

## 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:

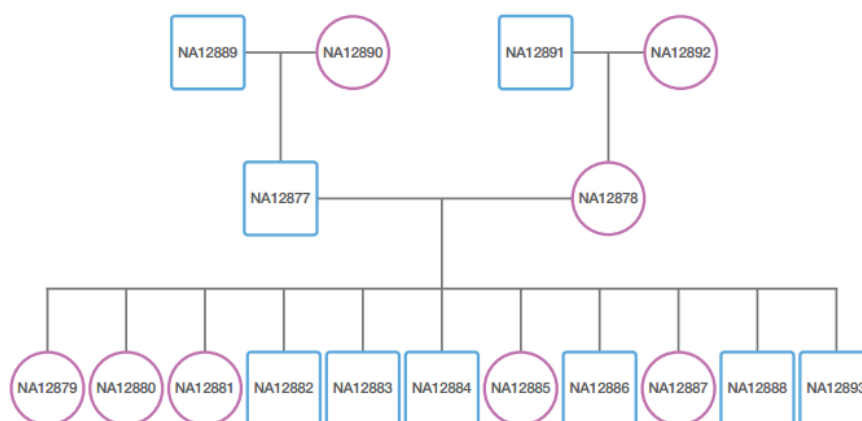
1. Use a range of software (GATK, Varscan, SAMtools, and BCFtools) to call small variants (SNVs and indels)
2. Describe the contents of pileup and variant call files
3. Generate and interpret variant quality control parameters (quality score, genotype quality, sequence context, strand bias, base quality 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](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

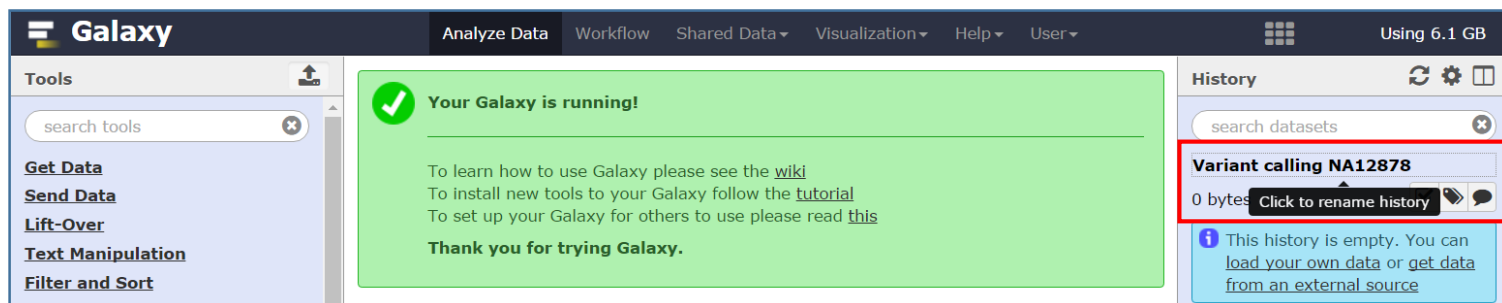
1. Login to the University of Southampton Galaxy server <http://galaxyngs-hpc.soton.ac.uk> using your iSolutions username and password.



The screenshot shows the University of Southampton Logon page. It features a dark blue header with the university's name and logo. Below the header, there is a white box containing a photograph of students in a classroom and a login form. The form includes fields for 'Username' and 'Password', and a 'Logon' button. Text instructions prompt the user to enter their University of Southampton user name and password.

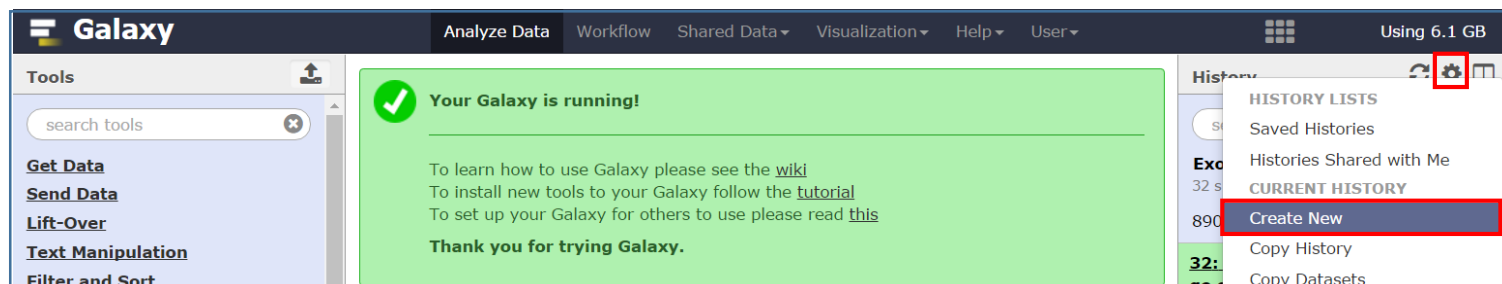
Depending on the computers cache, login will take you to an empty Unnamed history OR the last active history (Exome analysis of WES01). Follow step 2 if login arrives at Unnamed history or step 3 if login reveals the last active history.

2. If the history is called 'Unnamed history' rename it for the variant calling project (Click on name, type new name, press return). If the history is called 'Exome analysis of WES01' jump to step 3.



The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The left sidebar contains a 'Tools' section with a search bar and links for 'Get Data', 'Send Data', 'Lift-Over', 'Text Manipulation', and 'Filter and Sort'. The main content area displays a green message: 'Your Galaxy is running!' with links to the 'wiki', 'tutorial', and 'this' page. The right sidebar shows the 'History' section with a search bar and a list of datasets. A red box highlights the 'Variant calling NA12878' history entry, which is currently empty (0 bytes). Below the entry, there is a link to 'Click to rename history'.

3. If the history is called 'Exome analysis of WES01' create a new history by clicking on the cog icon in the history pane and selecting Create New. Rename the history 'Variant calling NA12878' (click on name, type new name, press return).



The screenshot shows the Galaxy web interface with the 'History' pane open. A red box highlights the 'Create New' option in the 'CURRENT HISTORY' section. The 'History' pane also shows a list of 'HISTORY LISTS' including 'Saved Histories' and 'Histories Shared with Me'. The 'Create New' option is highlighted with a red box, indicating the next step in the process.

