Overview

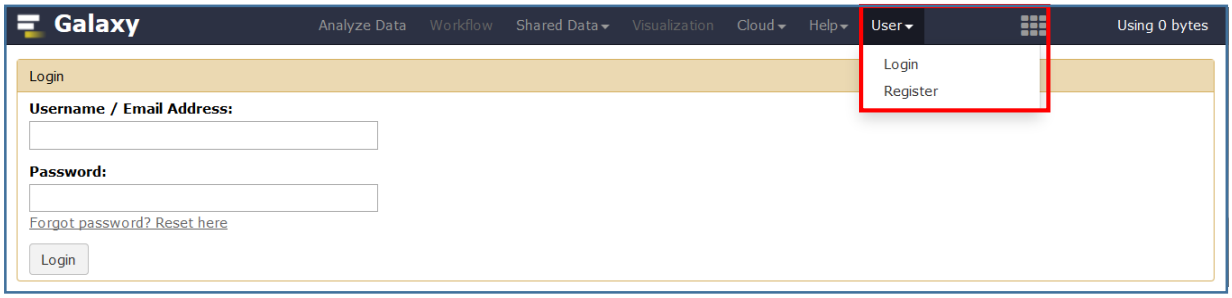
To generate a shortlist of potentially pathogenic variants we have filtered a variant call file (VCF) against public databases of known variation (1000 genomes, dbSNP, Exome Sequencing Project), estimates of deleteriousness such as SIFT and PolyPhen2, relevant candidate genes based on the patients phenotype and mutation databases (ClinVar and OMIM). Identifying the causal mutation(s) is then a straightforward process when it occurs in a known disease-causing gene or the mutation has previously been described. In situations involving novel mutations and genes, additional filtering against related and/or unrelated individuals is often used to identify genes that are recurrently mutated in cases and are not mutated in controls. Comparison with parents may also be used to filter variants and to determine the mode of inheritance. The aim of this practical is to further reduce the number of putative causal variants by comparing VCF files between individuals. At the end of this exercise you will be able to:

1. Combine VCF files into one multisample file
2. Compare VCF files to identify shared variants and infer relatedness
2. Use GEMINI to analyse trio data and identify Mendelian errors and de-novo mutations
3. Calculate and plot alternate read percentages and use them to identify potential regions of acquired uniparental disomy (aUPD)
4. Filter VCF files using COSMIC and regions of aUPD to identify the underlying somatic driver mutation

Let's begin

For this session we will be using tools that are only available on the public version of galaxy (https://usegalaxy.org/).

1. Register for a public galaxy account and login



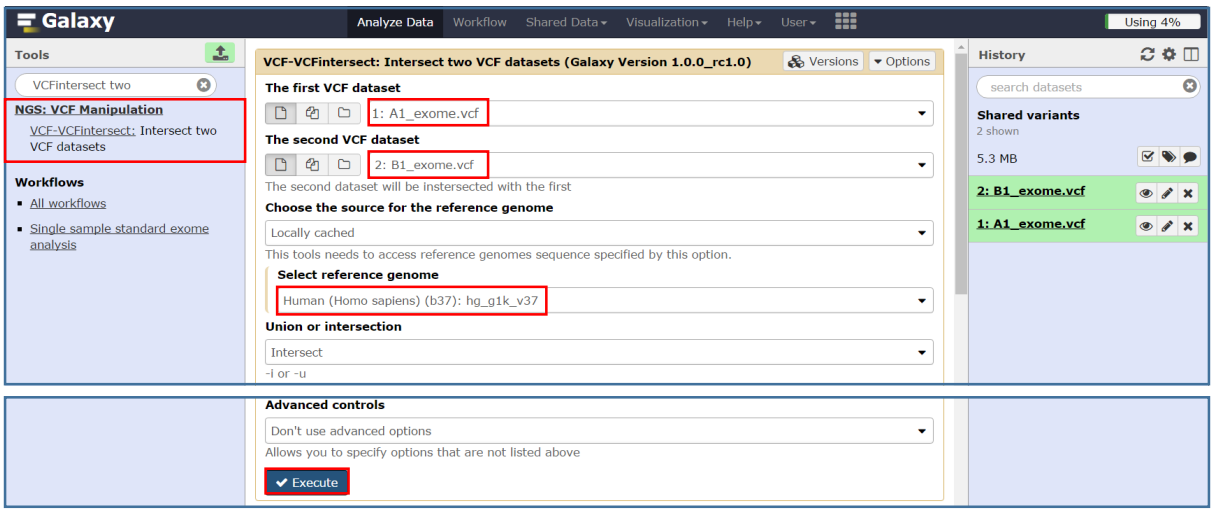
Identifying shared variants

In the first example whole exome sequencing was performed in two unrelated patients with intellectual disability, delayed speech and language development. The sequence data were analysed separately to produce two VCF files of exonic variants using BWA-MEM for alignment to hg19, samtools for variant calling and selection of variants with a minimum read depth of 4 and minimum QUAL score of phred 20. Our aim is to identify potentially causal variants that are present in both patients.

1. Rename the history for example 1 (Click on name, type 'Shared variants', press return)
2. Upload the two VCFs from the USB key in the 'FILTERING STRATEGIES' folder (A1_exome.vcf and B1_exome.vcf)) to galaxy (https://usegalaxy.org/). Select VCF for type and Human Feb. 2009 (GRCh37/hg19)(hg19) for reference genome.

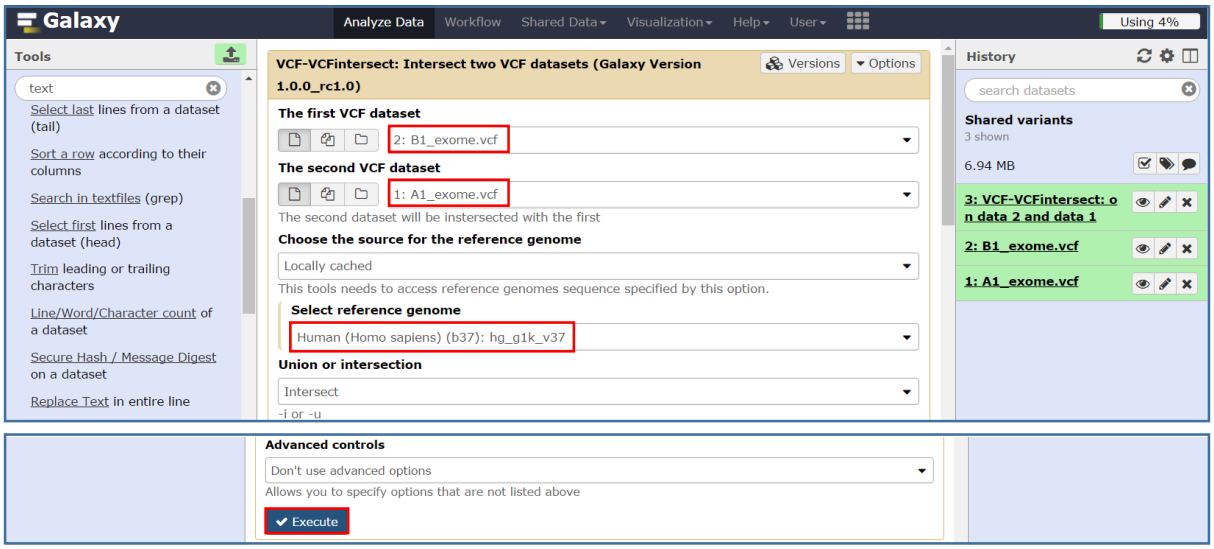
To identify shared variants we will intersect the two VCFs.

3. In **Tool Pane**: Go to **NGS: VCF Manipulation** > VCF-VCFintersect: Intersect two VCF datasets



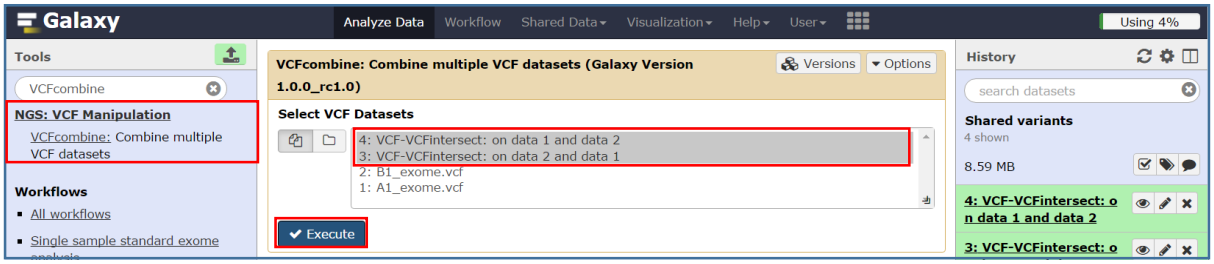
The result from VCF-VCF intersect only lists genotype calls present in sample B1. Repeat the intersection but this time compare sample B1 versus sample A1.

4. In **Tool Pane**: Go to **NGS: VCF Manipulation** > VCF-VCFintersect: Intersect two VCF datasets



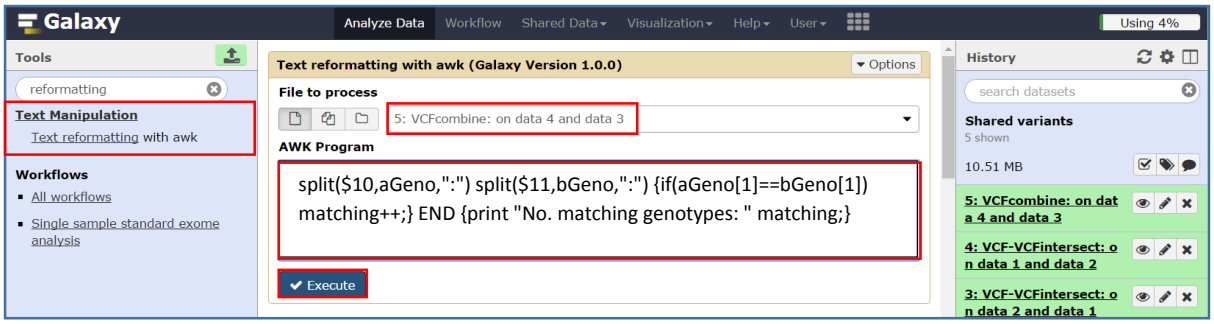
Now combine the two VCFs to make one VCF for variants shared between samples A1 and B1 and their genotype in both samples.

