**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

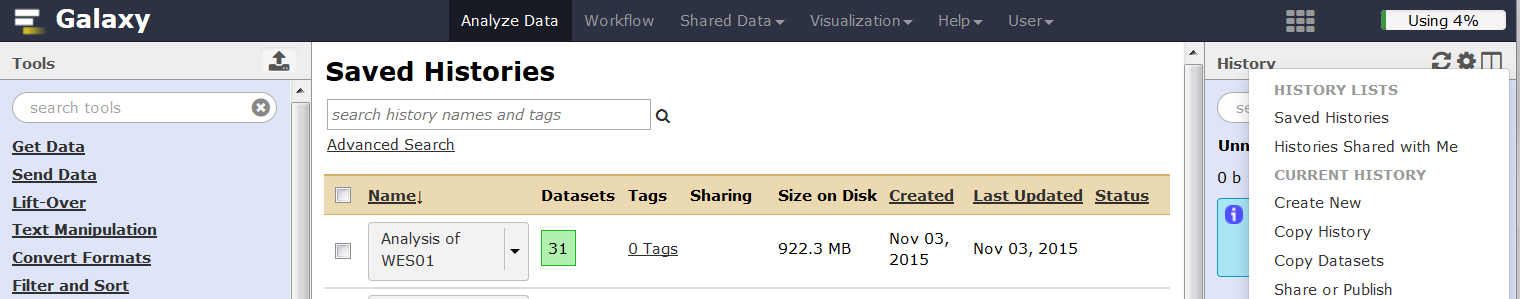
**4.** Investigate phenotype-genotype relationships using OMIM, ClinVar, and PubMed

**5.** Create and apply a workflow from your Galaxy history

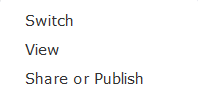
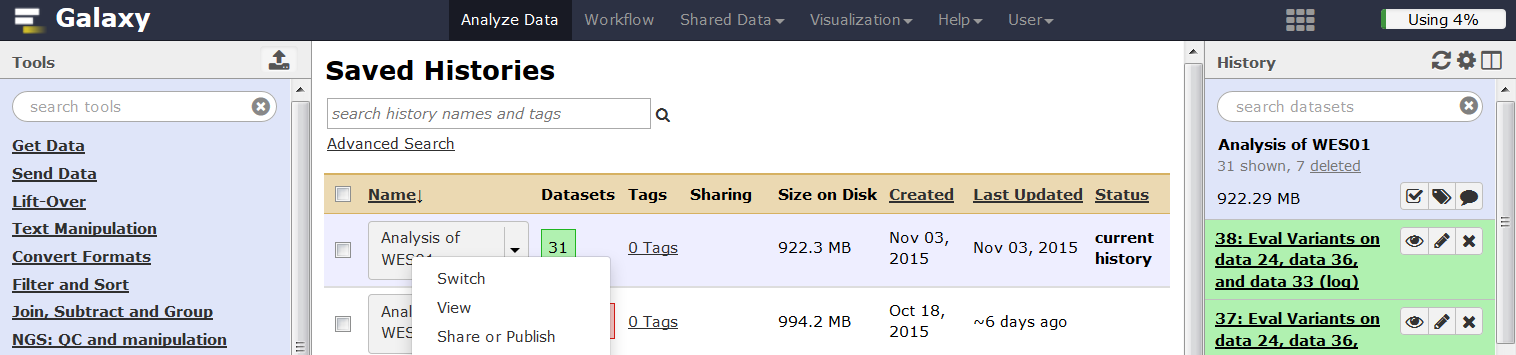
**Let’s begin**

Open the saved history for the analysis of WES01.

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



**2.** Click the saved history for ‘Analysis of WES01’ and select switch



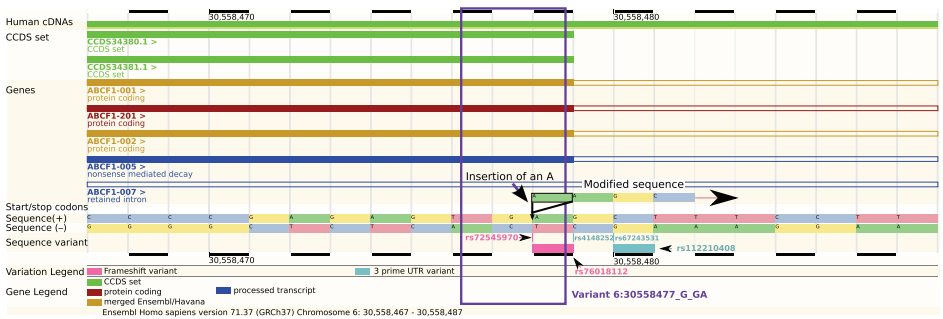
When the correct history has been loaded, the history pane should be titled ‘Analysis of WES01’.

**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](http://annovar.openbioinformatics.org/en/latest/) (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](http://www.1000genomes.org/), [dbSNP](http://www.ncbi.nlm.nih.gov/SNP/), [Exome Sequencing Project](http://evs.gs.washington.edu/EVS/) and [COSMIC](http://cancer.sanger.ac.uk/cosmic)) and predictions of functional significance ([avsift](http://sift.jcvi.org/) and [conservation scores](http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/README.txt)).

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](http://snpeff.sourceforge.net/) Cingolani et al 2012 and [Variant Effect Predictor VEP](http://grch37.ensembl.org/Homo_sapiens/Tools/VEP) McLaren et al 2010) and definition of transcripts ([Ensemble](http://www.ensembl.org/Homo_sapiens/Info/Annotation) Flicek et al 2012, [RefSeq](http://www.ncbi.nlm.nih.gov/refseq/about/) Pruitt et al 2012, [Gencode](http://www.gencodegenes.org/about.html) 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](http://annovar.openbioinformatics.org/en/latest/user-guide/gene/#output-file-2-refseq-gene-annotation)) 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).

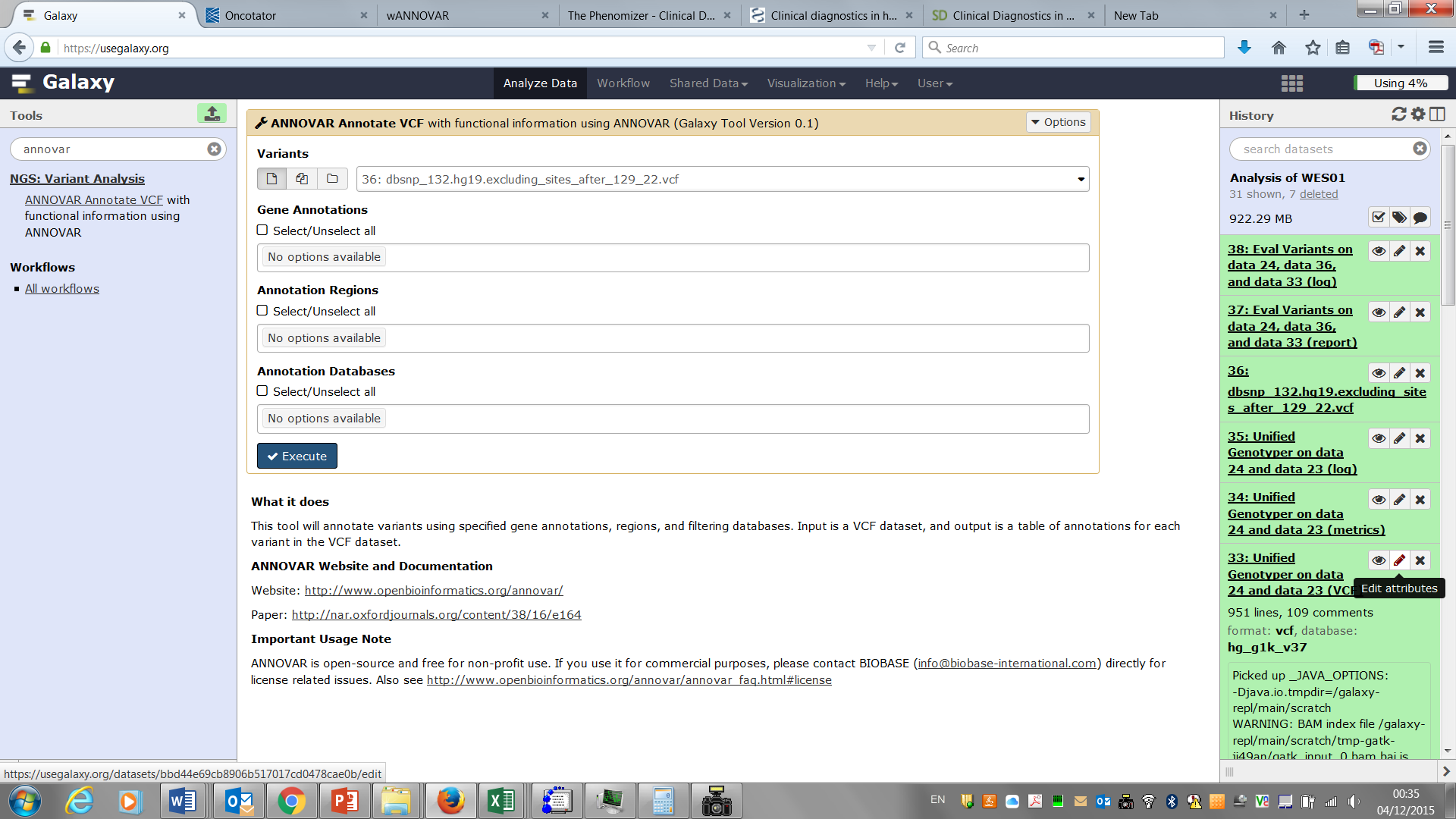
**Figure 1.** Annotation examples from McCarthy et al 2014



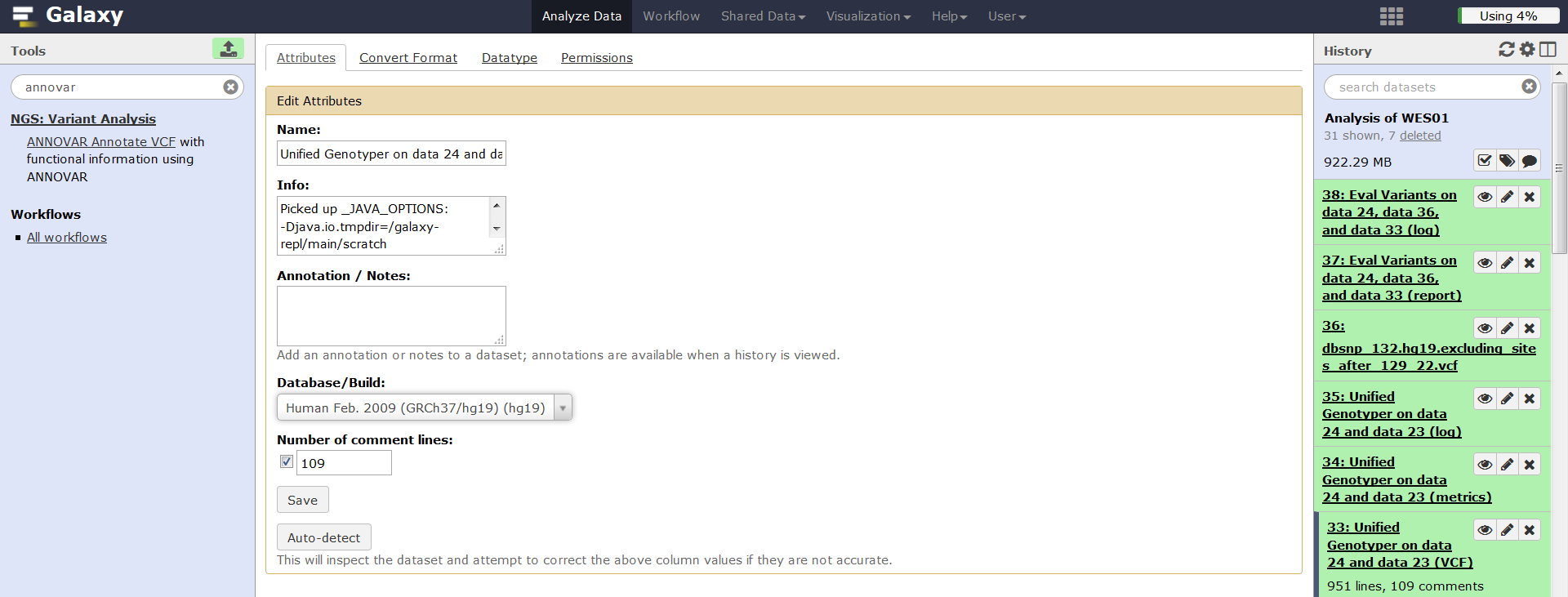
In figure 1, the variant is an insertion of an A in a stop codon (TGA) in the penultimate base of the exon in all transcripts except one. It is possible to interpret this variant as a frameshift insertion or a stop-loss but it is actually a synonymous variant because the stop codon remains in the same 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](http://www.sanger.ac.uk/science/projects/manual-annotation)), 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.

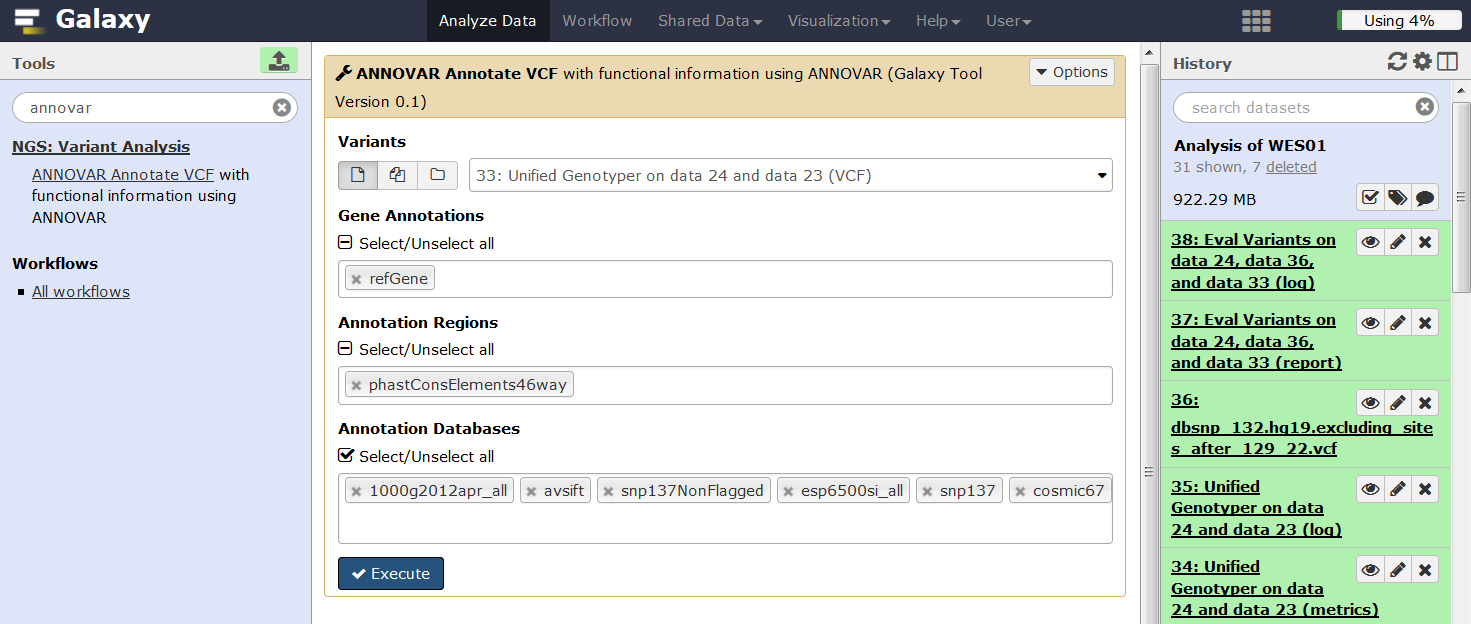


**2.** Select the hg19 Database/Build and click save.

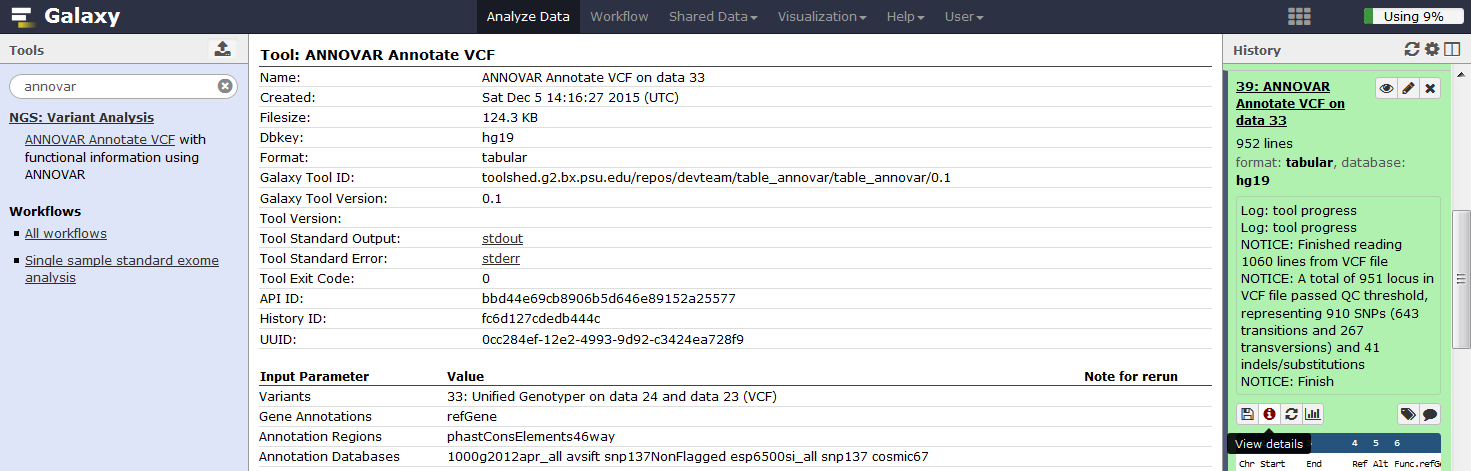


**3.** In **Tool Pane**: Go to **NGS: Variant Analysis** > ANNOVAR Annotate VCF

For Gene Annotation we will use [refGene](http://varianttools.sourceforge.net/Annotation/RefGene) only, which is based on refSeq, since this is the most curated dataset with fewest errors. [Segmental duplications](http://blog.goldenhelix.com/bchristensen/why-you-should-care-about-segmental-duplications/) (genomicSuperDups) are associated with false positives but we will ignore them in this analysis. Select all Annotation Databases.