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

- In **Tool Pane**: Go to **Get Data** > **Upload File** from your computer  
Choose local file, select type 'bam', select genome 'b37', click start

**Download data directly from web or upload files from your disk**

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

Name	Size	Type	Genome	Settings	Status
NA12878_chr20_2mb_filtered.bam	39.4 MB	bam	unspecified (?)		0%

Type (set all): bam Genome (set all): b37

Buttons: Choose local file, Paste/Fetch, Start, Pause, Reset, Close

## 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 it 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.

- In **Tool Pane**: Go to **NGS: SAMtools** > **IdxStats**

**IdxStats** tabulate mapping statistics for BAM dataset (Galaxy Tool Version 2.0)

BAM file: 1: NA12878\_chr20\_2mb\_filtered.bam

Execute

History: Variant calling NA12878, 1 shown, 37.6 MB

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

	1	2	3	4
MT		16569	1	0
1		249250621	235	8
2		243199373	215	7
3		198022430	152	6
4		191154276	154	5
5		180915260	114	3
6		171115067	151	5
7		159138663	90	2
8		146364022	98	4
9		141213431	76	6

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

**Create single interval** as a new dataset (Galaxy Tool Version 1.0.0)

Chromosome:

Start position:

End position:

Name:

Strand:

☒ Execute

Enter 20 for chromosome because it matches the contig format of the reference genome (B37)

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

**Depth of Coverage** on BAM files (Galaxy Tool Version 0.0.2)

Choose the source for the reference list:

BAM file:

Using reference genome:

☒ Execute

**Tools**

depthof

**NGS: GATK Tools**

Depth of Coverage on BAM files

**Workflows**

All workflows

**Output format**

table

--outputFormat <outputFormat>

**Basic or Advanced GATK options**

Advanced

**Pedigree file**

+ Insert Pedigree file

-ped,--pedigree <pedigree>

**Pedigree string**

+ Insert Pedigree string

-pedString,--pedigreeString <pedigreeString>

**How strict should we be in validating the pedigree information**

STRICT

-pedValidationType,--pedigreeValidationType <pedigreeValidationType>

**Read Filter**

+ Insert Read Filter

-rf,--read\_filter <read\_filter>

**Operate on Genomic intervals**

1: Operate on Genomic intervals

**Genomic intervals**

3: Create single interval

+ Insert Operate on Genomic intervals

-L,--intervals <intervals>

**Basic or Advanced Analysis options**

Basic

Execute

**History**

search datasets

**Variant calling NA12878**

3 shown

37.6 MB

3: Create single interval

2: IdxStats on data 1

1: NA12878\_chr20\_2mb\_filtered.bam

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

**Galaxy**

Analyze Data Workflow Shared Data Visualization Help User

Using 7.8 GB

**Tools**

depthof

**NGS: GATK Tools**

Depth of Coverage on BAM files

1	2	3	4	5	6
sample_id	total	mean	granular_third_quartile	granular_median	granular_first_quartile
sample-a	48268640	24.13	31	26	21
Total	48268640	24.13	N/A	N/A	N/A

**History**

5: Depth of Coverage on data 3 and data 1 (output summary sample)

4: Depth of Coverage on data 3 and data 1

Q1. What is the mean coverage?

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

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

**Galaxy**

Analyze Data Workflow Shared Data Visualization Help User

Using 7.8 GB

**Tools**

collectinsert

**NGS: Picard**

CollectInsertSizeMetrics plots distribution of insert sizes

**Workflows**

All workflows

**CollectInsertSizeMetrics** plots distribution of insert sizes (Galaxy Tool)

Version 1.126.0

**Select SAM/BAM dataset or dataset collection**

1: NA12878\_chr20\_2mb\_filtered.bam

If empty, upload or import a SAM/BAM dataset.

**Load reference genome from**

Local cache

**Using reference genome**

Human (Homo sapiens) (b37)

REFERENCE\_SEQUENCE

**History**

search datasets

**Variant calling NA12878**

11 shown

80.1 MB

11: Depth of Coverage on data 3 and data 1 (log)

10: Depth of Coverage on data 3 and data 1 (output cumulative coverage proportions sample)

**Select validation stringency**

Lenient

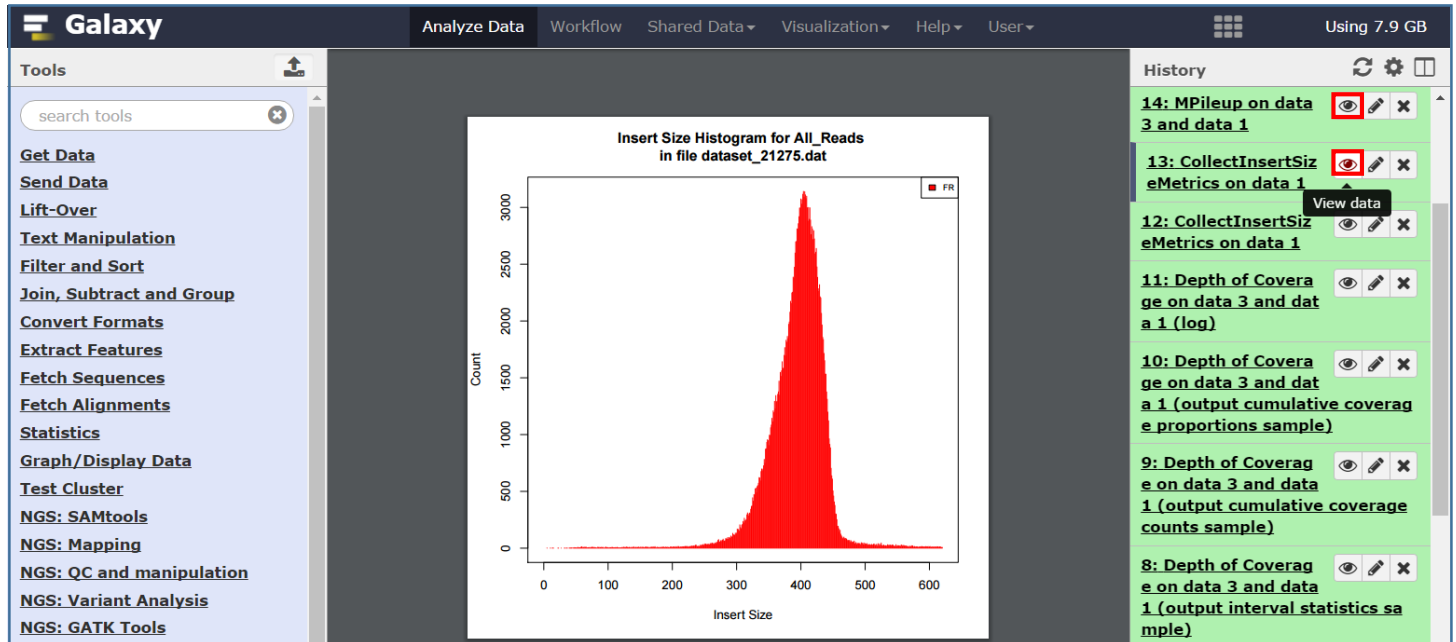
Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.

✓ Execute

**8: Depth of Coverage on data 3 and data 1 (output interval statistics sample)**

**7: Depth of Coverage on data 3 and data 1 (output interval summary sa**

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.

