**Overview**

Accurate and consistent variant calling requires statistical modelling and is essential for the clinical implementation of NGS. However, many programs are available for calling variants and their concordance varies. Furthermore, variants have different levels of confidence due to differences in data quality. For variants with intermediate confidence levels, it is difficult to separate true variation from artefacts that arise from many factors such as sequencing error, misalignment and inaccurate base quality scores. As a result, the evidence for variant calls requires scrutiny and caution should be used when interpreting positive and negative findings especially for indels which are more error prone. At the end of this exercise you will be able to:

**1.** Use a range of software (GATK, Varscan, SAMtools, and BCFtools)to call small variants (SNVs and indels)

**2.** Describe the contents of pileup and variant call files

**3.** Generate and interpret variant quality control parameters (quality score, genotype quality, sequence context, strand bias, base quality bias, mapping bias, tail bias, variant density, concordance with known variation dbSNP, heterozygous to homozygous ratio and transition to transversion ratio)

**4.** Use quality control filters to exclude or flag variants with low confidence

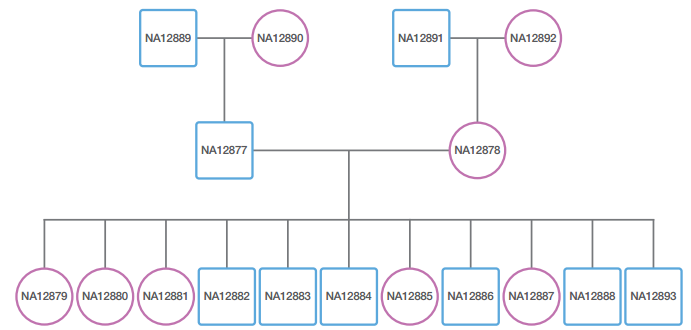
**5.** Calculate the concordance between VCF files

**6.** Assess the sensitivity and precision of variant callers by comparison with a catalog of highly accurate whole-genome variant calls

**Trial data**

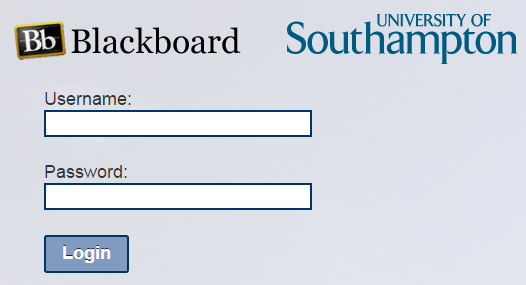
The data for analysis is from a healthy Caucasian woman (NA12878) belonging to CEPH pedigree 1463 (Figure 1, <https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=GM12878>). As part of the Platinum Genome Project (<http://www.illumina.com/platinumgenomes/>), all 17 members of this pedigree have been whole-genome sequenced (WGS) at 50x coverage on an Illumina HiSeq 2000. Several pipelines were used to analyse this data and account for the inheritance structure in order to identify a set of highly accurate whole-genome variant calls for individual NA12878.

Starting from aligned WGS data for individual NA12878 (BAM file), our aim is to use a range of software to call variants and to assess the sensitivity and specificity of these programs by comparing their variant calls with the high quality variant calls from Platinum Genomes. The data we will analyse was generated as part of the 1000 genomes project (<http://www.1000genomes.org>) and was not used by the Platinum Genome project to create the high confidence calls. To make the data manageable, it has been restricted to a 2Mb region of chromosome 20 between 1 to 2Mb.

**Figure 1.** CEPH pedigree 1463

**Let’s begin**

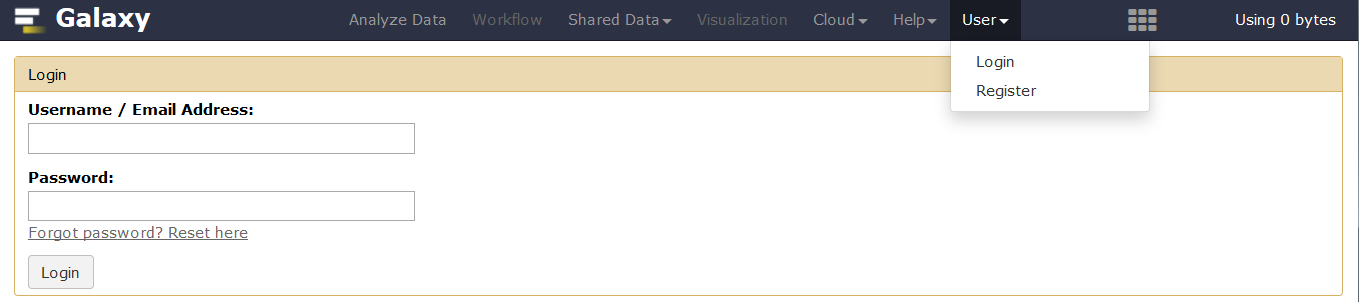
**1.** Login to blackboard: <https://blackboard.soton.ac.uk/>



**2.** Navigate to the course resources and download these files to your computer;

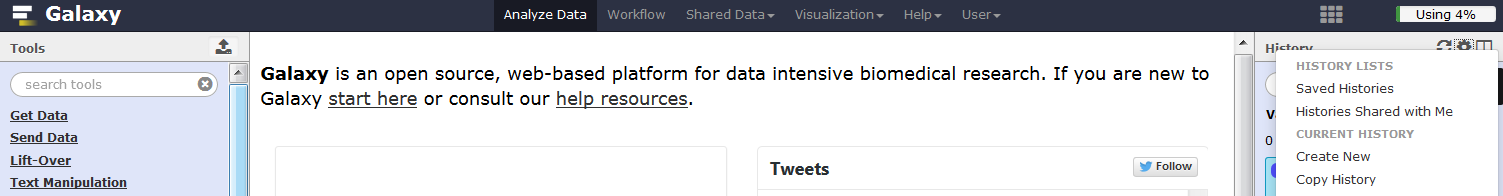
* Aligned data (NA12878\_chr20\_2mb\_filtered\_bam.bam)
* High quality variant calls (NA12878\_20\_2mb\_IlluminaPlatinum.vcf)
* Polymorphic sites from dbSNP (dbsnp\_132.hg19.excluding\_sites\_after\_129\_22.vcf)

**3.** Go to <https://usegalaxy.org/> and login to your Galaxy account

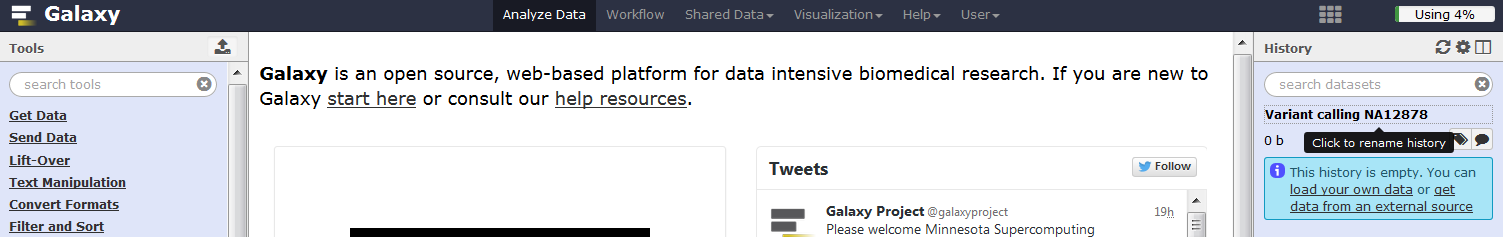


Login will take you to your last active history (Analysis of WES01).

**4**. Create a new history



**5.** Rename the history for this session (Click on name, type new name, press return)



Upload the aligned data (NA12878\_chr20\_2mb\_filtered\_bam.bam) from your computer to Galaxy

**6.** In **Tool Pane**: Go to **Get Data** > Upload File from your computer

Choose local file, select type ‘bam’, select genome ‘b37’, click start

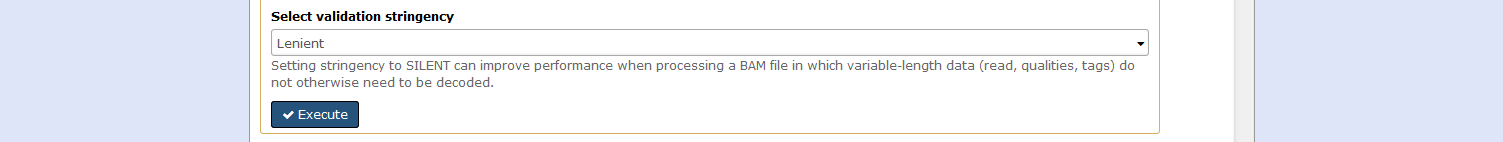
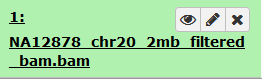
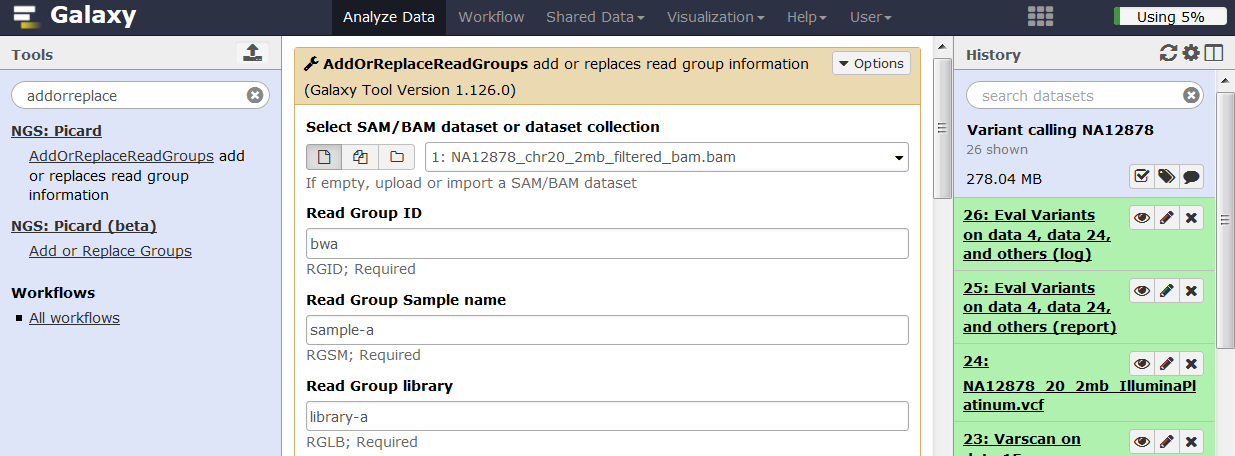


**Add read group information to the BAM file**

To use GATK programmes we need to add some read group information to the bam file.

**1.** In **Tool Pane**: Go to **NGS: Picard** > AddOrReplaceReadGroups

Change Read Group ID and Read Group Library, keep the other settings and click execute



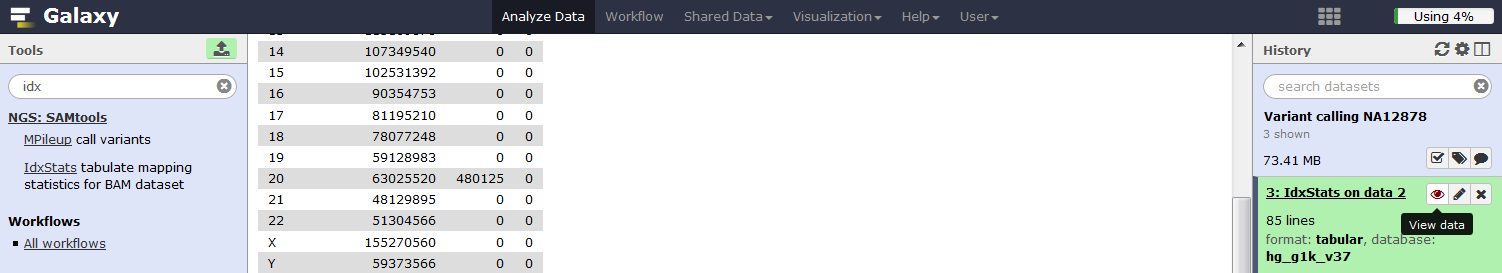
**Check the aligned data before calling variants**

The data were generated by paired-end sequencing using an Illumina HiSeq 2000 at 24x coverage with read lengths of 30bp. To make the data more manageable is has been reduced to a 2Mb region of chromosome 20. Use IdxStats to check that the data is mapped to chromosome 20 and determine the number of mapped reads.