5. In **Tool Pane**: Go to **NGS: VCF Manipulation** > VCFcombine: Combine multiple VCF datasets



The proportion of variants with the same genotype between a pair of individuals offers a crude measure of relatedness although this will vary according to the sequencing approach (eg gene panel, whole exome, or whole genome) and capture kit (Agilent version 4 versus 5). Use text reformatting with awk to count the number of shared genotypes.

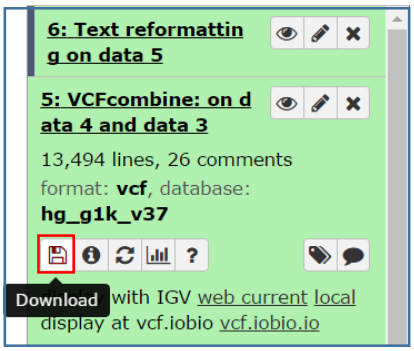
6. In **Tool Pane**: Go to **Text Manipulation** > Text reformatting with awk

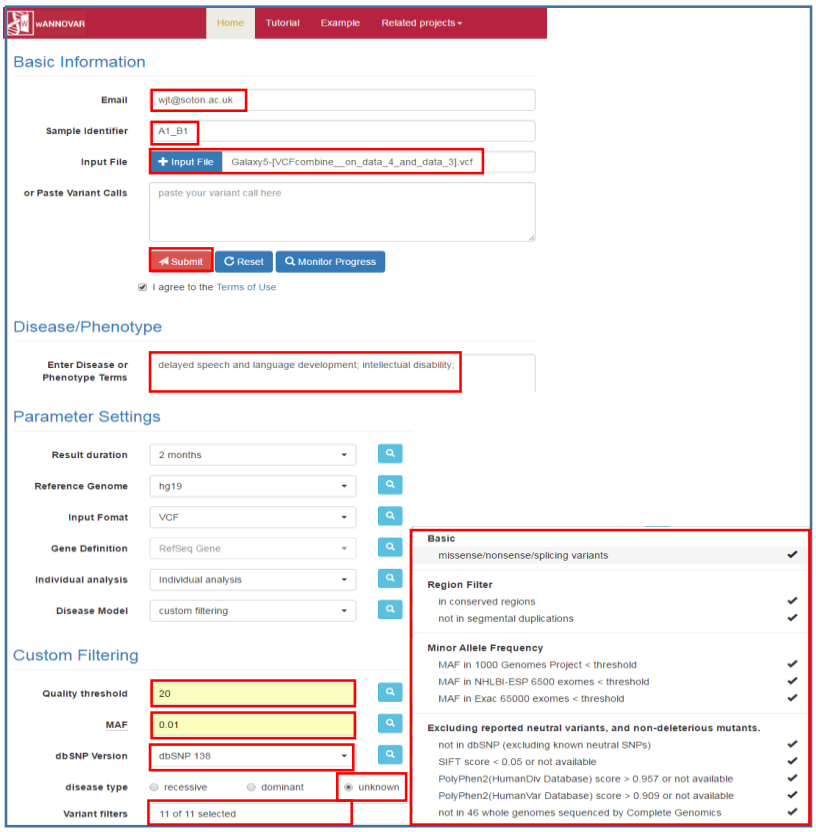


Awk is a powerful programming language for working on large data files and is often applied to NGS data. In VCF files, sample genotypes are recorded in a string of variables each separated by a colon. For example, 1/1:255,84,0:28:0:99 in the format GT:PL:DP:SP:GQ where GT is the genotype, PL is the phred scaled likelihoods, DP is the depth, SP is the phred-scaled strand bias P-value and GQ is the phred-scaled genotype quality. The awk command above splits this string into a list of separate variables for sample A1 (column 11) and sample B1 (column 10), counts every time the genotypes match and prints the number of matches when the end of the VCF file is reached.

Q1: What percentage of all variants in sample A1 have the same genotype in sample B1 and based on this value do you think these samples unrelated, at least to the 3rd degree, or not?

7. Save the combined VCF and use wANNOVAR to annotate and filter the variants

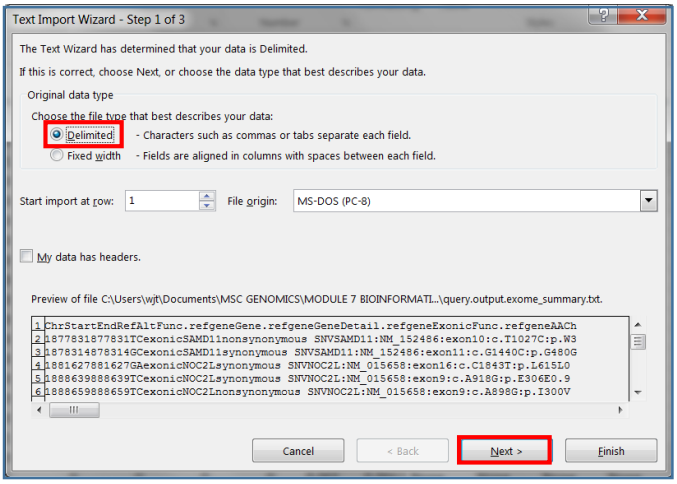


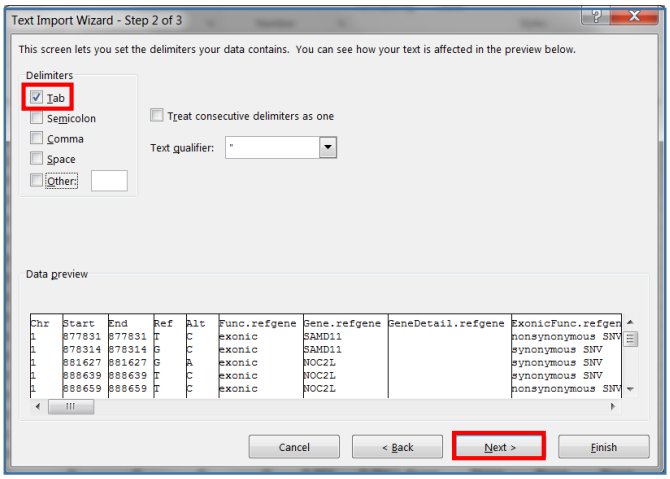


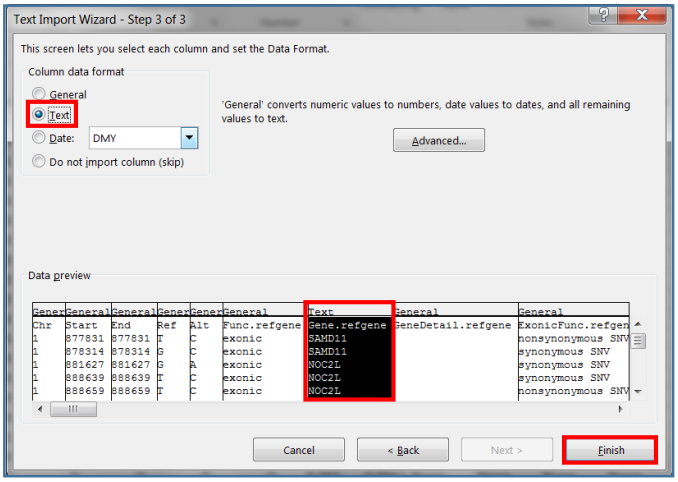
8. Save the annotated exome summary results as a TXT file and the 9 filtered variants from step 11 from wANNOVAR.

Unfortunately the filtered results are not annotated and Galaxy is not very good at displaying files with numerous columns. We will therefore use Excel to open and compare the annotated and filtered results.

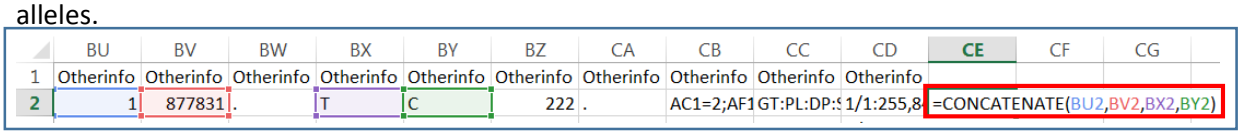
10. Open the annotated results (query.output.exome_summary.txt) in excel



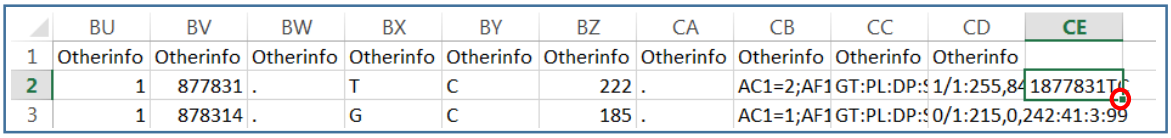




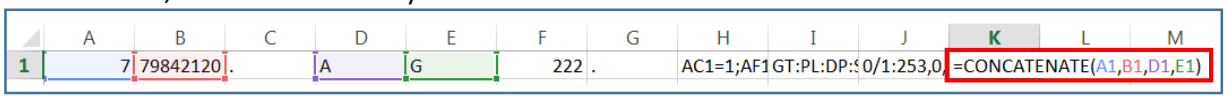
11. Create a key to compare files by merging the chromosome, location, reference and alternate alleles.