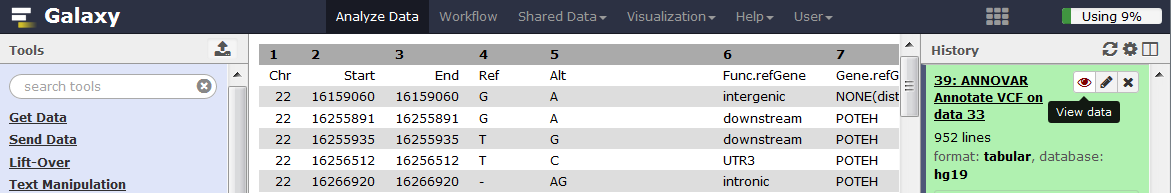


**4.** Before we look at the ANNOVAR results, let’s check the log file. Click on the info icon ‘i’. ANNOVAR prints its log file information to the Tool Standard Error: stderr, click this link to view the log file.



**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.



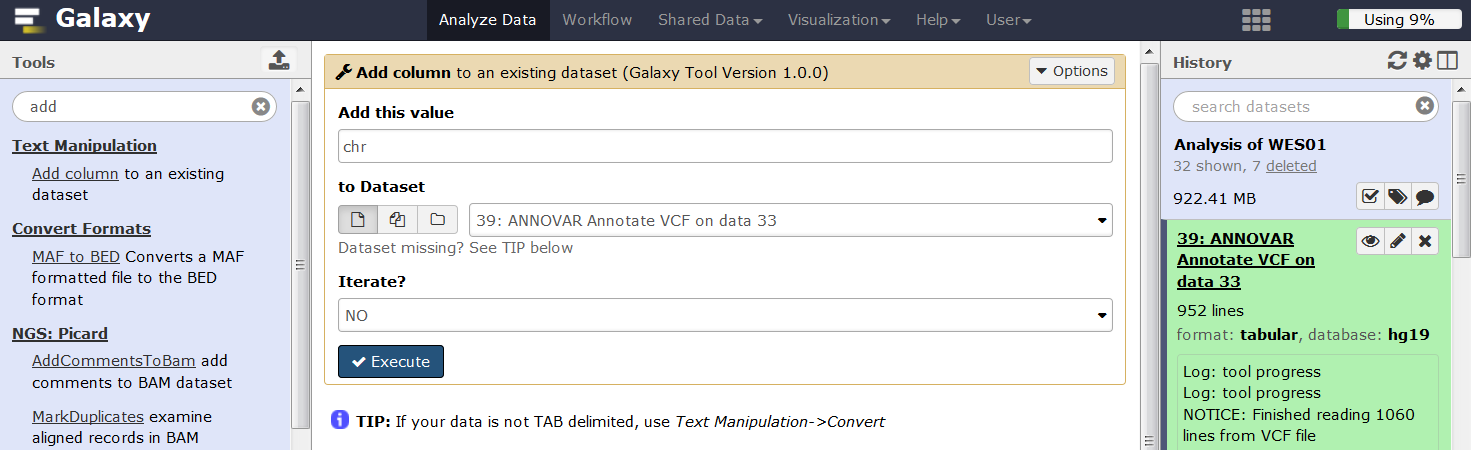
Note that missing data is listed as NA in some columns and -1 in others.

**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.



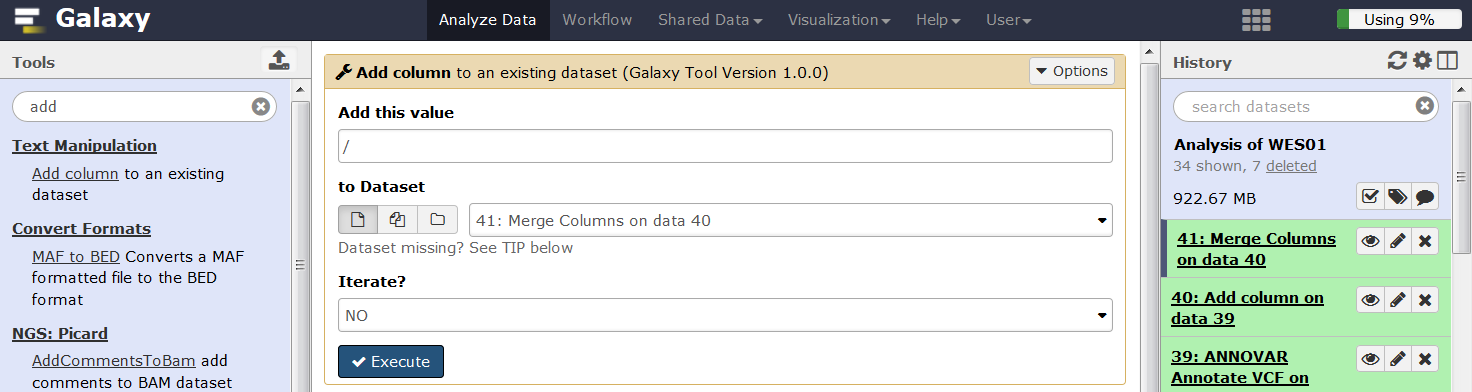
**7.** In **Tool Pane**: Go to **Text Manipulation** > Merge Columns

Merge columns 17 and 1 to make a new column of chromosome locations in the format chr#.



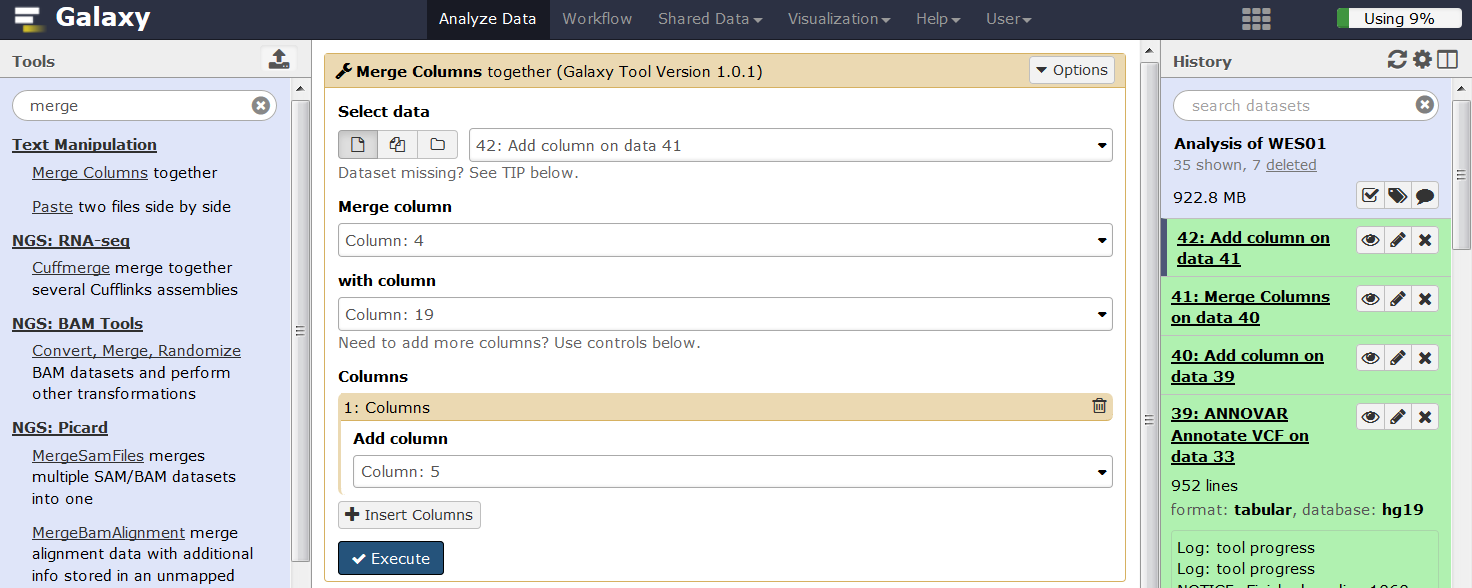
**8.** In **Tool Pane**: Go to **Text Manipulation** > Add column

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.



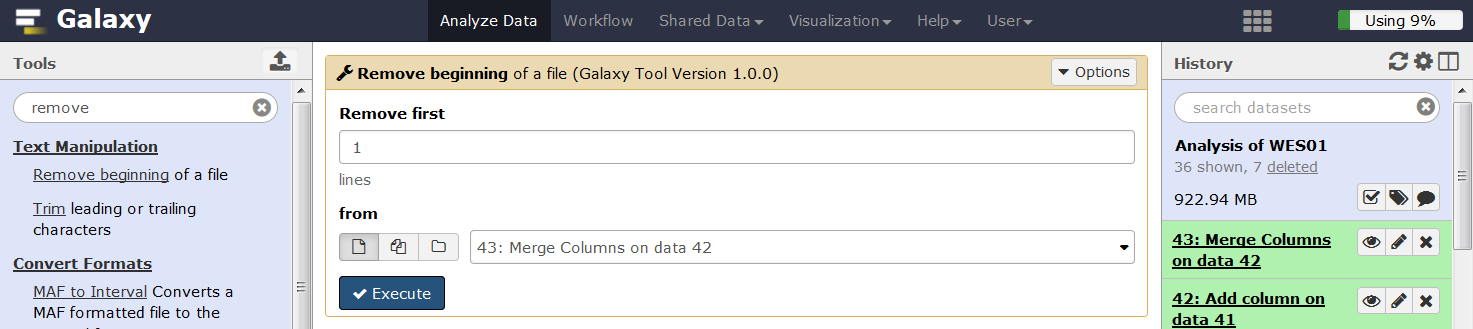
**9.** In **Tool Pane**: Go to **Text Manipulation** > Merge Columns

Merge columns 4 and 19 to make a new column of reference and alternate alleles in the format ref/alt.



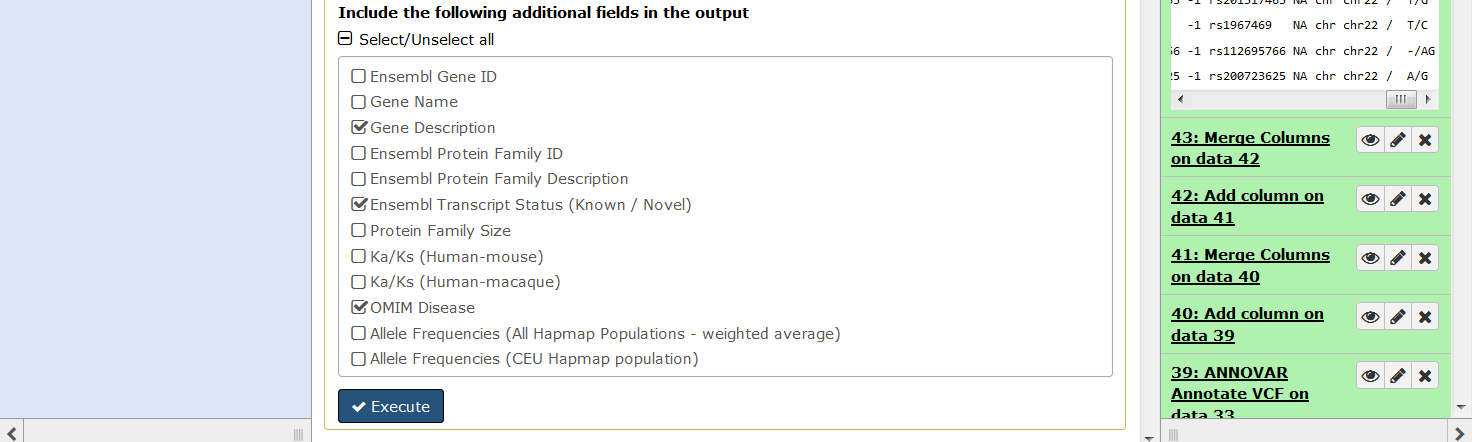
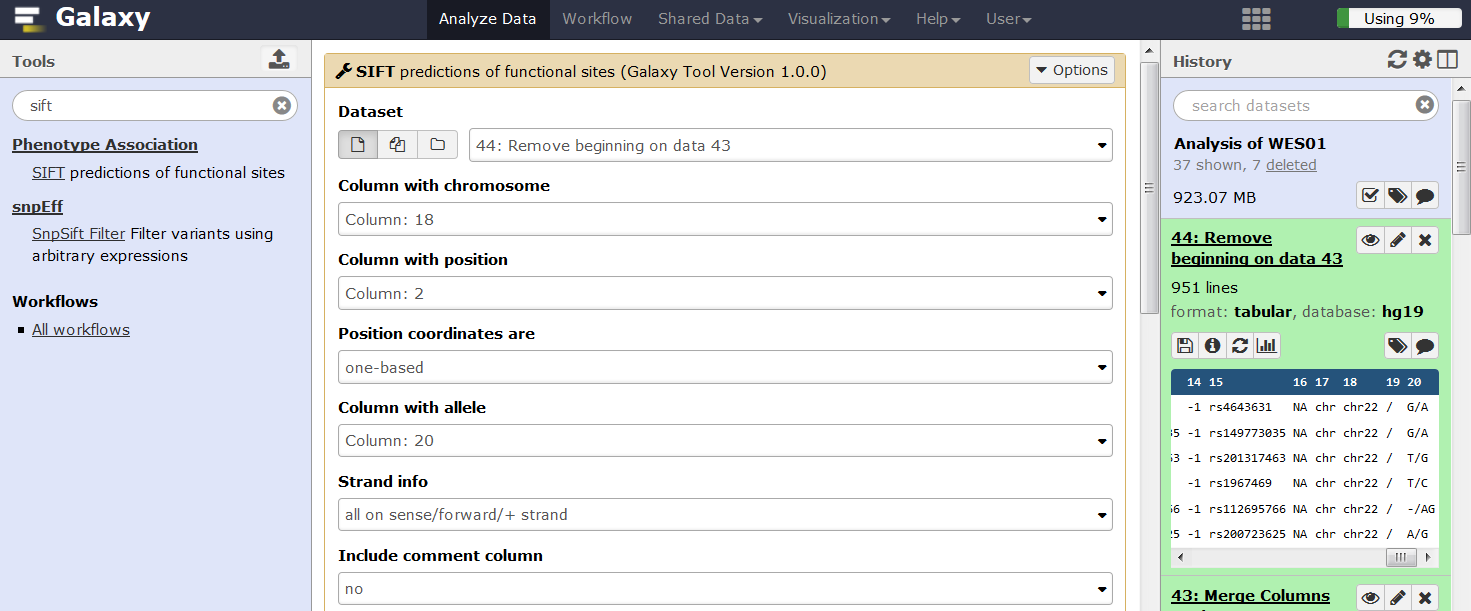
**10.** In **Tool Pane**: Go to **Text Manipulation** > Remove beginning

Delete the first line of the VCF file.



