

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:

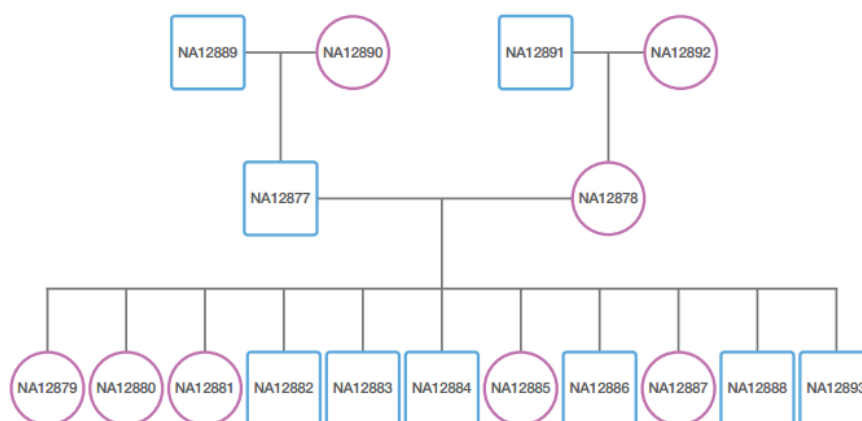
1. Use a range of software (GATK, Varscan, SAMtools, and BCFtools) to call small variants (SNVs and indels)
2. Describe the contents of pileup and variant call files
3. Generate and interpret variant quality control parameters (quality score, genotype quality, sequence context, strand bias, base quality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

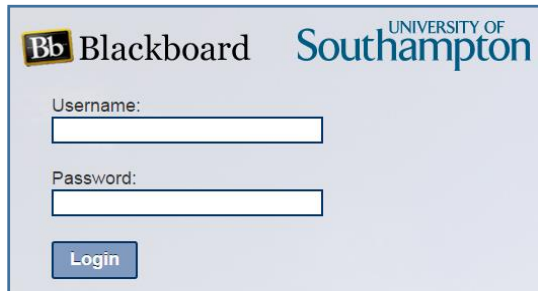
Starting from aligned WGS data for individual NA12878 (BAM file), our aim is to use a range of software to call variants and to assess the sensitivity and specificity of these programs by comparing their variant calls with the high quality variant calls from Platinum Genomes. The data we will analyse 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/>

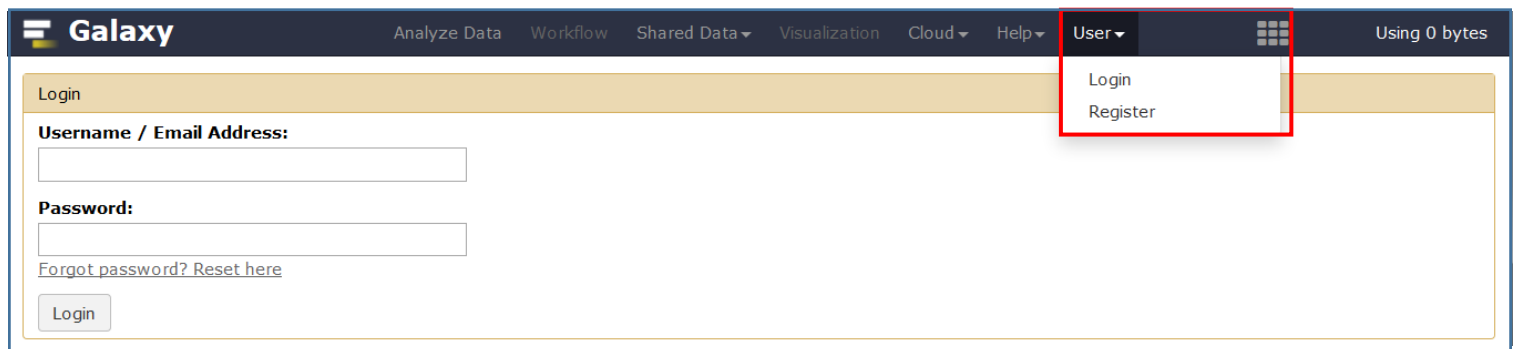


The image shows the Blackboard login page for the University of Southampton. It features the Blackboard logo and the University of Southampton name. There are two input fields: 'Username:' and 'Password:'. Below these fields is a 'Login' button.

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)

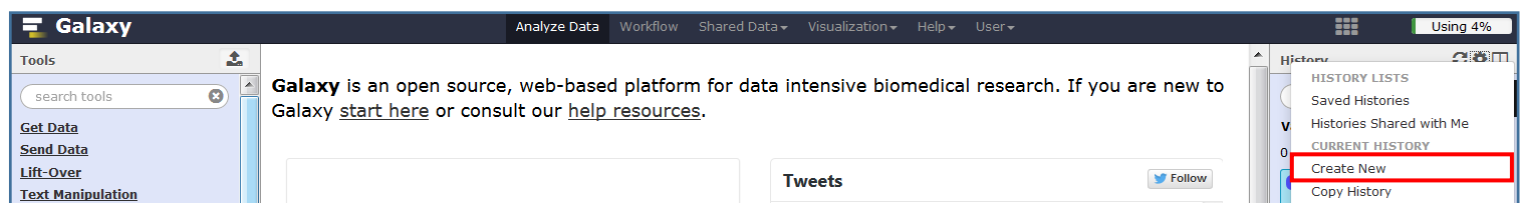
3. Go to <https://usegalaxy.org/> and login to your Galaxy account



The image shows the Galaxy login page. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. The 'User' dropdown menu is open, showing 'Login' and 'Register' options. The main content area has a 'Login' section with 'Username / Email Address:' and 'Password:' fields, a 'Forgot password? Reset here' link, and a 'Login' button.