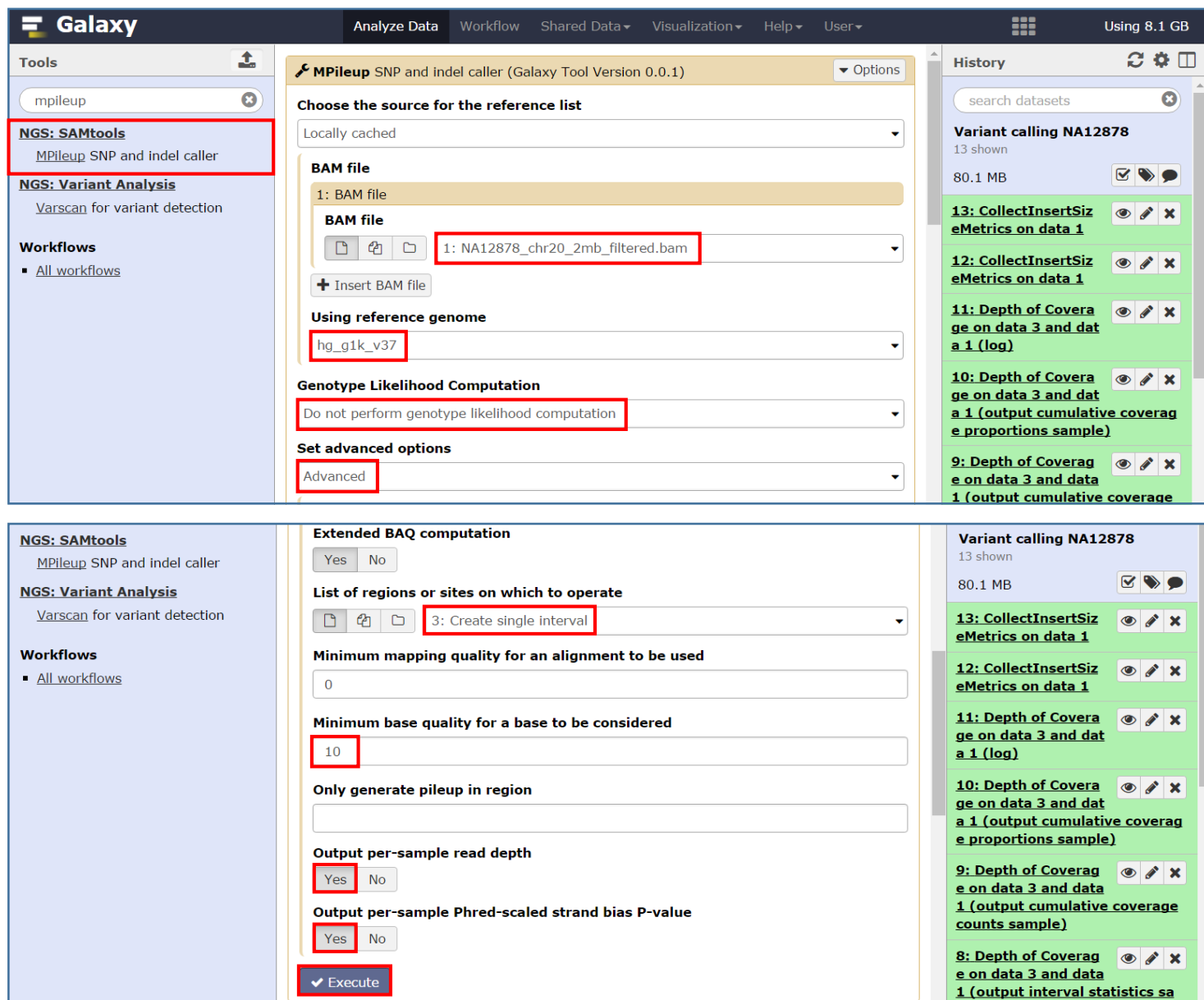
## Call variants using SAMtools MPileup and bcftools

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**



The screenshot shows the Galaxy web interface with the MPileup tool configuration. The left sidebar shows the 'Tools' panel with 'NGS: SAMtools' selected. The main panel shows the 'MPileup SNP and indel caller' tool configuration. The right sidebar shows the 'History' panel with 'Variant calling NA12878' selected.

**MPileup SNP and indel caller (Galaxy Tool Version 0.0.1)**

Choose the source for the reference list: Locally cached

**BAM file**

1: BAM file

BAM file: 1: NA12878\_chr20\_2mb\_filtered.bam

Using reference genome: hg\_g1k\_v37

Genotype Likelihood Computation: Do not perform genotype likelihood computation

Set advanced options: Advanced

**Extended BAQ computation**

Yes No

List of regions or sites on which to operate: 3: Create single interval

Minimum mapping quality for an alignment to be used: 0

Minimum base quality for a base to be considered: 10

Only generate pileup in region:

Output per-sample read depth: Yes No

Output per-sample Phred-scaled strand bias P-value: Yes No

Execute

**Variant calling NA12878**

13 shown

80.1 MB

13: CollectInsertSizeMetrics on data 1

12: CollectInsertSizeMetrics on data 1

11: Depth of Coverage on data 3 and data 1 (log)

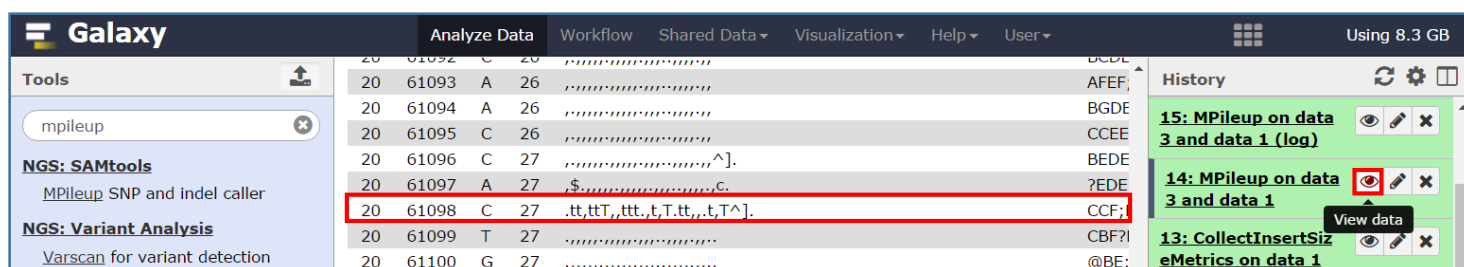
10: Depth of Coverage on data 3 and data 1 (output cumulative coverage proportions sample)

9: Depth of Coverage on data 3 and data 1 (output cumulative coverage counts sample)

8: Depth of Coverage on data 3 and data 1 (output interval statistics sample)

## 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.



The screenshot shows the Galaxy web interface with the MPileup tool output. The left sidebar shows the 'Tools' panel with 'NGS: SAMtools' selected. The main panel shows the MPileup tool output. The right sidebar shows the 'History' panel with 'Variant calling NA12878' selected.