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

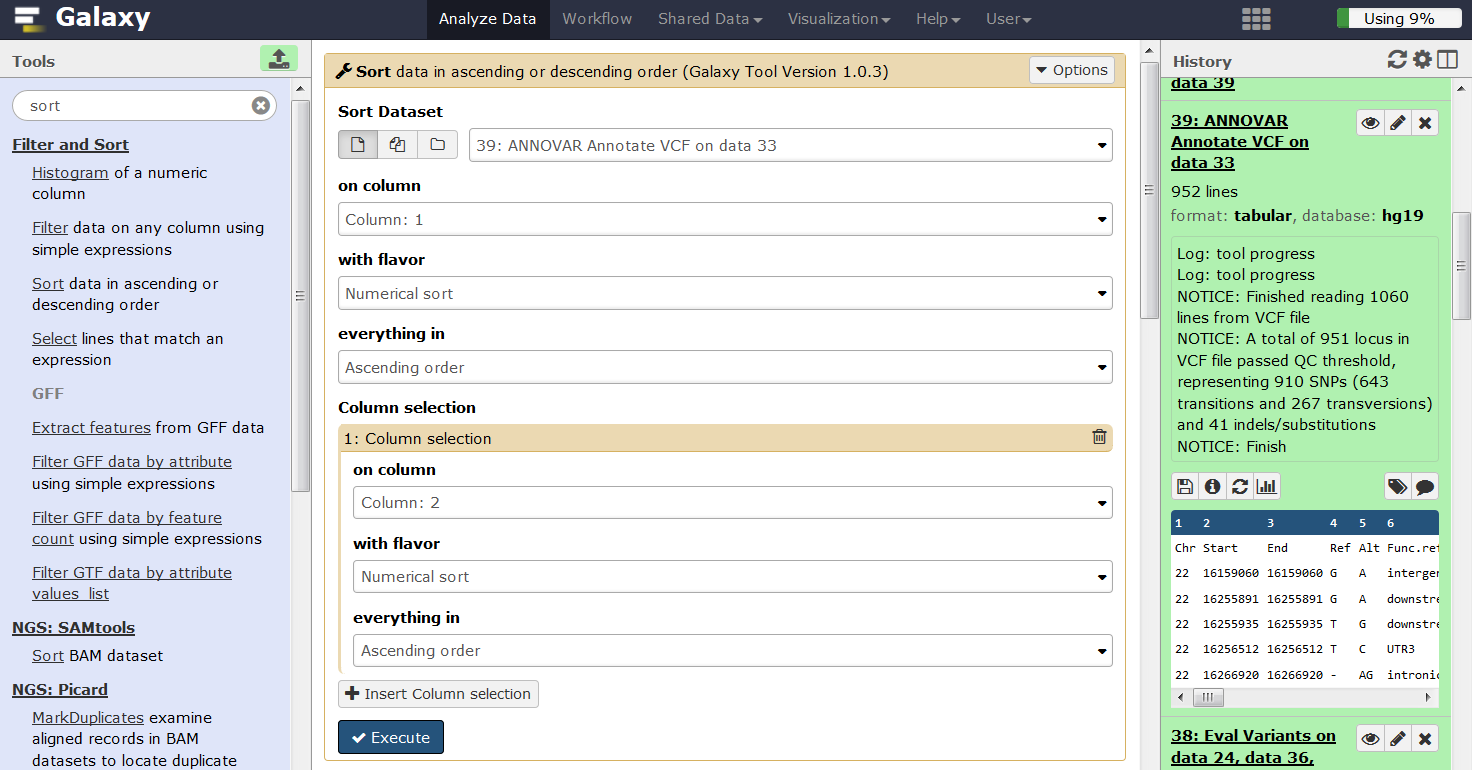
**11.** In **Tool Pane**: Go to **Phenotype Association** > SIFT



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

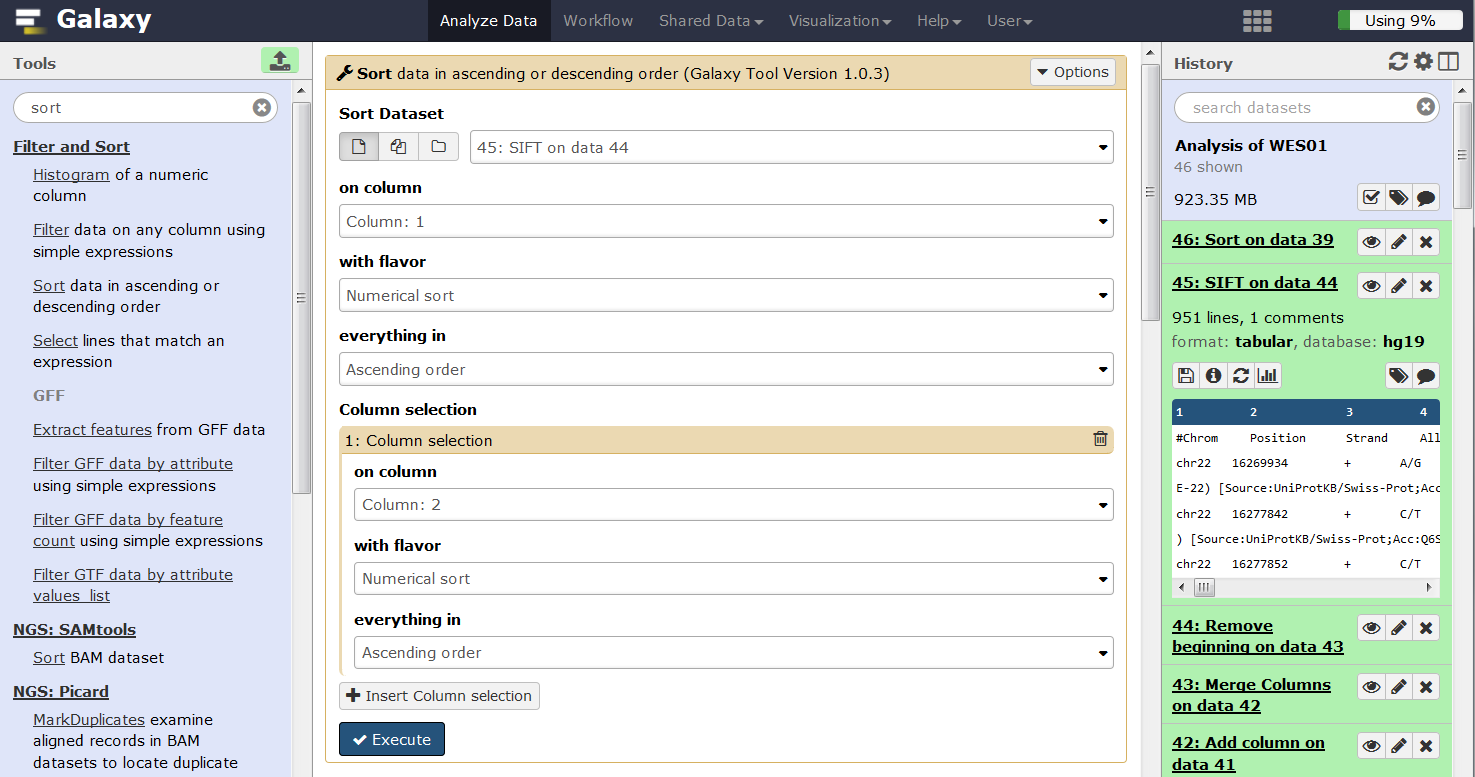
**12.** In **Tool Pane**: Go to **Filter and Sort** > Sort

Sort the ANNOVAR result by chromosome (column 1) and base pair location (column 2) in ascending order.



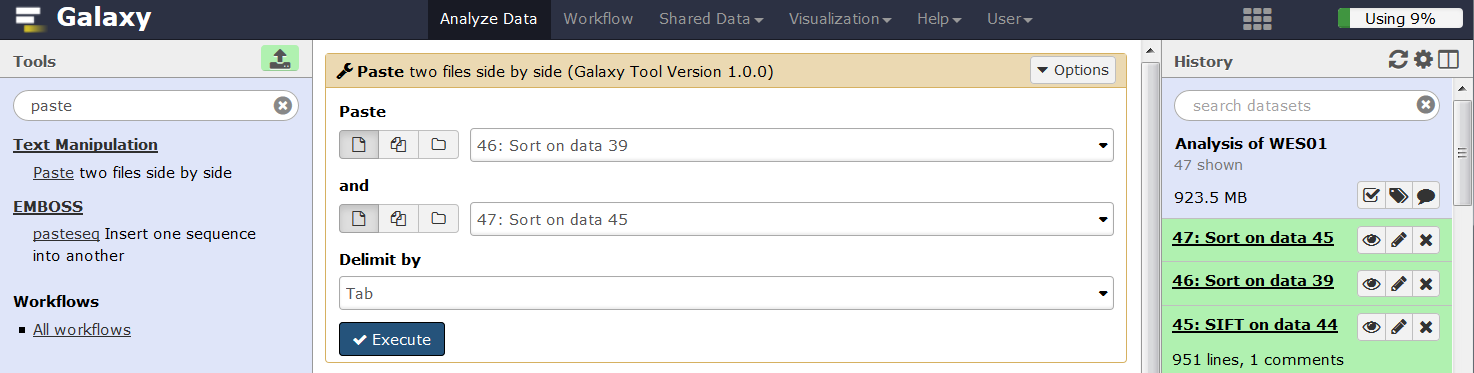
**13.** In **Tool Pane**: Go to **Filter and Sort** > Sort

Sort the SIFT result by chromosome (column 1) and base pair location (column 2) in ascending order.



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

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



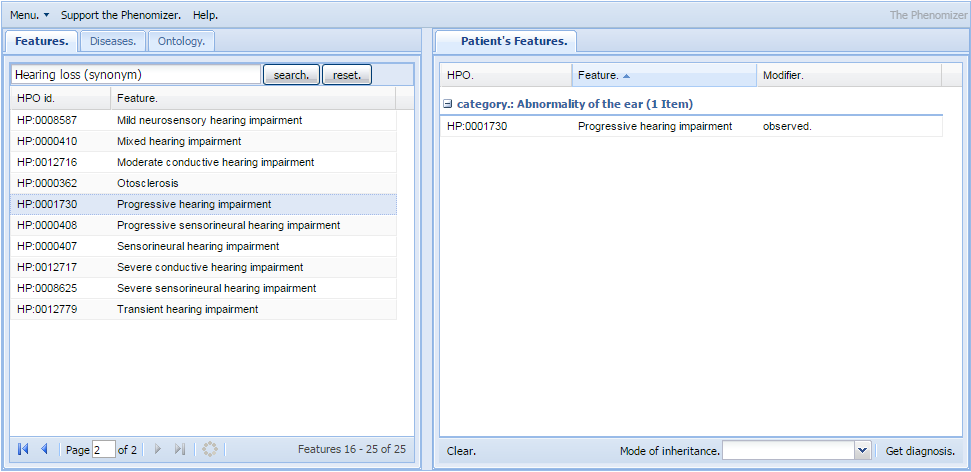
**Generate a list of candidate genes**

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.

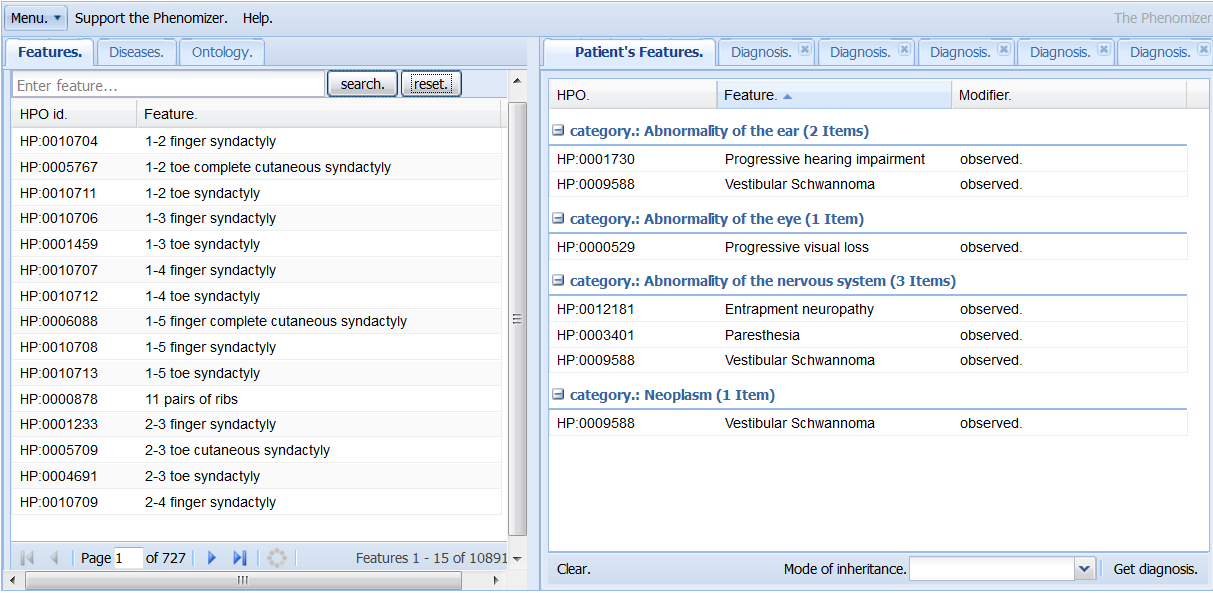


For a definition of specific HPO terms, modify the HPO id at the end of this link:

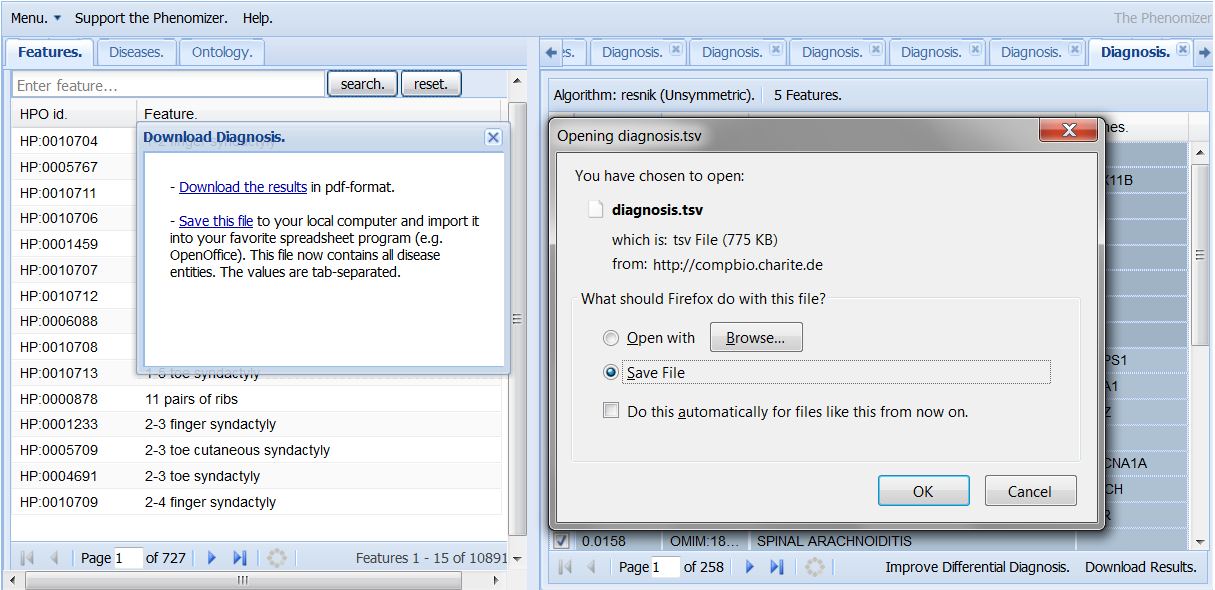
<http://purl.obolibrary.org/obo/HP_0000365>

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’.

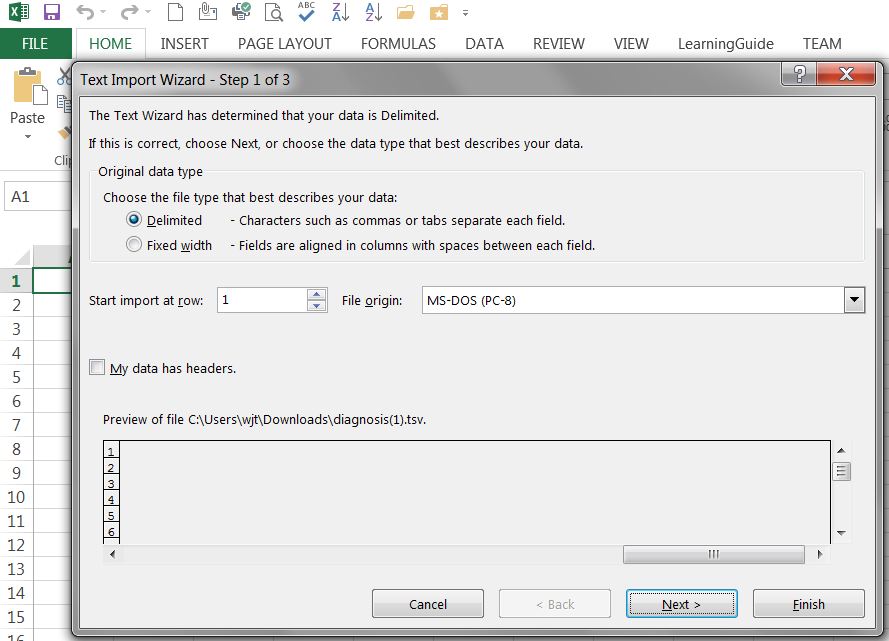
**3.** Having entered all of the patient’s phenotypes click ‘Get diagnosis’ to generate a list of associated diseases and genes.