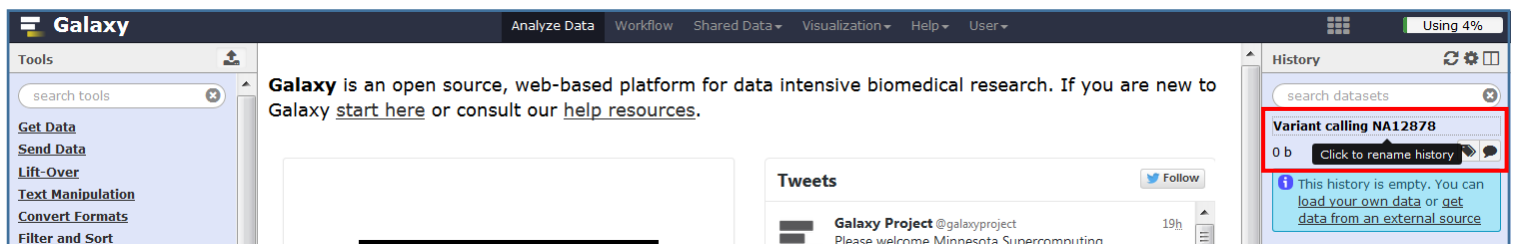
Login will take you to your last active history (Analysis of WES01).

4. Create a new history



The image shows the Galaxy main page. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'User' dropdown menu is open, showing 'Login', 'Register', and 'Create New' options. The 'Create New' option is highlighted with a red box. The main content area has a 'Tools' sidebar on the left, a central text area with a description of Galaxy, and a 'Tweets' section on the right.

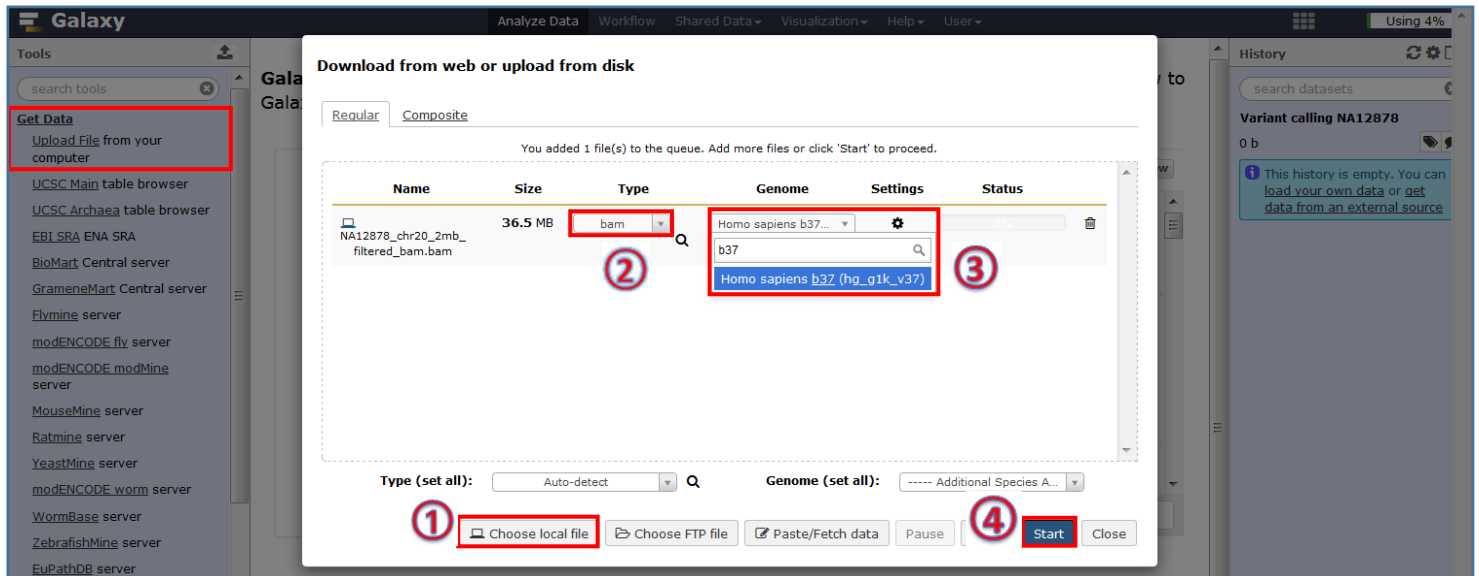
5. Rename the history for this session (Click on name, type new name, press return)



The image shows the Galaxy main page. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'User' dropdown menu is open, showing 'Login', 'Register', and 'Create New' options. The 'Create New' option is highlighted with a red box. The main content area has a 'Tools' sidebar on the left, a central text area with a description of Galaxy, and a 'Tweets' section on the right. The 'History' sidebar on the right shows a list of histories, with 'Variant calling NA12878' highlighted by a red box. A tooltip is visible over the history name, showing 'Click to rename history'.

Upload the aligned data (NA12878_chr20_2mb_filtered_bam.bam) from your computer to Galaxy

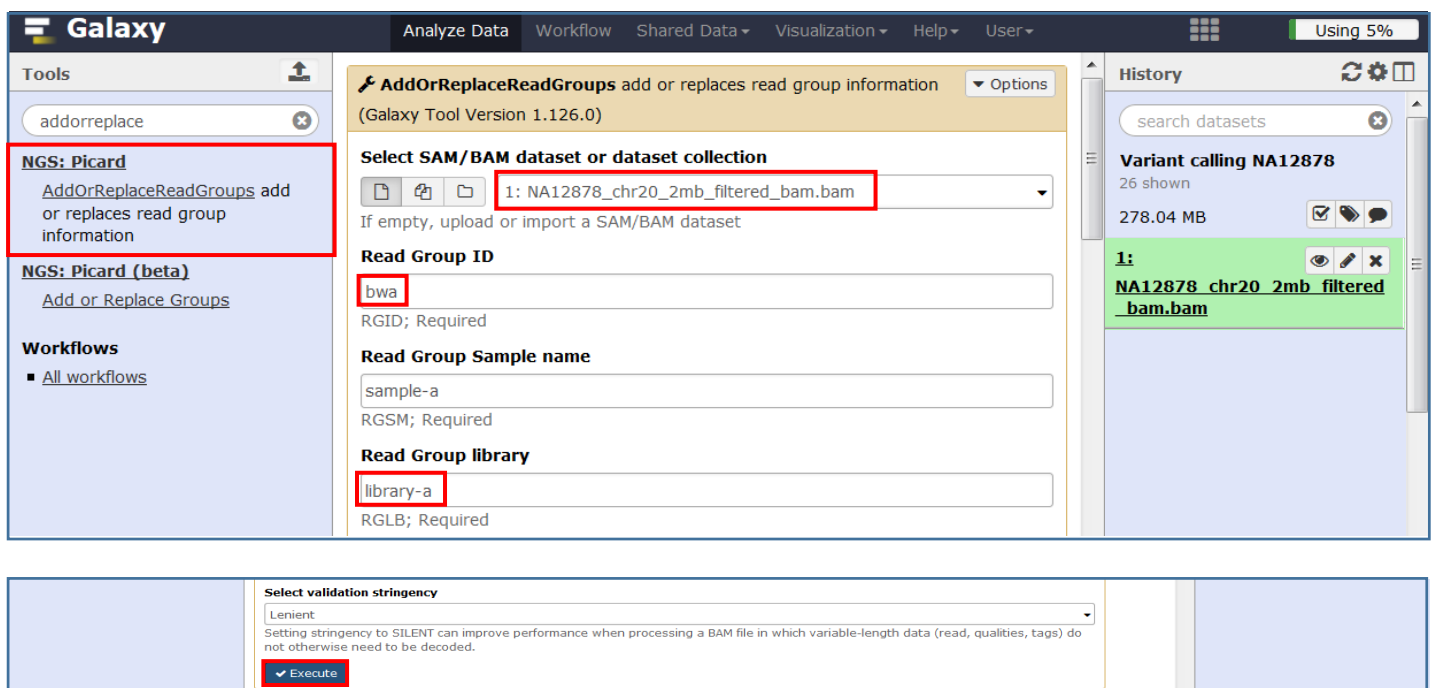
6. In Tool Pane: Go to **Get Data > Upload File** from your computer
Choose local file, select type 'bam', select genome 'b37', click start