**MPileup**

Chromosome	Position	Reference Base	Number of Reads	Read Bases	Base Qualities
20	61092	C	26	.....	.....
20	61093	A	26	.....	.....
20	61094	A	26	.....	.....
20	61095	C	26	.....	.....
20	61096	C	27	.....^.	.....
20	61097	A	27	.,\$.....C.	.....
20	61098	C	27	.tt,tT,,tt,,t,T.tt,,t,T^.	.....
20	61099	T	27	.....	.....
20	61100	G	27	.....	.....

**Variant calling NA12878**

13 shown

80.1 MB

15: MPileup on data 3 and data 1 (log)

14: MPileup on data 3 and data 1

13: CollectInsertSizeMetrics on data 1

**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.** Click the 'Run this job again' icon in the history pane, which will keep the previous settings/options, and select 'Perform genotype likelihood computation'.

The screenshot shows the Galaxy web interface. The main panel displays the 'MPileup SNP and indel caller (Galaxy Tool Version 0.0.1)' tool configuration. The 'Choose the source for the reference list' dropdown is set to 'Locally cached'. The 'BAM file' section shows a file named '1: NA12878\_chr20\_2mb\_filtered.bam'. The 'Using reference genome' dropdown is set to 'hg\_g1k\_v37'. The 'Genotype Likelihood Computation' dropdown is set to 'Perform genotype likelihood computation', which is highlighted with a red box. The 'Phred-scaled gap extension sequencing error probability' is set to '20'. At the bottom, there is a red 'Execute' button. The right-hand 'History' pane shows a list of jobs, with the most recent job '14: MPileup on data 3 and data 1' highlighted. Below the job list, there is a 'Run this job again' button, also highlighted with a red box. The bottom of the interface shows a 'Workflows' section with a 'VCF Manipulation' workflow.

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



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

**Galaxy** Analyze Data Workflow Shared Data Visualization Help User Using 7.9 GB

**Tools**

bcftools

**NGS: SAMtools**

bcftools view Convert, filter, subset VCF/BCF files

**Workflows**

All workflows

**bcftools view** Convert, filter, subset VCF/BCF files (Galaxy Tool Version 0.1.19.0) Options

Choose a bcf file to view

16: MPileup on data 3 and data 1

Choose the output format

VCF

-b

List of chromosome names for conversion

No selection

(-D)

Use alternate INDEL-to-SNP mutation rate

-1

defaults to 0.15 (-i)

Mutation rate for variant calling

0.001

default to 0.001 (-t)

Retain all possible alternate alleles at variant sites

Yes No

-A

Output Potential Variant Sites Only

Yes No

**SNP calling**

Yes No

Forces -e the max-likelihood inference parameter. (-c)

Suppress all individual genotype information

Yes No

-G

Skip sites where the REF field is not A/C/G/T

Yes No

-N

Perform max-likelihood inference only

Yes No

Including estimating the site allele frequency, testing Hardy-Weinberg equilibrium and testing associations with LRT. (-e)

Call per-sample genotypes at variant sites

Yes No

-g

Execute

**History**

search datasets

**Variant calling NA12878**

17 shown

240.6 MB

17: MPileup on data 3 and data 1 (log)

16: MPileup on data 3 and data 1

15: MPileup on data 3 and data 1 (log)

14: MPileup on data 3 and data 1

13: CollectInsertSizeMetrics on data 1

12: CollectInsertSizeMetrics on data 1

11: Depth of Coverage on data 3 and data 1 (log)

10: Depth of Coverage on data 3 and data 1 (output cumulative coverage)

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>.

## 5. View the VCF file created by bcftools.

The screenshot shows the Galaxy web interface. On the left, the 'Tools' panel lists 'bcftools' under 'NGS: SAMtools'. The main panel displays the output of the 'bcftools view' tool, showing a table of variant data. The table has columns: QUAL, FILTER, and INFO. The first row is highlighted with a red box. The 'History' panel on the right shows a search for 'Variant calling NA12878' and lists several datasets. The dataset '18: bcftools view o n data 16' is highlighted with a red box, showing '3,621 lines, 34 comments' and 'format: vcf, database: hg\_g1k\_v37'.

QUAL	FILTER	INFO
173	.	DP=27;VDB=0.0365;AF1=0.5;AC1=1;DP4=5,4,3,9;MQ=60;FQ=135;PV4=0.2,
183	.	DP=41;VDB=0.0400;AF1=0.5;AC1=1;DP4=18,8,8,6;MQ=60;FQ=186;PV4=0.5
210	.	DP=29;VDB=0.0305;AF1=0.5;AC1=1;DP4=8,5,6,10;MQ=60;FQ=205;PV4=0.2
135	.	DP=27;VDB=0.0398;AF1=0.5;AC1=1;DP4=7,10,5,4;MQ=60;FQ=138;PV4=0.6
209	.	DP=34;VDB=0.0099;AF1=0.5;AC1=1;DP4=12,7,6,8;MQ=60;FQ=200;PV4=0.3
225	.	DP=43;VDB=0.0373;AF1=0.5;AC1=1;DP4=9,11,12,11;MQ=60;FQ=225;PV4=C
225	.	DP=30;VDB=0.0345;AF1=0.5;AC1=1;DP4=5,7,9,8;MQ=60;FQ=192;PV4=0.71
217	.	INDEL;DP=33;VDB=0.0365;AF1=0.5;AC1=1;DP4=10,7,6,7;MQ=60;FQ=217;P
222	.	DP=30;VDB=0.0359;AF1=1;AC1=2;DP4=0,0,18,9;MQ=60;FQ=-108
93.5	.	INDEL;DP=26;VDB=0.0392;AF1=0.5;AC1=1;DP4=0,2,3,7;MQ=60;FQ=17.6;PV

The history panel shows the VCF file contains 3,621 variants and 34 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.

**Galaxy** Analyze Data Workflow Shared Data Visualization Help User Using 8.4 GB

**Tools** unifiedgeno

**NGS: GATK Tools**  
Unified Genotyper SNP and indel caller

**Workflows**  
All workflows

**Unified Genotyper** SNP and indel caller (Galaxy Tool Version 0.0.6) Options

**Choose the source for the reference list**  
Locally cached

**BAM file**  
1: BAM file  
BAM file  
1: NA12878\_chr20\_2mb\_filtered.bam

**Using reference genome**  
Human (Homo sapiens) (b37)

**Basic or Advanced GATK options**  
Advanced

**Pedigree file**  
+ Insert Pedigree file  
-ped,--pedigree <pedigree>

**Pedigree string**  
+ Insert Pedigree string  
-pedString,--pedigreeString <pedigreeString>

**How strict should we be in validating the pedigree information**  
STRICT  
-pedValidationType,--pedigreeValidationType <pedigreeValidationType>

