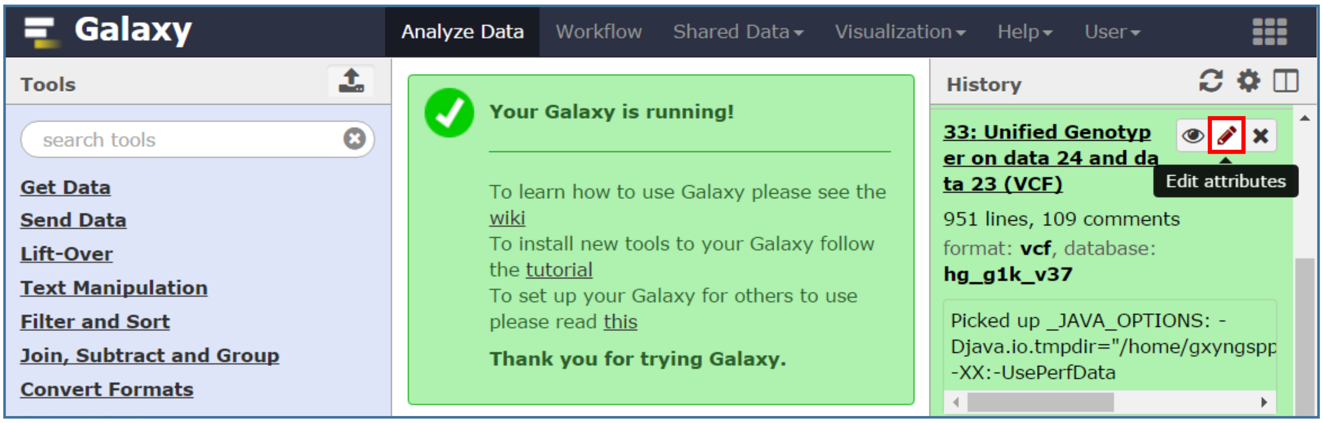
Figure 1. Annotation examples from McCarthy et al 2014



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.



Your Galaxy is running!

To learn how to use Galaxy please see the wiki
To install new tools to your Galaxy follow the tutorial
To set up your Galaxy for others to use please read this

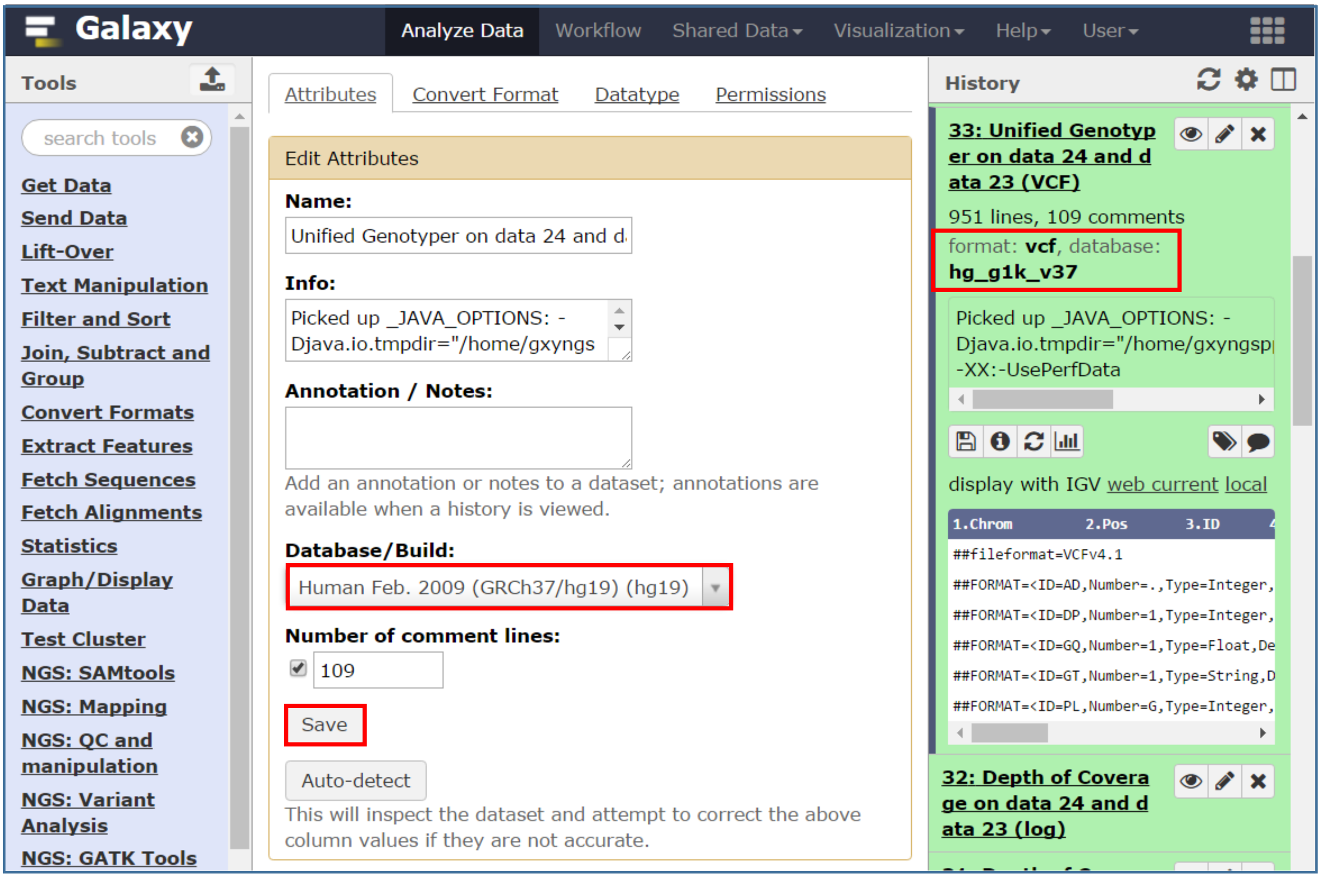
Thank you for trying Galaxy.

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

951 lines, 109 comments
format: vcf, database: hg_g1k_v37

Picked up _JAVA_OPTIONS: -Djava.io.tmpdir="/home/gxyngspp -XX:-UsePerfData

2. Select the hg19 Database/Build and click save. Note that the database associated with this step changes from hg_g1k_v37 to hg19 when you click save.



Edit Attributes

Name: Unified Genotyper on data 24 and d

Info: Picked up _JAVA_OPTIONS: -Djava.io.tmpdir="/home/gxyngs

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.

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

951 lines, 109 comments
format: vcf, database: hg_g1k_v37

Picked up _JAVA_OPTIONS: -Djava.io.tmpdir="/home/gxyngsp -XX:-UsePerfData

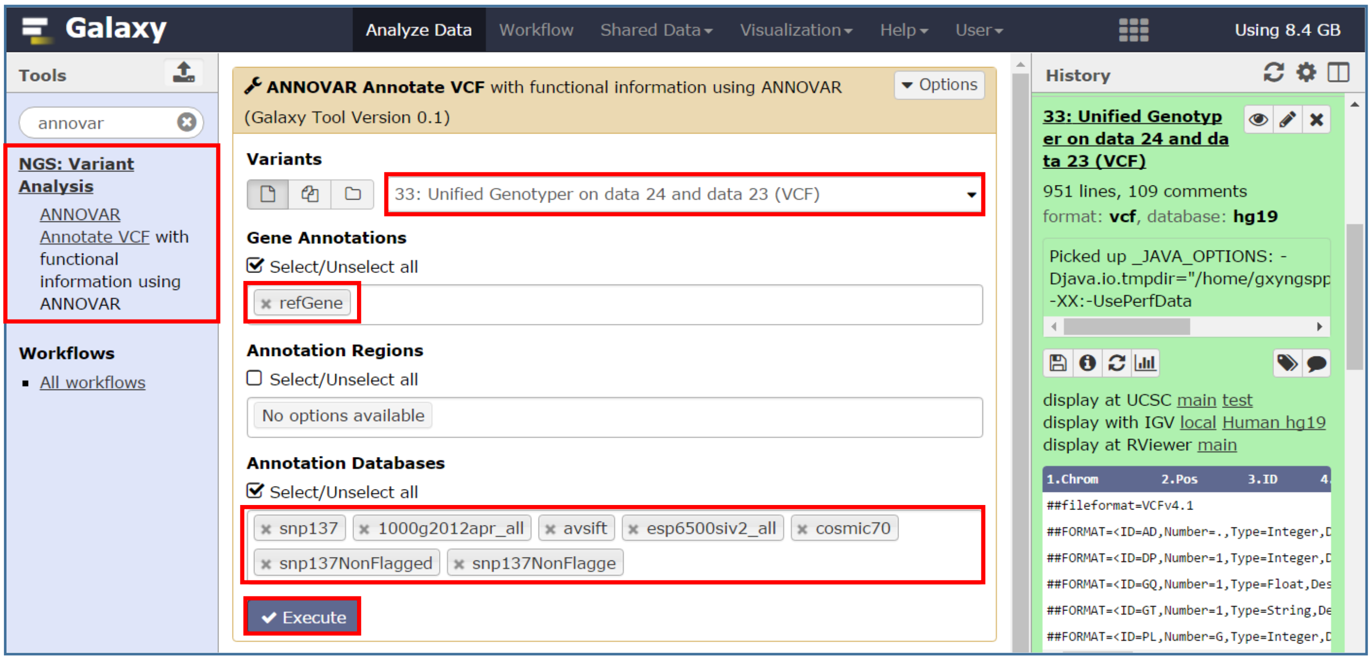
display with IGV web current local

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

32: Depth of Coverage on data 24 and data 23 (log)

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. Select all Annotation Databases.



NGS: Variant Analysis

ANNOVAR Annotate VCF with functional information using ANNOVAR

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

No options available

Annotation Databases

☑ Select/Unselect all

snp137 | 1000g2012apr_all | avsift | esp6500siv2_all | cosmic70 | snp137NonFlagged | snp137NonFlagge

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

951 lines, 109 comments

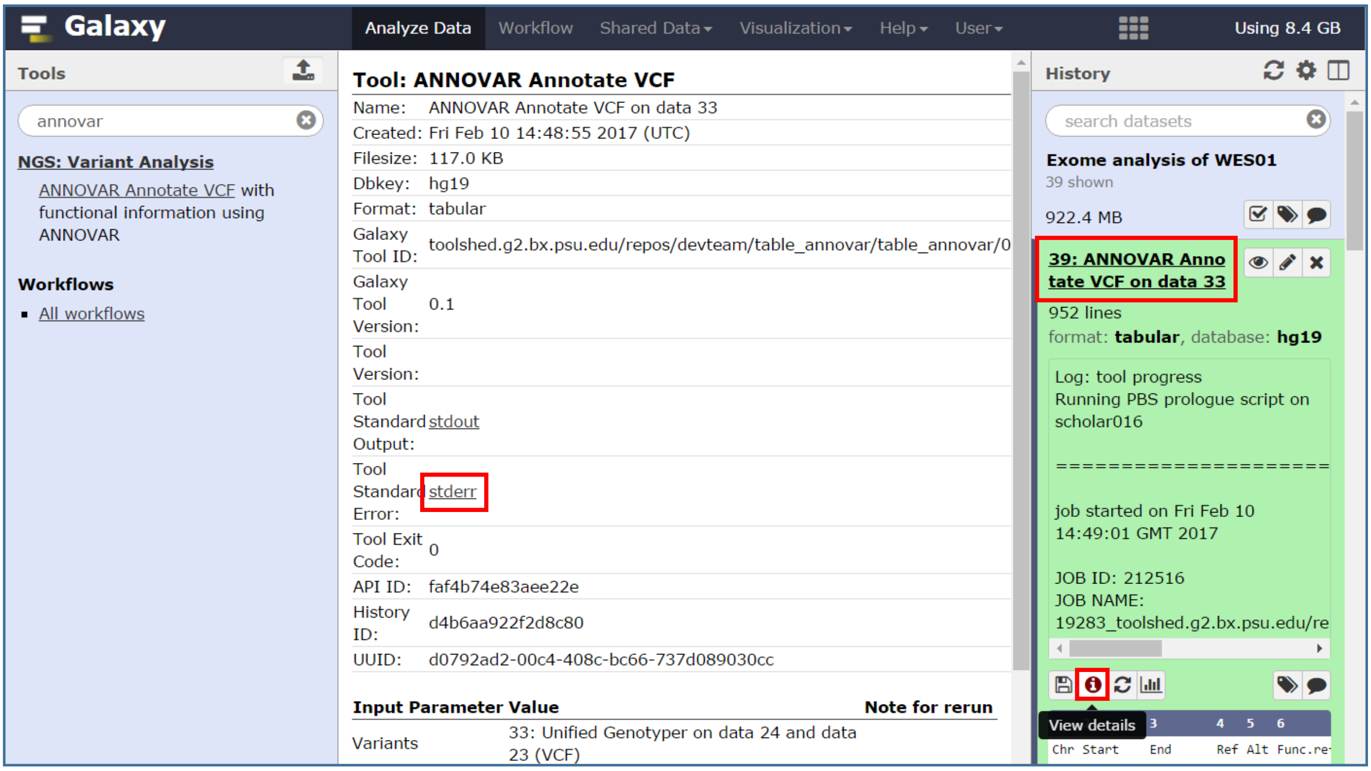
format: **vcf**, database: **hg19**

Picked up _JAVA_OPTIONS: -Djava.io.tmpdir="/home/gxyngspp -XX:-UsePerfData

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

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

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.



NGS: Variant Analysis

ANNOVAR Annotate VCF with functional information using ANNOVAR

Tool: ANNOVAR Annotate VCF

Name:	ANNOVAR Annotate VCF on data 33
Created:	Fri Feb 10 14:48:55 2017 (UTC)
Filesize:	117.0 KB
Dbkey:	hg19
Format:	tabular
Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/devteam/table_annovar/table_annovar/0
Galaxy Tool Version:	0.1
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	faf4b74e83aee22e
History ID:	d4b6aa922f2d8c80
UUID:	d0792ad2-00c4-408c-bc66-737d089030cc

Input Parameter	Value	Note for rerun
Variants	33: Unified Genotyper on data 24 and data 23 (VCF)	

Exome analysis of WES01
39 shown

922.4 MB

39: ANNOVAR Annotate VCF on data 33

952 lines

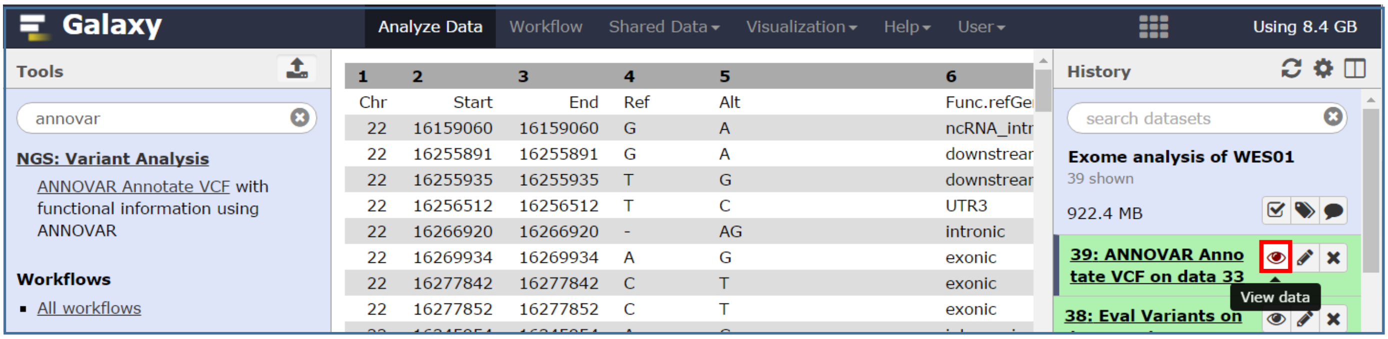
format: **tabular**, database: **hg19**

```
Log: tool progress
Running PBS prologue script on
scholar016

====================

job started on Fri Feb 10
14:49:01 GMT 2017

JOB ID: 212516
JOB NAME:
19283_toolshed.g2.bx.psu.edu/re
```

```
3      4    5    6
Chr Start    End    Ref  Alt  Func.re
```

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
Chr	Start	End	Ref	Alt	Func.refGe
22	16159060	16159060	G	A	ncRNA_intr
22	16255891	16255891	G	A	downstrear
22	16255935	16255935	T	G	downstrear
22	16256512	16256512	T	C	UTR3
22	16266920	16266920	-	AG	intronic
22	16269934	16269934	A	G	exonic
22	16277842	16277842	C	T	exonic
22	16277852	16277852	C	T	exonic

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?