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 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.

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

Galaxy interface showing the **IdxStats** tool configuration. The tool is selected in the **Tools** panel under **NGS: SAMtools**. The **BAM file** dropdown is set to **2: AddOrReplaceReadGroups on data 1: BAM with replaced/modified readgroups**. The **Execute** button is visible.

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

Galaxy interface showing the **IdxStats** results. The results are displayed in a table:

Chromosome	Reads	Pairs	Unmapped
14	107349540	0	0
15	102531392	0	0
16	90354753	0	0
17	81195210	0	0
18	78077248	0	0
19	59128983	0	0
20	63025520	480125	0
21	48129895	0	0
22	51304566	0	0
X	155270560	0	0
Y	59373566	0	0

The right sidebar shows the history of the tool run, with the entry **3: IdxStats on data 2** highlighted.

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

Galaxy interface showing the **Create single interval** tool configuration. The tool is selected in the **Tools** panel under **Text Manipulation**. The configuration is as follows:

- Chromosome**: 20
- Start position**: 0
- End position**: 2000000
- Name**: chr20_2mb
- Strand**: plus

The **Execute** button is visible.

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

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 4%

Tools depthof

NGS: GATK Tools (beta)
Depth of Coverage on BAM files

Workflows
All workflows

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

Choose the source for the reference list
Locally cached

BAM file
1: BAM file
BAM file
2: AddOrReplaceReadGroups on data 1: BAM with replaced/modified readgroups

Insert BAM file
-I,--input_file <input_file>

Using reference genome
Human (Homo sapiens) (b37): hg_g1k_v37
-R,--reference_sequence <reference_sequence>

RefSeq Rod
Nothing selected
-geneList,--calculateCoverageOverGenes <calculateCoverageOverGenes>

Partition type for depth of coverage
Select/Unselect all
☒ sample
☐ readgroup
☐ library
-pt,--partitionType <partitionType>

Summary coverage threshold
Insert Summary coverage threshold
-ct,--summaryCoverageThreshold <summaryCoverageThreshold>

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
4: Create single interval
Insert Operate on Genomic intervals
-L,--intervals <intervals>

Basic or Advanced Analysis options
Basic
Execute

History
search datasets

Variant calling NA12878
4 shown
73.41 MB
4: Create single interval
3: IdxStats on data 2
85 lines
format: tabular, database: hg_g1k_v37
1 2 3 4
1 249250621 0 0
2 243199373 0 0
3 190802430 0 0
4 191154276 0 0
5 100915260 0 0
6 171115067 0 0
2: AddOrReplaceReadGroups on data 1: BAM with replaced/modified readgroups
1: NA12878 chr20 2mb filtered b am.bam
36.5 MB
format: bam, database: hg_g1k_v37
uploaded bam file

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 4%

Tools depthof

NGS: GATK Tools (beta)
Depth of Coverage on BAM files

History
search datasets

6: Depth of Coverage on data 4 and data 2 (output summary sample)
View data

1	2	3	4	5	6	7
sample_id	total	mean	granular_third_quartile	granular_median	granular_first_quartile	%_bases_above_15

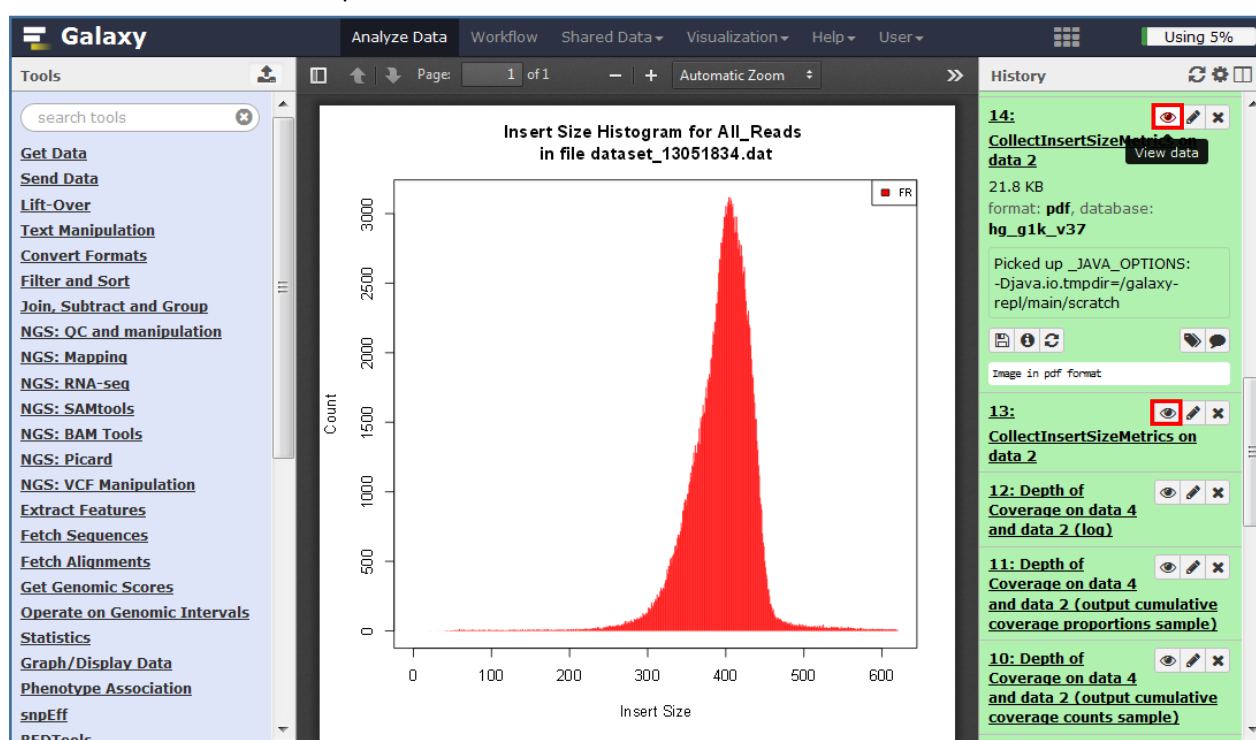
Q1. What is the mean coverage?