**Read Filter**  
+ Insert Read Filter  
-rf,--read\_filter <read\_filter>

**Operate on Genomic intervals**  
1: Operate on Genomic intervals  
Genomic intervals  
3: Create single interval

**Basic or Advanced Analysis options**  
Advanced

**Minimum base quality required to consider a base for calling**  
10  
-mbq,--min\_base\_quality\_score <min\_base\_quality\_score>

**Execute**

**History** search datasets

**Variant calling NA12878**  
18 shown  
242.5 MB

18: bcftools view on data 16

17: MPileup on data 3 and data 1 (log)

16: MPileup on data 3 and data 1

15: MPileup on data 3 and data 1 (log)

14: MPileup on data 3 and data 1

13: CollectInsertSizeMetrics on data 1

12: CollectInsertSizeMetrics on data 1

11: Depth of Coverage on data 3 and data 1 (log)

1: NA12878\_chr20\_2mb\_filtered.bam

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 screenshot shows the Galaxy web interface. In the 'Tools' panel on the left, 'unifiedgeno' is selected. The main panel displays the 'INFO' column of a VCF file, with several lines of header information highlighted in red. The header information includes:
   
AC=1;AF=0.50;AN=2;BaseQRankSum=2.404;DP=28;Dels=0.00;FS=4.154;HRun=1;HaplotypeScore=233.87
   
AC=1;AF=0.50;AN=2;BaseQRankSum=0.228;DP=29;FS=0.000;HRun=12;HaplotypeScore=233.87
   
AC=1;AF=0.50;AN=2;BaseQRankSum=-0.666;DP=41;Dels=0.00;FS=2.999;HRun=0;HaplotypeScore=233.87
   
AC=1;AF=0.50;AN=2;BaseQRankSum=-1.688;DP=29;Dels=0.00;FS=5.649;HRun=1;HaplotypeScore=233.87
   
AC=1;AF=0.50;AN=2;BaseQRankSum=1.582;DP=27;Dels=0.00;FS=1.510;HRun=2;HaplotypeScore=233.87
   
The 'History' panel on the right shows a list of jobs, with '19: Unified Genotyper on data 3 and data 1 (VCF)' highlighted in red.

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 increase the 'Minimum variant allele frequency threshold' to 0.1 then click execute.

The screenshot shows the Galaxy web interface with the 'Varscan for variant detection (Galaxy Tool Version 0.1)' tool selected. The 'Tools' panel on the left shows 'NGS: Variant Analysis' and 'Varscan for variant detection' highlighted in red. The main panel displays the tool configuration:
   
Pileup dataset: 14: MPileup on data 3 and data 1 (highlighted in red)
   
Analysis type: consensus genotype (highlighted in red)
   
Minimum read depth: 8
   
Minimum supporting reads: 2
   
Minimum base quality at a position to count a read: 10 (highlighted in red)
   
Minimum variant allele frequency threshold: 0.1 (highlighted in red)
   
The 'History' panel on the right shows a list of jobs, with '21: Unified Genotyper on data 3 and data 1 (log)' highlighted in red.

Ignore variants with >90% support on one strand

yes

sample\_names

Separate sample names by comma; leave blank to use default sample names.

Execute

13: CollectInsertSizeMetrics on data 1

12: CollectInsertSizeMetrics on data 1

11: Depth of Coverage on data 3 and data 1 (log)

10: Depth of Coverage

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

Galaxy									
Analyze Data Workflow Shared Data Visualization Help User Using 8.4 GB									
Tools									
varscan									
NGS: Variant Analysis									
VarScan for variant detection									
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO		
20	61098	.	C	T	.	PASS	ADP=21;WT=0;HET=1;HOM=0;NC		
20	61795	.	G	T	.	PASS	ADP=40;WT=0;HET=1;HOM=0;NC		
20	63244	.	A	C	.	PASS	ADP=29;WT=0;HET=1;HOM=0;NC		
20	63799	.	C	T	.	PASS	ADP=26;WT=0;HET=1;HOM=0;NC		
20	65900	.	G	A	.	PASS	ADP=33;WT=0;HET=1;HOM=0;NC		

**ADP:** Average per-sample depth of bases with Phred score  $\geq 10$

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

**DP:** Quality Read Depth of bases with Phred score  $\geq 10$

**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)

**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.

Download data directly from web or upload files from your disk

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

Name	Size	Type	Genome	Settings	Status
NA12878_20_2mb_IlluminaPlatinum.vcf	0.7 MB	vcf	unspecified (?) b37 Human (Homo sapiens) (b37)		

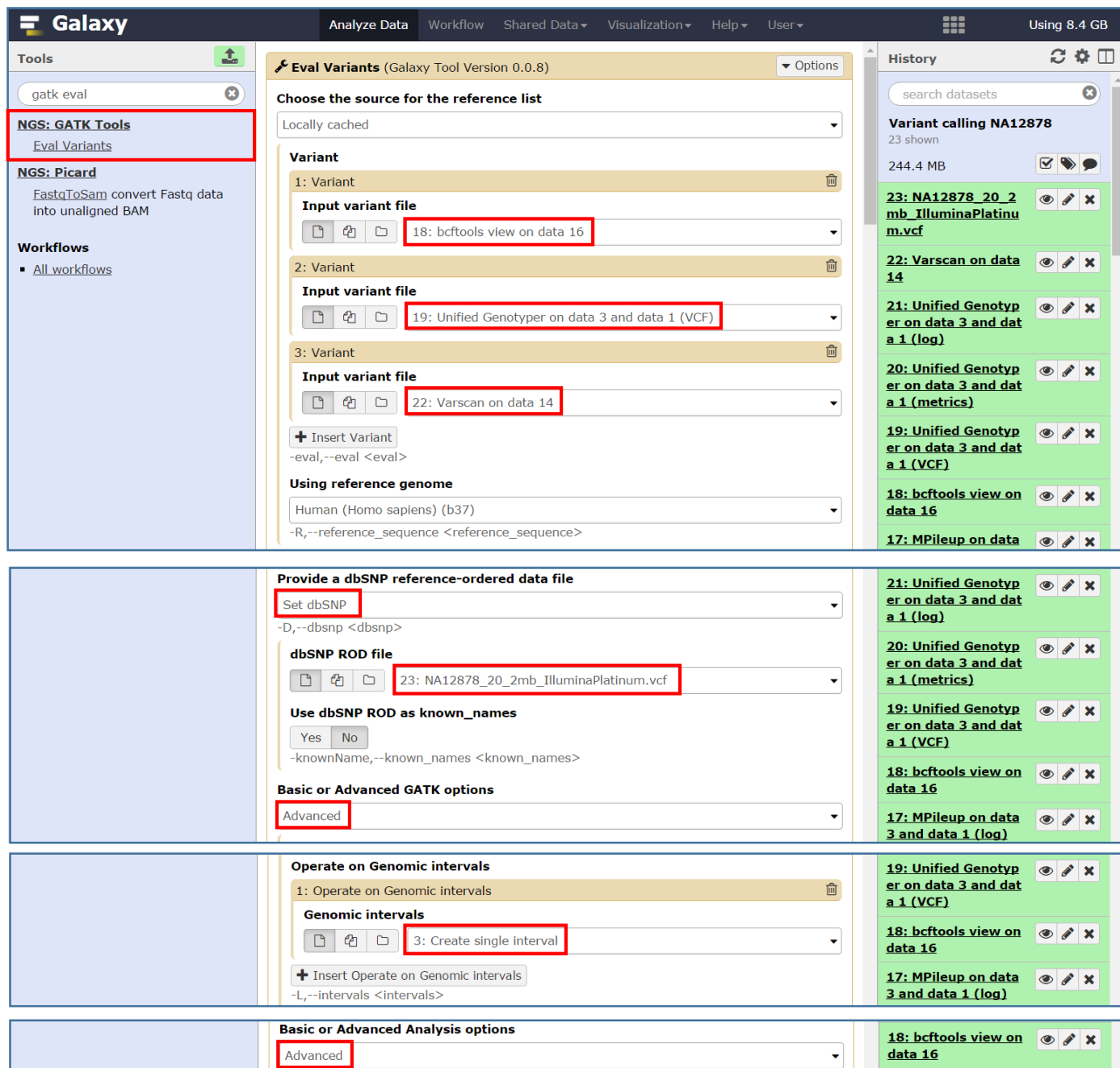
Type (set all): Auto-detect
Genome (set all): unspecified (?)