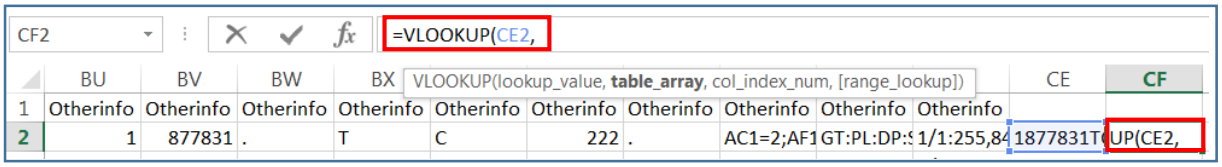
12. Mouse over the bottom right hand corner of the cell (CE2) and double click to auto fill column CE



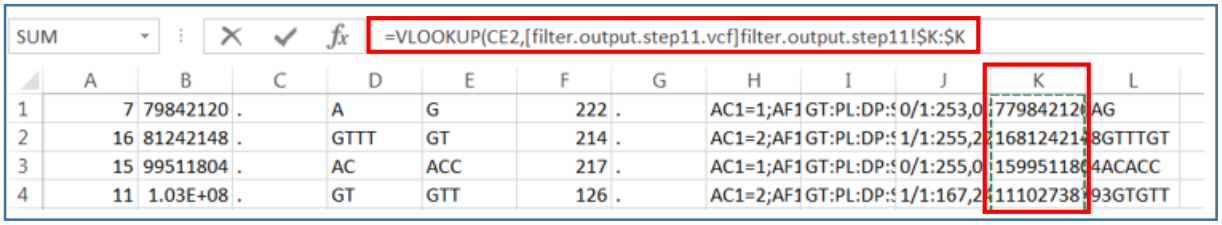
13. Open the filtered results file from wANNOVAR step 11 (filter.output.step11.vcf) in Excel as a tab delimited file, create the same key and auto fill column K



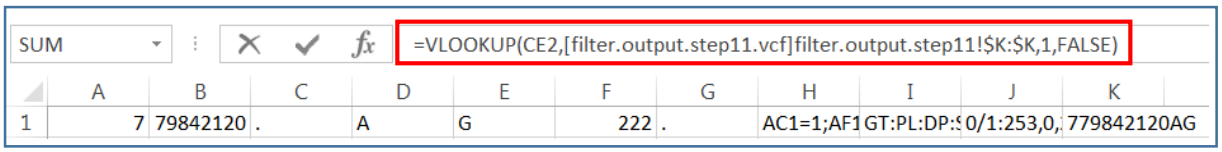
14. Now use the keys to compare the annotated and filtered results.
i) Go to the annotated file (query.output.exome_summary.txt) and type "=vlookup(CE2," in cell CF2



ii) For the table_array argument, switch to the filtered results file (filter.output.step11.vcf) and highlight column K which contains the key

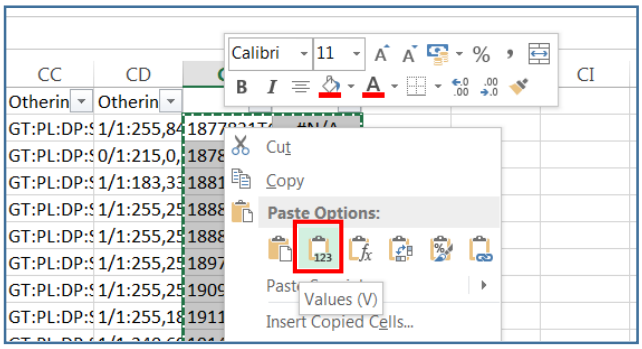


iii) Type "1" for the col_index_num, "FALSE" for [range_lookup] and press return then autofill column CF

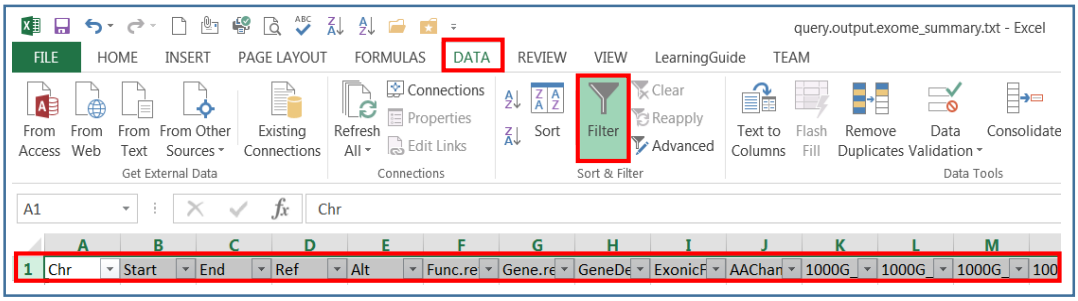


The complete function means search column K for the value in CE2 and return the value in column K when an exact match is found.

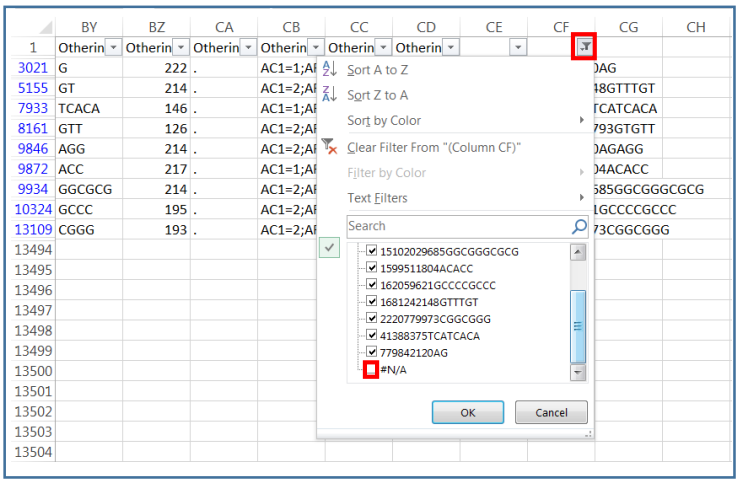
17. Before filtering the annotated results, select cell CE2, then hold down Shift and Ctrl and use the arrow keys to highlight all cells in columns CE and CF. Once highlighted, use Ctrl c to copy the cells then right click and select paste values. This will replace all formulas with text which will speed up filtering.



18. Select the DATA tab, highlight row 1 and click Filter



19. Use the autofilters in columns BG and CF to select the 9 filtered variants and sort them in descending order of pathogenicity.



The missense variant in GNAI1 (NM_001256414, p.K218R) is predicted to be the most deleterious shortlisted variant, although the others were not scored because they are either frameshifts or have unknown consequences. The variant is confidently called as heterozygous in both patients (GQ=99) and there are no QC issues recorded in the VFC file (high depth and no strand bias). In a recent study consisting of 7,580 whole exomes from patients with developmental disorders, de-novo GNAI1 mutations were identified in 8 patients including two with the same p.K218R mutation (Deciphering Developmental Disorders consortium 2017). Using this data, the authors demonstrated that GNAI1 was significantly enriched for deleterious de-novo mutations ($P < 5x10^{-7}$) and concluded that it may be a novel cause of intellectual disability. Based on this and the observation that GNAI1 is mutated in both patients, we can conclude that this mutation is likely to cause the disease in patients A1 and B1.

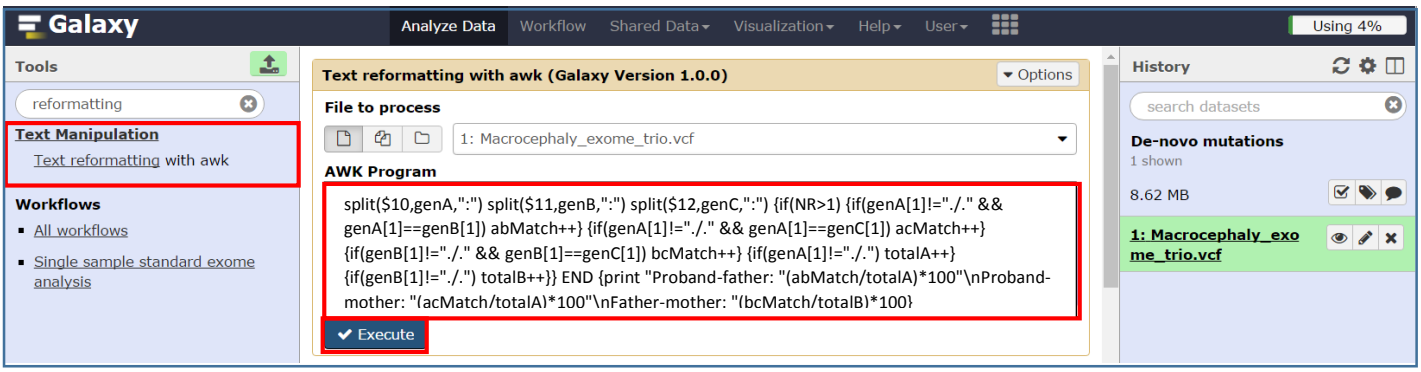
Trio sequencing for de-novo mutations

Whole exome trio sequencing of a child affected with a severe disorder and both unaffected biological parents is an effective strategy to determine diagnoses based on mutations in known genes and to generate evidence for the involvement of novel genes (Zhu et al 2015). The aim of the next example is to identify potentially causal de-novo mutations by comparing whole exome VCF files from an affected child and healthy parents. The child had macrocephaly at birth, capillary malformations on the nose and philtrum, cutis marmorata and polymicrogyria were identified by MRI. The VCF files were generated using BWA-MEM for alignment against hg19/build 37 and multi-sample GATK HaplotypeCaller for variant calling and selection of variants with a minimum QUAL score of phred 30.

1. Create a new history for example 2 (Click on name, type 'De-novo mutations', press return)

2. Upload the trio VCF from the USB key in the 'FILTERING STRATEGIES' folder (Macrocephaly_exome_trio.vcf) to galaxy (https://usegalaxy.org/), select VCF for type and Human Feb. 2009 (GRCh37/hg19)(hg19) for reference genome.
As before use text reformatting with awk to calculate the proportion of variants shared between all pairs of individuals as a simple check for relatedness.