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**

The screenshot displays the Galaxy web interface for the MPileup tool. The left sidebar shows the 'Tools' panel with 'NGS: SAMtools' and 'MPileup call variants' highlighted. The main panel shows the tool configuration for 'MPileup SNP and indel caller (Galaxy Tool Version 0.0.1)'. Key settings include:

- Choose the source for the reference list:** 'Locally cached'.
- BAM file:** '1: BAM file' and '2: AddOrReplaceReadGroups on data 1: BAM with replaced/modified readgroups'.
- Using reference genome:** 'hg_1k_v37'.
- Genotype Likelihood Computation:** 'Do not perform genotype likelihood computation'.
- Set advanced options:** 'Advanced'.
- Extended BAQ computation:** 'Yes'.
- List of regions or sites on which to operate:** '4: 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:** (empty field).
- Output per-sample read depth:** 'Yes'.
- Output per-sample Phred-scaled strand bias P-value:** 'Yes'.
- Execute:** (checked button).

The right sidebar shows the 'History' panel with a list of datasets, including 'Variant calling NA12878' and '14: CollectInsertSizeMetrics on data 2'.

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

The screenshot displays the Galaxy web interface showing the output of the MPileup tool. The main panel shows a table of variants. The variant at 61098 bp is highlighted with a red box, indicating it is the first high confidence 'platinum' variant.

Position	Read Depth	Reference	Alternate	Quality	Phred-scaled strand bias P-value
20 61091	A 26^].	BDEE<EEF=5EAD#D#FE#;6#F#		
20 61092	C 26^].	BCDE?EECA6EDDC#D#CD#B>#C##		
20 61093	A 26^].	AFEF;EFFC@ECFE#F#ED#A8#F#		
20 61094	A 26^].	BGDE9EEF7<E=DF#9#FE#B9#F#		
20 61095	C 26^].	CCCE4EECECEBEC#E#CA#@<#C##		
20 61096	C 27^].	BEDE8EEEB(EEDE#D#EE#E>#E#/#		
20 61097	A 27^].	?EDE:EEF83EDE#C#ED#7>#F#/#		
20 61098	C 27	.tt,ttT,ttt,,t,T,tt,,t,T^].	CCF;FFFC7F@FC#B#EC#DE#C#/?		
20 61099	T 27^].	CBF?FF?AEE?FF#E#BB#C=#F#<B		
20 61100	G 27^].	@BE:EEEDADCEDDDCE?#EE#E#AB		
20 61101	G 27^].	.BEEEE=CAEDBBDCE@#EE#E#BC		

The right sidebar shows the 'History' panel with a list of datasets, including 'Variant calling NA12878' and '16: MPileup on data 4 and data 2 (log)'.

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**

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 4%

Tools mpileup

NGS: SAMtools
MPileup call variants

NGS: Variant Analysis
VarScan for variant detection

Workflows
All workflows

MPileup SNP and indel caller (Galaxy Tool Version 0.0.1) Versions Options

Choose the source for the reference list

Locally cached ☒ Switch to 0.0.1 ☐ Switch to 2.0

BAM file

1: BAM file

BAM file

2: AddOrReplaceReadGroups on data 1: BAM with replaced/modified readgroups

Insert BAM file

Using reference genome

hg_g1k_v37

Genotype Likelihood Computation

Perform genotype likelihood computation

Set advanced options

Advanced

Extended BAQ computation

☒ Yes ☐ No

List of regions or sites on which to operate

4: 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

History search datasets

Variant calling NA12878
16 shown
239.33 MB

16: MPileup on data 4 and data 2 (log)

15: MPileup on data 4 and data 2

14: CollectInsertSizeMetrics on data 2

13: CollectInsertSizeMetrics on data 2

12: Depth of Coverage on data 4 and data 2 (log)

11: Depth of Coverage on data 4 and data 2 (output cumulative coverage proportions sample)

10: Depth of Coverage on data 4 and data 2 (output cumulative coverage counts sample)

9: Depth of Coverage on data 4 and data 2 (output interval statistics sample)

8: Depth of Coverage on data 4 and data 2 (output interval summary sample)

7: Depth of Coverage on data 4 and data 2 (output statistics sample)

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.

The screenshot shows the Galaxy web interface. In the 'Tools' panel, 'bcftools' is selected. The 'Attributes' tab is active, and the 'Datatype' sub-tab is selected. A 'Change data type' dialog is open, showing 'New Type' set to 'vcf'. A 'Save' button is at the bottom. The history panel on the right shows a dataset named '19: bcftools view on data 17' with 3,621 lines, 27 comments, and a format of 'tabular'.

7. View the VCF file created by bcftools.

The screenshot shows the Galaxy web interface with the 'bcftools view' tool selected. The output is displayed in a table with columns: QUAL, FILTER, INFO, FORMAT, and sample-a. The first row is highlighted in red. The history panel on the right shows the dataset '19: bcftools view on data 17' with 3,621 lines, 27 comments, and a format of 'vcf'.