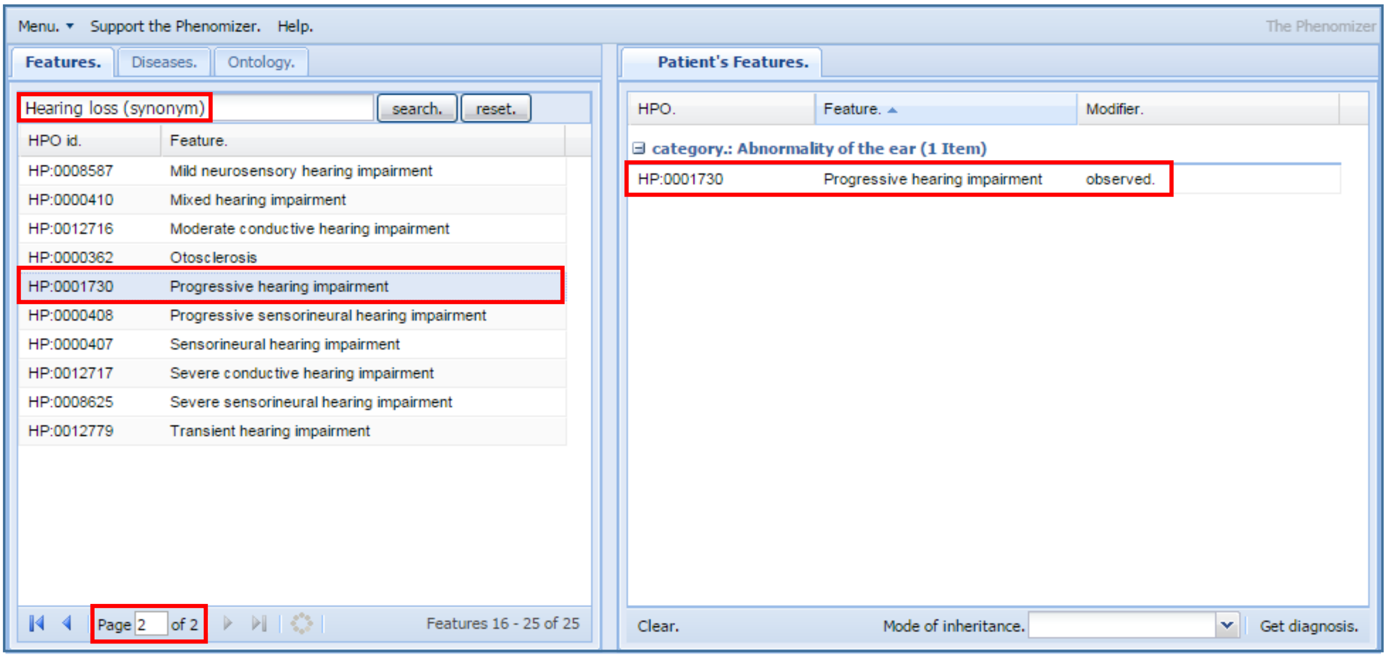
Generate a list of candidate genes

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.



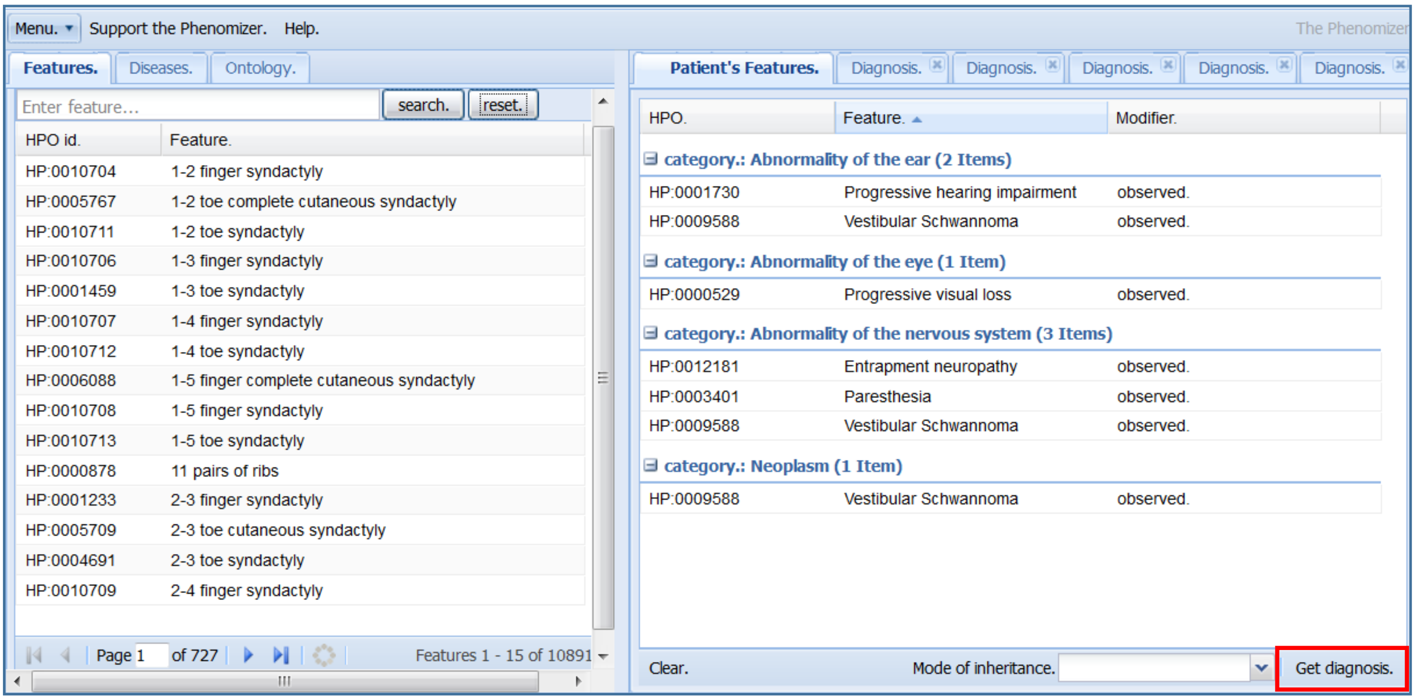
Hearing loss (synonym)

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

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.

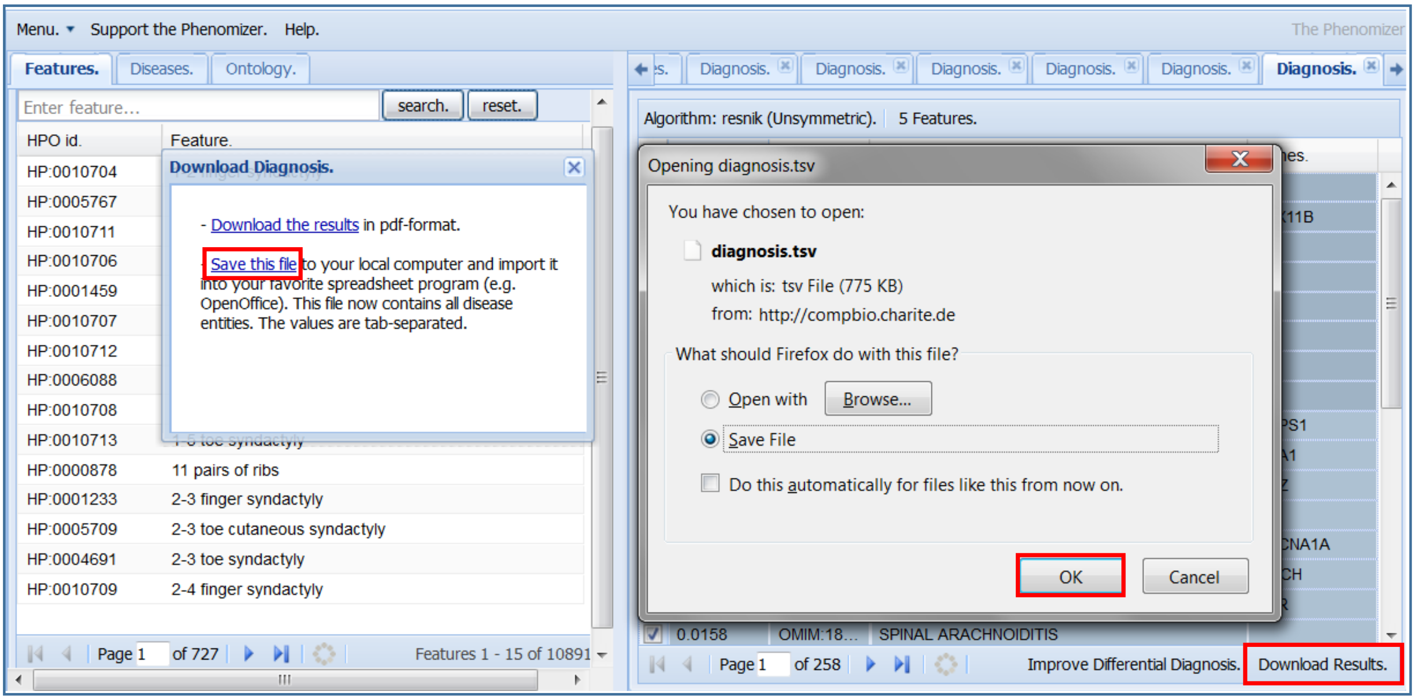


HPO id.	Feature.
HP:0010704	1-2 finger syndactyly
HP:0005767	1-2 toe complete cutaneous syndactyly
HP:0010711	1-2 toe syndactyly
HP:0010706	1-3 finger syndactyly
HP:0001459	1-3 toe syndactyly
HP:0010707	1-4 finger syndactyly
HP:0010712	1-4 toe syndactyly
HP:0006088	1-5 finger complete cutaneous syndactyly
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

Page 1 of 727 Features 1 - 15 of 10891

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.



HPO id.	Feature.
HP:0010704	
HP:0005767	
HP:0010711	
HP:0010706	
HP:0001459	
HP:0010707	
HP:0010712	
HP:0006088	
HP:0010708	
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

Page 1 of 727 Features 1 - 15 of 10891

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.

Algorithm: resnik (Unsymmetric). 5 Features.

Opening diagnosis.tsv

You have chosen to open:

diagnosis.tsv

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

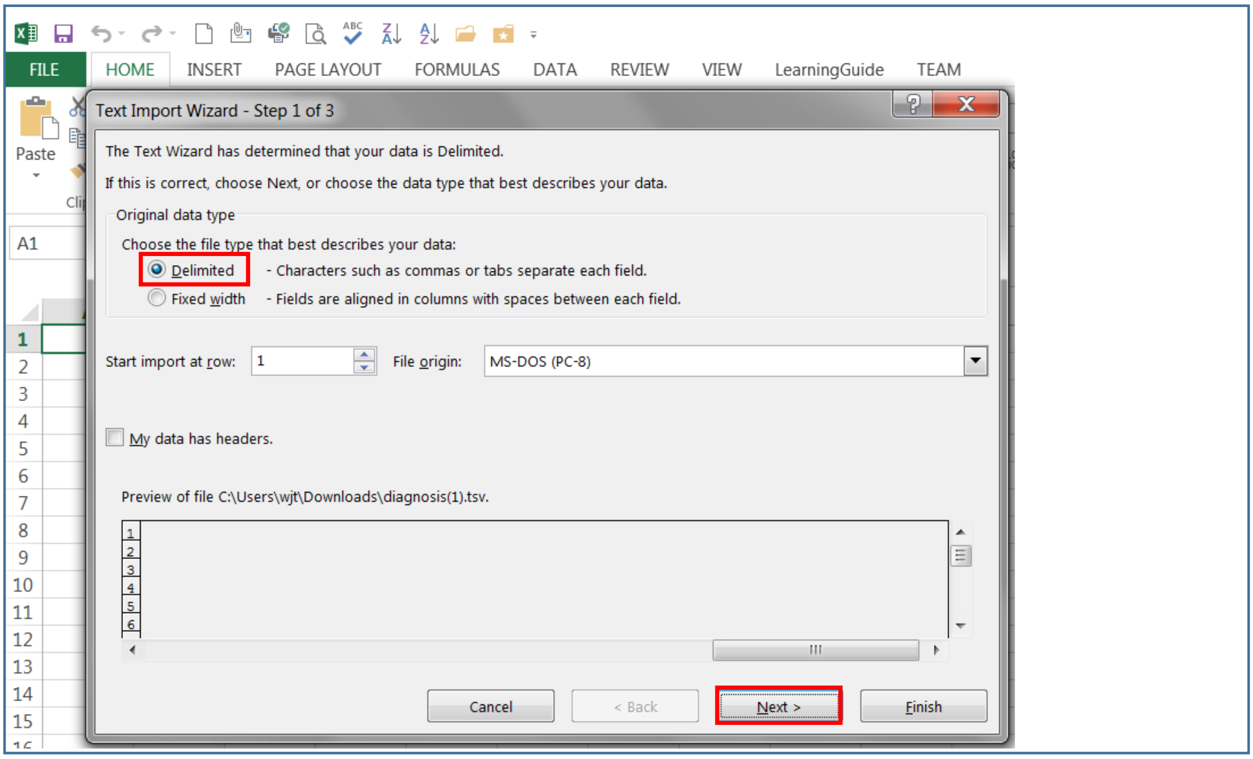
What should Firefox do with this file?

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

☑ 0.0158 OMIM:18... SPINAL ARACHNOIDITIS

Page 1 of 258

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.

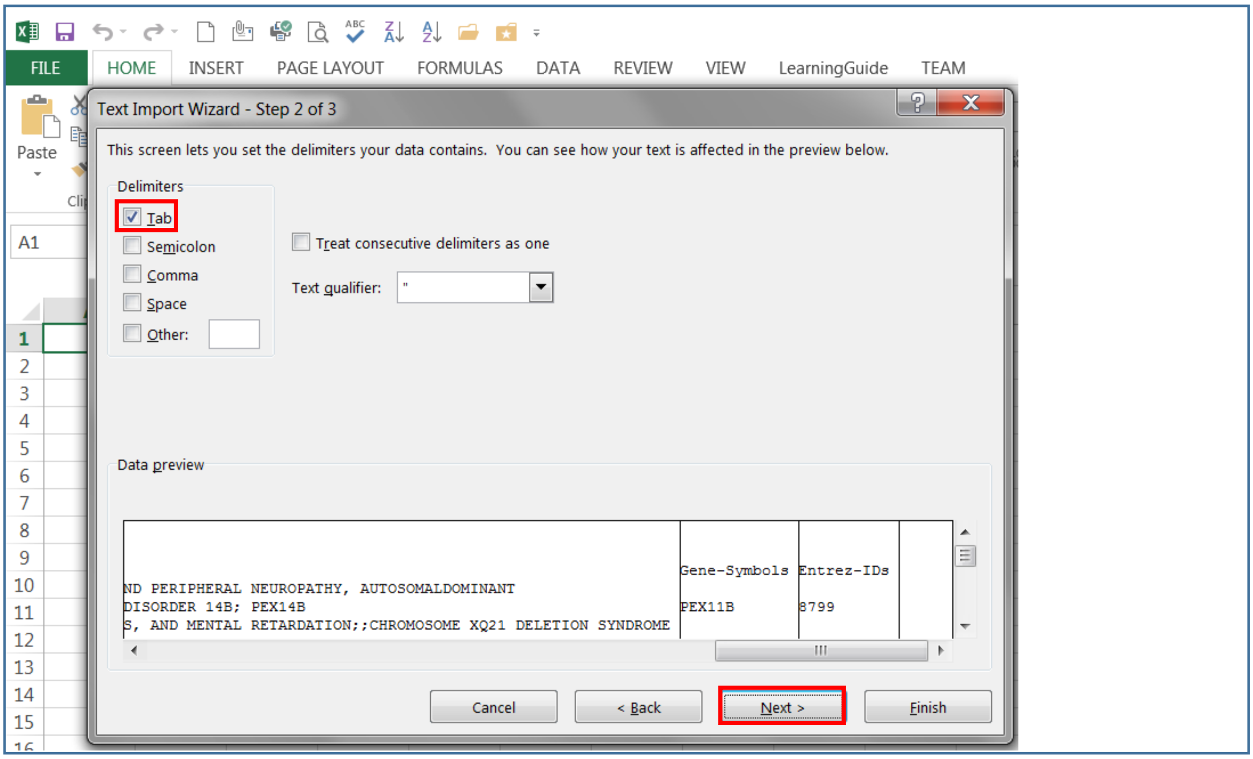
○ Fixed width - Fields are aligned in columns with spaces between each field.