**1.** In **Tool Pane**: Go to **NGS: SAMtools** > IdxStats

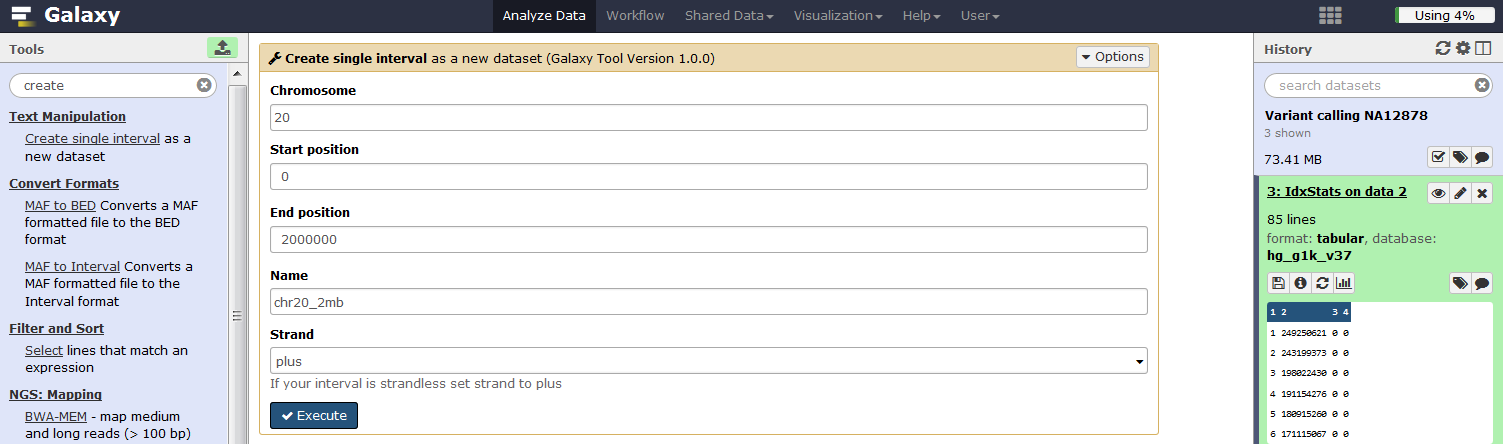


**2.** View the IdxStats result and make a note of the number of reads mapped to chromosome 20

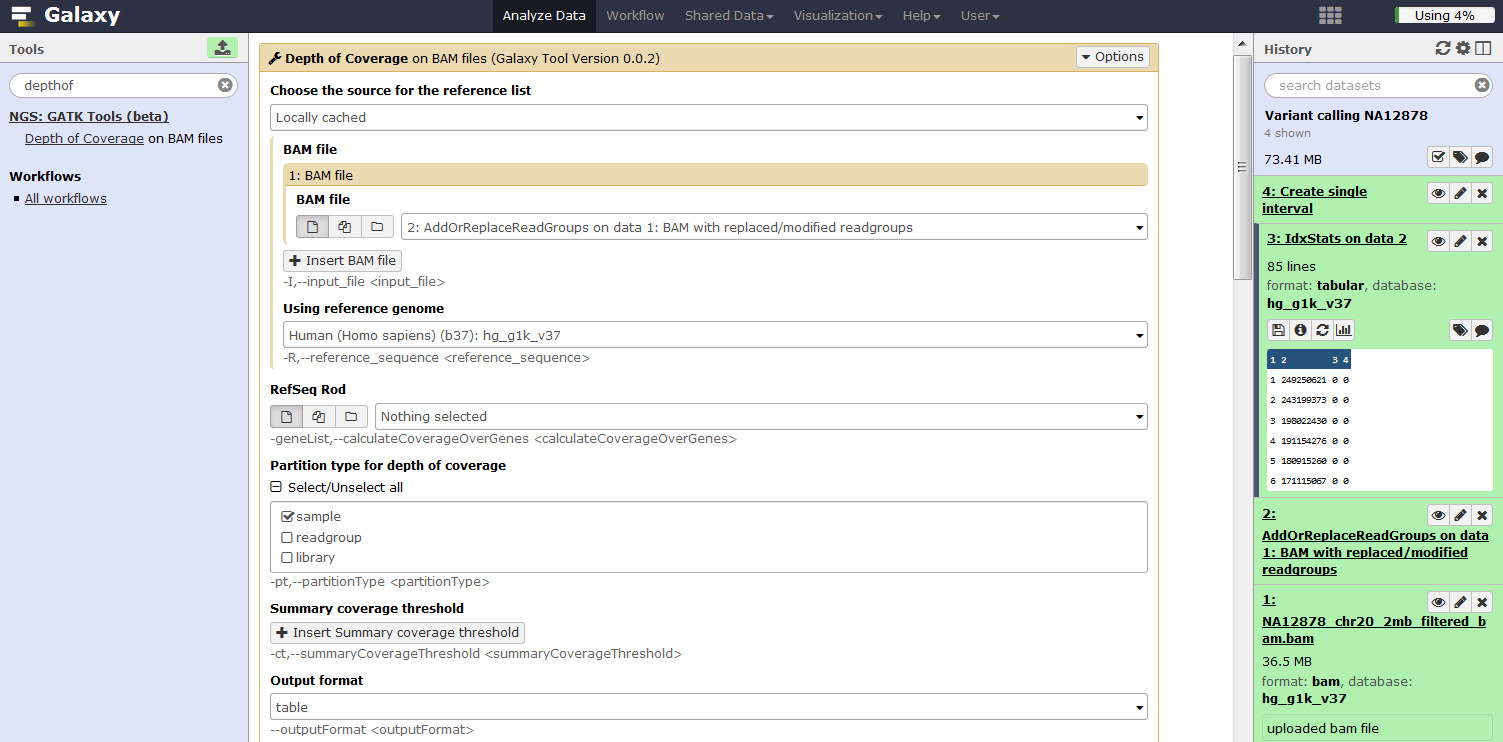
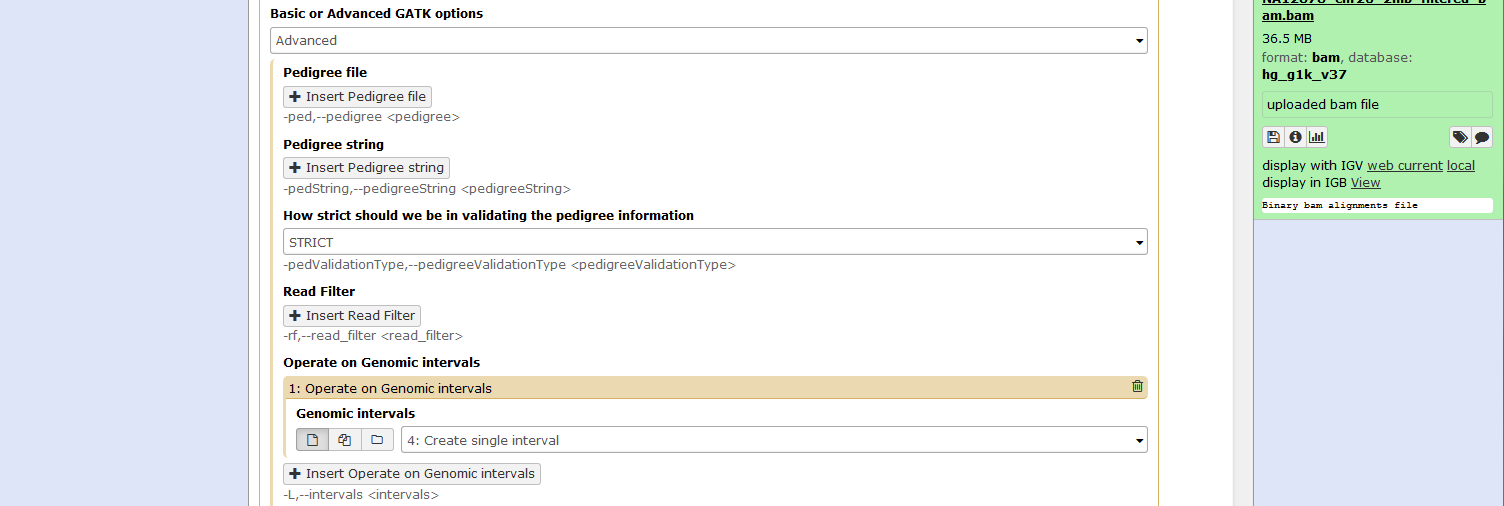
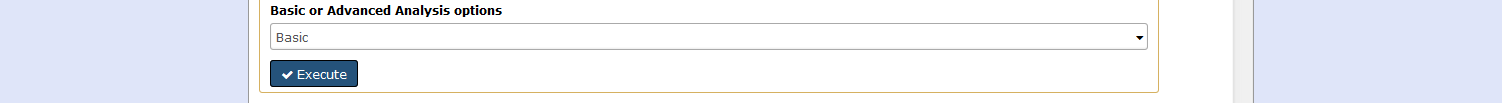


To calculate the depth of coverage you will need to create a bed file that describes the location (chromosome and base pair coordinate) of the sequenced region.

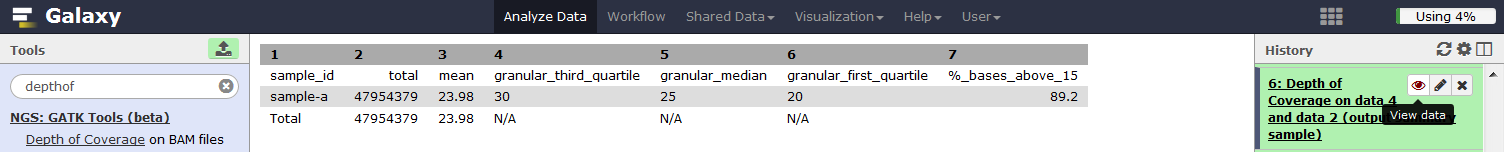
**3.** In **Tool Pane**: Go to **Text Manipulation** > Create single interval