3. In **Tool Pane**: Go to **Text Manipulation** > Text reformatting with awk

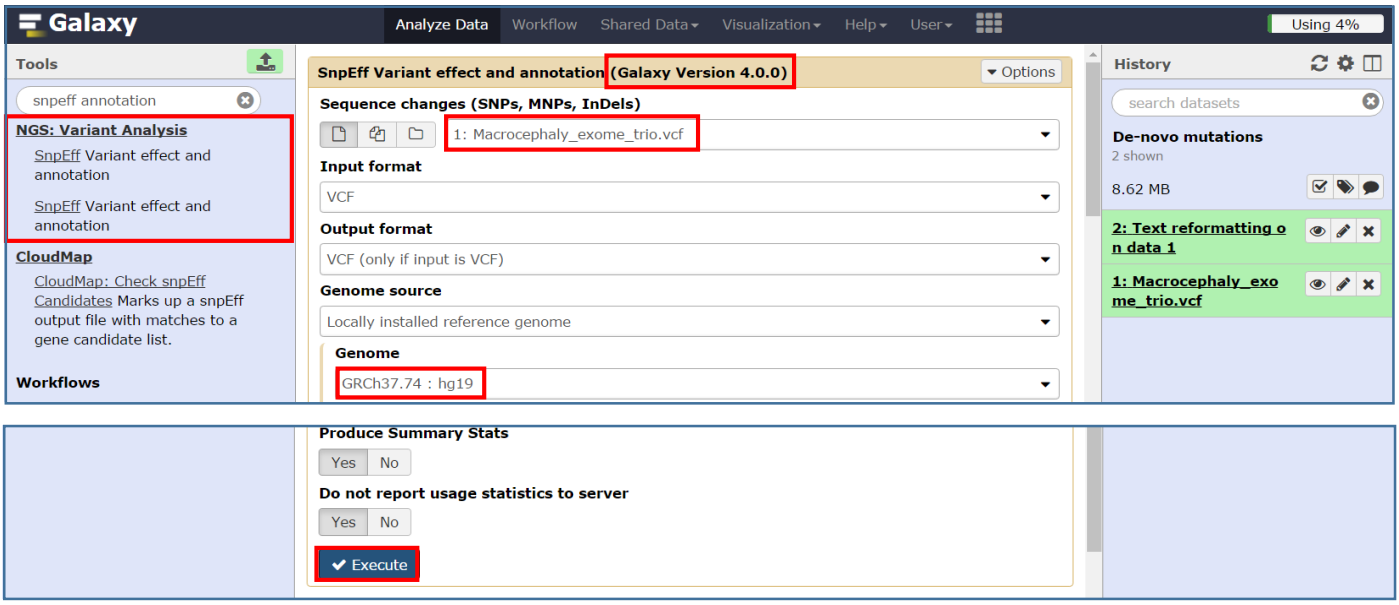


View the results and answer these questions.
Q2: Do the percentages of shared variants agree with reported pedigree?
Q3: Is there any evidence to suggest that the parents are related?
Q4: Compare the proportion of shared variants between unrelated individuals with the result from example 1 and give one reason which accounts for most of the difference.

To identify de-novo mutations we will use the GEMINI (GEnome MINIng) software. GEMINI annotates VCF files with data from multiple sources (ENCODE tracks, UCSC tracks, OMIM, dbSNP, KEGG, and HPRD) and loads this information along with familial relationships into a SQLite database. The data can then be queried based on sample genotypes, inheritance patterns and the annotated fields.

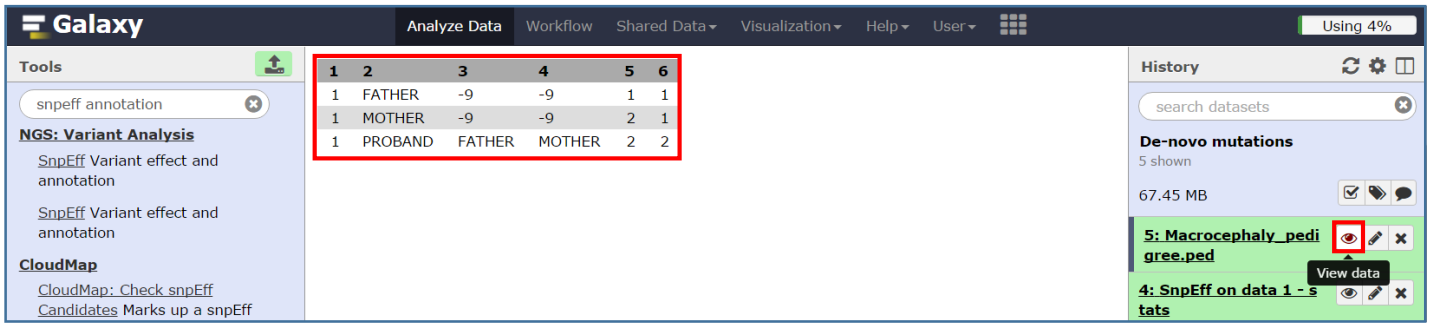
To use GEMINI the first step is to load a VCF file and a pedigree file into the GEMINI database. The VCF file must be annotated using either VEP or SnpEff before loading. We will therefore use SnpEff to annotate the VCF.

4. In **Tool Pane**: Go to **NGS: Variant Analysis** > SnpEff Variant effect and annotation
There are two versions of SnpEff available, make sure you choose version 4.0.0

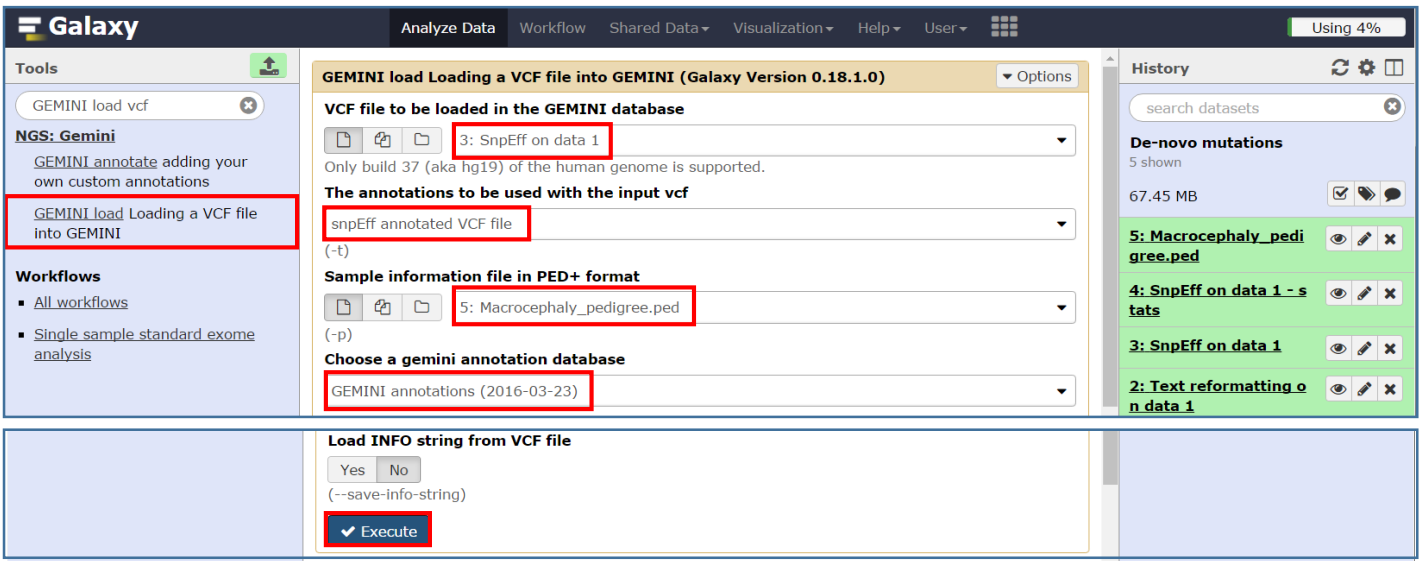


5. Now upload the trio pedigree from the USB key in the 'FILTERING STRATEGIES' folder (Macrocephaly_pedigree.ped) to galaxy (https://usegalaxy.org/), select tabular for type and Human Feb. 2009 (GRCh37/hg19)(hg19) for reference genome.

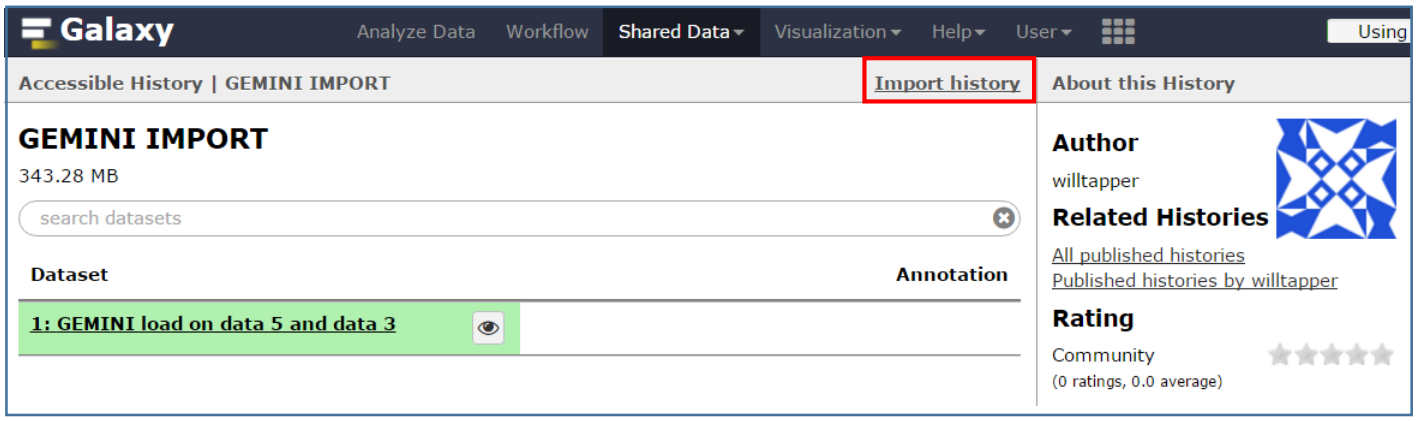
The pedigree file consists of six columns which describe the family_id, sample name, paternal_id, maternal_id, sex and phenotype where -9 is missing, 1=male, 2=female, 1=unaffected and 2=affected.