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:

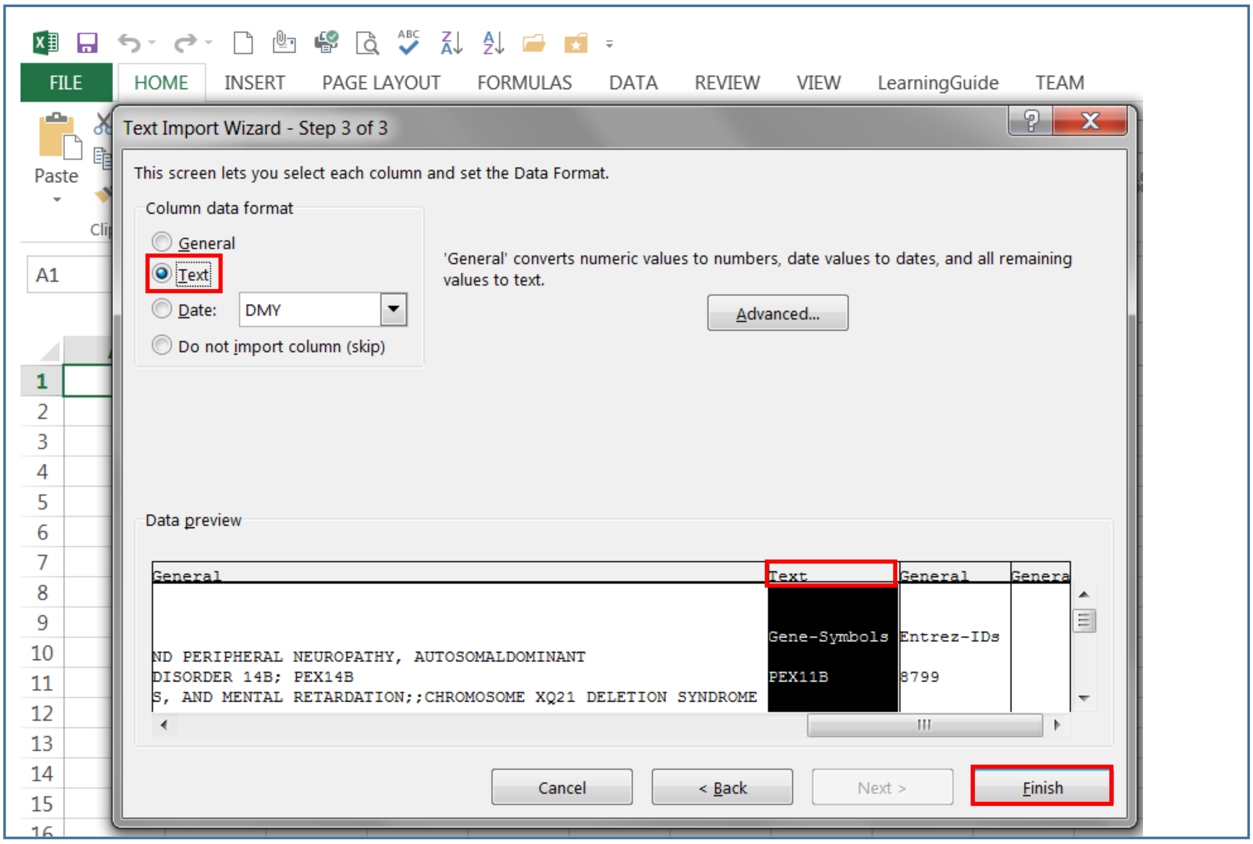
□ Treat consecutive delimiters as one

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.

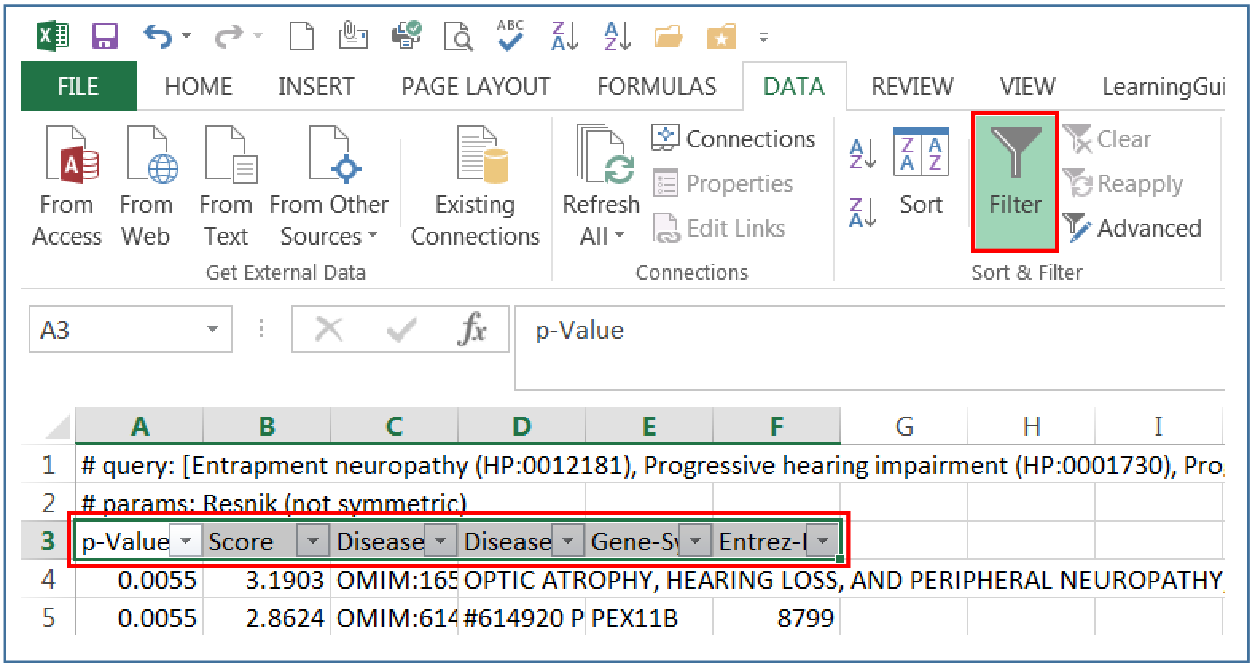
Advanced...