**4.** In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > Depth of Coverage



**5.** View output 6 ‘Depth of coverage on data… (output summary sample) and answer the following;

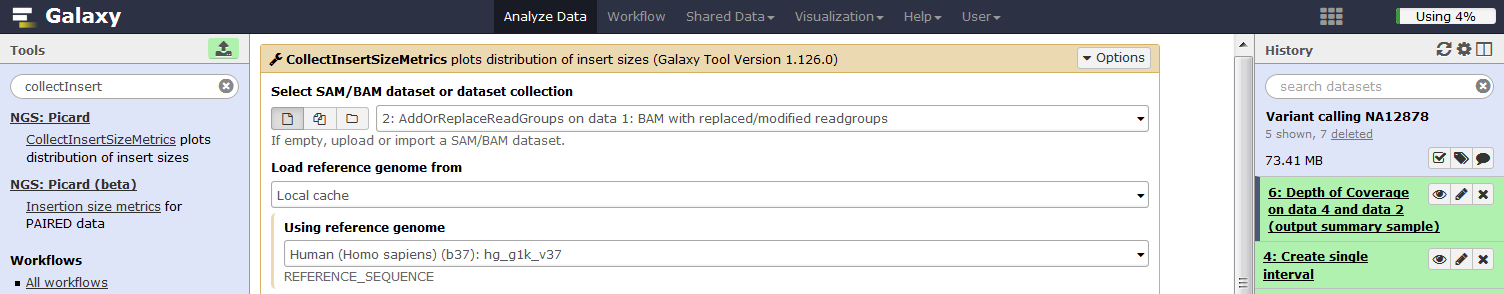


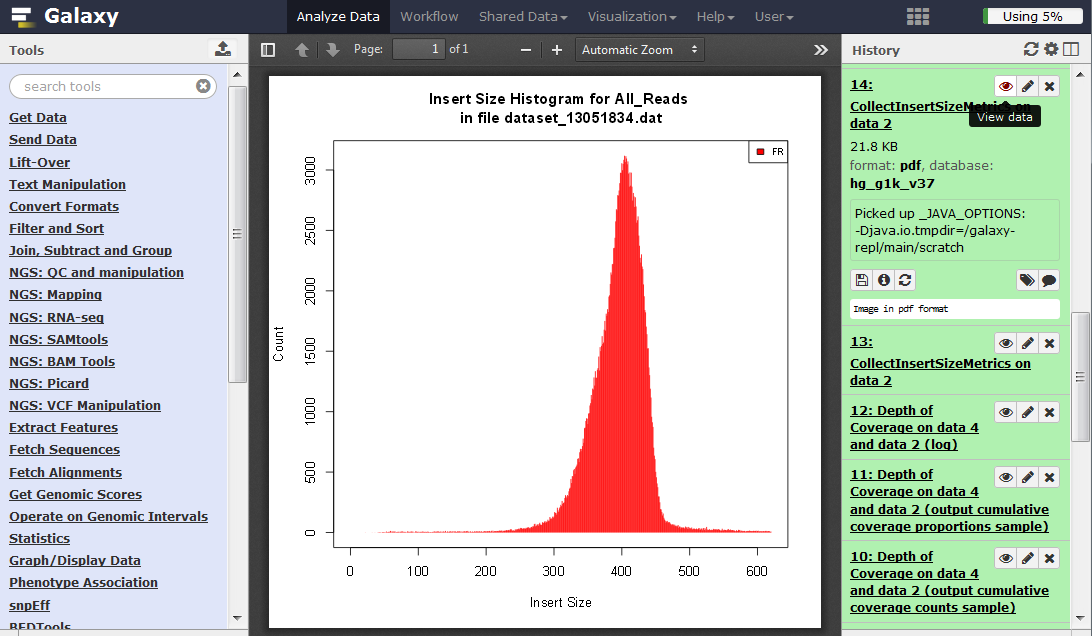
**Q1.** What is the mean coverage?

**Q2.** What percentage of target bases are covered by 15 or more reads?

Depth of coverage gives other statistics broken down by regions etc. We’re not interested in these stats so to keep the history clean use the ‘X’ button to delete the outputs for step 5 and steps 7 to 12.

**6.** Find the distribution of insert sizes. In **Tool Pane**: Go to **NGS: Picard** > CollectInsertSizeMetrics



**7.** View the results for steps 13 and 14.

**Q3.** What is the mean insert size and how does it compare with the Whole-Exome sequence data?

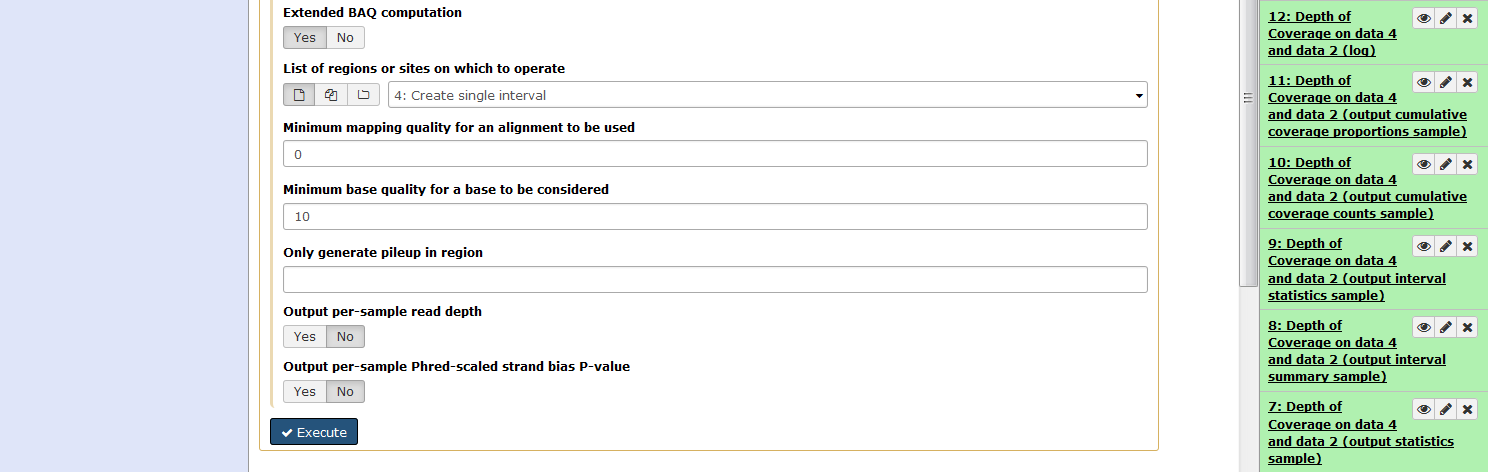
The bam file we are using has been QC filtered (duplicate, non-primary, and unmapped reads removed), the reads are sorted by chromosome and base pair location and read group information has been added so it is ready for variant calling. We will now use SAMtools to identify variants.

**Call variants using SAMtools MPileup and bcftools**

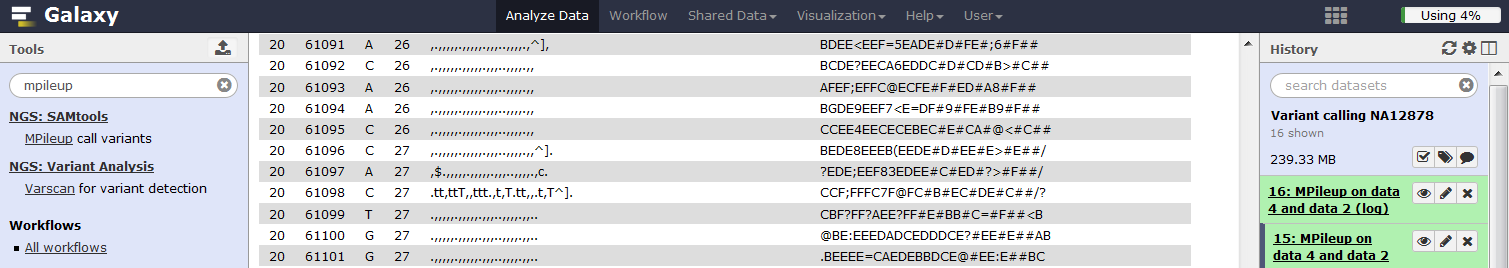
Calling variants with SAMtools Galaxy tool version 0.0.1 (Li et al 2009) is a two-step process, which uses a general Bayesian framework. In the first step, MPileup is used to compute the likelihood of data given each possible genotype. In the second step, the view command of bcftools is used to call variants by picking the base that maximises the posterior probability with the highest Phred quality score.

Use SAMtools MPileup to generate a ‘pileup’ file that describes the raw data for variant calling consisting of the read bases for reference and alternate alleles and their sequence qualities. Pileup files facilitate SNP/indel calling and manual viewing of the data. Use the ‘Set advanced options’ and select ‘Yes’ for ‘Extended BAQ computation’ so that Base Alignment Quality is calculated by probabilistic realignment. BAQ represents the probability that a read base is mis-aligned. In general, this setting increases sensitivity and helps to exclude false positives due to alignment errors caused by nearby indels but decreases specificity. To reduce computing time restrict the analysis to the sequenced region using the ‘List of regions or sites on which to operate’ option. Reduce the ‘Minimum base quality for a base to be considered’ to 10 and click execute.

**1.** In **Tool Pane**: Go to **NGS: SAMtools** > MPileup



**2.** View the MPileup file and scroll to 61098 bp which is the first high confidence ‘platinum’ variant.



The columns in the pileup file are **chromosome, location, reference base, number of reads covering the site, read bases and base qualities**. In the read base column, a dot stands for a match to the reference base on the forward strand, a comma for a match on the reverse strand, ‘ACGTN’ for a mismatch on the forward strand and ‘acgtn’ for a mismatch on the reverse strand.

**Q4.** Use the pileup file to complete the table below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chr. | Bp | No. reference reads | | No. alternate reads | |
| Forward strand (.) | Reverse strand (,) | Forward strand (ACGTN) | Reverse strand (acgtn) |
| 20 | 61098 |  |  |  |  |

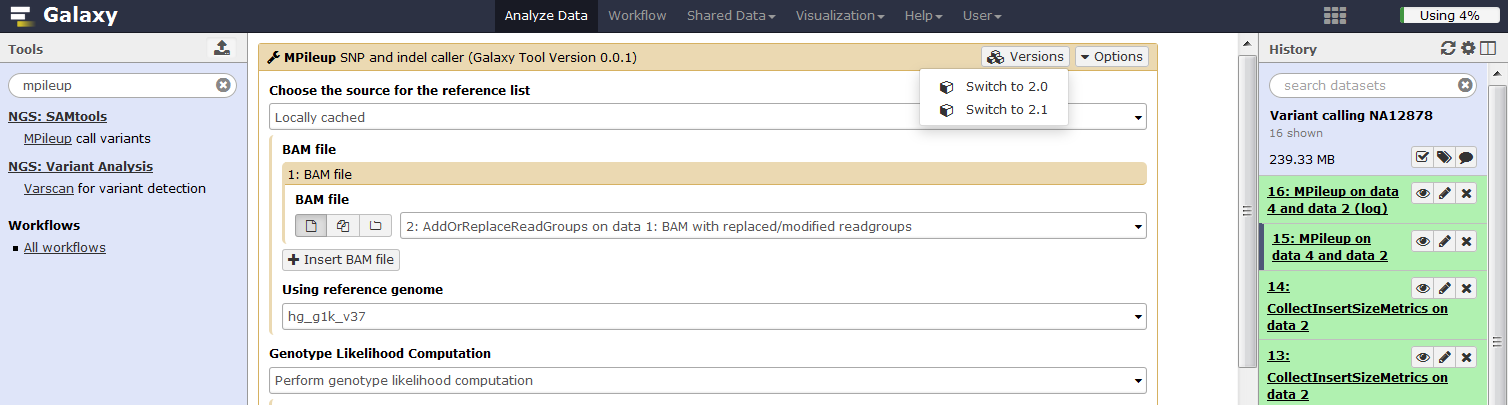
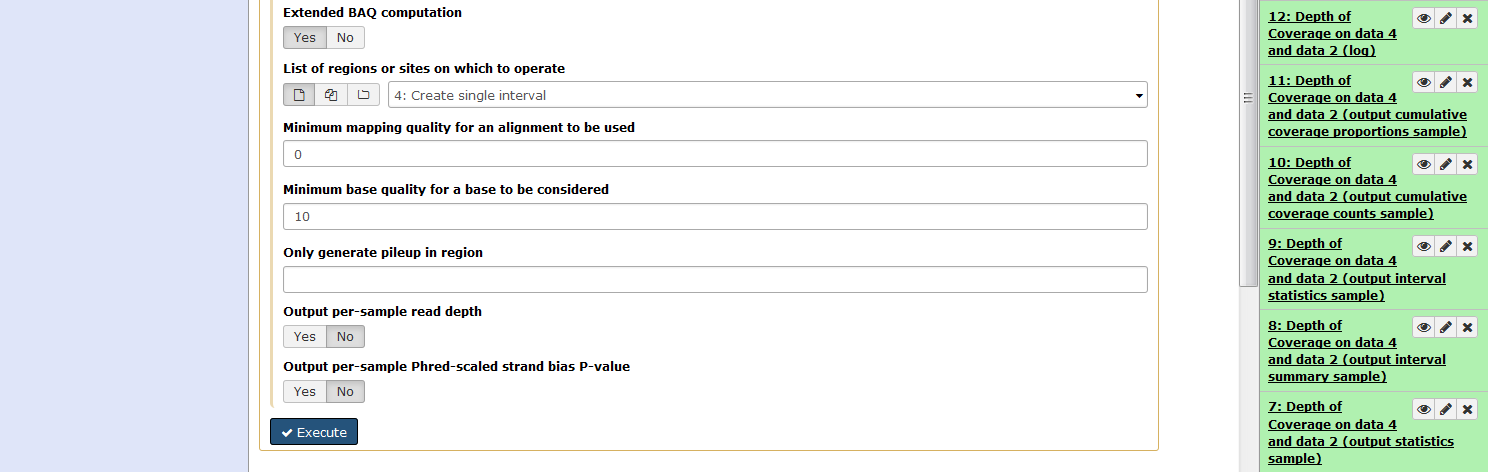
In the pileup file, insertions are represented as +[0-9ACGTNacgtn] where the integer gives insertion length followed by the sequence on either the positive or negative strand. For example, two reads with a 2bp insertion of AG one on the forward strand and one on the reverse strand is represented as +2AG+2ag. Deletions are shown by a minus sign.