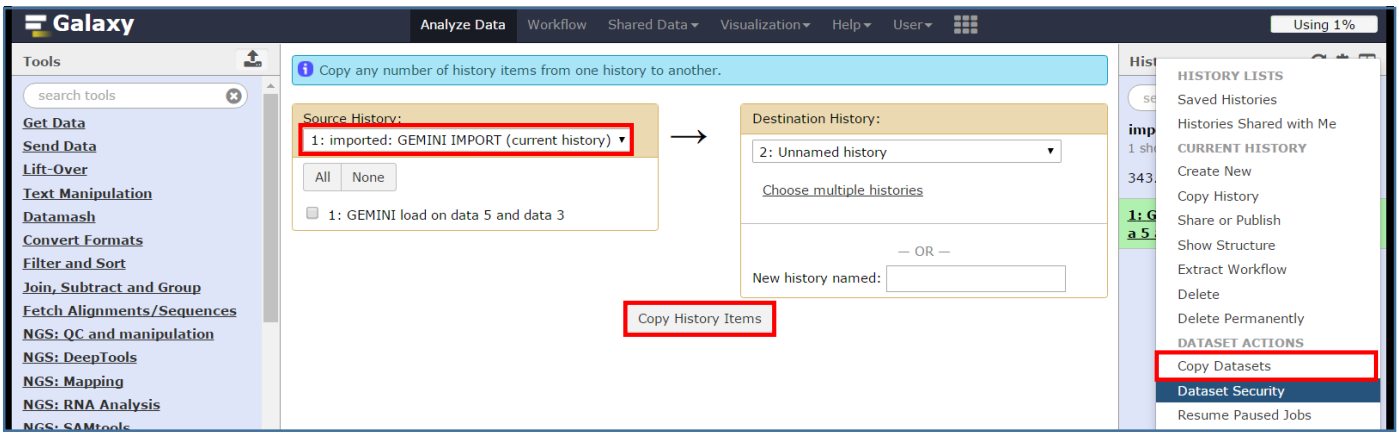
6. We're now ready to load the data into GEMINI using the options below. However, the loading step is computationally intensive and therefore very slow. To save time we will import a preloaded dataset so skip to step 7.



7. In a new browser tab enter this link https://usegalaxy.org/u/willtapper/h/gemini-import and click Import history. This will create a new history in your Galaxy account consisting of the GEMINI database for the macrocephaly trio.

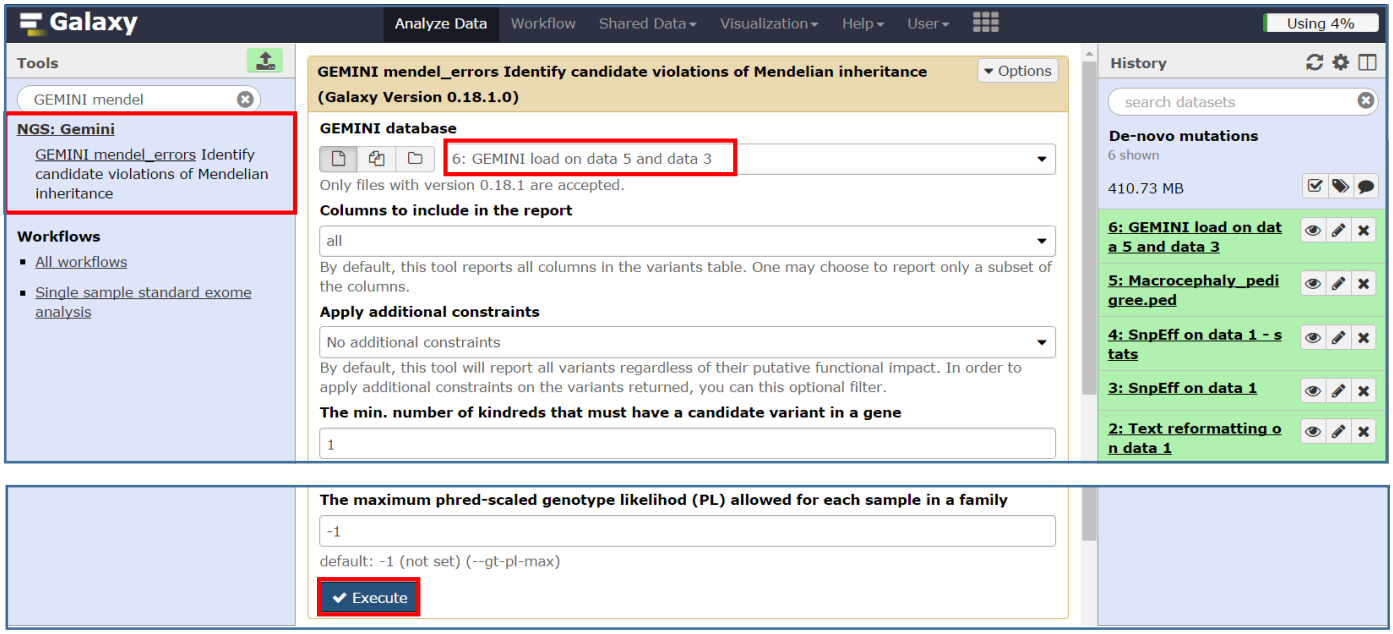


8. Click the cog icon in the history pane, select 'Copy Datasets', choose 'GEMINI IMPORT' as the source history and 'De-novo mutations' as the destination then click 'Copy History Items'.



We will now use GEMINI to identify all mendelian errors in the trio whereby an allele in the child could not have been received from either of its biological parents by Mendelian inheritance.

9. In **Tool Pane**: Go to **NGS: Gemini** > GEMINI mendel_errors Identify candidate violations of Mendelian Inheritance

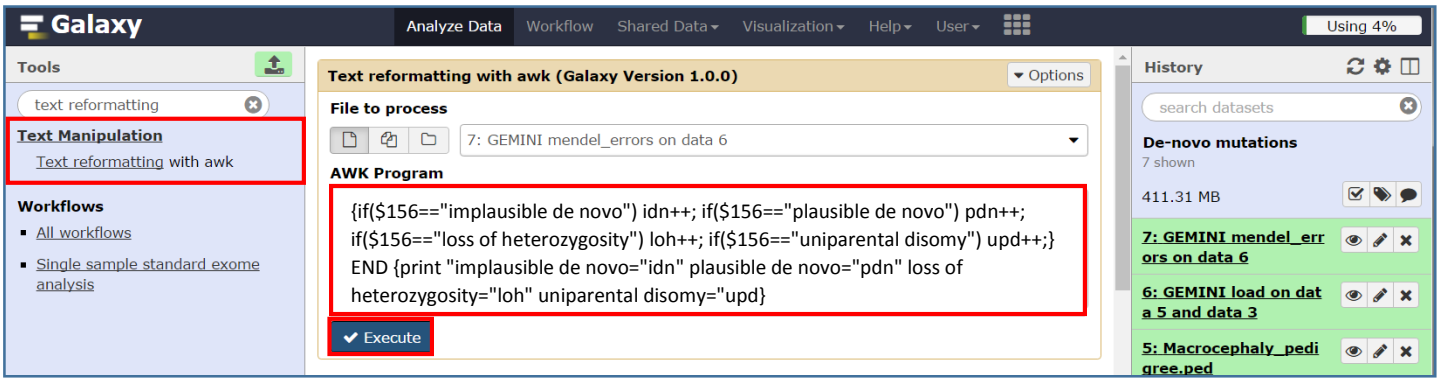


Categories of Mendelian error:

Category	Description	Child	Parent A	Parent B
Loss of heterozygosity (LOH)	Child and one parent are opposite homozygotes and the other parent is heterozygous	AA	BB	AB
Uniparental disomy (UPD)	Child is homozygous and the parents are opposite homozygotes	AA	AA	BB
Implausible de-novo mutations	Child is homozygous and both parents are homozygous for the opposite genotype to child	AA	BB	BB
Plausible de-novo	Child is heterozygous and parents have the same homozygous genotype as each other	AB	AA	AA

Use text reformatting with awk to count the number of variants in each category of Mendelian errors.

10. In **Tool Pane**: Go to **Text Manipulation** > Text reformatting with awk

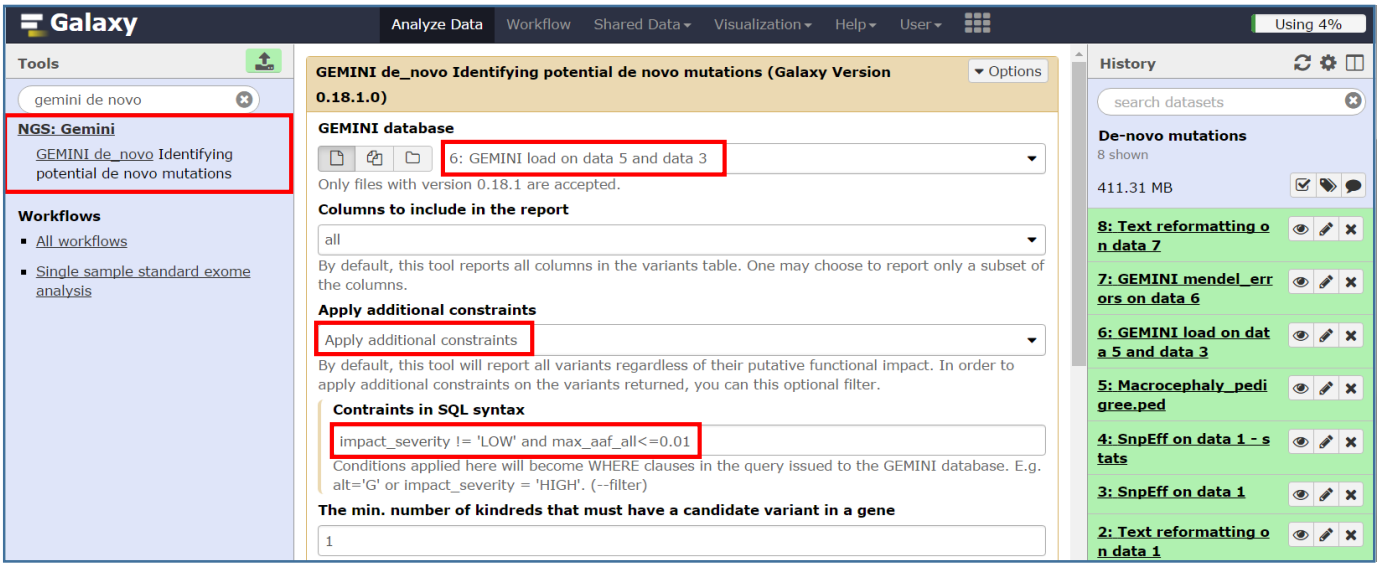


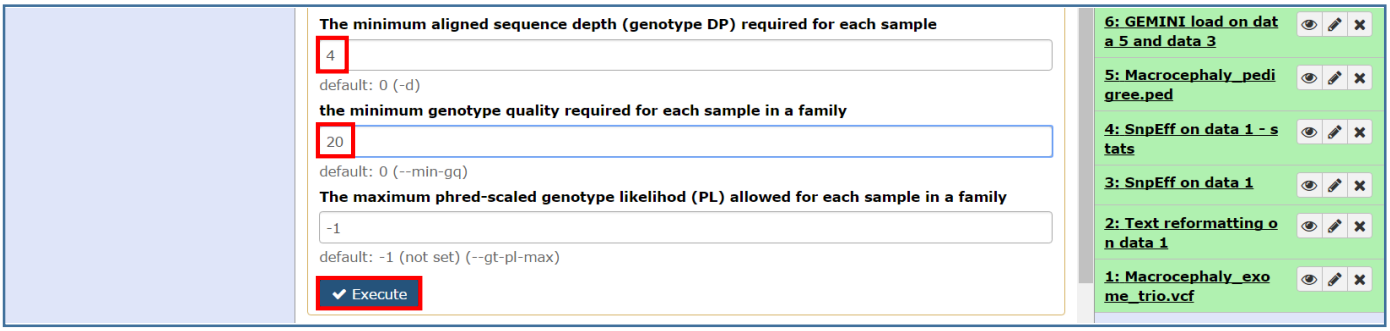
View the results and answer these questions.
Q5. How many variants are identified in each type of Mendelian error?
Q6. How does the number of de-novo mutations compare with expectation?