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
```

Cancel | < Back | Next > | Finish

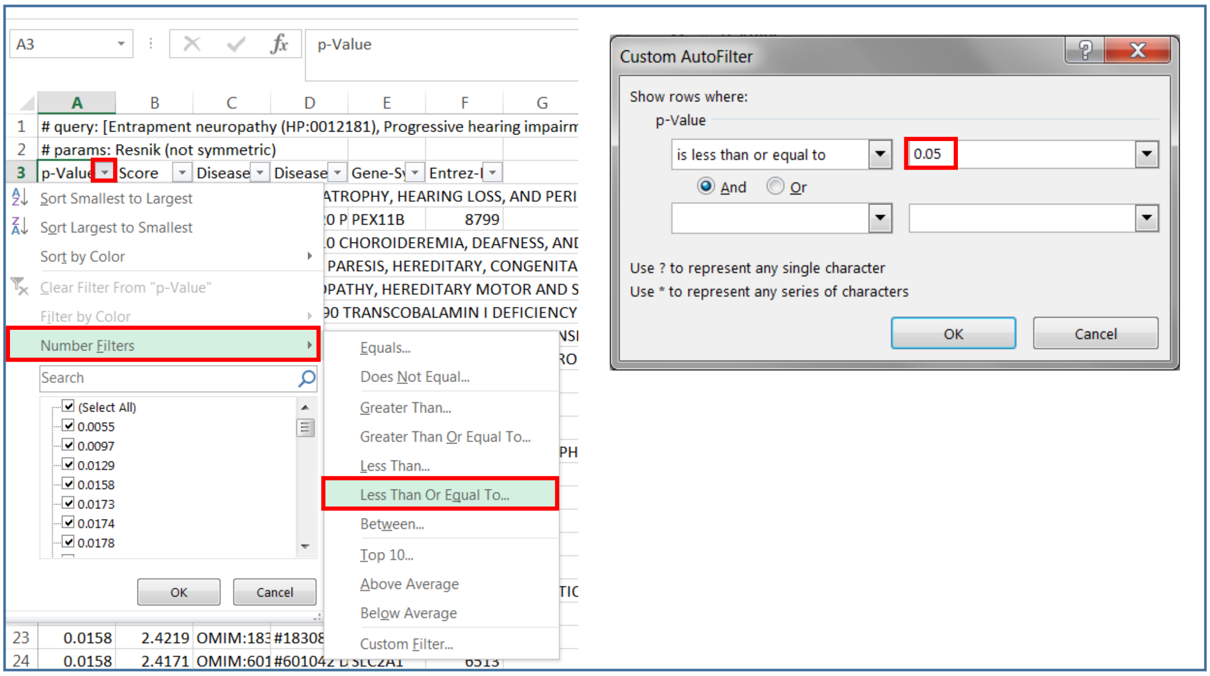
8. Add filters to the columns by selecting the column titles in row 3, hold down left click and mouse over all 6 columns, then select Filter. This will add a drop down box to each column title that enables filtering and sorting based on the column contents.



A3 | p-Value

	A	B	C	D	E	F	G	H	I
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.



Custom AutoFilter

Show rows where:

p-Value

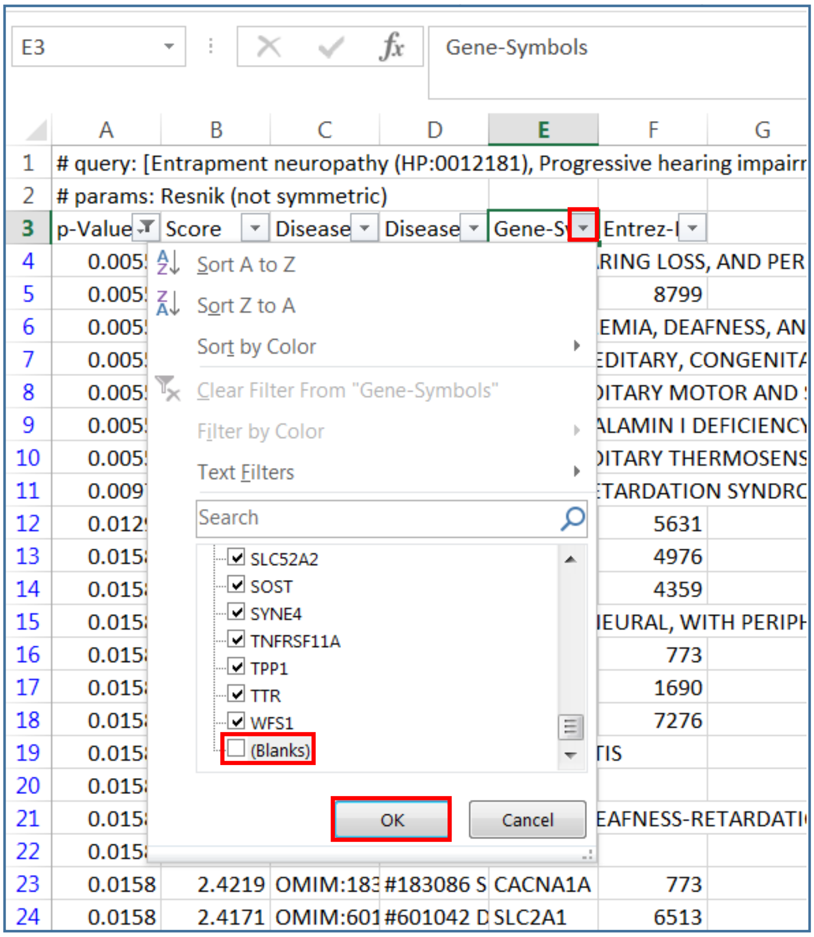
is less than or equal to 0.05

● And ○ Or

Use ? to represent any single character

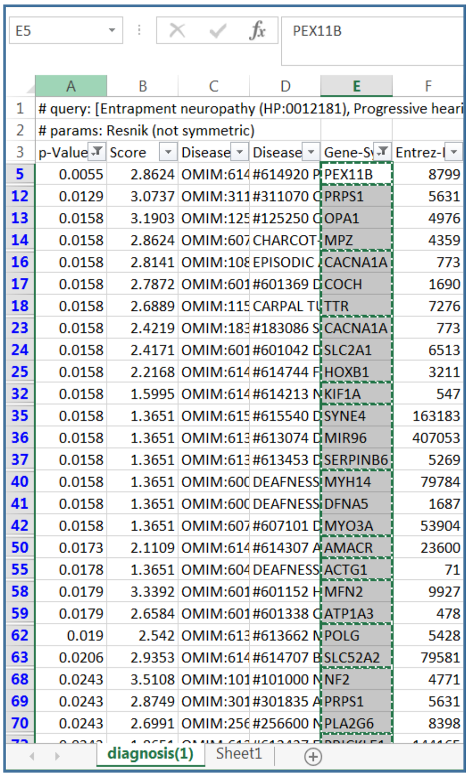
Use * to represent any series of characters

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.

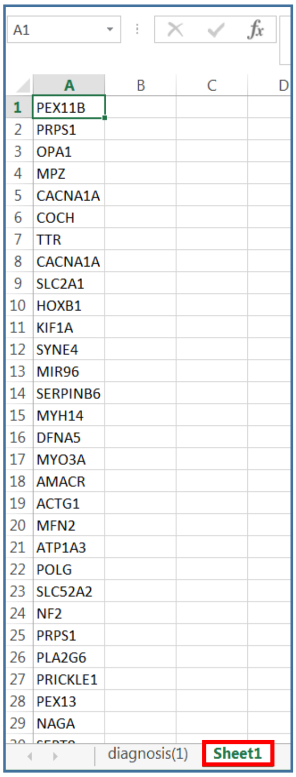


E3
Gene-Symbols
1 # query: [Entrapment neuropathy (HP:0012181), Progressive hearing impairr
2 # params: Resnik (not symmetric)
3 p-Value Score Disease Disease Gene-S Entrez-I
Sort A to Z
Sort Z to A
Sort by Color
Clear Filter From "Gene-Symbols"
Filter by Color
Text Filters
Search
☑ SLC52A2
☑ SOST
☑ SYNE4
☑ TNFRSF11A
☑ TPP1
☑ TTR
☑ WFS1
☐ (Blanks)
OK
Cancel
23 0.0158 2.4219 OMIM:18 #183086 S CACNA1A 773
24 0.0158 2.4171 OMIM:601 #601042 D SLC2A1 6513

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.

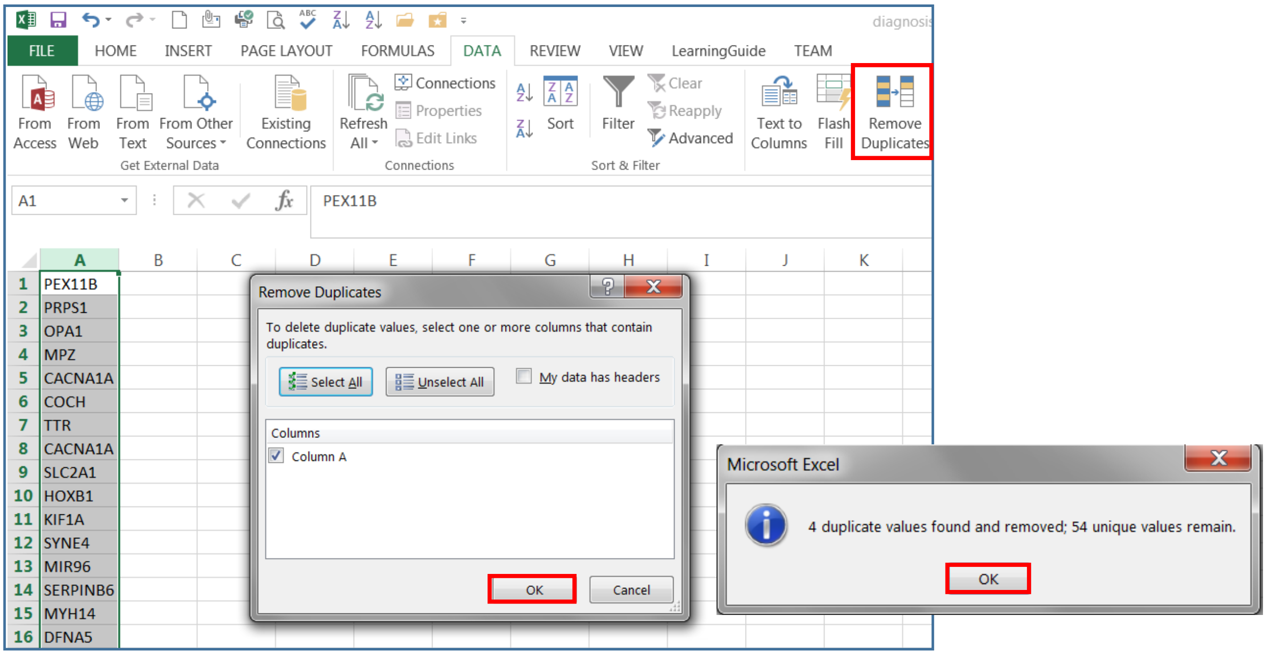


	A	B	C	D	E	F
1	# query: [Entrapment neuropathy (HP:0012181), Progressive heari					
2	# params: Resnik (not symmetric)					
3	p-Value	Score	Disease	Disease	Gene-Sy	Entrez-I
5	0.0055	2.8624	OMIM:614	#614920 P	PEX11B	8799
12	0.0129	3.0737	OMIM:311	#311070 C	PRPS1	5631
13	0.0158	3.1903	OMIM:125	#125250 C	OPA1	4976
14	0.0158	2.8624	OMIM:607	CHARCOT-	MPZ	4359
16	0.0158	2.8141	OMIM:108	EPISODIC	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



	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
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.



	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

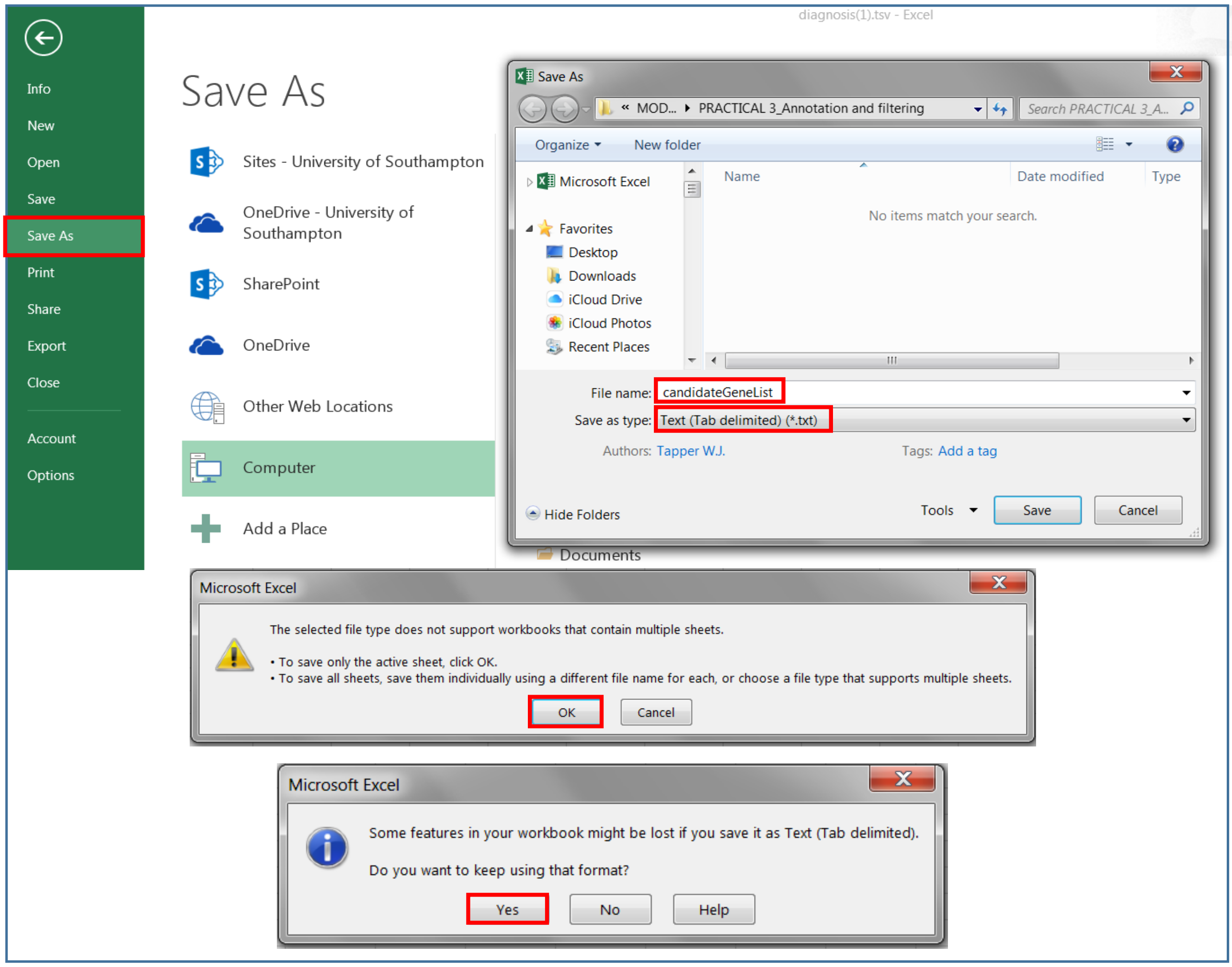
Microsoft Excel

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.



No items match your search.

File name: candidateGeneList

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

Authors: Tapper W.J.

Tags: Add a tag

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

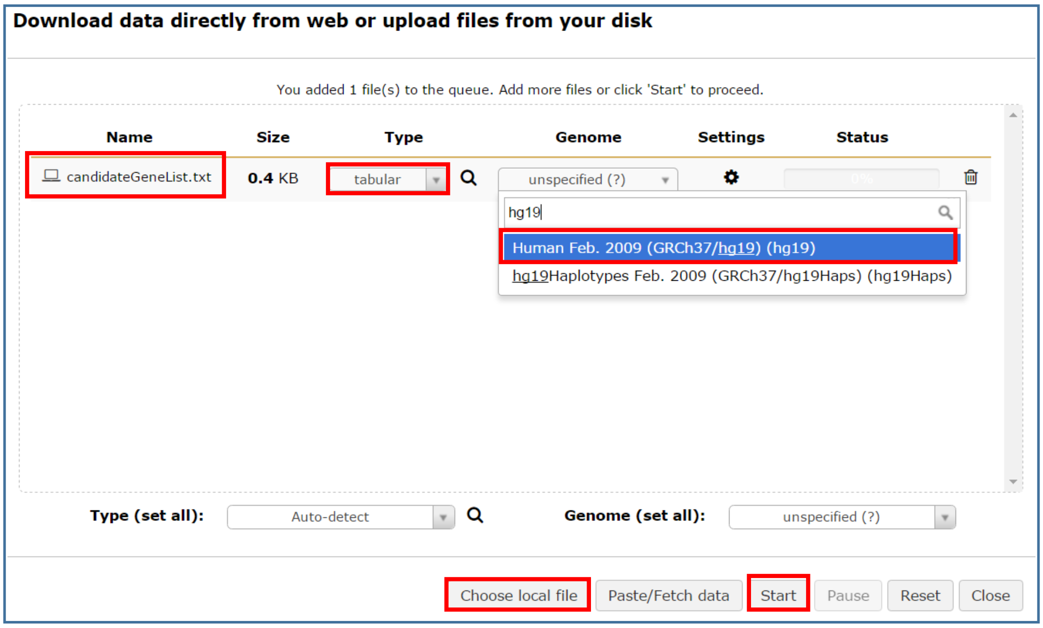
- To save only the active sheet, click OK.
- To save all sheets, save them individually using a different file name for each, or choose a file type that supports multiple sheets.

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.



Download data directly from web or upload files from your disk

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
candidateGeneList.txt	0.4 KB	tabular	unspecified (?)		

hg19

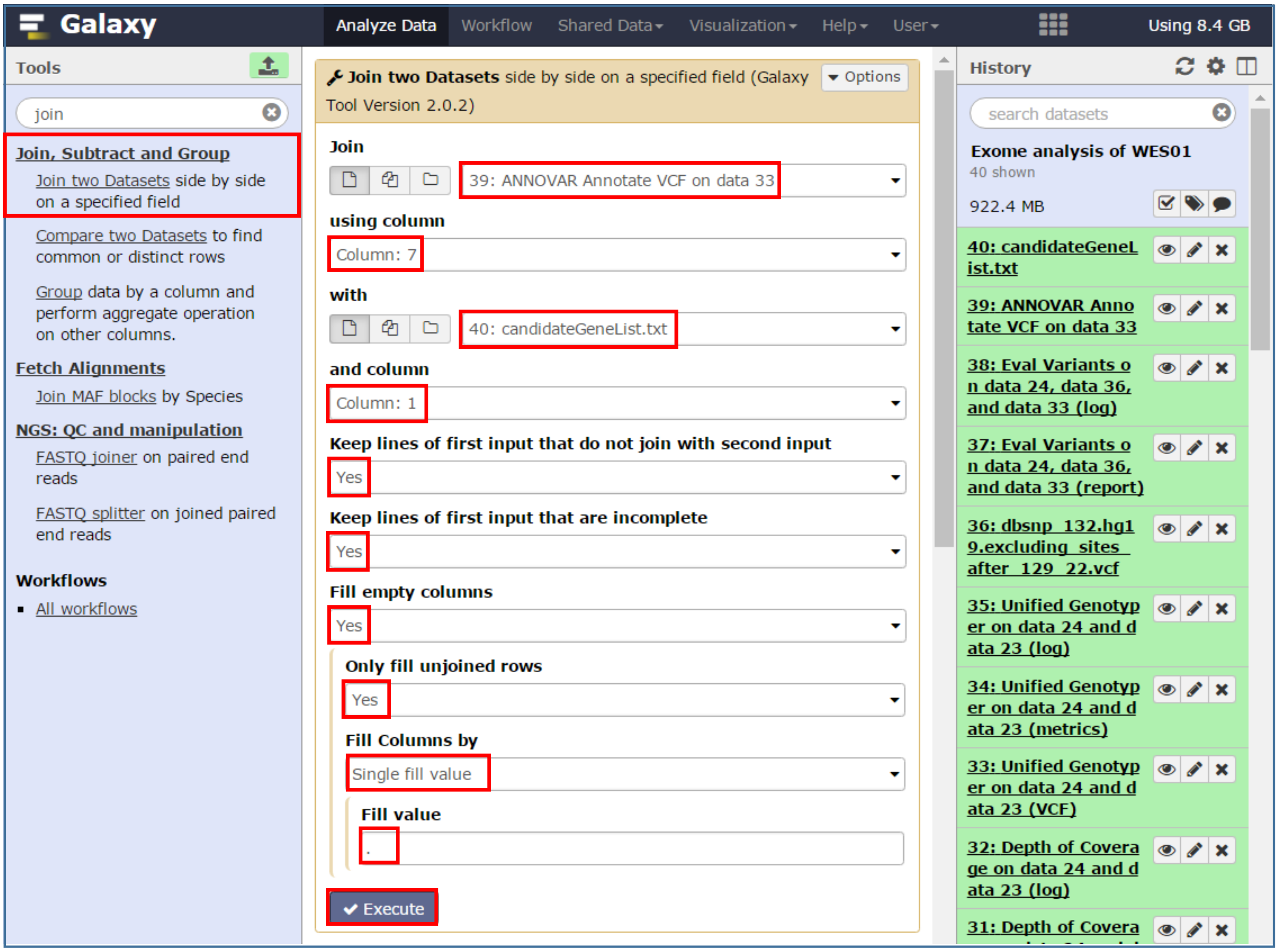
Human Feb. 2009 (GRCh37/hg19) (hg19)

hg19Haplotypes Feb. 2009 (GRCh37/hg19Haps) (hg19Haps)

Type (set all): Auto-detect

Genome (set all): unspecified (?)

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



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

Join: 39: ANNOVAR Annotate VCF on data 33

using column: Column: 7

with: 40: 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: .

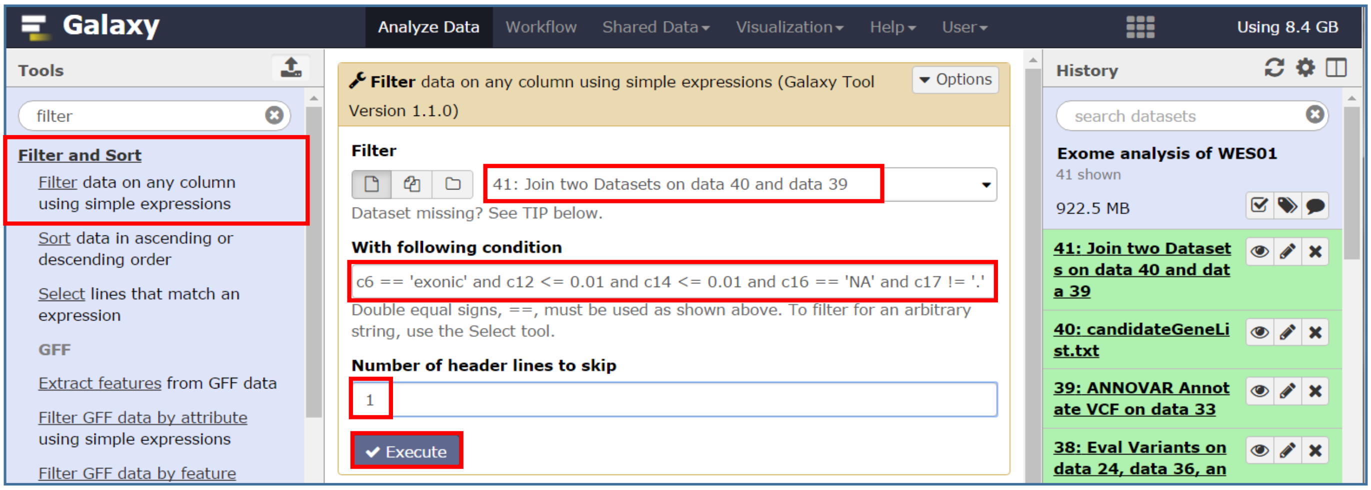
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 8.4 GB
Tools
join
Join, Subtract and Group
Join two Datasets side by side on a specified field
Compare two Datasets to find common or distinct rows
Group data by a column and perform aggregate operation on other columns.
Fetch Alignments
Join MAF blocks by Species
NGS: QC and manipulation
FASTQ joiner on paired end reads
FASTQ splitter on joined paired end reads
Workflows
All workflows
Options
Execute
History
search datasets
Exome analysis of WES01
40 shown
922.4 MB
40: candidateGeneList.txt
39: ANNOVAR Annotate VCF on data 33
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)
32: Depth of Coverage on data 24 and data 23 (log)
31: Depth of Covera

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 (c12 <= 0.01), the Exome Sequencing Project (c14 <= 0.01) and the non flagged version of dbSNP (c16 == ‘NA’), and are located in a candidate gene (c17 != ‘.’).


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 8.4 GB
Tools
filter
Filter and Sort
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
Options
History
search datasets
Execute

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

Filter

41: Join two Datasets on data 40 and data 39

Dataset missing? See TIP below.

With following condition

c6 == 'exonic' and c12 <= 0.01 and c14 <= 0.01 and c16 == 'NA' and c17 != '.'

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

1

Exome analysis of WES01

41 shown

922.5 MB

41: Join two Datasets on data 40 and data 39

40: candidateGeneList.txt

39: ANNOVAR Annotate VCF on data 33

38: Eval Variants on data 24, data 36, an

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 and GATK website (https://software.broadinstitute.org/gatk/guide/article?id=3225) 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.

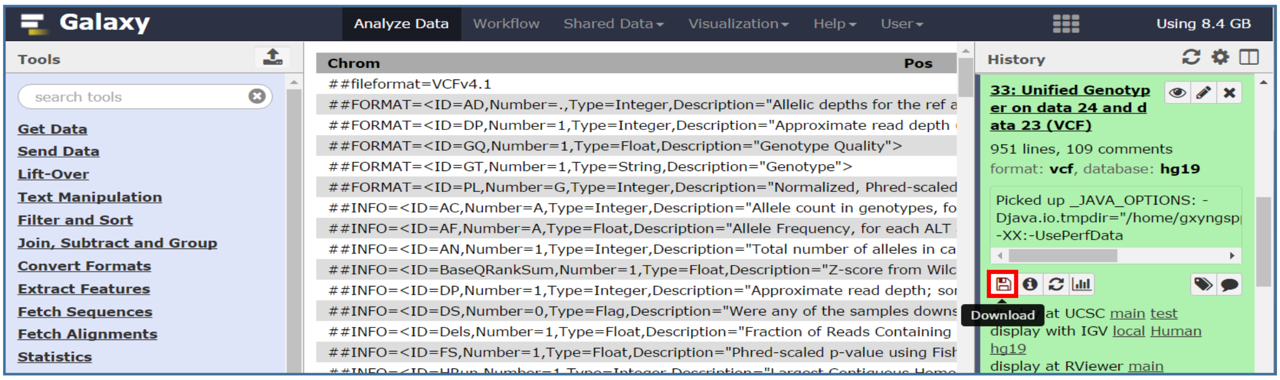
Filtering criteria	No. variants
None	
Exonic	
Exonic and absent from public databases (dbSNP non flagged, 1000 genomes, ESP)	
Exonic, absent from public databases and located in a candidate gene	

Alternatively, we could identify variants that are flagged by dbSNP as clinically associated ie they have an rsid in dbSNP137 (column 11) and no rsid in the non flagged version of dbSNP (column 16).

Q10. How many variants are flagged by dbSNP as clinically associated?

The online version of ANNOVAR offers many additional resources for annotation and filtering. Download your VCF file and use wANNOVAR to perform annotation and prioritisation of variants.

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