Other characters in the read base string indicate:

* ^ (caret) marks the start of a read segment, the following character gives mapping quality
* $ (dollar) marks the end of a read segment
* \* (asterisk) is a placeholder for a deleted base in a multiple basepair deletion

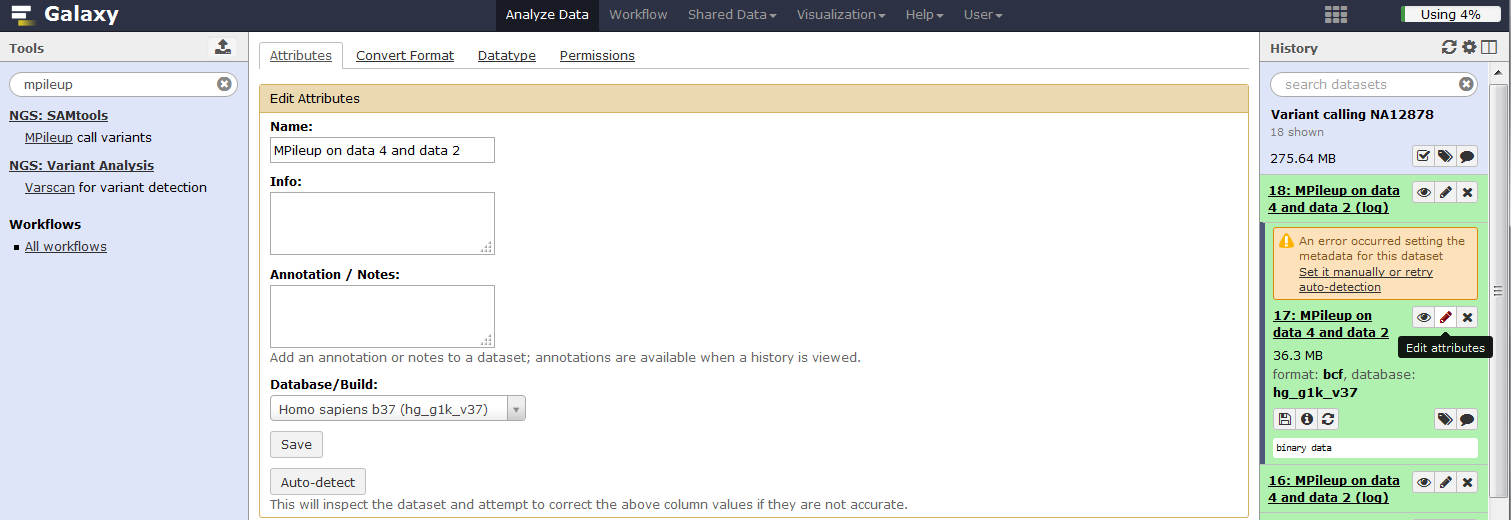
Repeat MPileup but this time perform genotype likelihood calculation to give the likelihood of data given each possible genotype.

**3.** In **Tool Pane**: Go to **NGS: SAMtools** > MPileup



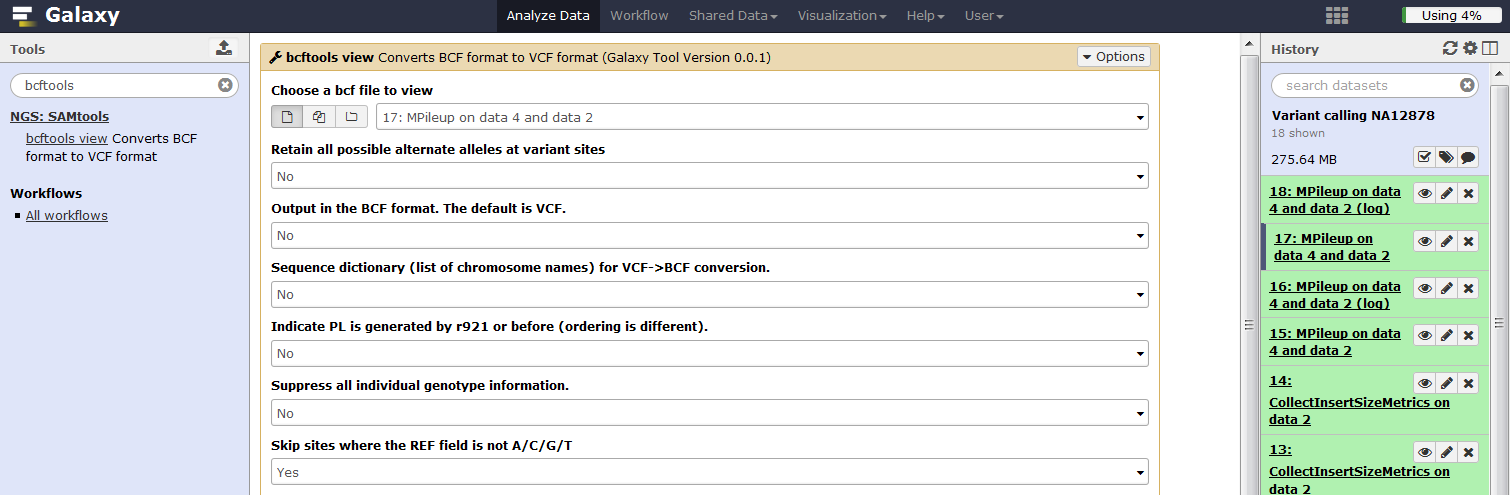
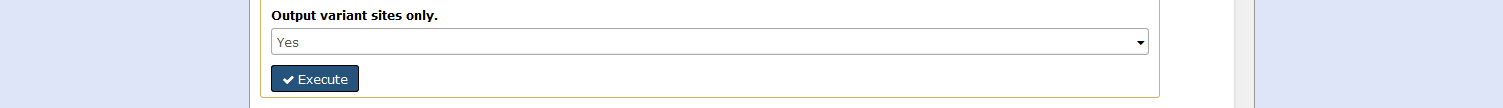
You will notice that MPileup gives an error message “An error occurred setting the metadata for this dataset”. Use the Edit attributes option to set the metadata as shown below.

**4.** Edit the attributes of the MPileup output.



Now use bcftools with default settings to assess the genotype likelihoods from MPileup and to call variants.

**5.** In **Tool Pane**: Go to **NGS: SAMtools** > bcftools view

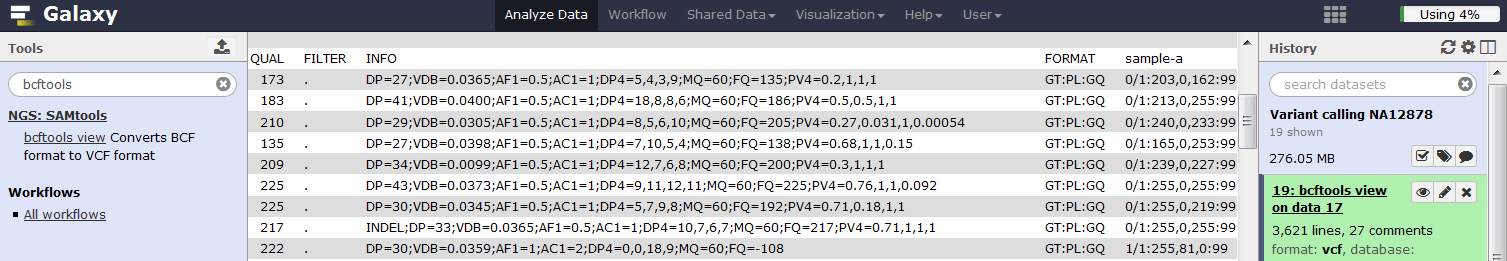


The output of bcftools view is in Variant Call Format (VCF), which is a standard way of encoding genetic variation including SNVs and indels. The VCF format is described in detail here: <http://samtools.github.io/hts-specs/VCFv4.1.pdf>.

**6.** Use edit attributes to change the datatype of the bcftools output to VCF.



**7.** View the VCF file created by bcftools.



The history panel shows the VCF file contains 3,621 variants and 27 comments. The comments appear at the top of the VCF file (lines begin with a ‘#’ character) and explain the format of the info and sample columns. For the first high confidence ‘platinum’ variant at 61098 bp, important fields in the qual, info and sample columns are as follows;

* Qual=173; Phred scaled evidence level for the alternate allele.
* DP=27; the variant is covered by 27 reads
* DP4=5,4,3,9; reads used for variant calling, 5 on the forward strand and 4 on the negative strand with the reference allele, 3 on the forward strand and 9 on the negative strand with the alternate allele. Six reads with base qualities less than phred 10 were not used.
* PV4=0.2,1,1,1; p-values for strand bias, base quality bias, mapping quality bias and tail bias.
* GT=0/1; Genotype, the variant is heterozygous.
* PL=203,0,162; Phred-scaled likelihoods for the three possible genotypes (0/0, 0/1, and 1/1). The values are normalized so that the most likely genotype scores 0 and the others are scaled relative to the most likely genotype.
* GQ=99; Genotype quality is the Phred-scaled confidence that the genotype is correct, with a maximum of 99 because larger values are not more informative.

Variants with low Qual (<20), DP (<10), GQ (<20), or significant PV4 values (<0.05) can be flagged or filtered as potential false positives. For example;

**Strand bias:** uses a Fishers 2x2 exact test to evaluate the distribution of reads mapping to the forward and reverse strand for the reference and alternate allele, which should be similar. A significant strand bias is suggestive of a sequencing error that could exaggerate the amount of evidence for a particular allele resulting in a false positive variant.

**Base quality bias:** uses a t-test to determine if the average sequence quality is similar between reads with reference and alternate alleles. False positives are more likely to have alternate reads with significantly lower sequence qualities.

**Mapping quality bias:** uses a t-test to test if the average mapping quality is similar for reads with the reference and alternate allele. False positives are more likely to have alternate reads with significantly lower mapping qualities.