Choose local file Paste/Fetch data Start Pause Reset Close

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.



**Galaxy** Analyze Data Workflow Shared Data Visualization Help User Using 8.4 GB

**Tools** gatk eval

**NGS: GATK Tools** Eval Variants

**NGS: Picard** FastqToSam convert Fastq data into unaligned BAM

**Workflows** All workflows

**Eval Variants** (Galaxy Tool Version 0.0.8) Options

**Choose the source for the reference list** Locally cached

**Variant**

1: Variant

**Input variant file** 18: bcftools view on data 16

2: Variant

**Input variant file** 19: Unified Genotyper on data 3 and data 1 (VCF)

3: Variant

**Input variant file** 22: Varscan on data 14

**Using reference genome** Human (Homo sapiens) (b37)

**Provide a dbSNP reference-ordered data file** Set dbSNP

**dbSNP ROD file** 23: NA12878\_20\_2mb\_IlluminaPlatinum.vcf

**Use dbSNP ROD as known\_names** Yes No

**Basic or Advanced GATK options** Advanced

**Operate on Genomic intervals** 1: Operate on Genomic Intervals

**Genomic intervals** 3: Create single interval

**Basic or Advanced Analysis options** Advanced

**History** search datasets

**Variant calling NA12878** 23 shown

244.4 MB

23: NA12878\_20\_2mb\_IlluminaPlatinum.vcf

22: Varscan on data 14

21: Unified Genotyper on data 3 and data 1 (log)

20: Unified Genotyper on data 3 and data 1 (metrics)

19: Unified Genotyper on data 3 and data 1 (VCF)

18: bcftools view on data 16

17: MPileup on data 3 and data 1 (log)



**NGS: GATK Tools**  
[Eval Variants](#)  
**NGS: Picard**  
[FastqToSam](#) convert Fastq data into unaligned BAM  
**Workflows**  

- All workflows

**Eval modules to apply to the eval track(s)**  
☐ Select/Unselect all  
☐ ACTransitionTable  
☐ AlleleFrequencyComparison  
☐ AminoAcidTransition  
☒ CompOverlap  
☒ CountVariants  
☐ GenotypeConcordance  
☐ GenotypePhasingEvaluator  
☐ IndelMetricsByAC  
☐ IndelStatistics  
☐ MendelianViolationEvaluator  
☐ PrintMissingComp  
☐ PrivatePermutations  
☐ SimpleMetricsByAC  
☐ ThetaVariantEvaluator  
☒ TTVVariantEvaluator  
☐ VariantQualityScore  
 -EV,--evalModule <evalModule>  
**Do not use the standard eval modules by default**  
☒ Yes ☐ No  
 -noEV,--doNotUseAllStandardModules

**Variant calling NA12878**  
 23 shown  
 244.4 MB  
 23: NA12878\_20\_2mb\_IlluminaPlatinum.vcf  
 22: Varscan on data 14  
 21: Unified Genotyper on data 3 and data 1 (log)  
 20: Unified Genotyper on data 3 and data 1 (metrics)  
 19: Unified Genotyper on data 3 and data 1 (VCF)  
 18: bcftools view on data 16  
 17: MPileup on data 3 and data 1 (log)

☒ Execute

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

**Galaxy**  
 Analyze Data Workflow Shared Data Visualization Help User  
 Using 8.4 GB

**Tools**  
 gatk eval  
**NGS: GATK Tools**  
[Eval Variants](#)  
**NGS: Picard**  
[FastqToSam](#) convert Fastq data into unaligned BAM  
**Workflows**  

- All workflows

```

##:GATKReport.v0.2 CompOverlap : The overlap between eval and comp sites
#CompOverlap CompRod EvalRod JexlExpression Novelty nEvalVariants
#CompOverlap dbsnp input_0 none all 3621
#CompOverlap dbsnp input_0 none known 3456
#CompOverlap dbsnp input_0 none novel 165
#CompOverlap dbsnp input_1 none all 3663
#CompOverlap dbsnp input_1 none known 3501
#CompOverlap dbsnp input_1 none novel 162
#CompOverlap dbsnp input_2 none all 3913
#CompOverlap dbsnp input_2 none known 3480
#CompOverlap dbsnp input_2 none novel 433
      
```

**History**  
 24: Eval Variants on data 3, data 23, and others (report) View data  
 23: NA12878\_20\_2mb\_IlluminaPlatinum.vcf  
 22: Varscan on data 14  
 21: Unified Genotyper on data 3 and data 1 (log)

Table of column definitions from the CompOverlap report

Column	Definition
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						
GATK						
Varscan						

**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

EvalRod	Novelty	nProcessedLoci	nCalledLoci	variantRatePerBp	nSNPs	nInsertions	nDeletions	nHets	nHomVar	hetHcmRatio
input_0	all	2000000	3621	552	3173	200	203	2237	1384	1.62
input_0	known	2000000	3456	578	3111	156	153	2128	1328	1.60
input_0	novel	2000000	165	12121	62	44	50	109	56	1.95
input_1	all	2000000	3663	546	3207	222	234	2317	1346	1.72
input_1	known	2000000	3501	571	3107	201	193	2200	1301	1.69
input_1	novel	2000000	162	12345	100	21	41	117	45	2.60
input_2	all	2000000	3913	511	3281	279	353	2678	1235	2.17
input_2	known	2000000	3480	574	3044	223	213	2260	1220	1.85
input_2	novel	2000000	433	4618	237	56	140	418	15	27.87