```
##fileformat=VCFv4.1
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref a
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (
##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
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, fo
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in ca
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilc
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; soı
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downs
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisł
```

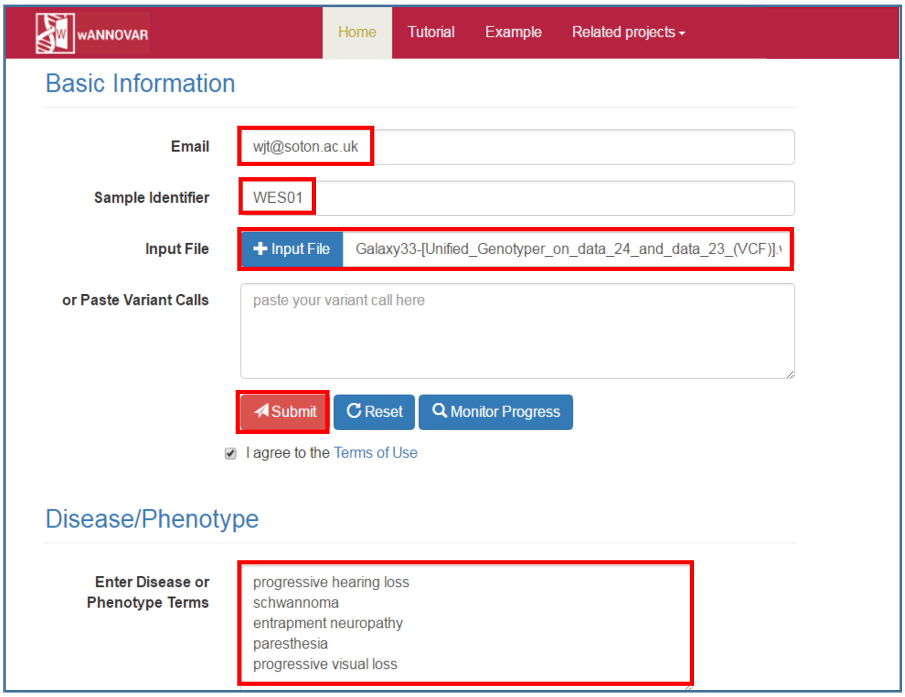
33: Unified Genotyper on data 24 and data 23 (VCF)

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