QUAL	FILTER	INFO	FORMAT	sample-a
173	.	DP=27;VDB=0.0365;AF1=0.5;AC1=1;DP4=5,4,3,9;MQ=60;FQ=135;PV4=0.2,1,1,1	GT:PL:GQ	0/1:203,0,162:99
183	.	DP=41;VDB=0.0400;AF1=0.5;AC1=1;DP4=18,8,8,6;MQ=60;FQ=186;PV4=0.5,0.5,1,1	GT:PL:GQ	0/1:213,0,255:99
210	.	DP=29;VDB=0.0305;AF1=0.5;AC1=1;DP4=8,5,6,10;MQ=60;FQ=205;PV4=0.27,0.031,1,0.00054	GT:PL:GQ	0/1:240,0,233:99
135	.	DP=27;VDB=0.0398;AF1=0.5;AC1=1;DP4=7,10,5,4;MQ=60;FQ=138;PV4=0.68,1,1,0.15	GT:PL:GQ	0/1:165,0,253:99
209	.	DP=34;VDB=0.0099;AF1=0.5;AC1=1;DP4=12,7,6,8;MQ=60;FQ=200;PV4=0.3,1,1,1	GT:PL:GQ	0/1:239,0,227:99
225	.	DP=43;VDB=0.0373;AF1=0.5;AC1=1;DP4=9,11,12,11;MQ=60;FQ=225;PV4=0.76,1,1,0.092	GT:PL:GQ	0/1:255,0,255:99
225	.	DP=30;VDB=0.0345;AF1=0.5;AC1=1;DP4=5,7,9,8;MQ=60;FQ=192;PV4=0.71,0.18,1,1	GT:PL:GQ	0/1:255,0,219:99
217	.	INDEL;DP=33;VDB=0.0365;AF1=0.5;AC1=1;DP4=10,7,6,7;MQ=60;FQ=217;PV4=0.71,1,1,1	GT:PL:GQ	0/1:255,0,255:99
222	.	DP=30;VDB=0.0359;AF1=1;AC1=2;DP4=0,0,18,9;MQ=60;FQ=108	GT:PL:GQ	1/1:255,81,0:99

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/>.

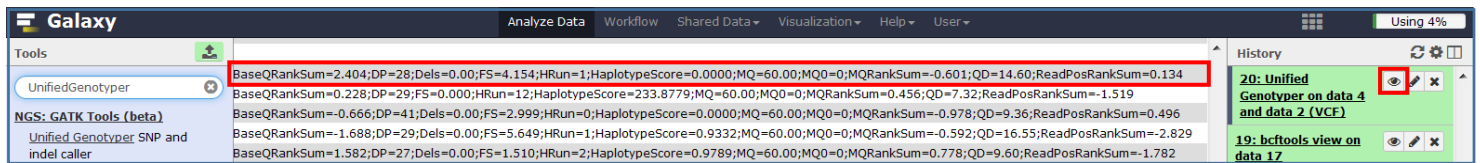
1. In **Tool Pane**: Go to **NGS: GATK Tools (beta)** > **Unified Genotyper**

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.

The screenshot displays the Galaxy web interface for the GATK Unified Genotyper tool. The interface is organized into several panels:

- Tools Panel:** Shows the 'unifiedGenotyper' tool under the 'NGS: GATK Tools (beta)' category.
- Workflows Panel:** Shows the 'Unified Genotyper SNP and indel caller' workflow.
- Configuration Panel:**
 - Basic or Advanced GATK options:** Set to 'Advanced'. Includes options for Pedigree file, Pedigree string, How strict should we be in validating the pedigree information (set to 'STRICT'), Read Filter, and Operate on Genomic intervals (set to '4: Create single interval').
 - Basic or Advanced Analysis options:** Set to 'Advanced'. Includes the 'Minimum base quality required to consider a base for calling' set to '10'.
 - Execute Button:** A red button labeled 'Execute'.
- History Panel:** Shows a list of previous runs, including 'Variant calling NA12878' and various 'MPileup' and 'bcftools view' operations.

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.

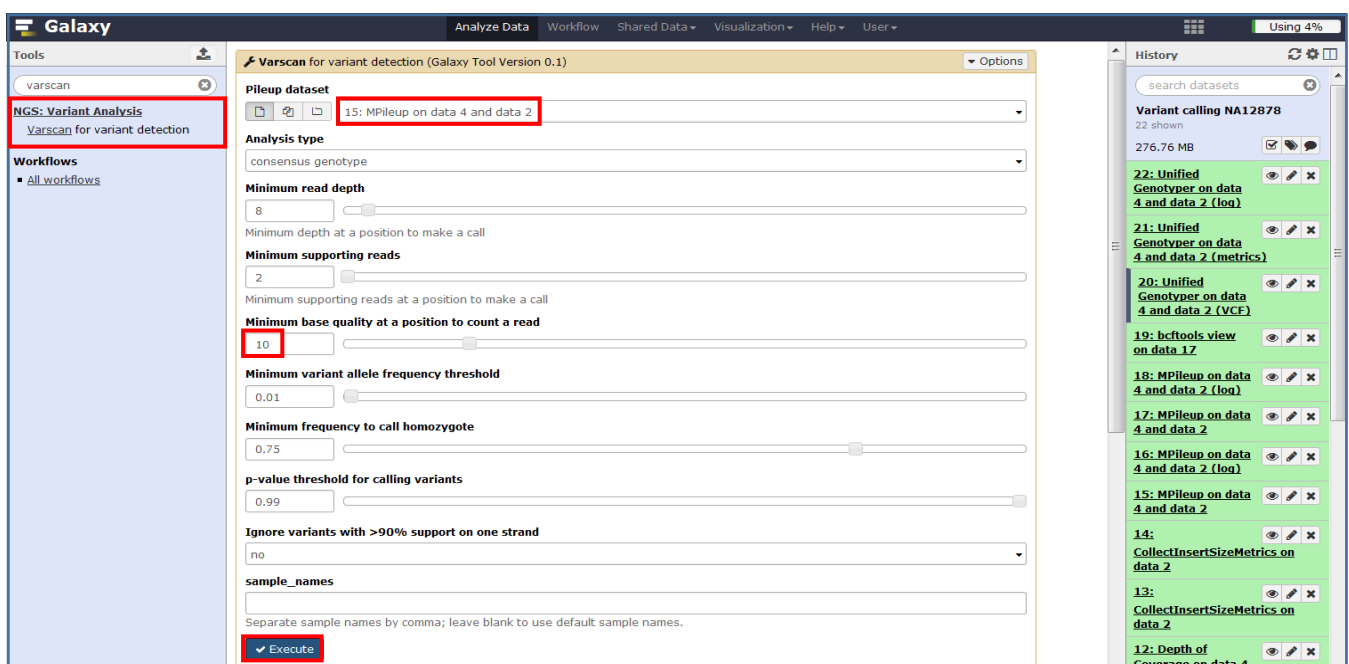
ReadPosRankSum (equivalent of tail bias): Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias

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.