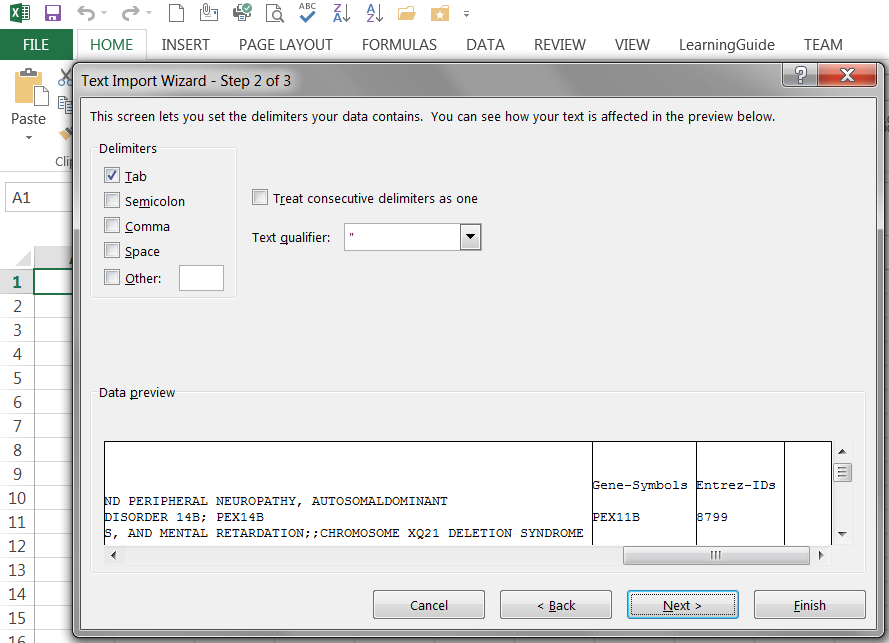
**4.** Click ‘Download Results’, then click the ‘Save this file’ link to save the ‘diagnosis.tsv’ file to your computer.



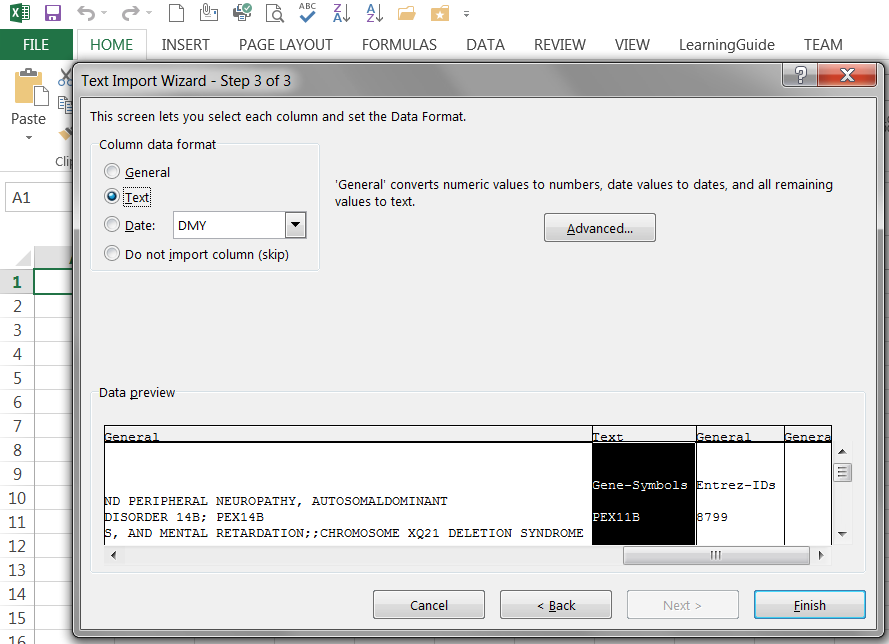
**5.** Open the Phenomizer results file ‘diagnosis.tsv’ in Excel, choose ‘Delimited’ in the Text Import Wizard and click next.



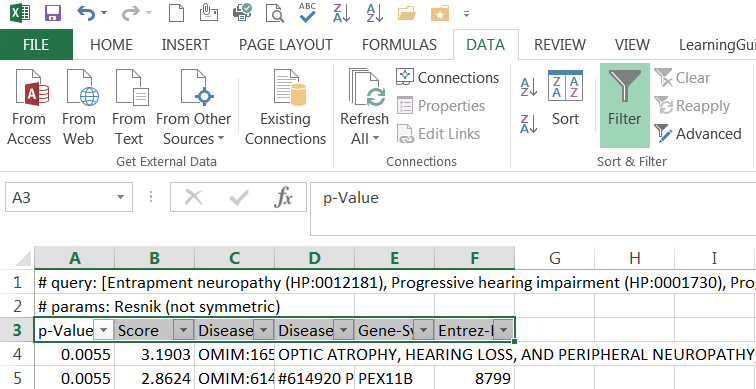
**6.** Choose ‘Tab’ as the delimiter and click next



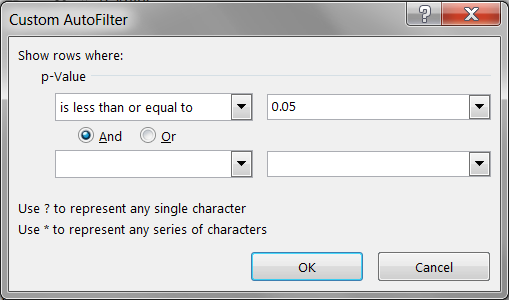
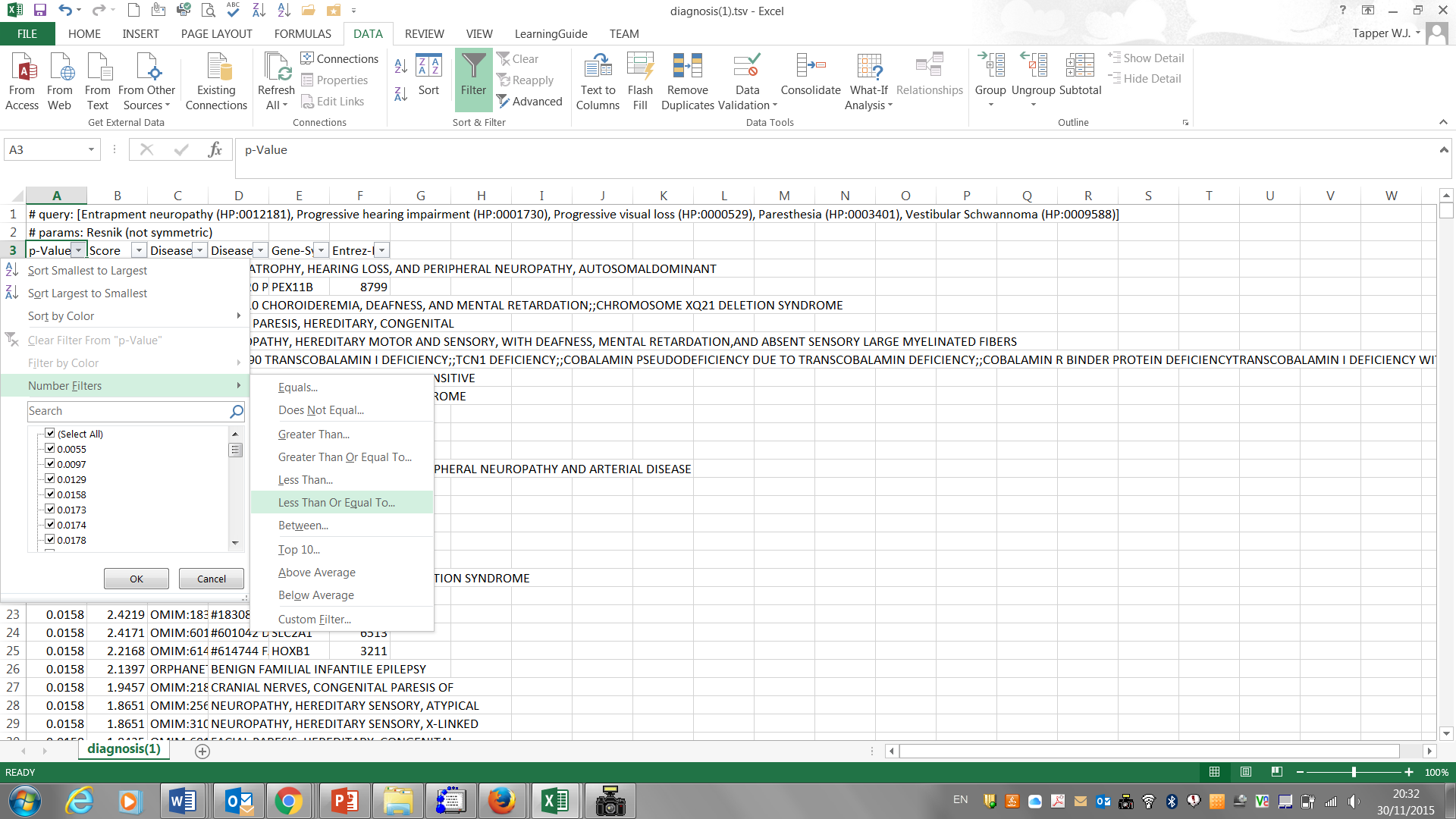
**7.** Use the scroll bar to navigate to the Gene-Symbol column and select ‘Text’ as the column data format, then click Finish. Setting ‘Text’ as the data format for gene symbols is important because it will prevent Excel from automatically interpreting some gene names as dates, eg ‘SEPT12’ will otherwise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



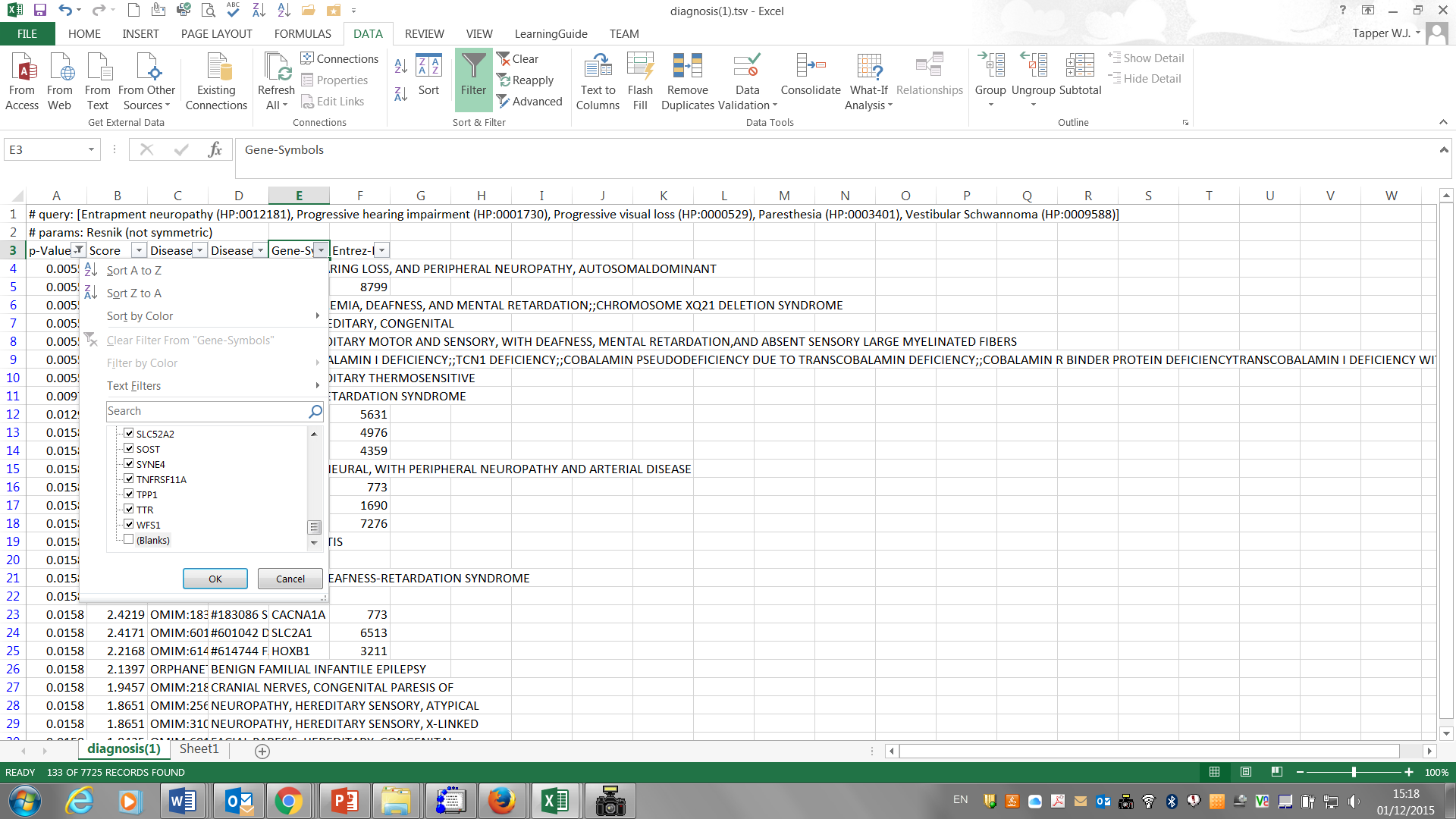
**8.** Add filtersto the columns by selecting the column titles in row 3, hold down left click and mouse over all 6 columns, then select Filter. This will add a drop down box to each column title that enables filtering and sorting based on the column contents.



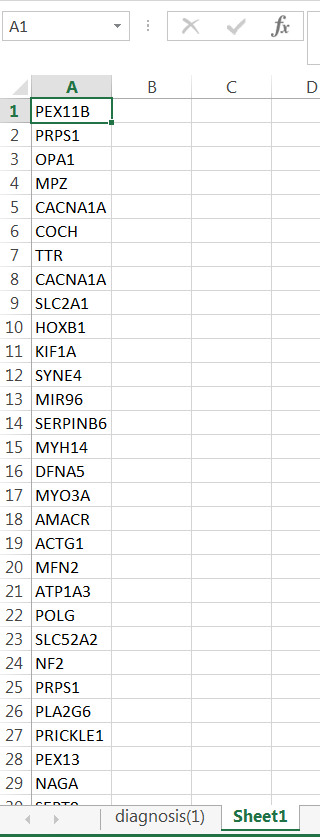
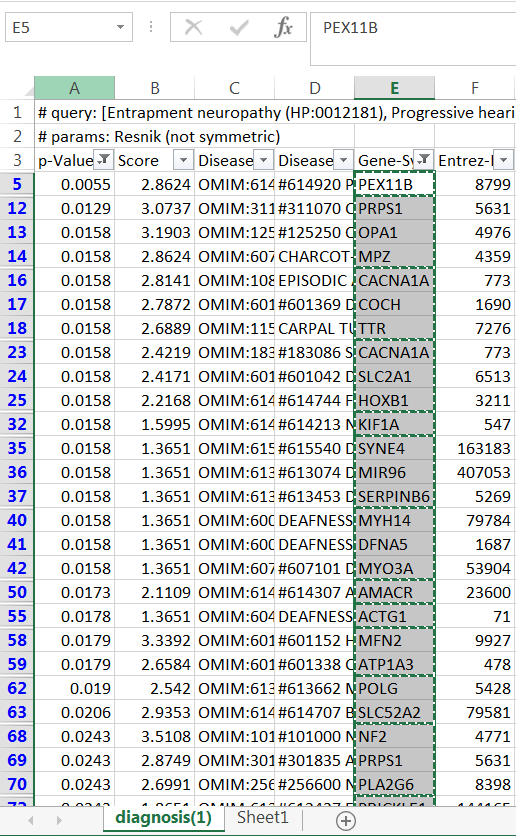
**9.** Use the p-value filter to select genes that are associated with the phenotype with a p-value less than or equal to 0.05.