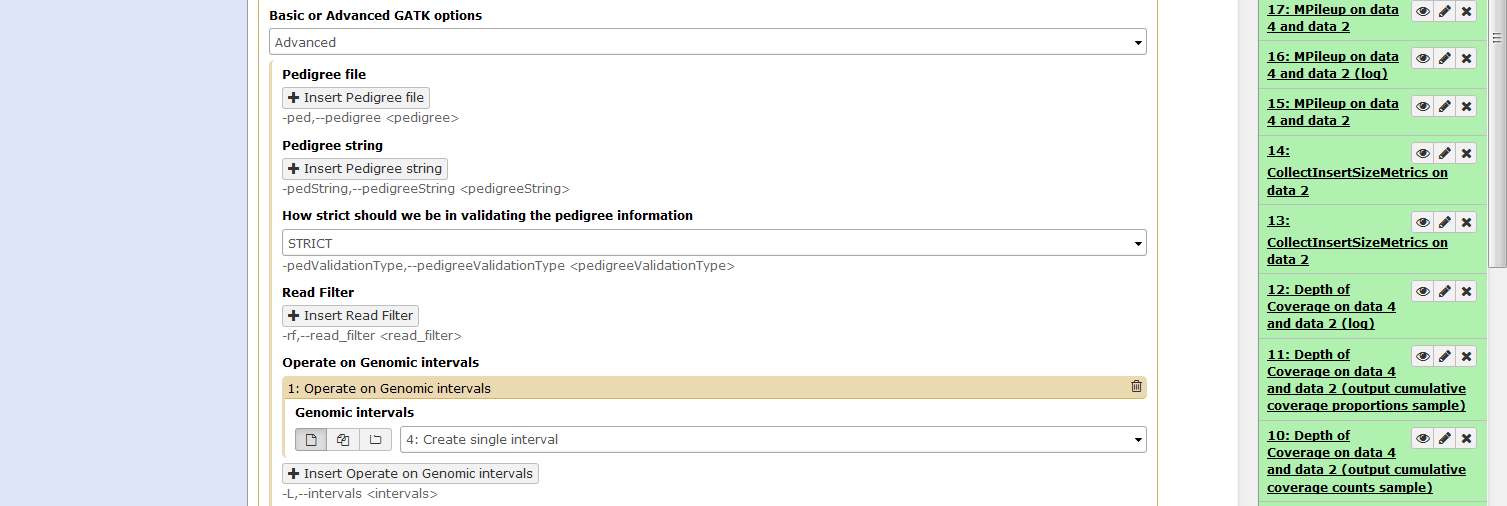
**Tail bias:** uses a t-test to test if alternate alleles are located evenly throughout the reads. False positives are more likely to have alternate alleles towards the end of reads where sequence quality diminishes.

**Call variants using GATK Unified Genotyper**

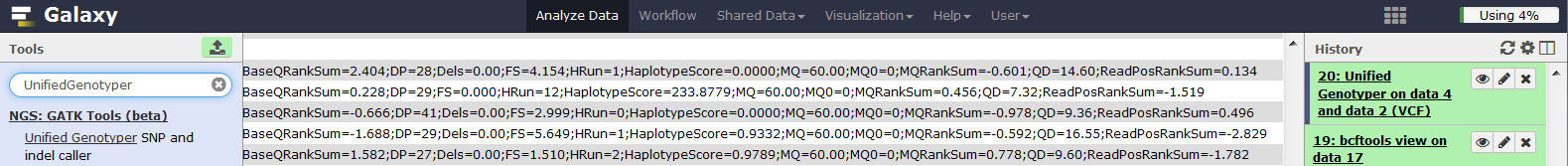
The GATK Unified Genotyper uses a general Bayesian framework to call variants and an error correction model based on expected characteristics of human variation to refine the variant calls (DiPristo et al 2011). More details on GATK are available here: <https://www.broadinstitute.org/gatk/>.

**1.** In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > Unified Genotyper

To reduce computing time, use advanced GATK options to restrict the analysis to the sequenced region by selecting operate on genomic intervals. In the advanced analysis options, lower the minimum base quality required to call to phred=10 to keep parity with the SAMtools analysis. Keep the other default settings and click execute.



**2.** View the VCF file created by GATK Unified Genotyper and read the comments in the header section to become familiar with the variables in the VCF file.



The info column contains several variables, described below, that can be used to flag or exclude low quality variants.

**BaseQRankSum** (equivalent of base quality bias): compares the base qualities between reads with the reference and alternate allele. Values are; close to zero if there is little difference; negative if alternate alleles have lower quality; positive if alternate allele have higher quality. Significant differences either way suggests that the sequencing process may have been biased or affected by an artefact.

**FS** (equivalent of Strand bias): Phred-scaled p-value using Fisher's exact test to detect strand bias.

**HRun:** Largest contiguous homopolymer run of variant allele in either direction.

**HaplotypeScore**: Consistency of the site with at most two segregating haplotypes.

**MQ:** Mapping quality.

**MQRankSum** (equivalent of mapping bias): Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities.

**QD:** Variant Confidence/Quality by Depth.

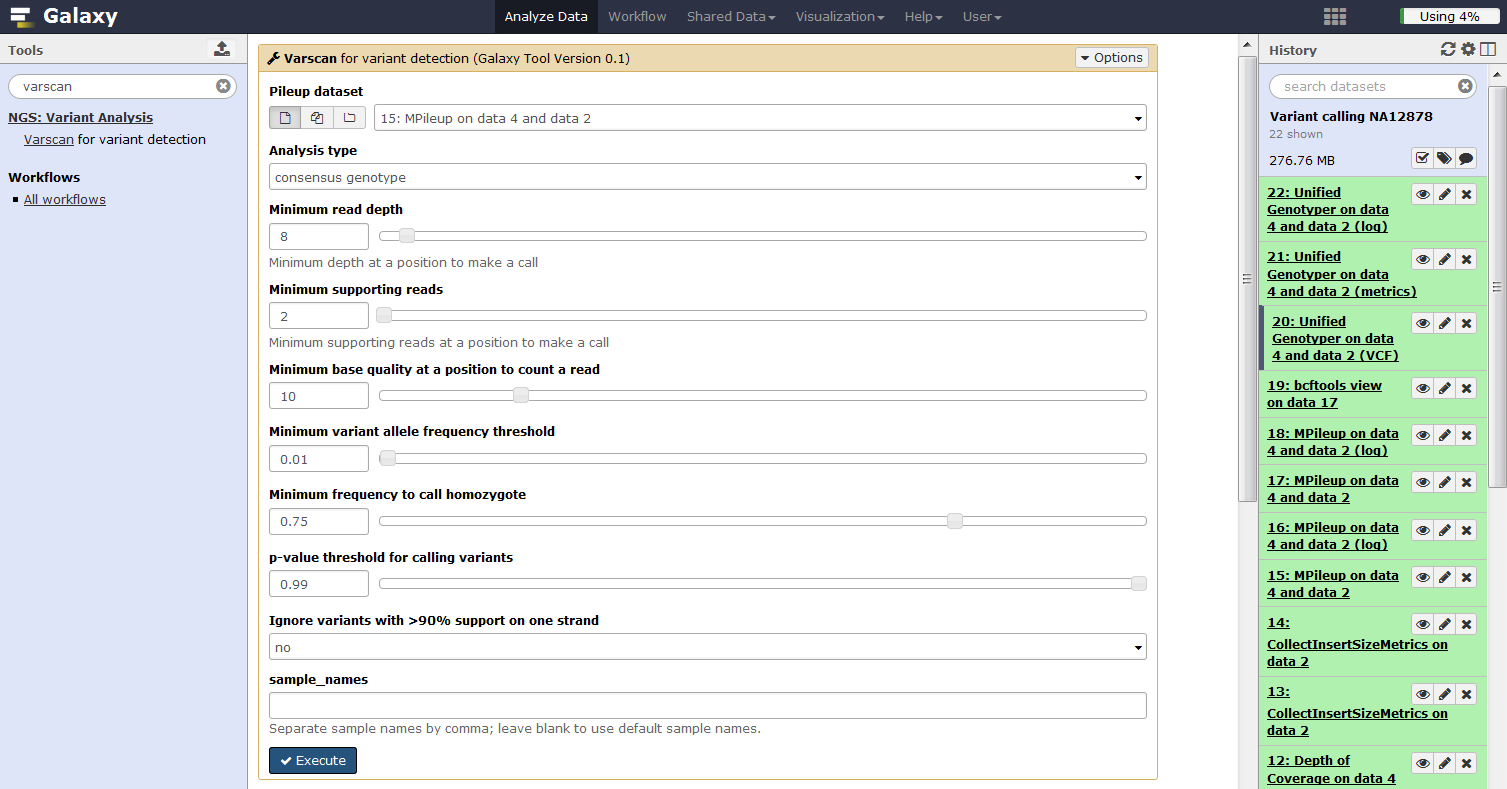
**ReadPosRankSum** (equivalent of tail bias): Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias

**Q5.** How many variants are called by GATK unified genotyper?

**Call variants using Varscan**

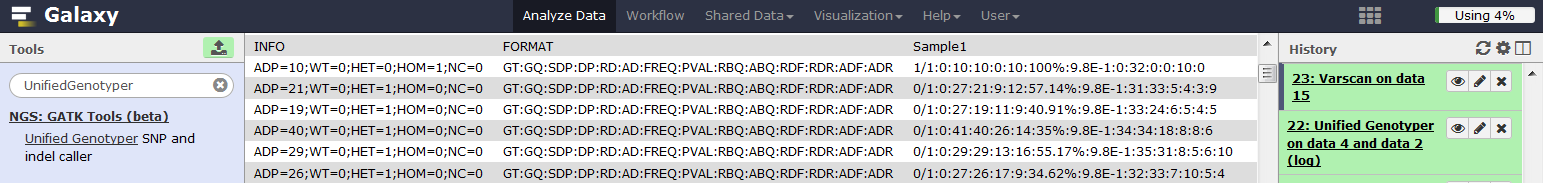
VarScan calls germline variants (SNPs and indels) using a heuristic method and a statistical test based on the number of aligned reads supporting each allele.

**1.** In **Tool Pane**: Go to **NGS: Variant Analysis** > Varscan



Reduce the ‘Minimum base quality at a position to count a read’ to 10 and click execute.

**2.** View the VCF file created by Varscan and familiarise yourself with the variables by reading the comments in the header section.