Based on the human mutation rate (~1.1×10^{-8} per bp per generation) and genome size o(3.1×10^{9}) there should be approximately 31 de-novo mutations per haploid genome and 0-2 mutations per exome since the exome is approximately 2% of the genome. Given the low mutation rate, the number of Mendelian errors can be used to estimate the genotyping error rate. Errors occur due to factors such as PCR artefacts, incorrect alignment and low read depth and are unevenly distributed throughout the genome. In a study by Patel et al 2014, approximately 95% of Mendelian errors were removed while retaining 80% of the called variants by applying filters based on read depth (DP<15 removed), genotype quality (GQ<20 removed) and alternate allele ratio (homozygous reference variants with alternate ratio >0.15 were removed, homozygous alternate variants with alternate ratio <0.85 were removed, and heterozygous variants with alternate ratio <0.3 or >0.7 were removed).

Use GEMINI to select all plausible de-novo mutations with: a maximum alternate allele frequency in public databases of <=0.01, an estimated impact severity of medium or high, a minimum read depth of 4 in all samples and a minimum genotype quality of 20 in all samples.

11. In **Tool Pane**: Go to **NGS: Gemini** > Gemini de_novo Identifying potential de novo mutations





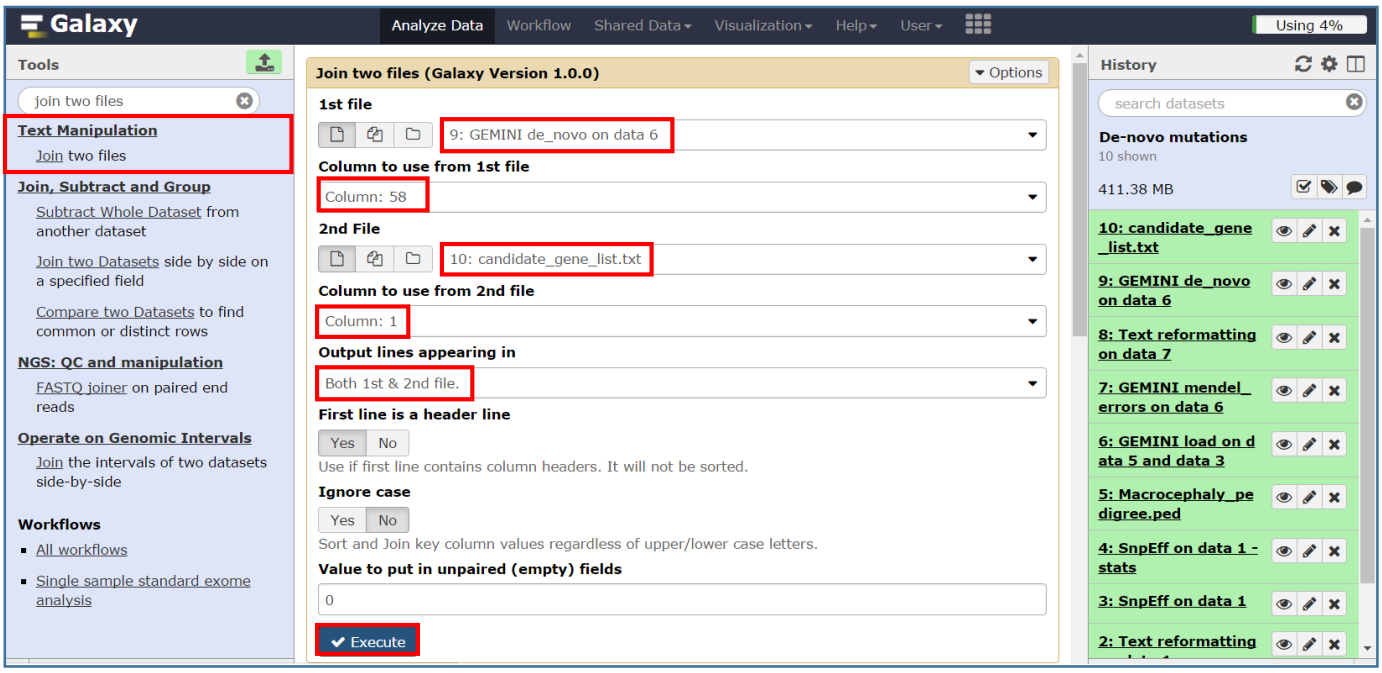
View the results and answer the question below.
Q7. How many plausible de-novo variants meet the search criteria?

A list of candidate has been generated using Phenomizer to assess the patients phenotype. Upload the

12. Upload the list of candidate genes from the USB key in the 'FILTERING STRATEGIES' folder (candidate_gene_list.txt) to galaxy, select tabular for type and leave the genome blank.

Join the list of plausible de-novo variants with the candidate gene list.

13. In **Tool Pane:** Go to **Text Manipulation** > Join two files



Download the joined file and open it in Excel to view the results.
Q8. Which variant is most likely to cause the disease in the child?
Q9. What proportion of reads have the mutant allele (mutant allele frequency)?
Q10. Is there anything unusual about the mutant allele frequency and if so how could it be related to the disease?

Identifying genetic targets of acquired uniparental disomy

The concept of uniparental disomy (UPD) was introduced in the de-novo mutation section as a type of Mendelian error whereby the child is homozygous (AA) and the parents are opposite homozygotes (AA and BB). Inherited UPD can occur when a person receives both copies of a chromosome pair, or parts of chromosomes, from one parent. Constitutional UPD is associated with developmental disorders caused by the abnormal expression of imprinted genes, i.e. genes that are differentially expressed depending on whether they have been maternally or paternally inherited. For example, Prader-Willi syndrome and Angelman syndrome can be caused by UPD or other errors in imprinting involving genes on the long arm of chromosome 15.

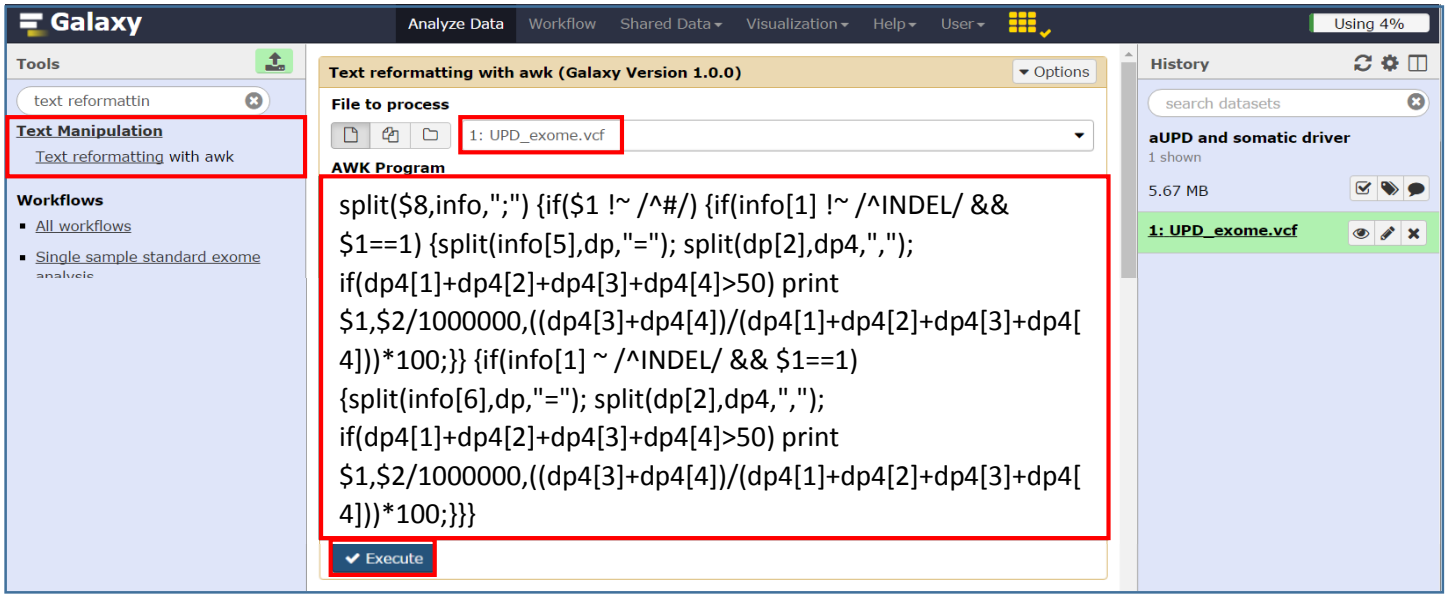
By contrast, somatically acquired UPD (aUPD) in cancer is a mechanism by which a pre-existing driver mutation (usually somatically acquired) is converted to homozygosity, thereby providing an additional clonal advantage. aUPD may involve whole chromosomes as a result of non-disjunction or, more commonly, whole chromosome arms or terminal segments as a consequence of mitotic recombination. Regions of aUPD can be identified by NGS as long tracts of allelic imbalance (ie deviation from the expected percentage of alternate reads for heterozygous variants which should be approximately 50%) and many genes involved in a wide range of cancers have been identified by detecting minimal regions of recurrent aUPD and searching these regions for functionally relevant genes.

The aim of this example is to identify a region of aUPD and the underlying somatic driver mutation from whole exome sequencing of tumour DNA in a patient with polycythemia vera. The VCF file was generated using BWA-MEM for alignment to hg19, samtools for variant calling and selection of variants with a minimum read depth of 4 and minimum QUAL score of phred 20.