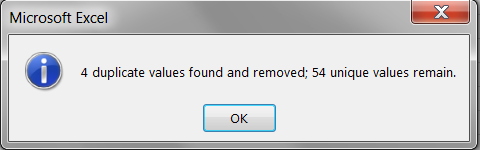
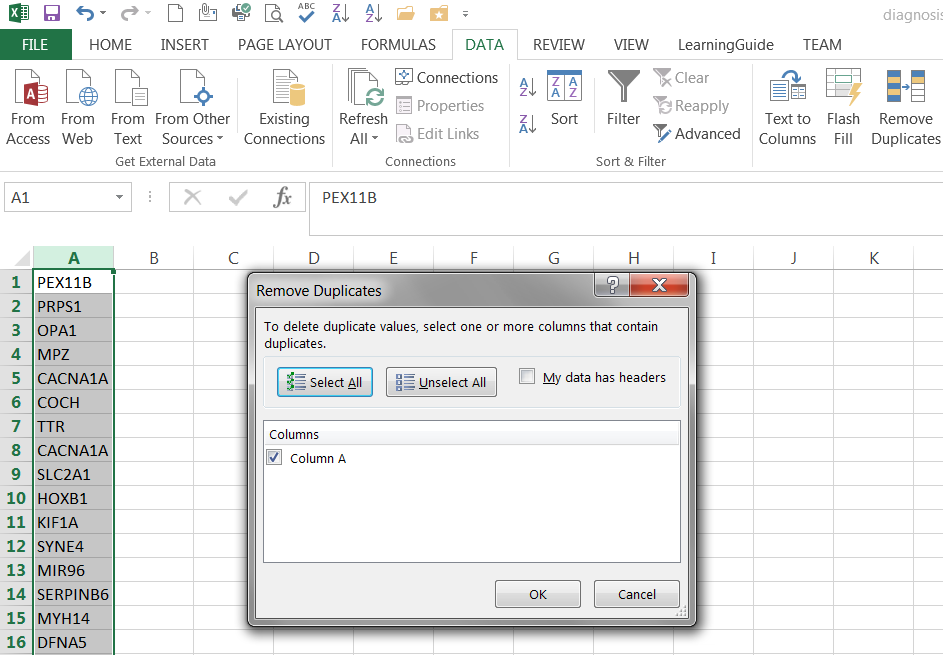
**10.** Now click on the ‘Gene-Symbol’ filter, scroll to the bottom of the list and uncheck the ‘(Blanks)’ box to exclude rows that do not have a gene symbol.



**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.

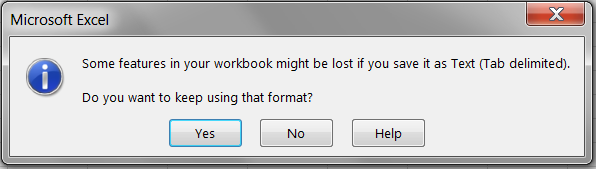
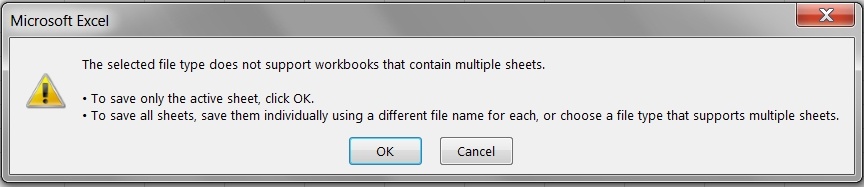
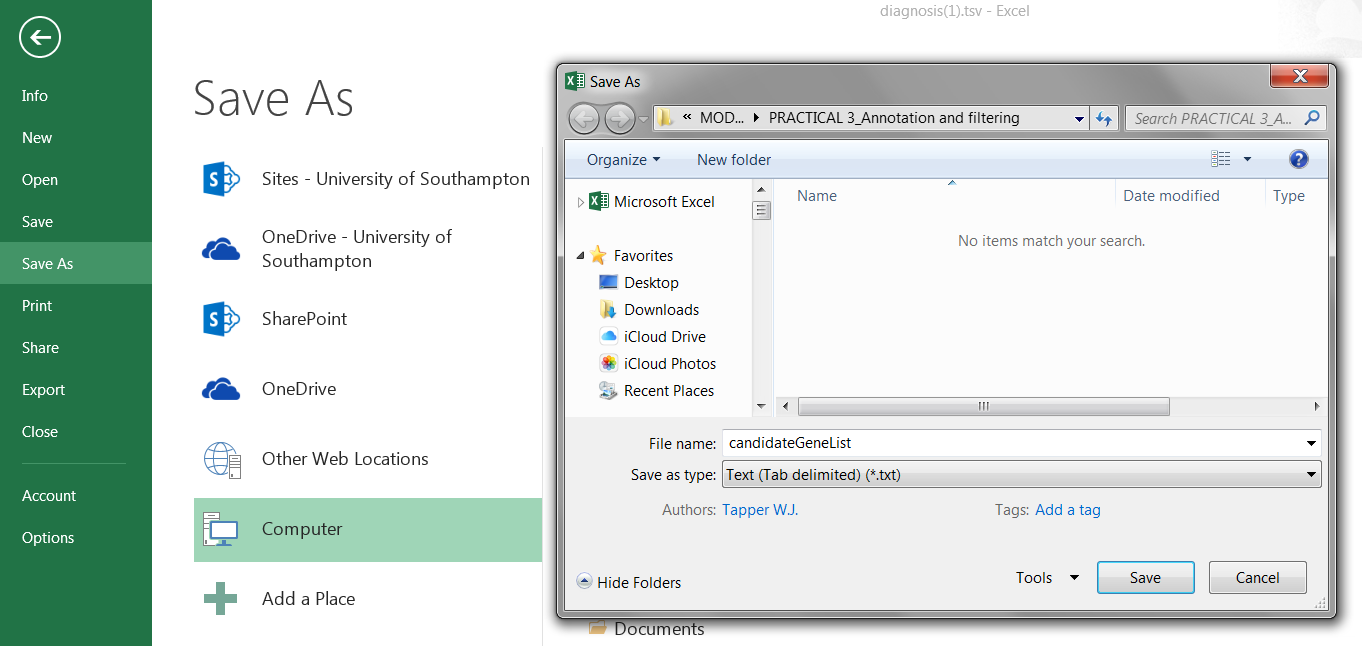


**12.** Now highlight the gene list (column A in Sheet 1), select ‘Remove Duplicates’ and click OK to generate a unique list of genes.



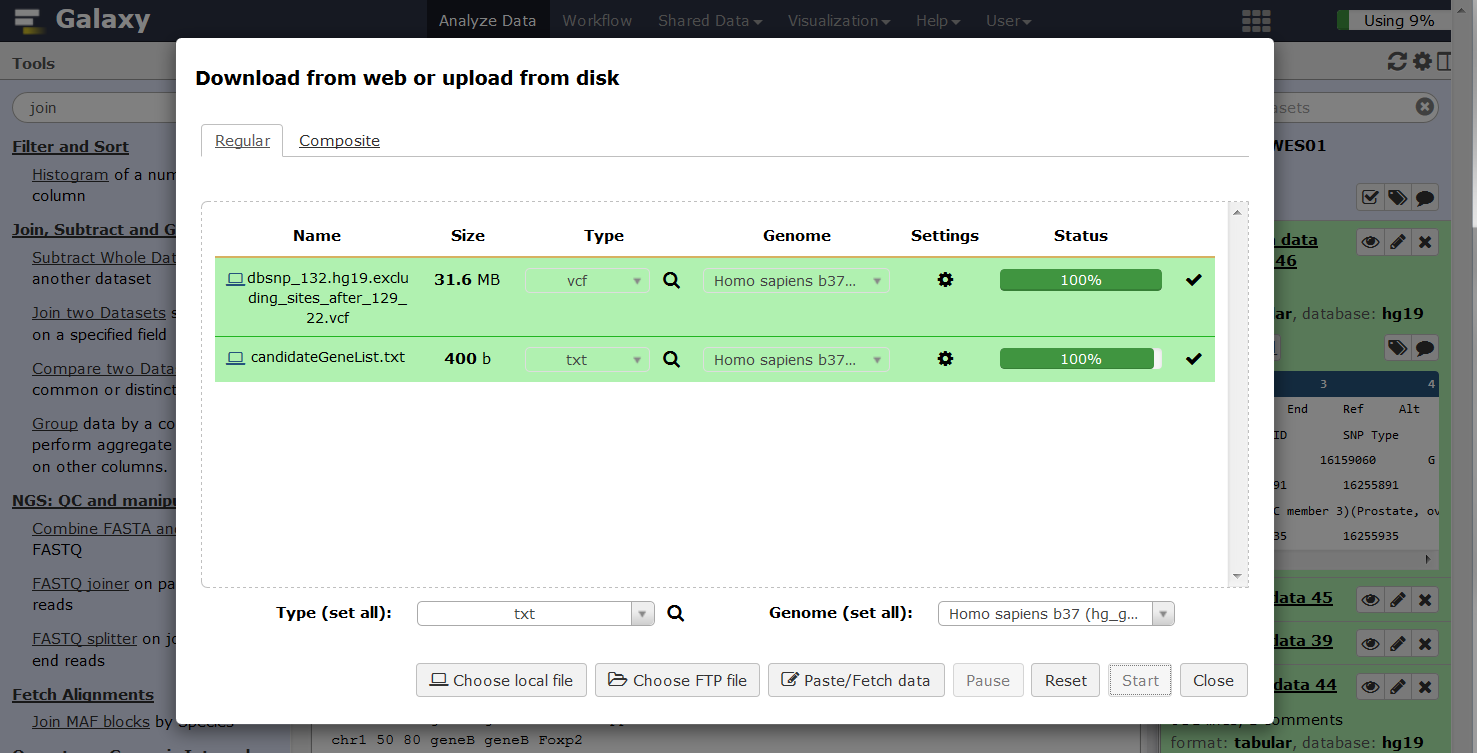
Depending on the HPO ids that were selected in Phenomizer, this process should yield a unique list of 54 candidate genes that are associated with the patients phenotype (p-value 0.05).

**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.

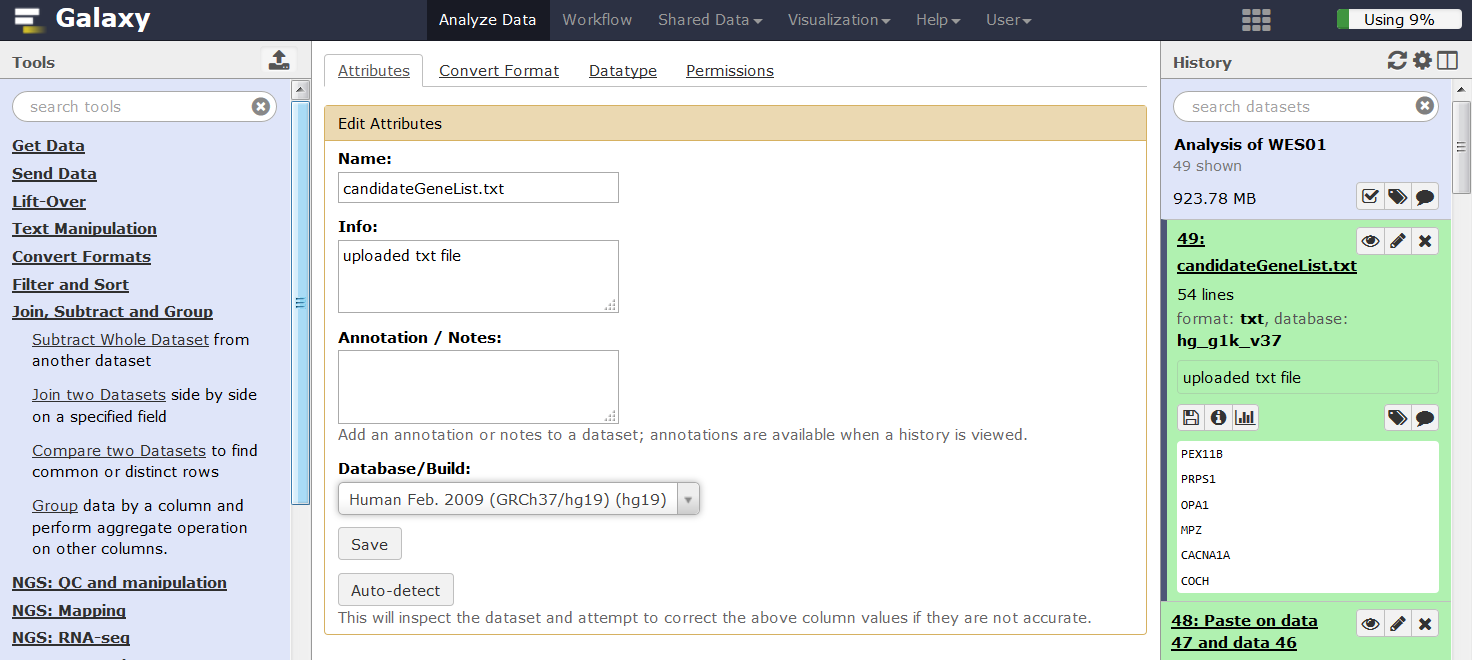


We will now cross reference the list of candidate genes with the annotated VCF file.

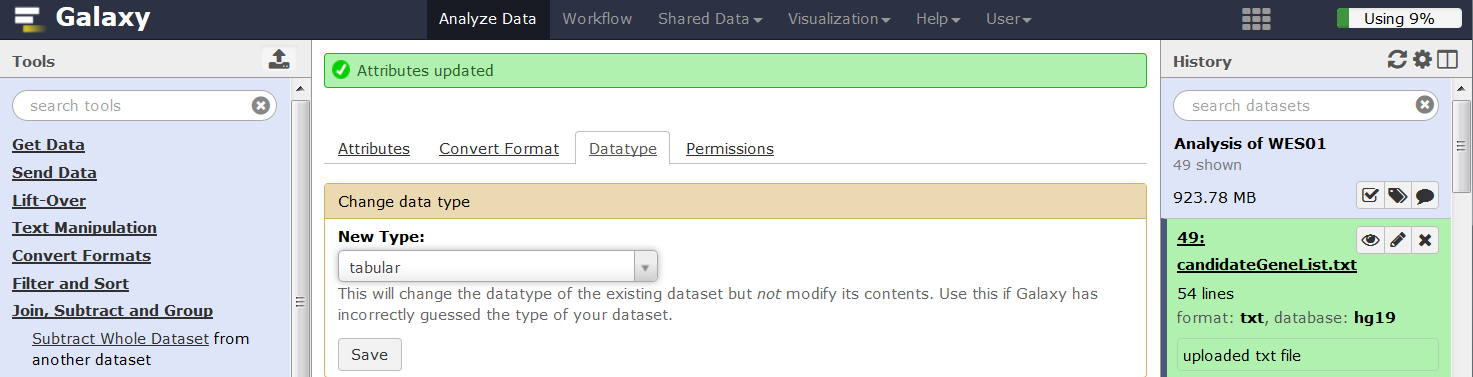
**14.** Upload the list of candidate genes to Galaxy.