Galaxy			Analyze Data	Workflow	Shared Data	Visualization	Help	User	Using 4%
Tools			INFO			FORMAT			Sample1
UnifiedGenotyper			ADP=10;WT=0;HET=0;HOM=1;NC=0			GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR			1/1:0:10:10:0:10:100%;9.8E-1:0:32:0:0:10:0
NGS: GATK Tools (beta)			ADP=21;WT=0;HET=1;HOM=0;NC=0			GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR			0/1:0:27:21:9:12:57.14%;9.8E-1:31:33:5:4:3:9
Unified Genotyper SNP and indel caller			ADP=19;WT=0;HET=1;HOM=0;NC=0			GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR			0/1:0:27:19:11:9:40.91%;9.8E-1:33:24:6:5:4:5
			ADP=40;WT=0;HET=1;HOM=0;NC=0			GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR			0/1:0:41:40:26:14:35%;9.8E-1:34:18:8:8:6
			ADP=29;WT=0;HET=1;HOM=0;NC=0			GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR			0/1:0:29:29:13:16:55.17%;9.8E-1:35:31:8:5:6:10
			ADP=26;WT=0;HET=1;HOM=0;NC=0			GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR			0/1:0:27:26:17:9:34.62%;9.8E-1:32:33:7:10:5:4

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.

The screenshot displays the Galaxy web interface for the 'Eval Variants' tool (version 0.0.8). The interface is divided into several sections:

- Tools Panel (Left):** Shows 'NGS: GATK Tools (beta)' and 'Eval Variants' under the 'Tools' section.
- Options Panel (Top Right):** Shows 'Using 4%' of memory.
- History Panel (Right):** Lists previous jobs, including 'Variant calling NA12878', '24: NA12878_20_2mb_IlluminaPlatinum.vcf', '23: Varscan on data 15', '22: Unified Genotyper on data 4 and data 2 (log)', '21: Unified Genotyper on data 4 and data 2 (metrics)', '20: Unified Genotyper on data 4 and data 2 (VCF)', and '19: bcftools view on data 17'.
- Main Configuration Area:**
 - Choose the source for the reference list:** Set to 'Locally cached'.
 - Variant Section:**
 - 1: Variant:** Input variant file is '19: bcftools view on data 17'.
 - 2: Variant:** Input variant file is '20: Unified Genotyper on data 4 and data 2 (VCF)'.
 - 3: Variant:** Input variant file is '23: Varscan on data 15'.
 - Using reference genome:** Set to 'Human (Homo sapiens) (b37): hg_g1k_v37'.
 - Provide a dbSNP reference-ordered data file:**
 - Set dbSNP:** '-D,-dbSNP <dbSNP>'.
 - dbSNP ROD file:** '24: NA12878_20_2mb_IlluminaPlatinum.vcf'.
 - Use dbSNP ROD as known_names:** 'Yes'.
 - Basic or Advanced GATK options:** Set to 'Advanced'.
 - Operate on Genomic Intervals:**
 - 1: Operate on Genomic intervals:** '4: Create single interval'.
 - Basic or Advanced Analysis options:** Set to 'Advanced'.
 - Eval modules to apply to the eval track(s):**
 - Select/Unselect all:** 'Yes'.
 - Modules:** 'CompOverlap', 'CountVariants', and 'TiTvVariantEvaluator' are selected.
 - Do not use the standard eval modules by default:** 'Yes'.
 - Execute:** A red 'Execute' button is visible at the bottom.

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

##:GATKReport.v0.2 CompOverlap : The overlap between eval and comp sites											
CompOverlap	CompRod	EvalRod	JexlExpression	Novelty	nEvalVariants	novelSites	nVariantsAtComp	compRate	nConcordant	concordantRate	
CompOverlap	dbSNP	input_0	none	all	3621	241	3380	93.34	3354	99.23	
CompOverlap	dbSNP	input_0	none	known	3453	73	3380	97.89	3354	99.23	
CompOverlap	dbSNP	input_0	none	novel	168	168	0	0.00	0	0.00	
CompOverlap	dbSNP	input_1	none	all	3656	223	3433	93.90	3429	99.88	
CompOverlap	dbSNP	input_1	none	known	3500	67	3433	98.09	3429	99.88	
CompOverlap	dbSNP	input_1	none	novel	156	156	0	0.00	0	0.00	
CompOverlap	dbSNP	input_2	none	all	4154	769	3385	81.49	3377	99.76	
CompOverlap	dbSNP	input_2	none	known	3466	81	3385	97.66	3377	99.76	
CompOverlap	dbSNP	input_2	none	novel	688	688	0	0.00	0	0.00	

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 ($n\text{VariantsAtComp}/n\text{EvalVariants}$)
 nConcordant: number of EvalRod variants with the same alleles as CompRod
 concordantRate: % of variants in both EvalRod and CompRod with the same alleles ($n\text{Concordant}/n\text{VariantsAtComp}$)

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 % = $100 * (\text{false positive} / [\text{false positive} + \text{true positive}])$

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

##:GATKReport.v0.2 CountVariants : Counts different classes of variants in the sample													
CountVariants	CompRod	EvalRod	JexlExpression	Novelty	nProcessedLoci	nCalledLoci	variantRatePerBp	nSNPs	nInsertions	nDeletions	nHets	nHomVar	hetHomRatio
CountVariants	db SNP	input_0	none	all	2000000	3621	552.00000000	3171	202	203	2237	1384	1.62
CountVariants	db SNP	input_0	none	known	2000000	3453	579.00000000	3108	156	153	2125	1328	1.60
CountVariants	db SNP	input_0	none	novel	2000000	168	11904.00000000	63	46	50	112	56	2.00
CountVariants	db SNP	input_1	none	all	2000000	3656	547.00000000	3202	221	233	2313	1343	1.72
CountVariants	db SNP	input_1	none	known	2000000	3500	571.00000000	3106	201	193	2199	1301	1.69
CountVariants	db SNP	input_1	none	novel	2000000	156	12820.00000000	96	20	40	114	42	2.71
CountVariants	db SNP	input_2	none	all	2000000	4154	481.00000000	3463	295	396	2921	1233	2.37
CountVariants	db SNP	input_2	none	known	2000000	3466	577.00000000	3029	223	214	2247	1219	1.84
CountVariants	db SNP	input_2	none	novel	2000000	688	2906.00000000	434	72	182	674	14	48.14