```
Picked up _JAVA_OPTIONS: -
Djava.io.tmpdir="/home/gxyngsp
-XX:-UsePerfData
```

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

5. Upload the VCF file to wANNOVAR, enter phenotype, select filters and click Submit



Basic Information

Email: wjt@soton.ac.uk

Sample Identifier: WES01

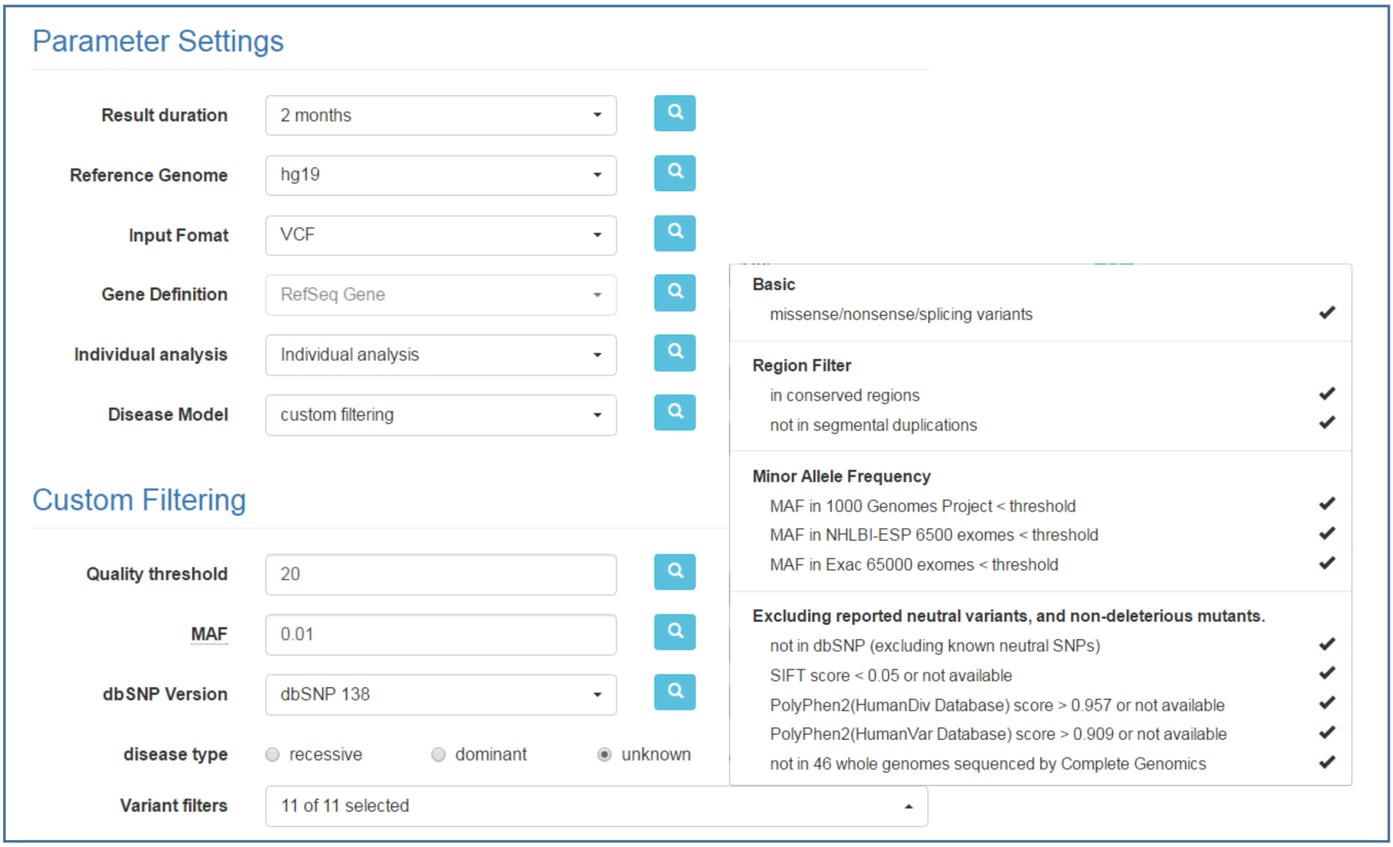
Input File: Galaxy33-[Unified_Genotyper_on_data_24_and_data_23_(VCF)]

or Paste Variant Calls: paste your variant call here

☑ I agree to the Terms of Use

Disease/Phenotype

Enter Disease or Phenotype Terms:
progressive hearing loss
schwannoma
entrapment neuropathy
paresthesia
progressive visual loss



Parameter Settings

Result duration: 2 months

Reference Genome: hg19

Input Fomat: VCF

Gene Definition: RefSeq Gene

Individual analysis: Individual analysis

Disease Model: custom filtering

Custom Filtering

Quality threshold: 20

MAF: 0.01

dbSNP Version: dbSNP 138

disease type: ○ recessive ○ dominant ● unknown

Variant filters: 11 of 11 selected

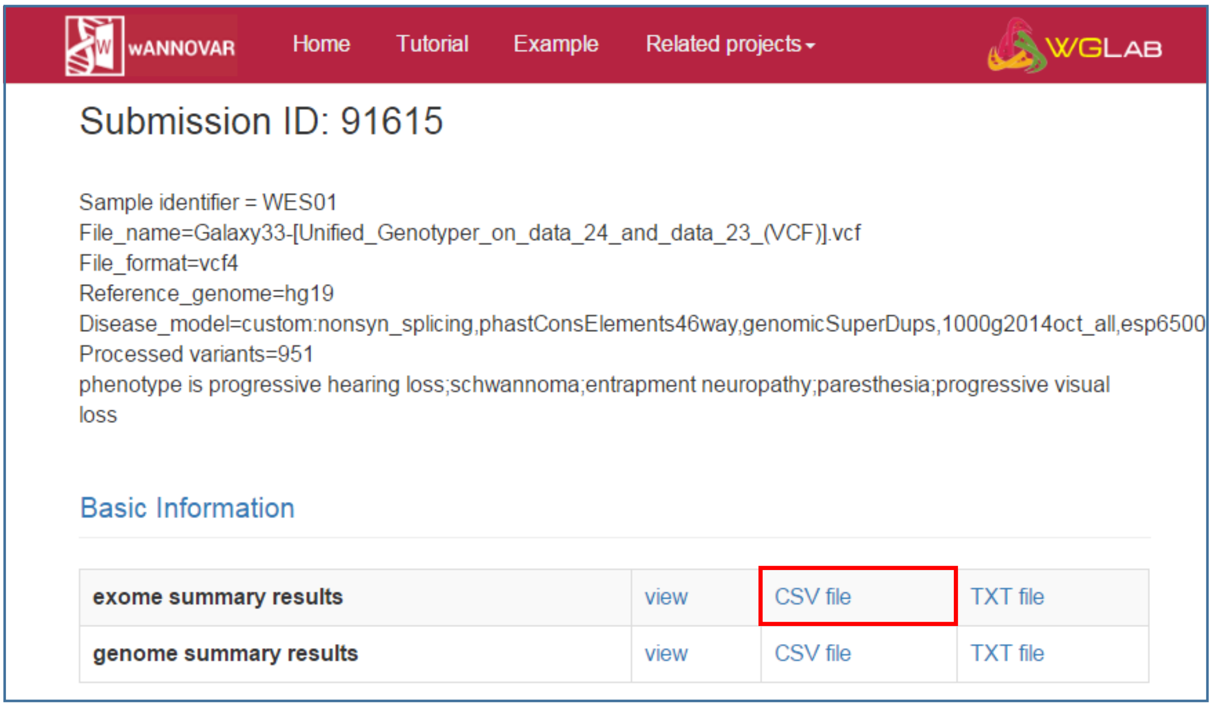
Basic
- missense/nonsense/splicing variants ✓

Region Filter
- in conserved regions ✓
- not in segmental duplications ✓

Minor Allele Frequency
- MAF in 1000 Genomes Project < threshold ✓
- MAF in NHLBI-ESP 6500 exomes < threshold ✓
- MAF in Exac 65000 exomes < threshold ✓

Excluding reported neutral variants, and non-deleterious mutants.
- not in dbSNP (excluding known neutral SNPs) ✓
- SIFT score < 0.05 or not available ✓
- PolyPhen2(HumanDiv Database) score > 0.957 or not available ✓
- PolyPhen2(HumanVar Database) score > 0.909 or not available ✓
- not in 46 whole genomes sequenced by Complete Genomics ✓

6. Download the exome summary results



Submission ID: 91615

Sample identifier = WES01
File_name=Galaxy33-[Unified_Genotyper_on_data_24_and_data_23_(VCF)].vcf
File_format=vcf4
Reference_genome=hg19
Disease_model=custom:nonsyn_splicing,phastConsElements46way,genomicSuperDups,1000g2014oct_all,esp6500
Processed variants=951
phenotype is progressive hearing loss;schwannoma;entrapment neuropathy;paresthesia;progressive visual loss

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.

Q11. 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.

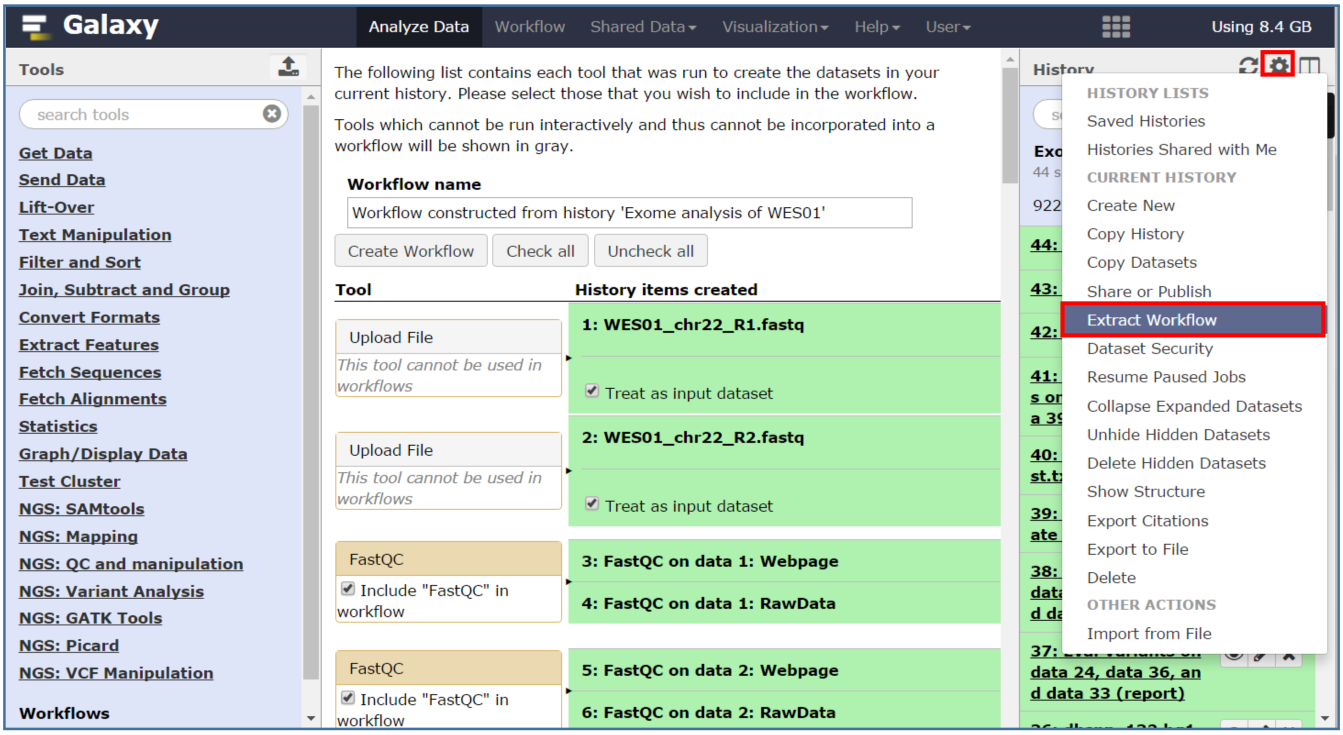
Q12. Does wANNOVAR filtering prioritize the same variant?

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.

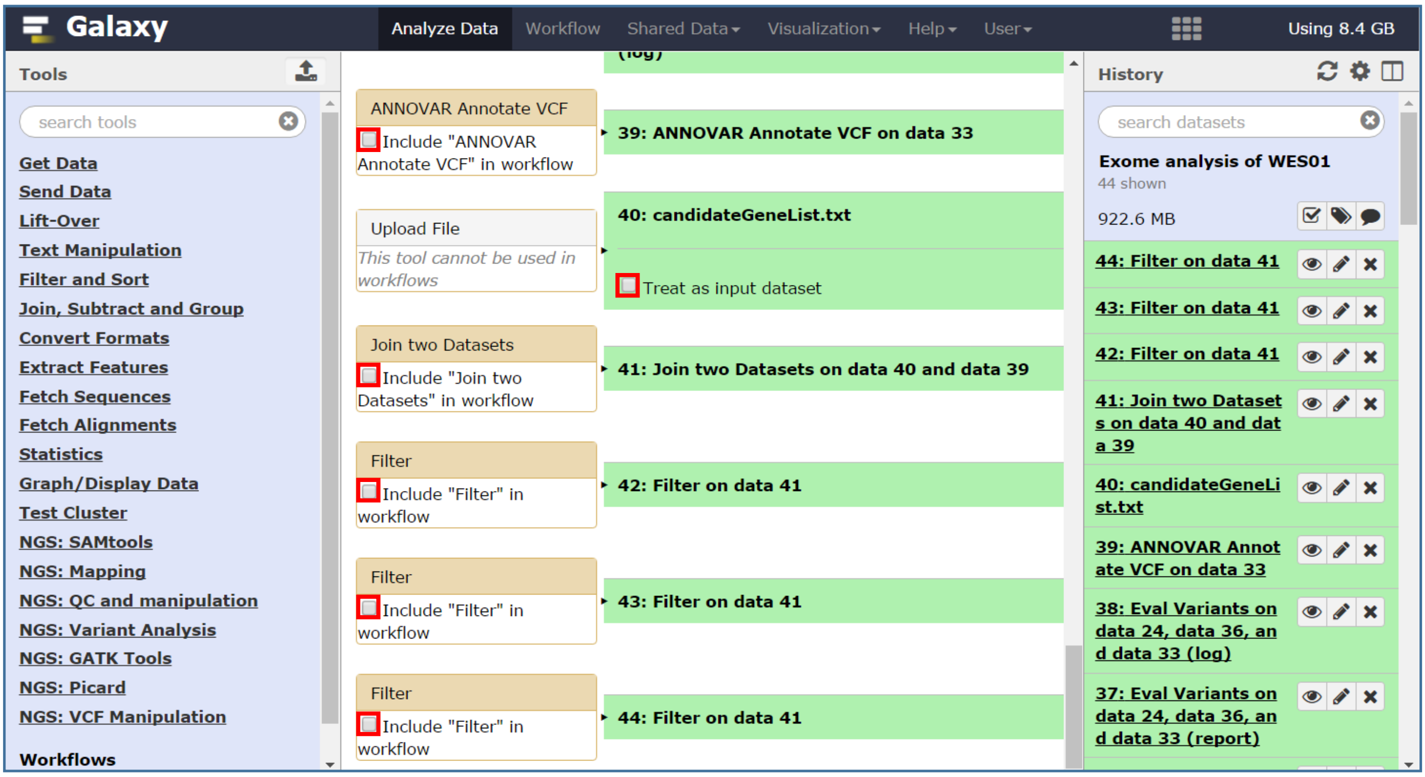
Workflow name
Workflow constructed from history 'Exome analysis of WES01'

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 1: Webpage** **4: FastQC on data 1: RawData**
FastQC ☑ Include "FastQC" in workflow	**5: FastQC on data 2: Webpage** **6: FastQC on data 2: RawData**

Exo
44 s
922
44:
43:
42:
41: s or a 39
40: st.t
39: ate
38: data d da
37: data 24, data 36, and data 33 (report)

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

Upload File
This tool cannot be used in workflows
40: candidateGeneList.txt
□ Treat as input dataset

Join two Datasets
□ Include "Join two Datasets" in workflow
41: Join two Datasets on data 40 and data 39

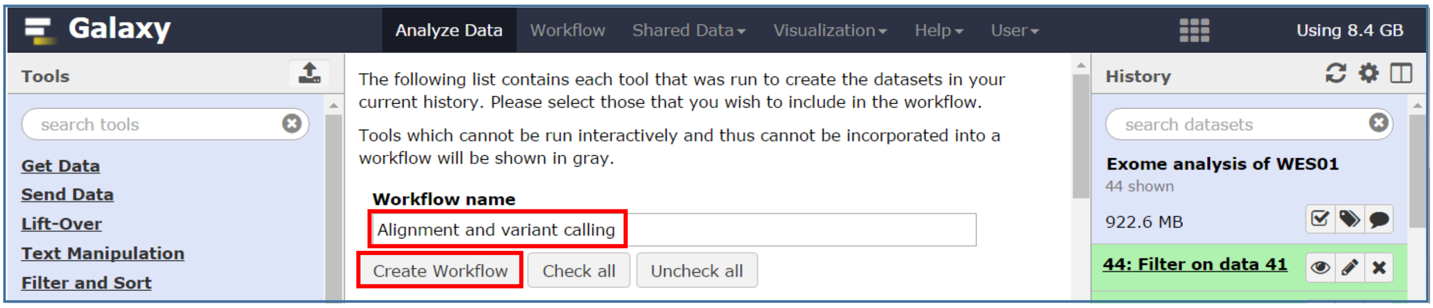
Filter
□ Include "Filter" in workflow
42: Filter on data 41

Filter
□ Include "Filter" in workflow
43: Filter on data 41

Filter
□ Include "Filter" in workflow
44: Filter on data 41

44: Filter on data 41
43: Filter on data 41
42: Filter on data 41
41: Join two Datasets on data 40 and data 39
40: candidateGeneList.txt
39: ANNOVAR Annotate VCF on data 33
38: Eval Variants on data 24, data 36, and data 33 (log)
37: Eval Variants on data 24, data 36, and data 33 (report)

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

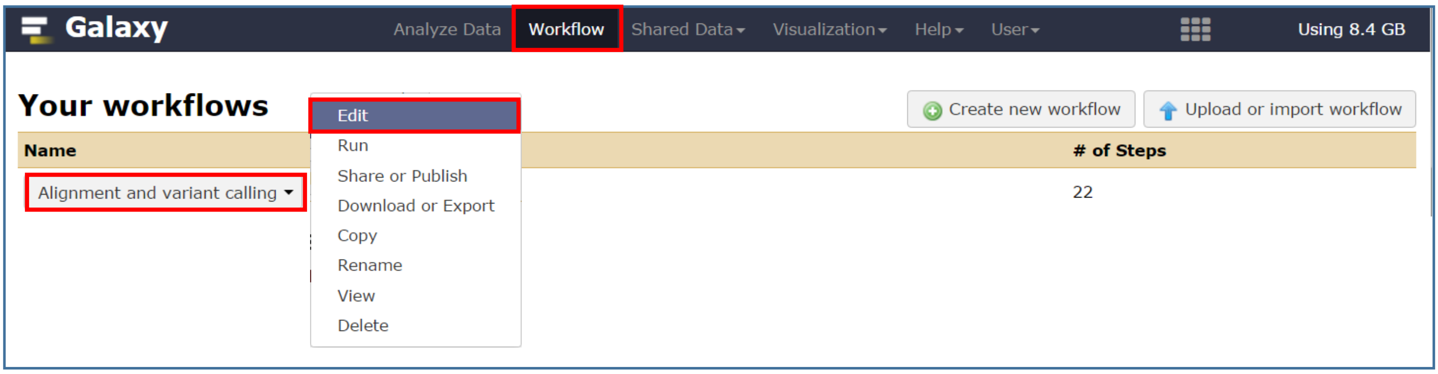


The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

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

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

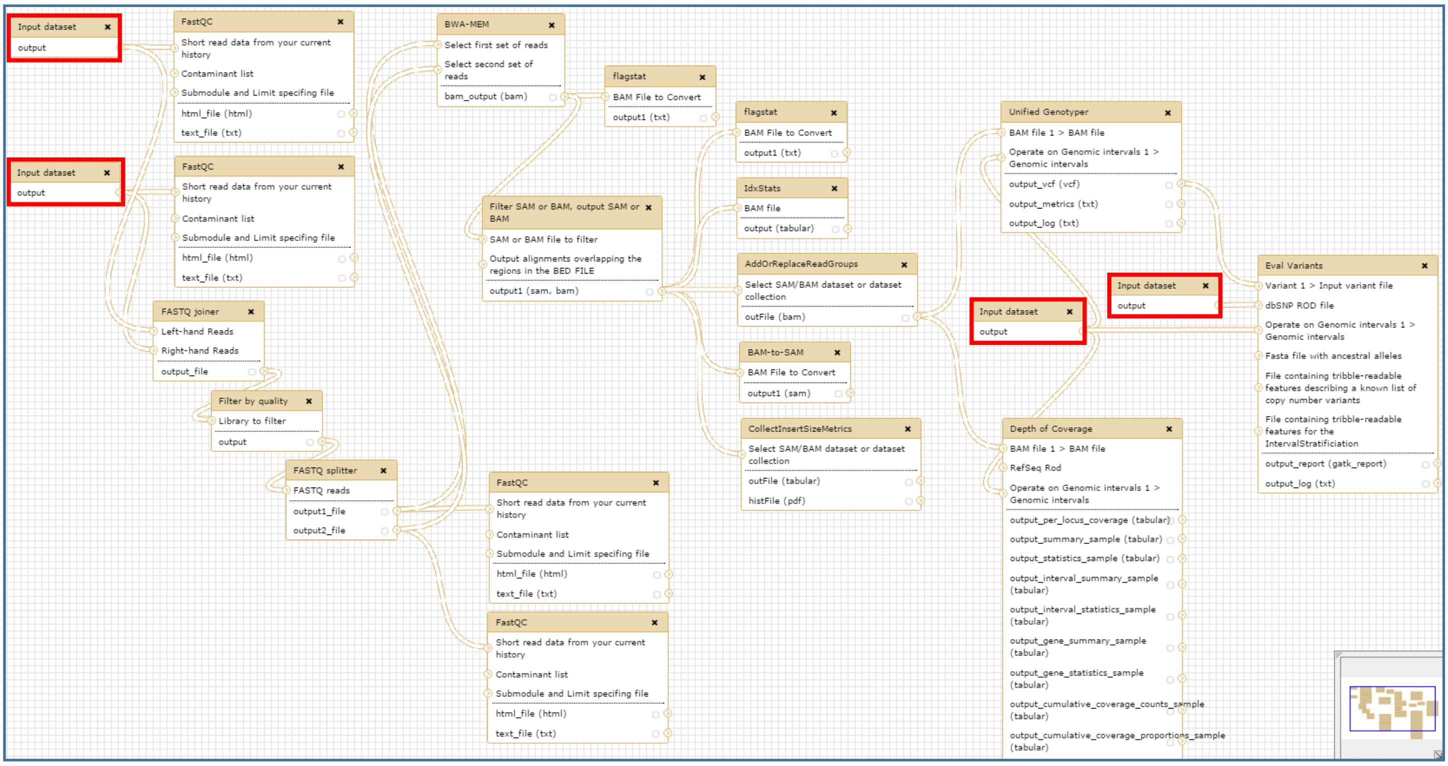


Your workflows