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

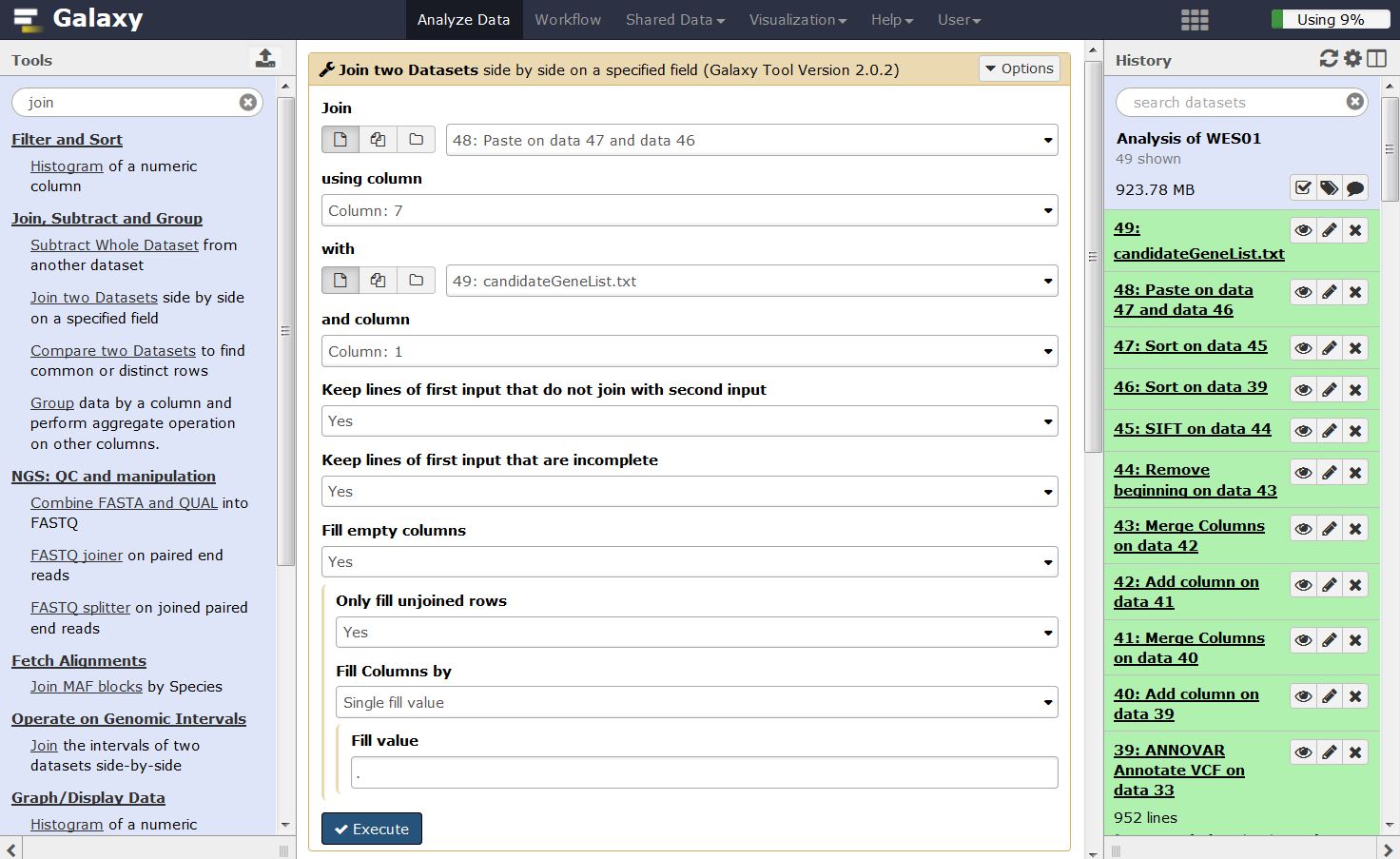


**16.** Click edit attributes, select the Datatype tab and change the datatype to tabular.



**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.

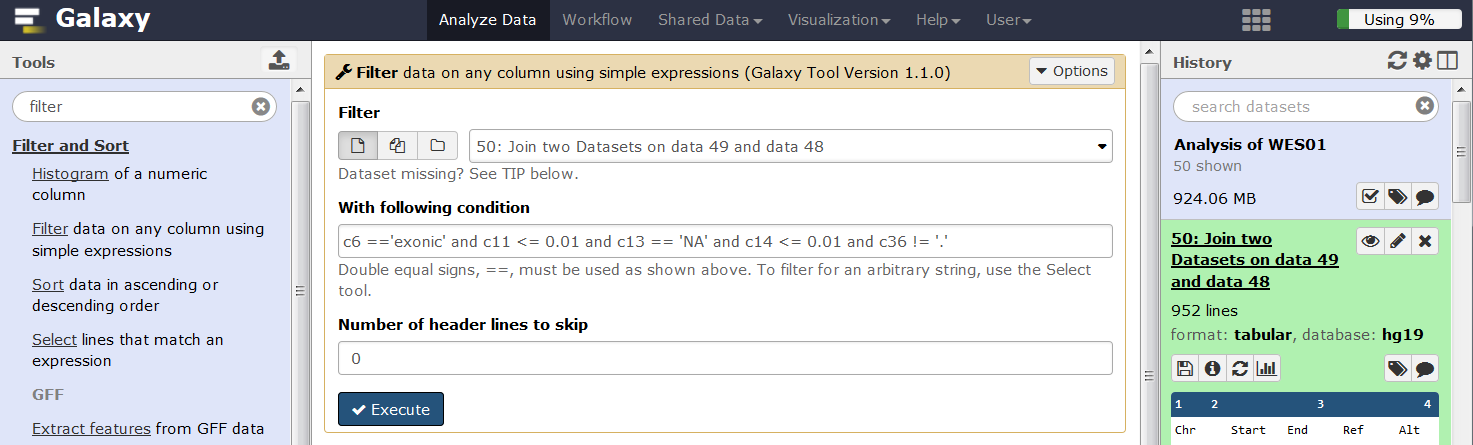


**Filter variants**

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](http://www.1000genomes.org/) (c11 <= 0.01), the non flagged version of [dbSNP](http://www.ncbi.nlm.nih.gov/SNP/) (c13 == ‘NA’), or [Exome Sequencing Project](http://evs.gs.washington.edu/EVS/) (c14 <= 0.01) and are located in a candidate gene (c36 != ‘.’).



**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](http://www.ncbi.nlm.nih.gov/clinvar/) and [OMIM](http://www.omim.org/) 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](http://www.genecards.org/). 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 |  |

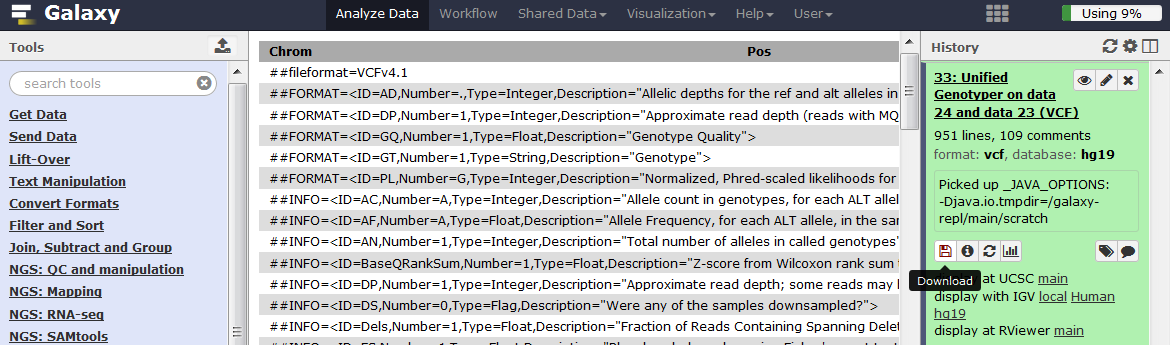
Alternatively, we could identify variants that are flagged by dbSNP as clinically associated or are present in OMIM.

**Q10.** How manyvariants are flagged by dbSNP as clinically associated?

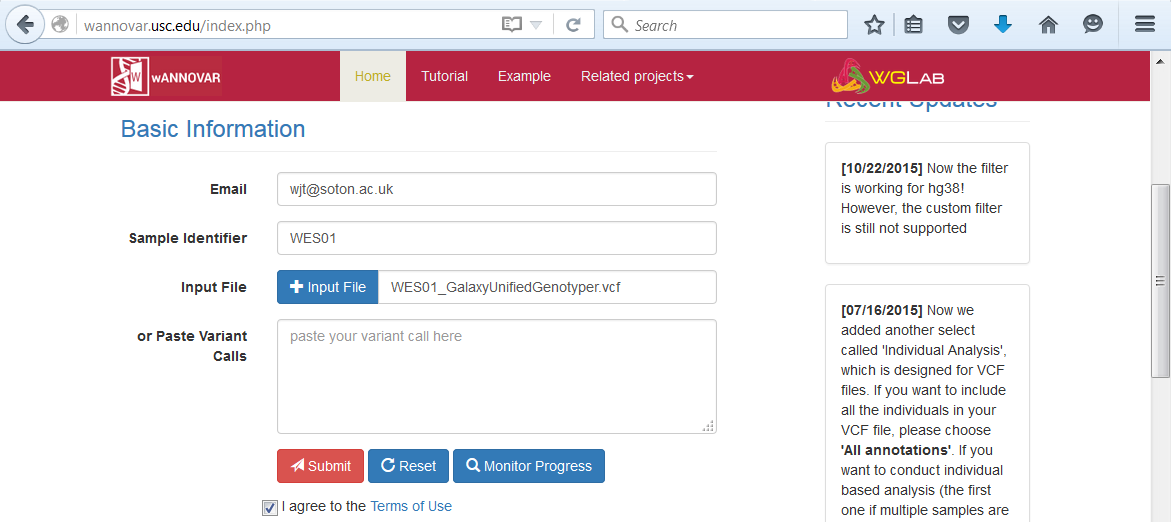
**Q11.** How many variants are present in OMIM?

The online version of ANNOVAR offers some additional annotation. Download your VCF file and use [wANNOVAR](http://wannovar.usc.edu/) to perform annotation.

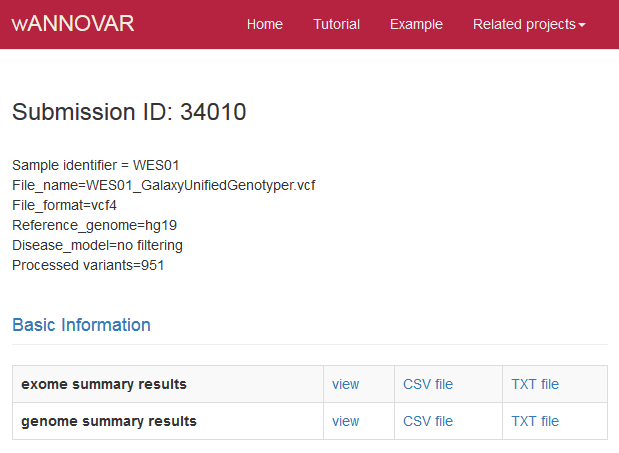
**4.** Click the download icon and save the Unified Genotyper VCF file to your computer



**5.** Upload the VCF file to wANNOVAR and click Submit



**6.** Download the genome summary results



**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](http://sift.jcvi.org/) |  | Deleterious/Tolerated/Missing |
| [Polyphen-2 HDIV](http://genetics.bwh.harvard.edu/pph2/) |  | Probably damaging/Possible damaging/Benign/Missing |
| Polyphen-2 HVAR |  | Probably damaging/Possible damaging/Benign/Missing |
| [LRT score](http://www.genetics.wustl.edu/jflab/lrt_query.html) |  | Deleterious/Neutral/Unknown/Missing |
| [Mutation Taster](http://www.mutationtaster.org/info/documentation.html) |  | Disease\_causing\_automatic/Disease\_causing/Polymorphism/ Polymorphism\_automatic/Missing |
| [Mutation Assessor](http://mutationassessor.org/) |  | Predicted functional (high, medium)/Predicted non-functional (low, neutral)/Missing |
| [FATHMM score](http://fathmm.biocompute.org.uk/) |  | Damaging/Tolerated/Missing |
| [RadialSVM](http://genomics.usc.edu/members/15-member-detail/36-coco-dong) |  | Deleterious/Tolerated/Missing |
| [LR score](http://genomics.usc.edu/members/15-member-detail/36-coco-dong) |  | Deleterious/Tolerated/Missing |
| [VEST3 score](http://karchinlab.org/apps/appVest.html) |  | None, likelihood of functional effect increase with score |
| [CADD raw](http://cadd.gs.washington.edu/home) |  | None, likelihood of damaging effect increase with score |
| [CADD phred](http://cadd.gs.washington.edu/home) |  | None, ranked and phred scaled CADD score |
| [GERP++RS](http://mendel.stanford.edu/SidowLab/downloads/gerp/) |  | None, conservation increases with score (range -12.3 to 6.17). Scores >2, high sensitivity for truly constrained sites |
| [phyloP46way placental](http://compgen.cshl.edu/phast/help-pages/phyloP.txt) |  | None, the larger the score, the more conserved the site |
| [phyloP100way vertebrate](http://compgen.cshl.edu/phast/help-pages/phyloP.txt) |  | None, the larger the score, the more conserved the site |
| [SiPhy 29way logOdds](https://www.broadinstitute.org/genome_bio/siphy/index.html) |  | 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**](http://compgen.cshl.edu/phast/help-pages/phyloP.txt)

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

**SiPhy 29way logOdds**

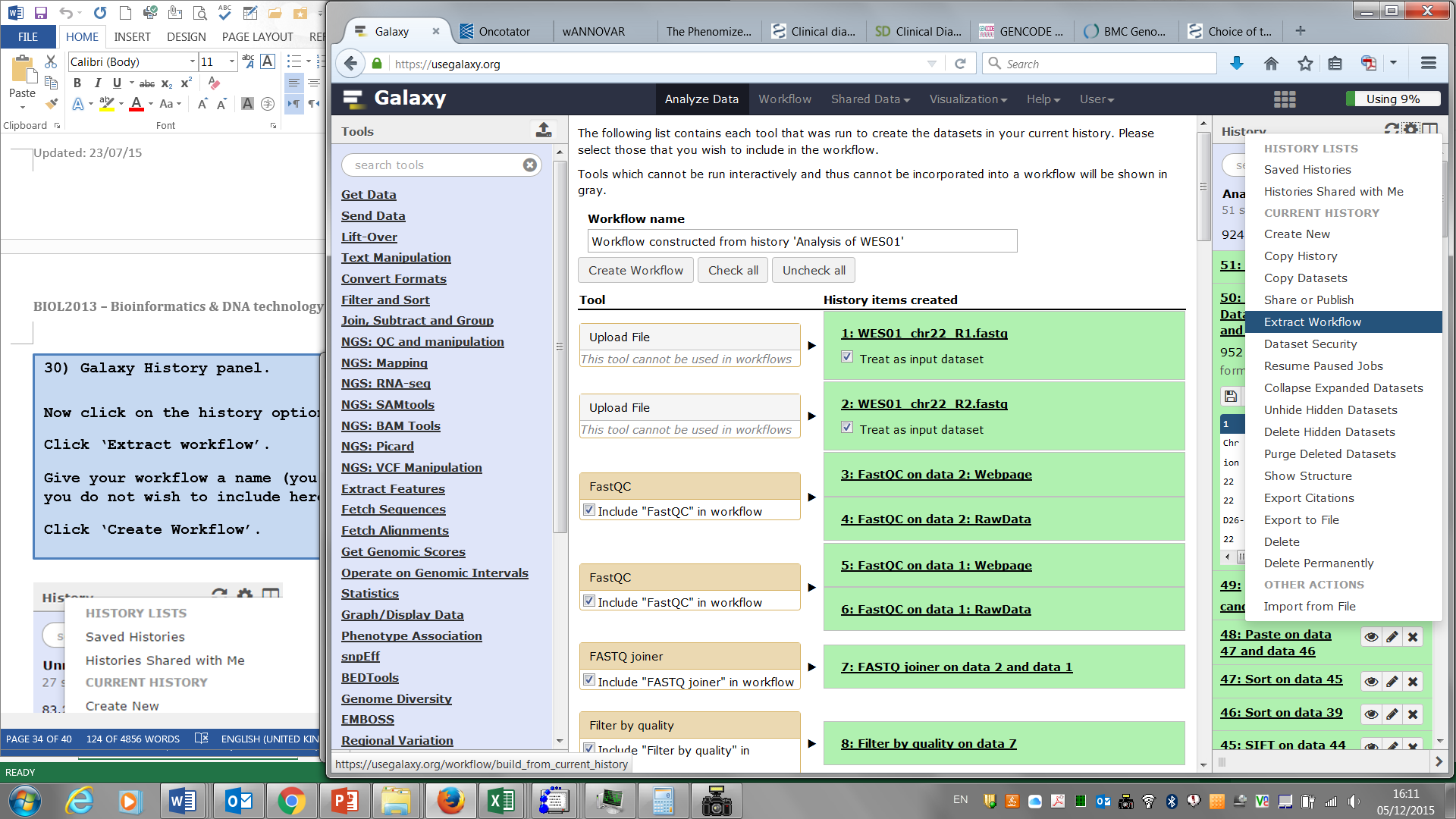
The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes.

**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.

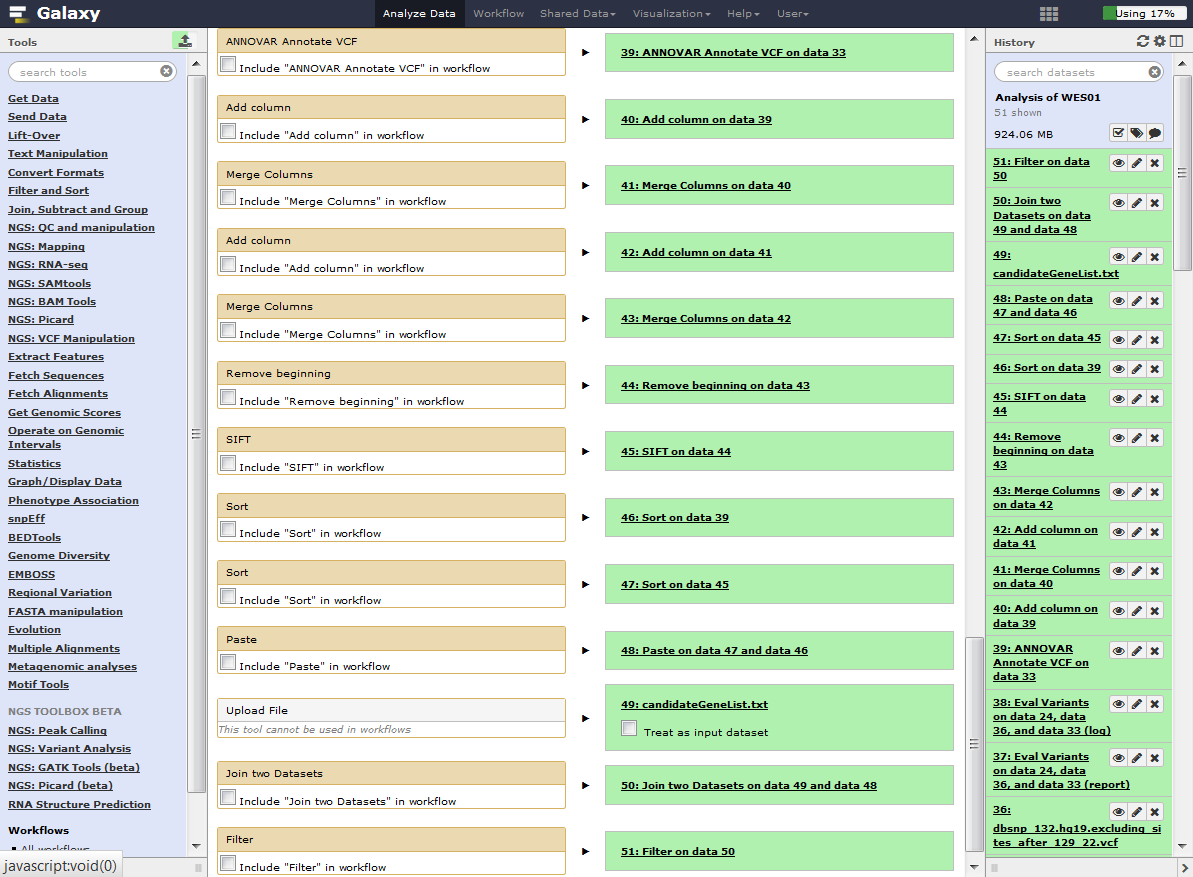
**1.** Before making your workflow delete any unwanted steps from the history so that it only includes the steps needed to make the final result.