**ADP:** Average per-sample depth of bases with Phred score >= 10

**SDP:** Raw Read Depth as reported by SAMtools

**DP:** Quality Read Depth of bases with Phred score >= 10

**RD:** Depth of reference-supporting bases (reads1)

**AD:** Depth of variant-supporting bases (reads2)

**FREQ:** Variant allele frequency

**RBQ:** Average quality of reference-supporting bases (qual1)">

**ABQ:** Average quality of variant-supporting bases (qual2)">

**RDF:** Depth of reference-supporting bases on forward strand (reads1plus)

**RDR:** Depth of reference-supporting bases on reverse strand (reads1minus)

**ADF:** Depth of variant-supporting bases on forward strand (reads2plus)

**ADR:** Depth of variant-supporting bases on reverse strand (reads2minus)

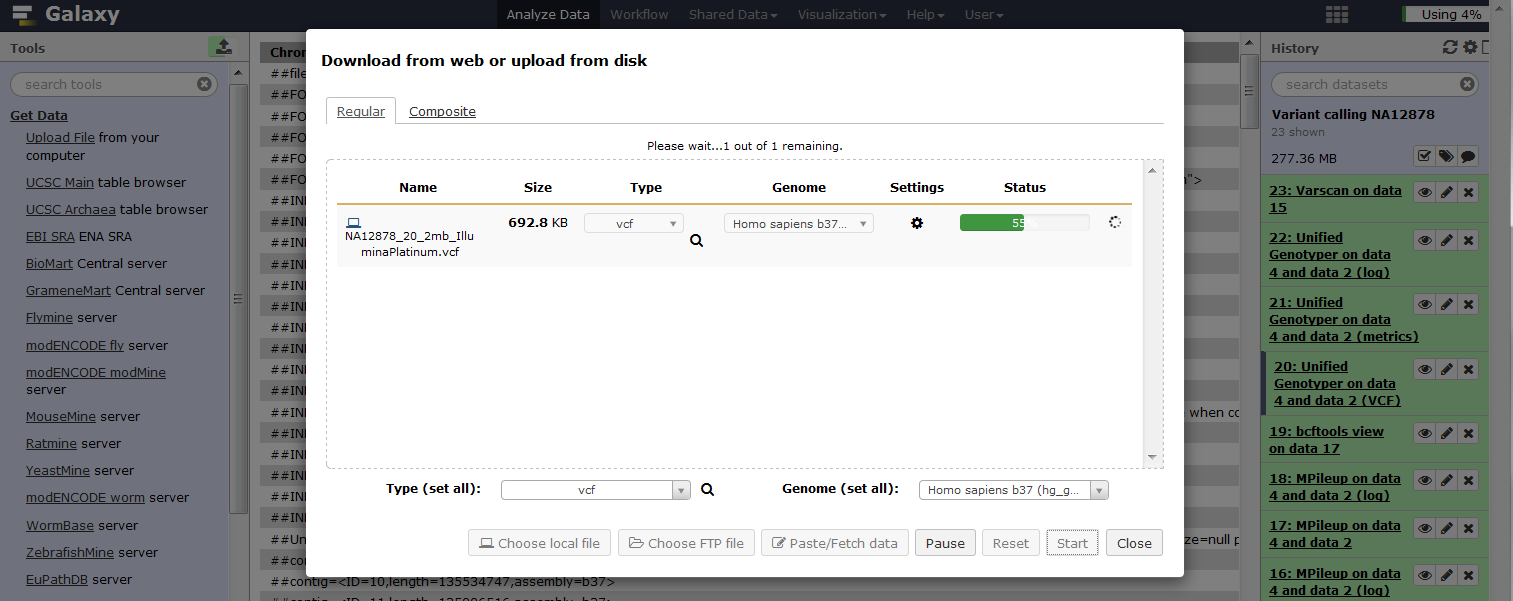
**Q6.** How many variants are called by Varscan?

**Evaluate variant callers by comparison with high confidence calls**

We now have three lists of variants generated by SAMtools/bcftools, GATK, and Varscan analysis of the same dataset. To evaluate the performance of these variant callers, we will use the GATK tool ‘Eval Variants’ to compare the VCF files with a set of high confidence variant calls.

**1.** In **Tool Pane**: Go to **Get Data** > Upload File

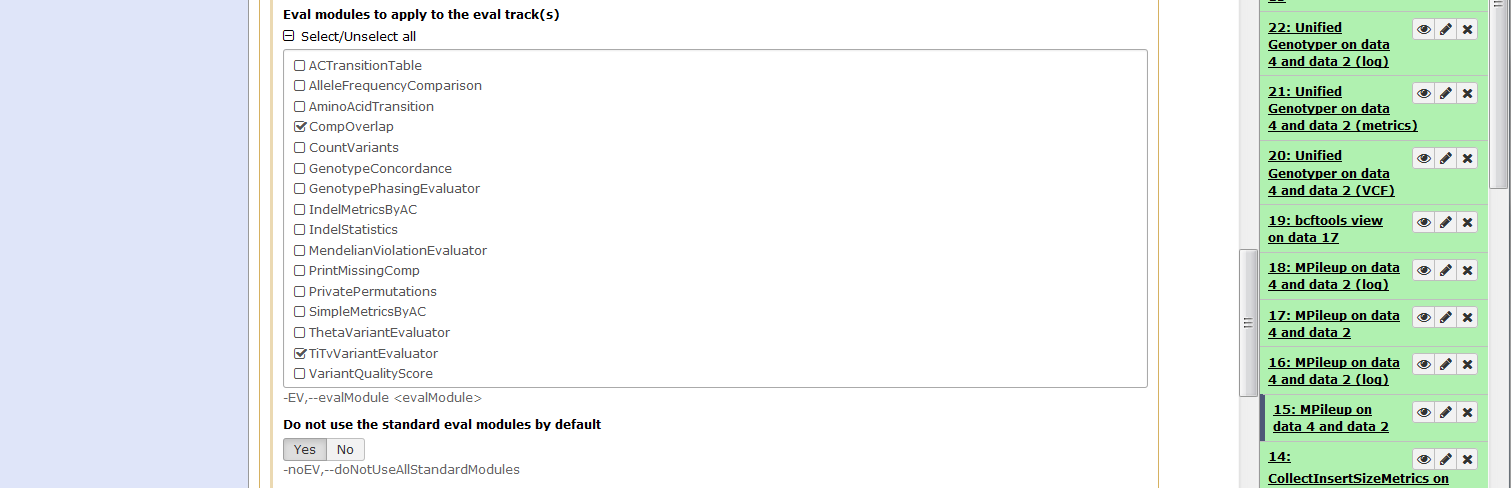
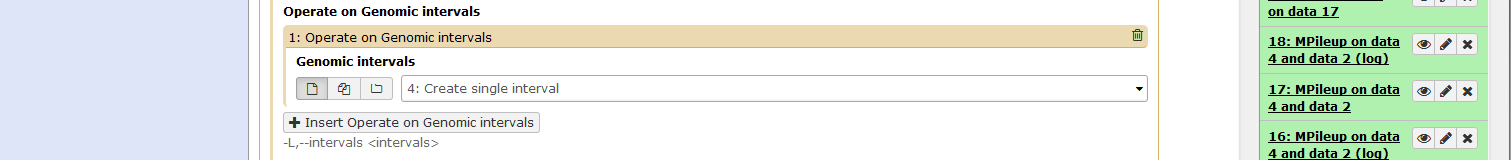
First, upload the high confidence variant calls ‘NA12878\_20\_2mb\_IlluminaPlatinum.vcf, select type ‘vcf’ and genome ‘b37’, click start.

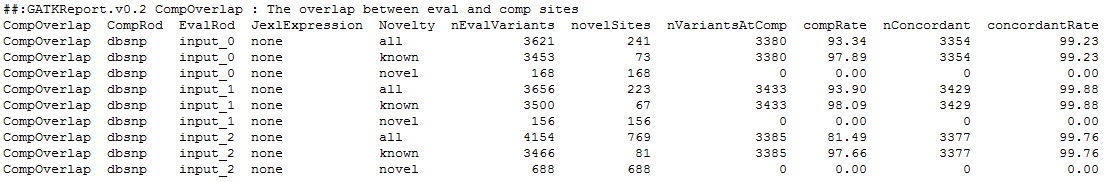


Use the GATK Eval Variants tool to compare the variant callers with the set of high confidence and to calculate the rate of transition to transversions.

**2.** In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > Eval Variants

Enter the VCF files in the order made (1. Bcftools, 2. Unified Genotyper, 3. Varscan), select the high confidence variant calls as a dbSNP ROD file, use the advanced GATK options to restrict the analysis to the genomic interval sequenced and use the advanced analysis options to select CompOverlap, CountVariants and TiTvVariantEvaluator as the evaluation modules then click execute.



**3.** Look at the CompOverlap table in the Eval Variants report.

CompRod: file used for comparison, here dbSNP refers to the Platinum high confidence calls

EvalRod: file being evaluated, input\_0 = MPileup/bcftools, input\_1 = GATK Unified Genotyper, input\_2 = Varscan

Novelty: is the variant in CompRod (Platinum), known = yes, Novel = no

nEvalVariants: number of variants in EvalRod which meet evaluation criteria

novelSites: number of variants in EvalRod and not in CompRod (Platinum)

nVariantsAtComp: number of variants in both EvalRod and CompRod (Platinum)

compRate: % of EvalRod variants in CompRod (nVariantsAtComp/nEvalVariants)

nConcordant: number of EvalRod variants with the same alleles as CompRod

concordantRate: % of variants in both EvalRod and CompRod with the same alleles (nConcordant/nVariantsAtComp)

The high confidence Platinum calls for the 2Mb region consist of 3,810 variants but only 3,629 which meet the EvalVariants criteria were considered. Use the CompOverlap table and number of high confidence variants considered (n=3,629) to fill in the table and answer the questions below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Caller** | **No. variants (all)** | **True positive (nConcordant)** | **False negative**  **(3629-nConcordant)** | **Sensitivity** | **False positives**  **(No. variants -nConcordant)** | **False**  **positive %** |
| MPileup/bcftools | 3621 | 3354 | 275 | 92.4 | 267 | 7.4 |
| GATK | 3656 | 3429 | 200 | 94.5 | 227 | 6.2 |
| Varscan | 4154 | 3377 | 252 | 93.1 | 777 | 18.7 |

**Q7:** Which variant caller (MPileup/bcftools, GATK and Varscan) has the highest true positive rate/sensitivity? Sensitivity = (true positive/[true positive + false negative])

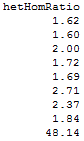
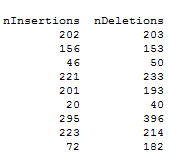
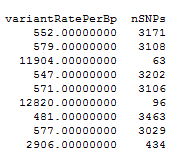
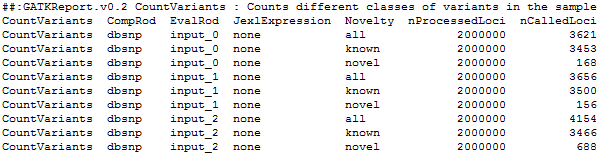
**Q8:** Which variant caller (MPileup/bcftools, GATK and Varscan) has the lowest percentage of false positives? False positive % = 100\*(false positive/[false positive + true positive])

Studies such as the 1000 Genomes project and Platinum Genomes have provided a lot of information about human variation that enable predictions to be made about the variation we expect to see in a new sample:

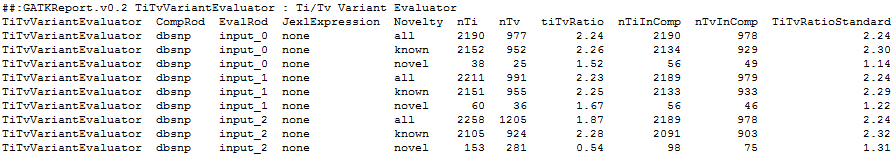
* For whole genome sequencing, true variation occurs at a rate of about 1 variant per 650bp.
* The exome is roughly 2 times more conserved than non-coding regions, which corresponds to a lower rate of approximately 1 variant per 1250bp.
* Approximately 83% of variation will be present in dbSNP version 129, which is the last 'clean' version that does not include variation, some of which is causal, from the 1000 Genomes Project and other large-scale next-generation sequencing projects. The number of variants has increased from 13.6 million in dbSNP129 to 63.3 million in dbSNP138!
* The ratio of transition (A<>G or C<>T) to transversions (A<>C, G<>T, A<>T, C<>G) in the genome is expected to be greater than 2 and close to 3 in the exome. Transitions are more common due to the molecular process involved, the bases having similar shape and the changes being less deleterious as they are less likely to result in an amino acid substitution.
* The ratio of heterozygous to homozygous variants should be around 1.6. Excess levels of heterozygosity could relate to sample contamination or recent admixture while deficiencies could occur due to inbreeding, large deletions, loss of a whole chromosome or acquired uniparental disomy (both copies of a chromosome are from one parent due to loss of either the paternal or maternal copy).
* Average percentage of heterozygous variants on chromosome X is 20% for males and 65% for females

**4.** Look at the CountVariants and Transition (Ti) / Transversion (Tv) Variant Evaluator tables in the Eval Variants report.