Selected columns from Ti/Tv Variant Evaluator

##:GATKReport.v0.2 TiTvVariantEvaluator : Ti/Tv Variant Evaluator										
TiTvVariantEvaluator	CompRod	EvalRod	JexlExpression	Novelty	nTi	nTv	tiTvRatio	nTiInComp	nTvInComp	TiTvRatioStandard
TiTvVariantEvaluator	db SNP	input_0	none	all	2190	977	2.24	2190	978	2.24
TiTvVariantEvaluator	db SNP	input_0	none	known	2152	952	2.26	2134	929	2.30
TiTvVariantEvaluator	db SNP	input_0	none	novel	38	25	1.52	56	49	1.14
TiTvVariantEvaluator	db SNP	input_1	none	all	2211	991	2.23	2189	979	2.24
TiTvVariantEvaluator	db SNP	input_1	none	known	2151	955	2.25	2133	933	2.29
TiTvVariantEvaluator	db SNP	input_1	none	novel	60	36	1.67	56	46	1.22
TiTvVariantEvaluator	db SNP	input_2	none	all	2258	1205	1.87	2189	978	2.24
TiTvVariantEvaluator	db SNP	input_2	none	known	2105	924	2.28	2091	903	2.32
TiTvVariantEvaluator	db SNP	input_2	none	novel	153	281	0.54	98	75	1.31

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:

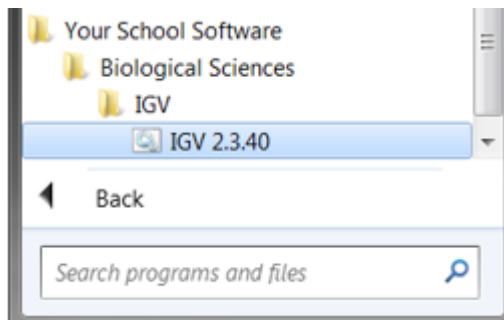
<https://www.broadinstitute.org/gatk/guide/article?id=6309>

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

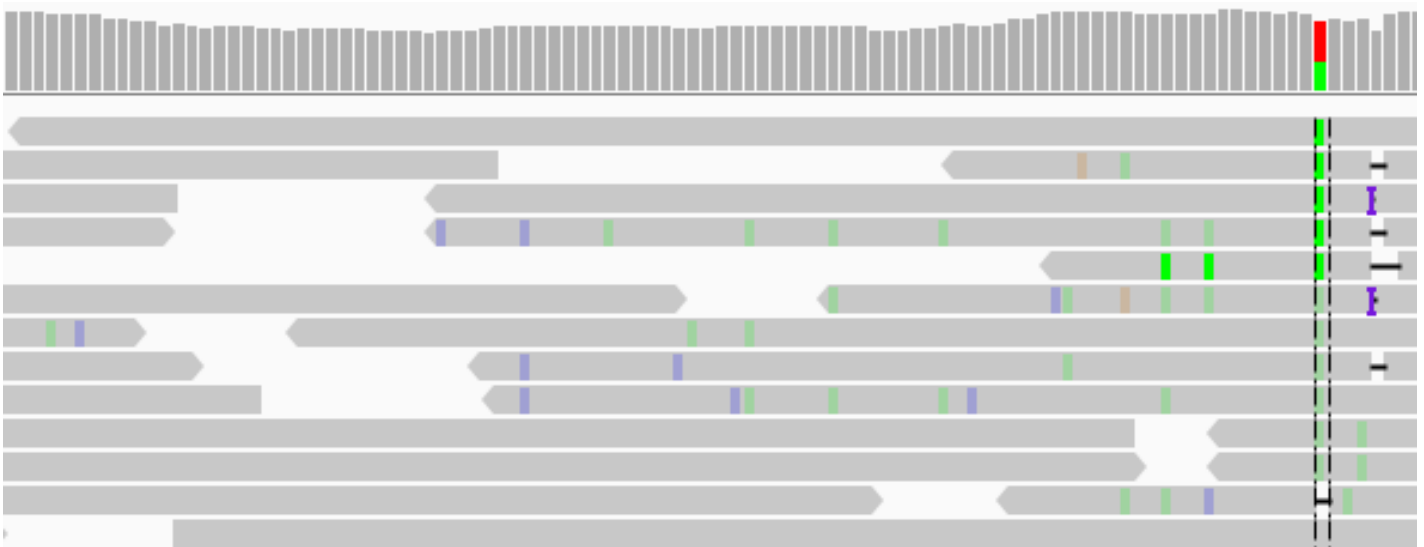
Each of the variant calling tools identifies variants that others do not, and the accuracy of these discordant variants is expected to be low. To help spot false positive variants, we will now use IGV to look at a unique-to-MPileup/bcftools variant with significant strand bias and base quality bias,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

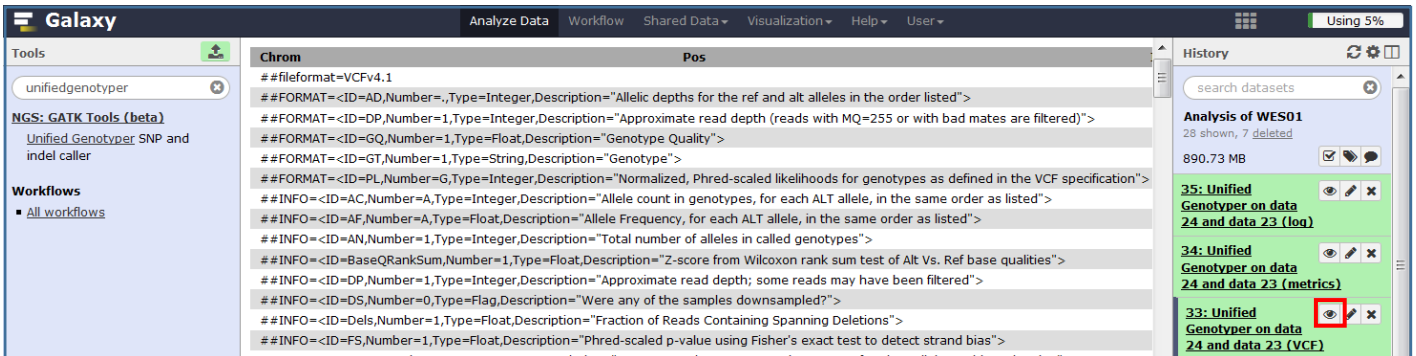
1. Click the cog icon in the history pane and select saved histories.