**2.** Go to History panel, click the cog icon and select ‘Extract workflow’.

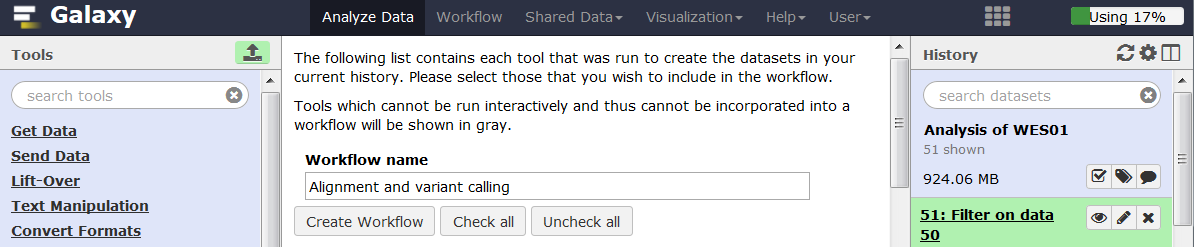


We’re going to create two workflows, one for alignment and variant calling and the other for annotation and filtering because the VCF reference genome has to be manually converted from hg\_g1k\_v37 to hg19.

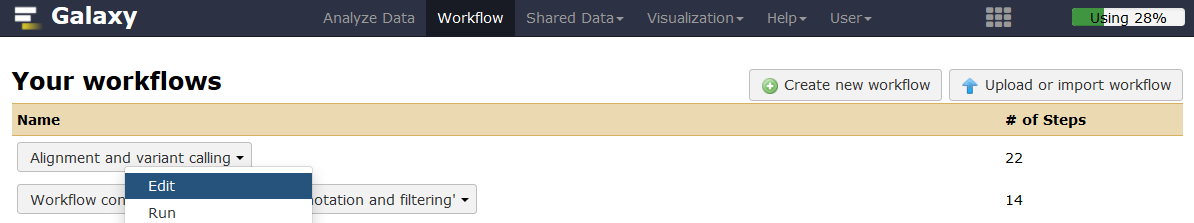
**3.** We will create the alignment and variant calling workflow first. Scroll down and uncheck all steps from ANNOVAR onwards.



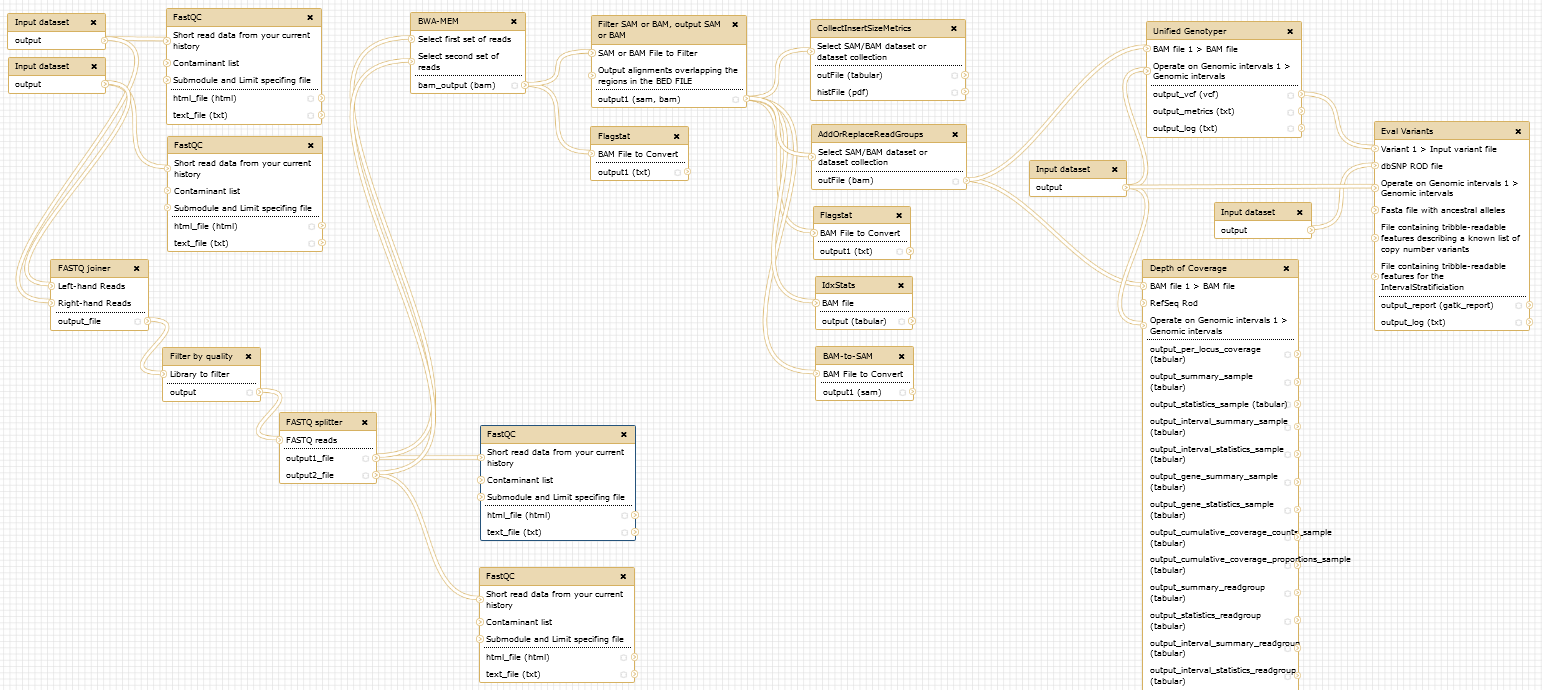
**4.** Rename your workflow as ‘Alignment and variant calling’ and then click ‘Create Workflow’.



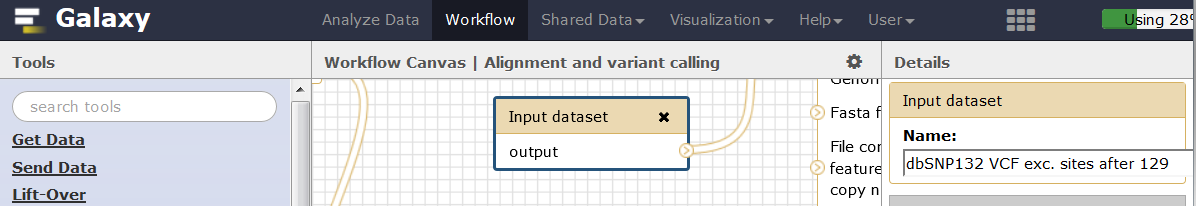
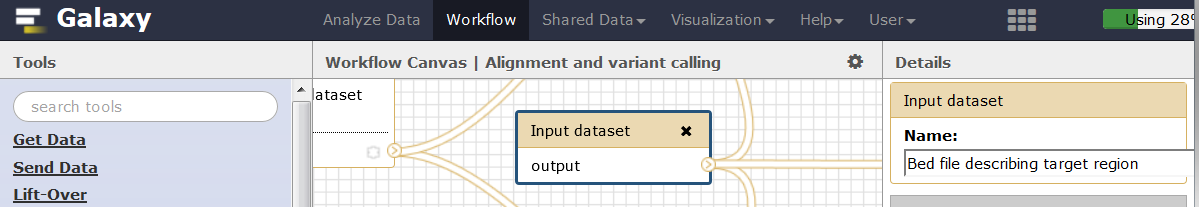
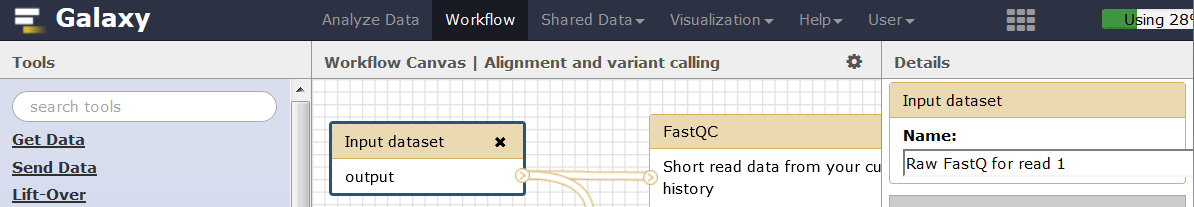
**5.** Edit the workflow by clicking on the ‘Workflow’ tab, then click on the workflow and select Edit.



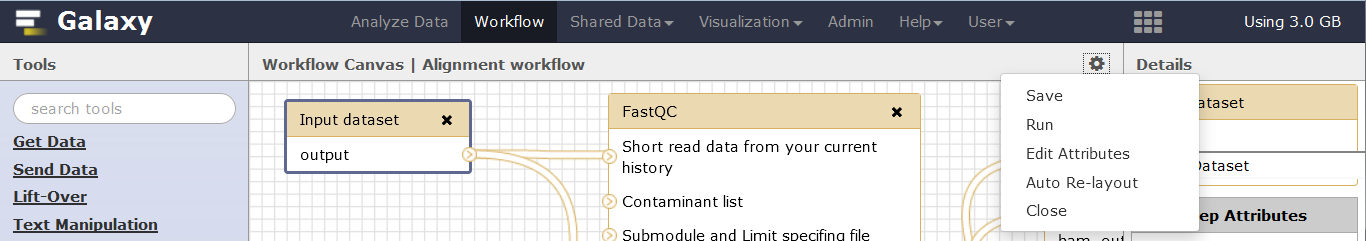
This will bring up the workflow canvas showing a flow diagram of your work. The boxes represent the steps and they are joined together to represent the inputs and outputs of each step. You can drag the boxes around to organise them in the clearest way. If you click on a box, its details will appear in 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.



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



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

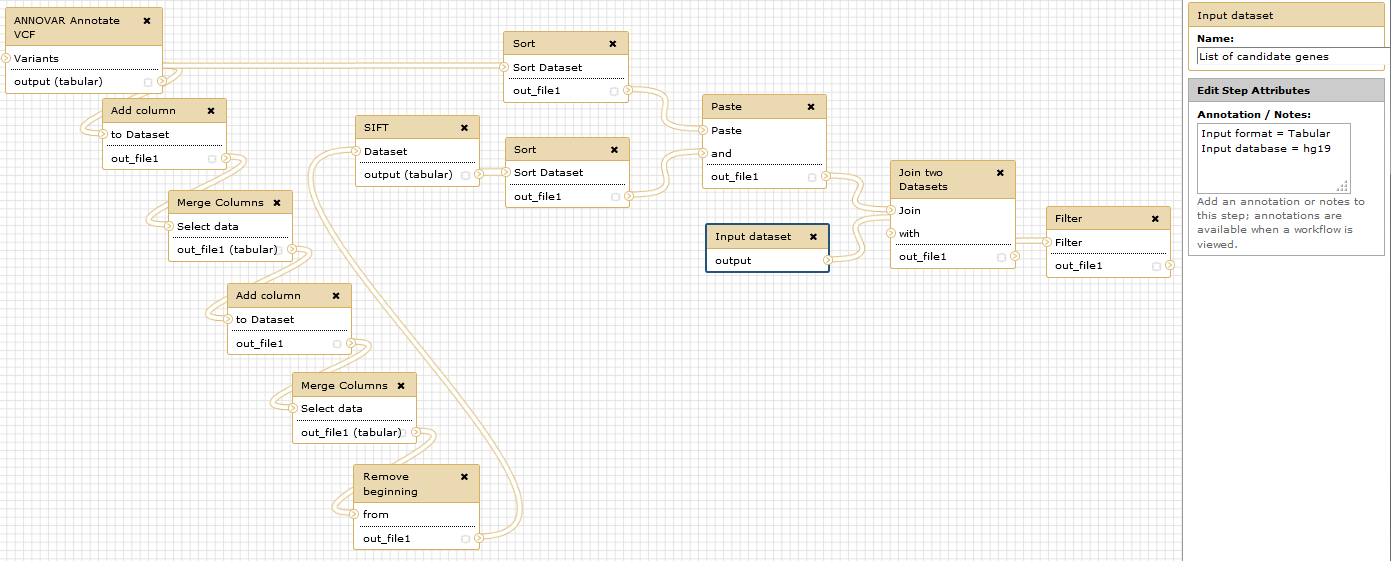


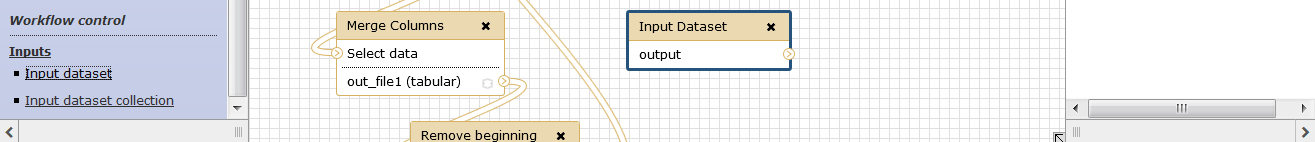
Now close the Workflow editor by clicking the cog icon and selecting ‘Close’.

**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.

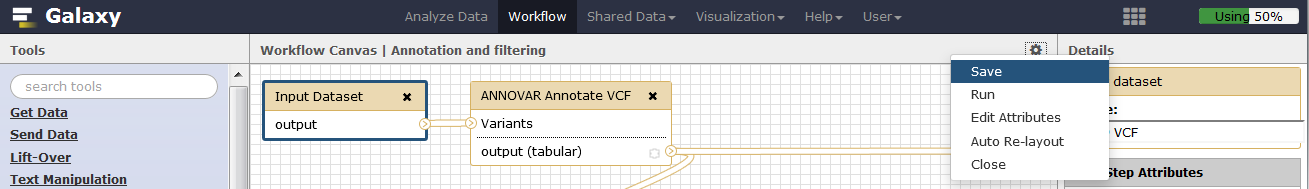


**10.** Now add an Input dataset by clicking the ‘Input dataset’ link at the bottom left of the tool pane.

**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.



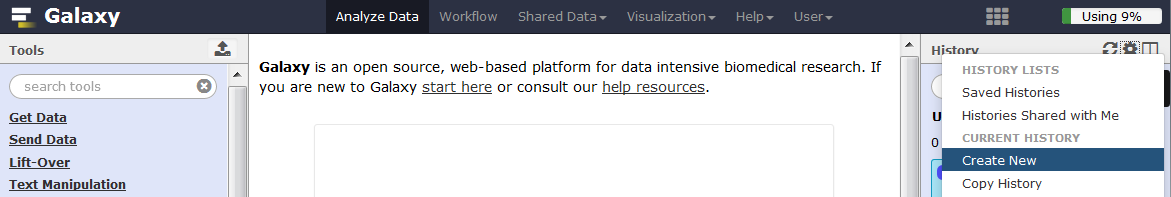
**12.** Save and Close the annotation and filtering workflow.



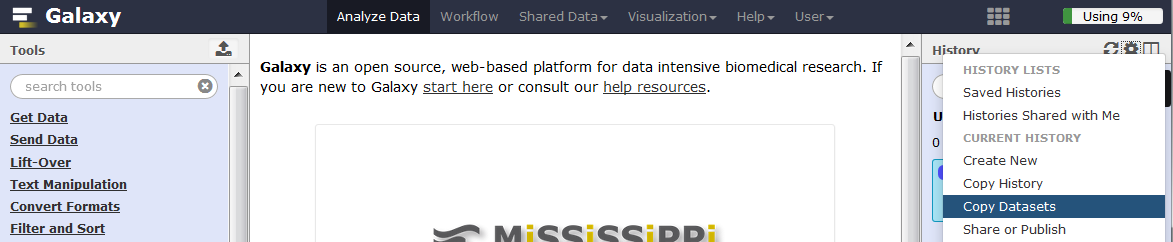
**Running workflows**

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.