Selected columns from CountVariants



Slected columns from Ti/Tv Variant Evaluator



The format of CountVariants and Ti/Tv Variant Evaluator are similar to the CompOverlap report. For Ti/Tv, the most important columns are ‘tiTvRatio’ and ‘TiTvRatioStandard’ which should be similar to each other. More details on the EvalVariant output is available here:

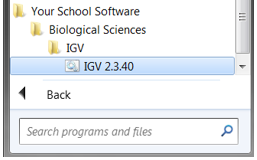
<https://www.broadinstitute.org/gatk/guide/article?id=6309>

**Q9.** How does the rate of variation per bp, het:hom ratio and tiTvRatio compare with the expected genome wide values from Platinum Genomes (1 variant per 650bp, 1.6, and 2 respectively)?

**Visualise a suspected false positive variant in IGV**

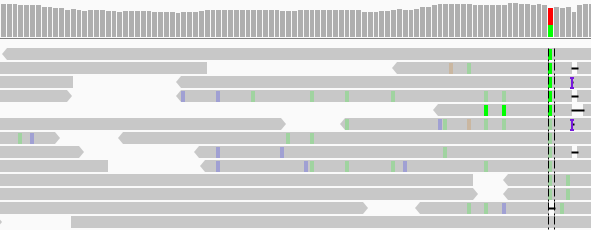
Each of the variant calling tools identifies variants that others do not, and the accuracy of these discordant variants is expected to be low. To help spot false positive variants, we will now use IGV to look at a unique-to-MPileup/bcftools variant with significant strand bias and base quality bias, which is probably an error.

**1.** Launch IGV from Start menu > All Programs > Your School Software > Biological Sciences > IGV > IGV 2.3.40 (This will take some time <5mins as IGV has to load the whole genome, a black window will appear with messages, check this and be patient).



**2.** When IGV opens, make sure the reference genome is set to hg19 (Figure 6). From the file tab select ‘load from file’, navigate to the folder with your data, select your bam file and select open.

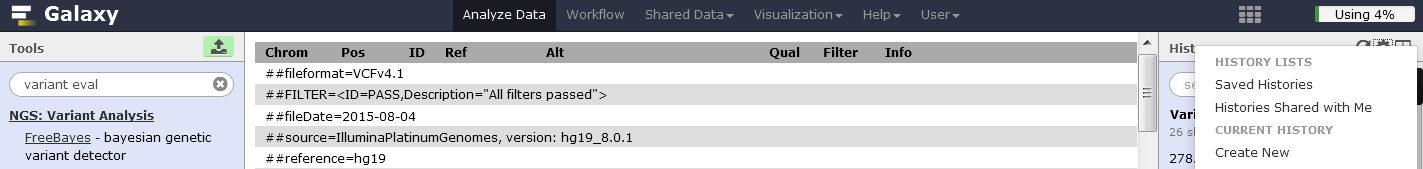
**3.** Navigate to ‘chr20:1,707,746’, which marks the location of a unique-to-MPileup/bcftools variant with significant bias in strand (p=0.0014, all reads with the alternate allele are on the negative strand) and base quality (p=0.0000054, alternate alleles have lower average sequence quality than reference alleles). Right click and select sort alignments by base and these biases are clearly visible in IGV as reads with the alternate allele pointing in the same direction and alternate alleles with lighter shading. In addition, IGV shows that reads with the alternate allele contain many other variants. These features strongly suggest that the variant is an artefact and could be excluded from a tiered analysis. However, it is important to bear in mind that many unique-to-caller variants have been validated.



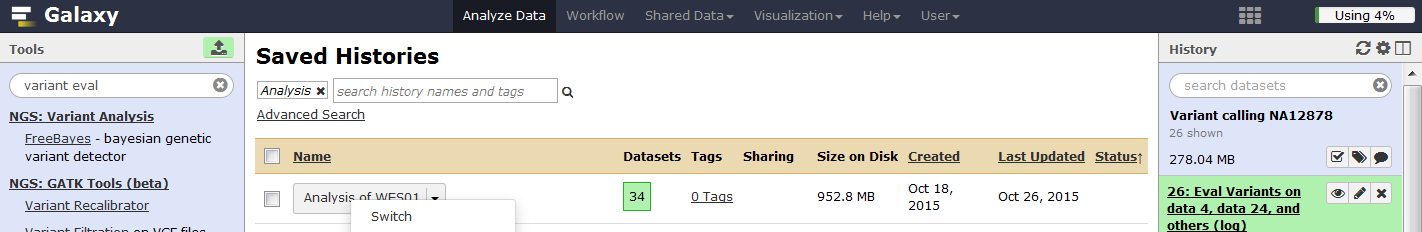
**Call variants in WES01**

Having established that GATK Unified Genotyper has the highest sensitivity and lowest false positive rate, we will now use this program to call variants in the trial exome data for patient WES01.

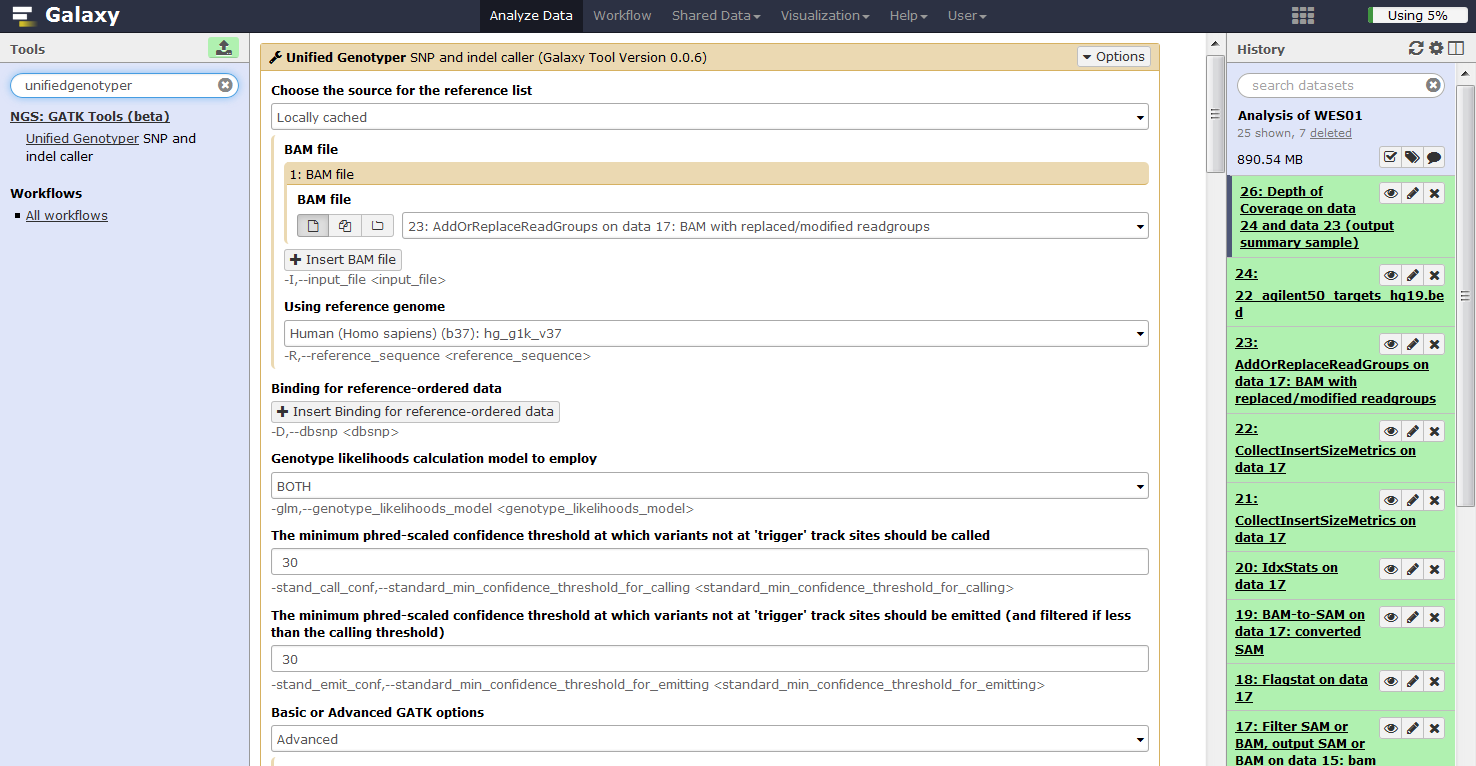
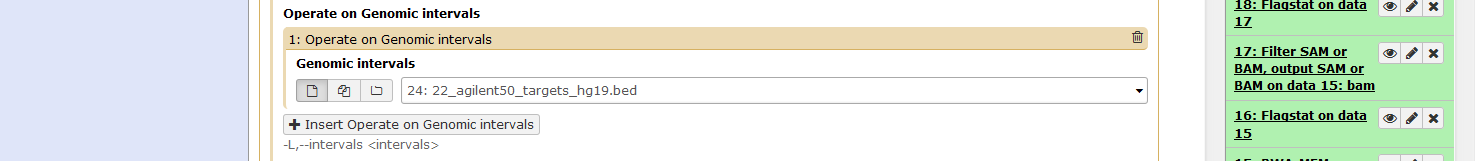
**1.** Click the cog icon in the history pane and select saved histories.