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.

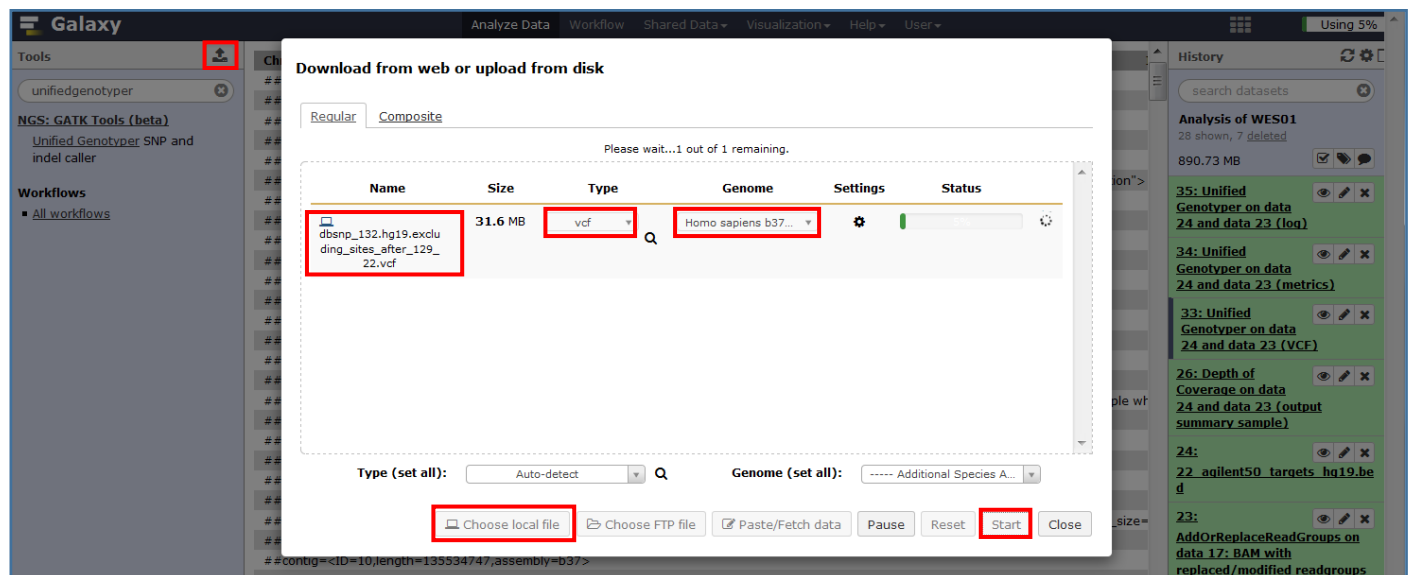


Q10. Considering that the targeted region on chromosome 22 spans 1,183,396 bp what is the rate of variation and does it come close to the prediction from Platinum Genomes for coding regions (1 variant per 1400bp)?

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**

Galaxy Analyze Data Workflow Shared Data Visualization Help User Using 5%

Tools

eval variants

NGS: GATK Tools (beta)

Eval Variants

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)

+ Insert Variant

-eval,--eval <eval>

Using reference genome

Human (Homo sapiens) (b37): hg_g1k_v37

-R,--reference_sequence <reference_sequence>

Binding for reference-ordered comparison data

+ Insert Binding for reference-ordered comparison data

-comp,--comp <comp>

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

Operate on Genomic intervals

1: Operate on Genomic intervals

Genomic intervals

24: 22_agilent50_targets_hg19.bed

+ Insert Operate on Genomic intervals

-L,--intervals <intervals>

Basic or Advanced Analysis options

Advanced

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
☒ TITvVariantEvaluator
☐ VariantQualityScore

-EV,--evalModule <evalModule>

Do not use the standard eval modules by default

Yes No

-noEV,--doNotUseAllStandardModules

History

Analysis of WES01

29 shown, 7 deleted

922.29 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)

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

26: Depth of Coverage on data 24 and data 23 (output summary sample)

24: 22_agilent50_targets_hg19.bed

23: AddOrReplaceReadGroups on data 17: BAM with replaced/modified readgroups

22: CollectInsertSizeMetrics on data 17

24: 22_agilent50_targets_hg19.bed

23: AddOrReplaceReadGroups on data 17: BAM with replaced/modified readgroups

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

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

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

26: Depth of Coverage on data 24 and data 23 (output summary sample)

24: 22_agilent50_targets_hg19.bed

23: AddOrReplaceReadGroups on data 17: BAM with replaced/modified readgroups

22: CollectInsertSizeMetrics on data 17

Execute

3. View the Eval Variants report

Galaxy interface showing the Eval Variants report. The main panel displays two tables of variant statistics.

##:GATKReport.v0.2 CompOverlap : The overlap between eval and comp sites

CompOverlap	CompRod	EvalRod	JexlExpression	Novelty	nEvalVariants	novelSites	nVariantsAtComp	compRate	nConcordant	concordantRate
CompOverlap	dbsnp	input_0	none	all	951	166	785	82.54	784	99.87
CompOverlap	dbsnp	input_0	none	known	785	0	785	100.00	784	99.87
CompOverlap	dbsnp	input_0	none	novel	166	166	0	0.00	0	0.00

##:GATKReport.v0.2 TiTvVariantEvaluator : Ti/Tv Variant Evaluator

TiTvVariantEvaluator	CompRod	EvalRod	JexlExpression	Novelty	nTi	nTv	tiTvRatio	nTiInComp	nTvInComp	TiTvRatioStandard	nTiDerived
TiTvVariantEvaluator	dbsnp	input_0	none	all	643	267	2.41	4406	1852	2.38	0
TiTvVariantEvaluator	dbsnp	input_0	none	Known	537	220	2.44	521	210	2.48	0
TiTvVariantEvaluator	dbsnp	input_0	none	novel	106	47	2.26	3885	1642	2.37	0

The right sidebar shows a history of analyses, with '37: Eval Variants on data 24, data 36, and data 33 (report)' highlighted.

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.