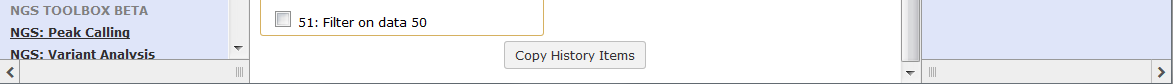
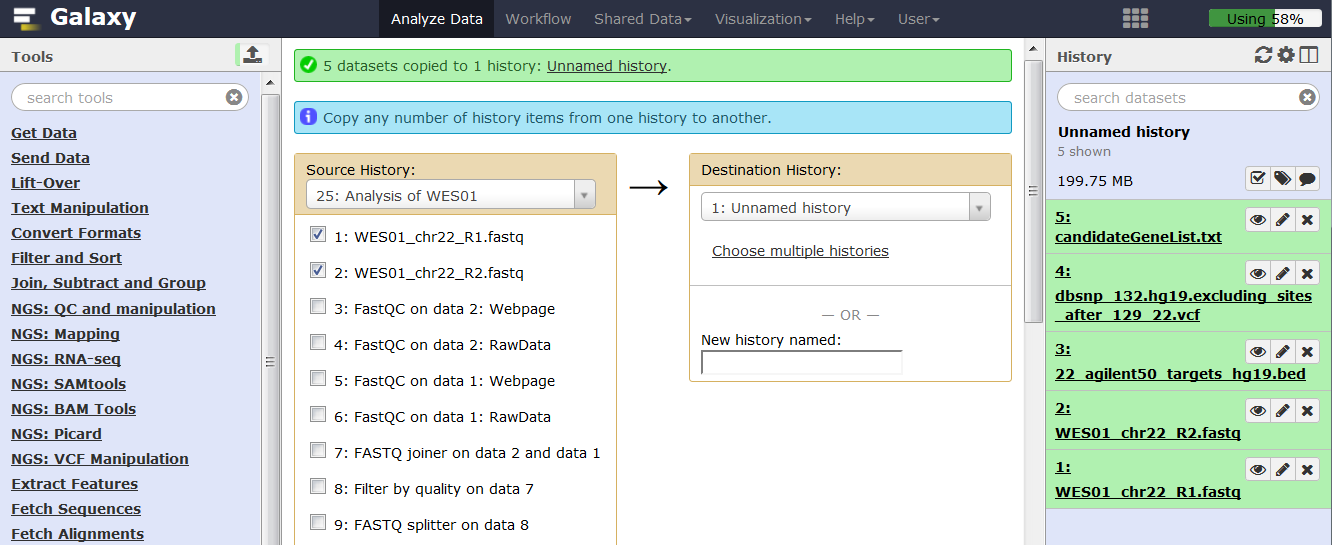
**1.** In the history pane click the cog icon and select ‘Create New’ to make a new history.



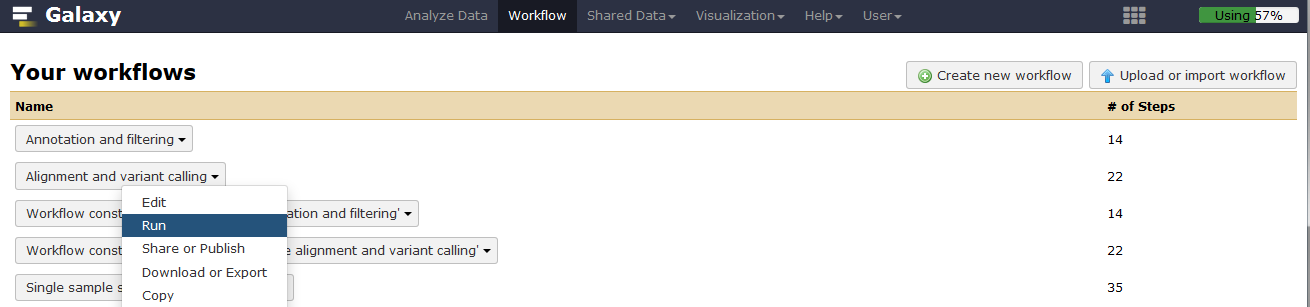
**2.** Before running the workflows, we need to put the raw data files into the new history. You can do this by uploading or copying the data from an existing dataset. We will copy the input files by clicking the cog icon and selecting ‘Copy Datasets’



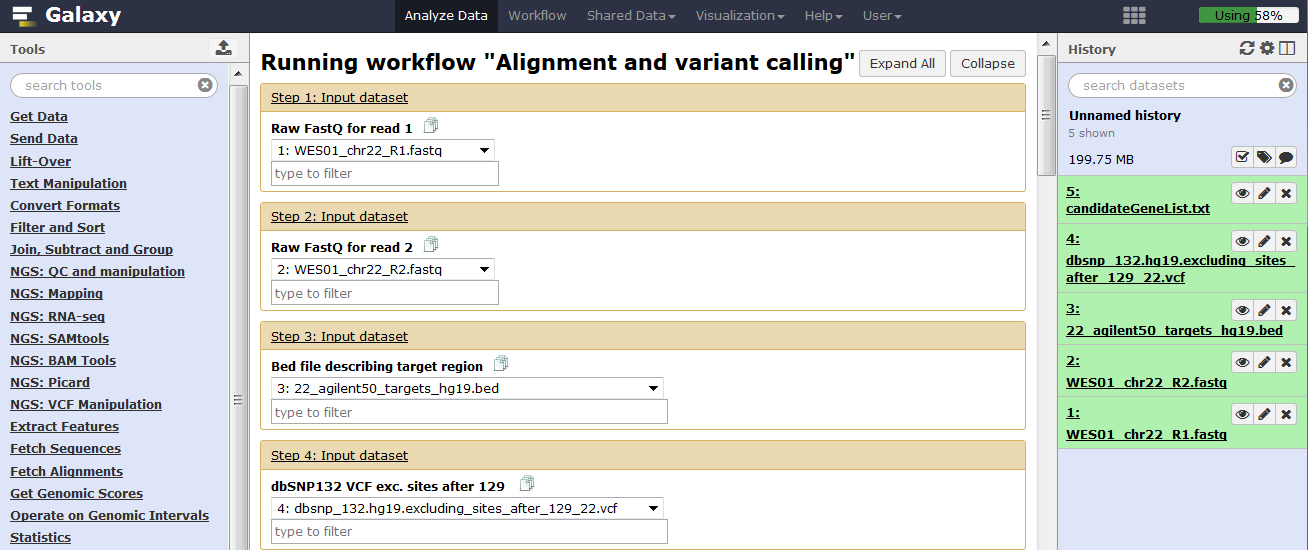
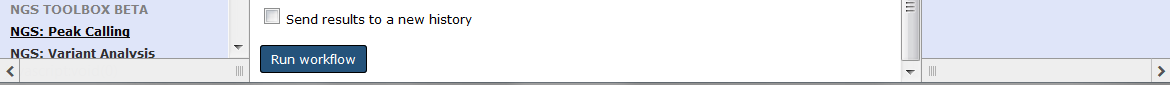
**3.** Select ‘Analysis of WES01’ as the source history and the new ‘Unnamed history’ as the destination. Check the boxes for the 5 input files that we need (raw fastqs: WES01\_chr22\_R1.fastq, WES01\_chr22\_R2.fastq, target bed file: 22\_agilent50\_targets\_hg19.bed, dbSNP rod file: dbsnp\_132.hg19.excluding\_sites\_after\_129\_22.vcf, list of candidate genes: candidateGeneList.txt). Scroll to the bottom of the page and click ‘Copy History Items’.



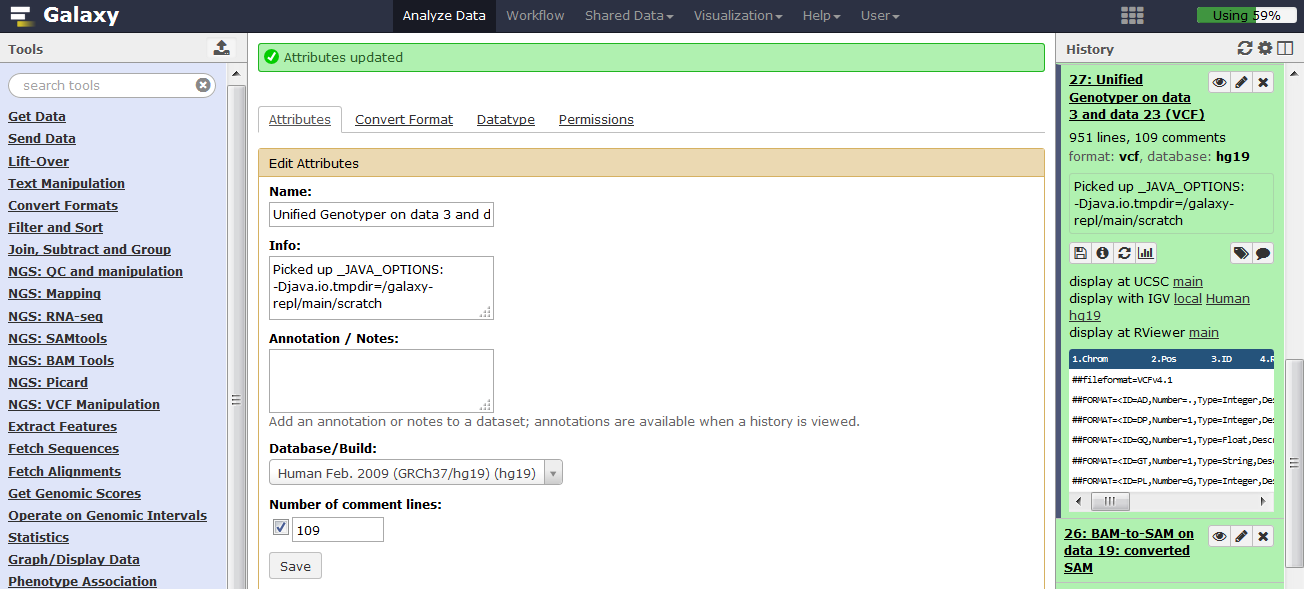
**4.** Click the workflow tab then click your ‘Alignment and variant calling’ workflow and select run from the drop down menu.



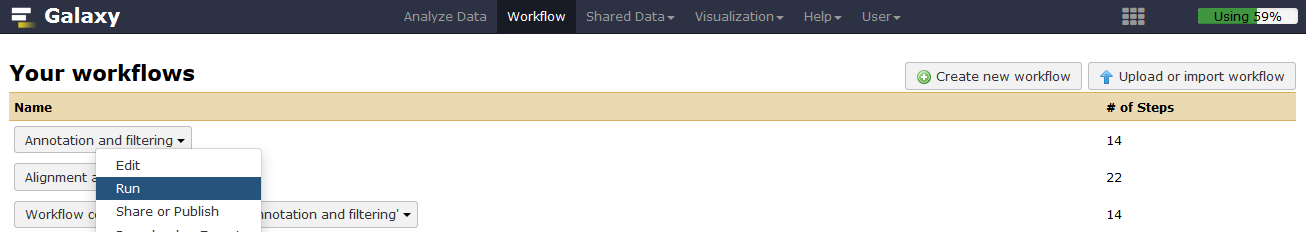
**5.** Now choose the relevant input files and click ‘Run workflow’ to analyse the data.



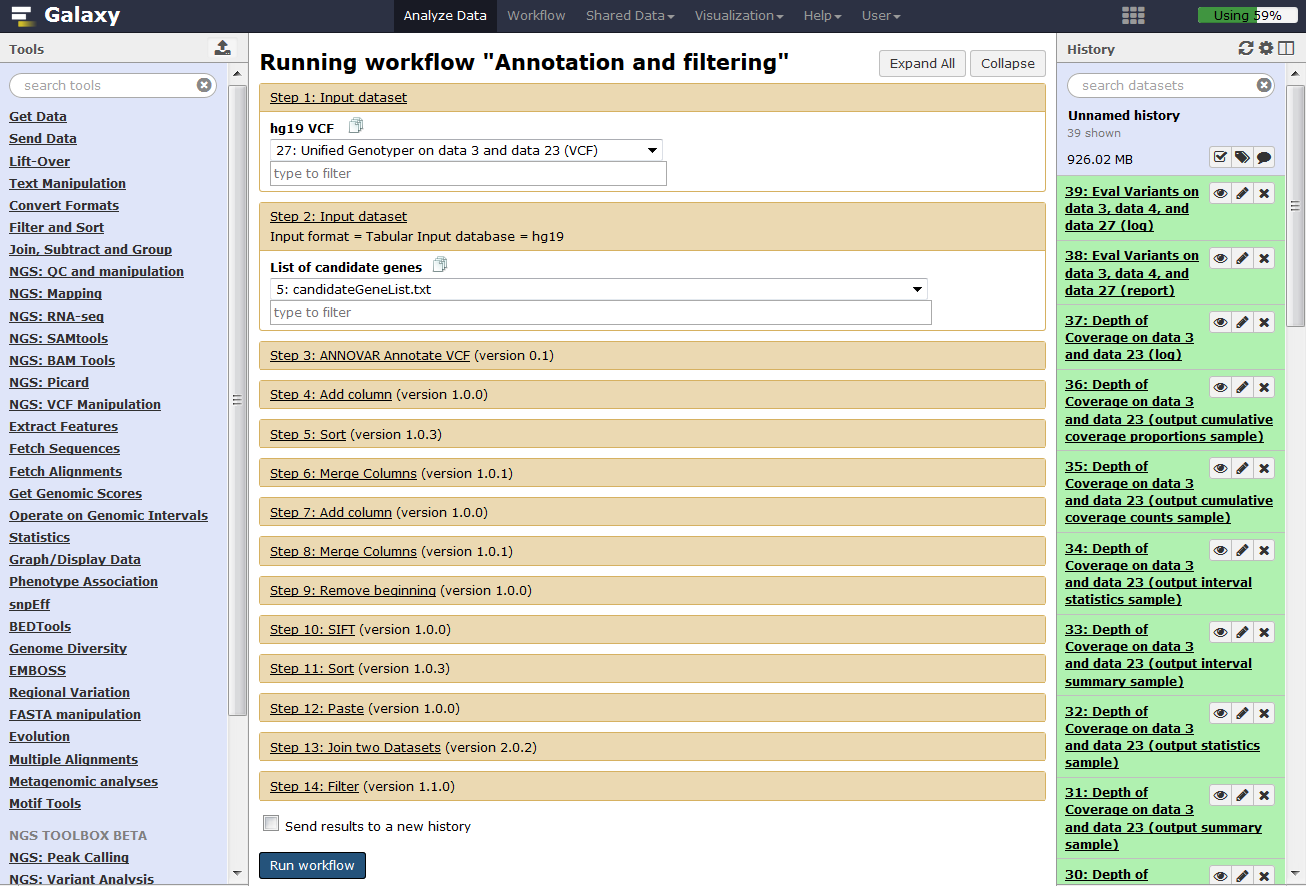
**6.** When Unified Genotyper has finished, use edit attributes to change the VCF database from hg\_g1k\_v37 to hg19.



**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.



**8.** Selectthe appropriate input files and click ‘Run workflow’ to analyse the data.

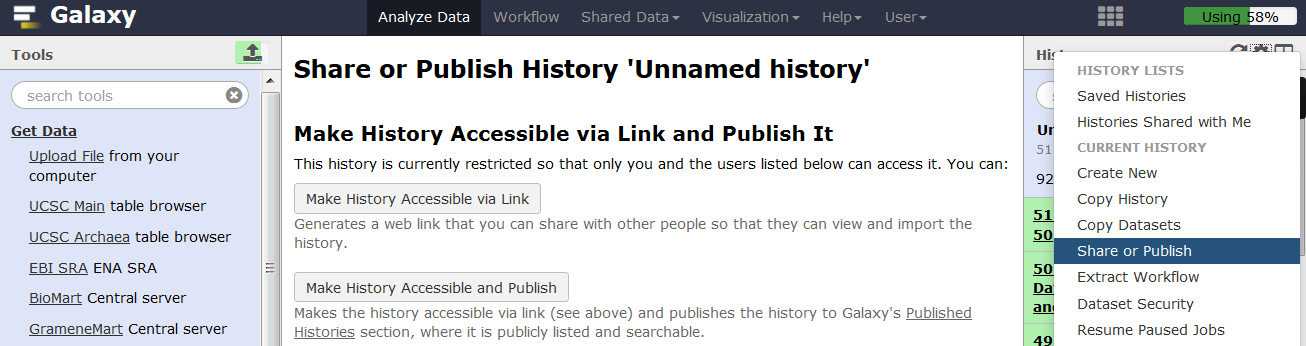


When the annotation and filtering workflow has finished view the output from the last step and check that the result matches your original analysis.

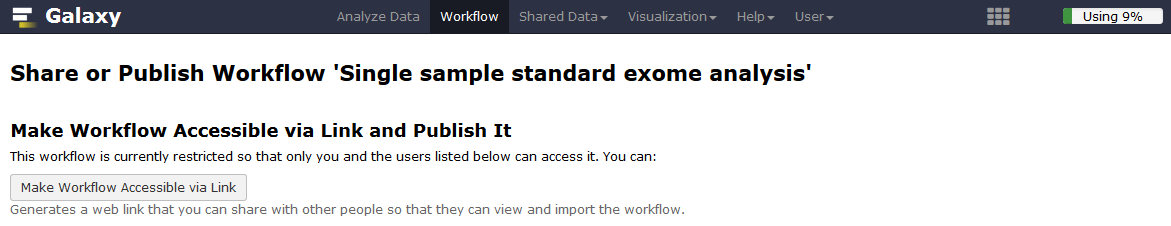
**Sharing and publishing your work**

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.



**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.



**Well done Bioinformaticians you finished the exercise!**