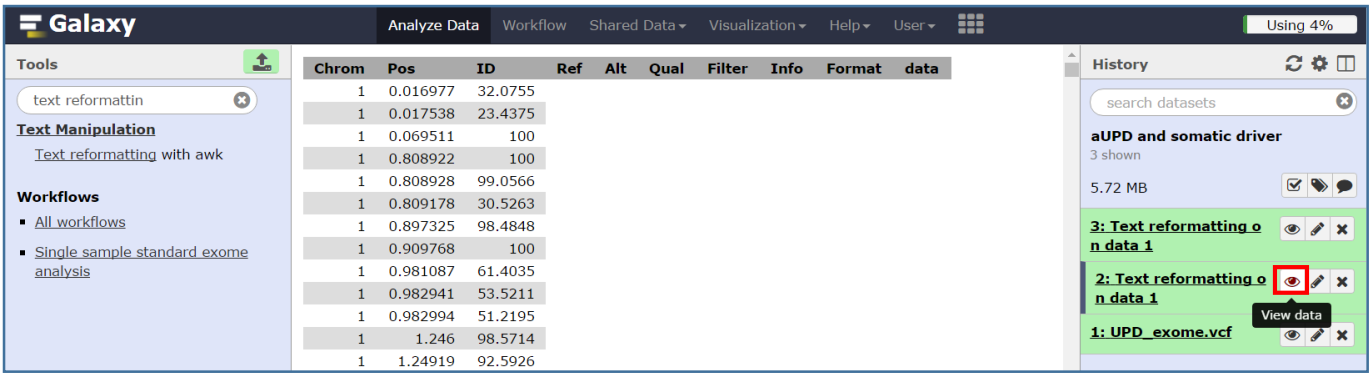
1. Create a new history for example 3 (Click on name, type 'aUPD and somatic driver', press return)
2. Upload the tumour VCF from the USB key in the 'FILTERING STRATEGIES' folder (UPD_exome.vcf) to galaxy (https://usegalaxy.org/), select VCF for type and Human Feb. 2009 (GRCh37/hg19)(hg19) for reference genome.

Use awk to determine the percentage of alternate reads for all variants on chromosome 1 with a read depth greater than 50.

3. In **Tool Pane**: Go to **Text Manipulation** > Text reformatting with awk

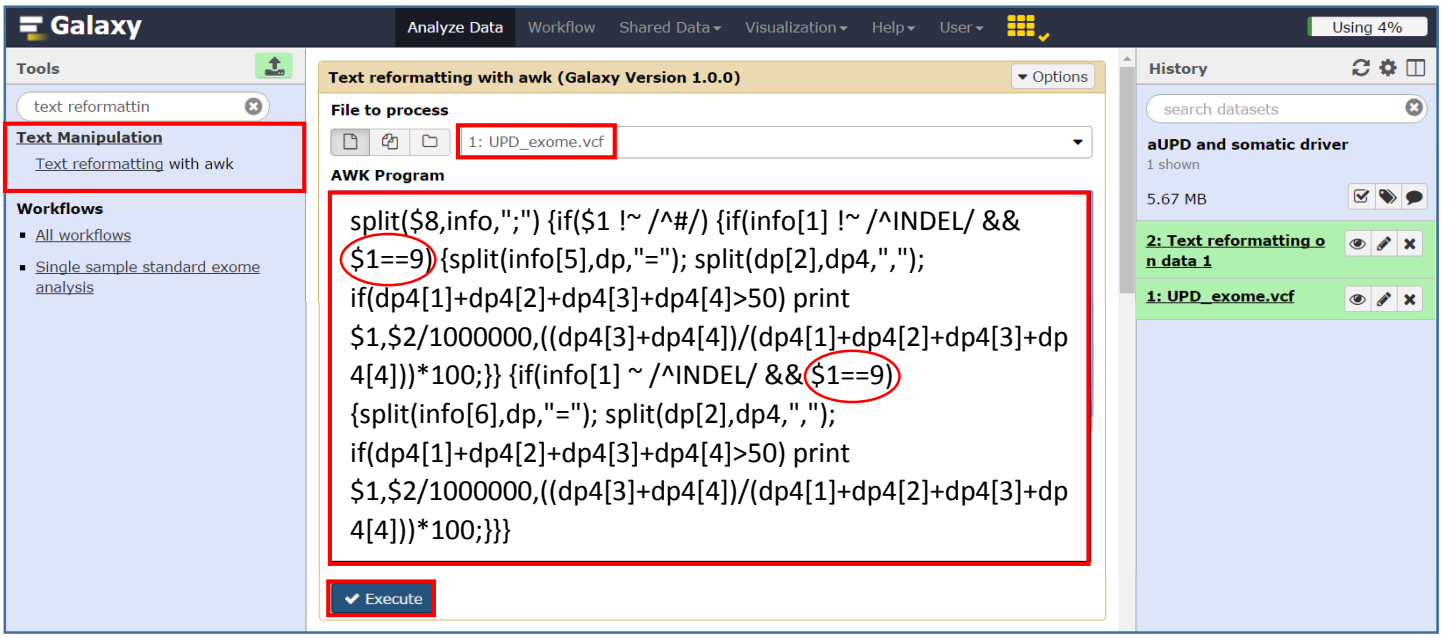


The output produced by this awk code consists of three columns of data for chromosome, sequence location in megabases and percentage of alternate reads for each variant as rows.



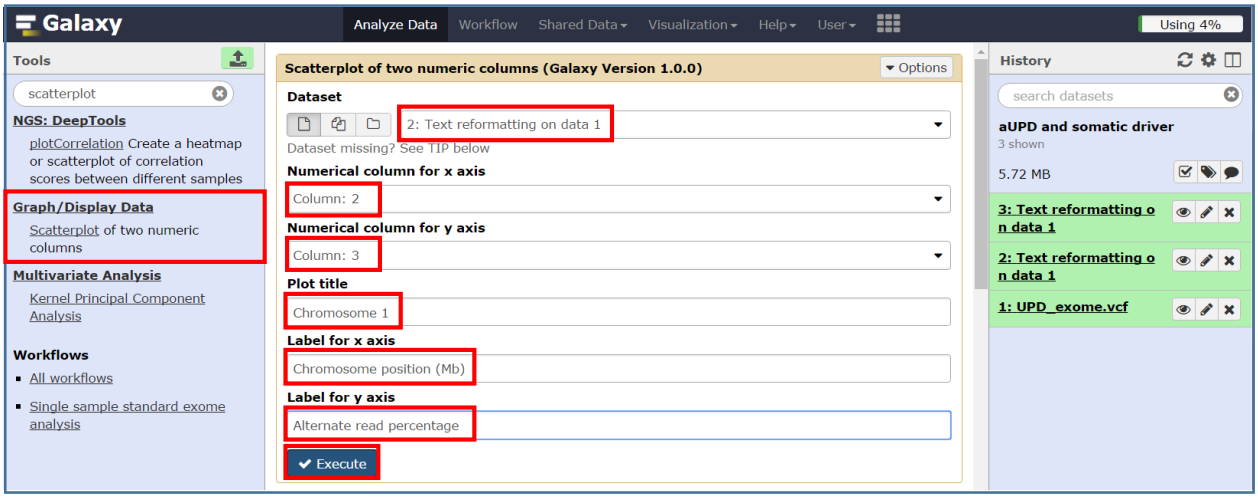
Repeat the text reformatting for chromosome 9 by changing the two instances of "$1==1" to "$1==9".

4. In **Tool Pane**: Go to **Text Manipulation** > Text reformatting with awk

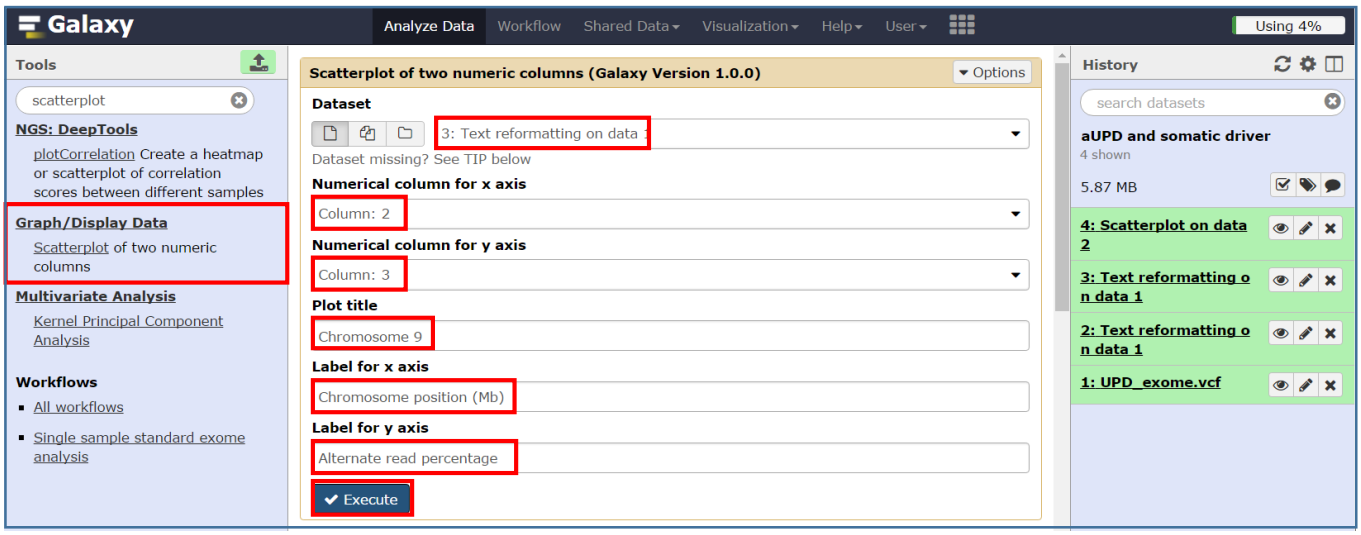


Make a scatter plot for chromosome 1 and chromosome 9 with the alternate read percentages on the Y-axis and megabase location on the X-axis.