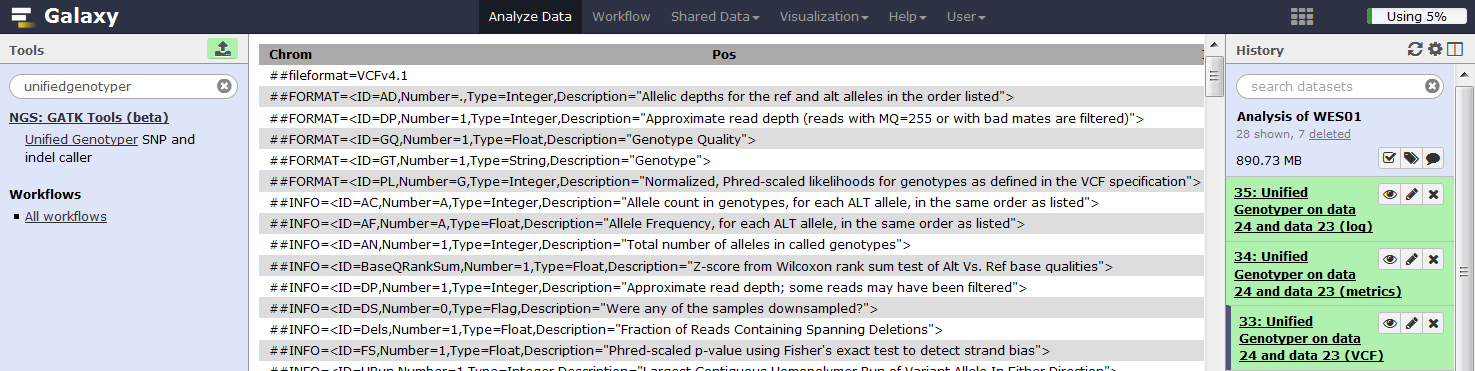


**2.** Either click on the history or select ‘Switch’ from the dropdown menu to change histories to ‘Analysis of WES01’.



**3.** In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > Unified Genotyper



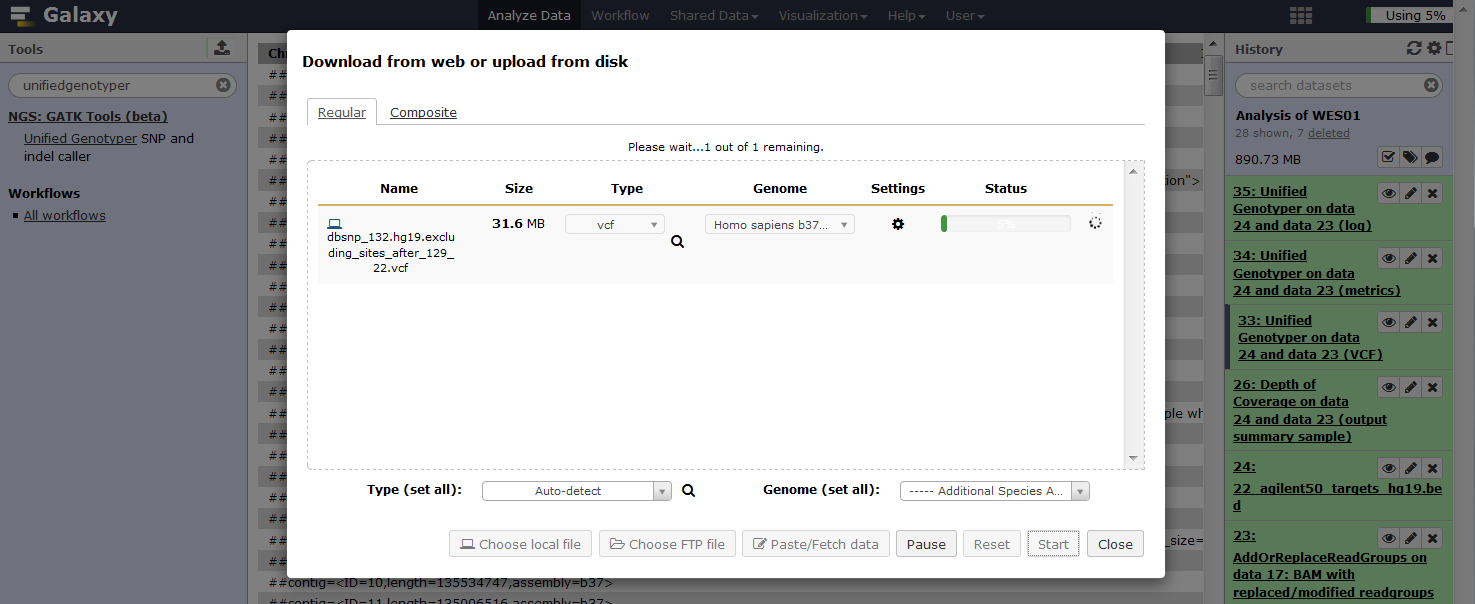
**4.** View the VCF file.

**Q10.** Considering that the targeted region on chromosome 22 spans 1,183,396 bp what is the rate of variation and does it come close to the prediction from Platinum Genomes for coding regions (1 variant per 1400bp)?

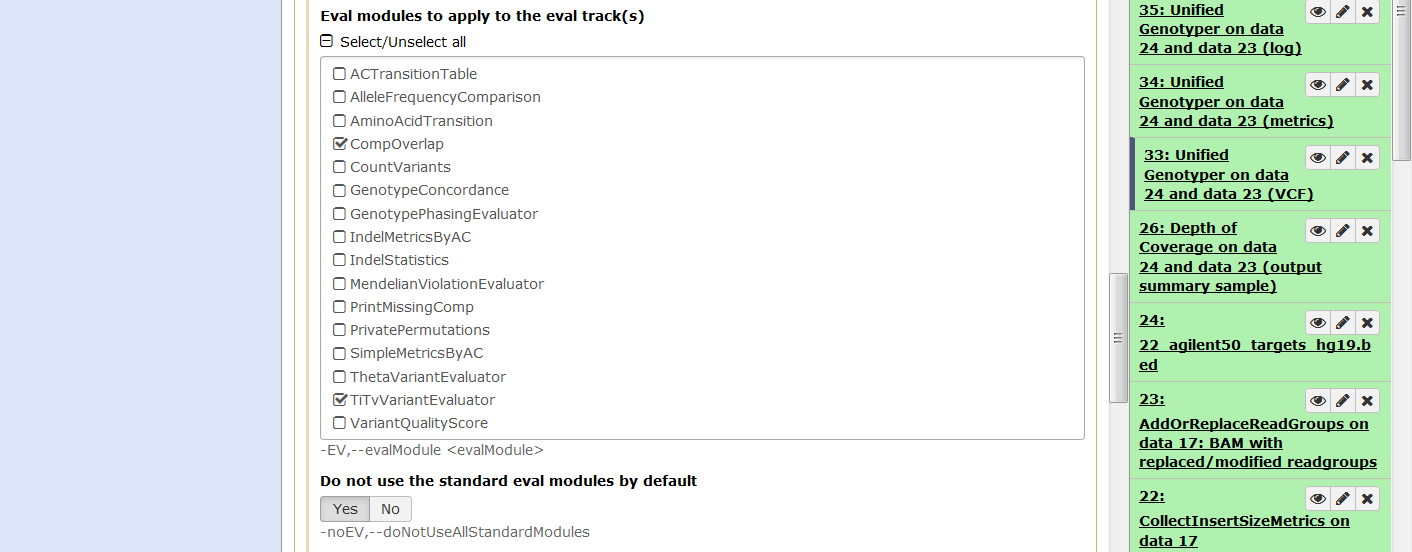
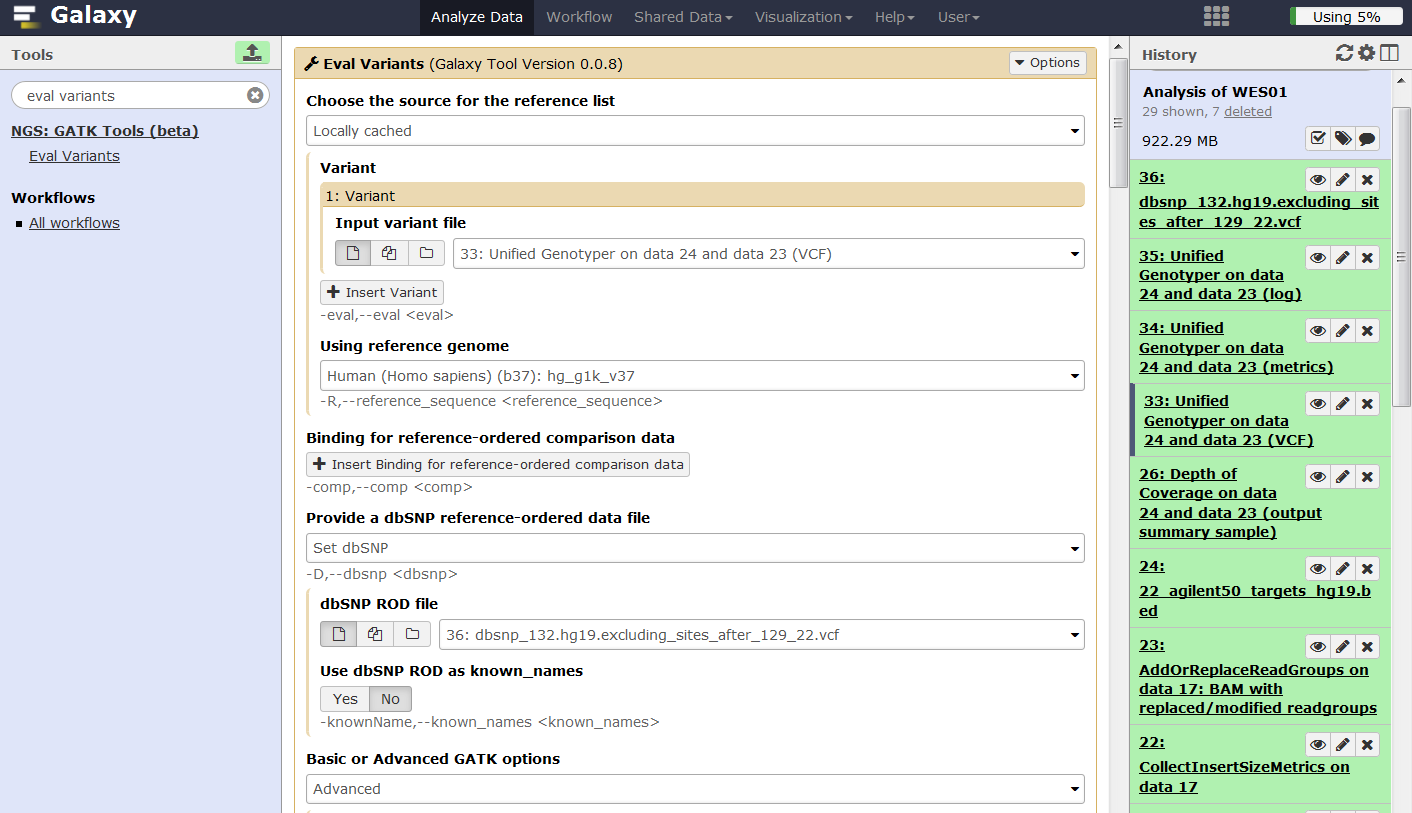
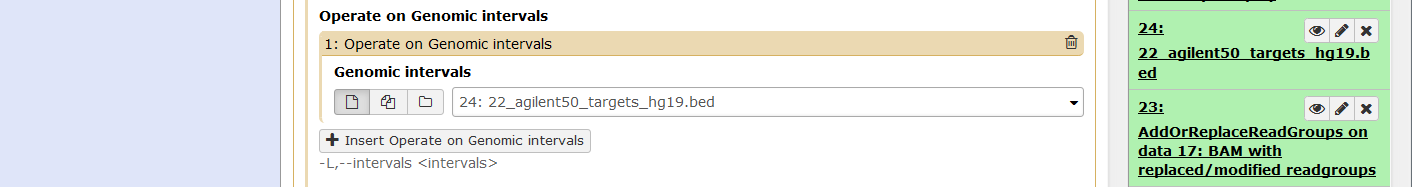
**Evaluate the variant calls by comparison with dbSNP**

**1.** In **Tool Pane**: Click the upload icon

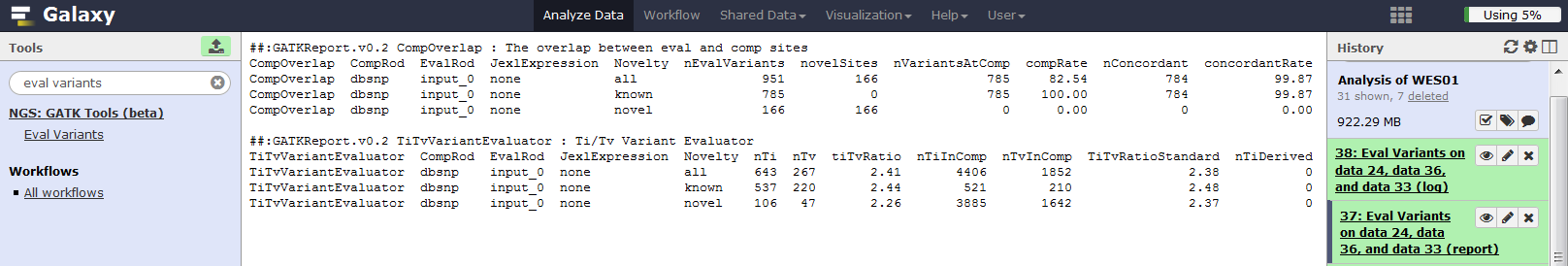
Choose local file, select the ‘dbsnp\_132.hg19.excluding\_sites\_After\_129\_22.vcf’ file, set type to ‘vcf’ and genome to ‘b37’ then click start.



2. In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > Eval Variants



**3.** View the Eval Variants report



**Q11.** Comment on the amount of variation that is present in dbSNP129 and how well it agree with expectation (~83% of variation is usually present in dbSNP version 129)?

**Q12.** What is the ratio of transitions to transversions and is it in line with the predicted value of 3?

**Congratulations you finished the exercise!**In the next practical we will annotate the variants with respect genes, major databases of normal variation, and predictors of pathogenicity and use filtering strategies to search for potentially causal variants.