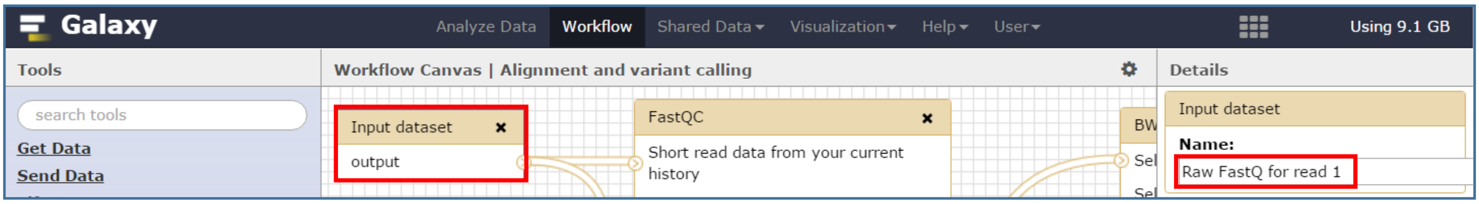
Name	# of Steps
Alignment and variant calling	22

This will bring up the workflow canvas showing a flow diagram of the selected process. The boxes indicate the input and output of each step and they can be moved around to show the workflow more clearly. Click on a box to show the step details in the right hand panel where annotation or notes for each step can be added.

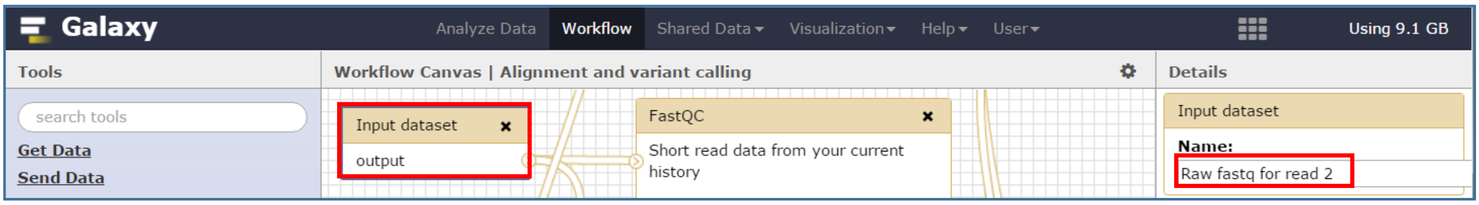
6. For the Input datasets (highlighted in red below), add a name to describe the input file which will make the workflow much easier to use.


Input dataset
output
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Input dataset
output
FastQC
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)
flagstat
BAM File to Convert
output1 (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)
IdxStats
BAM file
output (tabular)
AddOrReplaceReadGroups
Select SAM/BAM dataset or dataset collection
outFile (bam)
BAM-to-SAM
output1 (sam)
CollectInsertSizeMetrics
outFile (tabular)
histFile (pdf)
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
Eval Variants
Variant 1 > Input variant file
dbSNP ROD file
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)
Depth of Coverage
RefSeq Rod
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)

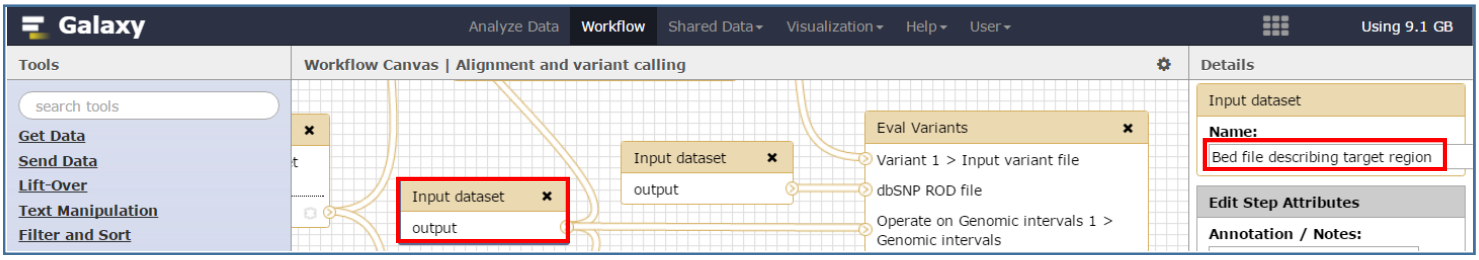
7. Click on the box for each Input dataset and add a name to describe the input file.



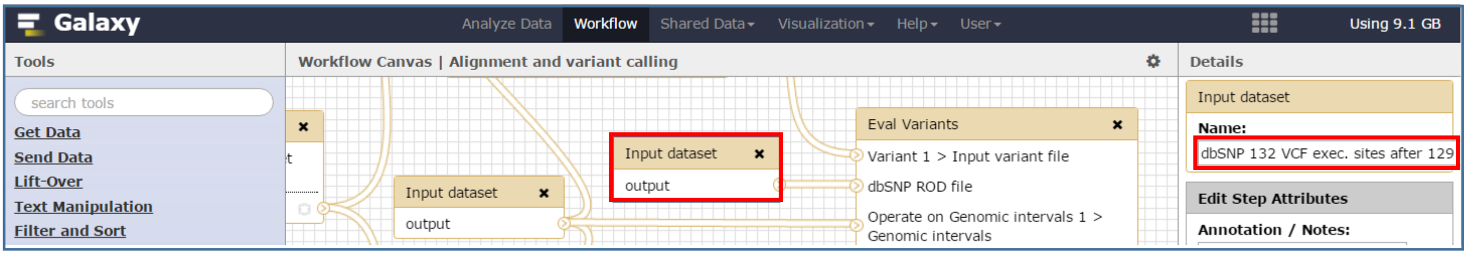
Name: Raw FastQ for read 1



Name: Raw fastq for read 2

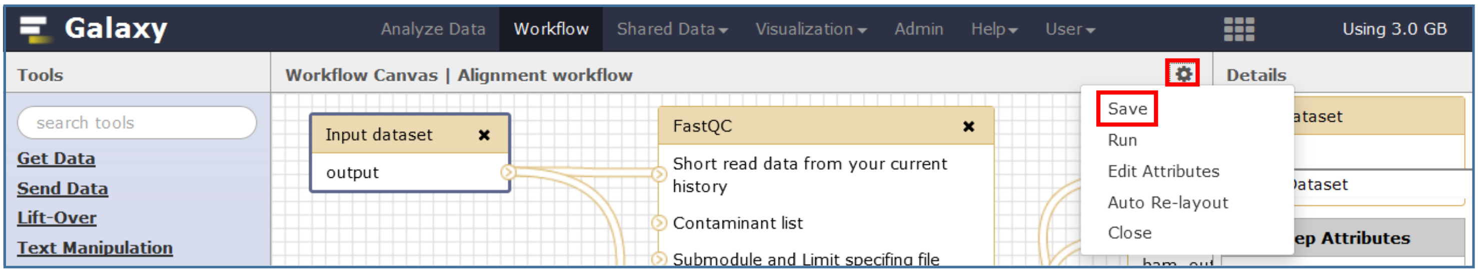


Name: Bed file describing target region



Name: dbSNP 132 VCF exec. sites after 129

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

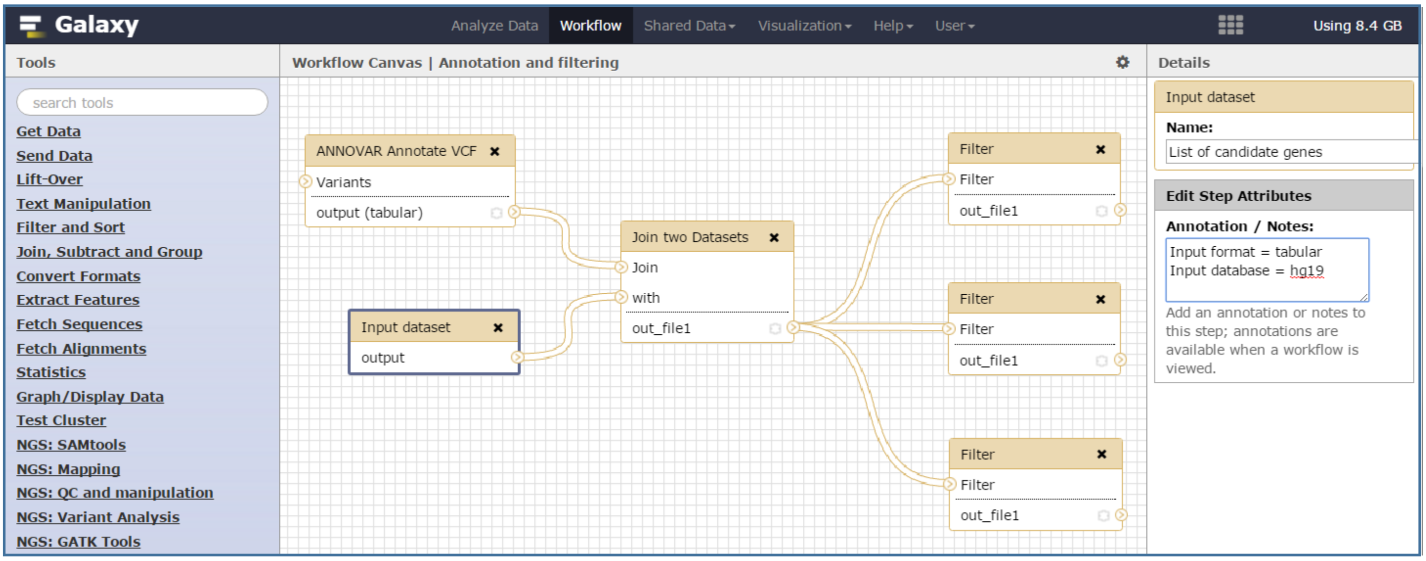


Now close the Workflow editor by clicking the cog icon and selecting 'Close'.

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

- Go back to your history for Exome 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 and annotate the input datasets.

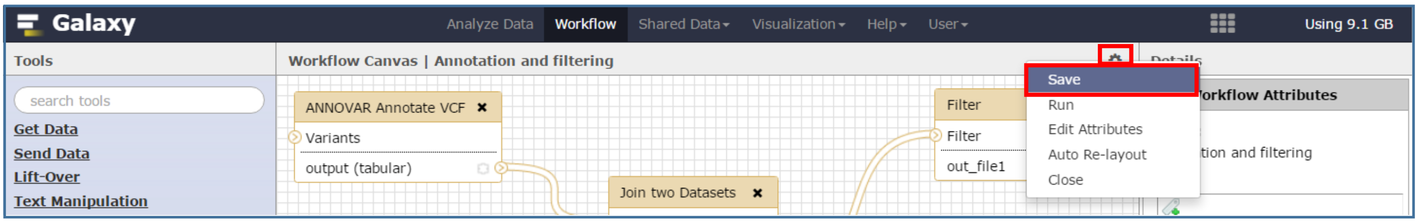
10. The Annotation and filtering workflow should look like this. Give the Input dataset a name to describe the file and its format.



Name: List of candidate genes

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

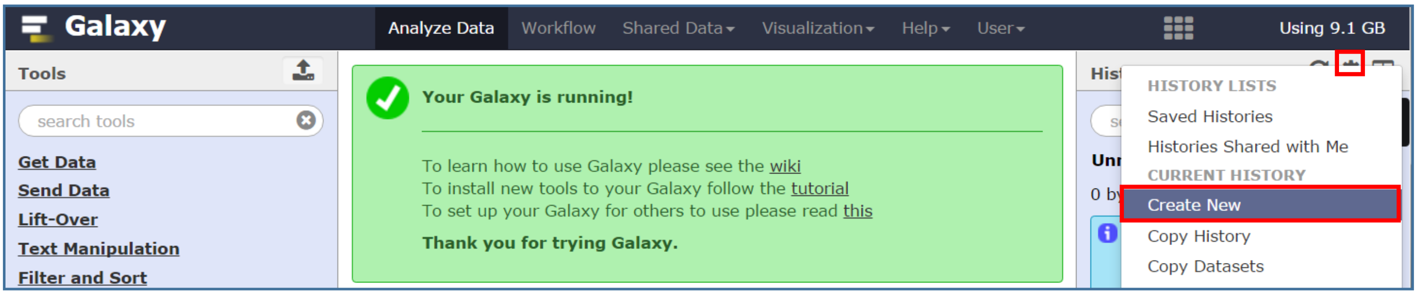
11. Save and Close the annotation and filtering workflow.



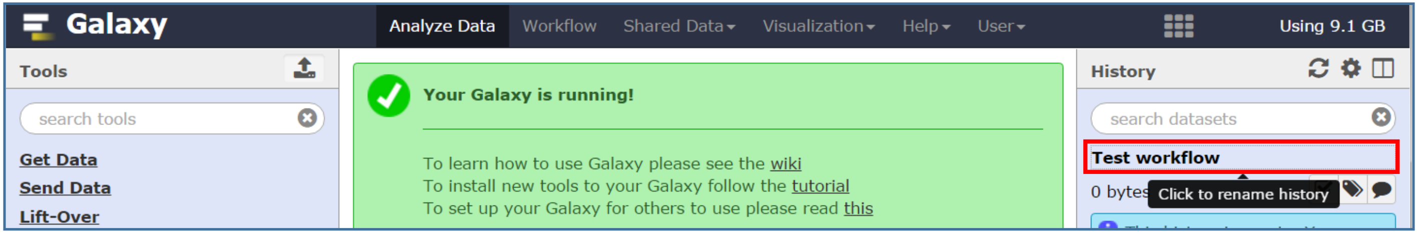
Running workflows

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.

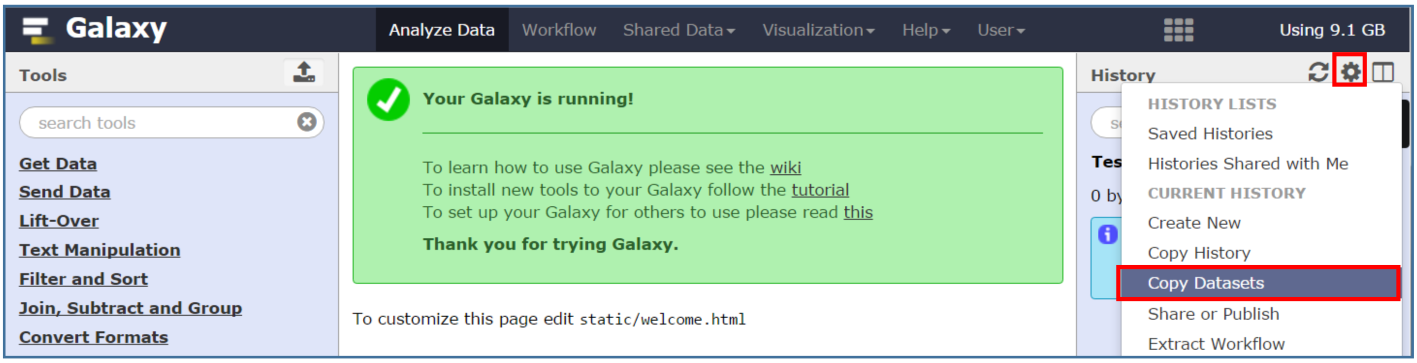
1. In the history pane click the cog icon and select 'Create New' to make a new history.



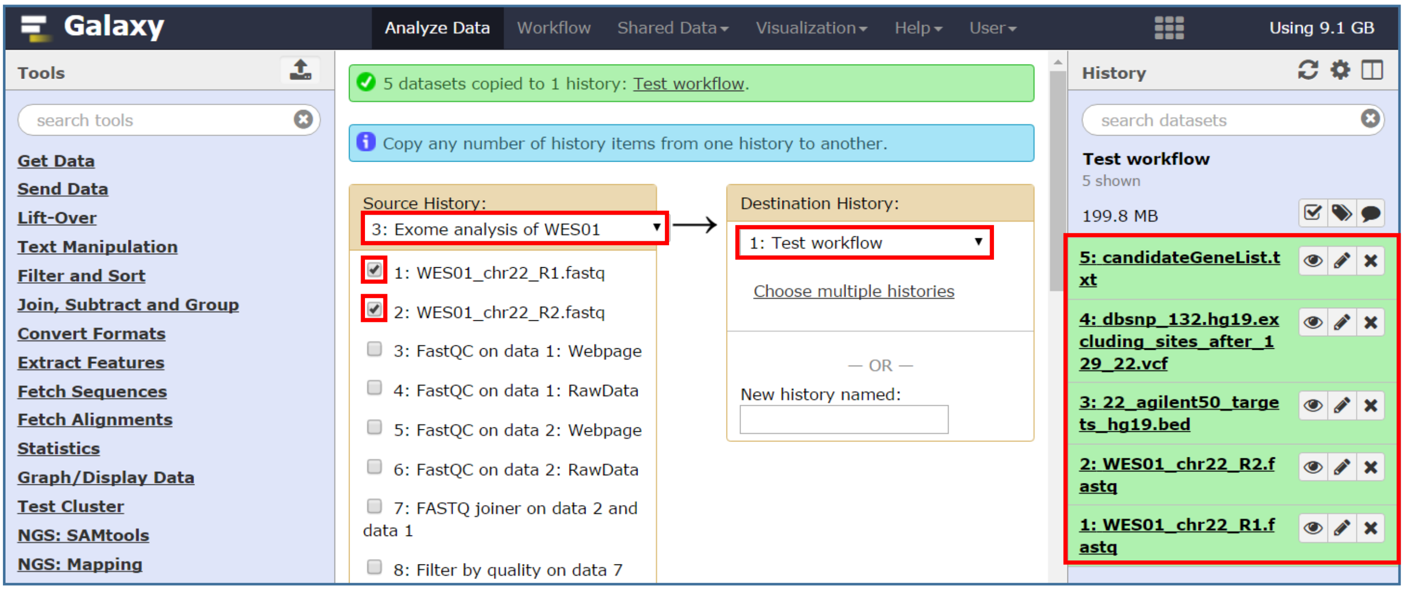
2. Rename the history.



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



4. Select ‘Exome analysis of WES01’ as the source history and the new ‘Test workflow’ 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 and list of candidate genes: candidateGeneList.txt). Scroll to the bottom of the page and click ‘Copy History Items’.



5 datasets copied to 1 history: Test workflow.

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

Source History: 3: Exome analysis of WES01

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

Destination History: 1: Test workflow

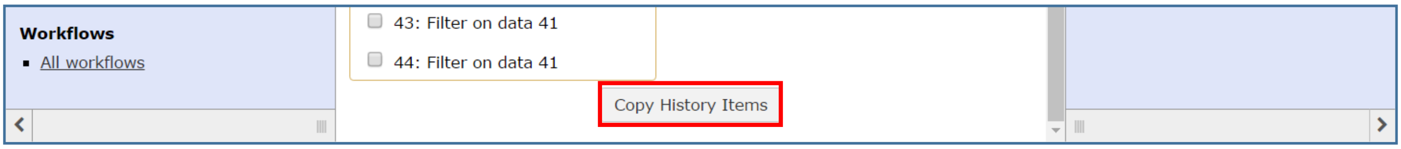
Choose multiple histories

— OR —

New history named:

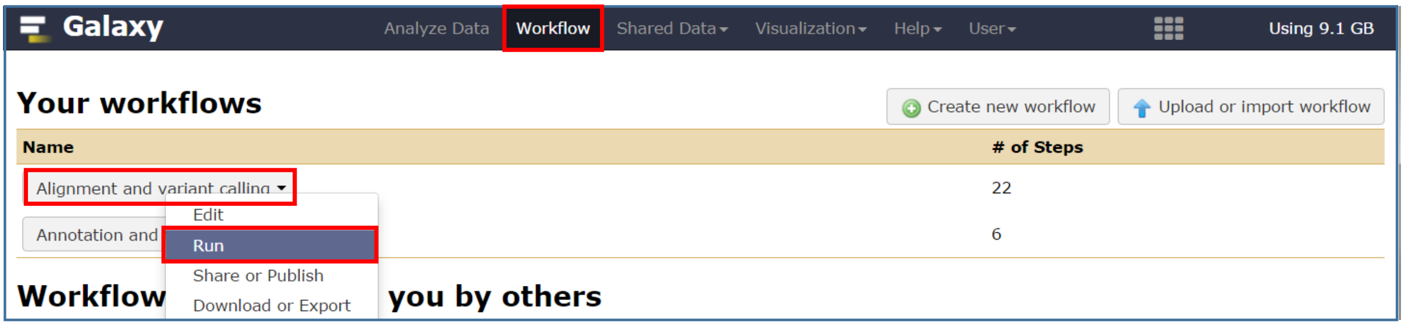
History: Test workflow, 5 shown, 199.8 MB

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



□ 43: Filter on data 41
□ 44: Filter on data 41

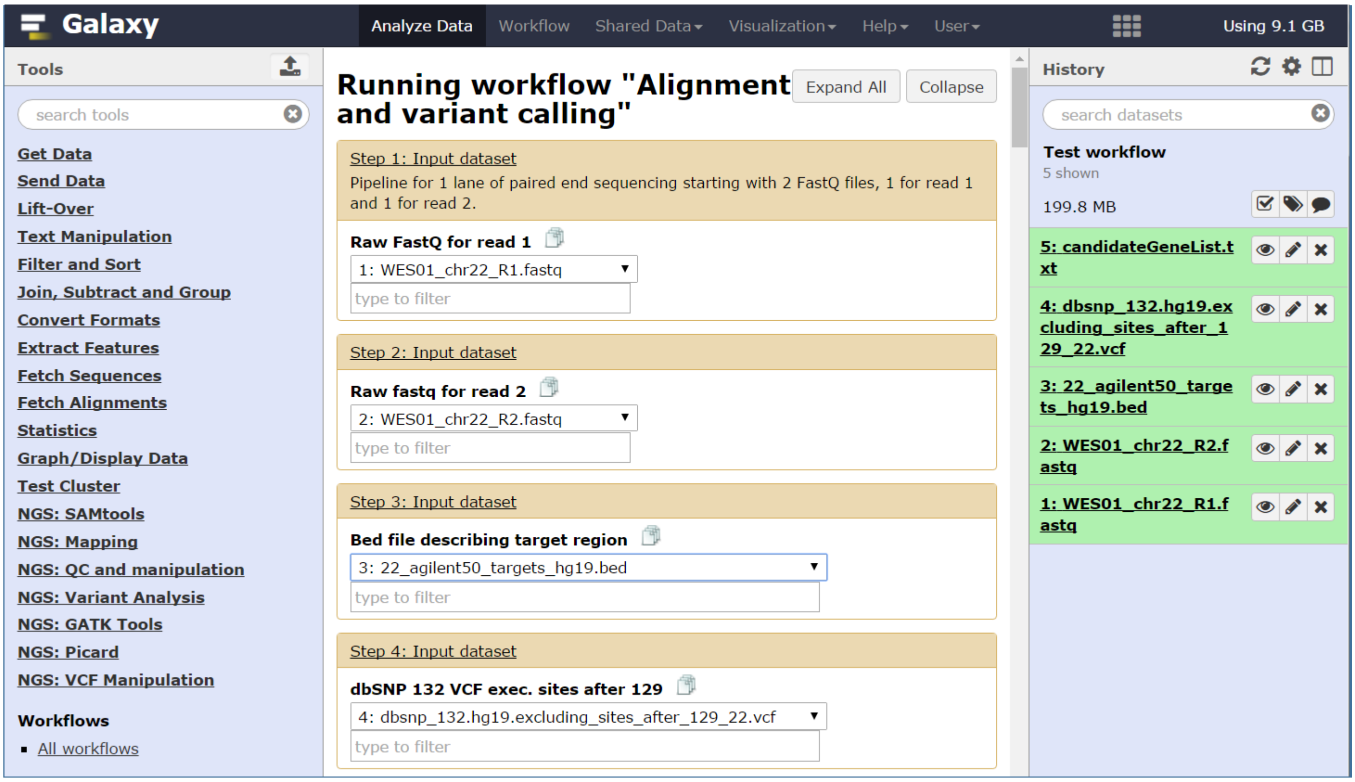
5. Click the workflow tab then click your ‘Alignment and variant calling’ workflow and select run from the drop down menu.



Your workflows

Name	# of Steps
Alignment and variant calling	22
Annotation and	6

6. Now choose the relevant input files and click ‘Run workflow’ to analyse the data.



Running workflow "Alignment and variant calling"