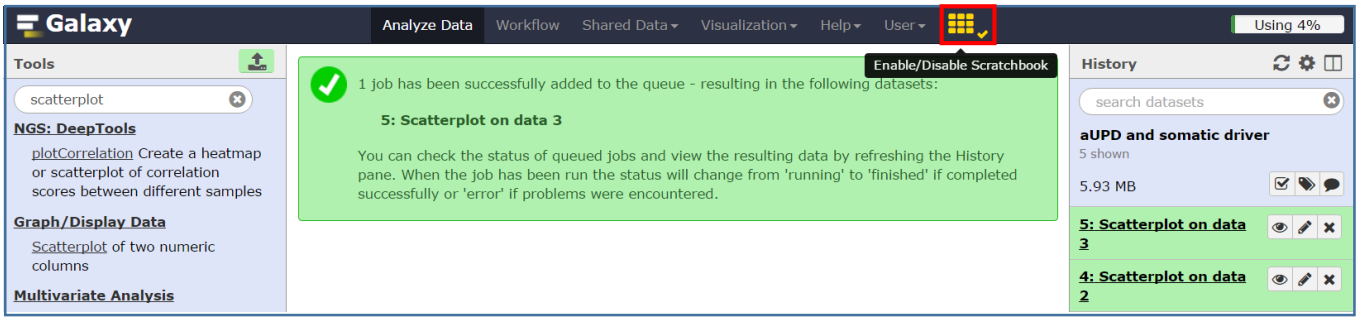
5. In **Tool Pane**: Go to **Graph/Display Data** > Scatterplot of two numeric columns



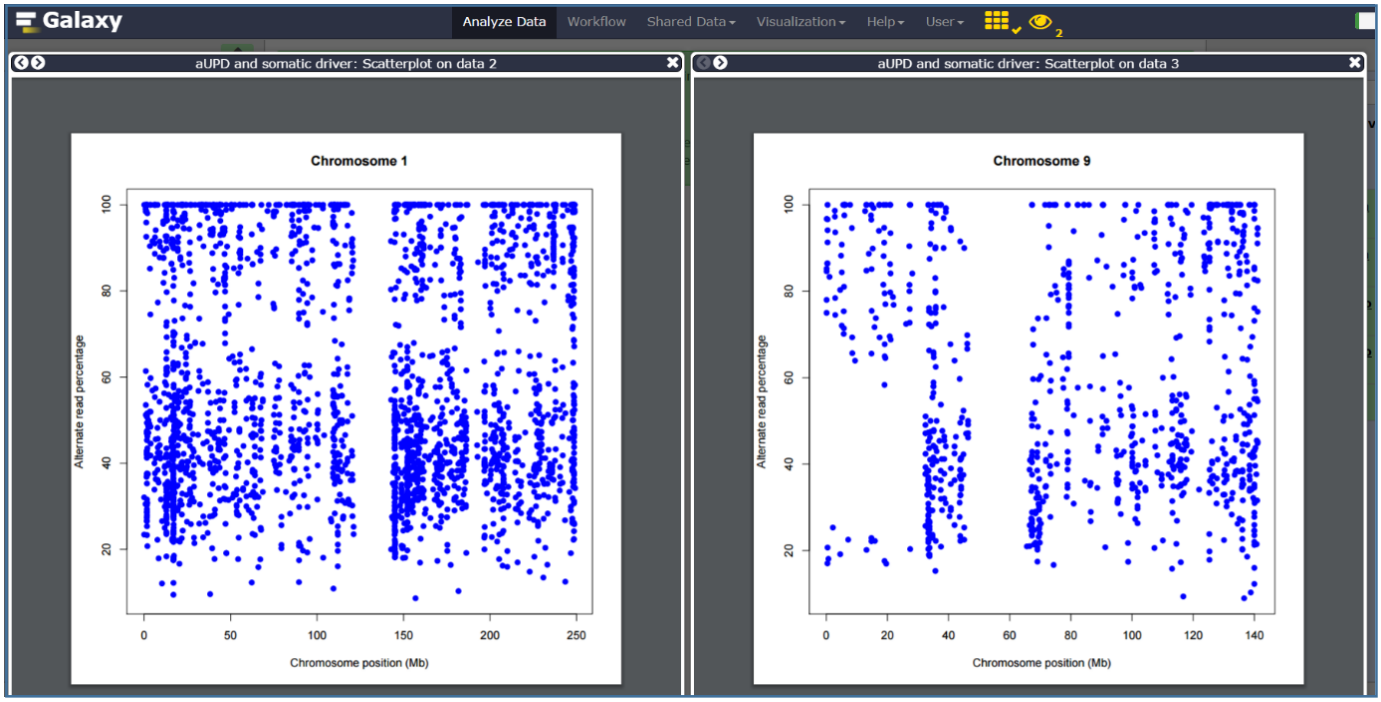
6. In **Tool Pane**: Go to **Graph/Display Data** > Scatterplot of two numeric columns



Enable the Scratchbook and view both alternate read frequency plots side by side for comparison by clicking the view icon for steps 4 and 5.



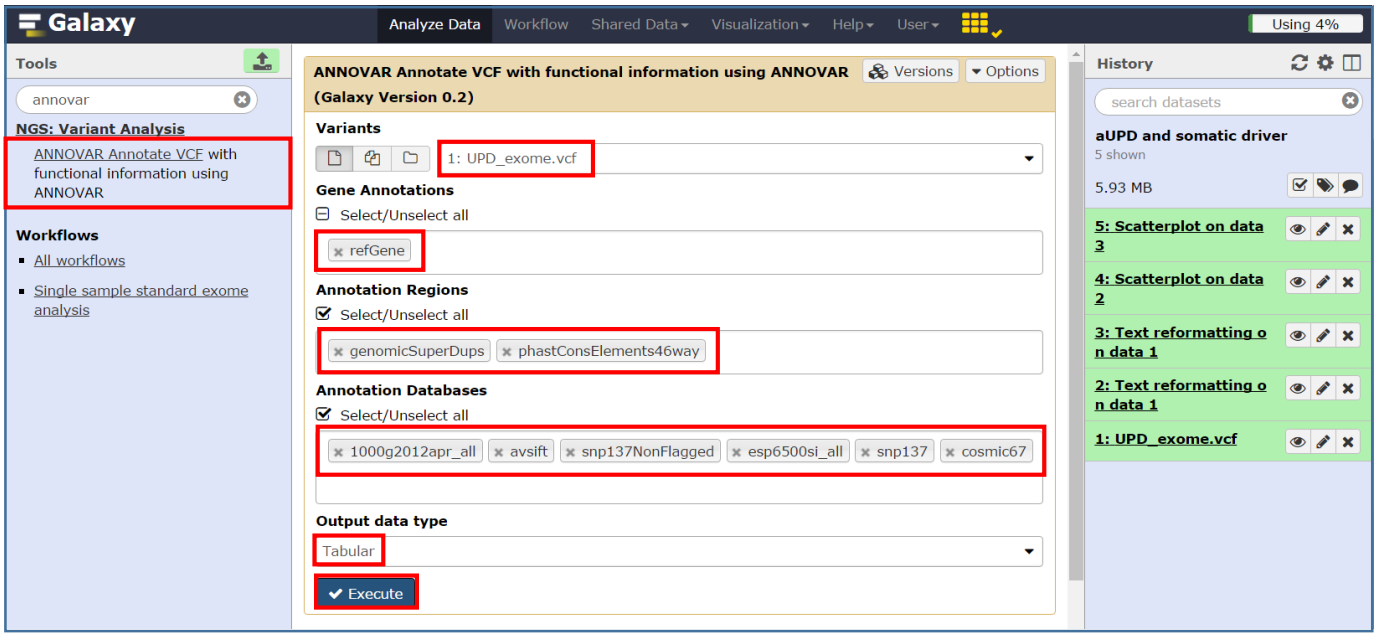
The plots should look like this.



Q11. Does either chromosome have any evidence for aUPD and if so what are the approximate genomic coordinates in megabses?

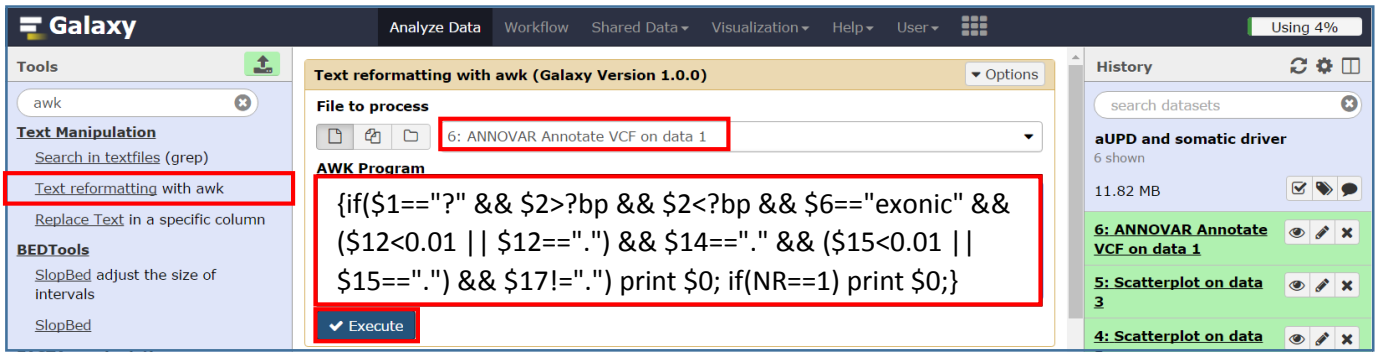
Annotate the VCF file using ANNOVAR.

7. In **Tool Pane**: Go to **NGS: Variant Analysis** > ANNOVAR Annotate VCF with functional information using ANNOVAR



Use awk to filter the annotated VCF file. Replacing the question marks with coordinated for the potential aUPD region ($1==”?chr” && $2>?bp and $2<?bp). Also select coding variants ($6 == “exonic”) that are either rare or absent from databases of known variation such as 1000 genomes ($12 <0.01 || $12==”.”), the non flagged version of dbSNP ($14 ==”.”), the Exome Sequencing Project ($15<0.01 || $12==”.”) and are present in the COSMIC database of somatic mutations ($17 != ‘.’).

8. In **Tool Pane**: Go to **NGS: Variant Analysis** > ANNOVAR Annotate VCF with functional information using ANNOVAR



View the filtered results and answer question 12.
Q12. Which variant is most likely to cause the clonal proliferation of erythrocytes (polycythemia vera) in this patient?

Well done Bioinformaticians you finished all exercises!