**Slected columns from Ti/Tv Variant Evaluator**

EvalRod	Novelty	nTi	nTv	tiTvRatio	nTiInComp	nTvInComp	TiTvRatioStandard
input_0	all	2192	977	2.24	2190	978	2.24
input_0	known	2155	952	2.26	2137	929	2.30
input_0	novel	37	25	1.48	53	49	1.08
input_1	all	2216	991	2.24	2189	979	2.24
input_1	known	2152	955	2.25	2134	933	2.29
input_1	novel	64	36	1.78	55	46	1.20
input_2	all	2204	1077	2.05	2189	978	2.24
input_2	known	2115	929	2.28	2095	904	2.32
input_2	novel	89	148	0.60	94	74	1.27

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:

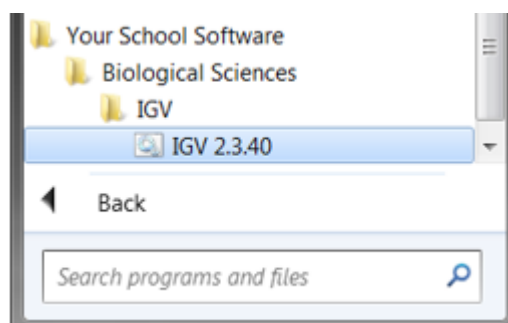
<https://www.broadinstitute.org/gatk/guide/article?id=6309>

**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.

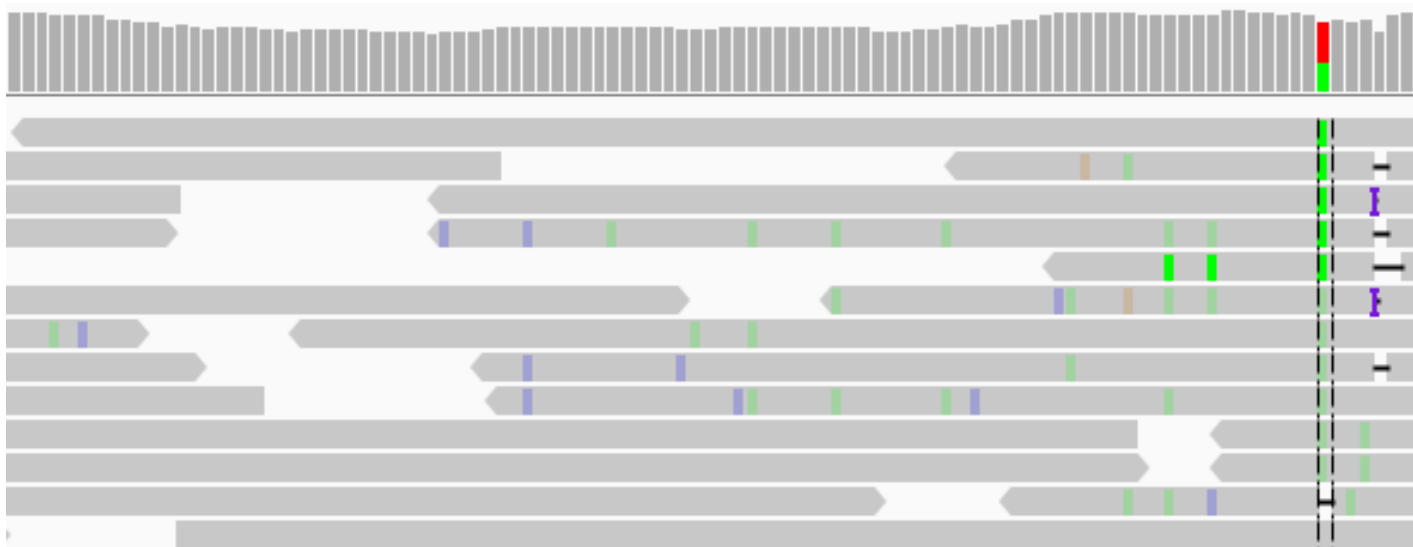
**1.** Launch IGV from Start menu > All Programs > Your School Software > Biological Sciences > IGV > IGV 2.3.40 (This will take some time <5mins as IGV has to load the whole genome, a black window will appear with messages, check this and be patient).



**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

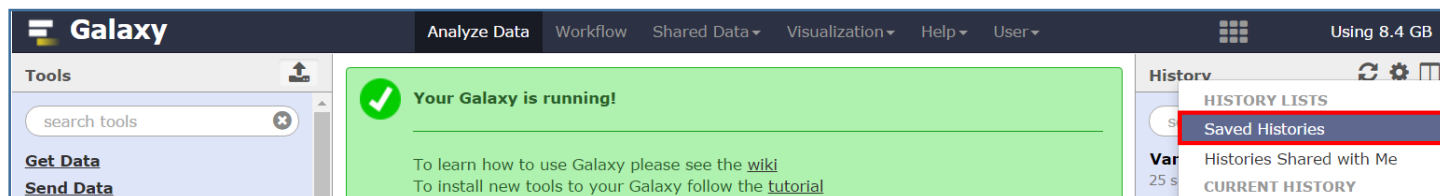
strongly suggest that the variant is an artefact and could be excluded from a tiered analysis. However, it is important to bear in mind that many unique-to-caller variants have been validated.



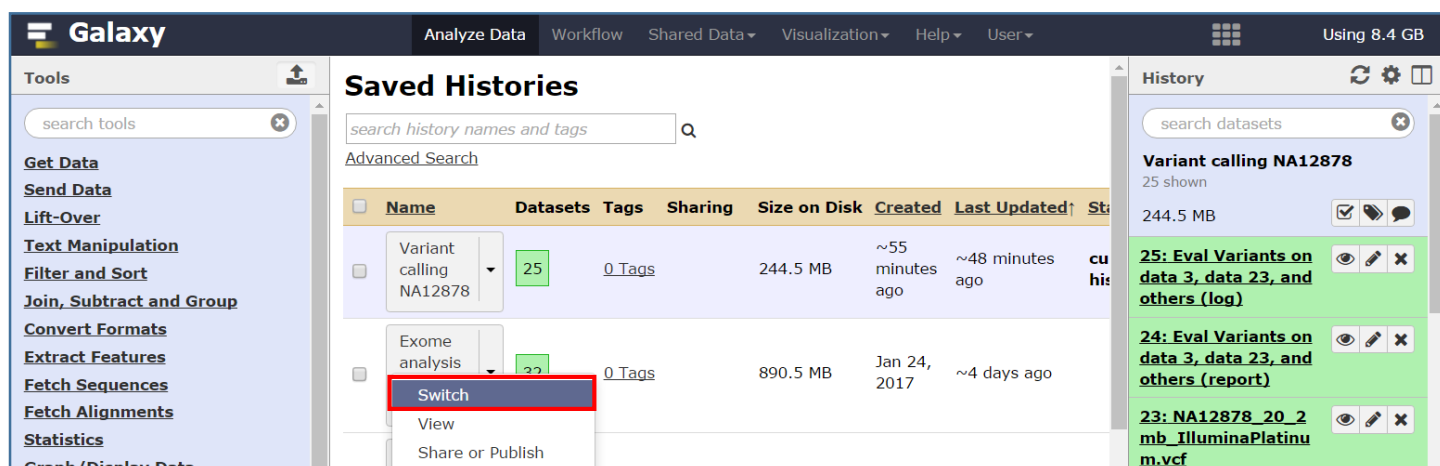
## Call variants in WES01

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**

The screenshot displays the Galaxy web interface for the Unified Genotyper tool. The interface is divided into several panes:

- Tools Pane (Left):** Shows the search results for 'unifiedgenotyper' under the 'NGS: GATK Tools' category. The tool is highlighted with a red box.
- Tool Configuration Pane (Center):**
  - Unified Genotyper SNP and indel caller (Galaxy Tool Version 0.0.6):** The tool title and version.
  - Choose the source for the reference list:** A dropdown menu set to 'Locally cached'.
  - BAM file:** A section for selecting the BAM file. A red box highlights the option '23: AddOrReplaceReadGroups on data 17: BAM with replaced/...'.
  - Using reference genome:** A dropdown menu set to 'Human (Homo sapiens) (b37)'. A red box highlights this option.
  - Basic or Advanced GATK options:** A dropdown menu set to 'Advanced'. A red box highlights this option.
  - Pedigree file:** A section for selecting the pedigree file. A red box highlights the option '24: 22\_agilent50\_targets\_hg19.bed'.
  - How strict should we be in validating the pedigree information:** A dropdown menu set to 'STRICT'.
  - Read Filter:** A section for selecting the read filter.
  - Operate on Genomic intervals:** A section for selecting the genomic intervals. A red box highlights the option '24: 22\_agilent50\_targets\_hg19.bed'.
  - Basic or Advanced Analysis options:** A dropdown menu set to 'Advanced'. A red box highlights this option.
  - Minimum base quality required to consider a base for calling:** A text input field set to '10'. A red box highlights this value.
  - Execute:** A button to run the tool. A red box highlights this button.
- History Pane (Right):** Shows a list of datasets, including 'Exome analysis of WES01' and '32: Depth of Coverage on data 24 and data 23 (log)'. A red box highlights the '32: Depth of Coverage on data 24 and data 23 (log)' dataset.

## 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.

**Download data directly from web or upload files from your disk**

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

Name	Size	Type	Genome	Settings	Status
dbsnp_132.hg19.excluding_sites_after_129_22.vcf	33.1 MB	vcf	unspecified (?) b37 Human (Homo sapiens) (b37)		0%

Type (set all): Auto-detect Genome (set all): unspecified (?)

Choose local file Paste/Fetch data Start Pause Reset Close

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

**Galaxy** Analyze Data Workflow Shared Data Visualization Help User Using 8.4 GB

**Tools** gatk eval

**NGS: GATK Tools** Eval Variants

**NGS: Picard** FastqToSam convert Fastq data into unaligned BAM

**Workflows** All workflows

**Eval Variants (Galaxy Tool Version 0.0.8)** Options

Choose the source for the reference list  
Locally cached

**Variant**  
1: Variant  
Input variant file  
33: Unified Genotyper on data 24 and data 23 (VCF)

Using reference genome  
Human (Homo sapiens) (b37)

Provide a dbSNP reference-ordered data file  
Set dbSNP  
-D,--dbsnp <dbsnp>  
dbSNP ROD file  
36: dbsnp\_132.hg19.excluding\_sites\_after\_129\_22.vcf  
Use dbSNP ROD as known\_names  
Yes No  
-knownName,--known\_names <known\_names>

Basic or Advanced GATK options  
Advanced

**History** search datasets

**Exome analysis of WES01**  
36 shown  
922.3 MB

36: dbsnp\_132.hg19.excluding\_sites\_after\_129\_22.vcf

35: Unified Genotyper on data 24 and data 23 (log)

34: Unified Genotyper on data 24 and data 23 (metrics)

34: Unified Genotyper on data 24 and data 23 (metrics)

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

951 lines, 109 comments  
format: vcf, database: hg\_g1k\_v37

Picked up \_JAVA\_OPTIONS: -Djava.io.tmpdir="/home/gxyngspp -XX: -UsePerfData

	<b>Operate on Genomic intervals</b> 1: Operate on Genomic intervals <b>Genomic intervals</b> 24: 22_agilent50_targets_hg19.bed + Insert Operate on Genomic intervals -L,--intervals <intervals>	<b>31: Depth of Coverage on data 24 and data 23 (output cumulative coverage proportions sample)</b> <b>30: Depth of Coverage on data 24 and data 23 (output cumulative coverage counts sample)</b>
	<b>Basic or Advanced Analysis options</b> Advanced	<b>18: bcftools view on data 16</b>
<b>NGS: GATK Tools</b> Eval Variants <b>NGS: Picard</b> FastqToSam convert Fastq data into unaligned BAM <b>Workflows</b> All workflows	<b>Eval modules to apply to the eval track(s)</b> Select/Unselect all <input type="checkbox"/> ACTransitionTable <input type="checkbox"/> AlleleFrequencyComparison <input type="checkbox"/> AminoAcidTransition <input checked="" type="checkbox"/> CompOverlap <input checked="" type="checkbox"/> CountVariants <input type="checkbox"/> GenotypeConcordance <input type="checkbox"/> GenotypePhasingEvaluator <input type="checkbox"/> IndelMetricsByAC <input type="checkbox"/> IndelStatistics <input type="checkbox"/> MendelianViolationEvaluator <input type="checkbox"/> PrintMissingComp <input type="checkbox"/> PrivatePermutations <input type="checkbox"/> SimpleMetricsByAC <input type="checkbox"/> ThetaVariantEvaluator <input checked="" type="checkbox"/> TiTvVariantEvaluator <input type="checkbox"/> VariantQualityScore -EV,--evalModule <evalModule> <b>Do not use the standard eval modules by default</b> Yes No -noEV,--doNotUseAllStandardModules	<b>Variant calling NA12878</b> 23 shown 244.4 MB <b>23: NA12878_20_2mb_IlluminaPlatinum.vcf</b> <b>22: Varscan on data 14</b> <b>21: Unified Genotyper on data 3 and data 1 (log)</b> <b>20: Unified Genotyper on data 3 and data 1 (metrics)</b> <b>19: Unified Genotyper on data 3 and data 1 (VCF)</b> <b>18: bcftools view on data 16</b> <b>17: MPileup on data 3 and data 1 (log)</b>
	Execute	<b>17: MPileup on data 3 and data 1 (log)</b>

Now view the Eval Variants report and answer the following questions.

**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)?

**Q11.** Comment on the amount of variation that is present in dbSNP129 and how well it agrees 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.