Step 1: Input dataset
Pipeline for 1 lane of paired end sequencing starting with 2 FastQ files, 1 for read 1 and 1 for read 2.

Raw FastQ for read 1
1: WES01_chr22_R1.fastq
type to filter

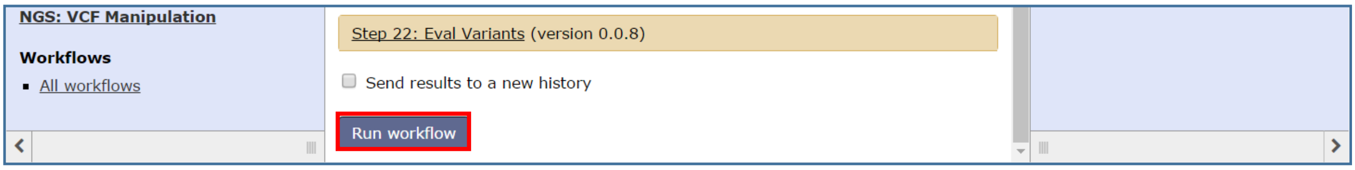
Step 2: Input dataset
Raw fastq for read 2
2: WES01_chr22_R2.fastq
type to filter

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

Step 4: Input dataset
dbSNP 132 VCF exec. sites after 129
4: dbsnp_132.hg19.excluding_sites_after_129_22.vcf
type to filter

Test workflow
5 shown
199.8 MB

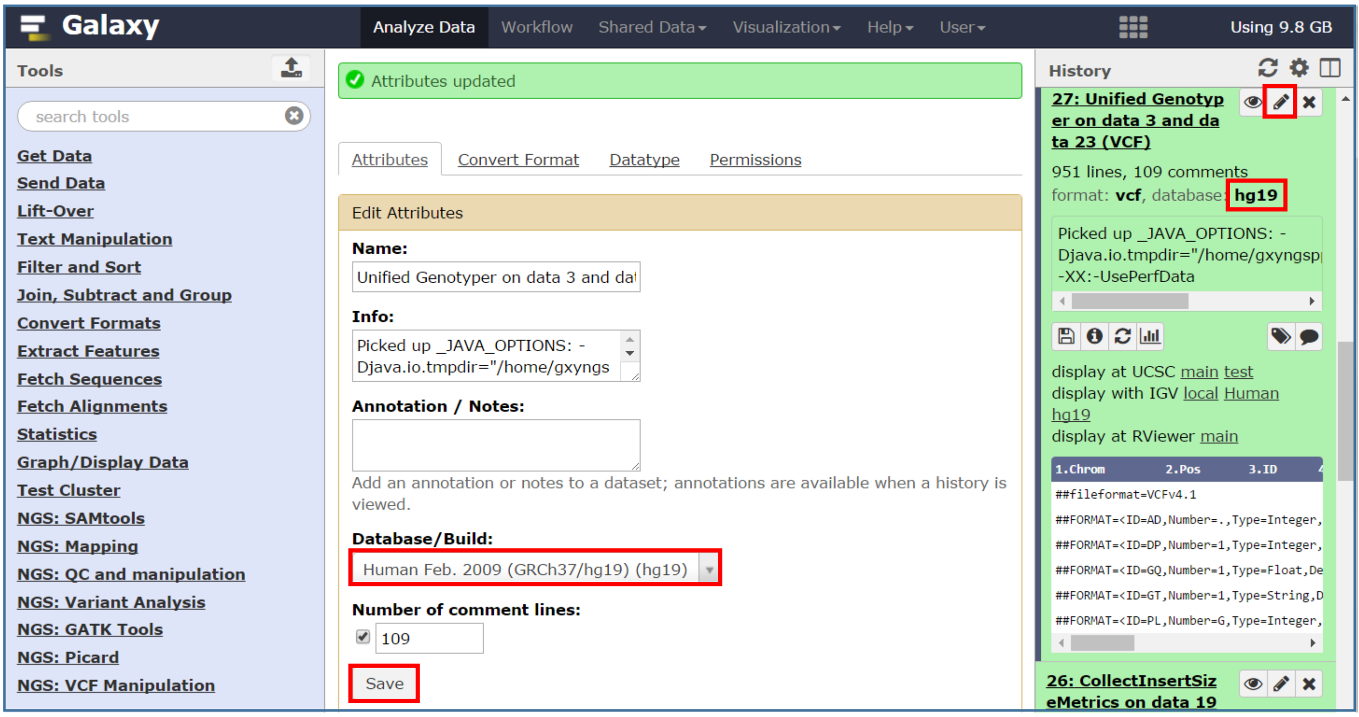
5: candidateGeneList.txt
4: dbsnp_132.hg19.excluding_sites_after_129_22.vcf
3: 22_agilent50_targets_hg19.bed
2: WES01_chr22_R2.fastq
1: WES01_chr22_R1.fastq



Step 22: Eval Variants (version 0.0.8)

☐ Send results to a new history

7. 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 dat

Info:
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir="/home/gxyngs

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

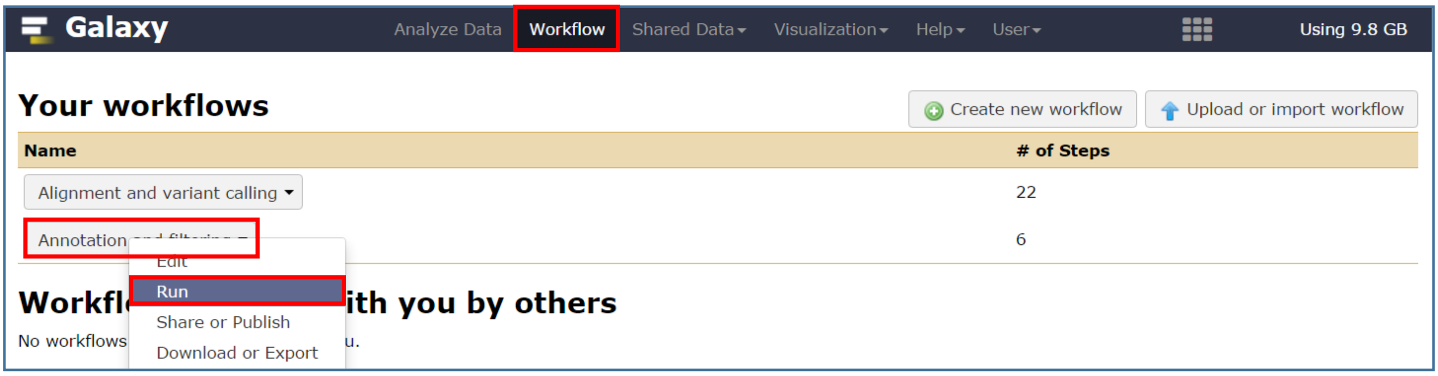
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir="/home/gxyngsp -XX:-UsePerfData

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

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

26: CollectInsertSizeMetrics on data 19

8. When all of the Alignment and variant calling steps have finished and the VCF database has been changed from hg_g1k_v37 to hg19, click the workflow tab then click your 'Annotation and filtering' workflow and select run from the drop down menu.



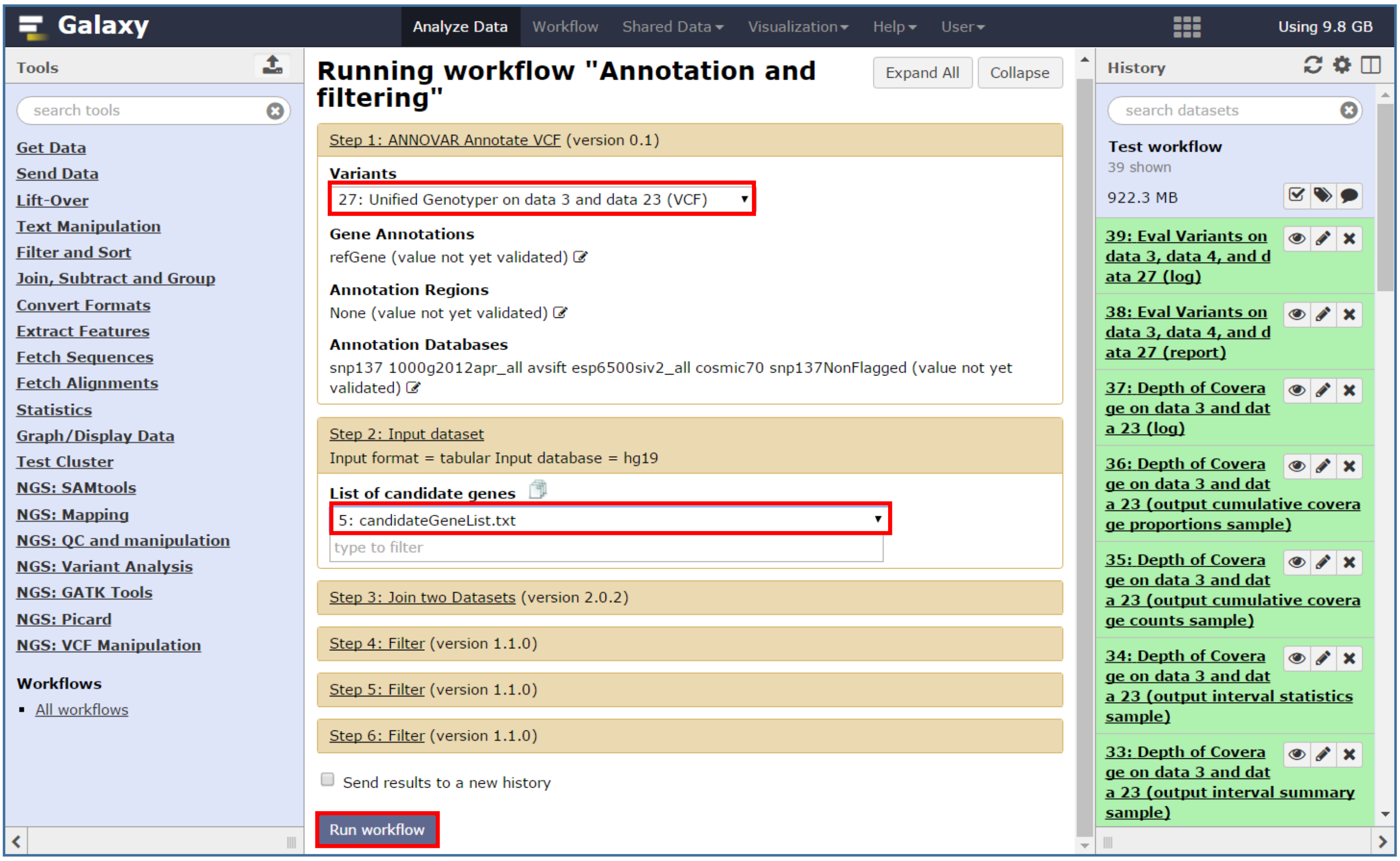
Your workflows

Name	# of Steps
Alignment and variant calling	22
Annotation and filtering	6

Workfl... ...ith you by others

No workflows ... u.

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


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 9.8 GB
Tools
search tools
Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
Test Cluster
NGS: SAMtools
NGS: Mapping
NGS: QC and manipulation
NGS: Variant Analysis
NGS: GATK Tools
NGS: Picard
NGS: VCF Manipulation
Workflows
All workflows
Expand All
Collapse
Run workflow
History
search datasets
Test workflow
39 shown
922.3 MB
39: Eval Variants on data 3, data 4, and data 27 (log)
38: Eval Variants on data 3, data 4, and data 27 (report)
37: Depth of Coverage on data 3 and data 23 (log)
36: Depth of Coverage on data 3 and data 23 (output cumulative coverage proportions sample)
35: Depth of Coverage on data 3 and data 23 (output cumulative coverage counts sample)
34: Depth of Coverage on data 3 and data 23 (output interval statistics sample)
33: Depth of Coverage on data 3 and data 23 (output interval summary sample)

Running workflow "Annotation and filtering"

Step 1: ANNOVAR Annotate VCF (version 0.1)

Variants
27: Unified Genotyper on data 3 and data 23 (VCF)

Gene Annotations
refGene (value not yet validated)

Annotation Regions
None (value not yet validated)

Annotation Databases
snp137 1000g2012apr_all avsift esp6500siv2_all cosmic70 snp137NonFlagged (value not yet validated)

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

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

Step 3: Join two Datasets (version 2.0.2)

Step 4: Filter (version 1.1.0)

Step 5: Filter (version 1.1.0)

Step 6: Filter (version 1.1.0)

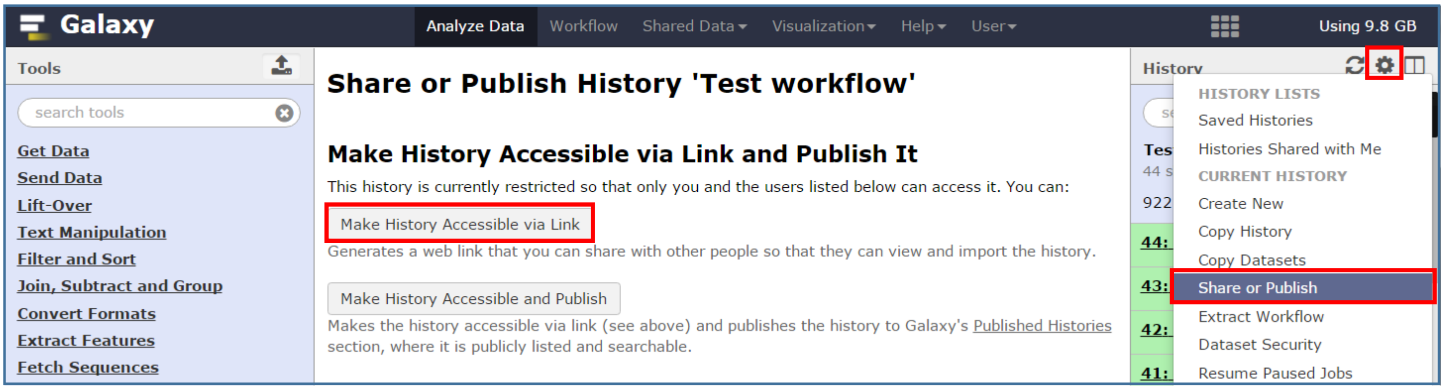
□ Send results to a new history

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 'Test workflow'

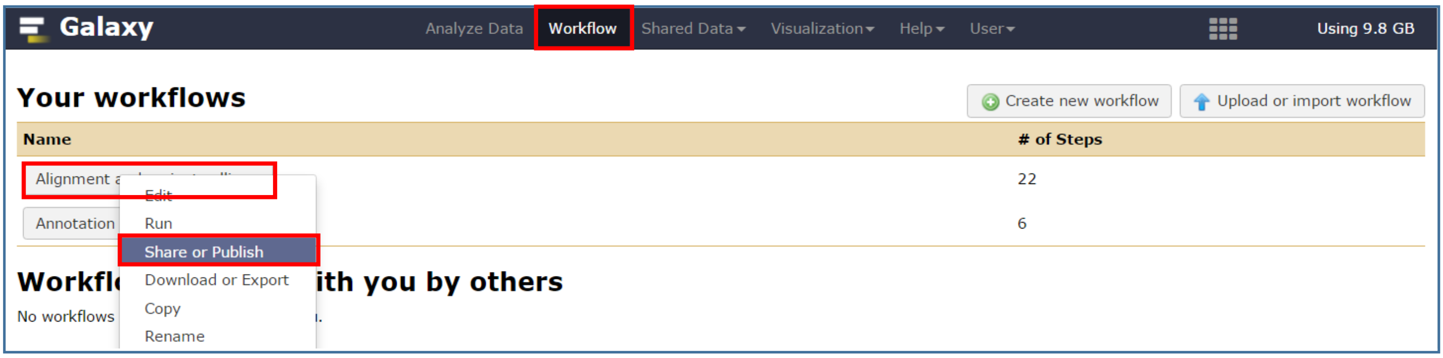
Make History Accessible via Link and Publish It

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..



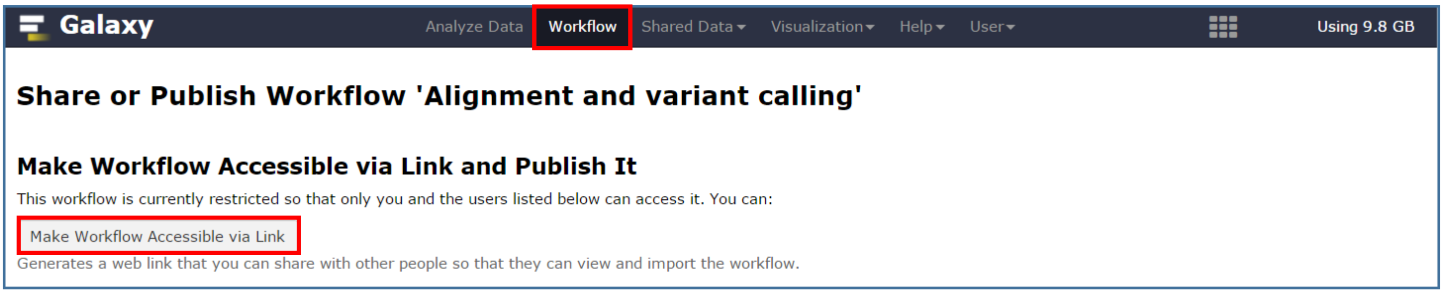
Your workflows

Name	# of Steps
Alignment a	22
Annotation	6

Workfl ith you by others

No workflows .

3. Click ‘Make Workflow 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 'Alignment and variant calling'

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 you finished the exercise!